THORLABS

Optical Microscopy Course

Neil A. Switz ● Daniel A. Fletcher

Based on the Course at

UC Berkeley

Course Notes

Optical Microscopy Course

Course Notes Preface

Inexpensive image sensors (in webcams, and now cellphones) have changed the possibilities for optics labs. Suddenly the real power and elegance of optical imaging has become accessible at a lower level of cost and higher degree of complexity, allowing for vastly more exciting and useful undergraduate imaging labs. Furthermore, it is now commonplace to build custom microscopes for research projects – especially those involving fluorescence microscopy, a technique at the vanguard of the current revolution in biology and biomedicine – using the exact techniques taught in this course.

We have built many such systems in both industry and academia, and this is the class we wished we had had as students. A relevant vignette: one of us, as a junior graduate student in a lab internationally known for its microscopy research, asked, "what is NA?" and was told only that "it's n sin(θ)." Nobody seemed to have an intuitive feel for what it really was, or why it influenced resolution; at least not one they could (or would) convey to a new graduate student.

This course is the opposite of that: there is a strong emphasis on building physical intuition, and on the simple rules of thumb that practicing scientists and engineers use to understand optical systems. The use of actual cameras and lenses allow you to see (literally!) what is being discussed. Things that once took hours of drawing tedious diagrams to understand now become obvious in seconds as you look firsthand at where the light actually goes, or how the image changes when an iris is adjusted. Learning optics this way is also a lot more fun – a regular favorite is the moment when people first set up imaging with their cameras (during the second lab session): even when three groups have just done the same thing, the fourth still finds it exciting, and it never becomes boring for the instructors either.

We keep the mathematical level at extremely simple algebra, and purposefully so: practicing engineers and scientists do not perform a bunch of integrals, or draw complex ray diagrams, when roughing out a system design – they use easy rules of thumb to get a quick idea of what will happen. These rules usually contain all the physical intuition, and provide (for less than 20% of the effort) 80% or more of the benefit one would achieve through a longer calculation. We acknowledge here a pedagogical debt to Paul Horowitz and Winfried Hill who beautifully demonstrated a similar approach in their (vastly more comprehensive) *Art of Electronics*, which inspired a generation of experimentalists. There is no question that optics can benefit from the same methodology: undergraduates make up most of our typical class, while the graduate students also find the material useful and engaging. In fact, the class works wonderfully – and without a change in the level of the material – for graduate students and postdocs in the physical sciences. Those with additional training readily see more deeply into the basis and extensions of the concepts, but nonetheless appreciate the practical nature of the presentation.

Pedagogically, the course starts with lenses, and moves quickly (in Lab 2) to actual imaging with real cameras. Labs 3 and 4 introduce resolution, aberrations and the importance of illumination, setting the stage for constructing a complete microscope (with Köhler illumination) in Lab 5. The following three labs, 6 through 8, are the theoretical heart of the course and make use of a second camera imaging the objective back focal plane to introduce the Abbe theory of image formation in substantial detail, including deeper investigations of resolution and contrast, covering the MTF, darkfield and phase contrast imaging. The final two labs, 9 and 10, involve fluorescence imaging and include a quantitative introduction to filter selection using Excel – a critical skill for those working with fluorescence, and simple enough that there is no reason not to include it at this level.

The course is designed around two-person groups, with each week consisting of one 3-hour lab session, one 90-minute lecture, and one lab write-up. The 10 labs should fit well in a quarter-long course; in our semester-long version, the last several weeks are reserved for independent student projects using the equipment. Some project examples include polarization and Rheinberg microscopy, and additional explorations of fluorescence, aberrations, and the PSF. We have also successfully compressed the material into a 2-week "bootcamp" for advanced students, running one lecture and one lab per day.

When we teach, lab groups keep the same experimental rig for the entire course so that they benefit from using extra care in alignment and start each lab from exactly where they left off. It is difficult to have multiple groups using the same rig on different days, since switching between users inevitably results in substantial time lost to realignments at the start of each lab. Separately, while cost considerations make it desirable to have three students per rig, our experience is that this generally results in one student not really learning how to handle the equipment. A student instructor is required to support the lab sections.

As anyone who has put together a lab course will know, getting the equipment together and maintaining it over multiple years provide some of the largest challenges. Sourcing from a single company makes both of these more tractable, with one-stop ordering of known-compatible parts, and the security of continued availability for replacements.

With this in mind we approached Thorlabs with the hope of creating a kit – in which we have no financial interest – to simplify dissemination of the course to other institutions. However, Thorlabs' engagement with us has far exceeded that simple idea. Their efforts to accommodate our instructional requirements and to modify and produce new hardware to improve the course experience have gone well beyond the commercially justifiable and constitute a real service to the optics community.

As a final note, we have been fortunate to have had a supportive and collaborative environment at the University of California, Berkeley in which to develop this course. In an era where the national investment in education is being steadily challenged and reduced, we hope that whatever utility these materials provide will serve as a reminder of the public service such universities deliver.

Neil Switz and Dan Fletcher

Berkeley, CA

May 2019

Course Notes Table of Contents

THORLABS

Lab 1 Notes: Introduction to Optical Imaging (I)

Optical Microscopy
Course

Course Goals

Optical microscopy is extremely powerful: take, for example, the vast number of university labs in all science departments that have microscopes, many of which cost more than $500k. There are labs performing sum-frequency microscopy using femtosecond laser pulses, labs achieving ~3 nanometer resolution using fluorescence (far below the "Rayleigh resolution limit"), and biologists who use micron-scale 3D optical reconstruction on a regular basis. Optical imaging is a hot field and there are new, sophisticated techniques being developed regularly.

The idea behind this course is to introduce the main concepts that optical engineers and microscopy experts use to design, specify, and build their own systems, and to give you experience building them yourself. Many of the optical concepts that can be so hard to understand on paper come alive when you can simply ***see*** them, and watch what happens as you adjust the system.

We have several major goals for this course.

To introduce you to:

1) Optical prototyping using standard "optical breadboarding" parts, so that you feel comfortable building a (simple) custom optical system if you need to do something unusual.

2) The Abbe theory of image formation and, more generally, to give you an intuitive feel for the manipulation of "spatial frequencies" in an optical system. For those of you interested in pursuing optics further, this will give you a good base to understand optics from the unbelievably elegant and powerful Fourier perspective. In this course, we approach this from a conceptual and experimental standpoint, rather than a heavily mathematical one.

3) Fluorescence imaging, especially the details of filter choice. It is often easy to buy $500 worth of optical filters and double the sensitivity of a $250k microscope if you know what you are doing. Few people do, but you could soon be one of them!

4) Optical rules of thumb. It is rare for anyone to perform a complicated calculation while building a basic optical system. Rather, if one knows the basics and a few rules of thumb used by the pros, one can quickly get an idea of what should work. We want to introduce you to, and give you practice with, those simple formulae so that, by the end of the course, you have developed the confidence and skill set to try things on your own.

This is not a mathematically difficult course – the homework is not nasty, and only once do we use any math beyond algebra (a simple integral). However, if you do not keep up on the reading, and especially **if you have not read over the lab procedure and the lab objectives in advance, you will likely get lost and run out of time**. This is NOT a course where you can sleep in class, then cram for the final.

We will help you stay on top of things by providing an **in-class quiz on the upcoming lab instructions**. If you have read them carefully, you will be fine – there will be no trick questions. The quiz is solely to make sure you read the lab instructions ahead of time.

Lab 1 Course Notes:
Introduction to Optical Imaging (I)

Overview

We will introduce you to the basics of digital cameras, optical spectra and spectrometers, lenses and the measurement of focal length, as well as some of the optomechanical hardware included in this course. These topics will underlie the rest of the semester, and we will go into all of them in much more depth over time.

Optical imaging is one of those things that can seem so obvious – after all, your eyes come pre-installed – that the miracle that imaging *actually works* gets lost. Since this is a practical optics class, let's get started with experiments right away: hold a piece of paper near your computer screen. The paper will interrupt the path of the light from the screen and deflect it into many directions, just as a movie screen does in the theatre. If you focus your eyes on the movie screen, you see the movie; if you focus on the paper in front of your screen, you will see what the light from the screen **really** looks like at that plane. When you look directly at the screen, you see this text because your eyes have lenses that automatically adjust to focus that light into an image for you. If all you had was your retinas, without the attached lenses, you would still see the light from everything around you but it would look like what you see on the paper. What lenses do is actually quite remarkable.

Try it!

- Hold a piece of white paper near your computer screen, and look at the side of it closest to the screen. What you see on the paper is what the light actually looks like that far from the screen.
- Try expanding the text, or resizing the window (especially making a small light window against a dark desktop, or vice-versa) to see if it makes any difference.
- What sized objects can you start to see traces of in the light that hits the paper? Does it matter how far the paper is from the screen?

Of course, our eyes have a limited range of things they can help us see, and the advent of devices which can help us see beyond those limits has ushered in scientific breakthroughs for centuries: Galileo used a telescope to discover the planets and their moons, shaking the cosmology of his time; van Leeuwenhoek invented a high-resolution microscope (in the 1670s) and first saw cells and single-celled organisms, a discovery that provided a basis for modern biology and medicine. This is also research at its most practical: in discovering bacteria ("germs"), van Leeuwenhoek's work laid the groundwork for the germ theory of disease, and microscopy, in a direct path leading through such figures as Pasteur, has become a central medical diagnostic used in virtually every hospital in the world. This progress has not slowed; fluorescence imaging techniques currently serve as central tools for the revolution in biotechnology, underlying most gene sequencing, as well as the use of the genetically expressible Green Fluorescent Protein (subject of the 2008 Nobel Prize in chemistry).

Because both telescopes and microscopes usually magnify images for us, many people think of optical systems, especially microscopes, in terms of their magnification – with more being considered better. Now is a good time to debunk this misconception: more magnification is not necessarily better. To see this, simply blow this document up on your computer screen – as the letters get bigger and bigger, does your view become better? Probably not, unless you need glasses. In fact, the bigger the text, the fewer words you can see on the screen at one time. In optics, the number of words you can see is known as the "field of view," and naturally the area of an object that you can see gets smaller as you blow the image up more and more. So if more magnification costs you field of view; what does it give you? Only convenience – it makes sense to magnify an image to the point where it can be adequately assessed by the detector, whether that is your eye or a camera sensor.

But wait: more magnification should allow us to see smaller things, right? Sometimes. Try this: look at Figure 1 and blow it up on your screen while looking at the small bars in the target. At some point, you will not be able to see finer features, even though you are magnifying the image more and more. This phenomenon is known as "empty magnification" because you are gaining nothing as you blow the image up further. Of course, that very term should suggest that sometimes magnification is not "empty." Magnification is good (called "useful magnification") as long as it is helping your detector (possibly your eye) to see more of the details *already present* in the image.

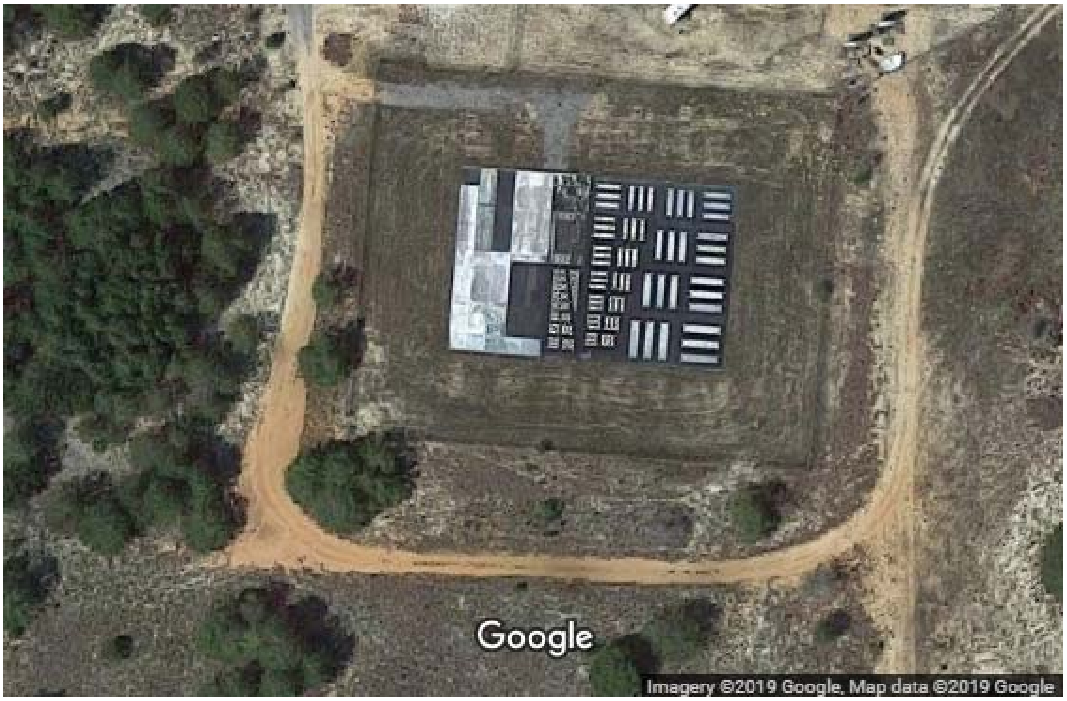


Figure 1: An eye chart for satellites: this huge (50 meter) resolution target, used for testing satellite and aircraft surveillance cameras, is laid out on the ground at Eglin Air Force Base in Florida. If you blow the image up (magnify it) on your computer, can you tell the bars apart in the smallest 3-bar sets on the left in the image? Is more *magnification* helping, or do you really need better *resolution*? (Image Source: https://goo.gl/maps/4L6hPbFamM72 - also available under the *Reference Links* tab at www.thorlabs.com/OMC)

What determines the ultimate level of detail present in an image? The resolution! Of course, we have not explained **why** the resolution is what it is; we will cover that in various ways in this course. However, a surprising number of medical doctors and professors will confuse high magnification with good resolution.

Those who work with images a lot (including good microscopists) know that good resolution is not very useful alone; one also needs ***contrast***. Generating contrast is so important that it has been directly and indirectly the subject of Nobel Prizes – physics in 1956, for phase contrast, and chemistry in 2008, for the aforementioned green fluorescent protein. Contrast is simply the relative difference between light and dark in an image. Contrast is especially important in biological microscopy because cells are small bags of water, and water is (generally) clear – which is to say, it has low contrast. See Figure 2:

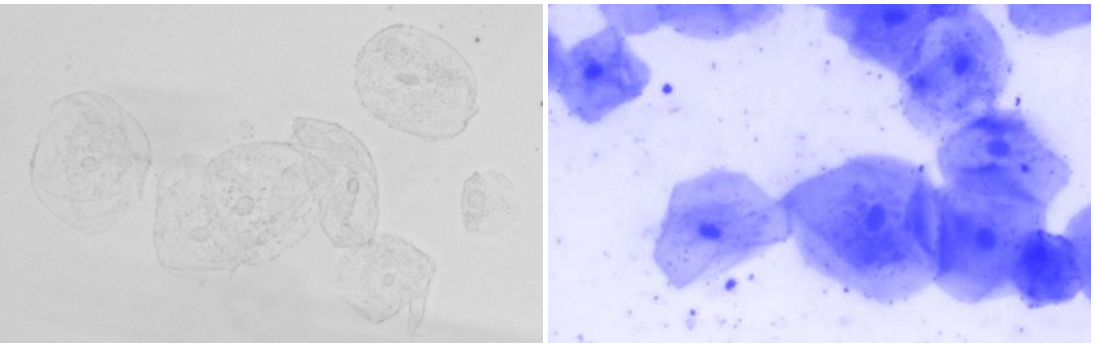


Figure 2: Human cheek epithelial cells imaged using brightfield with unstained cells (left) and stained cells (right). In both images, the resolution and magnification are similar. Note that the unstained (clear) cells are hard to see, whereas the cell stained with a blue dye has much clearer features. The blue dye has served to increase the contrast. (Photos Courtesy of Neil Switz.)

Organic chemistry, in the form of the industrial dye industry, was developing at about the same time (1800's) that microscopy and biology were making huge strides. It would not be too much to say that the industrial dyes (e.g. methylene blue) were the green fluorescent proteins of their day – they made previously unknown features of cells visible to the microscopist's eye. The importance of this discovery is still present in the biological and medical nomenclature. Take eosinophils, for example, blood cells named for eosin, the dye used to stain them, and the important classes of Gram-positive and Gram-negative bacteria, which either stain or remain transparent with the application of Gram's stain.

The problem with organic dyes is that the staining process usually kills the cells. Often one wants to see how cells are behaving while they are still alive – a good example might be if one wants to watch the process of cell division, or to observe an immune cell attacking and eating (phagocytosing) a foreign cell. How can one watch clear cells in water? It can be tricky – hence the Nobel Prizes – but you will learn three of the techniques during the course. The first two, darkfield contrast (usually just called "darkfield") and phase contrast (often just called "phase") are shown in Figure 3:

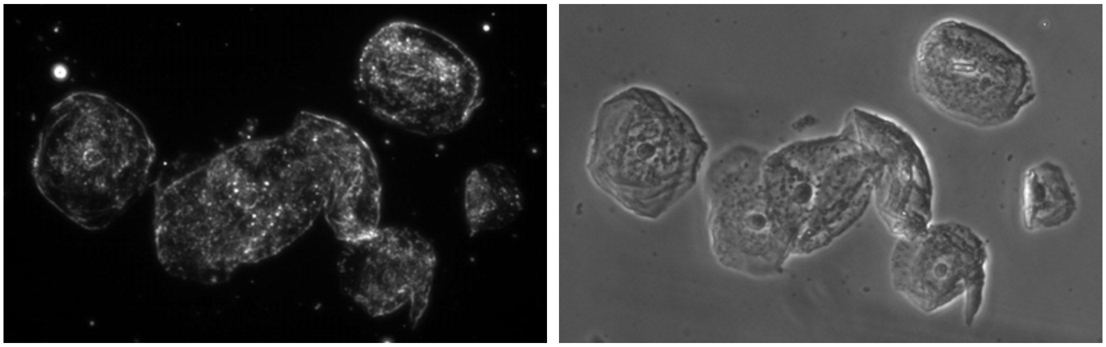


Figure 3: Human cheek epithelial cells imaged using darkfield (left) and phase contrast (right). Magnification and resolution are similar in both images. (Photos Courtesy of Neil Switz.)

Brightfield and phase contrast are used regularly in biology (and daily in hospitals), and staining with chemical dyes is still used on a massive scale for medical diagnosis. The cutting edge of biological imaging now typically involves the use of fluorescent dyes to obtain contrast. The most notable of these dyes is the green fluorescent protein, which can be genetically expressed to label individual cellular proteins and structures, giving exquisite insight into cellular processes. An example of this is shown in Figure 4, where a living cell is imaged using three dyes, each attached to a different component of the cell. You will learn about fluorescence imaging and configure your microscopes to do it in the final labs of this course.

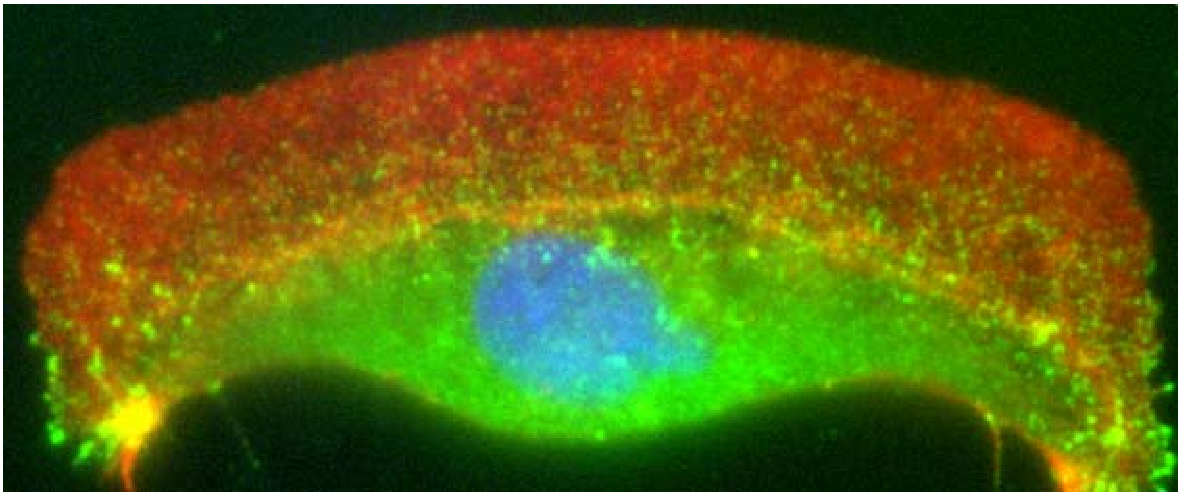


Figure 4: A crawling fish keratocyte, imaged using fluorescence microscopy. Here the different parts of the cell are labeled with different fluorescent dyes: the DNA in the nucleus in blue, the F-actin molecules in the crawling edge ("lamellipodium") in red, and the protein vinculin in the body in green. (Photo Courtesy of Daniel A. Fletcher and Martjin van Duijn.)

Note that in each of the figures above (2 – 4) different features of the cells are noticeable. Depending on what you are interested in observing, different contrast techniques will be most useful.

The intent of this course is to give you familiarity with microscopy (especially microscopy using digital cameras as detectors) such that you can get the resolution you need in an image, with sufficient contrast to see the features, using only the magnification necessary to make use of your resolution without sacrificing your field of view (the amount of the sample you can see) due to "empty" magnification.

Along the way, you will also learn to build a microscope using standard optomechanical research tools and a number of tricks so that you get the most out of even more sophisticated systems. It should be fun!

As a final note, this course only assumes knowledge of very basic algebra and trigonometry (though there is a single integral late in the course, which we do for you). It is assumed that you have had the optics portion of a basic physics course, so that you are familiar with concepts such as light being a wave, interference, and how the wavelength of light is related to its color. Regardless of whether you have had this material, read over the following "Background Notes on Light and Lenses" in order to refresh your knowledge. You do not need to memorize this material, but do need to be familiar with the basic ideas. Information you are required to know will be in the standard weekly Course Notes and Lab Notes, so concentrate your efforts on those.

Review Questions:

1. Why does increasing the magnification of an image decrease your field of view?
2. What (roughly) is resolution?
3. When is it no longer helpful to magnify an image more?
4. What is contrast?

Introduction to Light and Lenses

Comment on Lab 1 Course Notes: This material is designed to be helpful, rather than as additional required material. There will be parts of this section that will show up in the required material, and some may be found on the "Equations to Memorize" sheet, but in general the material below simply provides useful background for those without a prior optics course.

You do not need to memorize anything, or do any problems, listed in this document (unless, of course, it shows up elsewhere, like the "Equations to Memorize" sheet).

These Course Notes are intended to be helpful; if you find them confusing, a) keep reading so you see what is there, and b) do not worry – there is relatively little math (other than simple algebra) in the course, and many things are much easier to understand once you start seeing them in practice, which is why this is a lab course!

Light is a Wave

Physical Picture

To start at the beginning, light is a wave. This has profound consequences for optics, so it is worth covering some of the basics of waves. Though the analogy is not perfect, water surface waves are much easier to visualize, so we will use them as an example.

Waves have a number of properties: amplitude, phase, speed, frequency, wavelength, and direction. Not all of these are independent, as we will see.

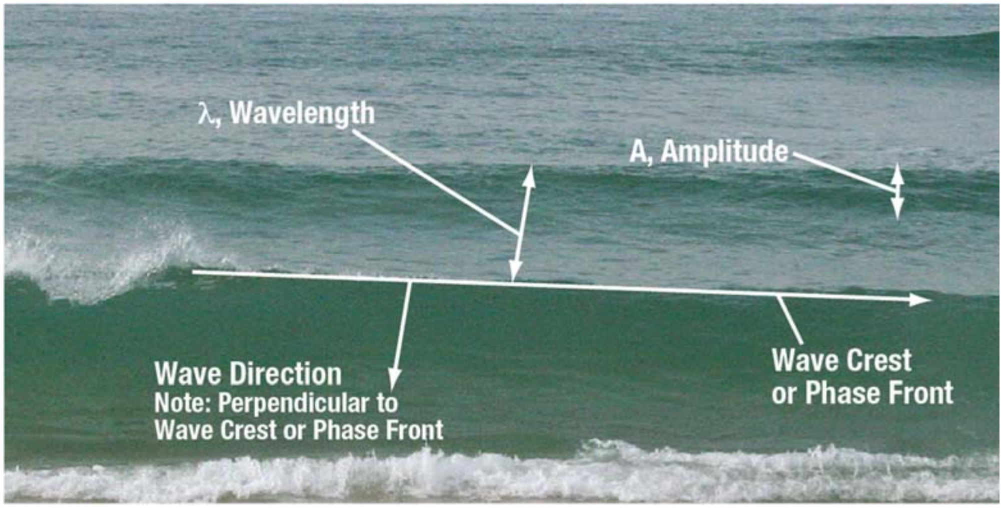

λ, Wavelength
A, Amplitude
Wave Direction
Note: Perpendicular to Wave Crest or Phase Front
Wave Crest or Phase Front

Figure 5: Waves at a beach. The crest of each wave is a “phase front.” These waves are not sinusoidal, nor are the phase fronts completely flat (i.e., the waves are not “plane waves”), but they are illustrative. Image Source: https://commons.wikimedia.org/wiki/File:Gentle_waves_come_in_at_a_sandy_beach.JPG. Image is licensed under the GNU Free Documentation License: https://www.gnu.org/licenses/fdl.html, Note: This image was adapted; text and arrows were added to original image.)

Amplitude is easiest to understand. It is just the height of the wave, measured from the midpoint. For a water wave, for instance, the midpoint is just normal sea level.

Phase is trickier to understand. The easiest way to think about phase is to imagine a surfer surfing on the crest of a wave – the surfer is moving on a “point of constant phase” on the wave. That point moves forward at the speed of the wave[1]. Imagine a second surfer who started on the same wave at the same time, but whose friend is secretly holding onto the back of their surfboard, being dragged through the water such that the surfer is slowed down. Even though they are still going forward compared to the first surfer, this surfer will slowly get farther behind, as s/he cannot keep up with the wave crest and slides down the back side of the wave, then up the next crest, etc. The first surfer will get to the beach sooner than the second. We say that surfer #1 became “phase advanced” compared to the slower one, or equivalently that surfer #2 became “phase delayed” compared to the faster one.

Velocity is simply how fast the surfer would be moving forward, which is to say the speed of the wave crest. Of course, we need not actually be going along the crest – if we moved along with the trough of the wave we would still be going the same speed. We are really talking about traveling with some point on the wave which always has the same height, which is what “constant phase” means.

Frequency. If we stand in one place in the water and use our watch to time how long it takes for two successive wave crests to pass over us, we will have the “period” of the wave, usually denoted as T. The frequency of a wave is simply the number of wave crests that pass over us in a given amount of time, so $f = \frac{1}{T}$. The frequency times the time (i.e., f * t) is equal to the number of waves that have passed over us in that time.

[1] For our purposes all waves are sinusoidal, and we will ignore the difference between group and phase velocities.

Wavelength, usually denoted λ, is the distance between wave crests, if we froze the wave and measured with a tape measure. Just as frequency is related to the wave period by $f = \frac{1}{T}$, it can be handy to think of a "spatial frequency" (usually denoted by k) related to the wavelength, where k = $\frac{1}{\lambda}$. Similar to frequency and time, the product k * x is equal to the number of wave crests we would pass over if we walked a distance (x) through the water, perpendicular to the waves.

Direction is relatively simple. Consider which way a surfer being pushed forward by the wave would move: that is the wave direction. It turns out that this direction is always perpendicular (aka "normal") to the crest of the wave, which should make some sense. We will return to this point later.

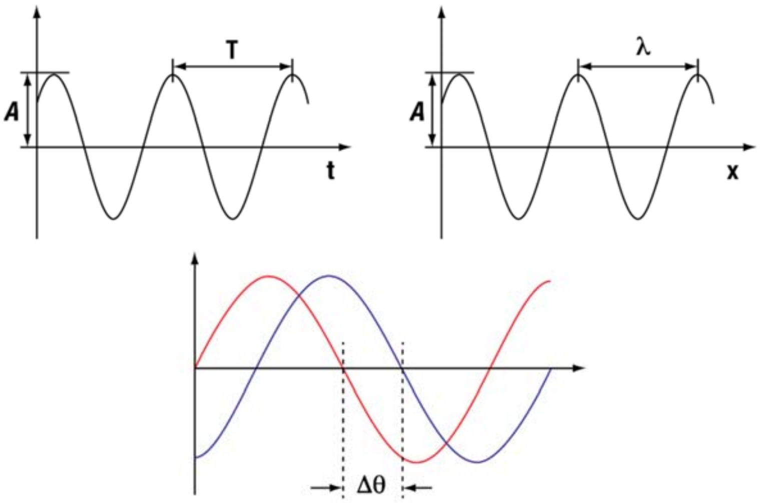


Figure 6: Waves as functions. Note that the top left figure is graphed vs. *time*, so period, T, is the *time* it takes for two wave crests to pass. The top right corresponds to a wave frozen in time, where we are walking along it and the wavelength, λ, is the *distance* between wave crests. The bottom figure shows one wave ahead of the other, which means it has a phase difference (Δθ).

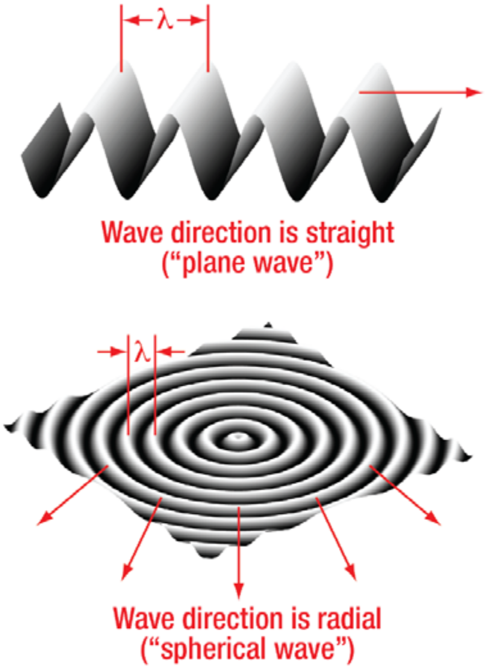


Figure 7: Two types of waves often discussed in optics: plane waves, where the wave crests fall along parallel lines and the wave has a single direction, and spherical waves (a spherical wave is a 3-D radial wave). Note that spherical waves have no single direction – rather, the wave expands in *all* directions, moving locally always perpendicular to the phase front (wave crest). Hot atoms (e.g. in a lightbulb filament), as well as fluorescent molecules, emit light in spherical waves centered on the atom or molecule. Getting ahead of ourselves, the arrows in the figure are the light *rays* which correspond to the waves shown.

Some Math

Later on it will help to be able to refer to waves using math, so it is helpful to state the equation for a wave here so we can refer back to it. There is nothing new here that was not already discussed above, but it is worth your time to think through these equations and make sure you have a sense for how they relate to the actual behavior of the wave.

Remember, a wave is a function that repeats in time. For this course we will choose our waves to be sinusoidal – sine or cosine functions, since sin(θ) and cos(θ) return to the same value every time θ increases by 2π, which is to say cos(2π + θ) = cos(θ). Remember: 2π radians just equals 360°.

If we want the wave to go up and down in time, we can make θ change in time – for instance, if we want the wave to go up and down and back up to the same spot in a time, T, we could write the amplitude of the wave as:

Equation 1:

$$\boldsymbol{A \cos\left(2\pi \frac{t}{T}\right)}$$

Notice that the wave will go *up* to a height A and *down* to a height –A, and it will do this every time t changes by an amount T, since if t increases by T, $\frac{t}{T}$ increases by 1 and the argument of the cosine increase by 2π * 1 = 2π.

We could do the same thing if we wanted to have the wave change with distance, like a set of rolling hills:

Equation 2: $$\boldsymbol{A\cos\left(2\pi\,\frac{x}{\lambda}\right)}$$

Notice that now the wave goes up and down the same way, but as we walk along in the x direction, and every time x changes by an amount λ, the argument of the cosine changes by 2π and the height of the hill (or wave) is back to where it began.

Of course, we can put these together:

Equation 3: $$\boldsymbol{A\cos\left(2\pi\,\left[\frac{x}{\lambda}-\frac{t}{T}\right]\right)}$$

The “–” sign could also be a “+,” which would merely change the direction of the wave. Think about being at the crest of the wave: at the top, the cosine must equal one, cos(θ) = 1, so the argument θ must equal zero (or a multiple of 2π; let’s choose zero for simplicity). When does

Equation 4: $$\boldsymbol{\left[\frac{x}{\lambda}-\frac{t}{T}\right]=0\ ?}$$

The answer is when

Equation 5: $$\boldsymbol{\frac{x}{t}=\frac{\lambda}{T}}$$

But $\frac{x}{t}$ is just the velocity, v, $\frac{\text{distance}}{\text{time}}$. Similarly, as we mentioned above, $\frac{1}{T}=f$, the frequency. Putting all this together gives us the fundamental relationship for a wave,

Equation 6: $$\boldsymbol{v=\lambda f}$$

Since we could have started the wave at some height other than cos(θ) = 1, it is usual to write the 1-dimensional wave equations as

Equation 7: $$\boldsymbol{A\cos\left(2\pi\,\left[\frac{x}{\lambda}-\frac{t}{T}\right]+\phi\right)}$$

where ϕ (pronounced “fie”) is the “phase constant” or “phase shift” and the entire argument in the curved brackets is the “total phase.”

Just as $\frac{1}{T}$ = f, we will often write $\frac{1}{\lambda}$ = k, where k = “spatial frequency”[2] just like $\frac{1}{T}$ = “temporal frequency.” If we let time pass for 1 second, a wave crest will pass us f times. If we walk 1 meter, we will cross k wave crests. Spatial frequency is a useful concept, and we will come back to it in Lab 6 when we get to the Abbe theory of image formation.

Writing the wave in this fashion, we have

Equation 8: $$\boldsymbol{A\cos(2\pi\,[kx-ft]+\phi)}$$

[2] Note: We define spatial frequency k = $\frac{1}{\lambda}$, NOT as $\frac{2\pi}{\lambda}$. There are good reasons for this, but regardless this is the convention we use. k here is thus analogous to the temporal frequency f, and not the angular frequency ω.

For those familiar with Euler's relation, a similar oscillatory function can be written as

Equation 9: $\boldsymbol{A\,e^{i\,(2\pi\,[kx-ft]+\phi)}}$

This form can be handy because it is so easy to multiply exponentials, whereas one must memorize a lot of trigonometric relationships to know what $(\cos\theta)^2$ is. There are numerous ways to write the equation for a wave (e.g. using a sine function), these are just some popular ones. Also, many symbols are used for the phase shift – it is not always φ or ϕ.

If you have not seen this before, review it later so you are a bit more familiar.

To recap what we have covered: in these equations,

Amplitude. Since the cosine function goes from +1 to –1 and back as the angle (the argument inside the brackets) increases, and averages zero, we have A = amplitude (height of the wave) in the equations above.

Phase. The whole argument of the cosine or exponential, $[2\pi\,(kx - ft) + \varphi]$, is the phase. When the term inside these brackets is constant, then the value of the cosine function stays the same, and we are moving along with the wave. For example, if $kx = ft$ and $\varphi = 0$, then $[2\pi\,(kx - ft) + \varphi] = 0$ always, therefore $\cos[0] = 1$, and we will always be on the wave crest, surfing along as the wave moves forward. Choosing $\varphi \neq 0$ just means we have chosen to move along with a different point on the wave – for instance, choosing $\varphi = \pi$ means we are moving along with the wave trough instead of the crest, since $\cos(\pi) = -1$.

Speed. How fast would we be surfing if we were moving along with the wave crest? Speed = distance / time, and since we are moving along with the crest (so the wave height is never changing for us) we have $\boldsymbol{kx = ft}$. We can rearrange that into $\frac{x}{t} = \frac{f}{k} = \lambda f = \frac{\lambda}{T}$ (since $\boldsymbol{k} = \frac{1}{\lambda}$ and $\boldsymbol{f} = \frac{1}{T}$). This gives us the fundamental wave relationship $\boldsymbol{v = \lambda f}$.

The wavelength (λ) of the waves, times the frequency (f) with which the waves go by, gives their speed (v). So if we surf along the wave crest, we will move at speed $\boldsymbol{v = \lambda f}$. Looking at it another way, the speed = $\dfrac{\textbf{distance between wave crests}}{\textbf{time between wave crests}}$, or $\boldsymbol{v} = \frac{\lambda}{T}$.

Frequency, period, wavelength, and spatial frequency were defined above:

Equation 10: $\boldsymbol{f = \frac{1}{T}}$

Equation 11: $\boldsymbol{k = \frac{1}{\lambda}}$

Direction. Since we chose a 1-D equation above, direction does not really show up – of course we have to move in x. For 2-D or 3-D, the equation looks very similar, but k becomes a vector related to how much the wave's direction lies along x, y, or z. What we get is $\boldsymbol{k_x = k\cos(\theta_x)}$, $\boldsymbol{k_y = k\cos(\theta_y)}$, and $\boldsymbol{k_z = k\cos(\theta_z)}$, where the θ angles are the angles between the wave direction and the x, y, and z axes. The equation for the wave then becomes:

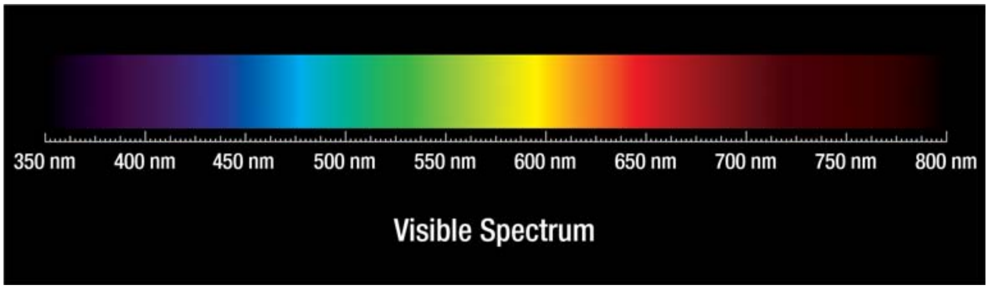
Equation 12: $\boldsymbol{A\cos\left(2\pi\left[k_x x + k_y y + k_z z - ft\right] + \phi\right)}$, or

Equation 13: $\boldsymbol{Ae^{i(2\pi\,[k_x x + k_y y + k_z z - ft] + \phi)}}$

Note that Equations 12, 13 reduce to equations 8, 9 if the wave is moving in the x direction, since then $\theta_x = 0$ (the direction of motion is parallel to x), while θ_y and $\theta_z = \frac{\pi}{2}$ (remember $\frac{\pi}{2} = 90°$) and so $k_x = k$, while $k_y = k_z = 0$. Although it is not as obvious stated this way, the direction of wave motion is still perpendicular the wave crest at each point.

Light Waves

Light waves have a number of important traits. One of the most notable is that we can see them, and perceive different frequencies (or, equivalently, wavelengths) as different colors:



Color	Wavelength Interval	Frequency Interval
Red	~ 700 - 635 nm	~ 430 - 480 THz
Orange	~ 635 - 590 nm	~ 480 - 510 THz
Yellow	~ 590 - 560 nm	~ 510 - 540 THz
Green	~ 560 - 490 nm	~ 540 - 610 THz
Blue	~ 490 - 450 nm	~ 610 - 670 THz
Violet	~ 450 - 400 nm	~ 670 - 750 THz

Figure 8: Color of light, as a function of wavelength (λ).

This will be very important, and so you should look at it carefully, and memorize some common reference points. The human eye is most sensitive in the yellow-green, so when people choose a "typical" wavelength for imaging, they usually choose something around ~ 500 or 550 nm, or about ½ μm. For reference, a human hair is ~ 25 μm dia., and aluminum foil is usually ~ 1.5 μm thick.

Study Point 1: Memorize the wavelengths vs. color table (you can skip the frequencies).

There is a reason you do not really need to know the frequency of light – it is too high to notice. The fastest computer chip is a few GHz, FM radio signals are ~ 100 MHz, the highest audible frequency is ~ 20 kHz, and your eye can respond at ~ 30 Hz. The fastest oscilloscope is ~ 1000x slower than the oscillation frequency of light. As a result, we only ever see the *average* effects of light. This does not mean we do not see wave properties, just that the wave properties we see (e.g. interference fringes) must be pretty steady – a plane wave incident on the wall will not appear to be flashing on and off, but just look like constant illumination.

It is important to note that when light goes from one material (say, air) into another (say, glass), its *frequency* does not change. This makes sense if you consider that the oscillations must "match up" at the boundary. The light wave in the air is causing the electrons to oscillate in the glass, which is what makes up the subsequent wave traveling in the glass. Those electrons must be "going up and down" at the same rate as the incident wave, since the incident wave is what makes them vibrate in the first place. Everything oscillates at the same frequency.

Study Point 2: Think about this, and return after the next section to make sure you understand.

Light usually slows down in denser materials; the speed of light in different materials is given by $v = \frac{c}{n}$, where c is the speed of light in vacuum and n is the index of refraction. Since light cannot go faster than c (= 3e8 m/s), n ≥ 1 for all cases of practical interest.

If the speed of light changes, but the frequency does not, then the wavelength ***must*** change, since $\boldsymbol{v = \lambda f}$.

Being more formal about this, when people talk about the wavelength of light, they always mean the wavelength of light with that frequency traveling in vacuum (where n = 1, or, equivalently, v = c). We will call this λ_{vac}, for "lambda vacuum," the vacuum wavelength, and

Equation 14: $$\lambda_{\mathbf{vac}} = \frac{\boldsymbol{c}}{\boldsymbol{f}}$$

If the light goes into glass, which typically has n ≈ 1.5 (i.e., light in glass has slowed down to $\frac{1}{1.5} = \frac{2}{3}$ its normal speed, since $v = \frac{c}{n}$), then

Equation 15: $$\boldsymbol{\lambda = \frac{v}{f} = \frac{\frac{c}{n}}{f} = \frac{\lambda_{vac}}{n}}$$

So the wavelength of light in glass is only ~ $\frac{2}{3}$ of its wavelength in vacuum. This definitely matters; the wavelength of light limits the smallest objects one can resolve in a light microscope. One can increase resolution by using bluer light (i.e., light of shorter vacuum wavelength), but one can also increase resolution by putting the sample in a substance (e.g. oil) where light goes more slowly, and so the wavelength *in the material* is shorter.

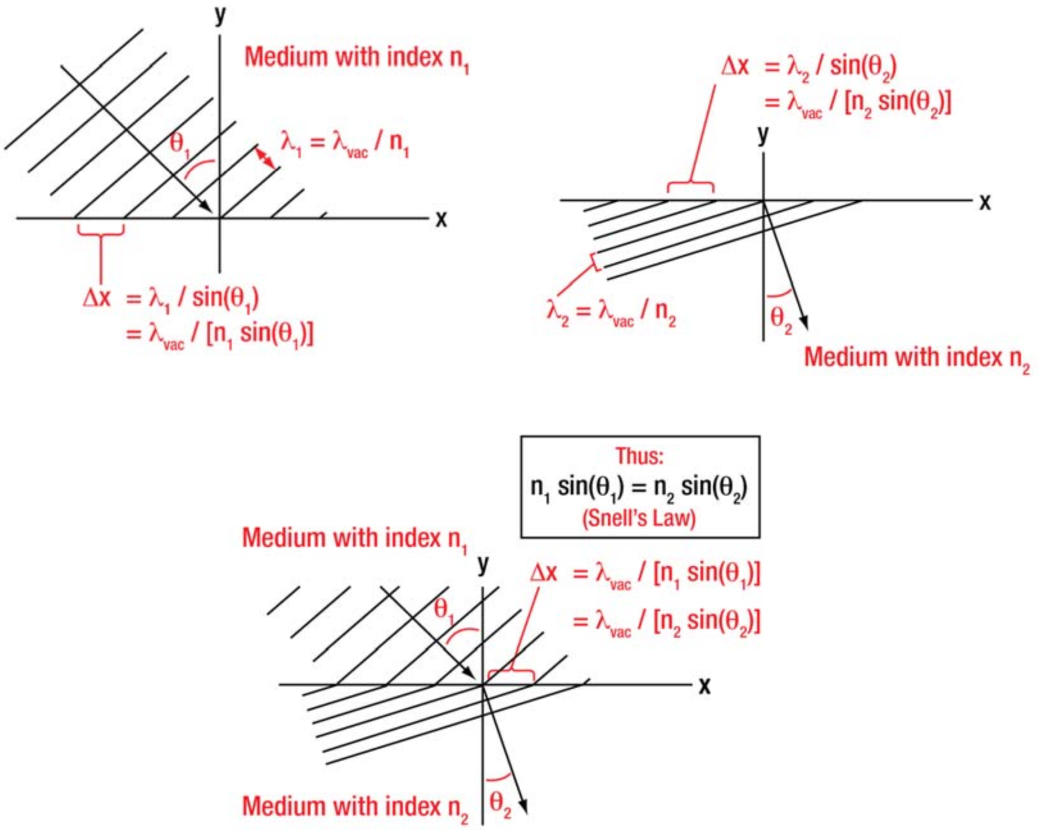
By the way, this is not unique to light – it is a general wave property. When ocean waves come into the shallows near a beach, they slow down (their speed decreases) and they crowd closer together (their wavelength gets shorter), something you can easily see if you look down at a beach (e.g. from the Marin Headlands near San Francisco).

Wavefront Bending and Rays: Snell's Law

We are now ready to get down to some real optics. Let's consider what happens when a plane wave in air (index n_1 = 1.0) hits a glass interface (n_2 = 1.5) at some angle θ_1.

To begin with, what do the "n" values mean? Remember from the last section that the speed of the wave is $v = \frac{c}{n}$, so the wave in air (n = 1.0) must be going at the speed of light, while the wave in glass (n = 1.5) has apparently slowed down to about $\frac{2}{3}$ the speed of light. Also recall that the frequency does not change (if this does not make sense, reread the last section, and/or ask in class). If the frequency does not change, the only way the speed could have changed is for the wavelength to change, which is what we saw in Eq. 15. In this case the wavelength λ (in air) will drop to $\frac{2}{3}\lambda$ when the wave goes into the glass.

The frequencies of the waves must be the same, so the wave crests must match up:



y
Medium with index n_1
θ_1
$\lambda_1 = \lambda_{vac} / n_1$
x
$\Delta x = \lambda_1 / \sin(\theta_1)$
$= \lambda_{vac} / [n_1 \sin(\theta_1)]$
$\Delta x = \lambda_2 / \sin(\theta_2)$
$= \lambda_{vac} / [n_2 \sin(\theta_2)]$
y
x
$\lambda_2 = \lambda_{vac} / n_2$
θ_2
Medium with index n_2
Thus:
$n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$
(Snell's Law)
Medium with index n_1
y
θ_1
$\Delta x = \lambda_{vac} / [n_1 \sin(\theta_1)]$
$= \lambda_{vac} / [n_2 \sin(\theta_2)]$
x
Medium with index n_2
θ_2

Figure 9: Snell's law. Because the frequencies are the same, the wave crests need to line up at the interface between two materials (known as "meeting the boundary conditions"). The result is that the wave directions on either side are related by Snell's Law.

Note that the wavefronts in the medium with index n_2 are closer together – the wavelength is shorter (so you know $n_1 < n_2$, right?) What happens to the wavelength when the wave goes into a medium of *lower* index?

Also look at what happens when the wave hits the interface at an angle: since the wavelengths are different in the two materials, the distance between wave crests cannot be the same unless the wave changes direction – the wavefronts *have* to be at different angles in order to make things work out. This result is known as Snell's Law.

In Figure 9, notice the direction of the arrows; the direction of the wavefronts is equivalent to a light "ray." This is the link between ray optics and wave optics, and we can see that the bending of light "rays" at interfaces actually results from the wave nature of light and the shift in direction of the wavefronts. Traveling along a ray is the same as surfing down the wave crest, and talking about "rays" really only makes sense when the phase fronts are relatively smooth, which is to say when it is easy to define the normal to the wave crests (i.e., the direction you would surf down one). It turns out this is generally practical when the distances under consideration are all >> λ. Note that this analysis of wave bending holds true for all waves – sound waves, water waves, light waves, etc.; there is nothing special about light here.

Study Point 3: Convince yourself, using drawings similar to those in Fig. 9, that any *reflected* light wave (ray) will be reflected at an equal and opposite angle from the incoming wave (ray).

Your argument will give the reflection angle, but will not tell you *how much* light is reflected. That is harder to derive, so we will just state it: the fraction of light reflected from an interface, when a wave hits it perpendicularly ("at normal incidence"), is given by the reflection coefficient R, where[3]

Equation 16: $$\boldsymbol{R = \left[\frac{n_1-n_2}{n_1+n_2}\right]^2}$$

The proportion of transmitted light is (rather reasonably) given by T = (1 – R), but nobody memorizes this – they just derive it when they need it. It should make sense: if, say, 8% of the light is reflected, then 92% must be transmitted.

> Aside: the amount of light reflected varies a lot with angle – you have probably noticed that when the sun is low, it reflects off the road very well, making it hard to drive facing into the sun. The details of that variation are beyond the scope of this course, but you can look them up easily if you need them.

For simplicity, we will restate Snell's law as well, and then suggest some study questions:

Equation 17: $$\boldsymbol{n_1\, sin(\theta_1) \;=\; n_2\, sin(\theta_2)}$$

Study Point 4: Memorize the reflection coefficient at normal incidence, and also Snell's Law.

Study Point 5: Figure out which way light bends when it goes into a higher index (usually, more dense) medium. Does the angle get bigger or smaller?

Study Point 6: Calculate (and memorize) the reflection coefficient R for light going from air (n = 1) into glass (n = 1.5), and for light going from glass into air. What is the total reflection loss for light going through a plate of glass (two sides)? Fancy microscope objective lenses may have as many as *eight* air-glass interfaces, as well as some interfaces between different index glasses. What fraction of light would you expect to get through such an objective lens, if no special steps were taken?

[3] For those tracking such things, this is the *intensity* coefficient of reflection, not the amplitude coefficient.

Lenses: Phase Shifts, Wavefront Shaping, and Ray Bending

We are now armed with everything we need to know to discuss lenses. Before we start, though, it is worth asking a simple question which has complex ramifications: what does it mean to image? After all, the whole point of lenses in microscopy is to image – to know what we want a lens to be doing, we have to define what we want.

In one sense, the answer is simple: we want an image, a re-creation of the light distribution at the sample (perhaps scaled up or down by some magnification M).

To make it slightly more complicated, why can we not put a piece of film (or a digital sensor) in front of the sample, and see the image?

Study Point 7: Try this: In a dark room, take a piece of white paper, and hold it up to your computer screen. Look at the side of the paper closest to the screen (you will have to angle the paper to see it). How close do you have to get the paper to the screen before you can see anything like what is on your screen? You may have to experiment with different images on your screen- for me, looking at the desktop (blue background with the yellow folders) worked best.

Obviously, if you could lay the paper (or film) directly on the screen, you would get a good image, though it does not take much distance before things get very fuzzy. The issue we are confronting is the *effect of the propagation of light.*

So really the main things we want from a lens are to:

a. Reverse the effects of light propagation (to make the image clear again)
b. Allow for scaling of the image (magnification).

What are the effects of propagation? Let's consider a point source (these are famous, but quite reasonable idealizations: for all practical purposes, a hot atom or a fluorescent molecule is a point source of light). Without getting into quantum mechanics (though the analogy is not so bad even there), we can think of a point source as a vibrating electron – any accelerating charge always generates a radiating electromagnetic field, and light *is* just a radiating electromagnetic field.

This is exactly the same as putting your finger in the middle of the water in the bathtub and wiggling it up and down – you will get spreading waves in a circle around your finger. This is the same for the vibrating electron, except that the waves are electromagnetic, moving a lot faster, oscillating up and down at 5×10^{14} cycles / sec., and are in 3-D.

What we have is something like Figure 10, below, where the oscillating electron is at the center of the expanding waves.

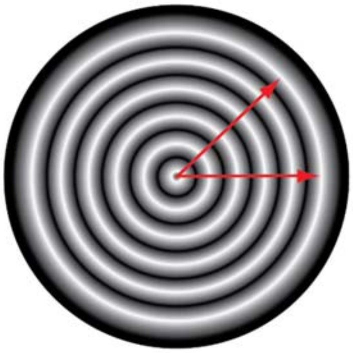


Figure 10: Expanding spherical waves. Note that the time it takes to reach any point on the wavefront (e.g. where the two red arrows go) is the same.

An important fact is that it will take a surfer on the crest of the wave the same amount of time to follow the wave crest in any direction. While this makes sense in the spherical wave case since everything is the same in all directions, it is more generally true since the wave crest is a phase-front: the height of the wave is always the same there. For a wave with the equation cos[2π (kx – ft)], this means that for a phase front,

Equation 18: $$(\boldsymbol{kx} - \boldsymbol{ft}) = \boldsymbol{constant}.$$

Let's multiply Eq. 18 by λ, remembering that k = $\frac{1}{\lambda}$ (Eq. 5). We then have

Equation 19: $$(\boldsymbol{x} - (\boldsymbol{\lambda ft})) = \boldsymbol{constant}.$$

And we know (from Eq. 3) that λ f = v, so for the phase front (or wave crest),

Equation 20: $$(\boldsymbol{x} - \boldsymbol{vt}) = \boldsymbol{constant}.$$

We already covered this in terms of changing λ for different indices, but the form (x – vt) is also illuminating (we will do this a lot in this course – looking at things from slightly different perspectives). In 3-D, x becomes r, and we can talk about the distance the wave crest has moved as having radius r, which is growing in time. As r increases, the curvature of the wave changes (in this case it is getting flatter), which is to say that ***propagation changes the radius of curvature of the phase front***.

We want a lens to reverse that increase in the radius of curvature. The easiest way to do this is to slow down some parts sooner than others. For instance, slow down the middle first, then after the edges have caught up, slow them down too. And what slows down light? Anything with a higher index, say glass, if the wave was initially in air. For simplicity, let's make the expanding wave into a plane wave, using a curved glass surface. If we do that, we can repeat the process to make it into a converging spherical wave, converging onto a point, which is what we started with, and thus we would be making an image!

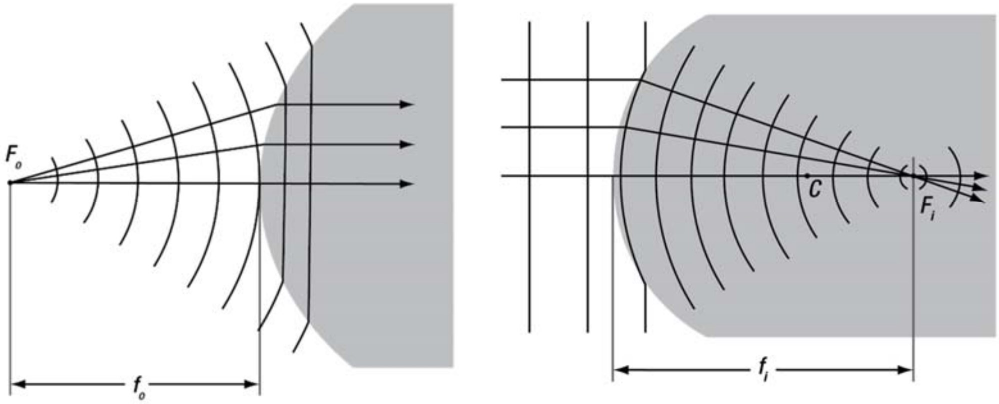


Figure 11: Left: A curved glass surface (shaded) is used to slow down the central part of an expanding spherical wave first, allowing the edges to catch up and form a plane wave. Right: Similarly, a plane wave is turned into a converging spherical wave, forming an image of the original point source.

Of course, there are some problems. To begin with, we started with an expanding *spherical* wave, but what we have on the right side of Fig. 11 is only a converging *cone* of waves, just part of a spherical wave, so we cannot expect to get a perfect replica of our original source: we are missing some of the bits, all the original spherical stuff that is not in the cone we have at the end. We will come back to that later, but for now there is a larger worry: how do we get the right shape for the curved surface?

Remember that every point on the wavefront has taken the same time to get there, so if the wavefronts all start at the same (source) point and end at the same (image) point, the time each part of the wavefront takes to get there is the same. But we know what the speeds of the waves are in different media: light goes at speed c in vacuum and at speed $v = \frac{c}{n}$ in media of a higher index. Assuming air has n ~ 1, and assigning glass n = n_2 (actually ~ 1.5), we know that the time taken by a surfer on any part of the wavefront is the distance *divided by* the speed for each leg of the journey, as shown in Fig. 12.

Why *divided* by? Velocity = $\frac{\text{distance}}{\text{time}}$, so $\frac{\text{distance}}{\text{velocity}}$, $\frac{d}{v} = \frac{d}{\frac{d}{t}}$ = t, time. But v = $\frac{c}{n}$, so time = $\frac{d}{v}$ = $\frac{nd}{c}$. The constant c (the speed of light) appears in *all* the times, so we can ignore it if all we care about is relative differences, which leaves only time ~ n d.

This formulation, where each distance is weighted by the index of the medium (since the speed depends on that) is known as the "optical path length," or OPL, where OPL = n d. As an example, 1 cm of glass has an optical path length of ~ 1.5 * 1 cm = 1.5 cm, since n = 1.5 for glass. It will take a wavefront the same time to go through 1 cm of glass as it takes to go through 1.5 cm of air (or vacuum).

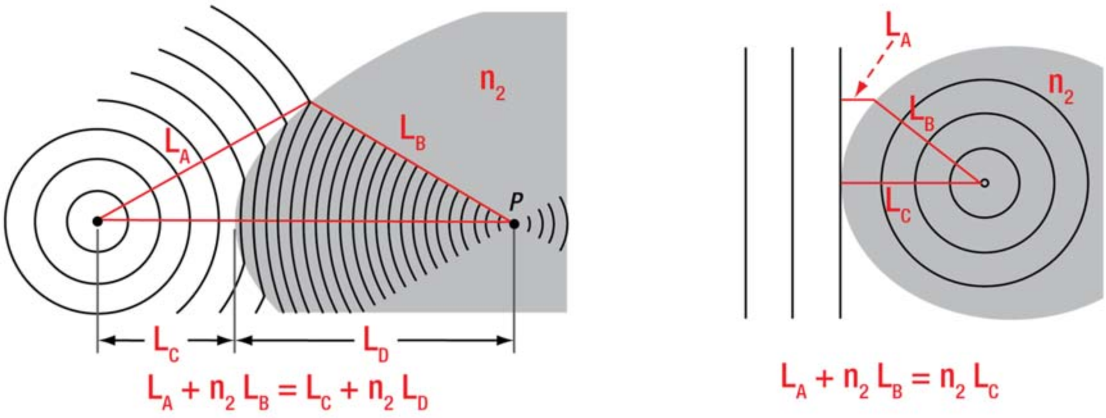


Figure 12: Equal time paths, defining the correct curved surface to form an image. Note that the formula (and hence the curve shape) for imaging one point to another (Left) is different than for focusing a plane wave down to a point (Right).

It is important to note that the required curve shape in Figure 12 changes depending on what sort of change you want to make in a wavefront – if you see in a catalog that something is designed to be the "best," you should immediately ask "for what?" Being the best at focusing a plane wave is not the same as the best at imaging one point to another. Remember this when reading the sections on "Lens Shape" and "Lens Combinations" in the CVI / Melles Griot Fundamental Optics Guide (see the *Reference Links* tab at www.thorlabs.com/OMC for the link).

Perhaps worse than the problem that a different shape is required for every different focusing task is that these shapes are not easy to make – which is to say, they are not spherical. By far the easiest surface to make is spherical – grind two surfaces together in a rotary motion and they *both* naturally become spherical (one hollow or concave, the other convex). This is no small issue: in the 1700's and 1800's a sphere was the best you could do, and even now aspheres (non-spherical shapes) are much more expensive to make than spherical lenses (when manufactured in similar quantities).

The spherical surface is, in general, not thick enough at the edges compared to the curves above (it "bends back" too quickly); as a result, the wavefronts "wrap around" a little too much for points far from the optic axis, and consequently light in the wavefronts at the margins come to a focus closer to the sample than the light closer to the axis. This is known as "spherical aberration" and is shown in Figure 13:

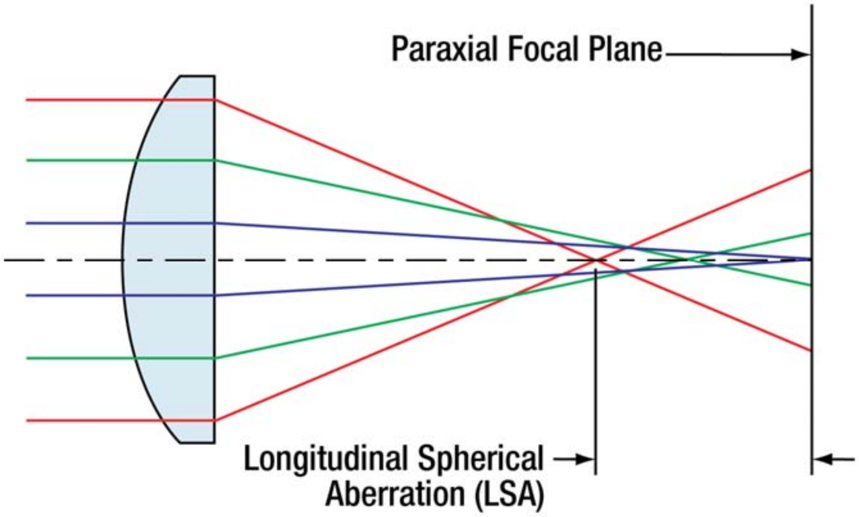


Figure 13: Spherical aberration. Note how the rays at the edge (red) come to a focus sooner than those near the axis (blue). One can view this as the edge rays bending "too much," or as not having enough glass at the edges so that the wavefronts out there "curl around" too far. Remember that rays are just the path perpendicular to the wavefronts, so the views are equivalent. Note: All the light is the same color for purposes of this discussion; the different line colors are just to help make sense of it.

There are a variety of different defects in optical systems due to the use of spherical surfaces (rather than the ideal curves), and we will return to them later. The usual (and often quickest) way of discussing them is in terms of ray bending, but it is worth remembering that there is an equivalent way of discussing them in terms of phase fronts. In fact, imaging lens aberrations are sometimes discussed in terms of the "phase error" they consist of. This view is especially important when considering adaptive optical systems for aberration correction and also for phase mask systems, e.g. where a liquid-crystal display can be used to change the phase of a wavefront, and hence its shape and focusing properties.

THORLABS

Lab 2 Notes: Introduction to Optical Imaging (II)

Optical Microscopy Course

Lab 2 Course Notes:
Introduction to Imaging (II)

Overview

For Lab 2, you will explore imaging further with your camera and a lens, set your system up to be the collection side of an "infinity optical system," and get further experience with the optomechanics by building the collection side of your optical train and the mechanical rail system you will use for the rest of the course. Using this system, you will examine the imaging quality of different lens orientations and types, and in different colors of light, as well as familiarize yourselves with a resolution test target and use it to calibrate your magnification; this latter set of tasks will likely spill over into Lab 3. The camera is covered in the optional part of the assigned reading, and will be discussed in more detail over the course of the next several labs, but at a basic level it just records the light intensity present on the sensor face.

Imaging at Infinite Conjugates

How do we get an image onto the sensor? Using a lens, or several lenses. For the purposes of Lab 2 (and for most of this course), we only need to know a few things about lenses, and the main ones are:

A lens will form an image of an object placed on one side of it if the distance to the image and object obey the following rule:

Equation 1: $$\frac{1}{f} = \frac{1}{S_{object}} + \frac{1}{S_{image}}$$ **(this is worth memorizing),**

where f is the focal length of the lens (usually you just order a lens with the focal length you want, so all you need to know is what f you need). S_{object} and S_{image} are the distances from the lens to the image or object respectively.

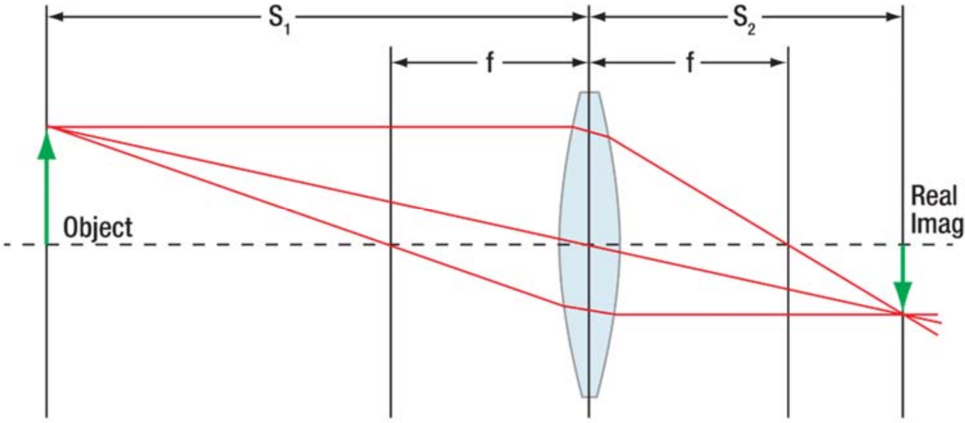


Figure 1: Imaging with a lens. $S_1 = S_{object}$ and $S_2 = S_{image}$.

It turns out that the magnification, M, of the image is given by Equation 2, below. The real image is upside down compared to the object; hence the "–" sign:

Equation 2: $$M = -\frac{S_{image}}{S_{object}}$$ **(this is worth memorizing).**

You can compute the focal length using the "lensmaker's equation," which is:

Equation 3: $$\frac{1}{f} = (n - 1)\left[\frac{1}{R_1} - \frac{1}{R_2}\right]$$ (no need to memorize this)

Here n is the index of refraction of glass; typically n ~ 1.5, but it varies a bit for different colors of light and types of glass. R_1 and R_2 are the radii of curvature of the sides of the lens, so if you know the index of your glass and the focal length you want, you can choose the radius to grind on each side of your piece of glass.

Technically Equation 3 assumes that the lens is rather thin and that there is air on either side of it (the formula is different if the lens is in, e.g., water, or for lenses thick compared to their focal lengths – for that see the CVI/Melles Griot Fundamental Optics Guide, pp. 29 – 35, see link in the *Reference Links* tab at www.thorlabs.com/OMC)

The most important thing to note about Equation 3 is just that if n varies for different colors of light then the focal length f will also vary, and so (using Equation 1) the location of your image will also change. This can be a problem, and you will explore it in Lab 3.

Conveniently, most of the optics we will do in this class involves using the lenses such that either the object we use or the image we form is infinitely far away (referred to as "at infinity"). If you plug infinity into Equation 1, things get very simple; for example, if $S_{object} = \infty$, then:

Equation 4: $$S_{image} = f$$ **(when $S_{object} = \infty$)**

So if we want to image something infinitely far away, all we need to do is put the lens one focal length away from the camera. We know roughly how far that is since we know what focal length lens we ordered, and we can check if we got it right by adjusting the distance using a focus mechanism until the image looks good on the camera. Usually we attach the lens to the camera with a tube, so we call this the "tube lens" since it is the lens that goes in the tube.

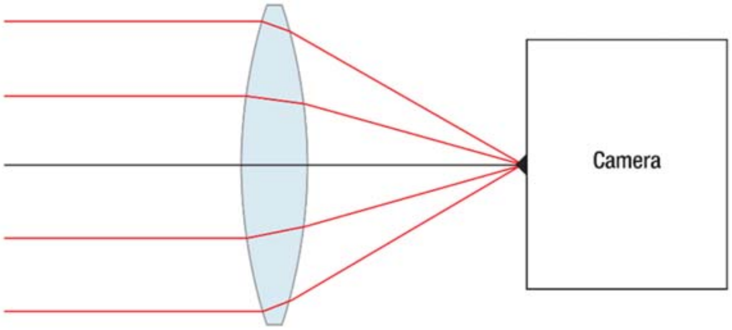


Figure 2: Imaging from infinity onto a camera.

What if we do not know the focal length because the lens was taken out of its package and some (terrible) person did not label the lens holder they put it into? You can figure this out yourself using the lens and a ruler; see if you can figure out how.

Using what we have just discussed, we can get an image of something infinitely far away to look good on our camera. That, however, is not microscopy; in microscopy things are usually very close. So what now?

One way (the old, more complicated way) to do this would be to figure out just what magnification you want, then decide how far away from a sample to put your lens, then get the camera to be exactly the right distance away from the other side of the lens. This is a hassle – the only way to do it is to position things at very precisely known distances, which is hard to do.

The new, better way is as follows: put a lens on your camera, and focus the system at infinity. Now put another lens in front of the camera and its lens, near the object you want to image (we will call this the "objective lens" since it is near the *object* we are imaging). If we move that lens until the object is one focal length away from it, the object will now be imaged at infinity (parallel rays).

The camera is already set up to image things infinitely far away, so we are all set: see Figure 3. The great part about this (well, one of the great parts; there are others) is that we did not have to measure a single thing, or position anything at a precisely known distance – we just focused the camera and its lens at infinity, then adjusted the distance between the other lens and the object until it was in focus on the camera. Most modern microscopes work this way (it is often called "infinity corrected," or "working at infinite conjugates").

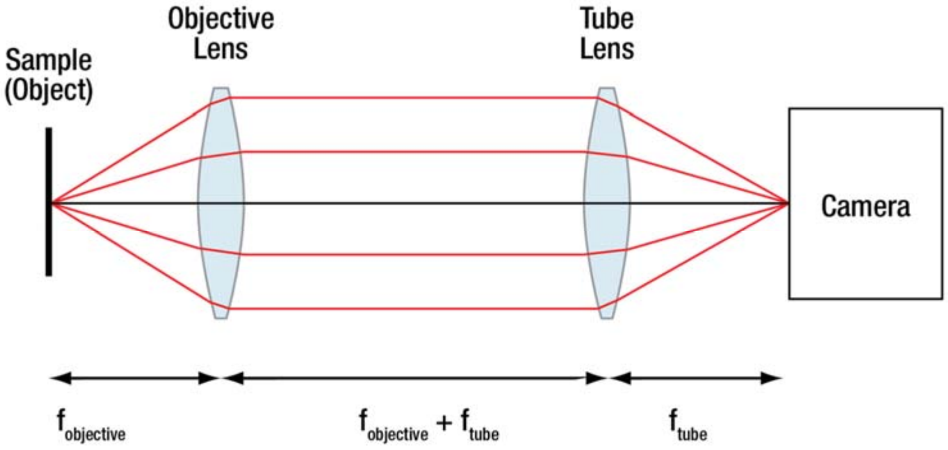


Figure 3: Imaging a sample to infinity, then imaging it from infinity onto a camera.

It might not be much surprise that the magnification of this system is very similar to Equation 2:

Equation 5: $$M = -\frac{f_{tube}}{f_{objective}}$$ (no need to memorize this)

Here the distance from the object to the objective lens (which is also the focal length) is equivalent to the object distance, S_{object}. Also, the distance from the tube lens to the camera (which is the focal length of the tube lens) is equivalent to the image distance, S_{image}. What about the distance between the lenses? It does not matter that much[1], which is extremely handy – we can change it if we want without worrying about messing up our image.

A theoretically convenient way to space the lenses is to use the sum of their focal lengths, so the distance between them is $d = f_{objective} + f_{tube}$; this arrangement is known as a "4-F" system, since the total length from the object to the image is the sum of four focal lengths. This is a good place to start, but in general there will be a specific distance that is best, and it will usually not be exactly the 4-F distance. That is

[1] Actually, a lens system will be designed assuming very specific distances between the lenses, and will work best if set up that way. We will touch on this (briefly) in later labs. The important thing for this class is that the "infinity corrected" approach allows you to set the magnification without precisely measuring distances, and is in general pretty insensitive to the exact distance between the two lenses used, which is extremely convenient.

often close, though, and good enough. If you want to err, make the lenses a little too close, rather than too far apart. We will discuss this more later in the class, but for now just remember that there is a good distance to put the lenses, it is not critical that it be very exact, and that "4F" is a decent guess for the maximum distance. Armed with all this, you can already make a surprisingly good microscope! You will do this, and start looking at its performance in Lab 2.

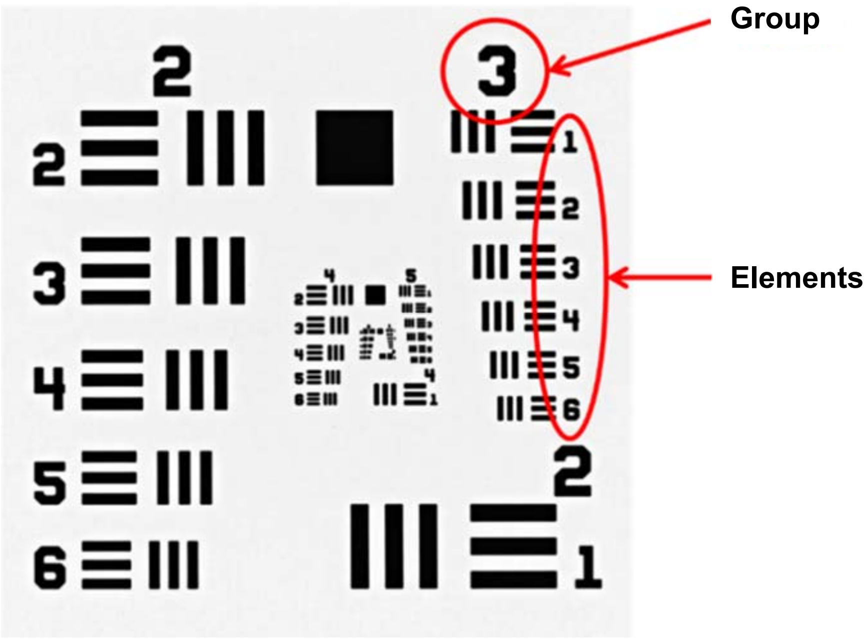
USAF 1951 Resolution Test Target

For this lab, you may want to understand what a USAF test target is (USAF stands for United States Air Force; these targets were originally used for testing aerial surveillance cameras). The target is chrome on glass (chrome blocks all the light, so the contrast can be very good). It has a series of smaller and smaller 3-bar test patterns, where the spacings between the bars are known. Roughly speaking (we will speak more precisely later in the course), the smallest set of bars you can resolve by eye give you the resolution of your system. There are 6 "Elements" (pairs of 3-bars) in each "Group" – for instance, in the image below Group 2, element 1 is in the lower right corner. Notice that sometimes the biggest element in a Group is on the other side of the slide from the rest.

For the lab, you should look over the pictures below – it will help you when you are trying to figure out where you are on an out of focus sample, and how to find the small sets of bars.

Your resolution will be roughly 1 / (line spacing), so if there are 8 "line pairs / mm" as given by the table for the set of bars you are looking at (i.e., the spacing between the smallest set of lines you can distinguish is 1 mm / 8 bars = 0.125 mm from bar center to bar center), then the resolution would be ~0.125 mm.

Example: In the target below, say the smallest set of bars you can make out is roughly Group 4, Element 5. The table indicates this corresponds to 25.39 line pairs / mm, so the spacing between the lines is 1 / [25.39 lines / mm] = 0.039 mm = 39 μm, and so your resolution would be about 39 μm.



$$\boldsymbol{Resolution}\left(\frac{\boldsymbol{line\ pair}}{\boldsymbol{mm}}\right) = 2^{\boldsymbol{Group} + \left(\frac{\boldsymbol{Element}-1}{6}\right)}$$

Element	Group Number									
	-2	-1	0	1	2	3	4	5	6	7
1	0.250	0.500	1.00	2.00	4.00	8.00	16.00	32.00	64.00	128.00
2	0.280	0.561	1.12	2.24	4.49	8.98	17.95	36.0	71.8	144.0
3	0.315	0.630	1.26	2.52	5.04	10.10	20.16	40.3	80.6	161.0
4	0.353	0.707	1.41	2.83	5.66	11.30	22.62	45.3	90.5	181.0
5	0.397	0.793	1.59	3.17	6.35	12.70	25.39	50.8	102.0	203.0
6	0.445	0.891	1.78	3.56	7.13	14.30	28.50	57.0	114.0	228.0

Values are in lp/mm.

Lab 3 Notes: Aberrations and Illumination

Optical Microscopy Course

Lab 3 Course Notes:
Aberrations and Illumination

Overview

Lab 3 involves imaging; it may be your first experience with imaging a resolution target if you did not have a chance to do so during Lab 2 (this timing varies a bit from year to year when we teach the material). This lab involves imaging of the resolution target, and then goes on to have you explore the impact of different lens aberrations (imperfections) on image quality, as well as the effects of "coherent" and "incoherent" illumination on the way an image looks.

References

For links to the following references, see the *Reference Links* tab at www.thorlabs.com/OMC.

1. The Wikipedia pages on "Spherical Aberration" and "Achromatic Lens" are both related to this lab.

2. First 36 pages of IDEX / CVI / Melles Griot Fundamental Optics Guide are extremely good – and more technical than the Wikipedia material. Particularly, the Fundamental Optics Guide section on lens "Performance Factors" is directly relevant to this lab.

3. The Molecular Expressions™ Optical Microscopy Primer website, put together by Michael Davidson and collaborators at Florida State University, is excellent, as are the related sites by Nikon and Olympus (as well as separate sites by Zeiss and Leica).

Aberrations and Illumination

To understand how lens imperfections and other issues affect image quality, it is worth starting by considering what it takes to actually make an image. In the Lab 2 Course Notes, we saw that we could make an imaging system by collecting light from the sample, imaging it to infinity (i.e., making all the rays from a point parallel after the first lens), and then focusing that light down onto the camera, as shown in Figure 1.

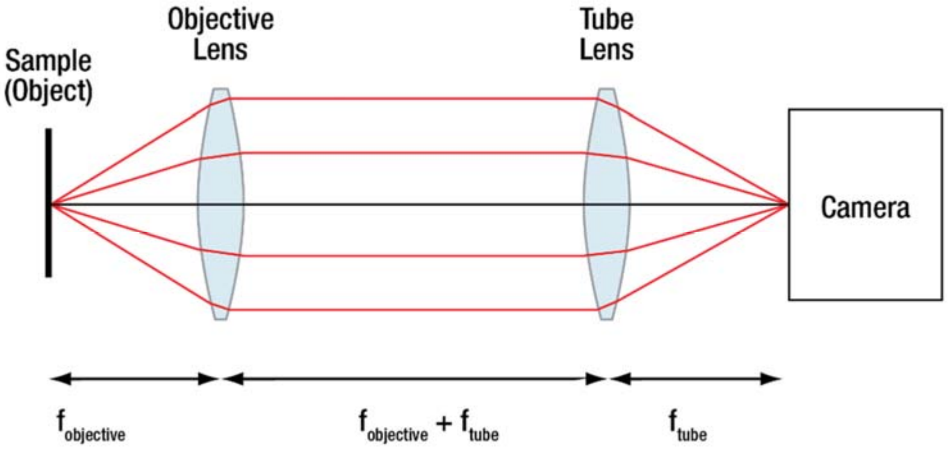


Figure 1: Imaging a sample to infinity, then imaging it from infinity onto a camera.

In principle this should give us a perfect image, since each infinitely small point in the sample will focus to an infinitely small point on the camera, as shown in the top image in Figure 2:

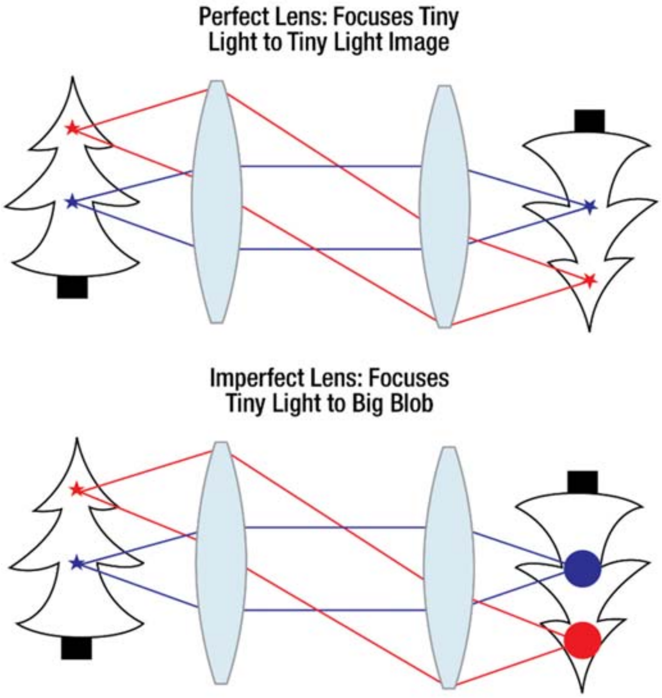

Perfect Lens: Focuses Tiny Light to Tiny Light Image
Imperfect Lens: Focuses Tiny Light to Big Blob

Figure 2: (Top) perfect imaging, with zero-blur; (Bottom) imperfect imaging with lens blur

However, lenses are all imperfect to one extent or another, so they focus a point source of light to a larger blur area as shown in the bottom image (Fig. 2). This blur is known as the Point Spread Function, or PSF, since it is the amount a point source of light is spread out into in the image.

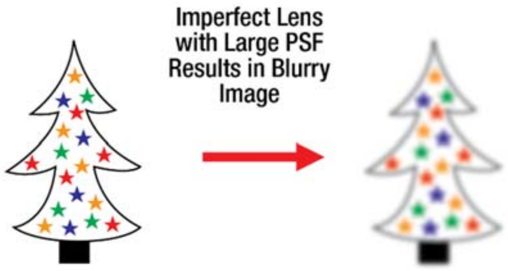


Figure 3: An object can be considered as a collection of infinitely small point sources of light; if the lens images each point source to a large blur – or PSF – then the image will be blurry and one will lose the ability to see (i.e. resolve) fine features like the points on the little stars.

Being out of focus is perhaps the simplest example of something that would increase the blur – or point-spread function, PSF – of a lens system, as Figure 4 shows.

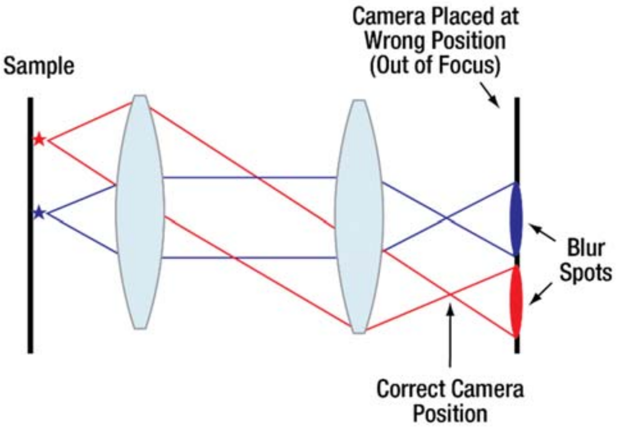


Figure 4: Focus error resulting in an increase in image blur (i.e. system PSF). Note that as the blur gets larger, eventually the light from the two separate stars will overlap, and it will no longer be possible to tell them apart (resolve them).

Focus is something one could correct, and in fact when focusing one usually adjusts things until the image is the *least blurry* – i.e., the system has the smallest point-spread function. However, there are traits of lenses one cannot usually adjust once they have been made, and that serves to increase the PSF. These are known as lens "aberrations." There are a number of different types of aberrations, but they fall primarily into three categories: aberrations due to the shape of the lens(es), aberrations due to the material of the lenses, and a performance limit due to the fact that lenses, not being infinite in extent, cannot capture all the light from a sample. This last one is known as diffraction and, while not technically considered an aberration, it nonetheless produces a similar effect in terms of increasing the PSF.

In this lab you will have a chance to explore two different lens aberrations: spherical aberration and chromatic aberration[1]. Spherical aberration is caused by the fact that it is by far cheapest to grind lenses in spherical shapes – each surface typically has a radius of curvature, instead of being some more complex ("aspheric") shape which is more expensive to produce. However, a sphere is not necessarily the best shape for imaging; you may already be familiar with the fact that many dish satellite antennas (and also telescope mirrors) are parabolic, since a parabola – not a sphere – focuses incoming parallel rays to a single focal point.

A spherical lens thus tends to bend light too much if it is a positive lens, or too little if it is a negative (diverging) lens. Because a parabola and a sphere are good approximations of each other over a small range, the rays near the axis focus correctly. However, the rays that hit the outer edges of the lens will be bent too much, and come to a focus too close to the lens, as shown in Figure 5.

[1] Technically, longitudinal chromatic aberration.

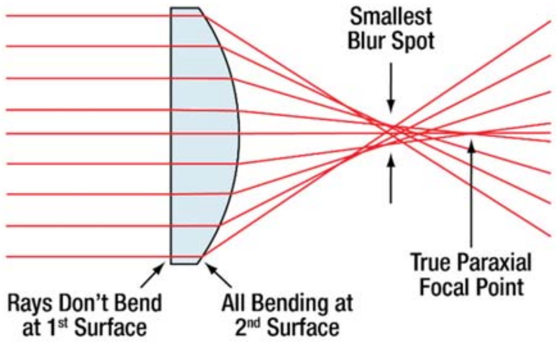

Smallest Blur Spot
True Paraxial Focal Point
Rays Don't Bend at 1st Surface
All Bending at 2nd Surface

Figure 5: Spherical aberration in a plano-convex (positive) lens. Note that the rays closest to the axis ("paraxial rays") focus farthest from the lens at the "correct" focal point. Rays farther from the axis bend too far, and thus focus too close to the lens. The sum total of all the rays produces a confused bunch of rays crossing the axis at different points; the position where the blur spot of rays is smallest is thus technically known as the "circle of least confusion."

The result of this is that the rays do not focus well. The smallest spot they form is larger than theoretically necessary, and thus the image will not be as sharp as it could be. One way to improve the performance of a lens system is to spread the bending of the rays over more surfaces. Note that in Figure 5 the rays coming in from the left hit the flat side of the lens at normal incidence, i.e. at a perpendicular angle, and thus do not bend at all. Consequently, the rays bend at the second lens surface. If one simply reverses the lens, the horizontal rays will hit the curved surface first, so that they bend there, and then bend again when they hit the subsequent planar surface. Lens performance improves, as shown in Figure 6.

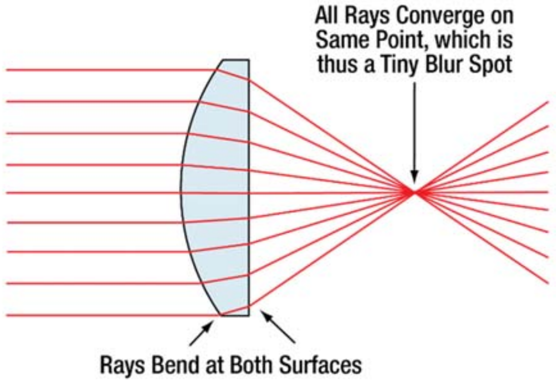

All Rays Converge on Same Point, which is thus a Tiny Blur Spot
Rays Bend at Both Surfaces

Figure 6: Proper plano-convex lens orientation results in much less spherical aberration, and thus better imaging performance.

It is beyond the scope of this class to cover the details of lens design, but it is possible to get a sense for why this splitting of the ray bending over multiple surfaces is helpful. The simple theory of lens behavior is based on the approximation that sin(θ) ≃ θ, i.e. that the system behaves in a linear way. Of course, sin(θ) does *not* actually equal θ, but rather:

Equation 1: $$\sin(\boldsymbol{\theta}) = \boldsymbol{\theta} - \frac{\boldsymbol{\theta}^3}{3!} + \frac{\boldsymbol{\theta}^5}{5!} - \frac{\boldsymbol{\theta}^7}{7!} + \frac{\boldsymbol{\theta}^9}{9!} - \cdots$$

Roughly speaking, the error in the assumption that sin(θ) = θ is given by the next term in the series, $\frac{\boldsymbol{\theta}^3}{3!}$. If we were to split the bending angle θ for a ray equally onto two surfaces, at each surface θ would be half its total value, and the correction, scaling at θ^3, would be one-eighth of its original value, with two surfaces bringing the total error to ¼ of the single-surface value. Flipping the lens around can reduce the error by a factor of four! It may not surprise you to learn that sophisticated lenses can have many, many surfaces (a high NA microscope objective might have about 15!) over which the ray bending is split. For now, the important take-home is this:

> **Always place the more curved side of a lens facing the more parallel rays; conversely, always place the less curved side of a lens facing the most converging rays. For a plano-convex lens, that means the flat side should face whatever is closest, whether it is the object or image plane. This will tend to split the ray bending most evenly.**

For high-NA microscope objectives, using the wrong thickness of coverslip, or focusing deep into a sample of different index of refraction can also cause significant spherical aberration. If you really need high resolution, be careful in your choice of coverslip thickness and immersion medium (for instance, the entire point of using silicone-immersion objectives is to avoid spherical aberration when focusing deep into an aqueous buffer such as salt water).

Lens shape is only one of the possible things to affect imaging; the material the lens is made of can also have a significant impact. These days most glass is quite uniform in density, but that does not mean behavior is fully uniform: typically the speed of light in glass varies for different colors (wavelengths) of light. The index of refraction, n, of a material is defined as

Equation 2: $$\boldsymbol{n} = \frac{\boldsymbol{c}}{\boldsymbol{v}}$$

Since nothing goes faster than light in a vacuum, at v = c, n ≥ 1 for all practical purposes. For water, n = 1.33, so the speed of light is only ¾ as fast in water as it is in vacuum.

The fact that n = n(λ), i.e. the index n varies with wavelength, starts to matter a lot when one notes the formula for the focal length of a lens:

Equation 3: $$\boldsymbol{\frac{1}{f} = (n - 1)\left[\frac{1}{R_1} - \frac{1}{R_2}\right]}$$ **(do not memorize this)**

Clearly the focal length depends on the index, n, and thus we expect different colors of light to focus at different distances from a lens.

One can compensate for this effect by using multiple lenses made of different glasses, such that, say, an increase in focal length in the red due to one lens is canceled by a decrease in red focal length due to a different glass and lens. Such a set of lenses is called an "achromat," roughly derived from *a-* (no) and *chroma* (color). Conveniently, multiple lenses also have multiple surfaces to spread bending across, so one typically corrects for both chromatic aberration and spherical aberration at the same time.

Even if all aberrations are fully corrected, such that a lens system is "perfect," the image blur – or the system point-spread function, PSF – is not zero, and imaging is thus not perfect. The remaining issue is due to a combination of the wave nature of light and the limited set of angles of light that the lens can collect from the sample. This will be discussed further in the Lab 6 Course Notes, but for now the thing to remember is that there is a lower limit to the size of the blur in any optical system, and that this limit depends on the lens NA – the range of angles of light collected by the lens.

The width of the blur is in fact directly related to the resolution limit of the optical system: just as one could expect based on the overlap of blur spots from two different sample light sources (as in Fig. 4).

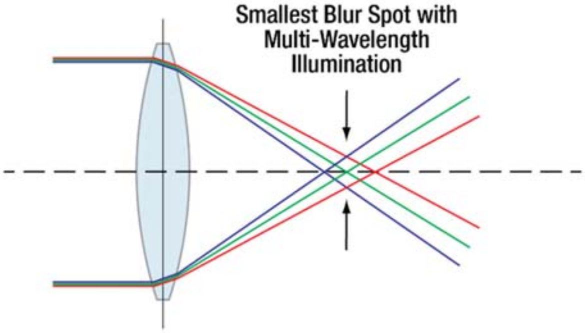


Figure 7: Chromatic aberration, the shift of focal length with wavelength (color)[2]. One could get a small focal spot – i.e. good imaging performance – at any *single* color, but if imaging in multiple colors (e.g. white light) then the smallest blur spot may be large. Note that this is analogous to the "circle of least confusion" from spherical aberration.

The exact resolution of an optical system depends not only on the size of the diffractive blur spot, but also on how the light from different points – and thus in different overlapping blurs – interferes. This is a complex subject, and depends on the coherence of the illumination (or emission). In this lab, you will see the effects of illumination coherence – which tends to make diffractive effects such as fringes around image features – when you use a light source placed far from the sample, providing fairly coherent light, and then see it change when you use a diffuser to illuminate the sample more incoherently. For this lab, seeing the effect (and understanding that there is always some blur) is the important part; the way in which illumination coherence affects resolution will be revisited in more detail as the topic of Lab 7.

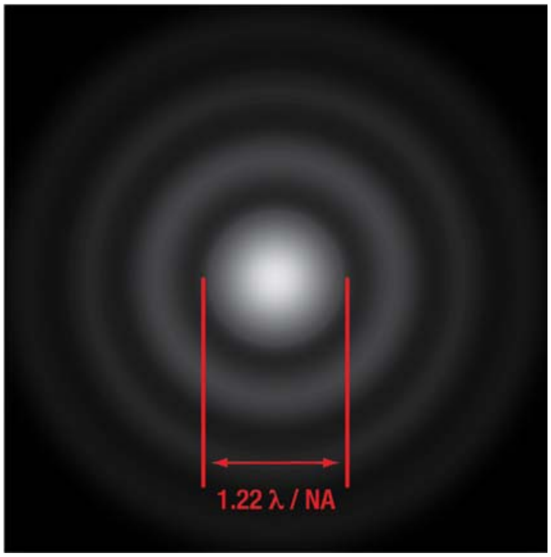


Figure 8: The diffraction-limited blur of an optical system, also known as the Airy disk or the diffraction-limited point spread function (PSF). Note that the width of the PSF depends on both the wavelength of light used and the NA (the range of angles of light collected by the system).

[2] Technically, this is only one of two forms chromatic aberration takes, so-called longitudinal chromatic aberration. A closely related effect is the lateral shift of an image with wavelength, called "lateral color". This is often most noticeable at short wavelengths; one may see purple fringes on one side of an image due to the shift of the shorter wavelength image off to one side of the mid-wavelength image.

Lab 4 Notes:
Köhler Illumination

Optical Microscopy
Course

Lab 4 Course Notes:
Köhler Illumination

Overview

The main goal this lab will be to set up a controlled illumination system for the collection optics you built last lab. Once you have done that, you will have an essentially complete microscope. We will explore the capabilities of this microscope configuration, and then move on to the second major part of the course in which we will exploit these capabilities to generate different modes of image contrast (such as darkfield and phase contrast).

Attention: in order to cover the material you will need to move fairly quickly in lab; to do this, everyone must have read the lab notes and know *ahead of time* what they are going to do that day. You will not be able to keep up if you are reading the notes for the first time during lab.

The effects of illumination on image quality should be quite obvious after Lab 3. For example, using a bare LED resulted in image containing many diffraction fringes, while trying to fix that problem with a diffuser resulted in very dim light and resultant poor images. This is an even bigger problem than it might first seem – the USAF sample is a very high contrast sample: the black lines block 100% of the light. Imagine how much worse things would look if the sample were small cells which absorb or scatter very little light (i.e., have low contrast to begin with).

There are several things we need to control in our microscope illumination:

1. Intensity (brightness)
2. Spectrum (color)
3. Location (what part of the sample is illuminated)
4. Uniformity (even illumination across the sample)
5. Angular distribution (from what angles the sample is illuminated)

Moreover, we want to be able to control these things independently – for example, when we change the intensity we want the color of the light to remain the same.

Unfortunately it would be too inefficient to have you walk through the development of this illumination by guided trial and error. Instead, we will cover the details of the design in the Course Notes and lecture, and you will build the complete system straightaway during lab. You will then explore the details of the system to gain an understanding of why it is put together the way it is.

Definitely Pay Attention: An excellent midterm or final exam might include having you set up a misaligned microscope for proper imaging and illumination.

More than that however, the first thing you should do <u>each time</u> you use a research microscope is to set it up for proper illumination. This typically takes less than five minutes, and is the difference between a novice getting poor images and an expert getting excellent images. At least one company has used it as an interview test when hiring technical personnel. Once you understand the basic design you will always find it easy to get it right, even on an unfamiliar microscope.

Köhler Illumination

The microscope illumination system is quite subtle and sophisticated, but also relatively simple. However, the simplicity is masked at first by the number of unfamiliar parts. In these notes, we will go through the various parts and the reasons for their arrangement in cursory detail. The supplementary reading and lecture will then fill in additional detail.

Let's return to the list of things we need to control in the illumination path:

1. Intensity (brightness)
2. Spectrum (color)
3. Location (what part of the sample is illuminated)
4. Uniformity (even illumination across the sample)
5. Angular distribution (from what angles the sample is illuminated)

The first two items are relatively straightforward and should be familiar. As you may have seen in the lab, for LEDs the light intensity can be adjusted by controlling the current through the LED, and the spectrum does not change much with this variation.

For an incandescent lamp, however, the spectrum does change with the temperature of the bulb filament – as you turn down the power, the filament gets cooler and the spectrum of the light emitted shifts toward the red. In order to control the intensity separately from the spectrum of the light, we will use "neutral density" (effectively, gray) filters to control the intensity without adjusting the power to (and hence the temperature and color of) the filament. Furthermore, since a hot filament emits light at all wavelengths, we can use a color filter to select only the wavelength region we want. As an example, a green color filter (like the one you have) is often used to limit the range of wavelengths one is imaging with and hence also to eliminate chromatic aberration. Green is usually chosen because it is near the peak sensitivity for the human eye. Conveniently many silicon-based cameras (e.g. CCD or CMOS) are also fairly sensitive in the green.

The remaining three properties – location, uniformity, and angular distribution, are more complicated. It makes sense to address these in order:

Location

There are actually two components to location: where the light source is and what part of the sample is illuminated. The location of the light source is simple – it is generally not convenient to put the actual light source up against the sample, especially if the light source is a hot filament. However, as you saw in the last lab, when the light source is far from the sample the light gets quite dim. As noted above, this problem is especially problematic for small or mostly transparent samples, which scatter very little of the incident light. The obvious thing to do is to use a lens to *collect* the light from the filament and focus it toward (or onto) the sample. Not surprisingly, the lens used to do this is called the "collector lens."

The simplest version of this type of limitation involves simply imaging the light source (e.g. the lamp filament) onto the sample, as shown in Figure 1.

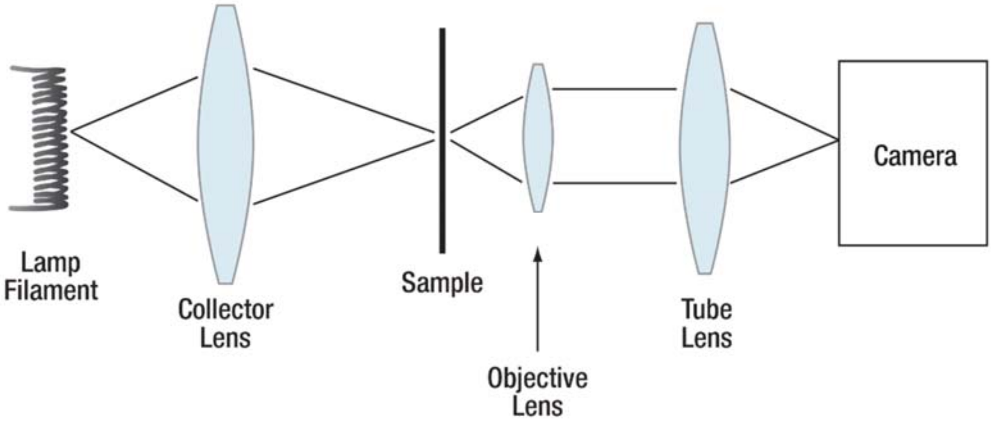


Figure 1: Critical Illumination

Note that we have also shown the collection path – the objective, tube lens, and camera. In addition, the collector lens is positioned such that it images the lamp filament onto the sample. Of course the sample is imaged onto the camera (by the objective and tube lens), so the filament must *also* be imaged onto the camera.

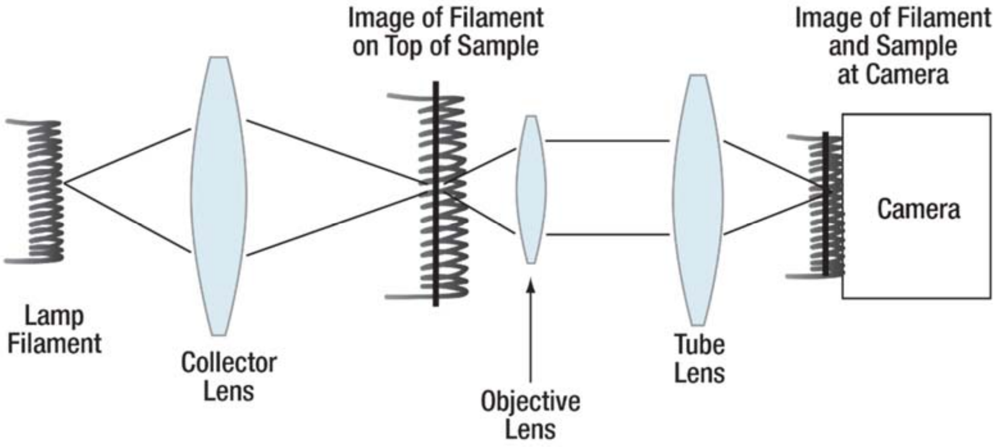


Image of Filament on Top of Sample
Image of Filament and Sample at Camera
Lamp Filament
Collector Lens
Objective Lens
Tube Lens
Camera

Figure 2: Conjugate Planes: The filament, sample, and camera face are all imaged onto each other, i.e. are "conjugate" (though the magnifications may vary).

The term for two planes that are imaged onto each other is “conjugate.” Conjugate planes are very important in microscopy, and we will discuss them often. The most obvious conjugate planes are the sample plane and the camera – if the image is in focus then those two planes are conjugate. This concept can be extended further: for instance, if you look through a microscope eyepiece and see the sample, then your retina and the sample plane are conjugate.

In Figure 2 you can see both some advantages and a disadvantage of critical illumination: on the plus side, we can use the lens to magnify the filament so it illuminates the whole sample, and the lens has allowed us to collect more light from the filament. On the minus side however, all the structure in the filament (the coils) will be immediately visible in the image; the light on the sample is not at all *uniform*.

Before we move on it is worth introducing an additional concept: the field stop. Usually in microscopy the light source is much bigger than the sample you are interested in looking at. In addition, there is a limit to how small one can make the image of the filament – we will return to this, but typically one cannot make the image of the filament much smaller than its actual size. In the last lab, we turned off the room lights to reduce the amount of background scattered light polluting your images. For exactly the same reason, we will not want to illuminate more of the sample than the area we are specifically interested in (i.e. the "field") – any additional illuminated area will simply allow more background scatter into our optical system and that will reduce our image contrast (i.e., make the image look "washed out").

If we cannot do it by demagnifying the size of the filament to illuminate only the area we care about, then we could introduce an iris (or aperture) right before the sample, which serves to limit the area of the sample that gets illuminated. In many cases it is inconvenient to have an iris right next to the sample, but we can use the concept of conjugate planes to solve a problem like this:

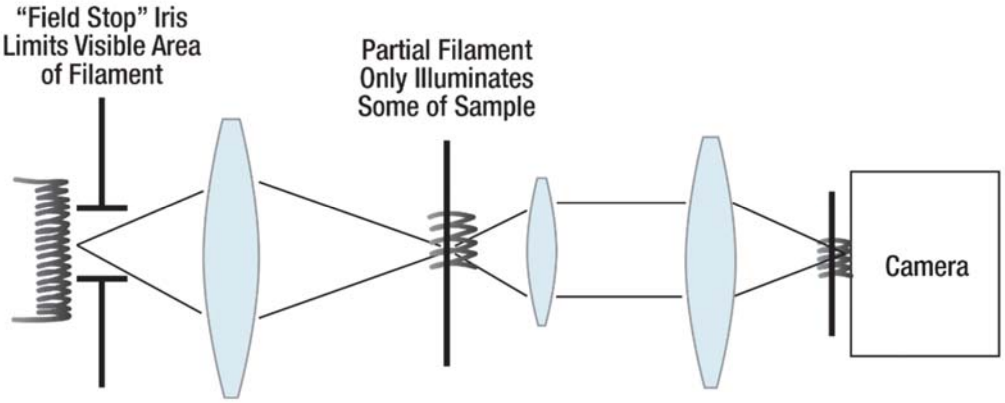


Figure 3: Field stop limits area of sample that gets illuminated. Compare this image to Figure 2.

We could achieve the same effect by putting the field stop next to the filament, next to the sample, or at the camera face. It is worth noting that by putting the field stop at the filament we block any excess light at the source – which is to say, far from the rest of the optical train. If we put the field stop down by the camera then any scattered light would already have polluted the rest of the image.

Important Note: The field stop is in a conjugate plane to the sample and the detector. (This is worth memorizing).

You may have noticed one problem in Figure 3: the filament of an incandescent light is actually enclosed in a lightbulb, so we cannot really put an iris there. One way to solve this would be to use yet another lens to make an image of the filament somewhere where we can put a field stop and then use a lens to image that onto the sample. In practice we will do this a little differently, but the principle is similar.

Important Note: New Terms to Learn (Memorize These):

- **Collector lens: the lens next to a light source that collects the light from it.**
- **Conjugate planes: planes which are images of each other.**
- **Field stop: an aperture which limits the illuminated field on the sample. By necessity, the field stop must be conjugate to both the sample plane and the camera plane.**

Uniformity

The biggest problem with imaging the lamp filament directly onto the sample (known as "critical illumination") is that any nonuniformities in the light source – e.g. the coils of the helical hot filament – are then very apparent in the image of the sample.

In the last lab, you positioned the LED some distance away from the sample in order to get more even illumination – imagine what the sample would have looked like if you slid the LED right up next to it (you could even try this in the next lab). If you wanted the illumination to be even more uniform you could slide the LED (or light bulb) even farther away from the sample. In fact, if the light source were infinitely far away, then the illumination would be extremely uniform. That last sentence should definitely have given you an idea for how we could make the light source more uniform.

➔ Stop and think about that for a moment if it is not already obvious to you.

A good way to make the filament look like it is infinitely far away without having to put it a long distance off (and so losing a lot of intensity) would be to put the filament one focal length behind a lens so that the filament would be focused at infinity on the other side. Even better, if we put the sample exactly one focal length on the *other side* of the lens then each point in the filament would illuminate the entire sample resulting in complete uniformity.

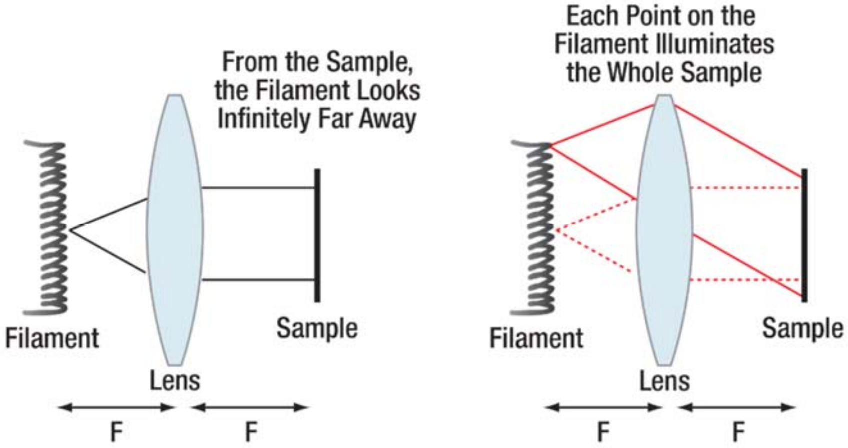

From the Sample, the Filament Looks Infinitely Far Away
Filament
Lens
Sample
F
F
Each Point on the Filament Illuminates the Whole Sample
Filament
Lens
Sample
F
F

Figure 4: Placing the filament in one focal plane of a lens and the sample in the other focal plane generates good illumination uniformity, since each point in the filament illuminates the whole sample, and from the sample position the filament looks "infinitely far away" (or, equivalently, "totally out of focus").

By now it should already be occurring to you that we do not have to put the filament *itself* in the front focal plane of this lens, but rather could (equivalently) put *an image* of the filament in that focal plane.

Notice the new terminology: "front focal plane" and "back focal plane." These are the planes one focal length in front and in back of a lens. Because in microscopy one usually follows the light starting from the illumination source and going step-by-step through all the optics to the detector, "front" is the side of a lens nearest the lamp, and "back" is the side farthest from the lamp. As an example, in the last lab you set up an iris in the back focal plane of the ½" achromat you are using as an objective.

Important Note: New Terminology (Worth Memorizing):

- **Front Focal Plane: The plane one focal length away from a lens and (in our case) *closest to the lamp.***
- **Back Focal Plane: The plane one focal length away from a lens and *farthest from the lamp.***

Angular Distribution

Angular distribution is the last of the things we would want to control about the illumination. Before getting into details of how we might control that, it is worth considering a simple example of why we would care: imagine trying to look at a faint fingerprint on a piece of glass. If you look straight through the glass at a light, the light from the light source will overwhelm any contrast from the fingerprint. However, if instead you look through the glass at something dark, with the light off to the side (out of your immediate field of view), then the fingerprint will be easily visible. Technically, this is called oblique illumination – and "oblique" refers to an angle; in this case the angle between the direction you are looking through the glass and the direction to the light source. By shifting the angle of the illumination, you have gone from being unable to see the fingerprint to being able to see it fairly clearly; that should help provide a sense of why it can be convenient to be able to manipulate the angle at which illumination is hitting the sample.

The manner in which we can manipulate the angular distribution of illumination is hinted at in Figure 5, where in the right-hand image it is apparent that light emitted from different parts of a filament is directed at the sample from different angles. This is a direct consequence of the fact that light collimated on one side of the lens is focused down to a point on the other side – so, conversely, light emitted from a point on a filament (if that filament is in a focal plane) will be collimated on the other side of the lens.

In fact, the height of a point above the axis in one focal plane of the lens corresponds directly to the angle of the parallel bundle of rays in the other focal plane.

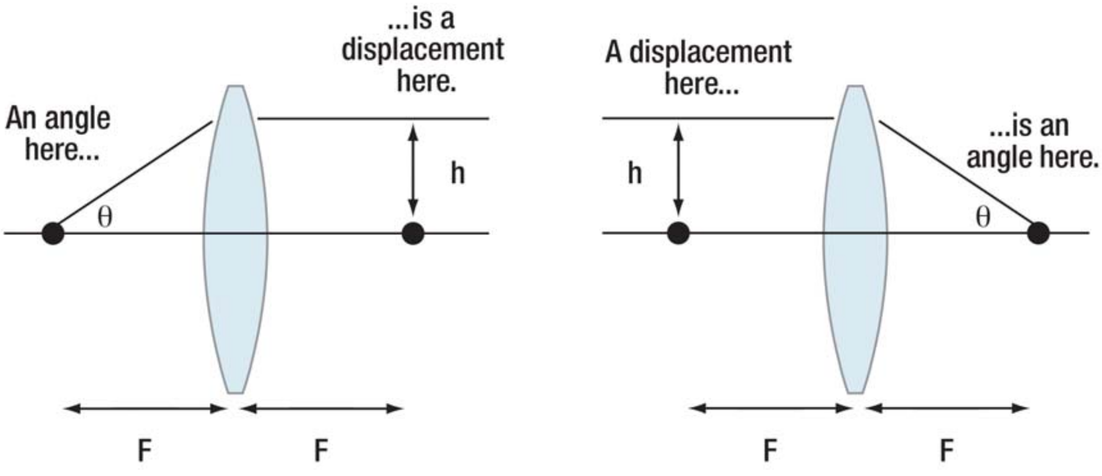


Figure 5: An angle in the front focal plane is a displacement in the back focal plane and vice-versa.

Important Note: In Figure 5, one can see from geometry that $\tan(\theta) = \frac{h}{F}$. Remember from the formula sheet that imaging lens systems are usually carefully designed so that the actual

relationship is given by $\boldsymbol{sin(\theta)} = \frac{h}{F}$. This requires a lot of work (and systems of multiple lenses) but is necessary for proper imaging. We will discuss this further in the middle part of this course; for now, be sure to remember the sin(θ) relationship for microscope lens systems.

Figure 6 shows how we can use an aperture in front of the filament in the front focal plane of a lens to implement control over the angular distribution of the illumination at the sample. Note that since we are using a circular aperture centered on the optical axis, the illumination appears to converge ("condenses") in a conical path onto the sample. As a result, this lens is called the "condenser lens."

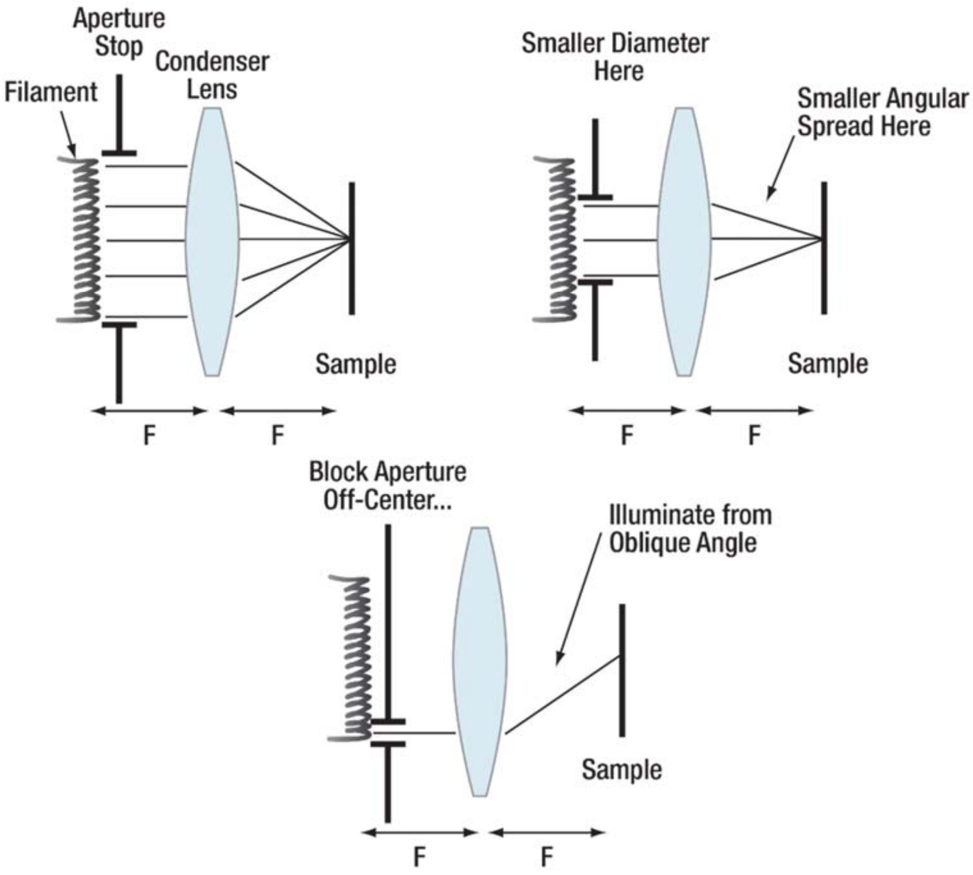

Aperture Stop
Condenser Lens
Filament
Sample
F
F
Smaller Diameter Here
Smaller Angular Spread Here
Sample
F
F
Block Aperture Off-Center...
Illuminate from Oblique Angle
Sample
F
F

Figure 6: Using an aperture in the front focal plane of the condenser lens to control the angular distribution of the illumination on the sample.

The aperture placed in the front focal plane of the condenser lens determines the angle or spread of the rays hitting the sample; in fact, using the relation given earlier, the diameter of the aperture determines the numerical aperture (NA) of the rays hitting the sample. As a result, it is referred to as the "aperture stop."

Important Note: New Terminology (Worth Memorizing):

- **Condenser Lens: The lens which focuses the illumination down onto the sample.**
- **Aperture Stop: The iris which limits the illumination NA (cone of rays) onto the sample.**

Putting it All Together

From the above discussion, we know we want several things:

- To be able to put the lamp some distance from the sample without losing intensity
- A field stop (which needs to be imaged at the sample plane)
- A condenser system that involves an image of the filament in the front focal plane of the lens and an adjustable aperture stop at that same plane, and which focuses the illumination down onto the sample.

There are number of ways to accomplish this; however, typically one wants the total optical path to be reasonably short (~ 30 cm) such that the instrument is not too large, and in addition it is preferable not to require too many lenses (to keep costs down). Since optimizing this is somewhat complicated, and there is an accepted standard for how it is done in typical microscope systems, we will have you implement the usual system directly. The typical configuration is shown in Figure 7.

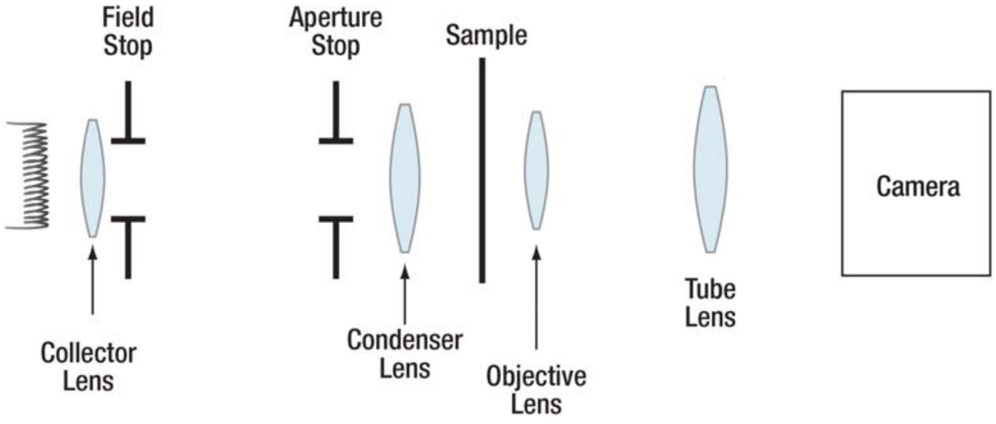
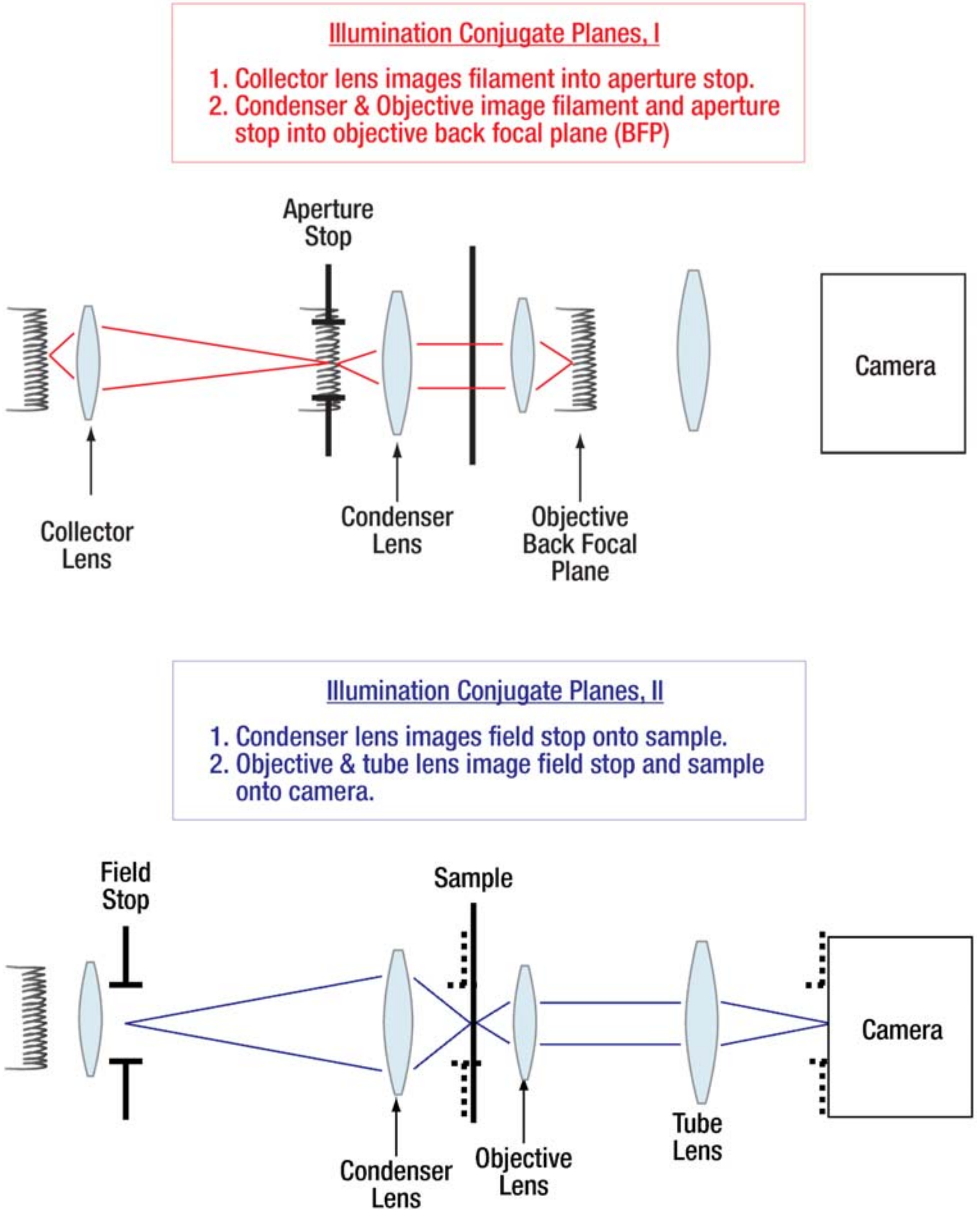


Figure 7: All the parts of the microscope, in order.



Illumination Conjugate Planes, I

1. Collector lens images filament into aperture stop.
2. Condenser & Objective image filament and aperture stop into objective back focal plane (BFP)

Illumination Conjugate Planes, II

1. Condenser lens images field stop onto sample.
2. Objective & tube lens image field stop and sample onto camera.

Figure 8: Conjugate planes:

Which lens is imaging what to where? Note that:

a. The filament is *not* imaged onto the sample or camera – so illumination is uniform.

b. The field stop *is* imaged onto the sample and camera, limiting the opportunity for stray/scattered light to pollute the image.

c. **The aperture stop (one focal length from the condenser lens, and so *not* imaged onto the sample) limits the size of the filament image in the condenser front focal plane, which in turn sets the angle of the cone of rays focused onto the sample (illumination NA).**

d. The iris in the objective back focal plane (not shown, but just like you built in the last lab) limits the cone of rays *collected* from the sample (i.e., limits the objective NA).

An important thing to note here is that in order for the condenser to image the field stop onto the sample, it is *impossible* for the sample to be in the back focal plane of the condenser (i.e., one focal length away from the condenser). Ideally that is where we would like it to be; in practice we can get fairly close if we place the field stop as far as possible from the condenser (so the object distance S_o is large).

There is a certain amount of additional detail in choosing the focal lengths, diameters, and distances between the lenses; for this lab, you will simply copy the layout provided in the Lab Notes. As the course goes on we will discuss some of the reasons behind the choices involved in this implementation.

Because it may be useful to have the steps laid out in terms of the specific parts in your microscope (rather than a research microscope), the sequence for setting up proper illumination is laid out below.

Setting Up Köhler Illumination when Building a Microscope

1. With your camera focused at infinity, put the camera approximately one tube lens focal length behind the objective (the best distance depends on lens design; if in doubt, closer than a focal length is better than farther, to avoid clipping rays on the edge of the lens, called "vignetting").
2. Turn on your lamp so that there is some illumination on the sample.
3. Position a sample in front of the objective such that it is in focus on your camera (a marker drawn on a slide works fine; even a fingerprint on a slide can work well).
4. With the aperture stop and field stop wide open, position the condenser roughly one focal length from the sample.
5. Position the lamp, collector lens, and field stop at the far end of the optical rail.
   a. Make sure everything is at the same height as the tube lens; this is hardest to do with the lamp, so adjust lamp height as best you can.
6. Adjust the collector lens position so that the filament is in focus at the aperture stop
   a. Adjust the lamp height to get the filament vertically centered on the aperture stop.
7. Position the field stop conveniently close to the collector.
8. Close down the field stop and adjust the condenser position along the optical axis in order to bring the field stop into focus at the sample (and hence on the camera image).
9. Reposition the collector if necessary to bring the filament back into focus at the aperture stop.
10. Open the field stop as much as desired – usually so no more area is illuminated than what is being viewed.
11. Close down the aperture stop until contrast is optimized (this is usually ~ 70% of the objective back aperture; we will return to this later).

Note: Filters (e.g. ND filters, color filters) or diffusers should be positioned *away* from the field stop so that they are not in focus at the detector – that way the diffuser texture and/or scratches or dust on the filters will not be visible.

Setting Up Köhler Illumination on a Standard Research Microscope

This is usually fast (~1-2 minutes) once you are experienced, but slow the first time.

- ➔ **The information below does not apply to this course; rather, it is to help you make the connections between the material in this course and what you are doing when you set up illumination on a laboratory microscope.**
- ➔ The microscope manuals usually have a detailed 1-3 page procedure that is easy to follow, with a diagram of where the relevant parts are on their microscope. Finding the manual (e.g. online) and following it can be the easiest way to get started.

1. Turn on the lamp.
2. Put in a low-magnification objective (20X or less).
3. Open the field stop and aperture stop fully.
    a. Remove any phase rings, polarizers, etc. from the light path (sometimes this requires rotating the condenser turret).
    b. You can tell when you have got this right because you can see light from the condenser hitting the sample.
4. Position a sample on the sample stage, so the light is hitting it.
    a. Marker drawn on a slide or coverslip works fine; even a fingerprint on a slide can work well.
    b. If you are using a high-NA objective, the sample may need to be on a coverslip, not a slide – high NA objectives often cannot focus all the way through a slide.
5. Adjust the condenser height so the light spot on the sample is as small as possible – this will be (roughly) the correct height for the condenser.
    a. If the condenser is hard to move, check to see if someone locked down the positioning lock for it (some microscopes have these). Similarly, if you cannot get it to come down far enough, check that there is not a block preventing it from moving farther down (some microscopes have these too) – if it is a problem, release that.
6. Adjusting the lamp filament and collector: Usually you cannot do this.
    a. If you change a lamp, sometimes you do need to position the filament – in that case, check the manual for the particular scope to see where the screws that adjust filament (or arc) position are.
7. Use ND filters or reduce the lamp intensity so you do not hurt your eyes when focusing.
8. Focus on the sample.
    a. You may want to do this by looking at the sample by eye (not through the eyepieces, but staring right at it) while bringing the objective up – that way it's far less likely that you'll accidentally ram the very expensive objective into the sample. Then, look through the eyepieces, or at the monitor if using a camera, and adjust the focus **by moving the objective AWAY** from the sample, until things come into focus.

   b. For low-NA objectives (magnification of 20X or lower, usually), the working distance can be quite long – sometimes you need to lower the objective a long way before things come into focus. Eventually you get a sense for this.

9. Close down the field stop to a small diameter and adjust the condenser height until the field stop is as sharply focused in the image as you can get it.
   a. Especially with high-NA objectives, the field stop may not come into very sharp focus; just do the best you can.
   b. If you cannot see the sample, open up the field stop until you can see the edge, and focus on that.
      i. If there are centering knobs on the condenser, adjust them so the field stop is centered in the field of view.
10. Flip in the Bertrand lens, or take off the eyepiece (usually they pull straight out of the tube) and look in to see the objective back focal plane.
    a. Adjust the aperture stop so it is ~ 70% of the size of the objective back aperture.
       i. The objective back aperture will look like a disk of light; close the aperture stop until it makes that disk about 2/3 as wide as it was when fully open).
    b. Note: if you are setting up Phase, you need the aperture stop all the way open! Then put in the phase ring and make sure the ring of light is centered over the phase ring in the objective back focal plane (this is a bit darker and easy to see). If any light is not hitting the objective phase ring, you need to adjust the position of the phase rings; this cannot be done using the condenser adjusters, so do not try those. Get the microscope manual; usually there are little setscrews you can adjust on the phase ring housing, easy to confuse with the screws that hold the housing on.
    c. Sometimes there is a focus knob on the Bertrand lens to allow you to get the aperture stop in best focus; this is a convenience but not necessary, since normally the Bertrand lens is not in the optical path.
11. If you will be doing microscopy by eye (and not with a camera), adjust the eyepieces for your eyes.
    a. One eyepiece usually has an adjustment (the "Diopter adjustment" ring on it. Use a business card to block your vision through that eyepiece (closing one eye perturbs the muscles around the other eye, so try use a card to block your vision instead).
    b. Looking with the unblocked eye, adjust the focus knob so the sample is sharpest.
       i. The best way to do this is to move the objective *a bit too far* from the sample, and to bring the sample slowly into focus. This way your eye will be focused on infinity (its most natural / comfortable state) when the image comes into focus. This also guarantees that the sample will be at the plane of best aberration adjustment for the objective when you are looking at it.
          1. To avoid ramming the objective into the sample, start by moving the objective close to the sample while watching the objective itself. Then while using the eyepieces moving the objective away from the sample until things come into focus, then go a bit farther away until you lose focus, wait for your eye to relax, then move the objective back to focus.

    ii. If you focus by bringing the objective *away* from the sample, then your eye can adjust to focus when the objective is still too close, and the sample will not be in the plane where the objective is best corrected, and your eye will be stressed and you will get a headache after long viewing.

c. DO NOT touch the focus knob! Now block the other eye, and adjust the Diopter Ring on the eyepiece until the sample is again in sharpest focus.

    i. As with focusing, there is a best direction to go: adjust the diopter ring to the highest "plus" position (e.g. +4), usually fully clockwise / all the way in, and then rotate it back out until the sample hits sharpest focus.

d. Adjust the spacing between the eyepieces (if they move at all, you can grab them and adjust the spacing by pushing/pulling) until they are best for your eyes.

    i. Adjust them a bit too wide, then, while looking through the eyepieces with both eyes, slowly bring the eyepieces closer together. At a certain point, the two separate disks of light you see will merge and you will see them become one. Keep going until they look like a single disk of light, but no farther. If viewing is uncomfortable, obviously adjust further.

Notes:

1. Adjusting the eyepieces is really worth it – it is like going from a normal theater to 3-D IMAX. Very impressive and cool when properly set up, and it usually takes only a minute or less once you know how to do it.
2. A good / easy sample for setting up complicated contrast methods (phase or DIC) is cheek epithelial cells: use a (clean!) pipette tip to gently scrape the inside of your cheek, and then touch the tip to a slide (try to get a little spit onto the slide too). Put on a coverslip, and seal the edges with wax or nail polish if you have it to keep the small volume of liquid from drying out. I like to put a marker stripe on the slide before putting on the cells, to have something easy to focus on.
3. Setting the eyepiece diopter can sometimes be a bit tricky; here are a few tips:

    a. If you have astigmatism, keep your glasses on – the eyepieces will not adjust for that. In general, if your glasses prescription is very powerful, chances are you should keep them on.

    b. [Rare issue – listed here only for completeness of coverage] Sometimes microscopes have diopter adjustments on **both** eyepieces, what then? Use the fact that depth of focus is smaller for a high-NA objective than a low-NA one:

        i. Set the diopter adjustments to the middle of their range.

        ii. Focus on the sample using a moderate NA (say, 20X 0.4 NA) objective.

        iii. Switch to a lower NA objective (say, 10X or 5X).

        iv. Adjust the diopters to the highest setting (usually the twisted all the way in / clockwise).

        v. Blocking one eye at a time, rotate the diopter rings back out until focus is best. Do this for each eye.

        vi. It is probably best to iterate this procedure by going back to the higher-NA objective, refocus using the focus knob, then repeating the steps above.

THORLABS

Lab 5 Notes: Köhler, Conjugate Planes, and Darkfield Imaging

Optical Microscopy Course

Lab 5 Course Notes:
Köhler, Conjugate Planes, and Darkfield Imaging

Transparent samples have the problem that they are transparent, so they do not absorb any light, and if they are very small (like a cell) they do not scatter much light either. As a result, what you will see when imaging a transparent sample is your bright illumination with very tiny modulations due to the small amount of scattering from the sample. Tiny changes in a bright background are very hard to see.

It would obviously be convenient if we could remove the bright background so that all we were seeing at the camera was the light which had interacted with the sample (i.e. been scattered). Then everything we saw would be "signal," with no background, thus providing excellent contrast.

Not surprisingly, our carefully set up illumination offers us some options for doing this. Remember that by placing an image of the filament one focal length away from the condenser, we have arranged things so that each point on the filament is illuminating the whole sample – in a sense, this is what it means to be "totally out of focus":

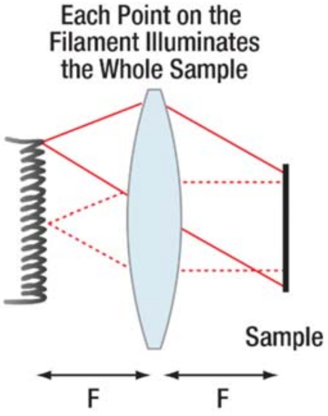


Figure 1: In Köhler illumination, each point on the filament illuminates the whole sample.

Furthermore, each point on the image of the filament in the aperture stop is imaged to a ***conjugate*** point in the objective back focal plane.

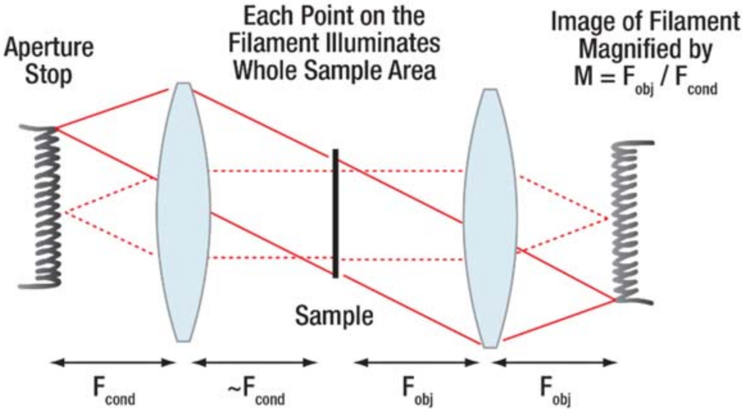

Aperture Stop
Each Point on the Filament Illuminates Whole Sample Area
Image of Filament Magnified by $M = F_{obj} / F_{cond}$
Sample
F_{cond}
$\sim F_{cond}$
F_{obj}
F_{obj}

Figure 2: Objective back focal plane has conjugate image of aperture stop.

Let's examine the implication of these conjugate images further: in particular, if I close down the aperture stop to a very small hole, then what happens to the image in the back focal point of the objective?

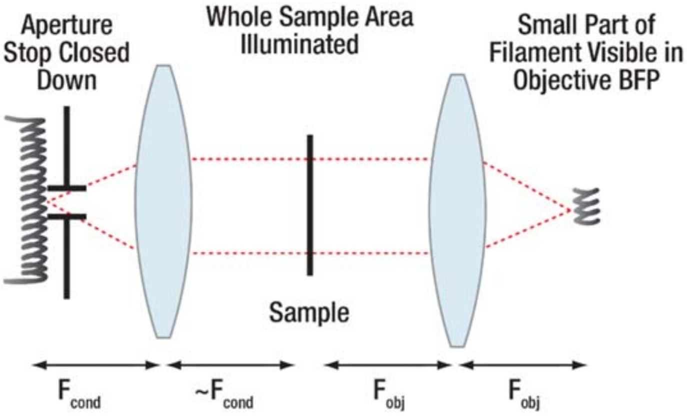

Aperture Stop Closed Down
Whole Sample Area Illuminated
Small Part of Filament Visible in Objective BFP
Sample
F_{cond}
$\sim F_{cond}$
F_{obj}
F_{obj}

Figure 3: Closing down the aperture stop.

If I now place a little block in the objective back focal plane, just downstream of the image of the filament, what would I see on the camera?

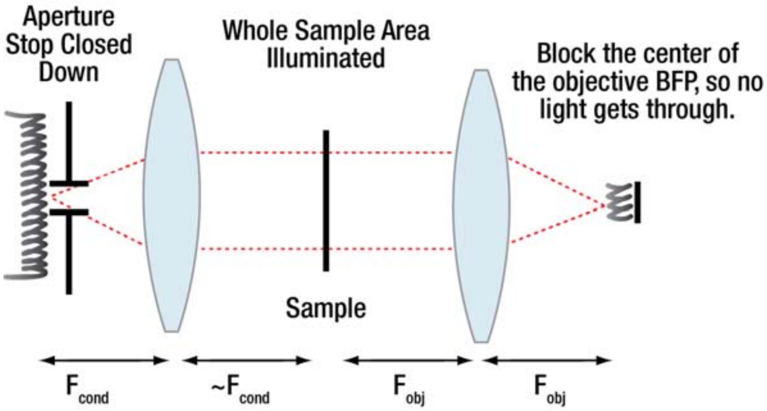

Aperture Stop Closed Down
Whole Sample Area Illuminated
Block the center of the objective BFP, so no light gets through.
Sample
F_{cond}
$\sim F_{cond}$
F_{obj}
F_{obj}

Figure 4: The method of central dark ground, a block at the center of the objective BFP, blocks all the illumination light.

Another way of thinking of this is to consider the shapes of the apertures in these two planes, as shown in Figure 5.

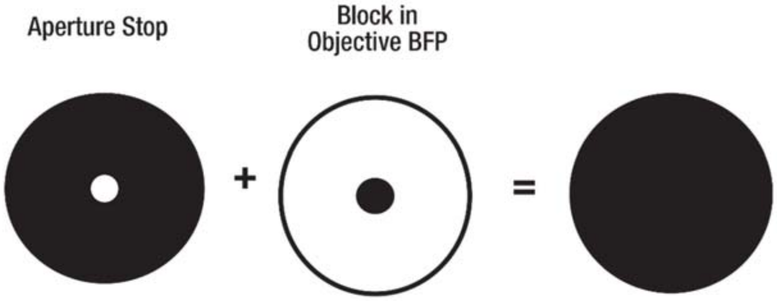


Figure 5: Superposition of images of the aperture stop in the condenser front focal plane and the objective back focal plane.

Looking at these figures, it should make sense that as long as I have no sample at the sample plane (which means that the light goes straight through, undeviated) then no light will get to the camera. So what happens if I ***do*** have a sample?

If the sample is transparent, then one can just imagine it as being a sort of little lens as it bends some of the light off of its original path. For small/thin samples, the amount of light which is bent is very small – so small that one can pretend the original illumination is essentially unchanged, and that the sample is essentially a little source of light superposed with the original illumination. It turns out that one can think of a sample which absorbs light in essentially the same fashion; however in this case, the light which is superposed is 180° out of phase with the original illumination, so that at the image plane the illumination and the light from the sample cancel, and consequently the sample features in the image look dark.

In both cases then, we can think about the illumination *separately* from the light scattered by the sample, which is very convenient! Let's do that in the case of the apertures discussed above.

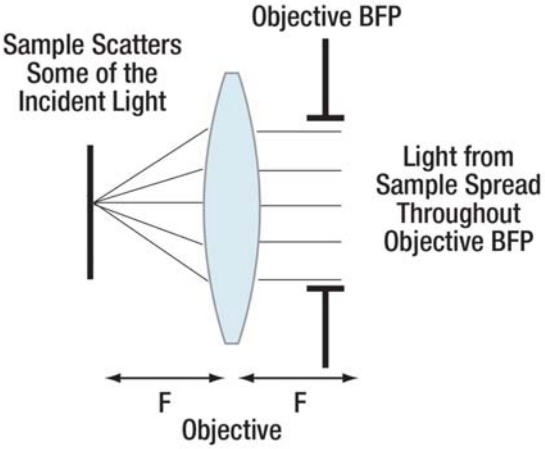

Objective BFP
Sample Scatters Some of the Incident Light
Light from Sample Spread Throughout Objective BFP
F
F
Objective

Figure 6: Light scattered by the sample becomes distributed throughout the objective BFP.

From the previous several labs, you will have seen that as you decrease the size of the aperture and the objective BFP (which is to say as you decrease the NA), the resolution of your image declines.

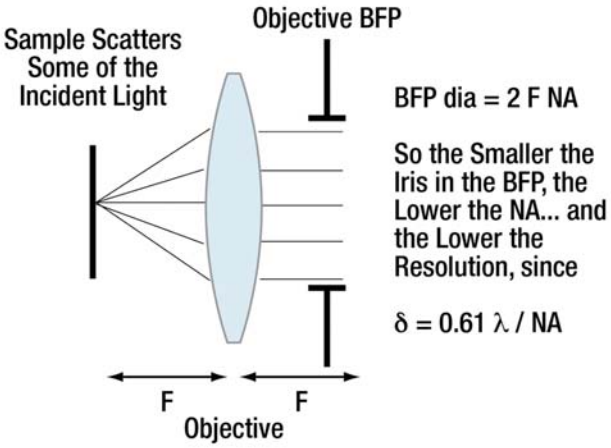

Objective BFP
Sample Scatters Some of the Incident Light
BFP dia = 2 F NA
So the Smaller the Iris in the BFP, the Lower the NA... and the Lower the Resolution, since
δ = 0.61 λ / NA
F
F
Objective

Figure 7: Resolution, objective BFP size, and light scattered from the sample.

This should tip you off that there is a relationship between location in the objective back focal plane and information about the sample – the light at the edges of the back focal plane apparently contains the higher resolution part of the image. For later, notice that the light at the edges of the back focal plane is also the light which is scattered at a higher angle by the sample; there must be some link between higher angle scattering and higher resolution.

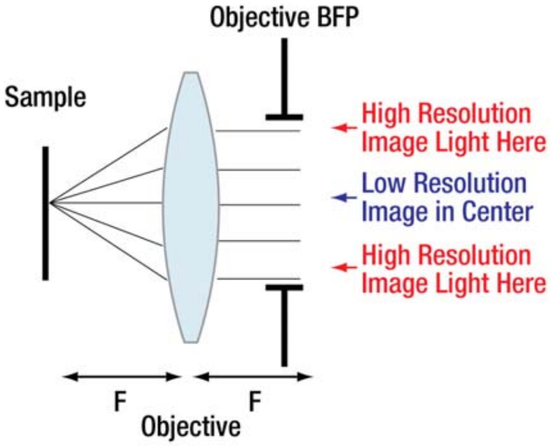

Objective BFP
Sample
High Resolution Image Light Here
Low Resolution Image in Center
High Resolution Image Light Here
F
F
Objective

Figure 8: Sample resolution information is spread in different parts of the objective BFP.

For the upcoming lab, you should think about the following: if I take an image with the objective back focal plane diameter set at 6 mm, take a second image with the BFP diameter set at 1 mm, and then *subtract* the images, is there a mask (or aperture shape) that I could put in the objective BFP which would give an equivalent effect?

THORLABS

Lab 6 Notes:

Abbe Theory of Image Formation (I)

Optical Microscopy Course

Lab 6 Course Notes:
The Abbe Theory of Image Formation (I)

Overview

The Abbe theory of image formation provides a powerful and frankly quite beautiful way to understand optical imaging. Unfortunately, it is almost never taught – not because it is hard, but because virtually all optics classes at the undergraduate level emphasize geometrical optics (which is what we have been doing for the past several labs). Partly for this reason, the ability to discuss imaging from an Abbe perspective tends to be a mark of an "optical expert."

Abbe theory is usually introduced as a special case in a more general Fourier optics class, which you probably have not had. That is no accident: part of the reason for this class is that we believe the concepts can be relatively simply introduced (and used) without grinding through the math. Even more than that, in many Fourier transform classes, students can do well and yet have no real idea how to apply the concepts to a physical system – and often have no clear idea of exactly what they have been doing for the past semester – making the courses poor preparation for an optics class.

This is not to say that an explicit Fourier approach to optics is not useful or fun; both of us find looking at systems from that perspective to be very enlightening. We will try and make the correspondence between the physical situations we demonstrate and the concepts in Fourier theory explicit as we go along, and your instructor can help you delve further into the mathematical detail with you if you want. We will repeat, however, that this class will not be approaching optics from an overly mathematical perspective; algebra and trigonometry will be enough for nearly everything.

On a separate note, often lost in discussions of the technical prowess of historical (and current) figures is their quality as human beings. One measure of a person's quality (in our experience usually quite accurate) is how someone treats their subordinates. In this respect, Abbe was perhaps even more remarkable as a person than he was as a scientist: he introduced for his workers an eight hour workday (in fact, Abbe *originated* the 8-hour day!), paid vacation, paid sick leave, profit sharing, and hiring without respect to race, religion, or political affiliation. Furthermore, the bulk of profits from the company (Carl Zeiss, Inc.) were rolled into a perpetual Foundation rather than merely enriching their heirs. Virtually all of this was unheard of in pre-1900's Europe and even now it is hardly universal.

References

1. Previous Course Notes.
2. To help you think about what you will see in this lab, the videos of a lecture on Abbe Theory by Dr. Peter Evennett are well worth watching; links can be found on the *Reference Links* tab at www.thorlabs.com/OMC. Although we learned about these videos long after we had started planning this course, we could not hope to do better in exploring Abbe theory. You will notice that we borrowed his idea of having Abbe's image in a conjugate plane directly from this lecture. Do not worry about the discussion of phase contrast, or other things which we have not covered yet.
3. Inoué's book, Video Microscopy, 2nd ed., is probably your best resource and contains a wealth of information; take a look at Ch 2, especially section 2.4, "Image resolution and wave optics."
4. For those of you interested in the explicit Fourier theory behind what we will be discussing, without doubt the best reference is Goodman's book, Introduction to Fourier optics, 4th edition. See section 3.10 ("Angular spectrum of plane waves") and Chapters 6 ("Wave optics analysis of coherent optical systems") and 7 ("Frequency analysis of optical imaging systems").

5. An interesting historical article on Abbe is by Volkmann, "Ernst Abbe and his work," Applied Optics, 1966.

What We Have Already Seen

As we will see shortly, the Abbe theory of imaging is intimately linked to resolution, contrast, and the condenser front focal plane (also known as the aperture stop, abbreviated here AS) and objective back focal plane (BFP). Let's review some of what we already know about these things:

1. **Objective NA / Objective back focal plane**
    a. **Opening the iris in the objective BFP increases resolution.**
        i. The size of the aperture of the objective back focal plane determines the NA of the lens, and hence its resolution (ignoring aberrations).
    b. **High-resolution image information seems to be in the outer part of the objective BFP.**
        i. Blocking the center of the BFP is equivalent to subtracting a low resolution image from the total image (per your central darkfield experiment in the last lab).
        ii. (b) is actually a restatement of (a), but it is worth considering from both viewpoints.
    c. **Resolution of horizontal and vertical edges is contained in different planes of the BFP.**
        i. Changing the axes of your oblique darkfield set up resulted in resolution of different parts of the 4-bar resolution targets.

2. **Condenser NA / Condenser aperture stop**
    a. **Opening the AS (the iris in the condenser FFP) increases resolution.**
        i. You have the formula and have seen the effect in lab, but we have not discussed why this happens.
    b. **Diffraction/interference fringes: visibility depends on condenser NA.**
        i. The LED actually qualifies as a very low NA condenser – a 0.1 × 0.1 mm square emitter 7" away from the sample gives an effective NA of ~ 0.05 mm / 175 mm ~ 0.0003.
        ii. With the lamp and condenser, the smallest condenser aperture (diameter ~ 1 mm, focal length 50 mm) results in an illumination NA ~ 0.01.
        iii. The larger condenser aperture (diameter ~ 10 mm, f = 50 mm) provides NA ~ 0.1.
        iv. Use of a diffuser in Lab 3 effectively also increased the NA; a ~ 1" spot on piece of paper, maybe 5" away from the sample amounts to an NA ~ 0.1, and resulted in no visible fringes.
        v. Fringes are very visible at illumination NA 0.003, somewhat so at NA 0.01, and mostly invisible by NA 0.2.

There is one further thing to notice about these observations: each modification of the aperture stop or back focal plane of the objective results in an effect on the image that is consistent across the entire sample. Thus, the resolution and contrast will be (in theory) identical for all points in the image, desirable for consistency. That this would be true should not necessarily be surprising, since we have already mentioned in class (and in the notes) how the location and size of apertures in a focal plane opposite the sample (i.e., the aperture stop or objective BFP) determine the angular distribution of light incident on, or light collected from, *all points* in the sample.

Before continuing, we will summarize all this in pictures.

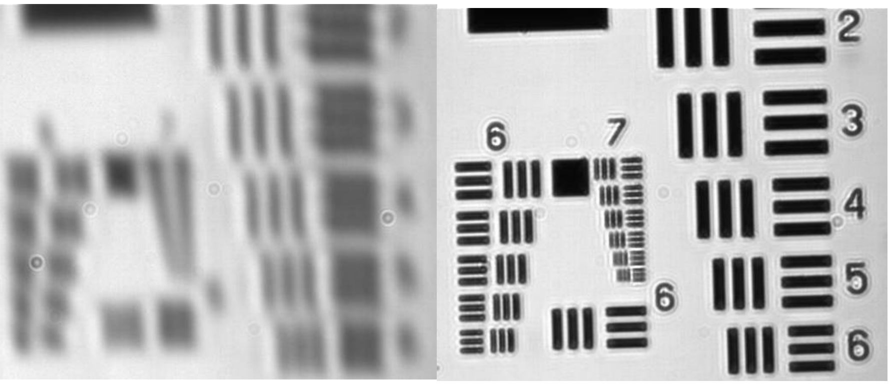


Figure 1: Objective BFP aperture 1 mm diameter (left; NA ~ 0.02) and ~ 8 mm diameter (right; NA ~ 0.15). Note the increase in resolution as the BFP aperture diameter increases; this implies that the high-resolution image information lies near the outer edge of the BFP (at least in the case of axial illumination).

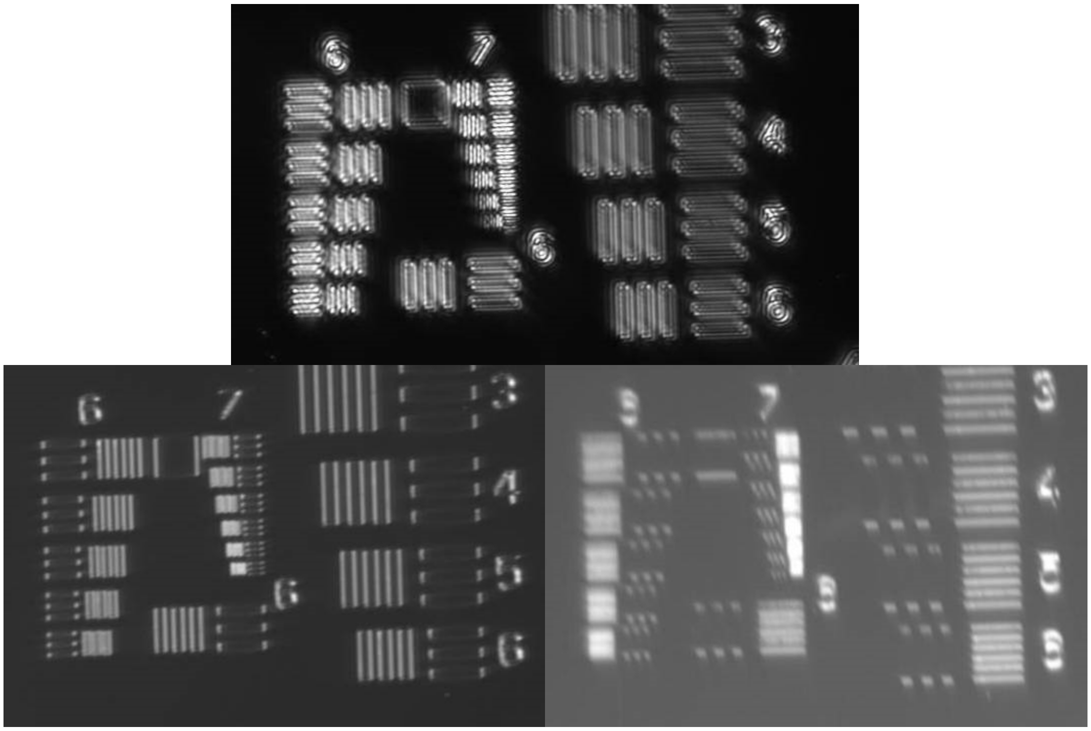


Figure 2: (Top) Central darkfield where all edge detail is present. Note effect of vertical- (lower left) and horizontal- (lower right) oblique darkfield. Note that occluding the focal planes along the vertical axis removes horizontal detail from the sample, while occluding on the horizontal axis removes vertical detail.

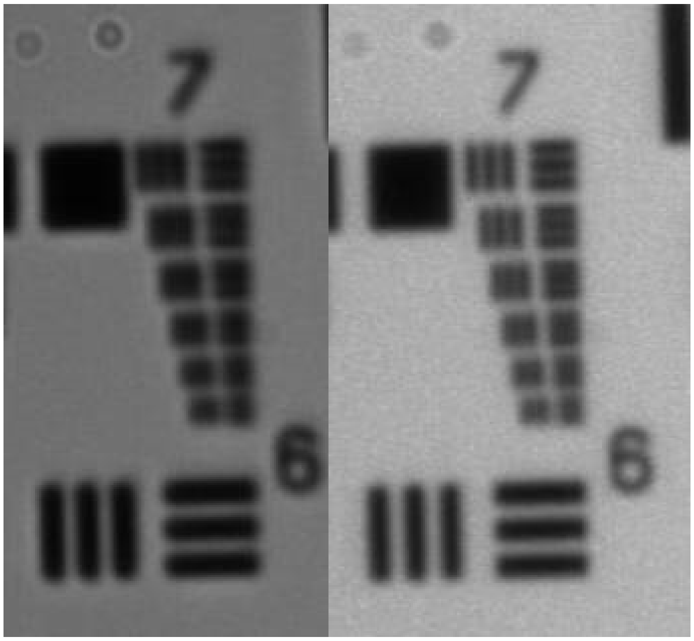


Figure 3: Condenser aperture closed (left) and wide open (right); note increase in resolution by ~ 2 elements. Exact improvement depends on both objective and condenser NA.

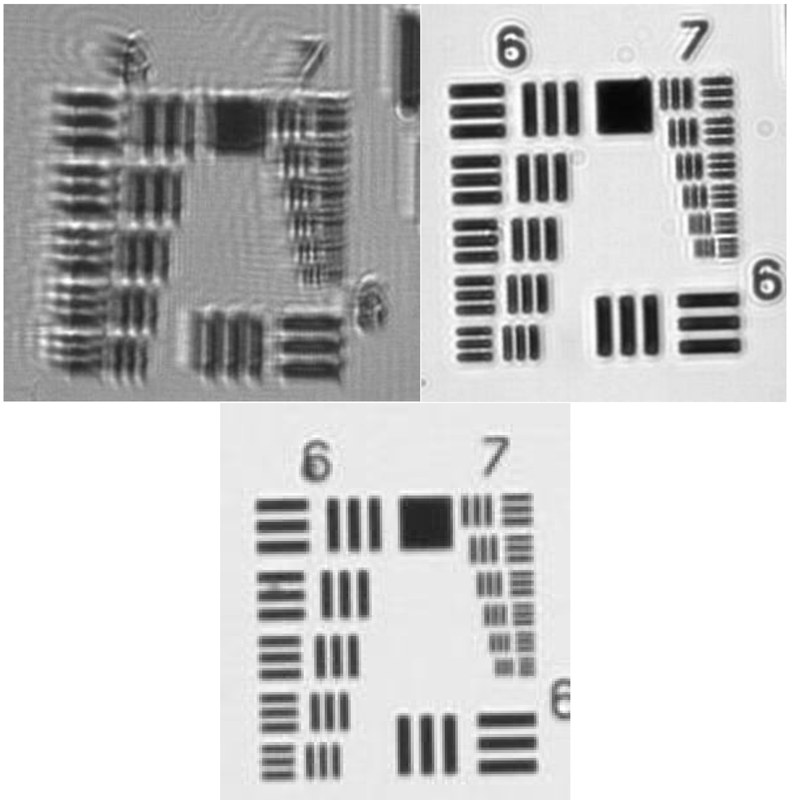


Figure 4: USAF target illuminated by an LED (upper left, NA ~ 0.003), a small condenser aperture (upper, right, NA ~ 0.01), and a diffuser (bottom, NA ~ 0.1). Note decreasing visibility of fringes.

Abbe Theory: Plane Waves; Another Way of Looking at the Sample

So far, we have not really talked about what happens when light interacts with a sample. Now would be a good time to start: for simplicity we will assume we are illuminating the sample with a plane wave on axis; later we will generalize this further.

To start with, how does one get a plane wave on axis? From previous Course Notes and from some of the previous Lab Notes you can probably guess:

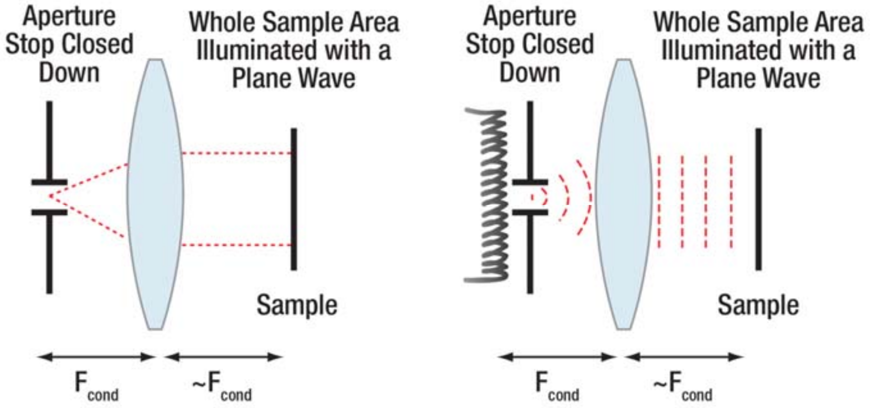

Aperture Stop Closed Down
Whole Sample Area Illuminated with a Plane Wave
Sample
F_{cond}
$\sim F_{cond}$
Aperture Stop Closed Down
Whole Sample Area Illuminated with a Plane Wave
Sample
F_{cond}
$\sim F_{cond}$

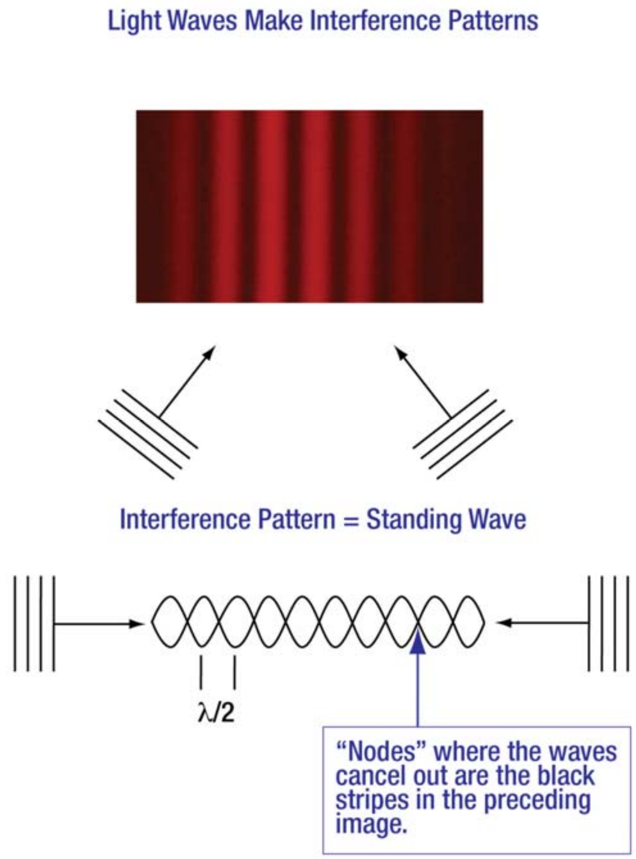
Figure 5: Left: ray picture. Right: wavefront picture. A point source in the front focal plane of the condenser (the aperture stop) results in an expanding spherical wave that is collimated by the lens into a plane wave that illuminates the whole sample.

So it is not hard to illuminate the sample with a plane wave, but what happens when the plane wave hits the sample? The sample will modify the plane wave – absorbing some light, or slowing some of the light down, or perhaps not modifying it at all (e.g. if no sample is present, or you have just a blank slide). Again for simplicity let's consider only *absorbing* samples; again, we can generalize later.

Take for example a perfectly absorbing sample (say coated with carbon black), where the absorbance goes from 0 to 100% in a sinusoidal pattern. Then when the wave crest of the plane wave hits the sample, immediately on the far side of the sample the wavefront will be modulated from 0 to 100% in the same sinusoidal pattern. It is of course not obvious what that means in terms of what the light does next.

At this point we invoke a trick: instead of trying to figure out what the light does next from first principles, we can instead see if there is a model system (one easier to understand) that produces the same electromagnetic field at the sample plane. Once we have that, we can figure out what happens next – it does not matter ***how*** we generate the field at the sample plane; if two different systems produce the same field there then what happens to the light afterwards ***must*** be the same.

Since light is a wave, setting up a model system that generates a sinusoidal pattern is not actually that hard: in fact, it is called an interference pattern. "Interference pattern" is just a fancy word for a standing wave – exactly the sort of thing one gets by plucking a guitar string or a rubber band. Of course a single plane wave is not a standing wave – the wave crests are moving in time – but *two* plane waves going in opposite directions ***do*** produce a standing wave.



Light Waves Make Interference Patterns

Interference Pattern = Standing Wave

“Nodes” where the waves cancel out are the black stripes in the preceding image.

Figure 6: Two plane waves can combine to make a standing wave, where certain points on the wavefront never move (and so have zero amplitude or intensity), while others combine to have twice the amplitude of either wave by itself (and oscillate between ±2A).
(Image Source for Interference Pattern: https://upload.wikimedia.org/wikipedia/commons/8/87/SodiumD_two_double_slits.jpg, Image is licensed under the GNU Free Documentation License: https://www.gnu.org/licenses/fdl.html, Note: This image was cropped from original image.)

Two cases are pretty simple: waves hitting head on, which generate an oscillating sinusoidal pattern with wavelength λ (and dark nodes every λ/2), and a single plane wave incident at 90°, which oscillates equally (and so is equally bright) at all points.

Slightly more complicated is the case of two plane waves intersecting at an angle θ.

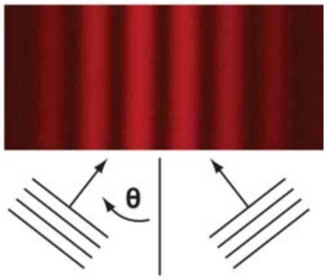


Figure 7: Two plane waves intersecting at an angle θ. (Image Source for Interference Pattern: https://upload.wikimedia.org/wikipedia/commons/8/87/SodiumD_two_double_slits.jpg, Image is licensed under the GNU Free Documentation License: https://www.gnu.org/licenses/fdl.html, Note: This image was cropped from original image.)

Since the waves are equal and opposite in angle, we know we will get a standing wave at the horizontal interface (see Fig. 7), and from geometry you should be able to convince yourself that the wavelength along the direction of the sample plane is λ / sin(θ):

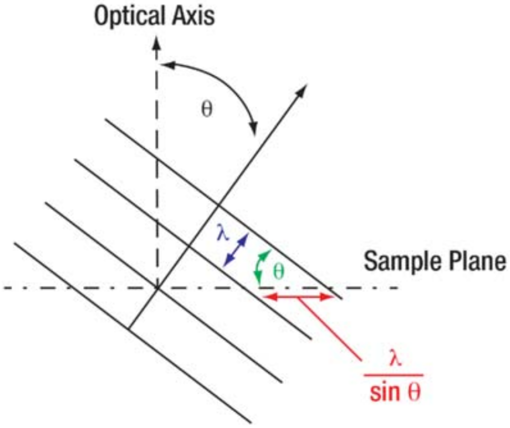


Figure 8: Wavelength for a plane wave *along the direction of the sample plane* is λ/sin(θ).

When two of these waves overlap at equal and opposite angles (as in Figure 7) we have a sinusoidal interference pattern (a.k.a. standing wave) with a "spatial wavelength" of λ / sin(θ).

Let's return to the question of our sinusoidal sample: in the case of Figure 5, a plane wave illuminates the sinusoidal absorbing sample (say, with transmission T = ½ [1 + sin(2π k x)]), leaving an electromagnetic field distribution on the far side of the sample of exactly the same thing (with an oscillating time term as well):

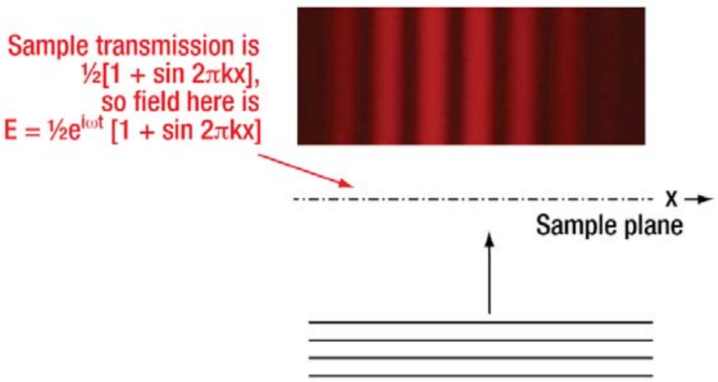


Figure 9: Field at sample plane for sinusoidal transmission sample illuminated by a plane wave at perpendicular ("normal") incidence. (Image Source for Interference Pattern: https://upload.wikimedia.org/wikipedia/commons/8/87/SodiumD_two_double_slits.jpg; Image is licensed under the GNU Free Documentation License: https://www.gnu.org/licenses/fdl.html; Note: This image was cropped from the original image.)

But we now know a second way to get exactly the same field distribution: all we have to do is choose the angle of incidence of two intersecting plane waves such that k = sin(θ) / λ. That will give us the sin(2π k x) term in Figure 9; adding a plane wave at normal incidence will give us the “1” term, and we are done!

Here's what it looks like:

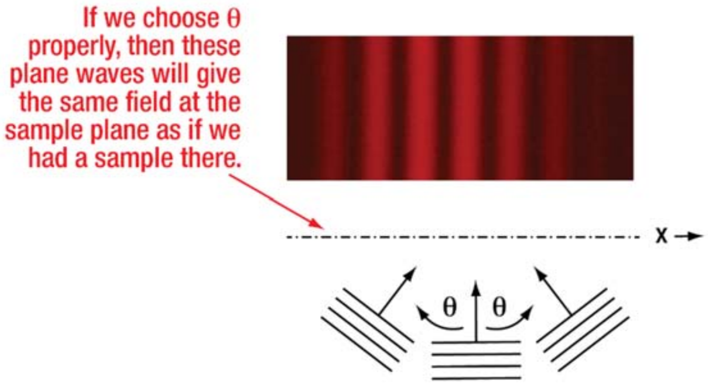


Figure 10: Superposition of plane waves to mimic an absorbing sample. (Image Source for Interference Pattern: https://upload.wikimedia.org/wikipedia/commons/8/87/SodiumD_two_double_slits.jpg, Image is licensed under the GNU Free Documentation License: https://www.gnu.org/licenses/fdl.html, Note: This image was cropped from original image.)

What good does this do us? Because we know what plane waves do in our optical system, now it is easy to tell what happens to the light from the sample. In fact, in general all we have to do is replace the plane-wave illuminated sample with a set of imaginary plane waves which would have given us the same electromagnetic field distribution at the sample plane. Since we already know how to handle plane waves

– they focus to points one focal length behind a lens – we can now figure out what happens to the light and the rest of our optical system.

Angular Spectrum: Building Up a Sample from Plane Waves

Let's look at a slightly more complicated example – a sample consisting of a series of parallel chrome lines on a glass slide, with line spacing $\boldsymbol{a}$; this is essentially a diffraction grating. Imagine illuminating the sample with a plane wave, which is to say with the condenser aperture closed down:

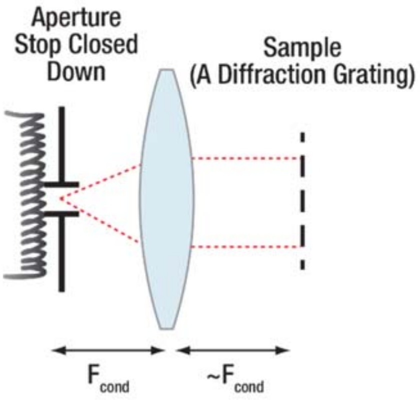


Figure 11: Plane wave illumination of a diffraction grating sample.

Using our above analysis, all we have to do is figure out what pattern of sine waves will give us a field equivalent to the plane wave illuminating the grating. Conveniently that is not hard to do for a function that is alternately zero or 100%: such a function is known as a square wave, and a series expansion for a square wave of spacing = $\boldsymbol{a}$ and amplitude running from 0 to 1 is known to be:

Equation 1:

$$\boldsymbol{F(x) = \frac{1}{2}\left[1 + \frac{4}{\pi}\left(sin(2\pi \cdot kx) + \frac{1}{3}sin(2\pi \cdot 3kx) + \frac{1}{5}sin(2\pi \cdot 5kx) + \ \dots\right)\right]\ ,}$$
$$\boldsymbol{with\ k = \frac{1}{a}}$$

Sums of the first four terms of the series are plotted below.

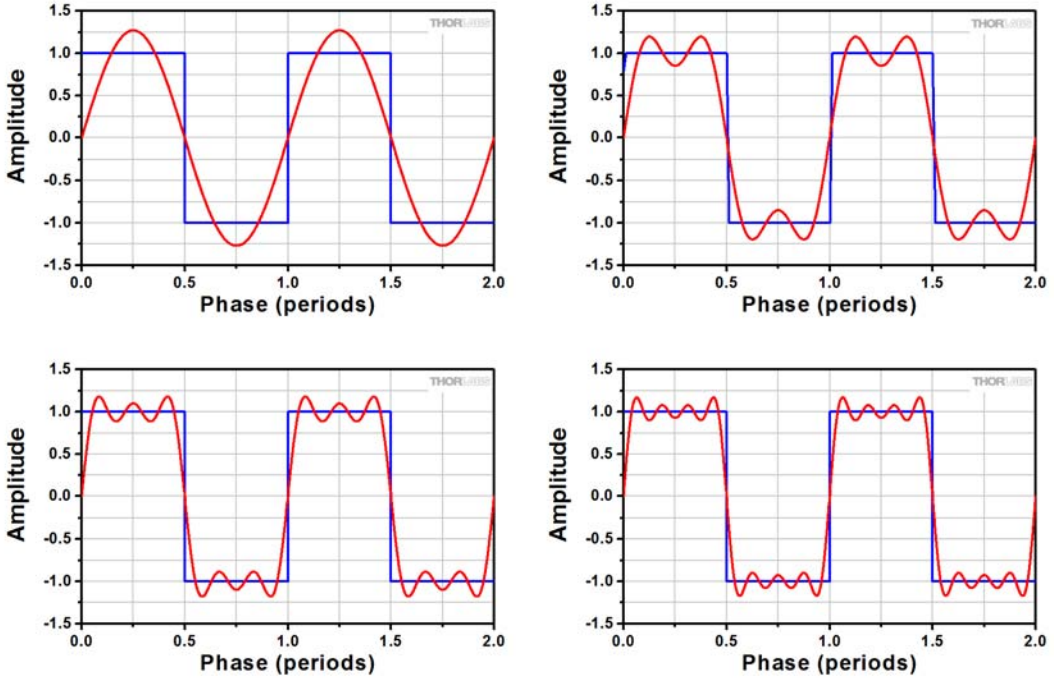


Figure 12: Sums of the first 4 terms of the Fourier series expansion for a square wave.

Since each sine waves corresponds to a pair of plane waves incident at some angle θ, all we have to do is pick θ so that k = 1/*a* = sin(θ) / λ and we are done:

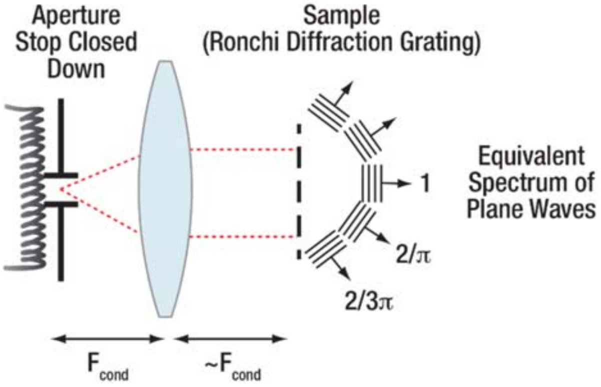


Figure 13: Superposition of plane waves producing a field equivalent to a diffraction grating sample. Note that the coefficients of the series give the amplitudes of the plane waves, and the angles come out to be sin(θ) = m λ / *a*, where m = 0, 1, 3, 5, etc. (odd harmonics), from Eq. 1.

It is worth pausing here to discuss this result a little bit, for it is actually rather remarkable. First, we now know how to figure out what angles look like scattering from a given sample – all we need to know is the Fourier series (or transform) for that sample, which is easy to get by hand, or for more complex samples, using software like Matlab. Second, from the same analysis, we get the amplitudes (and the light intensity is simply the square of the amplitude) of the scattered beams. Third, there is a direct relationship between NA (i.e., sin(θ)) and the (spatial) frequency of the interference pattern formed by plane waves at the sample, or equivalently between the spacing of features in the sample and the angles of the outgoing plane waves. Lastly, this analysis can sometimes reveal things which simpler models might miss; as an example, a square-wave grating apparently does not diffract into even harmonics, despite the fact that the standard grating equation suggests that there will be diffracted beams for every angle $\sin(\theta) = m \lambda / a$, where m = integer. As it happens, the even harmonics actually only show up for gratings with duty cycles other than 50% (i.e., where the width of the transmitting sections is different than for the absorbing sections), and the relative intensities of all the harmonics depend on these relative widths.

An additional (and important) conceptual point is that higher frequency sine waves correspond to both sharper features in the sample (see Fig. 12) and to higher scattering angles (see Fig. 13). That is to say, ***sharp or fine features in the sample scatter light into higher angles.***

Impact of the Objective Back Focal Plane:

Let's return to the physical picture of what happens as we image the sample:

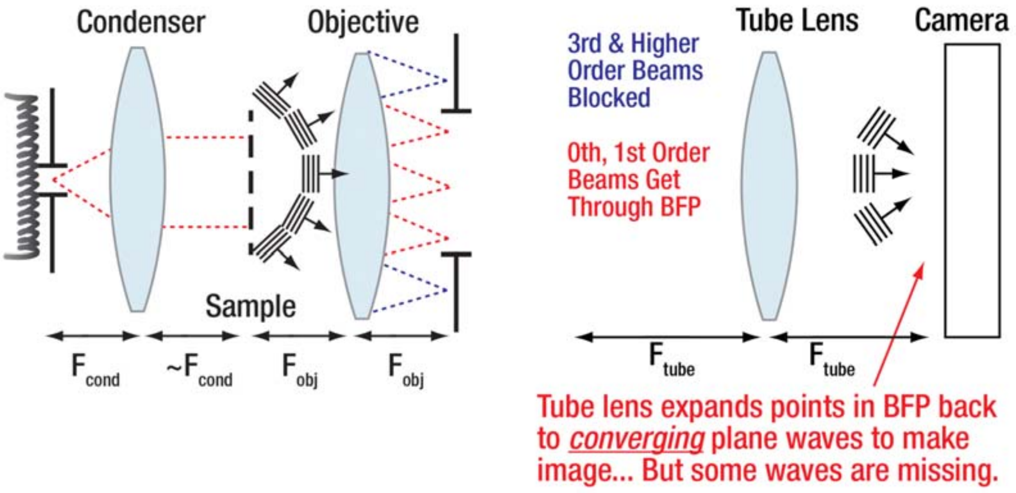

Condenser
Objective
Sample
Tube Lens
Camera
3rd & Higher Order Beams Blocked
0th, 1st Order Beams Get Through BFP
F_{cond}
$\sim F_{cond}$
F_{obj}
F_{obj}
F_{tube}
F_{tube}
Tube lens expands points in BFP back to *converging* plane waves to make image... But some waves are missing.

Figure 14: Diffracted beams in the objective back focal plane; some higher angle scatter (corresponding to higher frequency information in the image) does not make it through, so image reassembled by tube lens is missing some sine wave frequencies.

Figure 14 is one of the most important figures we will present in this course; it contains the entire essence of the Abbe theory of imaging. Moving from left to right, the small condenser aperture results in plane wave illumination at the sample. The sample diffracts light in a manner directly related to the sine waves (a.k.a. pairs of plane waves) required to make an equivalent electromagnetic field distribution, with higher frequency sine waves corresponding to plane waves at higher angles. These plane waves hit the objective and are focused down to points in the objective back focal plane, but the objective is a finite aperture, so not all the diffracted plane waves are collected by the objective, or (if they are collected) make it through the back aperture. As a result, when the tube lens expands the points in the objective back aperture into plane waves, which converge on the camera, the image formed is missing the information contained in the blocked plane waves. In normal brightfield imaging, the plane waves, which are blocked at the objective back aperture, are those which had scattered into higher angles (higher NA). As a result, the information which is ***missing*** from the reassembled image at the camera corresponds to the higher frequency sine waves required to make up the original sample distribution. This is the reason that resolution deteriorates as objective NA is reduced.

Another way of phrasing that last point is to say that the numerical aperture of the ***objective acts as a low-pass filter on the spatial frequencies present in the sample***. There is actually a more fundamental low-pass filter than the objective: the wavelength of light. This should make some sense – if one is working with waves of wavelength λ, then there is no way to reproduce sinusoids of finer spacing than that; hence λ itself sets upper bound on the reproducible frequencies. This is even more interesting than it might sound, and is related to bounds on the propagation of light and the nature of so-called nearfield radiation, though discussion of that is beyond the scope of this course.

A low-pass filter is one which only allows frequencies lower than a certain threshold through; there are of course other types of filters: one example would be the aperture shape you used for central darkfield – there the objective iris limited the maximum NA, but the ball driver you used to block the middle of the aperture blocked the light corresponding to low frequency sine waves in the sample. The combination of these two things amounts to a *bandpass filter*, allowing only a middle range of frequencies through. The lack of low frequencies meant that imaging depended on higher frequencies – those we have already seen are associated with edges (as shown in Fig. 12). The limit of higher frequencies due to the maximum NA of the objective is what determined resolution.

Given that we are now talking about filtering the various frequencies which make up the image, it makes sense to revisit why we have been paying so much attention to the front and back focal planes of various lenses.

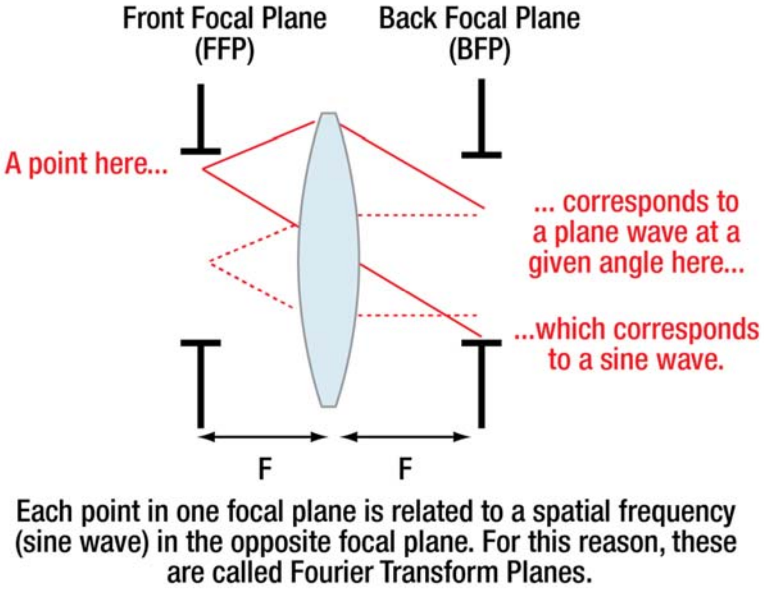

Front Focal Plane
(FFP)
Back Focal Plane
(BFP)
A point here...
... corresponds to a plane wave at a given angle here...
...which corresponds to a sine wave.
F
F
Each point in one focal plane is related to a spatial frequency (sine wave) in the opposite focal plane. For this reason, these are called Fourier Transform Planes.

Figure 15: The light distributions in the front and back focal planes of a lens are Fourier Transforms of each other.

By working in the focal planes of the lenses, and placing our apertures and making our adjustments there, we are affecting each part of the light distribution in the opposite "transform plane" equally – for instance, by limiting the numerical aperture by placing an iris in the BFP of the objective, we ensure that the resolution is the same for every point in the image.

The preceding pages should give you an enhanced appreciation for the concept of numerical aperture: the NA of diffracted light is directly related to the frequencies of the sine waves in the sample, since those frequencies are given by k = sin(θ) / λ = NA / λ. (Note: it is traditional to choose k to represent spatial frequencies; in this context it is usually called the wave number. As with ω and f, conventions vary as to whether k includes a factor of 2π. Our convention is that it does not – k in this case is similar to the frequency f). When applied to the limiting aperture of the objective, NA determines the highest frequency information from the sample which will be included in the image.

Given that we have decomposed the light distribution at our sample into a set of different sinusoids (related to the angles at which plane waves will be diffracted), it makes sense that in order to have our image be a scaled version of our sample, we will have to require that the frequencies of each sine wave be increased or decreased by the same factor during imaging. Unsurprisingly, that factor is called the magnification. Another way of saying this is to state that:

Equation 2: $$\frac{NA_{object}}{NA_{image}} = M$$

which is effectively the Abbe imaging criterion.

On a related note, the formula for the diameter of the back focal plane,

Equation 3: $$BFP\ diameter = 2\,f\,NA\,,$$

is not actually geometrically accurate (geometrically it should be a tan(θ), not a sin(θ), as in the NA). As a result, achieving this relationship in practice requires enormous effort on the part of lens designers[1]. The reason it is so important for it to be proportional to sin(θ) and not tan(θ) is that then the magnification is the same for all the sinusoids making up the object – they all scale the same way, and so when reassembled at the image they add up consistently to make the appropriate shape, just scaled by M. If the geometric tan(θ) relationship holds, then the M – and scaling for sinusoids – at small angles will be different than at larger angles, and all of the sinusoids from the sample will thus not be scaled by the same amount and so will not produce an ideal image when reassembled at the camera.

The effects of condenser aperture and details of the modulation transfer function (MTF) will be covered in subsequent Course Notes.

[1] The thing they need to accomplish is to have the principle planes of the lens system be spherical (i.e., they are principle surfaces, not principle planes), such that the sin(θ) relationship *does* hold geometrically. For details, see Juškaitis, Characterizing High Numerical Aperture Microscope Objective Lenses, in *Optical Imaging and Microscopy*. Springer Series in Optical Sciences, vol 87. 2003.

Addendum to Lab 6: Abbe Theory of Image Formation (I)

Due to the grid target/sample we use in class, the following question often arises. The question is both good and subtle; below is some discussion, with (calculated) pictures.

- ➔ **This level of detail is beyond the scope of what we expect you to know for the class; it is provided only to reduce any confusion if you noticed the odd features in the BFP image.**

You should nonetheless **read this (AFTER doing Lab 6** is probably best), and understand that **transmission through multiple filters or samples results in the *multiplication* of the effects of those filters or samples, not the addition of the effects.**

Question:

If the diffraction pattern from a sample consisting of vertical lines is a row of horizontal dots, and the pattern from a sample consisting of horizontal lines is a column of vertical dots, then when we use a sample that is horizontal lines on top of vertical lines, making a grid, how come we do not just get a "cross-shape" of a row of vertical dots on top of a row of horizontal dots when we look at the pattern in the objective back focal plane?

Answer:

The short answer is that a grid is NOT the *addition* of two linear gratings, even though it might seem like that is what it should be. Rather, it is the multiplication of the transmissions of the two gratings.

To think about how this works, imagine the plane wave coming toward the sample, with amplitude A = 1. Once it goes through the grating, right on the other side, it either has amplitude 1 if there was no metal line there, or amplitude 0 if there was a metal line that blocked it.

If I were to *add* two samples, then some places I would get amplitude A = 2 (where there were no lines in either grating – i.e., the "open" squares in a grid). Naturally, this cannot happen with the gratings – no light is added anywhere. What is really happening (and this is important for later in the course, when we get to spectra and fluorescence) is that **when light goes through one sample, and then goes through another, the effects are *multiplied.*** This makes sense if you think of what happens when you go through a 50% blocking filter, and then another 50% blocking filter – you do not end up with 0%, rather you get 25%, which is just 50% of 50%; the filter transmissions are *multiplied*. Similarly, two clear spots on the filter results in 100% * 100% = 100% of the light getting through at that point, not 200%.

If that was as obvious as it may sound, we would not be writing this. Many, many people (including one of us, when we were starting our first job in industry) have never thought about it. We encourage you to give it attention, since we will be returning to it in Lab 9 when we start covering fluorescence.

In thinking of the diffraction patterns, it *is true* that when you add two light (field) patterns at the sample, you get the sum of the individual diffraction patterns in the objective back focal plane, as shown in the following pages.

However, that is not what we have done when we effectively placed one grating on top of the other to get a grid – that actually multiplied the transmission functions. And multiplication of two sample functions results in a stranger behavior than just multiplication in the back focal plane – the original patterns are "convolved," which is a mathematical term we will cover only slightly, and which we are not going to worry much about now.

The most important point is this: for a diffraction pattern visible in the objective BFP, if we block out parts of it so that what is left corresponds to the same diffraction pattern made by some other sample, we will get an image that at the camera. Which should make sense, since those spots correspond to plane waves, and we are determining which plane waves make it to the image plane.

Pictures:

Note that a sample transmission of 1 + sin(2π k x) requires *three* plane waves: a constant (on axis) and two waves at equal and opposite angles. Hence we see three spots in the BFP in the figures below (explained in Figure 10 earlier as well).

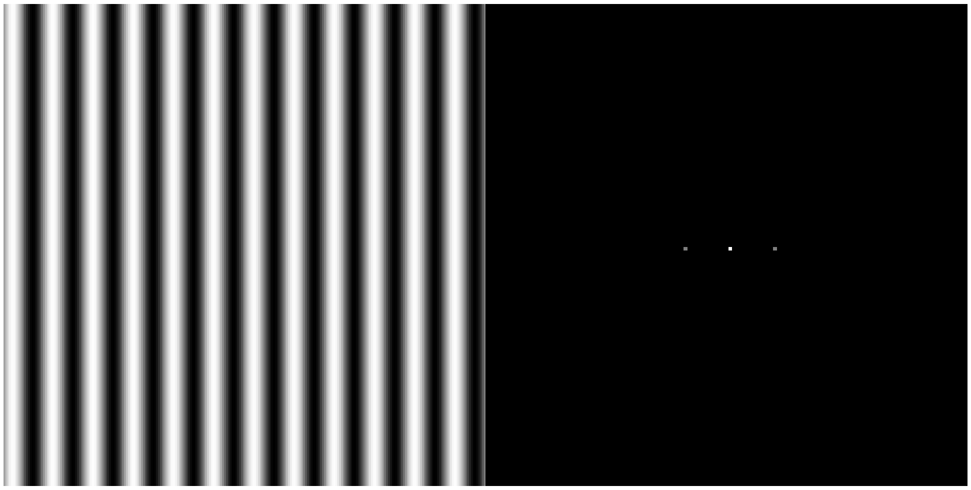


Figure 16: Vertical sine wave: f(x) = ½ [1 + sin(2π k x)]
Diffraction pattern is three dots, as discussed above and in Fig. 9

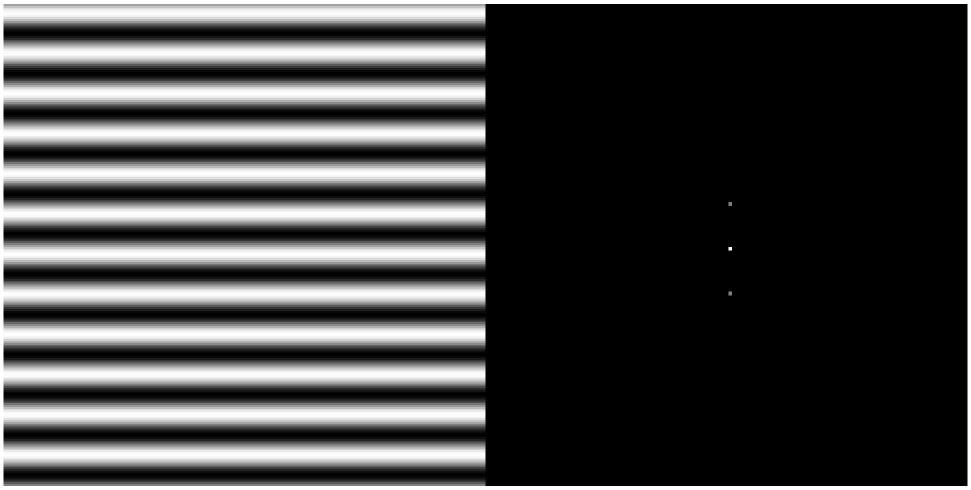


Figure 17: Horizontal sine wave: f(y) = ½ [1 + sin(2π k y)]
Diffraction pattern is three dots, as discussed above and in Fig. 9

If you add those two samples diffraction pattern is sum of previous patterns.

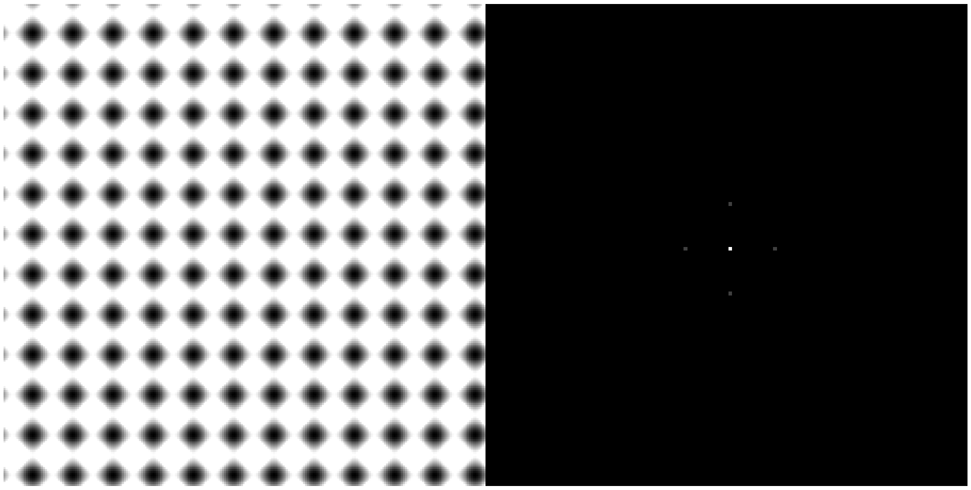


Figure 18: Sum of vertical and horizontal sine waves. Diffraction pattern is a sum of Fig. 16 and Fig. 17 diffraction patterns.

Note how the brightest areas in the sample must be brighter than before, since you have *added* two things each with a given brightness.

But if you *multiply* them, the diffracted pattern is weirder (becomes the *convolution* of previous patterns). Among other things, the maximum brightness does not change, since you have added nothing.

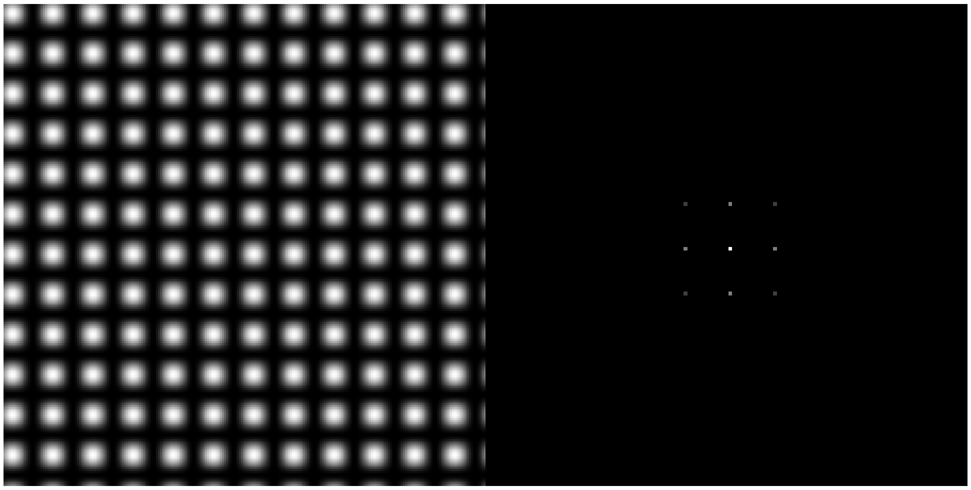


Figure 19: Multiplication of vertical and horizontal sine waves. Diffraction pattern is a "convolution" of Fig. 16 and Fig. 17 diffraction patterns.

Note: depending on the quality of the print copy of these notes you may be reading, the dimmer diffraction spots (in the right-hand images, above) may be hard to see. In that case, consider examining the PDF version of this document (available at www.thorlabs.com/OMC by clicking the "Download Manual and Educational Materials" button).

Similarly, for square wave patterns (like the parallel-line Ronchi rulings on the Thorlabs target), we get the same thing except that there are rows of diffracted spots since it takes more than three plane waves to make a square wave (remember the series expansion for the square wave given earlier in Equation 1):

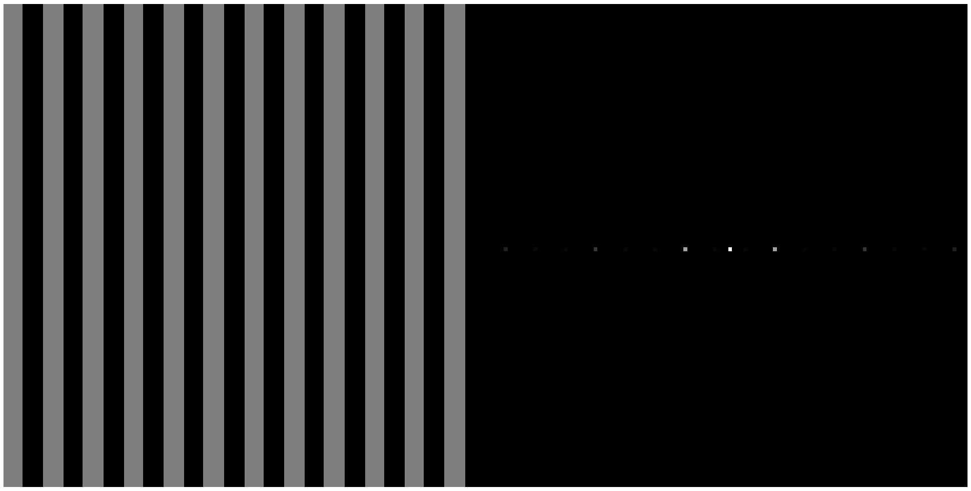


Figure 20: Vertical square wave:
$f(x) = ½ (1 + 4/\pi [\sin (2\pi k x) + 1/3 \sin(2\pi\cdot 3 k x) + 1/5 \sin(2\pi\cdot 5 k x) +\ldots])$

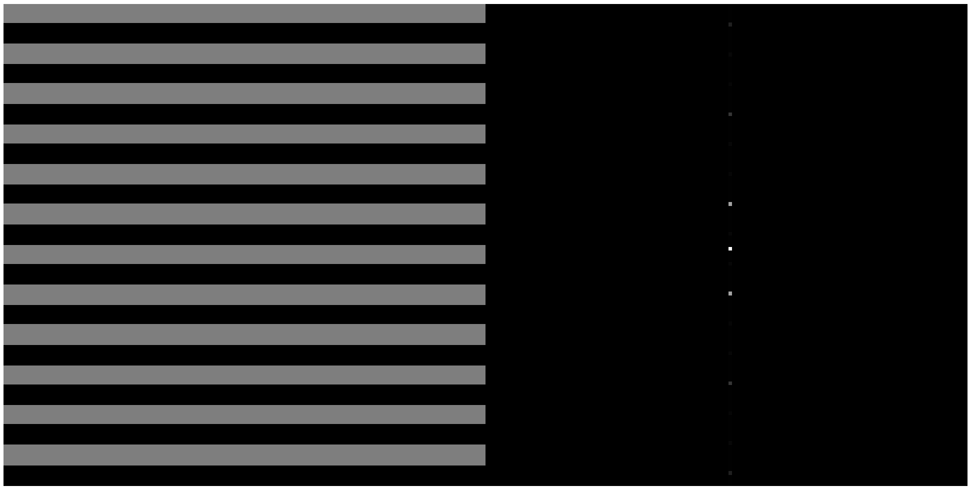


Figure 21: Horizontal square wave:
$f(y) = ½ (1 + 4/\pi [\sin (2\pi k y) + 1/3 \sin(2\pi\cdot 3 k y) + 1/5 \sin(2\pi\cdot 5 k y) +\ldots])$

Note how 2nd diffracted spots out are *farther* from the center than the 1st set of diffracted spots, just like you would expect from the equation – they are 3X as far out as the 1st spots (and fairly dim – see the comment earlier about the PDF version of these notes at www.thorlabs.com/OMC, or a high-quality print copy).

As before, *adding* the functions results in adding the diffraction patterns. However, adding the two functions does not give you just dark lines with white squares- note that where a dark line crosses a white line you get gray, so there are three intensities here, unlike in our grid sample, which has only black and white (0% and 100% transmission).

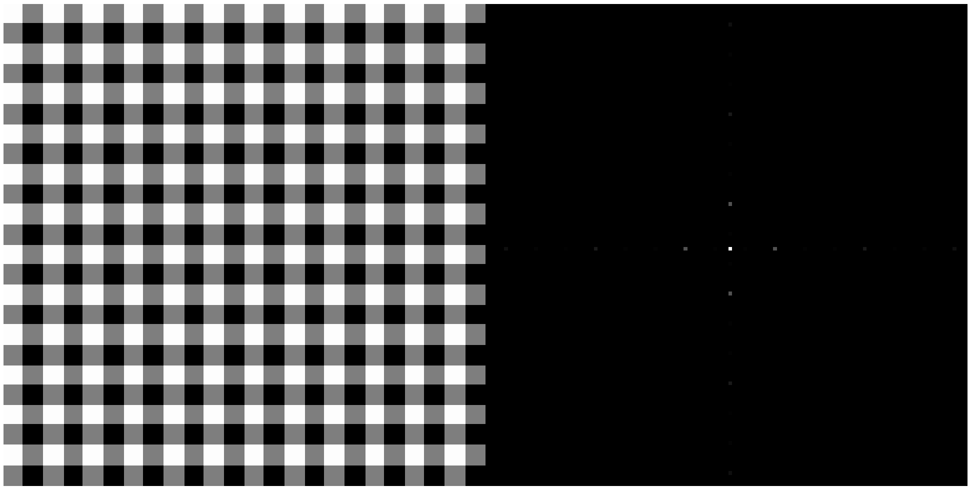


Figure 22: Sum of vertical and horizontal square waves. Diffraction pattern is sum of Fig. 20 and Fig. 21 diffraction patterns.

There are actually more than 5 dots in the right-hand image of Figure 22; they are faint, and you may need to adjust your screen angle or settings to see them well. Alternatively, you can cut and paste the left-hand image into ImageJ, go to process/FFT/FFT to get the diffraction pattern. To get rid of pixelation and edge effect issues (which are unimportant, so do not worry about them – they are more evident since ImageJ displays a log scale), then go to image/adjust/threshold, move the top slider (for the lower threshold) up until the crud disappears and you are left with the spots, then click "apply." This leaves them easy to see. Note: you can use this same process to look at the BFP image you will get for any sample; consider cutting and pasting a few images in and see what you get!

Below is the sample you are using in the lab – dark lines with white squares. However, that is obtained by *multiplying* the transmissions of the two gratings, and once you multiply, the diffraction pattern is no longer *the sum*, but rather *the convolution* of the two separate patterns. Note this produces extra spots on the diagonals:

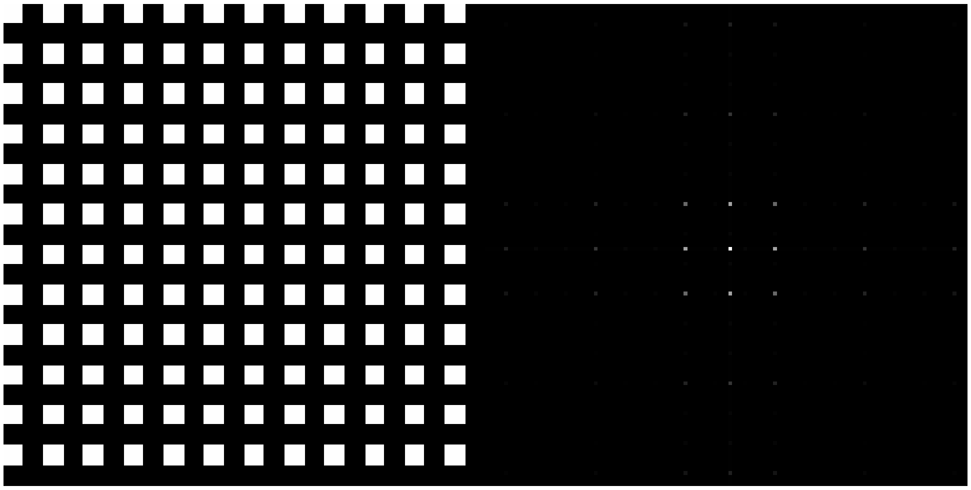


Figure 23: Multiplication of vertical and horizontal square waves. Diffraction pattern is a “convolution” of Fig. 20 and Fig. 21 diffraction patterns.

For this course, you do not need to know about convolution. Do not worry much about what the exact diffraction pattern looks like. **Rather, the important point is that when you *modify* the pattern in the BFP, you also modify the image.** For example, when you use a slit mask in the objective back focal plane (like in Peter Evennett’s videos), or block everything but one row of spots out (as in the lab), you are left with the pattern corresponding to a striped sample and that is exactly what you get in your image.

THORLABS

Lab 7 Notes: The Abbe Theory of Image Formation (II)

Optical Microscopy
Course

Lab 7 Course Notes:
The Abbe Theory of Image Formation (II):
Condenser NA and the Modulation Transfer Function

Overview

These notes will pick up directly from where the Lab 6 Course Notes left off; you should definitely review those notes before starting these.

In the Lab 6 Course Notes, we introduced the Abbe theory of image formation. For simplicity we assumed that the illumination consisted of a plane wave along the optic axis (which is what you get when you close the aperture stop down to a small pinhole), and saw that we could model the effect of the sample on the light as being equivalent to constructing the sample pattern from an appropriate set of incoming plane waves.

Looking at things from that perspective makes it clear why higher objective NA results in higher resolution: more of these scattered plane waves are collected by the objective, and consequently are able to subsequently interfere to form the image at the camera plane. The diffraction fringes you sometimes get are the result of blocking some of the plane waves, resulting in some missing sinusoidal patterns that would have been necessary to fully reconstruct the sample light distribution.

In these notes we will expand on that, extending the concepts to the case where you have the condenser aperture open wider. This will lead to the concept of the Modulation Transfer Function (MTF) – which is really just another way of looking at resolution, entirely analogous to asking about a stereo's frequency response instead of just whether it sounds good.

This, and phase contrast (which we will discuss in the following lab), will complete our discussion of Abbe theory. After that we will begin our discussion of fluorescence microscopy; among other things we will use that technique to revisit the concept of resolution from a different angle, closing the circle from Abbe theory back to the more familiar view of resolution in terms of the blurriness of an image.

Although not required, if you are interested in additional reading here are some suggested references:

References

1. (Easiest) The MicrsocopyU website has some nice figures (many taken directly from Inoue's book, #2 below); worth skimming. See *Modulation Transfer Function* under the *Reference Links* tab at www.thorlabs.com/OMC.
2. (Pretty accessible) Inoue's book, *Video Microscopy, 2nd ed.*, is a classic and contains a wealth of information; see Chapter 2, especially section 2.4, "Image resolution and wave optics."
3. (More mathematical) For those of you interested in the explicit Fourier theory behind what we will be discussing, without doubt the best reference (and much more readable than the equation content would make you think at first) is Goodman's book, *Introduction to Fourier optics, 4th edition.* See chapters 3, 6, and 7, especially section 3.10: "Angular spectrum of plane waves," section 6.1: "Thin lens as a phase transformation," section 6.3, "Image formation: monochromatic illumination," and 7: "Frequency analysis of optical imaging systems".
4. (In-depth coverage) A nice book with lots of good pictures and brief explanations, and almost no math (but which presumes a lot of prior knowledge in later chapters) is *Modulation Transfer*

Function in Optical and Electro-optical Systems, by Boreman, ISBN 0819441430. Worth skimming if you want a lot more detail, but you are better off starting with Inoue.

Recap

A quick review will be useful:

From the perspective of Abbe theory, a plane wave hitting the sample results in an electromagnetic field distribution just *after* the sample, which is given by the transmission of the sample multiplied by the incident plane wave field. If the plane wave is on-axis, then that is just a constant at any moment in time, so if we know what the sample is (e.g., a Ronchi grating) then we know what the field distribution is.

We can then figure out what sine wave patterns we would need to add together to get that same field distribution, and – since sine wave patterns can be created by interfering plane waves at different angles, with the sine wave spacing given by $\Delta x = \lambda / \sin(\theta)$, where θ is the angle of the incoming plane waves – this amounts to finding the angles of the plane waves the sample will generate.

Those plane waves will focus through the objective to be spots in the objective BFP; any that miss the objective, or are blocked in the BFP, will not continue on to the camera face.

The spots in the objective BFP are then focused back into plane waves by the tube lens, and those remaining plane waves interfere at the camera face to make the image. Since some plane waves are missing, the image is not perfect.

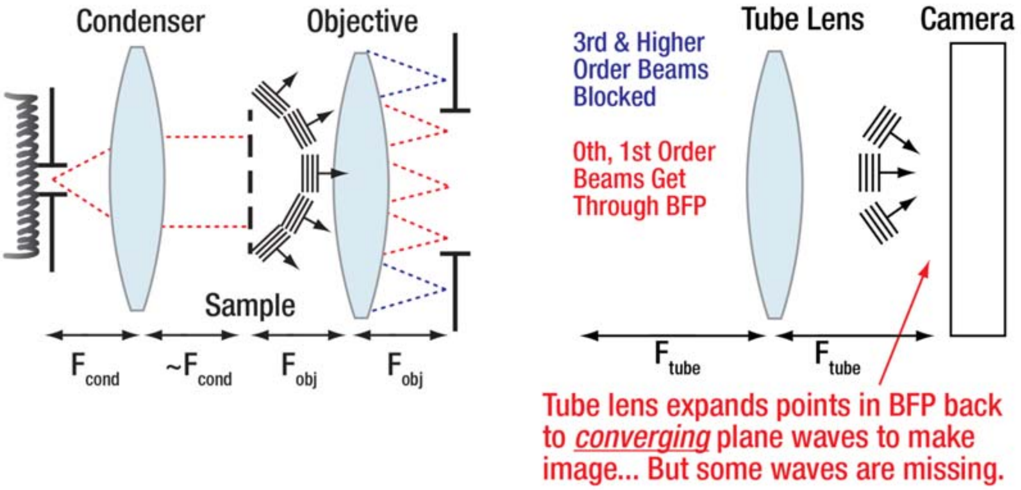

Condenser
Objective
Sample
F_{cond}
$\sim F_{cond}$
F_{obj}
F_{obj}
3rd & Higher Order Beams Blocked
0th, 1st Order Beams Get Through BFP
Tube Lens
Camera
F_{tube}
F_{tube}
Tube lens expands points in BFP back to *converging* plane waves to make image... But some waves are missing.

Figure 1: Diffracted beams in the objective back focal plane; some higher angle scatter (corresponding to higher frequency information in the image) does not make it through, so image reassembled by tube lens is missing some sine wave frequencies.

To gain intuition, let's look at the aperture stop and objective back focal plane (BFP) for some samples:

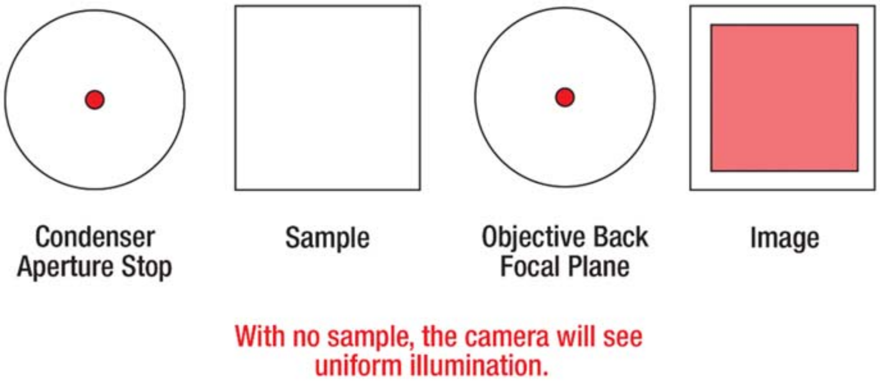

Condenser Aperture Stop
Sample
Objective Back Focal Plane
Image
With no sample, the camera will see uniform illumination.

Figure 2: With the AS closed down, we just get a spot in the objective BFP, which in turn (due to the tube lens) is a plane wave at the camera face. Hence, we see uniform illumination.

If we insert a grating sample, then as you saw in the previous lab, we get diffracted orders corresponding to the sine waves necessary to make up the square-wave transmission sample:

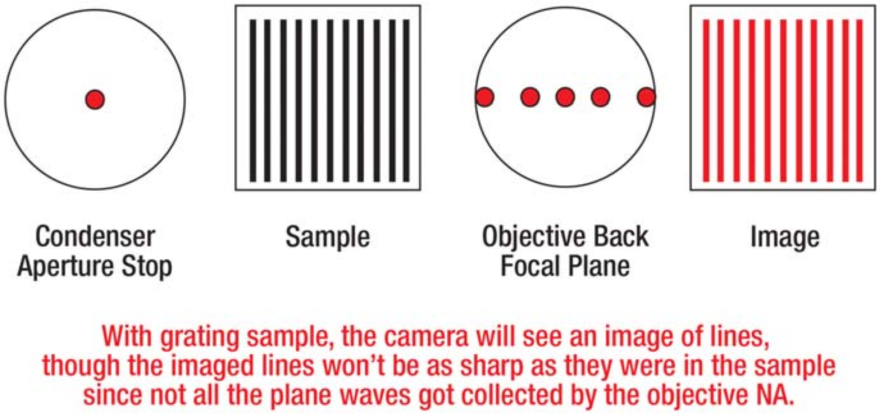
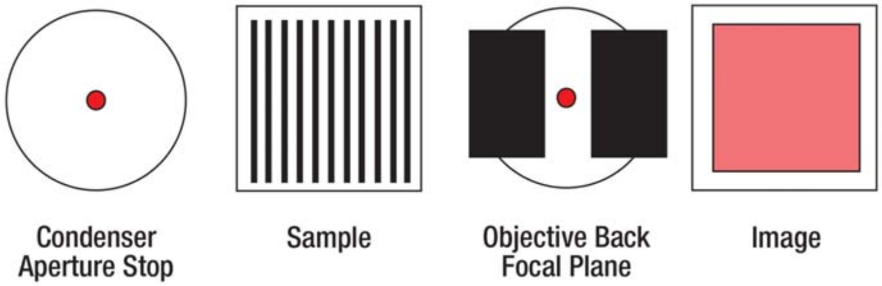

Condenser Aperture Stop
Sample
Objective Back Focal Plane
Image
With grating sample, the camera will see an image of lines, though the imaged lines won't be as sharp as they were in the sample since not all the plane waves got collected by the objective NA.

Figure 3: Diffracted spots from a square-wave grating sample.

We know that the tube lens refocuses the spots in the BFP to infinity, so that they are plane waves again at the camera face, where they interfere. Hopefully it is starting to make sense that the interference of the same set of plane waves we had at the sample would give us a perfect image, and if we are missing a few of the higher-angle ones we lose resolution and get some diffraction fringes. What if we purposefully modify things, and choose which plane waves get through? We should be able to construct any image we want (assuming we have the right spots in the BFP to choose from)! Let's start with a simple case…

Manipulating the Objective BFP

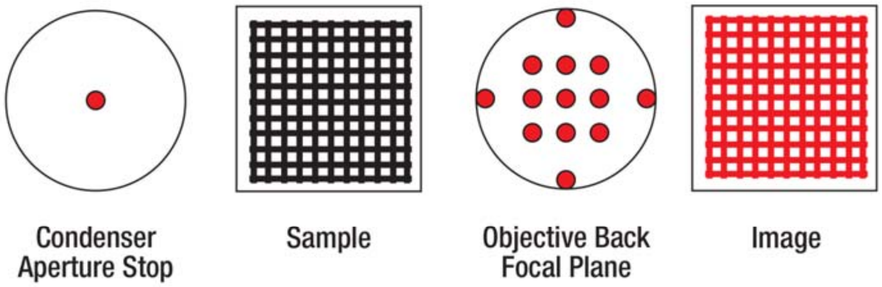


If we block everything but the central spot, we'll just have uniform illumination at the camera... even though the sample has lines.

Figure 4: Blocking all but the 0th order diffracted spot in the BFP, we get exactly what we would get with no sample: uniform illumination.

This may seem obvious by now, but it is actually rather subtle: we have made a different image by tailoring the light distribution in the objective back focal plane. And this example actually corresponds directly to what we have been doing when we set up central darkfield illumination: we have just been blocking the 0th order spot, which is to say, blocking the uniform background illumination at the sample – which makes the field of view *dark*. Hence "darkfield."

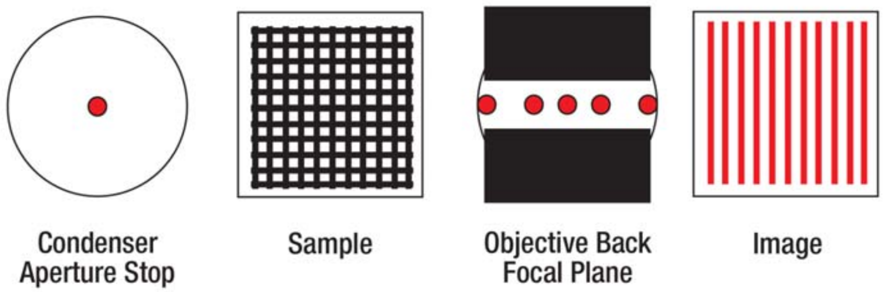
Let's examine a more sophisticated example:



Note: spot pattern is not a cross-shape; the reason for this is that we have not *added* two sample patterns (which would result in the sum of the two patterns in the BFP), but rather created a sample that is the *multiplication* of their separate transmissions... Which has a more complicated pattern in the BFP.

Figure 5: Diffraction pattern from a grid sample.

We will not spend time here discussing why the grid sample gives the pattern it does; this is discussed in the Addendum at the end of the Lab 6 Course Notes. Suffice it to say that the pattern does look roughly like what is shown above. What happens if we modify what we allow through the BFP to look like the pattern from the vertical lines grating sample (e.g. to be like Fig 3)?



If we block out spots so that only ones corresponding to a line pattern remain... then those are the only plane waves that get to the camera, and the image consequently looks like lines.

Figure 6: "Spatial filtering" of the diffracted spots in the objective BFP results in removal of the horizontal lines, turning the image of a grating sample into vertical lines.

In addition to building intuition, the point of the above examples is to help underscore the nature of image formation: the image consists of a set of plane waves which interfere with each other at the camera plane. If we remove some of the plane waves, we get a different image.

We will return to this point shortly, but first we will set up to think about resolution in terms of both the plane waves that make up the image (or, since the plane waves interfere to make sine-wave patterns with peaks spaced by some Δx, let's call them "spatial frequencies"), and the ***contrast*** we see at those spatial frequencies.

From Lab 6 Course Notes, we know that if we have a sinusoidal transmission sample with period Δx, illuminated by a plane wave on axis, the field just after it is:

Equation 1: $$\boldsymbol{f(x) = \frac{1}{2}\,[1 + sin(2\pi kx)]}\text{, where } \boldsymbol{k = \frac{1}{\Delta x}}$$

And we know we can make this up out of an on-axis plane wave of amplitude A =1 (giving the constant "1" in Eq. 1), and two symmetric plane waves coming in at equal and opposite angles, which provide the sine-wave component to the field. The necessary angle is given by

Equation 2: $$\boldsymbol{\Delta x = \frac{\lambda}{sin(\theta)}}$$ **Memorize this.**

Since the two waves combine to make the sine wave, their separate amplitudes are each A = ½. A critical point here is that we actually only need two waves to interfere- we already know this from darkfield- if we block out that central 0th order spot, we still get an image, right? Though it does look a bit different, it still results in interference and lines in the image. The same thing would be true of the 0th order and either of these 1st order spots. For now it does not matter, because the aperture is symmetrical, but later it will.

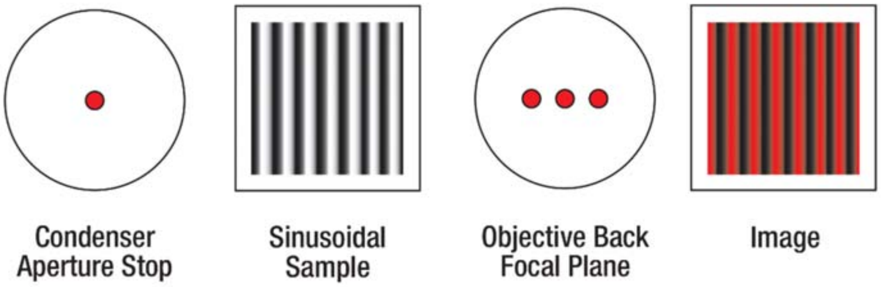
Contrast

So what is the *contrast* we have at the sample? Contrast is defined as:

Equation 3: $$Contrast = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}$$ **Memorize this.**

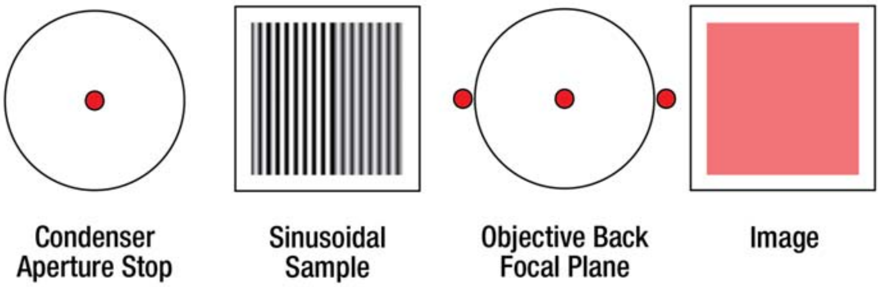
Where I = intensity (camera counts). The transmission of our sample goes from 0 to 1 and the maximum is 1, so in this case the *contrast at the sample* is 100% (since (1 – 0) / (1 + 0) = 100%).

What about the contrast in the image? Is it the same? Let's look:



A sinusoidally absorbing or reflecting sample results in 3 spots.

Figure 7: Diffraction and imaging from a sinusoidal sample; 100% contrast.



Higher frequency sinusoids require plane waves at higher angles... so the spots are farther apart in the BFP.

If they are too far apart (as above), they get blocked, and the sample loses all contrast

Figure 8: Diffraction and imaging from a sinusoidal sample, 0% contrast.

If *all* the diffracted spots make it through the objective BFP unobstructed, the image is a perfect replica of the sample. This holds true until the point when the spatial frequency of the sample sinusoid gets so high (i.e., the period of the sine wave gets so small) that the required angle for the plane waves results in them coming to a focus outside the objective BFP aperture, and they get blocked. At that point, only the background light, the 0th-order plane wave, is getting through so the image is not black, but it is completely uniform, and thus has 0% contrast.

Since we can make up any sample from a bunch of different sinusoids, it can be helpful to know what happens to the different sinusoids – which ones get through, and whether their contrast is undiminished. We can display this information in a slightly more informative way, by plotting the contrast for sinusoids of steadily increasing frequency.

The Modulation Transfer Function

Before you move on, try this yourself. To get you started, the graph will have the contrast on the y-axis, and on the x-axis the spatial frequency *k* (or, equivalently, diffracted spot spacing in the objective BFP, though it is always actually written in terms of *k*). Try sketching the graph. If you are feeling confident, sketch one for the case of central darkfield illumination too. Compare your sketches to Figures 9 and 10 to see if you are correct.

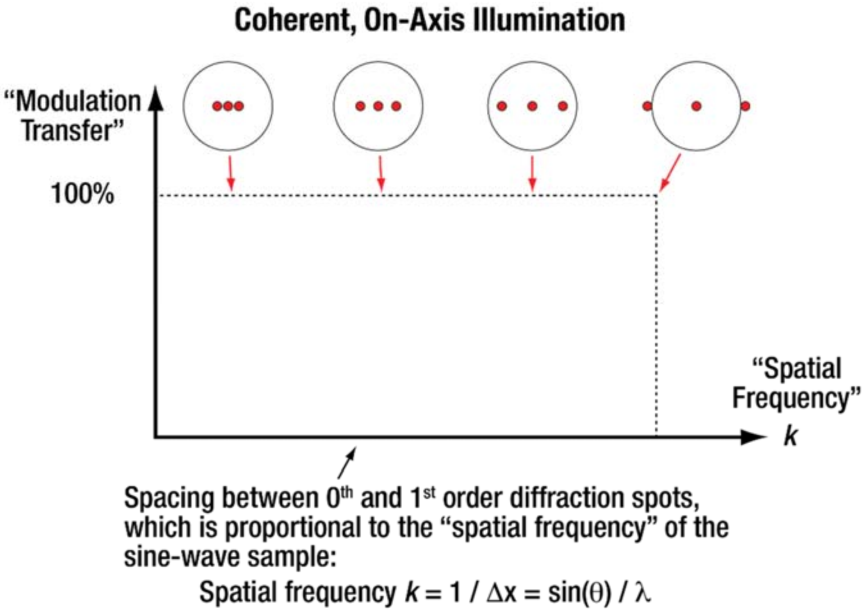


Figure 9: Modulation Transfer Function (MTF). Note that for coherent illumination (i.e., the aperture stop closed down to a pinhole, so only one point on the filament contributes light to the system) the contrast is 100% up until the cutoff where the spots fall outside the objective aperture – at which point the contrast becomes 0%.

For central darkfield, we block the middle of the aperture, where the spots would be close together. So the contrast *must* go to zero there, as seen in Figure 10.

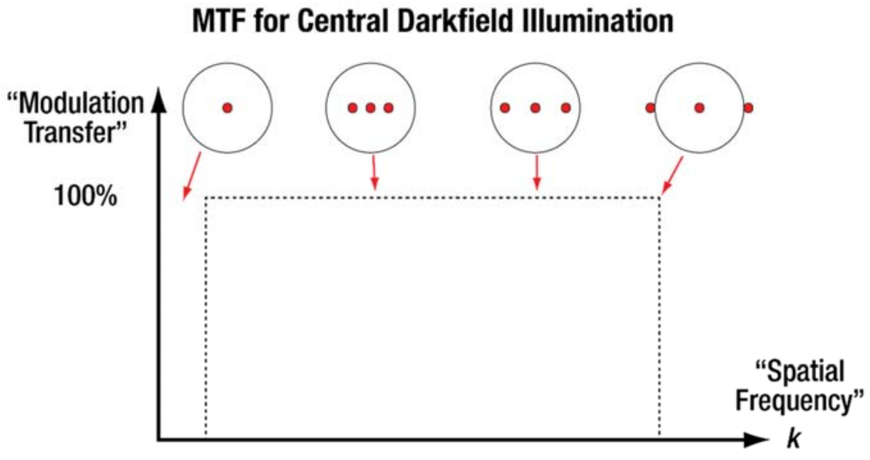


Figure 10: MTF for central darkfield. The lower limit depends on how much of the BFP aperture you block, and the upper limit is set by the objective NA (the iris in the BFP). Note: in each case, the center dot will be blocked by the mask in the objective BFP, as a result nothing at all will get through for low enough angles (low k).

You might ask if the contrast is really 100% where things are not blocked totally, and this is a good question. The short answer is "yes," since the definition in Eq. 3 essentially guarantees it in this case. More technically, with no constant term, Eq. 1 for the electric field goes both positive and negative, but one sees *intensity*, which is the square of the field and hence always positive. So the image will go from zero to some maximum, and by the definition of contrast this must be 100%, even though the maximum itself is lower than it would be if you had the constant term still there (that is why it is defined as a ratio).

For those of you with any electronics experience, there is a direct analogy here: Fig. 9 corresponds to a "brick-wall low pass filter" (called a brick wall filter because the transition is vertical when plotted, like a wall), and Fig. 10 to a "bandpass filter."

The above may seem rather simplistic – after all, we already knew the spots in the objective BFP got cut off abruptly. You could imagine cases where that is not true, though – an example would be if you put a so-called "apodizing" filter in the BFP, e.g. one that transmits less and less light as the radius increased, so that spots near the edge were attenuated compared to spots near the middle. In that case, the contrast would drop steadily as a function of frequency. In fact, there is at least one academic paper on doing exactly this; the reason for doing it is to get rid of the "fringes" that result from the abrupt frequency cut-off. There is a more typical way to do that, which we are about to discuss, but there are other reasons for using the apodizing filter just described. Those reasons are beyond the scope of this discussion; the important point is that the MTF curve need not be as simple as the one we just drew.

Coherence of Illumination

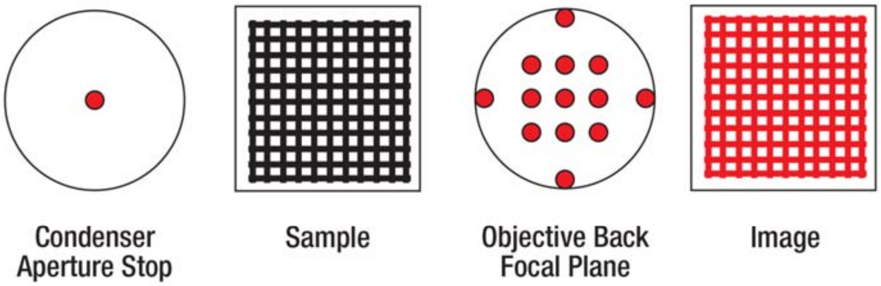
So far all our examples have involved having the aperture stop closed down to a pinhole. One effect of this is that all the light coming through the system originates in one part of the aperture stop and since the filament is imaged to the aperture stop, the light has all come from one tiny bit of the filament. Thermal light coming off a tiny (sub-wavelength) region of anything is spatially *coherent* – that is to say, it is an expanding spherical wave (discussed in Lab 1 Course Notes). If you collimate it, it becomes a plane wave. Either way, any part of the wave crest can *interfere* with any other part – a peak and a trough can add up to zero, or two peaks will add up to twice the height, etc.

This is not the case for light from two different places on the filament – say a millimeter apart. The reason for this is that, even if those two parts started out emitting in a synchronized fashion ("in phase"), the thermal vibrations (known as phonons) that exist in hot metal regularly jostle the emitting atoms, and every time that happens the oscillating electrons (if you take a classical view) get disturbed and the phase of the light they emit gets randomized. This does not matter if you are only imaging from one point, since all the light starting there has the same phase (even if that phase keeps changing) so any diffracted spots will still interfere with each other since their relative phases are correct. However, **the diffracted spots due to light from two different spots on the filament will not interfere on average** (any electromagnetic fields will interfere instantaneously, but if that interference is randomized over very short timescales (~ femtoseconds or less) then you will not see the interference since it will average out so fast).

What does that mean? In simple terms it means that if we illuminate from two or more points in the aperture stop, we will get two or more diffraction patterns in the objective BFP, but the light from different patterns will not interfere, which means the plane waves that hit the camera from those spots will also not interfere. Thus **the image we see will be what you get by adding up the images formed due to the illumination at each separate point in the condenser aperture**.

This has some important effects; it is worth looking at an example to get a sense for it:

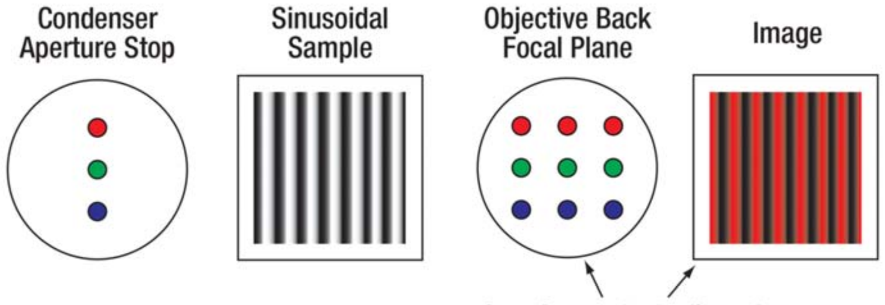
Recall the diffraction pattern for the grid sample:



Note: spot pattern is not a cross-shape; the reason for this is that we have not *added* two sample patterns (which would result in the sum of the two patterns in the BFP), but rather created a sample that is the *multiplication* of their separate transmissions...
Which has a more complicated pattern in the BFP.

Figure 11: Grid sample again.

Now let's put a mask with *three* pinholes in the aperture stop. We are coloring them red, green, and blue to distinguish them in the figure, but the light is all the same color. If we do that with the *sinusoidal sample,* we get the following:



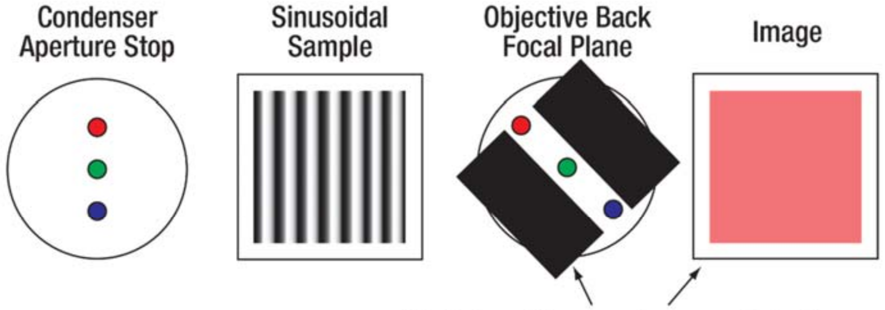
Set of spots looks like what we got for the grid sample.. so how come image isn't a grid?

- Light from different part of the filament (different spots in the condenser aperture) does not interfere.
- This is what "incoherent" means.
- Image is built up of separate images due to each single point in the condenser aperture.

Figure 12: Diffraction patterns from separate illumination points can *look* like the grid diffraction pattern, but they do not give a grid image.

Note: to simplify things conceptually, the figure above assumes the image of the AS in the BFP is right-side up (not inverted). As you probably have already seen, the image for our geometry (and any 4F system) is actually inverted. This makes no difference to the current discussion, and we note it only for completeness.

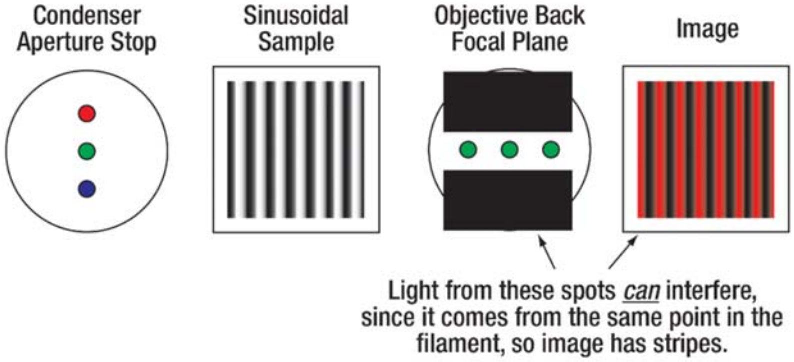
Not only does a similar-looking diffraction pattern in the objective BFP *not* give the same image, it also does not behave like we are used to if we block part of it out:



Light from these spots cannot interfere, so image is just constant background (due to three separate plane waves that don't interfere).

Figure 13: Lack of interference between diffraction spots from different illumination points.

If we turn the mask around so that the spots that get through all derive from the same illumination point, however, we see that those *do* produce plane waves at the camera plane which interfere with each other:



Light from these spots *can* interfere, since it comes from the same point in the filament, so image has stripes.

Figure 14: Diffracted light from the same illumination point *does* interfere to form image patterns.

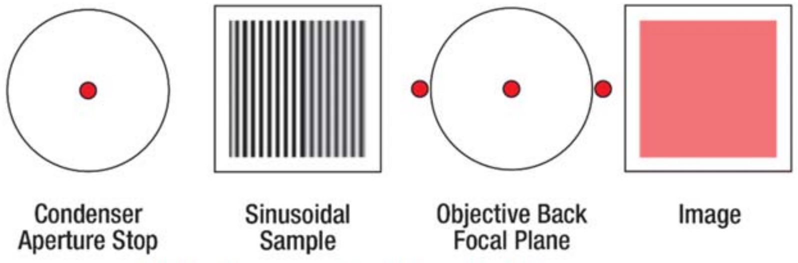
Because this is so important, we will restate it here:

Light (e.g. diffraction spots or plane waves) derived from different points in the illumination (aperture stop or, equivalently, the filament) *does not interfere* on average. Hence, the image consists of the sum of many separate images each formed by the light from one tiny region of the filament.

Non-Zero Condenser NA (i.e., Incoherent Illumination)

When we open up the aperture stop, we are essentially illuminating with more (individual) points of light. Before we try to cope with the effects of all those points, let's examine what happens when we take just one off-axis point.

Recall Figure 8:

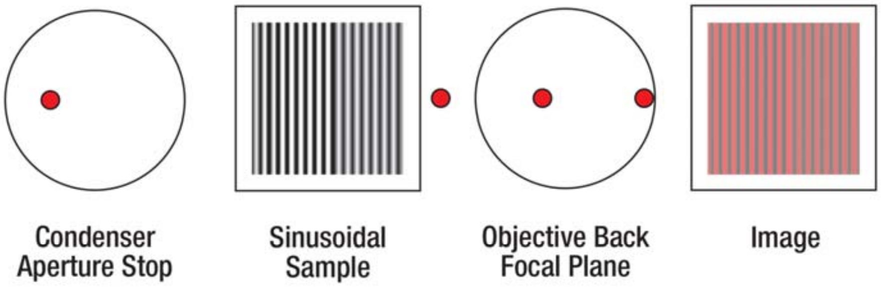


Higher frequency sinusoids require planes waves at higher angles... so the spots are farther apart in the BFP.

If they are too far apart (as above), they get blocked, and the sample loses all contrast

Figure 15: Diffracted spots from a sinusoidal grating that just miss getting through the objective aperture.

What happens if we illuminate from a point off-axis?

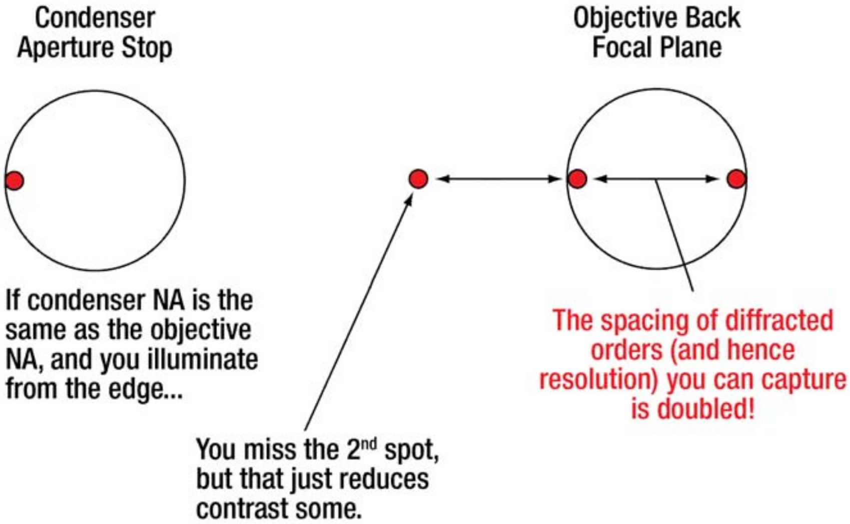


If we illuminate off-axis, we shift the center spot in the BFP. The diffracted spots stay the same distance from the center spot, and so one moves inside the aperture and is no longer blocked.

Since only two plane waves are required for interference, we get the pattern back in the image, though contrast is reduced.

Figure 16: Off-axis illumination allows us to capture higher-spatial-frequency sine waves, i.e., higher resolution information.

That ability to get extra resolution is a pretty big deal, so it is worth considering how much extra resolution we could hope to get. Practically speaking, if the distance between the 0th order spot and one of the 1st order diffracted spots is greater than the diameter of the objective BFP, you will not be able to get them both in there, so you will get no interference. So if that sets an upper limit, how can we get to that upper limit? By making our condenser aperture the same size as the objective aperture:



If condenser NA is the same as the objective NA, and you illuminate from the edge...

You miss the 2nd spot, but that just reduces contrast some.

The spacing of diffracted orders (and hence resolution) you can capture is doubled!

Figure 17: Achieving maximum resolution by illuminating such that the 0th order light hits the edge of the objective back aperture.

This should be an “aha!” moment – remember the Rayleigh resolution criterion:

Equation 4:
$$\delta x = \frac{1.22\,\lambda}{NA_{objective} + NA_{condenser}}.$$

Now we see that the extra resolution is present because, when we illuminate from the edge of a large condenser aperture (large condenser NA) we can capture diffracted spots from higher-angle plane waves. You may also notice a limitation to this formula: it only works if $\mathrm{NA}_{\mathrm{objective}} \geq \mathrm{NA}_{\mathrm{condenser}}$. To illustrate this, consider the case of $\mathrm{NA}_{\mathrm{objective}} = 0$. No diffracted spots can get through there no matter where you illuminate from, so even if $\mathrm{NA}_{\mathrm{condenser}} = 1$, the resolution δx must be 0.

Furthermore, there is little point in having the condenser NA be larger than the objective NA, since moving the 0th-order spot out of the objective back aperture does not do anything to increase resolution – you need it to interfere with the 1st-order spot. Generally, it is hard to get a condenser with NA > objective NA anyway – you spend all your money getting the highest NA objective you can. The one exception to this is darkfield.

Aside: Darkfield

Nobody actually does darkfield the way we do – the "method of central darkfield" is only a textbook thing, because it is easy to explain (and do). We could practically *double* our resolution if we just illuminated in a ring around the condenser aperture, like this:

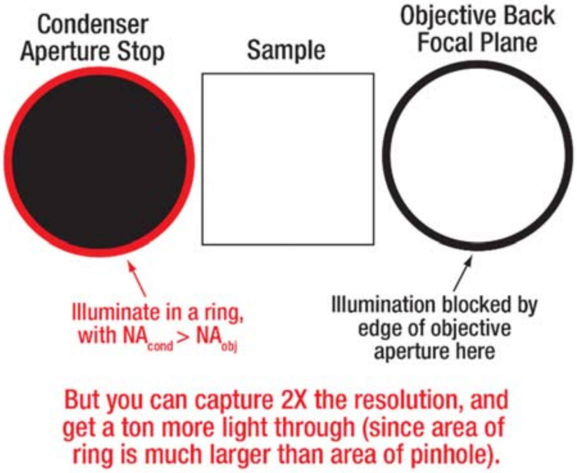

Condenser Aperture Stop
Sample
Objective Back Focal Plane
Illuminate in a ring, with NA$_{cond}$ > NA$_{obj}$
Illumination blocked by edge of objective aperture here
But you can capture 2X the resolution, and get a ton more light through (since area of ring is much larger than area of pinhole).

Figure 18: Typical darkfield illumination is done with an annular mask, so that the illumination is a cone of light that is at slightly too high an angle to be captured by the objective (either misses the lens or is blocked by the BFP aperture).

In addition to (nearly) doubling our resolution, the area of the annulus is much larger than the small pinhole we have been using, so you get a lot more light through – especially important when looking at tiny/dim samples that do not scatter much light.

Note: Ignore the following unless it is already bothering you. You may wonder, if in this version of darkfield the 0th-order and one of the two 1st-order diffracted spots will be missing, how will we get any contrast at all? The answer lies in the fact that a 1st- and a 2nd-order spot can also interfere, so if you imagine a square-wave grating sample, the absence of the 0th-order and all the orders on the other side of it is not the end of the world – there are still 1st, 2nd, 3rd, etc. orders that can interfere to give contrast. This also lies behind the fact that increasing the condenser NA above the objective NA does have a (small) effect on resolution. This is definitely beyond the scope of the course and is only noted in case some careful reader is getting confused.

Condenser NA and MTF

Returning to the subject of finite (meaning non-zero) condenser NA, let's examine what contrast we might expect. This is harder than before, since we need to find a way to take into account all the different points of illumination in our (now wide open) condenser aperture.

Conveniently, we know (roughly) what the maximum signal we will get is: it is just the light that would get through if there were no sample. That will be proportional to the area of the condenser aperture, shown below superposed on the objective back aperture (remember, these image onto each other):

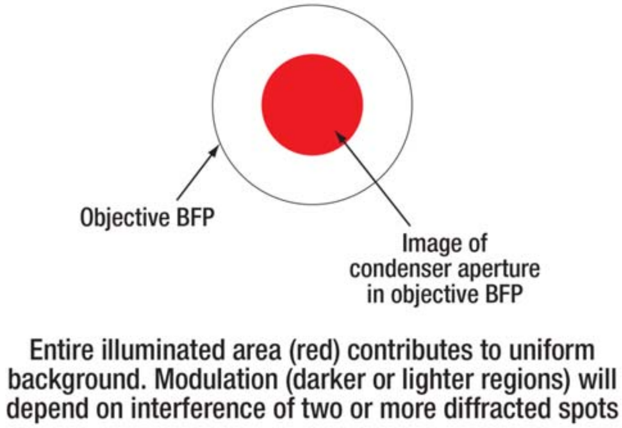

Objective BFP
Image of condenser aperture in objective BFP
Entire illuminated area (red) contributes to uniform background. Modulation (darker or lighter regions) will depend on interference of two or more diffracted spots

Figure 19: Condenser aperture imaged into objective back aperture. Note that as drawn, the condenser NA must be lower than the objective NA. The maximum light that can get to the camera is proportional to the red area.

So we know the denominator in the equation (Eq. 3) for the contrast. But what is the numerator? From our earlier analysis, we know that for any given sine wave, the contrast will be proportional to the number of diffracted orders we collect. And how many can we collect? Imagine a vertical sine wave grating – then the diffracted spots will be splayed out horizontally. Each spatial frequency in the sample corresponds to a spacing between a 0th-order and a 1st-order diffracted spot. We need to see how many points in the condenser aperture can generate a diffracted spot at a distance that still fits inside the objective aperture:

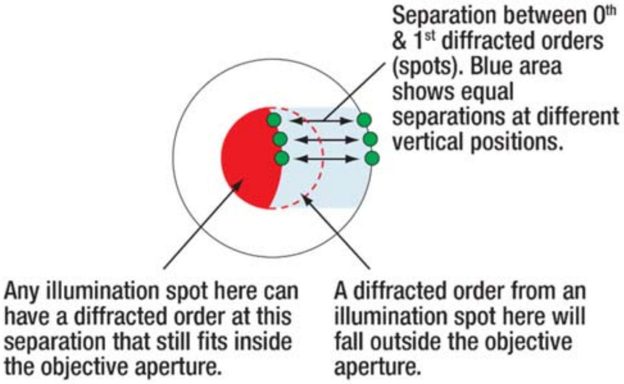

Separation between 0th & 1st diffracted orders (spots). Blue area shows equal separations at different vertical positions.
Any illumination spot here can have a diffracted order at this separation that still fits inside the objective aperture.
A diffracted order from an illumination spot here will fall outside the objective aperture.

Figure 20: Amount of illuminated area that can contribute diffracted spots (for a given spacing) that fit inside the objective aperture.

Let's review this, since it is important for understanding what comes next (though we will not require you to be able to reproduce the argument on any quizzes, etc):

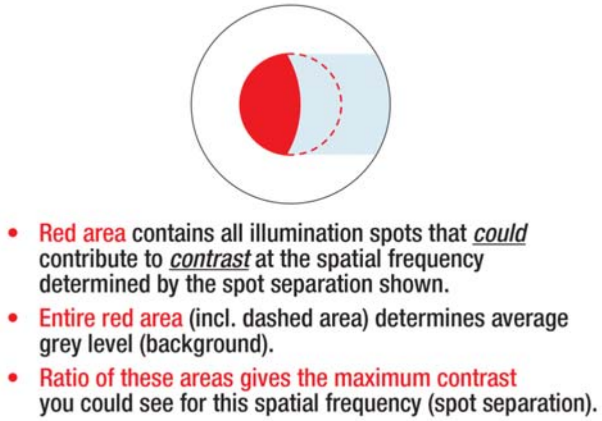

Red area contains all illumination spots that could contribute to contrast at the spatial frequency determined by the spot separation shown.
Entire red area (incl. dashed area) determines average grey level (background).
Ratio of these areas gives the maximum contrast you could see for this spatial frequency (spot separation).

Figure 21: Contrast ratio, explained graphically. The horizontal width of the blue area is the same, and represents a constant distance from the edge of the objective aperture (corresponding to a given diffracted spot spacing).

The argument I make above only takes into account spots on the right side of the illuminated points, but an identical argument holds for the ones on the left, so the only difference is that one must adjust the contrast by a factor of two somewhere.

Putting this all together gives us the modulation transfer function for the case of incoherent illumination, which is to say when we have the condenser aperture open:

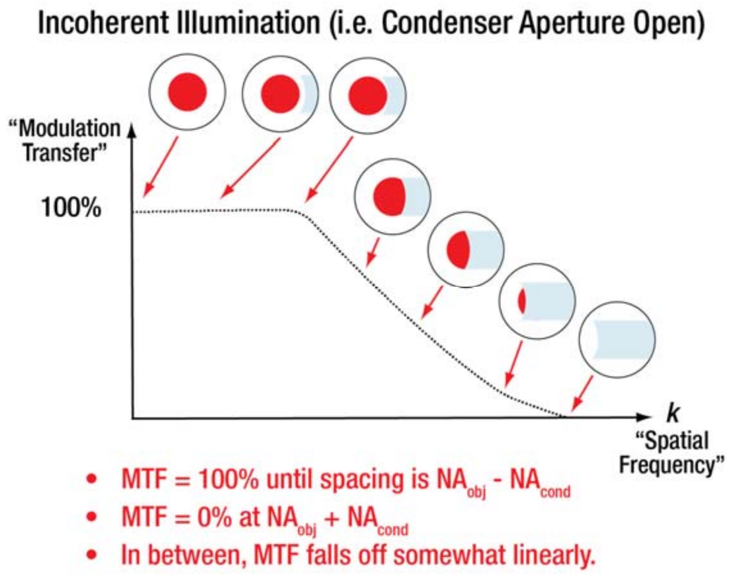


Figure 22: MTF for incoherent illumination, $NA_{condenser} < NA_{objective}$

Those of you mathematically inclined will recognize the geometric construction as being the autocorrelation of the objective and condenser apertures.

It is not critical that you understand completely the arguments that give the incoherent MTF, though you should be comfortable with the gist of them. What is important is that you understand the impact of opening the condenser aperture:

Opening the Condenser Aperture:

1. **Lets more light through.**
2. **Increases resolution by up to a factor of two.**
3. **Changes the shape of the MTF (modulation transfer function).**

The first two effects should already make sense – the larger aperture blocks less of the filament image at the aperture stop, so there is more light, and we have discussed the resolution increase above. The MTF changes shape as follows:

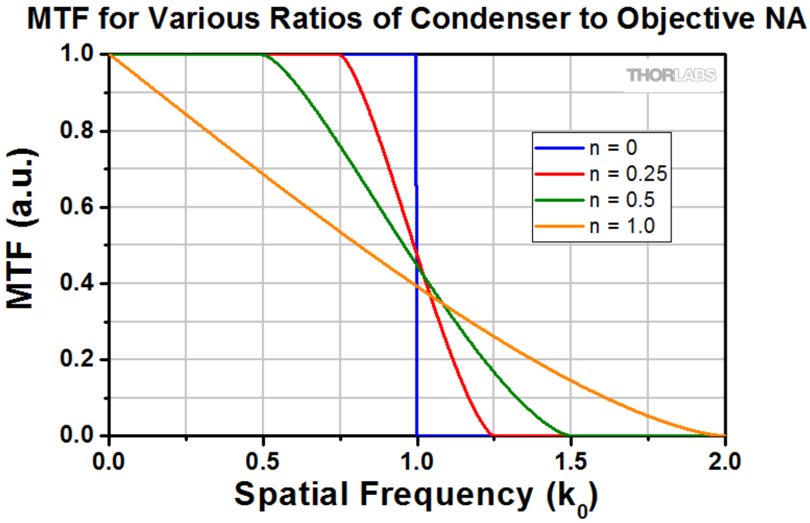


Figure 23: Effect on the MTF of opening the condenser aperture. With the aperture closed down (coherent illumination), we have the same thing we got in Fig. 9. When the condenser NA equals half the objective NA, we get what we had in Figure 22. As we open the aperture stop until it is the same as the objective NA, the MTF begins to drop off earlier and earlier, but reaches out farther and farther, eventually to 2X the coherent limit, k_0. However, while the contrast is 100% right up to the cutoff for the *coherent* case, the *incoherent* MTF can be much lower. X-axis is in units of the coherent cutoff frequency,

$k_0 = NA_{obj} / \lambda$.

(After GW Ellis via Inoue, Video Microscopy 2nd ed.)

What does all this mean? With the condenser aperture closed down, if we can see a sinusoidal sample at all, we can see it perfectly well – 100% contrast. However, if we open the condenser NA up to the same as the objective, our "resolution," by Eq. 4, increases by a factor of two. But at that limiting resolution in the figure above, we have a contrast of 0%, so we cannot make out anything! In fact, the Rayleigh resolution corresponds to an incoherent MTF of ~ 9%, and the limit of human eyesight is (depending on which study you read, and the observer, light level, etc.) 4% to 9% or so. That means that even though your nominal resolution might be nearly 5 cycles/μm (i.e., $\Delta x = 0.2$ μm), what you will be able to make out by eye will only be ~ 4 cycles/μm, or $\Delta x \sim 0.25$ μm, 20% worse than you expected.

That is one reason you should care a lot about minimizing your background from the room lights when possible (e.g. by using brighter illumination before turning up your camera exposure) – less background means more contrast, which means more resolution you can use, regardless of what the formula says. The comments above should suggest to you that you will rarely be able to make use of the entire theoretical resolution, at least if you have the condenser aperture open (such that maximum resolution occurs for very low contrast values).

Also, note in Figure 23 that the shape of the MTF curve changes depending on how far you have the condenser aperture open. Some of you have heard us say that when you set up Köhler properly, on a normal microscope, you actually look at the objective BFP (as you can do here with your color BFP camera) and adjust the aperture stop so it is about 70% of the size of the objective BFP aperture. The curve in the previous figure should show you what that implies: instead of going for maximum resolution, but with most of that resolution at very low contrast (and thus not much use to us), we are adjusting the illumination to get a ~65% improvement in resolution (instead of 100%), but the resolution we get is mostly at somewhat higher MTF than it would be for $NA_{condenser}$ = 100% $NA_{objective}$. In other words, the contrast is better so we can actually make use of the resolution we get.

From now on, adjust your aperture stop to 75% of the objective BFP when setting up Köhler. Be forewarned for the final.

What happens to the MTF as we increase the objective NA? It just reaches out to higher and higher spatial frequencies, as you might expect:

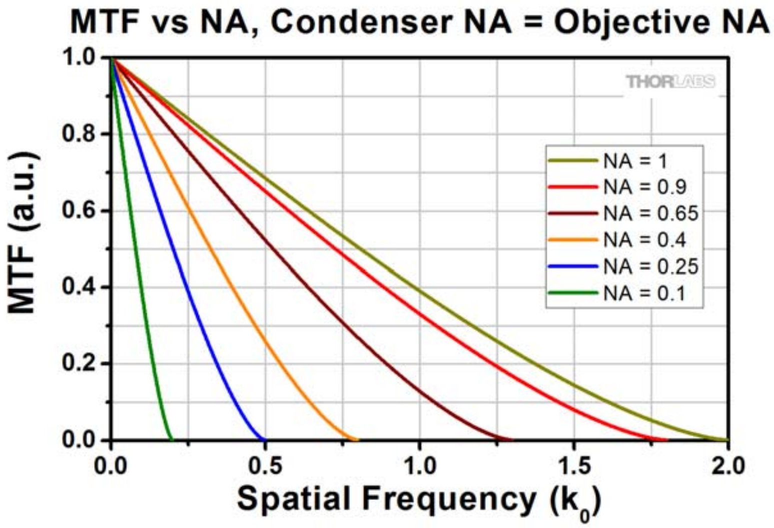


Figure 24: As NA increases, MTF reaches out to higher frequencies. For all traces, $NA_{condenser} = NA_{objective}$. As in the previous figure, the X-axis is in units of the coherent cutoff frequency, $k_0 = NA_{obj} / \lambda$.
(After GW Ellis via Inoue, Video Microscopy 2nd ed.)

Note that as the MTF reaches farther out, it is actually higher than the lower NA MTF all along the way; the total area under the curve is higher. It turns out that the area under the MTF curve is directly related to the height of the center peak of the Airy disk when you image a point source and roughly means the Airy disk is narrower, as you would expect since the resolution is higher.

Diffraction Fringes and Incoherent Illumination

You saw in lab that as you opened up the condenser aperture, the visibility of the diffraction fringes dropped:

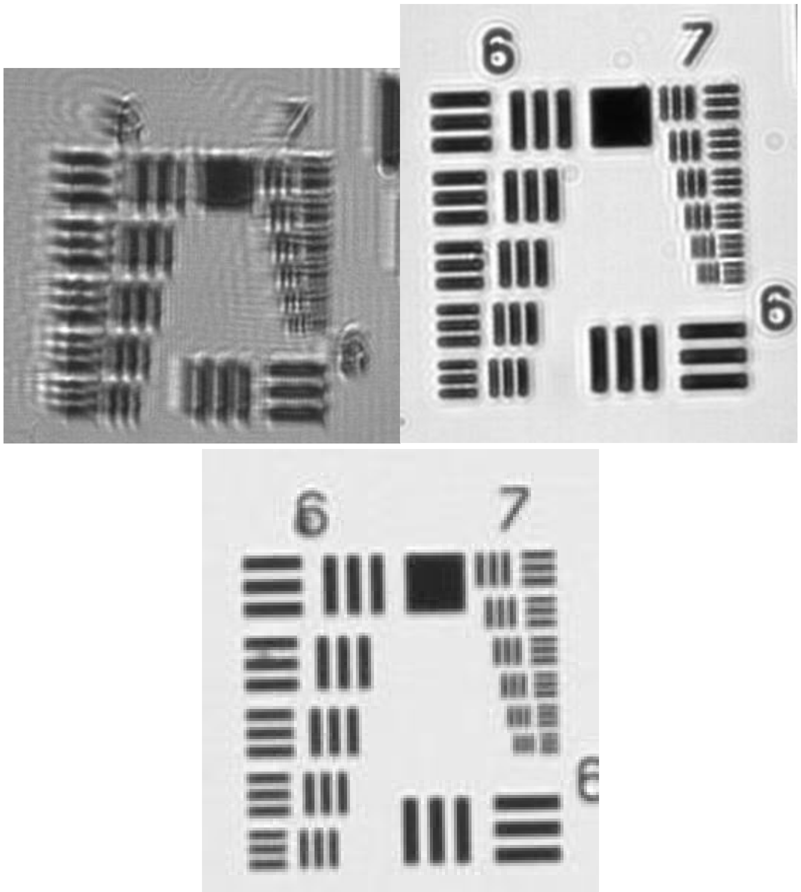


Figure 25: USAF target illuminated by an LED (top left, condenser NA ~ 0.003), a small condenser aperture (top right, NA ~ 0.01), and a diffuser (bottom, condenser NA~0.1). Note decreasing visibility of fringes.

The reason for this can be seen by looking at the effects of MTF on the series expansion for a square wave. That expansion is:

Equation 5:

$$\boldsymbol{F(x) = \frac{1}{2}[1 + \frac{4}{\pi}\,(sin(2\pi\, kx) + \frac{1}{3}\,sin(2\pi \cdot 3kx) + \frac{1}{5}sin(2\pi \cdot 5kx) + \ldots)]}$$

With the aperture stop closed down, the MTF = 100% right up to the cutoff, so that series might look exactly as in Eq. 5. For the incoherent MTF, which can drop roughly linearly from the start down to where it hits zero, we might get something closer to this in the image at the camera:

Equation 6:

$$\boldsymbol{F(x) = \frac{1}{2}[1 + \frac{4}{\pi}\,(0.8\,sin(2\pi\, kx) + \frac{0.4}{3}\,sin(2\pi \cdot 3kx) + \frac{0.1}{5}sin(2\pi \cdot 5kx) + \ldots)]}$$

where the red numbers are multipliers due to the dropping MTF. The fringes are less visible partly because we have more terms in the series (since the MTF reaches out to higher frequencies than in the coherent case) and partly due to the decreasing weighting on the terms, which results in a smoothing effect.

Before concluding, there is a nice example of where MTF can be useful. The following MTF curve shows what the effect of aberrations (in this case, primarily field curvature) can be – aberrations can only *lower* the MTF, but they do not do this uniformly.

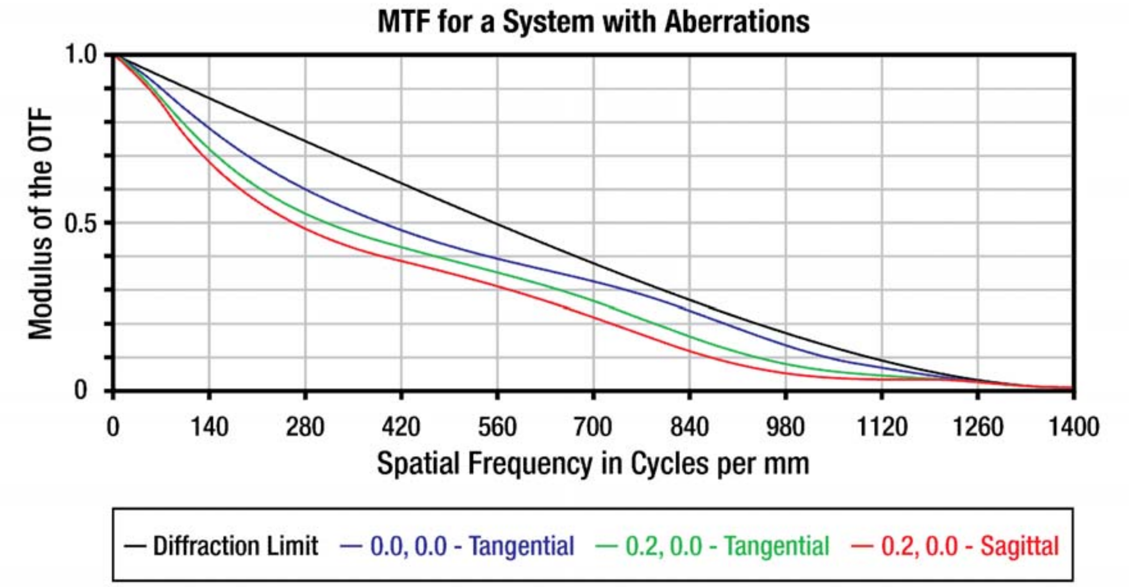


Figure 26: MTF for a system with aberrations. The upper traces are the diffraction limited MTF; the lower traces are actual performance due to geometric aberrations (mostly field curvature). Note that the MTF (the modulus of the Optical Transfer Function, or OTF) is worse than the diffraction limit on-axis (difference between black and blue lines) and falls off further at larger field radii (green lines). Curves were obtained using Zemax ray-tracing software.

Of particular note is that resolution can vary substantially depending on how you look for it. E.g., if we used a sample that contained primarily spatial frequencies near the contrast limit at ~ 5% at ~1300 lp/mm, this lens system would look almost as good as if it were diffraction limited, but it will actually perform worse than the diffraction limit across most of the spatial frequency range.

It is not hard to test the wrong part of the MTF curve – for example, the USAF 3-bar targets we use have specific spectral characteristics, favoring certain multiples of the line spacings, as shown in Figure 27.

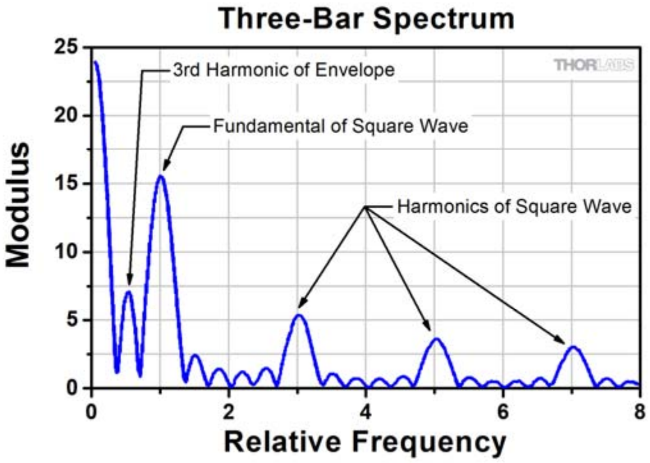


Figure 27: Spatial frequency content of USAF 3-bar target.
After work by John Rogers, Optical Research Associates (now Synopsys).
Rogers, "Three-bar resolution versus MTF: how different can they be anyway?," Proc. SPIE 7071, 2008.

If the 3-bar element you use has one of its harmonics near a spot where your lens system's MTF is comparatively good, it may mask the fact that the actual system performance is much worse elsewhere.

For this reason, people often test resolution by using a sub-resolution light source like a fluorescent bead. Point sources effectively contain all spatial frequencies in equal amounts, so one can obtain information about the full performance of the system. This gives the PSF; the MTF remains a useful way to quickly think about the effects of a lens on more complex images, however – the two approaches are complementary.

THORLABS

Lab 8 Notes: Contrast Methods and Abbe Theory

Optical Microscopy Course

Lab 8 Course Notes:
Contrast Methods and Abbe Theory

Overview

The point of optical microscopy is to allow us to see – that is, to provide images – of things we could not otherwise observe. It is easy to assume that what is really meant by this is a need for more magnification, or better resolution, or both, but neither of these matters if the sample remains invisible because the image has no *contrast*.

Blood cells provide a simple example of this problem: red blood cells are small, ~ 8 μm diameter, and many diseases (sickle cell anemia, malaria, etc.) can be diagnosed by looking at them. Let's say we use a 1.4 NA oil immersion objective to image a blood smear, and blow the resulting image up on our computer screen so that the cells look 8 cm in diameter – a magnification of 10,000 X, with a resolution of < 1 μm. The only problem is that we still cannot see anything: to a very large degree, blood cells are clear, so when we focus on the sample we see nothing at all (or, more likely, a little bit of dust on the slide). An example of this would be the brightfield image of the human cheek epithelial cell (Figure 1A on the following page).

The obvious way to fix this problem is to stain (dye) the cells so that they *do* absorb light, and historically this is what was done. In fact, red blood cells are known as erythrocytes because they stained well with the dye erythrosine, invented during the initial heyday of organic chemistry in the 1800s, and eosinophils (a kind of white blood cell) are so named because they stained well with the dye eosin. Figure 1B provides an example of a stained cell.

Often this is sufficient, and no more is needed; however, dyes may stain everything in a cell, or nothing, or only certain things. This can be useful for investigating different features, but the dyes may bind to the machinery of the cells, and this invariably kills them.[1] It would be nice to have a method that allows you to see everything without killing the cell – much of biology can only be done when the object of study is alive.

This brings us to the topic of this lab: contrast methods. In particular, we are going to go over brightfield, darkfield, and phase contrast; brightfield because it provides an easy-to-understand base from which to examine the other mechanisms, and darkfield and phase contrast because they flowed naturally from our investigation of conjugate planes and the Abbe theory of imaging. We will not cover DIC in any depth in these notes – it is more easily understood from the standpoint of the point spread function, and requires much more expensive equipment – but it too provides a way to look at clear cells with good contrast. Later in the course, we will explore a final contrast technique: fluorescence. See Figure 1 for examples of all of these.

It is worth underscoring the importance of these techniques; not only have they resulted in Nobel Prizes (Frits Zernike for phase contrast, in 1956; Roger Tsien, for green fluorescent protein, in 2008), but almost all modern research microscopes are outfitted with phase contrast, and phase and DIC are often used in conjunction with fluorescence. Darkfield remains one of the standard techniques in medical diagnosis. From this one can infer that contrast is a critical issue.

[1] This is not true for many fluorescent dyes, due to the way they are bound to (or incorporated in) the cells; however, these are more complicated to use. Fluorescence and fluorescence imaging are the subject of the next lab. Regardless, almost all fluorescence microscopy is done in conjunction with traditional brightfield, phase, or DIC imaging.

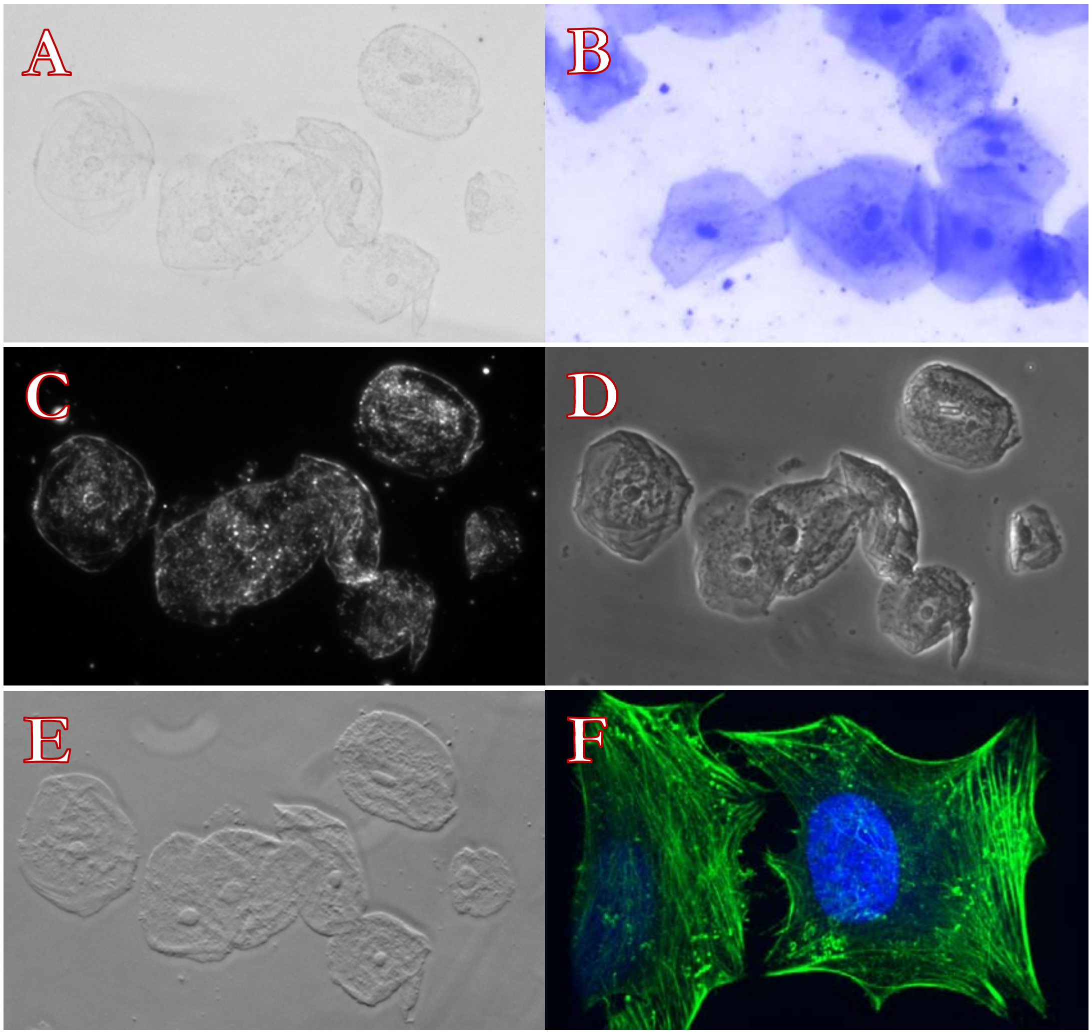


Figure 1: Human cheek epithelial cells using different contrast methods.
A) Brightfield; **B)** Brightfield, cells stained with methylene blue (and hence dead); **C)** Darkfield; **D)** Phase contrast; **E)** DIC (Differential Interference Contrast); **F)** Fluorescence, bacteria on epithelial cells stained with a dye which labels live bacteria in green and dead ones (with compromised membranes) in red. Large red nuclei are the epithelial cell nuclei. Note that the little dots in the darkfield, phase, and DIC images are bacilli similar to those labeled with dye in the fluorescence image. (Photos Courtesy of Neil Switz.)

Background Information

What is contrast?

Contrast is just the difference between the bright and dark features in an image:

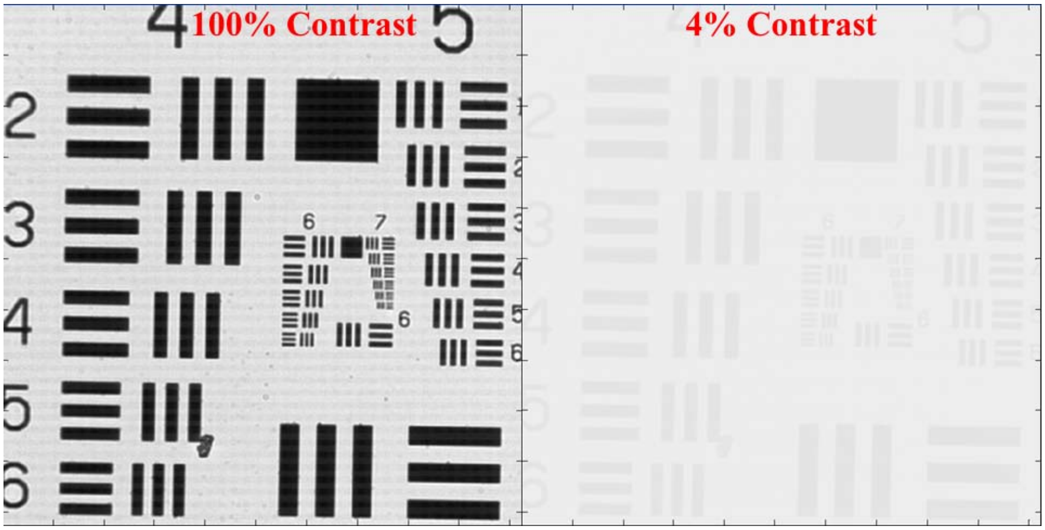


Figure 2: Image contrast. Left image of the USAF target has ~ 100% contrast (black ~ 0 counts, white ~ 240); Right image is identical but has ~4% contrast (black ~ 230 counts, white ~ 240).

Obviously higher contrast makes features easier to see; the human eye is not good at detecting contrast of < 4% or so[2]. When thinking of the MTF from last lab, it is worth considering that, by eye, one will rarely be able to detect the "full resolution" of a system because one will stop detecting image contrast by eye before the MTF actually reaches zero.

> **Note:** the Rayleigh resolution criterion corresponds to an MTF (in incoherent illumination) of ~ 9%, which is one reason it is a handy gauge of resolution: your eye can pick it up OK. The Sparrow criterion ($\delta = 0.5\ \lambda$ / NA, vs. the Rayleigh $\delta = 0.61\ \lambda$ / NA) corresponds nicely to the highest spatial frequency resolved – they are inverses of each other – but the MTF is zero there and so by eye you will never detect that level of resolution in a system. The Rayleigh criterion corresponds nicely to detail *you can see.*

Why might our contrast be bad? Major reasons include:

- **Poor Sample Absorption**: The sample may not absorb much light – if the sample absorbs only 1% of the light, you cannot get more than 1% contrast, no matter how much you turn up the lamp intensity. **Biological samples usually absorb very little light** – cells are mostly clear.

- **Low System MTF**: The MTF may be poor at the spatial frequency of the features in the sample – this will essentially always be true for features near the edge of resolution in incoherent

[2] Practically, this varies a good bit with the spatial frequency of the features in the image, age of the individual, luminance (brightness) of the image, etc. For some introductory discussion, see the links on *Visual Acuity*, *Contrast*, and *Campbell-Robson Sensitivity Chart* under the *Reference Links* tab at www.thorlabs.com/OMC.

illumination, since the incoherent MTF falls off fairly linearly toward zero. Low MTF *means* poor contrast for patterns with that spacing (spatial frequency).

- **Background**: Light leaking into the system without going through the sample (often room lights in our case) makes dark areas of the image less dark.

Regarding background, one might think that it would be simple to subtract off the background – then contrast is 100% – and then scale the remaining image so that it is bright enough to be easy to see (this is called "contrast stretching"). This is shown in Figure 3, below, which illustrates the problems with the technique. Lack of dynamic range in the image means small intensity differences between features get lost, and camera noise and noise from the intense background light start to dwarf the small signal from the sample.

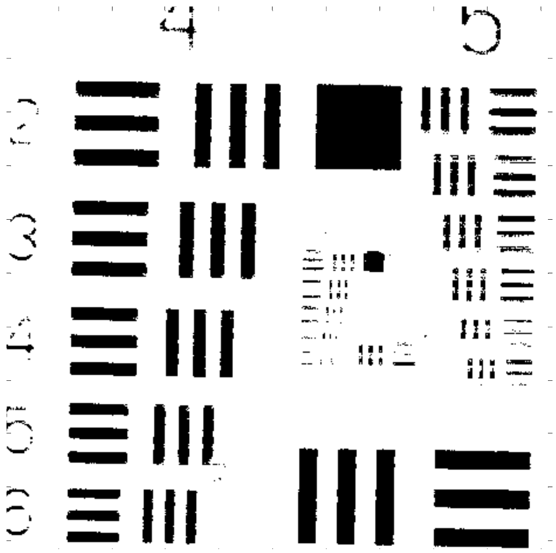


Figure 3: The problem with background subtraction. Image is the 4% contrast image from Fig. 2, with background subtracted off and pixel intensities rescaled. Contrast is now 100%, but noise and dynamic range issues in the original image (there were only 10 counts of difference between the lightest and darkest areas) result in poor quality in the processed image. All of these images were originally derived from the left-hand image in Fig. 2; compare the higher quality of that image.

Background subtraction and rescaling of images (sometimes referred to as "video enhanced" for video streams) can be very powerful when there are no other options – do not think this is a bad idea. Rather, it is not the best place to start; it is much better to improve the contrast before taking the image, and only then resort to processing tricks.

How do samples affect light?

It is worth considering how biological samples affect light, since that has a significant impact on how we achieve contrast for them. Basically, a sample can absorb light, reflect it out of the way, or slow it down. From a practical standpoint, absorption and reflection will look the same if you are imaging using transmitted light, since either way the light does not get to the camera. If one is imaging using reflected light microscopy, illuminating from the same side as the imaging optics, then obviously reflection is critical and the opposite of absorption in terms of contrast.

Let's look at these three mechanisms:

- **Reflection**: The coefficient of reflection is $R = [(n_1 - n_2) / (n_1 + n_2)]^2$. Cells are mostly salt water, which has n = 1.38 or so. Pure water has n = 1.33, so a less salty area of the cell would produce a reflection of R ~ $(0.05 / 2.71)^2$ = 0.0003 = 0.03%. However, the oily insides of some things, e.g. pollen, have an index n ~ 1.5, giving R ~ 4% if they are viewed in air. Diatoms, which you will look at in this lab, have glass skeletons with n ~ 1.5, so they generate a good bit of contrast just by reflecting light (in water, with n = 1.33, R ~ 0.4% – check the math yourself). All of this is increased for edges not parallel to the incoming light; reflection of unpolarized light is generally higher for large angles, especially angles > 45°.
    - **In Summary**: Cells do not reflect well, but larger features (diatom skeletons, oil storage areas in pollen, etc.) can.

- **Absorption**: Absorption depends on how well a molecule absorbs light, and how many molecules there are in a given area. For these reasons, biological cells do not absorb much. First, many of their molecules do not have electronic bond structures that result in them absorbing well in the visible range (though chlorophyll, rhodopsin, and hemoglobin are notable counterexamples). This is actually an advantage – excited molecules can react more easily, which can lead to damage – e.g. sunburn. Two of the three counterexamples above are molecules specifically designed for cases where the cell *wants* to detect or use light. Second, cells are very thin – even a large cell, especially when plated on glass, is ~ few µm thick. A piece of notebook paper (made up of wood cellular material) is ~ 200 µm thick, and if you hold it up to the light, not much light is blocked. Many cells together can block light acceptably, but single cells are so thin they are mostly clear. This can be changed by staining the cell with molecules that absorb light extremely well (such molecules are called dyes); however, these molecules usually bind in large quantities to various parts of the cell and in the process stop them from working, killing the cell.

- **Slowing Down Light**: This is a very important trait: cells are made of areas of slightly different index – saltier areas have indices a bit above 1.38, oily (e.g. lipid) areas have indices ~ 1.5, less salty areas have n ~ 1.33, etc. From the Lab 1 Course Notes, recall that the speed of light is v = c / n, so the higher the index, the slower light goes. When light goes more slowly, the wave crests also go more slowly and fall behind those which continued at the original speed.
    - **Example**: Let's take a 4 µm thick cell with index 1.38, and one organelle inside that is 1 µm thick and a bit oily (maybe lots of lipid), say with n = 1.52. Light going through the main cell will have a speed v = c /n = 3e8 m/s / 1.38 = 2.17e8 m/s, and take 4 µm / v = 18.4 fs (a femtosecond = 10^{-15} seconds, it takes light (in air) about 100 fs to go 1/1000th of an inch, about the width of a human hair). Light going through the organelle will take longer, since the index is higher: it takes $[3\ \mu m/v_{cell} + 1\mu m/v_{organelle}]$ = 18.9 fs. So if the light going through the organelle takes ~0.5 fs longer to get through the cell, where is the wave crest? Once the light gets out of the sample, the part of the wave crest of the light that went through the cell is still 0.5 fs behind the rest of the wave, and since light goes 3e8 m/s (roughly 1 foot / nanosecond), the delayed wave crest is 150 nm behind – for red light at 600 nm, that is 0.25 of a wavelength, ¼ λ. That may not seem like much, but remember that ½ a wavelength gives complete destructive interference. We will see that we can easily take advantage of ¼ λ to generate contrast and image cells.

It is worth looking at the nature of light a little bit more before we get into the details of contrast generation:

The Nature of Light

So far, we have mostly ignored the fact that light waves oscillate in time as well as space, but it is worth explaining why we have been able to get away with this. This material is helpful for understanding what follows, but is not critical to understanding the course in general; consider it background information.

In general, a light wave can be written as:

Equation 1: $$\boldsymbol{E(x,t) = E_0 \, cos\left(2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right)\right) = E_0 \, cos\left(\frac{2\pi}{\lambda}(x - vt)\right)},$$

where E is the electric field as a function of distance x and time t, E_0 is a constant, v = c / n, k = 1 / λ (the "spatial frequency"), and the frequency f = v / λ = c / λ_{vac}. The difference between the formulas is not actually important – they are all equivalent, so we will just use whichever is easiest for a given problem.

> **Note**: It turns out the frequency of light does not change – when light slows down, the wavelength gets shorter, so the ratio of v / λ stays the same. The details of this are beyond the scope of this course, but it should make some sense simply based on the fact that at an interface where the wave changes speed, the E-field still has to "match up," so its amplitude has to go up and down at the same rate (frequency) even though the wave speeds on either side of the interface are different. Since λ = v / f, when light slows down and the frequency stays constant, the wavelength *must* get shorter. This is part of the reason people use oil-immersion objectives in microscopy: the wavelength of light is shorter in oil (n = 1.51), so resolution gets better.

Usually an additional factor (called the "phase") is included in Eq. 1 which takes into account the possibility that the wave crest was not at x = 0 when you started timing things (i.e., at t = 0). Essentially, the phase allows you adjust things for whatever value the wave had at x = 0 and t = 0:

Equation 2: $$\boldsymbol{E(x,t) = E_0 \, cos(2\pi \, [\, k\,x - f\,t\,] + \varphi)}$$

If φ = 0, the wave was at a maximum at x = 0, t = 0; if φ = π, there was a 'trough' at x = 0, t = 0, etc.

The thing is, if we watch a wave, the *average* value is zero: it goes up and down lots of times, but the average level never changes from zero, since half the time it is positive and half the time it is negative. So what do we see?

For sound, a microphone detects the air pressure, which is like the electric field E. Radio antennas actually detect E directly. These things work because for radio or sound waves, the field is not changing very fast. At optical frequencies, the field changes much faster than the fastest detector anyone can build: f = c / λ, so for green light at 500 nm, f = 600 THz. The fastest detectors are ~ 10 GHz, which is not even close. How can they see light, then, without it averaging to zero? Conveniently, the energy of light is proportional to $|E|^2$, just as the energy in a spring is proportional to $k\,x^2$ (where k is the spring constant), and so the average of the intensity I (energy per unit time) is:

Equation 3:

$$I = \text{time average of } |E(x,t)|^2 = \text{time average of } |E_0 \, cos(2\pi \, [\, k \, x - f \, t] + \varphi)|^2$$

$$I = \frac{1}{2} |E_0|^2$$

Showing that the time average of $\cos^2(2\pi \, f \, t) = ½$ is something you can do on your own (hint: $\cos^2(\theta) + \sin^2(\theta) = 1$), but it should make sense given that the square of something is always positive – you know the average has to be some constant ≥ 0.

So let's take the case of a wave coming in at an angle to a surface (say, a camera face). What do we see?

Now we need to write the wave formula for *two* dimensions: z, along the optic axis, and y, along the camera face.

Equation 4: $$E(y,z,t) = E_0 \, cos(2\pi \, [\, k_y \, y + k_z \, z - f \, t] + \varphi)$$

where k_y and k_z are just the $1 / \lambda_y$ and $1 / \lambda_z$, the wavelengths as measured along the directions y and z:

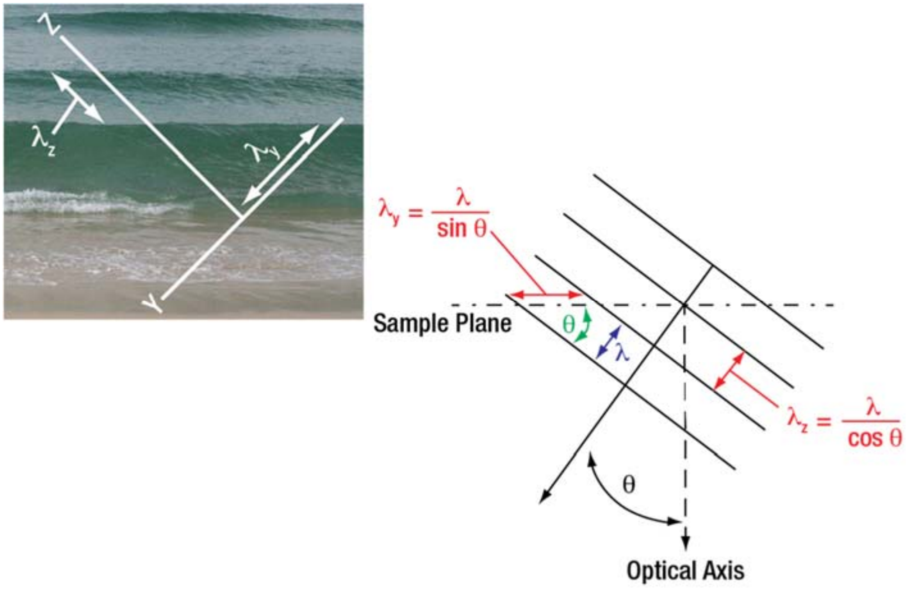


Figure 4: A wave incident on a plane. Z is the optical axis, Y lies in a plane perpendicular to the optic axis (e.g. the camera or sample plane). Note that the wavelength along any axis is the distance between wave crests as one walks along that axis. (Image Source: https://commons.wikimedia.org/wiki/File:Gentle_waves_come_in_at_a_sandy_beach.JPG, Image is licensed under the GNU Free Documentation License: https://www.gnu.org/licenses/fdl.html, Note: This image was adapted: text and arrows were added to original image.)

For convenience, let's pick the camera here as z = 0, and set φ = 0 too. Then Eq. 4 becomes:

Equation 5: $$E(y, z = 0, t) = E_0 \, cos(2\pi \, [\, k_y \, y - f \, t])$$

What intensity would we see on our camera? Per Eq. 3, I = ½ |E₀|², a constant! Of course, there are crests and troughs passing along the camera face all the time, but much too fast for us (or the camera) to detect them, so we just notice the average value.

The situation changes if we have *two* waves coming in at equal and opposite angles, though. In that case, the field becomes:

Equation 6: $$\boldsymbol{E(y,z\ =\ 0,t)\ = E_0\,[cos(2\pi\,[\,k_y\,y - f\,t]) \ +\ cos(2\pi\,[-\,k_y\,y - f\,t])}$$

$$\boldsymbol{=\ 2\,E_0\,cos(2\pi\,k_y\,y)\,sin\,(2\pi\,f\,t)}$$

We used trigonometric identities to simplify the equation. Eq. 6 has a very notable feature: when you take the time average, it is NOT a constant!

Equation 7: $$\boldsymbol{I\ =\ \text{time average of }|\,2\,E_0\,cos(2\pi k_y\,)\,sin\,(2\pi\,f\,t)|^2}$$

$$\boldsymbol{I =\ 2\ |E_0|^2\,cos^2(2\pi k_y\,y)}$$

This is why we have tended to ignore the issue of the time-dependence of light – even when it interferes, the time dependence averages out.

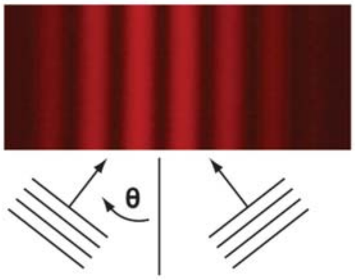


Figure 5: An interference pattern, as described by Eq. 7 (which, by the way, can be rewritten as I = 2 |E₀|² [1 + cos(π k$_y$ y)] – the pattern is sinusoidal even in intensity). (Image Source for Interference Pattern: https://upload.wikimedia.org/wikipedia/commons/8/87/SodiumD_two_double_slits.jpg, Image is licensed under the GNU Free Documentation License: https://www.gnu.org/licenses/fdl.html, Note: This image was cropped from original image.)

As described in the Lab 6 Course Notes, we can use this to figure out what angles of light come off of a sample. We will use this when we examine different contrast techniques.

Contrast Mechanisms

Preliminaries

In the discussions below, we will not need to consider all the different sinusoidal terms and different plane-wave angles coming from the sample. It is enough to remember that:

- Constant illumination (and light diffracted from large sample features) stays basically along the axis.
- Light diffracted from smaller sample features (and sharp edges) leaves the sample at higher angles.

In addition, since (as discussed previously) biological samples generally do not absorb or reflect much light, or cause the wave crests of light to get very far ahead or behind those from the rest of the light from the sample, we will represent light as follows:

- Incident illumination E = cos(ωt). Here ω = 2πf, simplifying the equation. Also, we will set $E_0 = 1$ to make the equations simpler; this does not matter in terms of contrast as long as my camera exposure is OK, because the constant E_0 multiples everything and thus divides out when we take the ratio that defines the contrast.

- Sample transmission T: this is how much of the incident E-field the sample allows through. For an absorbing sample, $0 \leq T \leq 1$ (note: $T > 1$ would require some sort of light amplification), and for a clear sample, T = 1, although we need to add a way to describe how much the light got slowed down.

  - Since cells do not affect light that much, we are going to write the sample transmission as $T = 1 + \delta\, f(y)$, where f(y) is some function (e.g. sin(ay), etc.) that has an *amplitude* of 1 (i.e., $-1 \leq f(y) \leq 1$). "δ" will be our "smallness factor," reminding us how much less than 1 the change in light due to the sample is. For us, $\delta << 1$, which is convenient since δ^2 is then so small that we can ignore it ($\delta^2 \sim 0$, since δ is small and δ^2 is small times small).

This can be summarized as follows:

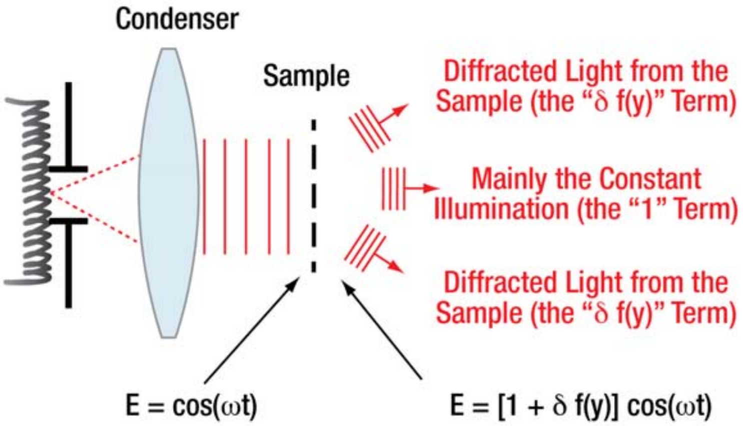

Condenser
Sample
Diffracted Light from the Sample (the "δ f(y)" Term)
Mainly the Constant Illumination (the "1" Term)
Diffracted Light from the Sample (the "δ f(y)" Term)
E = cos(ωt)
E = [1 + δ f(y)] cos(ωt)

Figure 6: Light scattered from a sample. For simplicity, the amplitude $E_0 = 1$.

Of course, for conservation of energy, we cannot really have light scattered into various angles and not also have some decrease in the light intensity along the axis, but since so little light is scattered ($\delta << 1$), we can ignore this.

Brightfield Contrast

We already basically know what happens in brightfield: intensity from a point on the sample is mapped (via the PSF) to some point in the image. So we expect to see that the image looks a lot like the sample transmission of light *intensity*, which is exactly what happens.

> **Note**: Remember, we defined the transmission T as how much of the *E-field* got transmitted; that definition is $\sqrt{T_{intensity}}$, since $I = |E|^2$. So the intensity after the sample is $I = [1 + \delta f(y)]^2 = [1 + 2\delta f(y) + (\delta f(y))^2]$; do not get confused by the squares floating around; those are expected given how we defined T.

Assuming we get all the light from the sample to the camera, with no losses or aberrations, and ignoring magnification, the E-field at the camera will then be basically the same as in Figure 6:

Equation 8: Brightfield E-Field at camera: $\boldsymbol{E = [1 + \delta f(y)] \cos(\omega t)}$,

where (again) we have set the amplitude E_0=1. Which then gives:

Equation 9: Brightfield Intensity at Camera:

$$\boldsymbol{I = time\ average\ of\ |E|^2 = 1 + 2\delta f(y) + [\delta f(y)]^2}$$

Where for simplicity we have ignored the factor of ½ from the time average of the cos(ωt). This result is just the same as the intensity distribution after the sample.

Now it is important to define *contrast* clearly. Contrast is the difference between the brightest thing and the darkest thing in the image, and our image is an image of the *intensity* of the light – that is, the square of the electric (E) field. So, since f(y) can be positive or negative, the contrast from Eq. 9 will involve finding the difference between the brightest and darkest features, i.e. between $+|\delta|$ and $-|\delta|$:

Equation 10: Brightfield Contrast $= \dfrac{\boldsymbol{I_{max} - I_{min}}}{\boldsymbol{I_{max} + I_{min}}}$

$$= \frac{\left[1+2\delta f(y)+(\delta f(y))^2\right] - \left[1 - 2\delta f(y)+(\delta f(y))^2\right]}{2}$$

$$\approx 2\,\delta\, f(y)$$

where we have ignored any factors of δ^2 or higher since they are very small (that is also why we set the denominator ($I_{max} + I_{min}$) = 2 in the equation above).

This is an important result: for weakly absorbing samples, brightfield intensity – what we see – *contrast* is proportional to how much of the electric field (E) is blocked by the sample. (Conveniently, in this limit, that turns out to also be essentially the same as the fraction of the intensity blocked by the sample, which is what we really expect.) To rephrase that: **brightfield contrast scales linearly with the sample E-field transmission.**

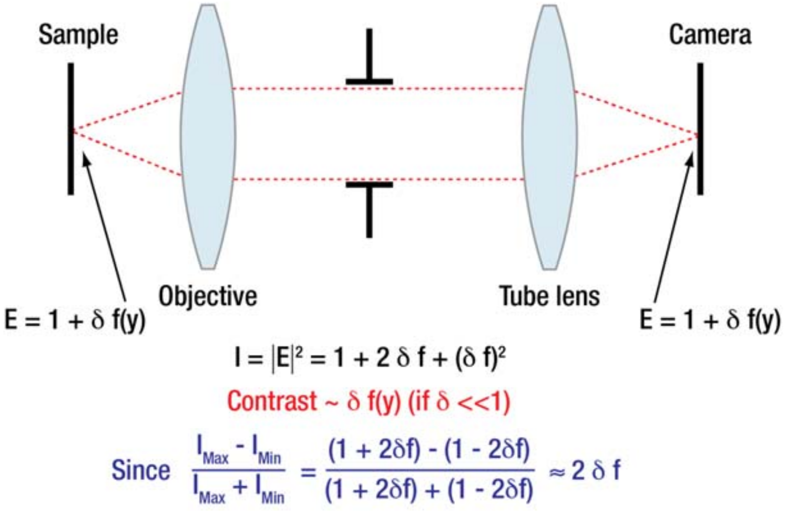

Sample
Camera
Objective
Tube lens
$E = 1 + \delta f(y)$
$E = 1 + \delta f(y)$
$I = |E|^2 = 1 + 2 \delta f + (\delta f)^2$
Contrast ~ $\delta f(y)$ (if $\delta << 1$)
Since $\frac{I_{Max} - I_{Min}}{I_{Max} + I_{Min}} = \frac{(1 + 2\delta f) - (1 - 2\delta f)}{(1 + 2\delta f) + (1 - 2\delta f)} \approx 2 \delta f$

Figure 7: Brightfield Contrast. Note contrast is linearly proportional to the sample transmission.

This is OK if the sample absorbs a good bit of light – i.e., if δ is reasonably large. But if δ ~ 0.04, then we get ~ 4% contrast, which is a bit hard to see (recall Fig. 2). How could we improve this?

Apodized Brightfield

Looking at Figure 2 should be a tip-off: if only we could reduce that *constant light*, the signal that is the same over the entire image. Of course, we do not want to make everything dimmer – dimming the overall illumination will not help, since the scattered light (the signal we want) is proportional to that in the first place. But we could reduce the constant term *after* the light had passed the sample:

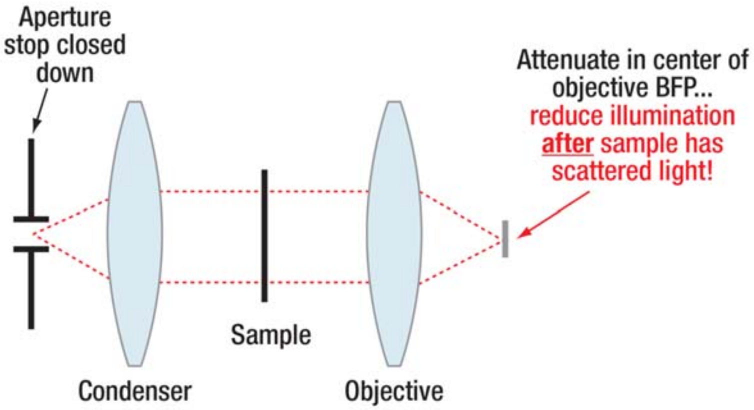

Aperture stop closed down
Attenuate in center of objective BFP...
reduce illumination after sample has scattered light!
Sample
Condenser
Objective

Figure 8: Reducing constant illumination *after* the sample.

Putting a light-attenuating filter in the BFP of a lens is known as "apodizing," which literally means "removing the foot" since it is often used to smooth out the sharp edges of the transmission (due to the iris) in the BFP that lead to the fringes (or rings of the Airy disk) – the "feet" at the edges of the main feature – in the image.

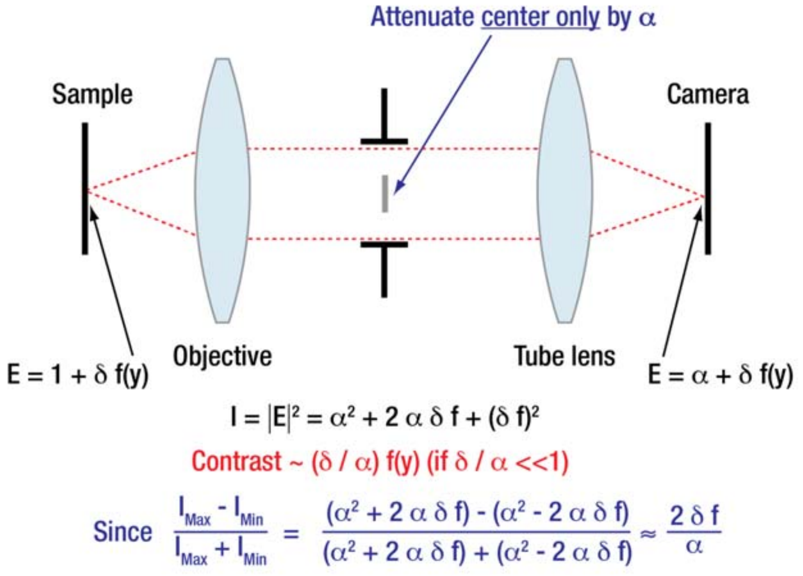

Attenuate center only by α
Sample
Camera
Objective
Tube lens
$E = 1 + \delta f(y)$
$E = \alpha + \delta f(y)$
$I = |E|^2 = \alpha^2 + 2\alpha\delta f + (\delta f)^2$
Contrast ~ $(\delta / \alpha) f(y)$ (if $\delta / \alpha << 1$)
Since $\frac{I_{Max} - I_{Min}}{I_{Max} + I_{Min}} = \frac{(\alpha^2 + 2\alpha\delta f) - (\alpha^2 - 2\alpha\delta f)}{(\alpha^2 + 2\alpha\delta f) + (\alpha^2 - 2\alpha\delta f)} \approx \frac{2\delta f}{\alpha}$

Figure 9: Apodized Brightfield: Contrast enhanced by 1 / α, i.e., 1 / (attenuation of the main illumination).

Doing this reduces the light incident on the whole camera face to a fraction α of the original, *without reducing the intensity of the light scattered from the sample* ***at all***. As long as α > δ (so we do not accidentally change the sign of parts of the E-field compared to others), this is the same as just subtracting the background off the final image, with the advantage that it is done *before* the camera senses the light – so nothing is lost in terms of camera dynamic range, or extra noise, etc.

As in the case above, where δ ~ 4%, if we choose to attenuate the middle of the BFP by a factor of 10 – allowing α = 10% of the light through – our contrast rises to ~ δ / α = 40%, very easy to see! Of course, there is less total light on the camera, so we have to turn up either the incident illumination (which also increases light scattered from the sample) or the exposure. But that is not a problem: turning up the illumination now helps, because that also increases the light scattered by the sample features, and then the background light is reduced by our apodization. Thus, the amount of light representing *changes* due to the sample – the signal we want – has been increased relative to the constant background light level.

Darkfield

Of course we can take this farther by blocking *all* the constant illumination in the BFP (setting α = 0). You have already done this in lab using the ball driver, and seen the resulting darkfield image (named darkfield because the background in the image is dark).

It is worth looking at this carefully to see how it affects the contrast:

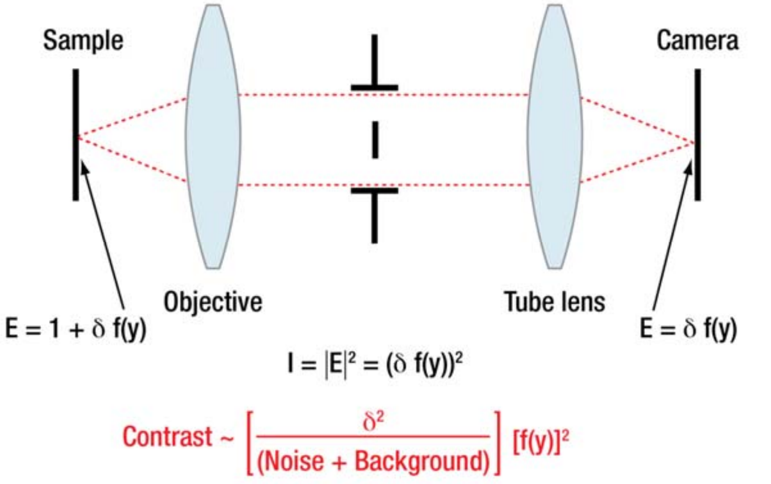


Figure 10: Darkfield: Contrast enhanced by 1 / α, i.e., 1 / (attenuation of the main illumination); darkfield uses α ~ 0, so in theory contrast approaches $\frac{1}{0}$ which goes to infinity. In reality, however, the background light limits how small the denominator can get.

Here there is no constant illumination, so technically the contrast should be $\delta^2 / 0 = \infty$, but nothing is perfect: electronic camera noise, leakage light (e.g. from room lights), etc. provide some small baseline signal on the camera. As always, but especially in darkfield, the better this is controlled, the better your contrast. Contrast of > 25 is not hard to do at all, even with the open-frame systems in the class. Compare that to 4%, which is amazing: a factor of ~ 600 improvement. Since we mostly use the USAF target, which has inherently great contrast (δ ~ 1), this may seem like no big deal; however, as we move on to look at smaller objects (e.g. cells, and also 100 nm diameter beads) the advantage will become obvious – these scatter very little light (objects smaller than a wavelength scatter an intensity proportional to r^6, so the amount of light scattered drops *very* fast as the particles get small) – and so would be completely invisible without using darkfield. Of course, the scattered light *is* small, and the contrast ~ δ^2, so one really needs to turn up the illumination. This is one reason it is important to use a collector lens, etc., to get as much light as possible from the lamp. For techniques like darkfield on small particles (and especially at high magnification – usual when looking at small things, but which spreads the light out over an M^2 larger area, making it proportionally dimmer), often an arc lamp or laser is necessary to get enough light.

There is a second important thing to notice about the contrast one gets from darkfield illumination: it is proportional to the *square* of the sample transmission i.e. δ^2, whereas we found brightfield contrast to be linearly proportional to the transmission – i.e. proportional to δ.

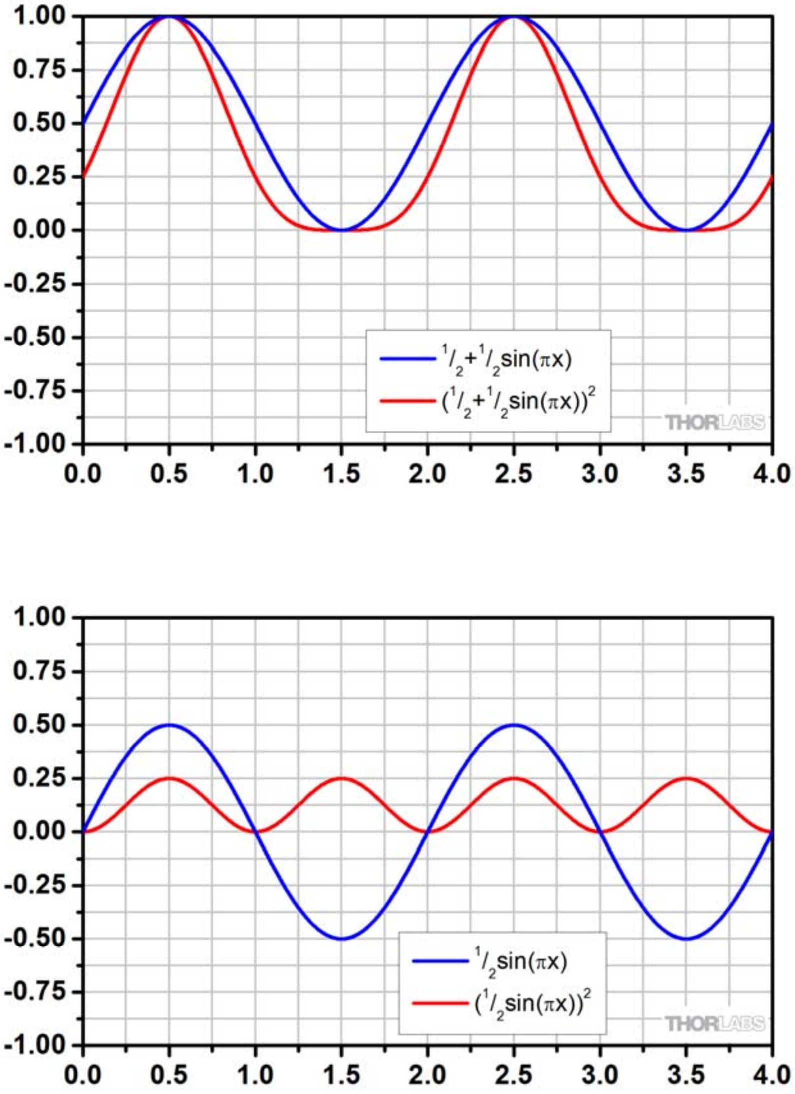


Figure 11: Effect of Squaring: The square of [½ + ½ sin(ax)] (Top, red trace) does not look like the square of [½ sin(ax)] (Bottom, red trace), even though both have the same initial amplitude (blue traces). Squaring rectifies the negative part of sin(ax) when no constant is added. This effect shows up in darkfield due to the $f(y)^2$ contrast term, though exact effects are hard to guess in advance. In the simple example above, notice that the bottom red trace has *twice* the frequency of the upper red trace – darkfield can make an image look different from "reality."

As noted, it is hard to guess in advance how this will affect the overall image; however, in general both regions brighter AND darker than the image average will show up as bright in a darkfield image, which can make such images harder to interpret.

One powerful advantage of darkfield (in addition to being fairly cheap to implement) is that it is also sensitive to *clear* objects, and since most cells are essentially clear, this is a huge advantage. We will cover why darkfield can make clear objects visible after our discussion of phase contrast, which is the more popular way to achieve visibility for clear samples.

Phase Contrast

If our sample is clear, then its transmission T = 1, which would seem to be bad news in all of the above analysis, since all of the contrasts go to zero if $\delta = 0$. Of course, there is a trick: as mentioned earlier, clear samples can slow light down by different amounts in different areas. Anywhere there is a difference in the time it takes light to get through a fixed distance – either because the sample is actually thicker, or because it contains material through which light goes slower or faster than the surrounding area – provides a change in the shape of the E-field coming out of the sample.

The easiest way to think of this is to consider some part of the sample where the light was slowed down enough that the wave crest of the light falls behind the rest of the wave by ½ λ; then as the wavefront comes out of the sample, that area might be negative (a trough in the wave) while the surrounding area will be positive (a crest of the wave). Of course, one can still "walk along" the top of the wave crest – see Figure 5 in the Lab 1 Course Notes – but the wavefront will no longer be a line (or plane); in some places it will bend ahead or drop behind. An analogy often drawn is to a marching band going across a field, with a big muddy patch in the middle. Because people in the formation who have to wade through the mud go more slowly than the others, they get behind, and the rows of people get bent – by the time the formation is across the field, the lines of people (analogous to the wave crests) are no longer straight.

Of course, what we care about most is what the E-field is right after the sample. At each point, are we on a crest, trough, or in between? This is very similar to the question we asked about absorbing samples; there the wave was, say, a crest all the way across, but the crest had different heights at different places, since some of the E-field had been attenuated by sample absorption. For a clear sample, the crests are always the same height – there has been no absorption – but some are delayed relative to others so as you walk along a plane just after the sample, sometimes you are on a crest and sometimes in a trough.

In both cases, the E-field varies along the plane just after the sample, so why can we not see clear samples well when we focus the light onto the camera?

Let's return to the case of the marching people: one way to describe how they move is to ask how fast each column of them can go at each point in the field, i.e., and figure out how long it takes each person to get across the field. Another (mathematically easier) way is to assume they all move the same speed, but to adjust the "effective distance" they cover based on how fast they can move there – if they march 3 mph on average, but in mud they go 1 mph, then another way to get the same final result for how long it takes them to cross the field is to say that mud counts for 3x the distance (= normal speed / mud speed): 15 m of walking through mud = 45 m of regular walking.

For light, this is easy: $n = c / v$, so if we multiply all distance traveled by n we will have the "effective distance" traveled. This is used so often it has a name: the "optical path length," or OPL.

Returning to the example from Part 1B, light going through a cell 4 μm thick with index n = 1.38 will have traveled an OPL of $n * \Delta z = 5.52$ μm. Light going through the oily organelle with n = 1.52 will have traveled an OPL of (1.38 * 3 μm + 1.52 * 1 μm) = 5.66 μm. The difference in the OPL is 0.14 μm, ~25% of a wavelength of 600 nm green light.[3] Hopefully it is clear how this way of working the problem is easier using the OPL than the original method involving travel time.

[3] The difference between this number and that in the earlier example in Part 1B is just due to rounding error based on significant digits. If you carry more digits, the answers turn out to be identical.

The critical thing to notice in all this is that the optical path along which light travels (the optical path length, OPL) will vary for different paths through the sample depending on whether the index is higher or lower there. Hence, let's refer to the total OPL at a given sample location y as the OPL(y).

> **Note**: Really this is OPL(x, y) – since the sample has two dimensions, x and y – just as f = f(x, y) above. In this document we are generally ignoring the x-dependence (thus assuming a 1-D sample) in order to keep notation simple.

Using this, we can look at what the E-field looks like at a given moment *after* the sample, assuming we illuminated with a plane wave, say $E_0 \cos(\omega t)$:

Equation 11: $\boldsymbol{E \;=\; E_0\, cos(2\pi\, [\, k\, z - \, f\, t])}$ **(Basic Illumination Wave)**

If we set z = 0 at the plane just before the sample, then after the sample z = OPL(y), the optical path length the light went through. Since we are looking at some moment in time, t is the same along the whole plane after the sample, so we have:

Equation 12: $\boldsymbol{E(y, z = \text{just after the sample}, t) \;=\; E_0\, cos(2\pi\, [\, k\, OPL(y) - \, f\, t])}$

Before, the cosine term did not change at all – the absorption only affected the ***size*** of E, which became $E(y) \sim E_0\, [1 + \delta(y)]$. Now it is a bit more complicated: the change shows up in the OPL, which is *inside* the cosine term.

We would like to get a simple expression like before, where we have a term which is basically the illumination, and some term representing a small change in the E-field due to the sample. We do that below, but **we do not expect you to be able to reproduce the math**. Rather, follow it, and then make sure you understand the results; the results are what we will be most concerned with.

To begin with, we can write the change in the OPL through the sample as

Equation 13: $\boldsymbol{OPL(y) \;=\; OPL \;+\; \Delta_{OPL}(y),}$

Where OPL = the average OPL through the sample. That allows us to rewrite Eq. 12 as

Equation 14:

$$\boldsymbol{E(y, z = \text{just after the sample}, t) \;=\; E_0\, cos(2\pi\, [\, k\, OPL \;+\; k\, \Delta_{OPL}(y) - \, f\, t])}$$

$$\boldsymbol{= E_0\, cos(2\pi\, [\, k\, OPL - \, f\, t] \;+\; 2\pi\, k\, \Delta_{OPL}(y))}$$

Note that the last term in Eq. 14, $2\pi\ k\ \Delta_{OPL}(y)$, looks just like the phase, φ, from Eq. 2 – in fact, it ***is*** a phase: the varying index in the sample has resulted in a phase variation in the wavefront as the light exits the sample. Not surprisingly, using this to somehow get contrast is called 'phase contrast'.

To make the phase nature of this more explicit, let's define the term as

Equation 15: $\boldsymbol{\delta\, \varphi(y) \;\equiv\; 2\pi\, k\, \Delta_{OPL}(y)}$

where δ is a smallness term, to remind us how small the changes are (like before), and φ(y) is a phase that varies in the sample with amplitude ~ 1. With this, Eq. 14 becomes:

Equation 16:

$$\boldsymbol{E(y, z = \text{just after the sample}, t) \;=\; E_0\, cos(2\pi\, [\, k\, OPL - f\, t] \;+\; \delta\, \varphi(y))}$$

Now we are ready to introduce the only derivative in the whole course: assuming the change in the argument of a function [say, g(x)] is small, then according to Taylor's theorem we can write:

$g(x + \Delta x) = g(x) + \frac{dg(x)}{dx} \cdot \Delta x + \cdots$ In our case, this becomes cos(argument + small change) ≈ cos(argument) – sin(argument) * (small change).

Using this to rewrite the E-field (Eq. 16), we have

Equation 17: $\boldsymbol{E_0\, cos(2\pi\, [\, k\, OPL - f\, t] \;+\; \delta\, \varphi(y))}$

$$\boldsymbol{\approx E_0\, [cos(2\pi\, [k\, OPL - f\, t]) - [\delta\, \varphi(y)\,]\, sin(2\pi\, [k\, OPL - f\, t])\,]}$$

Which looks like a big mess. But it is not so bad, actually: let's look at each term in turn:

1. **E$_0$:** Everything is proportional to the incident field strength. Big surprise.

2. **cos(2π [k OPL – f t]) :** This is a constant factor, independent of sample position y. It is the main illumination, which has not changed much. As before: constant = plane wave on axis.

So far, this is fairly familiar.

3. **[δ φ(y)] sin(2π [k OPL– f t]) :** this is the light affected by the sample.

    a. **[δ φ(y)] :** this is the "small change," which needs to be << 1 in order for the approximation to be valid. Essentially this represents all the features in the sample.

    b. **sin(2π [k OPL– f t]) :** the time evolution of the small term is now a sine instead of a cosine, so the wave crests always 90° out of phase (which is the same as π/2, or ¼ λ in distance) with the incident light. That will turn out to matter a lot.

Before we go on, though, let's investigate the requirement that δ φ(y) << 1. From Eq. 15, this is the same as saying 2π k Δ_{OPL}(y) << 1, and since k = 1 / λ, that is Δ_{OPL}(y) << λ / 2π. So now we know what is required from the changes to be "small enough" to get away with this approximation: the changes in the optical path length through the sample must be (much) smaller than a wavelength.

From the earlier example of the organelle, it should make sense that this will almost always be true for individual biological cells: 1 μm is a huge organelle, and the assumption that it is completely full of oil (instead of, say, salt water with an index of nearly the same as the rest of the cell) is pretty aggressive, yet the difference in OPL was only 25% of a wavelength. Even thicker things with big index variations – diatom skeletons, pollen grains, etc. – generally will not fit this requirement that the phase shift be small, and thus, are not ideal samples to look at in phase contrast.

Returning to the E-field, from now on we will not care that much about the arguments of the sine and cosine terms – they all change by the same amount as we go through the optical system to the camera. Furthermore, let's set E$_0$ = 1 again to simplify things further, and for similar reasons incorporate the factor of ½ into the δ. Then we can rewrite Eq. 17 as:

Equation 18: $\boldsymbol{E(y, z = \text{just after the sample}, t)\ \approx\ cos(\omega t) - \delta\, \varphi(y)\, sin(\omega t)}$

Look back at Fig. 6 and Eq. 8; this is very similar, except that instead of

Absorbing Sample: $\boldsymbol{E \approx cos(\omega t) + \delta\, f(y) cos(\omega t)}$

we have

Clear Sample: $\boldsymbol{E \approx cos(\omega t) - \delta\, \varphi(y)\, sin(\omega t)}$

So what are the main differences in the E-field from a clear sample instead of an absorbing sample? Although the approximations tend to hide the fact that absorbing samples reduce the field intensity and clear ones do not, the big differences between absorbing and clear samples are:

1. An absorbing sample reduces the overall field ***amplitude*** (and thus intensity) of the wave, but the whole wave stays in phase – the wave crest is still in a line, but the wave height changes.
2. A clear sample changes the ***phase*** of the wave crest exiting the sample – it becomes squiggly – but does not reduce the field amplitude (so the waves remain the same height).
    a. In fact, the field changed by the sample ends up being exactly 90° out of phase (wave crests ¼ λ behind) the illumination field – that is the sine vs. the cosine term above.

Naturally, most samples have a little of both – phase shift and absorption – but many samples fall closer to one extreme than the other. For instance, live cells tend to be pretty clear, and are called "phase objects," while stained cells are fairly absorbing, and are hence called "absorption objects."

We can now address the question of why clear samples do not produce contrast at the camera: in brightfield, the E-field from the sample is imaged to the camera, and what is detected there is the time-average of the square of the field. Using Eq. 18, this is:

Equation 19: Brightfield E-Field at Camera $\boldsymbol{E\ =\ cos(\omega t)\ +\ \delta\, \varphi(y)\, sin(\omega t)}$

Which then gives

Equation 20: $\boldsymbol{|E|^2\ =\ cos^2(\omega t)\ +\ 2\,\delta\, \varphi(y)\, cos(\omega t)\, sin(\omega t)\ +\ \delta^2\, |\varphi(y)|^2\, sin^2(\omega t)}$

Now the time average of $\sin^2(\omega t)$ and $\cos^2(\omega t)$ = ½, but what about the time average of $\cos(\omega t)\cdot\sin(\omega t)$? This is still an oscillatory function, going equally positive and negative, so over time is averages to zero. As a result,

Equation 21: Brightfield Intensity at Camera $\boldsymbol{|E|^2\ =\ 1\ +\ \delta^2\, |\varphi(y)|^2}$

where as usual we have ignored the factor of ½ from the time average of the sinusoidal terms.

Remember that $\delta << 1$, so δ^2 is so small we can ignore it – that is why clear objects are invisible: they only show up as δ^2 terms. Actually, they do not even do that: if you keep all the terms in the Taylor expansion, even the δ^2 term will cancel out, leaving you with nothing (as you might expect for samples that do not absorb at all).

So how can we see them? Well, first of all, in darkfield you can, which should make sense: if you block the center of the BFP (the constant illumination), Eq. 20 gives you the same result as for darkfield on an absorbing object[4] – so darkfield works the same way for clear samples as for absorbing samples; the contrast is the same, etc.

Of course, we did not get this far just to show that darkfield works; after all, virtually every biology lab has a phase contrast microscope (or many such scopes), and for good reason: we have already noted the disadvantages of darkfield. So what about phase contrast?

Well, the problem we have is that the light scattered from the sample does not interfere constructively or destructively with the main illumination light – it is always halfway in between. If we could do something to prevent that cos(ωt)·sin(ωt) term in Eq. 20 from averaging to zero, we would be effectively back to brightfield!

All that is necessary to do that is to change the terms in the E-field so they are both in phase (i.e., both cos(ωt) or both sin(ωt)).

It has probably occurred to you already that this could be done by introducing some extra material (say, glass) into the objective BFP, right on axis where the illumination light will go through it, but the scattered light from the sample will not. And that is basically how it is done.

Optional Exercise: If you put a glass plate ($n = 1.50$) in the BFP of the objective, how much extra thickness do you need on-axis to get a 90° phase shift compared to air ($n = 1$)? Does this depend on wavelength? Should you use your green filter?[5]

Once you shift the phase so the waves are all cos(ωt) functions (or sin(ωt) – it does not matter as long as they are all the same), the rest is exactly like brightfield. In fact, because clear samples slow the light down so little – our organelle example was actually a significant exaggeration – contrast that is proportional to δ is still hard to see. As a result, phase contrast objectives (they are special objectives because they have the phase plates built into the BFP, which is normally inside the objective where it is hard to get to) are almost always also apodized to further increase contrast.

Apodized brightfield is almost unheard of, but phase objectives usually have ~ ND 1.0 apodization (10% intensity transmission) in addition to the phase plate, increasing contrast by a factor of ~ 10 compared to no apodization. This is a problem if you also want to do fluorescence imaging, since that apodization blocks a lot of the weak (and thus precious) fluorescence emission. As a result, you generally never use phase contrast objectives for fluorescence imaging.

This is all summarized in Figure 12:

[4] This remains true even if you keep all the terms in the Taylor series; subtracting the constant term makes all the difference.

[5] Answers: glass in center must be ¾ λ thicker; yes; yes.

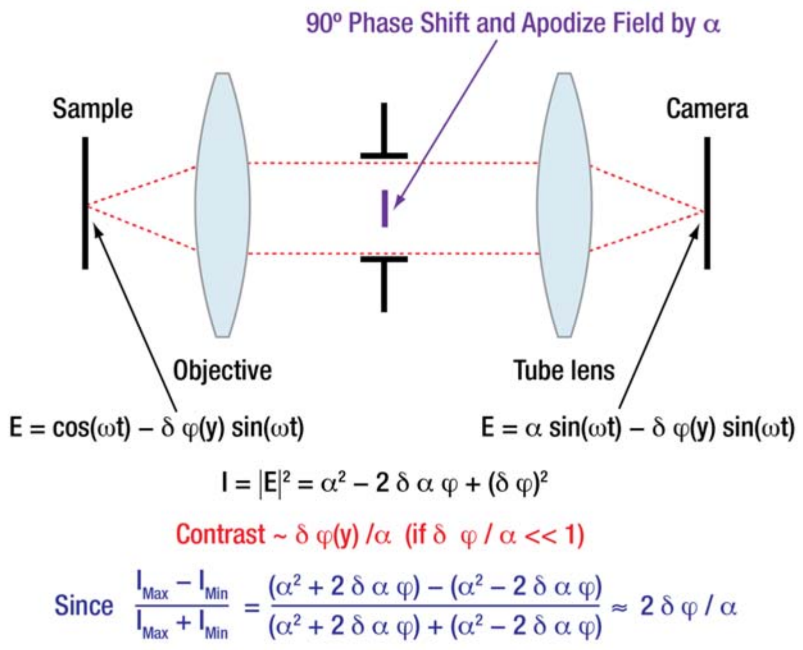

90º Phase Shift and Apodize Field by α
Sample
Camera
Objective
Tube lens
$E = \cos(\omega t) - \delta\,\varphi(y)\sin(\omega t)$
$E = \alpha\sin(\omega t) - \delta\,\varphi(y)\sin(\omega t)$
$I = |E|^2 = \alpha^2 - 2\,\delta\,\alpha\,\varphi + (\delta\,\varphi)^2$
Contrast ~ $\delta\,\varphi(y)/\alpha$ (if $\delta\ \varphi/\alpha << 1$)
Since $\frac{I_{Max} - I_{Min}}{I_{Max} + I_{Min}} = \frac{(\alpha^2 + 2\,\delta\,\alpha\,\varphi) - (\alpha^2 - 2\,\delta\,\alpha\,\varphi)}{(\alpha^2 + 2\,\delta\,\alpha\,\varphi) + (\alpha^2 - 2\,\delta\,\alpha\,\varphi)} \approx 2\,\delta\,\varphi/\alpha$

Figure 12: Phase Contrast: shifting the phase of the background (illumination) light by 90° allows it to interfere destructively and constructively with the light scattered by the clear sample. Apodizing further enhances *contrast*, which is linear in the phase shift from the sample and, thus, *proportional to the variations in the sample optical path length*.

Phase and Darkfield: Annular Illumination

Although the above discussion is accurate, the illumination method described is never used. There are two reasons for this:

1. Just as with darkfield, a tiny opening in the aperture stop means very little light gets through. As noted before, cells scatter very little light, so when using these techniques it is especially undesirable to make inefficient use of the one's lamp by closing down the aperture stop.

2. If the aperture stop is closed down, the condenser NA is essentially zero, which adversely affects resolution [since $\delta = 1.22\,\lambda\,/(NA_{obj} + NA_{cond})$].

With darkfield this is fairly easy to solve: one uses a donut-shaped (called "annular") mask in the condenser aperture, so that the ring of illumination imaged into the objective BFP is too large for the BFP aperture and all light gets blocked (see Fig. 18 of the Lab 7 Course Notes). Another way of saying the same thing is that all of the incident illumination is coming in at angles beyond the ability of the objective NA to collect. Since the annulus has a much larger area than a small spot, a lot more light gets through, and the illumination is furthermore much less coherent, so fringes are reduced. Resolution is also then at a maximum, since the condenser NA is very high (though our theory does not explain exactly what the MTF will look like).

For phase contrast, all the same things apply, with one exception: if the annulus is too large for the objective BFP, there is nowhere to put the phase plate! Given that, the obvious thing to do is to choose the annulus such that it falls just inside the objective BFP aperture, which would be great, except that objectives usually have nasty phase aberrations in the very last bit of their aperture (essentially, for the very highest NA rays), and having the phase correctly handled is the key to the technique. Consequently the annular image of the aperture stop mask is usually designed to fall around ~ 70% of the diameter of the objective back aperture.

Since the back aperture has different diameters for different objectives, one illumination annulus (aperture stop mask) is not sufficient: most microscopes have three different diameter rings for use with different NA objectives, and one can select the appropriate one

Question: Are the larger rings for use with higher or lower NA objectives?[6]

Phase objectives have the phase plate built in to the BFP, and the image of the illuminated aperture stop mask ***must*** overlap that exactly if the phase is to be shifted properly. This requires alignment: whenever you set up phase contrast, or darkfield, you should check the objective BFP to make sure things are properly aligned. You can do this either by flipping a lever that places an extra lens (called the "Bertrand lens") into the system which allows you to see the BFP, or by just pulling out an eyepiece and looking into the tube with your eye.

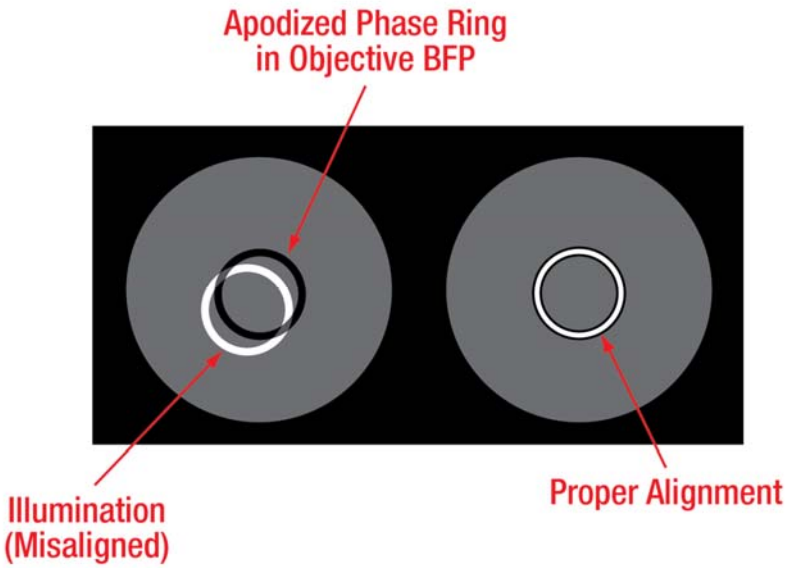

Apodized Phase Ring in Objective BFP
Illumination (Misaligned)
Proper Alignment

Figure 13: Aligning phase rings: when looking at the objective BFP, the annular illumination must fall on top of the darker apodized phase ring in the objective, or else contrast will be effectively brightfield instead of phase.

Phase "Halos"

If you look at the phase contrast image in Figure 1, you will notice a bright "halo" around both the edge of the cell and the nucleus. This halo feature does not appear in any other contrast technique shown, and is characteristic of phase contrast images. What causes it?

[6] Answer: higher; remember that rays of a given NA fall at a BFP diameter = 2·f·NA.

A tip-off is that these halos would not be noticeable if the illumination ring and phase annulus were infinitely thin; however, in practice they are not, both because one needs finite area to get enough light through, and because alignment is never perfect. So the phase ring in the objective is always sized a bit too big on purpose, to prevent any illumination from sneaking past without being phase-shifted (and thus spoiling the contrast).

It is perhaps easiest to return to the central-spot phase system shown in Figure 12. Remember that *larger* features in the sample diffract light into *smaller* angles (the angle of the diffracted light is NA = λ/a, where $1/a$ is the spatial frequency of the sample feature, so a is its approximate size). If the phase spot in the BFP is not infinitely small, some diffracted light corresponding to large sample features will go through it in addition to the transmitted illumination light; if the diffracted light from those features is phase shifted 90° at the same time the illumination is also shifted, then they will remain 90° out of phase and thus not interfere at the camera. Consequently, some low spatial frequencies are always missing from the image – it is analogous to subtracting off a very low-resolution (because the spot is not very big, and so corresponds to a very low NA) version of an image from the full-resolution version. The blurrier image naturally looks a bit larger (since it is blurred out), and subtracting it causes changes (halos) around the edge of the original image.

In phase contrast, these halos may be dark or light, depending on whether the illumination phase is advanced or retarded in the BFP; however, they are always present to some degree. Using annular illumination rather than the central spot does not change this, but in fact adds a glitch (droop) in the MTF at higher spatial frequency as well, where the diffracted spots from illumination on the far side of the ring fall into the opposite side of the phase annulus, again losing contrast. Despite this, the annular illumination is significantly superior to the central spot method, which is why nobody uses the spots.

THORLABS

Lab 9 Notes: Fluorescence Microscopy

Optical Microscopy Course

Lab 9 Course Notes:
Fluorescence Microscopy

Overview

We are now entering the final labs of the course, in which we will cover fluorescence. It is hard to overstate the importance of fluorescence for biological and medical research, and advances in fluorescence labeling techniques have resulted in a renaissance in microscopy over the past several decades. We will try to cover some of the details of those developments, but our main focus will continue to be the specifics of the actual imaging.

We will also need to move relatively quickly to cover the material while leaving time for projects before the end of the term. With that in mind, a note on readings:

References

1. The Lab 9 Course Notes (this document) and Lab 9 Lab Notes, will (as always) be the most important documents.
2. Revisit the "Equations to Memorize" document in Appendix A, and make sure you are up on all of them – most will now be very familiar, but some (especially the one regarding collection efficiency) are new.
3. Recommended reading: *Handbook of Optical Filters for Fluorescence Microscopy*, pages 1 – 23, and p. 29, is good background; see the *Reference Links* tab at www.thorlabs.com/OMC.

Why Fluorescence?

So far we have looked at a number of contrast mechanisms:

- **Brightfield**; best for *absorbing* or highly scattering transparent specimens – absorbing features are dark on a light background. Often dyes ("stains") are used to cause certain sample features to absorb. A particularly well-known and famous example is Gram's stain (hence the medical term "Gram-positive" bacteria). These stains usually kill the cells, and so are only good for examining dead samples. See the *Reference Links* tab at www.thorlabs.com/OMC for more information on Gram staining and Gram positive bacteria.
- **Phase contrast**; best for *transparent* specimens with some phase differences. Dyes are not required for phase; instead relative optical thickness of the sample appears as darkness or brightness against a grey background. The big advantage of phase is that one can examine *living* cells. This was revolutionary, resulting in the 1953 Nobel Prize in physics for its inventor, Zernike, and is now standard on the vast majority of tissue culture microscopes and biological research microscopes in universities and medical centers.
- **Darkfield**; best for *very weakly absorbing or scattering* (transparent) samples, typically very small objects. The main benefit of darkfield is that there is *no background* at all, so any sample feature shows up bright against a black background and is easier to see (higher contrast). Note that one cannot distinguish an absorbing sample from a phase sample using darkfield. Versions of darkfield are not uncommon to find on tissue culture microscopes, though phase contrast is used more frequently. To

illustrate the power of darkfield, it is easy to see 20 nm particles (1/25th the wavelength of green light) using a decent darkfield microscope!

In one sense, as we will cover shortly, fluorescence is simply another contrast mechanism: if something has been fluorescently labeled in some way, then by using the proper techniques, one can image the glowing fluorescence against a black background, providing astonishing (and beautiful) contrast.

The real power of fluorescence, however, lies in the fact that biochemical and genetic techniques have been developed to allow incredibly specific labeling of cellular features- one can now *label individual molecules* within a cell. Attaching fluorescent dyes to immune-system antibodies (which recognize very specific biomolecular features) was first done by Coons, et al., in 1941, but it was not until 1958 that Riggs et al. (in Berkeley) made the chemistry tractable and the technique really began to take off. Immunofluorescence contributed hugely to the revolution in biological imaging, since one could now image and localize specific molecules within a cell. However, it is very difficult to get antibodies *inside* a cell without killing it. Consequently, most imaging continued to be either of cell surface features or of dead cells, albeit with much higher specificity – one could now label exactly a specific protein-type of interest. The development of GFP (Green Fluorescent Protein; 2008 Nobel Prize, Chemistry) and its variants by Tsien, et al. (mostly at UC San Diego) provided yet another remarkable tool: the ability to genetically express a fluorescent marker attached to a protein of interest. That means one can track, in a living cell, exactly where a protein of interest is going, or what cellular assembly or process it is part of, or when it is produced, etc. – an astonishing capability.

Beyond simply allowing for labeling of specific cellular features (down to the molecular level), fluorescence is inherently a molecular event itself- the dye molecule is usually a few angstroms (Å) to a nanometer in diameter. As a result, the ability to localize the source of the light is not necessarily limited by the wavelength. As long as one is only looking at an isolated dye molecule, one can know almost exactly where it is; this advantage underlies some new techniques in "super-resolution imaging" which allow for nm-scale resolution (subject of the 2014 Nobel Prize in Chemistry), though this requires additional special chemical and imaging techniques.

So fluorescence is a technique, which allows for *molecular contrast*. Even if (as in most fluorescence imaging) resolution is no better than in brightfield, one knows that where there is any light at all, the specific feature of interest must be there.

Some Examples

It is impossible to cover the breadth of fluorescence methods in a quick overview, but some examples will be illustrative:

Lipid membranes are critical to cell existence – they define the boundary with the outside world, and all communications or actions by a cell ultimately must involve the membrane. Membranes, being made up of fatty acids, can exist in fluid or ordered (congealed) phases, and this affects the localization of receptor proteins on the cell surface, the ability of the cell to release neurotransmitters or other molecules, and many other membrane processes. Yet lipid bilayers, being only ~ 4 nm thick and absorbing or scattering little light, are very difficult to image in their (native) aqueous environment, let alone investigate in detail. Fluorescent labels, which preferentially localize to ordered or disordered areas of the lipid membrane, have opened the door to investigation of the effects of these domains and the details of the related membrane mechanics. An example is shown in Figure 1.

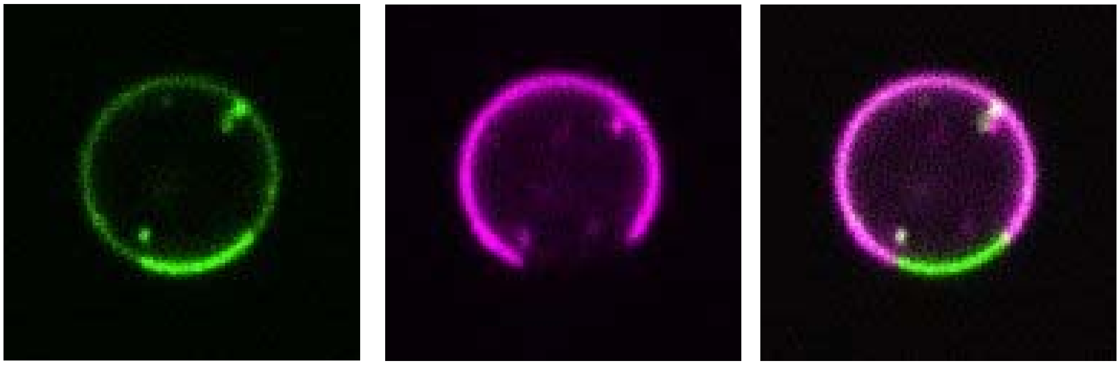


Figure 1: Phase-separated giant plasma membrane vesicles generated from HeLa cells. Imaged with confocal fluorescence microscopy, with (from left to right) liquid-ordered phase in green (labeled with cholera toxin subunit B-Alexa 488), liquid-disordered phase in magenta (labeled with Eff1-mCherry), and merged. (Photos courtesy of C. Chan, UC Berkeley.)

This ability to image at multiple wavelengths (colors) for different probes allows great flexibility in interrogating a cell. For example, the nucleus of a cell can be easily labeled separately from the (~10 nm wide) actin fibers that make up part of the cytoskeleton and which are required for cell motility (shown in Figure 2). Without fluorescent labels, the actin network would be completely invisible, and the dynamics of its organization (important for cell signaling, cancer metastases, and immune function) would remain unclear.

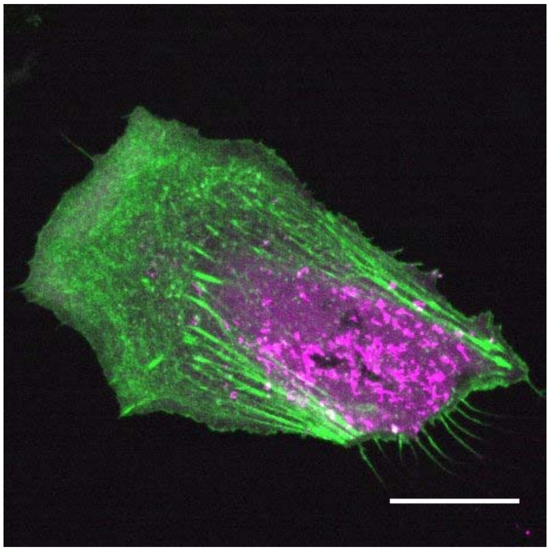


Figure 2. HeLa cell imaged with confocal fluorescence microscopy. Green: actin (labeled with Lifeact GFP). Magenta: cytosol and endosomes (labeled with fluorescent protein p14-mCherry). Scale bar 10 μm. (Photos courtesy of C. Chan, UC Berkeley.)

Fluorescence techniques are hardly limited to work with live cells, but the ability to image objects which may be below the normal optical resolution threshold, or to distinguish overlapping (co-localized) features in a time-resolved manner, while a cell is alive and functioning in its normal environment, are especially useful.

FRAP, Fluorescence Recovery After Photobleaching, provides an example of the use of fluorescence to obtain dynamic information about cellular behavior. The technique makes use of what is often a liability in fluorescence imaging: the fact that the excited state of the fluorophore is more reactive than the ground state and tends to photodestruct (usually due to oxidation while in the excited state) after a sufficient number of excitations. The mobility of molecules (e.g. cell-surface receptors) in and on a cell has major implications in biology. FRAP provides a method for quantitatively measuring the mobility of species *in vivo*: one simply bleaches all the fluorophores in a given area, and watches the recovery of signal in that area as unbleached molecules diffuse in from the unbleached surroundings. Figure 3 illustrates this: the lipid bilayer is bleached, and the fluorescent molecules then diffuse back in, showing that the bilayer is fluid and also allowing an estimate of molecular mobility in the membrane.

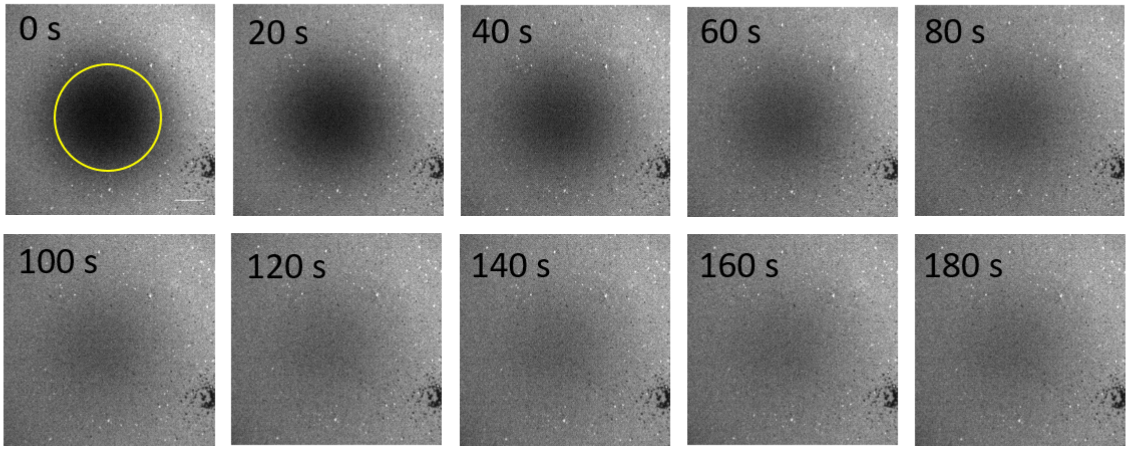


Figure 3: FRAP (Fluorescence Recovery After Photobleaching) experiment on supported lipid bilayer. Bilayer is composed of DOPC with 0.3% DOPE labelled with fluorescent dye Atto390. Photobleached area is shown in yellow circle. Scale bar 10 µm. (Photos courtesy of C. Chan, UC Berkeley.)

The range of fluorescence techniques gives some measure of the breadth of the field: there is regular widefield fluorescence microscopy, FRAP, TIRF (total internal reflection fluorescence), confocal and two-photon laser scanning fluorescence microscopies, fluorescence polarization/anisotropy and fluorescence lifetime imaging, as well as non-imaging techniques like ELISA (enzyme-linked immuno-sorbent assay), real-time PCR (polymerase chain reaction), FCS (fluorescence correlation spectroscopy), and more.

How Does Fluorescence Work?

Fluorescence is in many ways very simple: you shine light on a molecule, it absorbs some of that light and emits light of a slightly longer (redder) wavelength. You collect that light, using appropriate filters to block out the original excitation wavelengths, and detect the remaining fluorescent signal. Sometimes the molecules bleach, just as dyed cloth fades in the sun. Those basics – absorption, emission of a longer wavelength, filtering out of the excitation light, and collection of the fluorescence for detection – cover all the main issues.

It is worth going over these parts in a simplistic way before delving into the details:

Excitation

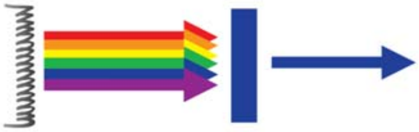
Fluorophores are dye molecules, and they absorb light just like any dye does, which is a good place to start:

Question: What color(s) does red food coloring absorb? How about blue food coloring?

Often one uses a laser to excite the fluorophore, and lasers (usually having a single wavelength) provide a very pure color excitation. However, sometimes it is useful to use a lamp (often a very bright arc lamp), since then you can easily choose what color you want to shine on your sample.

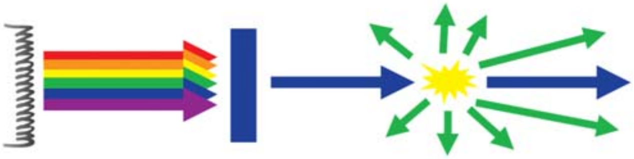
Question: What color light goes through blue stained glass?

So a simple excitation set-up looks like this:



Absorption and Emission

Assuming we shined the correct color of light on the dye, the molecule will absorb some of the light. If the dye is fluorescent, it will then emit light of its own – but with some energy lost to molecular vibrations, so the emitted light is *redder* than the absorbed light. So now the system looks like:



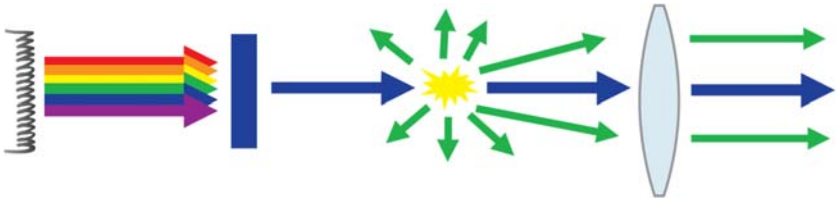
Dyes fade, which is to say, a fluorophore will only work for so long before it is dead. Consequently it is important to collect as much of the light as possible while the molecule is emitting (and before it bleaches).

Collection

Fluorophores emit light in random directions, so the more of these directions that eventually get focused onto our detector, the more signal we get. The main way of collecting a lot of light is to put as large a lens as possible as close to the sample as possible.

Question: If we put a big lens right up against the side of a molecule, what is the largest possible percentage of emitted fluorescence we can collect?

So now we have:



Filtering

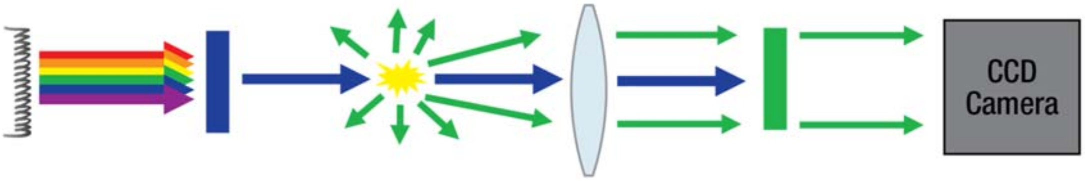
In order to see the fluorescence, we must separate it from the (usually much brighter) excitation light. Usually one uses a filter, just like before.

Question: What color(s) of glass will block blue light?

Detection

Not all detectors are equally efficient – some (like your eye, when dark-adapted) detect nearly all the light that hits them; others detect maybe 1% of the incident photons. The "Quantum Efficiency" of a detector is the percentage of incident light that it will convert into a signal. More is usually better, but also more expensive.

So the final set-up looks roughly like this:



Of course, in practice the details of the technique underlie both its power and limitations, so they are worth investigating in more detail.

Since the molecule's absorption and emission spectra determine much of the rest of the apparatus, we will begin by examining the chemistry involved in fluorescence:

Fluorescence Excitation

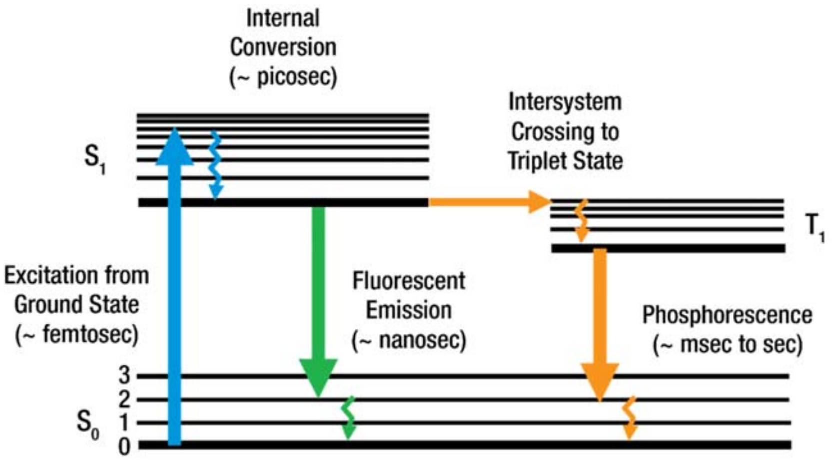


Figure 4: A Jablonski Diagram of the Energy-Level Transitions Involved in Fluorescence

At room temperature, virtually all molecules are in their electronic ground states (S_0) since the energy required for excitation to the first excited electronic state is usually ~2.5 eV, which is ~100 k_BT, giving a probability of thermal excitation (Boltzmann or Arrhenius factor) of ~ e^{-100}, effectively zero. To a less extreme extent, it is usually the vibrational ground state v_0 that is populated as well. When a photon of appropriate energy is absorbed by (hits) the molecule, it can excite an electron to a higher electronic energy level (blue arrow in Figure 4).

What is an "appropriate energy"? Any energy equaling the difference between an (occupied) S_0 vibrational state and an excited electronic state S_1 vibrational state will do. Since there are many vibrational states v_n to choose from, there are a variety of different photon energies that will do the job. This gives rise to the absorption spectrum of a molecule (see Figure 4), since the wavelength is related to the various acceptable energies E by $\frac{hc}{E}$, where h is Planck's constant and c is the speed of light. Since the electron tends to mainly be in the vibrational ground state to begin with, the excitation spectrum gives a measure of the vibrational energy states of S_1, which are numerous and densely distributed for fairly large molecules. Consequently, there are many slightly different energy photons which can be absorbed, which gives rise to the first major point:

➔ **Fluorophores typically have broad excitation spectra.**

Fluorescence Emission

Regardless of what vibrational state the electron got excited *to*, it rapidly (~ 10^{-12} sec) thermally relaxes (i.e., without radiating any light) to the lowest vibrational (v_0) state of S_1. As a result, most fluorescence emission happens from the same state (S_1, v_0), regardless of the excitation photon energy (wavelength). This is important, since it implies that:

➔ **Emission spectra do not depend much on excitation wavelength.**

After ~10^{-9} sec (called the fluorescence lifetime) in the excited S_1, v_0 state, the excited electron will drop back down to *some* (possibly excited) vibrational state of the electronic ground state S_0 (green arrow, Figure 4). Thermal relaxation from this vibrational state to the vibrational ground state v_0 is again quick. In a manner similar to the excitation spectrum, the emission spectrum will thus give information about the vibrational ground state energies. Also, the fluorescence emission will be of lower energy (and hence longer wavelength) than the absorbed photon, due to the thermal losses in the relaxations to the vibrational ground states of S_1 and S_0. This "red-shift" (since longer visible wavelengths are closer to the color red) is known after it is discoverer as the "Stokes Shift." Which is the second major point:

➔ **Emission spectra are also broad, and are always red-shifted relative to the excitation.**

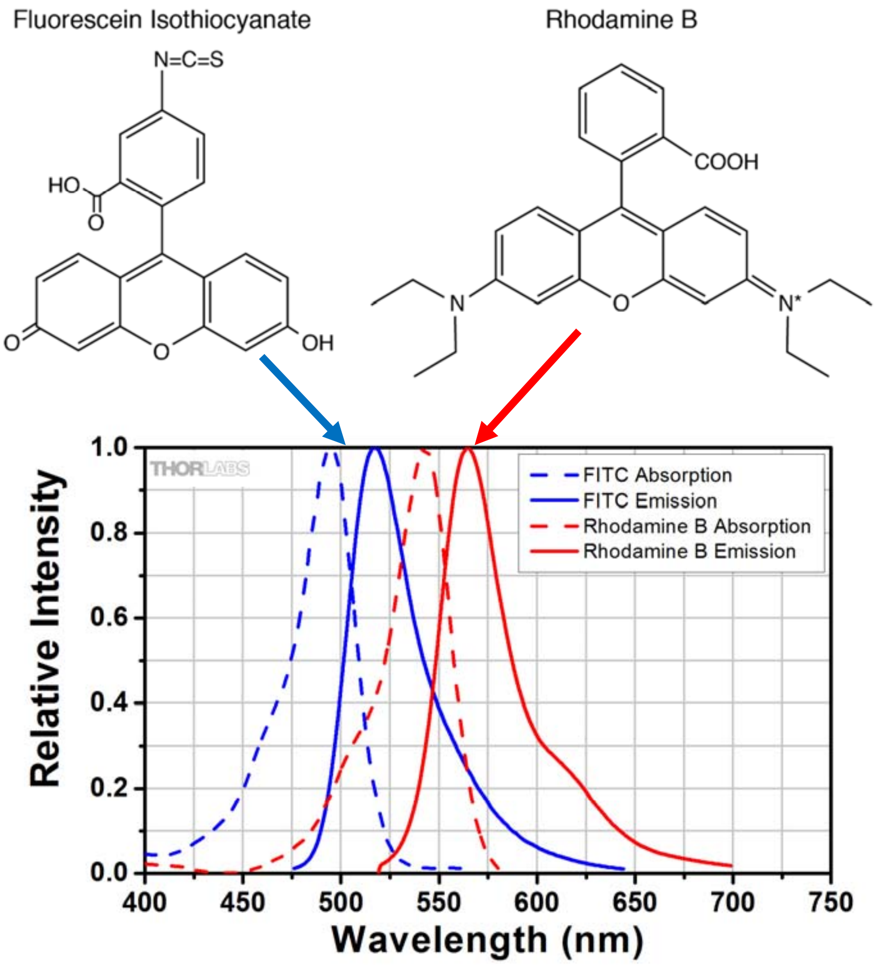


Figure 5: Two widely used fluorophores, Fluorescein Isothiocyanate (FITC) and Rhodamine B. Note the similarity in structures; the changed groups on Rhodamine B (compared to FITC) result in a shift in the absorption and emission spectrum toward longer wavelengths. Most common fluorophores are aromatic compounds.

Although fluorescence need not be in the visible wavelengths, the most widely used fluorophores are those which excite at wavelengths above 190 nm (where air starts to transmit), and often at 400 nm or above (roughly where inexpensive glass lenses used to route the excitation light start to transmit). Naturally emission wavelengths for those fluorophores in wide use are shorter than ~700 nm, which is the far red end of the spectrum visible to your eye.

While the breadth and shape of the absorption and emission spectra can be tuned by modifying the vibrational energy levels of the ground and excited electronic states, the rough position of the peaks is changed by modifying the underlying energy difference between the S_1 and S_0 states. While this may sound complex, it amounts in practice to (often clever) selection of modifying groups: see Figure 5.

Limits on Emission

A final note regarding fluorophores: there are de-excitation pathways for the excited electron that do not involve fluorescence emission. One of these is a "triplet-state-crossing" (see Fig. 4), where the electron shifts to a state from which radiative decay is quite slow. This (much slower) decay is known as phosphorescence, and typically has an even larger red-shift.

A more important de-excitation method is by chemical reaction of the molecule such that its ability to fluoresce is irreparably destroyed. Most typically this is due to oxidation, which occurs much more easily when the molecule is in an excited state.

One way of estimating when a fluorophore will bleach is to assume that it will happen when the time spent in the excited state multiplied by the rate constant for destructive reaction from that state equals unity, i.e. when $(n\ \tau_F)\ k_d \sim 1$, where n is the number of excitations before bleaching, τ_F is the fluorescence lifetime (i.e. the average time the electron spends in the excited state after each excitation), and k_d is the rate constant for oxidation or other photodestruction from the excited state. Approximate values for these numbers are $\tau_F \sim 2.5$ ns and $k_d \sim (600\ \mu s)^{-1}$, giving an average value for the number of excitations (and hence emitted fluorescence photons) before bleaching of n ~ 250,000. This number depends strongly on the molecular environment: for instance, oxygen scavengers like β-mercaptoethanol can significantly reduce k_d and hence increase the number of emitted photons before bleach, and dyes protected from oxygen entirely (e.g. when encased in plastic) are extremely hard to bleach. In aqueous solution, however, the number given is a good rule of thumb:

➔ **A fluorophore in solution emits 10^5 to 10^6 photons before irreversibly bleaching.**

How much does bleaching really matter? After all, the smallest resolvable spot in a standard light microscope has a volume of about 1 femtoliter. Even assuming a (typical) low concentration of fluorophores, say 1 μM, which still leaves 10-100 fluorophores in each resolvable volume, thanks to the largeness of Avogadro's number. Even if one needs many photons for an image (and you do), that is usually still plenty – after all, you get ~10^5 photons from each fluor, and so you might get ~100 images before bleaching.

This can be acceptable if you are using a camera to record a single image, and do not need to spend any time focusing, but viewing by eye requires >10 frames/second (video rate is 30 fps), and this adds up fast – 100 frames is less than 10 seconds of direct viewing, and each frame requires a bunch of (maybe 1000) emitted (and collected!) photons to form an image.

Worse, cellular structures are often much smaller than the limit of the optical focus – for instance, a cell membrane receptor may be <10 nm across. Even this smaller area often cannot be fully labeled – to avoid interfering with the biological function of the structure, one may only be able to label it sparsely, maybe say with 5-20 fluorophores. If one wants to be able to take enough pictures to focus and then track the dynamics of the receptor for some time, it becomes critical to get *and collect* as many photons as possible from each fluorophore before it bleaches.

Light Collection

While the excitation light is usually traveling in a specific direction through the sample, the photons emitted by the fluorophore do not preserve this directionality – they for the most part are emitted isotropically[1], i.e.:

➔ **Fluorescence emission occurs randomly in all directions.**

This reduces the issue of collection to one of *solid angle.*

What *is* "solid angle"? The solid angle is basically the fraction of the horizon your lens fills, as seen from the molecule. Usually lenses are round, and if we place a molecule at the focal point, the angles which intercept the lens will define a cone. This is easier to see than to explain:

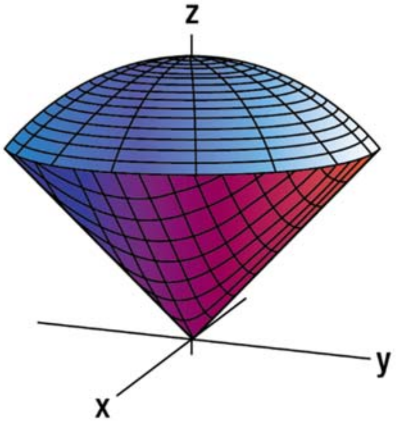


Figure 6: Picture of a cone; fluorescent molecule would be at the point of the cone, emitting light in all directions. The angles of light collected by the lens (objective) define a cone; if the light is emitted in all directions equally (called "isotropically") then the percentage of the light that is collected by the lens will be proportional to the area of a sphere that the cone intercepts (the blue cap in this figure).

You can see that the outer edge of the lens limits the rays that can be collected from the molecule. Or, more accurately, the angle defined by lines from the molecule to both the center and edge of the lens – it will not matter if you have a lens of radius *r* a distance *d* away, or whether you have a lens of radius *2r* a distance *2d* away, the total fraction of light (angles) will be the same.

We are, of course, already familiar with the angles of light collected by a lens, given by the Numerical Aperture:

Equation 1: **Numerical Aperture:** $\boldsymbol{NA = n\,sin(\theta)}$

where θ is the half-angle of the cone defined by the molecule and the lens, and *n* is the index of refraction of the medium the lens is in. The *n* turns out to be helpful when considering the resolving power of a

[1] Specialists will recognize that there are details here we are omitting.

lens, which depends on the index of the medium you are focusing into, but it does obscure the geometrical interpretation we care about here.

It is not possible to make a decent lens system that collects much more than NA ≈ 0.9 or so in air.

Question: What angle θ does NA = 0.9 (in air) correspond to? What is the full angle from one edge of the "horizon" to the other? How close is this to the maximum you could possibly get with a lens on only one side of a molecule?

Question: What is the angle θ for a 1.4 NA objective in oil or glass (n = 1.51)?

Question: What is the angle θ for a 1.2 NA objective in water (n = 1.33)? Any comments on how all these numbers compare?

It would clearly be useful to know how much of the light emitted (in all directions) by a fluorophore can be collected by a given lens. This is equivalent to asking, "what portion of the total horizon is occupied by the lens?"

The full horizon, of course, is the surface of a sphere with the fluorophore at the center, which has a surface area of $4\pi r^2$. So the question can be reduced to, "what portion of that area is chopped out by the cone defined by the lens?"

A picture can be helpful here:

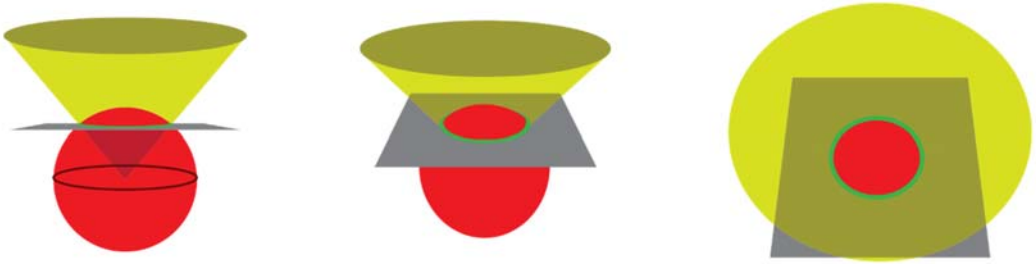


Figure 7: A cone intersecting a sphere. The percentage of collected light will be equivalent to the portion of the sphere surface inside the cone; if the cone angle is small (small NA approximation), the surface of the sphere inside the cone will be pretty flat, so one could approximate it by the area of the green circle.

For some this calculation may seem trivial, but few people actually know how to do it. Deriving it is much easier than looking it up (it is not tabulated many places), and the integral is surprisingly simple, so we will do it here:

The natural coordinate system is spherical coordinates, i.e., *r*, θ, and φ.

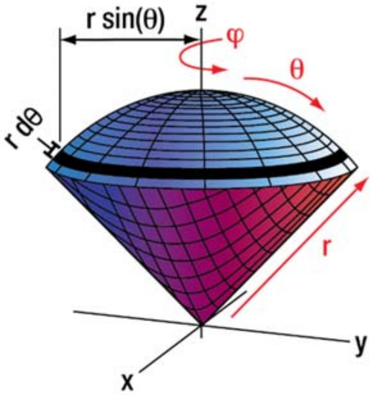


Figure 8: Coordinates for finding collection efficiency. Circular strip has width *r dθ* and radius *r sin(θ)*, for a total area of $dA = 2\pi r^2 \sin(\theta)\, d\theta$. This is derived below.

The first thing is to ask, "what is the area of a tiny bit of the surface?" Looking at Fig. 8, we can see several things:

1. We are talking about the surface of a sphere, which is *defined* by constant *r*, so we will not need to integrate over *r* at all.
2. The total area of the cap can be constructed of a bunch of circular stripes, so all we need to know is the area of each stripe and then add them all up.
3. Each stripe will have a width of *r* dθ and a circumference (note that this tacitly includes the integration over φ) of 2π r sin(θ), so its area *dA* will be $dA = 2\pi r^2 \sin(\theta)\, d\theta$

So the area of the cap is given by:

Equation 2: $$\boldsymbol{A = 2\pi r^2 \int_0^{\Theta} \sin(\theta)\, d\theta}$$

And that area divided by $4\pi r^2$ will be the fraction of the total area of the sphere this represents (and equivalently, the fraction of light collected), so the collection efficiency will be:

Equation 3: Fraction of the total area that is collected $$\boldsymbol{CE = \frac{1}{2}\int_0^{\Theta} \sin(\theta)\, d\theta}$$

Problem #1: Do this integral, and divide the result by the total area of the sphere to get the ratio. Then fill in the following table (magnifications are only to give an idea of typical values; n = 1). Leave the last two columns blank for now and just fill in the columns for the "Half-angle Θ" and "Collection %."

Magnification	NA	Half-angle Θ	Collection %	Small angle %	Relative Brightness
5X	0.15				1.0
10X	0.25				
20X	0.45				
40X	0.65				
60X	0.9				
100X	0.9				

Problem #2: Rearrange the result from your integration using the trigonometric identity $\frac{1}{2}[1-\cos(\theta)] = \sin^2\left(\frac{\theta}{2}\right)$ to get something in terms of a sine (rather than cosine) function. Then use the small-angle approximation $\sin(\theta) \approx \theta$ twice to cast your answer in terms of the NA (assume n ~ 1) and you should get something like the small-angle approximation from the "Equations to Memorize" document. Fill in the "Small angle %" column, above. Compare it with your exact results in the "Collection %" column: at what NA does the small-angle approximation start to break down?

Problem #3: If you double the NA, roughly how much extra light would you expect to collect? (If you used a calculator for that, or even looked at your table, you have missed the point of Problem #2). If you double the magnification, how much extra area will you spread your collected light over? (Think carefully about that.) The "relative brightness" of an image will be the ratio of the amount of light you collect to (i.e., divided by) the image area you spread that light over. Fill in the sixth column of the table above, calculating the brightness of the other objectives compared to the 5X objective (i.e., after calculating all the brightness numbers, divide them all by the brightness for the 5X), and comment on the results. Do you need a brighter light to use a high-magnification objective? Why do you think there are no 100X, NA 0.15 objectives? If you have a dim sample, would you rather use the 60X or 100X 0.9 NA objective?

Filters

Once you collect the light, you need to filter out the excitation. One can do this with colored glass, or a colored liquid – an early experiment on the blue fluorescence from quinine, an anti-malarial drug and a main component of tonic water, made use of a glass of yellow wine as a filter. Generally, though, one needs to be able to select narrow bands of light and to completely block ("reject") excitation light from getting to the detector. Usually colored glass will not do the trick well enough – almost everyone uses special filters called "interference filters." These work on the same principle as the antireflection coating on eyeglasses: on top of a glass surface there are a number of thin layers with different refractive indices; light reflects off each of these layers, and the reflected beams can combine to interfere in a way that either increases or decreases the net reflection (or, conversely, the net transmission, since T = 1 – R).

We will discuss filters briefly now, and in more detail next lab (enough detail that you can intelligently choose your own filters if you ever need to).

Sometimes you know the fluorophore you will use (e.g., it is the only one you can buy commercially already linked to the antibody you want), but sometimes you have several choices, and choose the fluorophore based on the laser or lamp you have available for excitation, etc. Let's assume we want to work with EGFP, a fluorescent protein. GFP and its variants are a very important set of fluorophores because one can use genetic engineering techniques to attach their DNA sequences to the DNA sequences of cellular proteins you might want to investigate. Then, when the cell expresses the gene for the protein you are interested in, it also makes a fluorescent part attached to it (these are called Chimeric proteins, after the monstrous Chimera from Greek mythology, whose body had the front of a lion, the middle of a goat, and the rear of a snake. Similarly, these proteins have parts from multiple different proteins all linked together). There are now a variety of these FP (Fluorescent Protein) constructs, mostly generated from an original jellyfish protein, the original GFP (Green Fluorescent Protein). EGFP is just a better (brighter, hard to bleach) version of GFP – the "E" stands for enhanced.

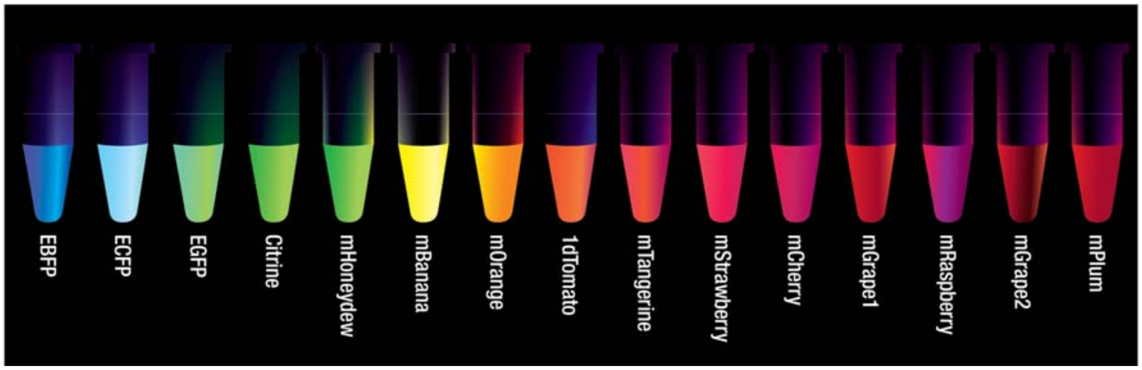


Figure 9: A variety of fluorescent protein variants.
(Image modeled after "Composite of Fluorescent Cells," Tsien Lab)

To choose a filter, first you need to examine the fluorophore's excitation and emission spectra. These can be found on many websites; see the References listed earlier, or just Google. For EGFP, the spectra are:

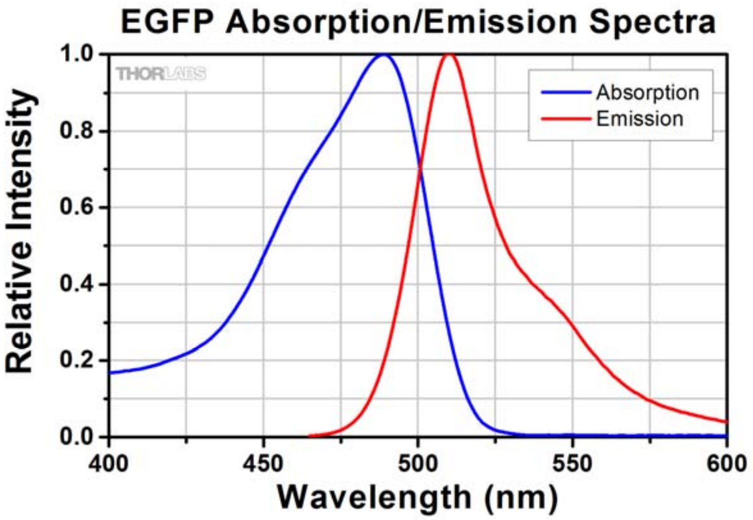


Figure 10: EGFP Absorption/Emission Spectra

Second, you need to decide how you will excite the fluorophore(s). If your source is a laser, then its spectrum will usually be quite narrow and you may not need to filter your excitation light. A lamp (e.g. an arc lamp), however, puts out light at many wavelengths (this is what makes the light look "white"), and you will need to filter it so that you can tell the emitted fluorescence apart from the lamp light. Let's say we have a Xenon arc lamp; its spectrum looks roughly like:

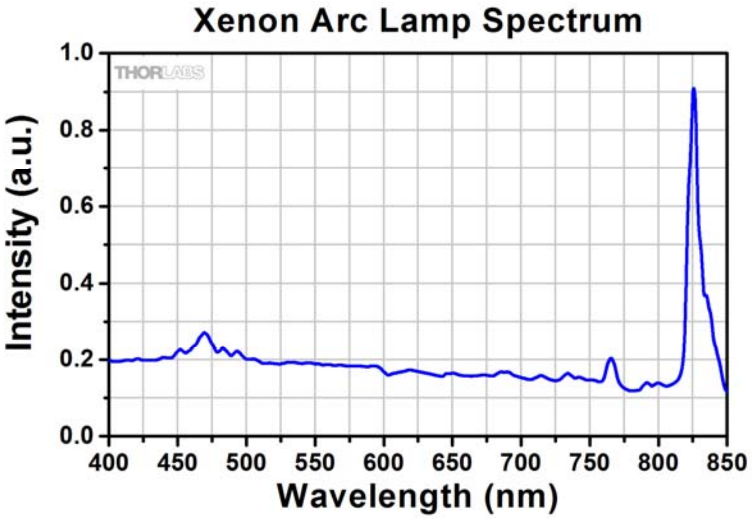


Figure 11: Xenon arc lamp spectrum (EXFO is the manufacturer).
(Source: McNamara-Boswell Spectra Compilation)

The nice thing about Xenon lamps is that their spectrum is rather flat in the visible (Hg lamps, in contrast, have very sharp peaks at specific wavelengths). You can see a "small" peak at ~ 830 nm in the figure above (sharp peaks are often 100+ times as bright as the surrounding spectrum) and naturally, you would not want this to get through to your camera and wash out the tiny fluorescence signal.

Looking at the EGFP spectrum (Figure 11), we can see that exciting between 450 and 490 nm will give the most excitation (since those wavelengths overlap the absorption) without overlapping the emission spectrum (the signal we want to detect). Similarly, collecting the emission light between 500 and 550 nm will give us most of the emitted fluorescence without picking up scattered light (or lamp light that has leaked through the first filter) in the region where the fluorescence signal is falling off. Consequently, we might choose filters like those below, where the filter curves show where the filters transmit light.

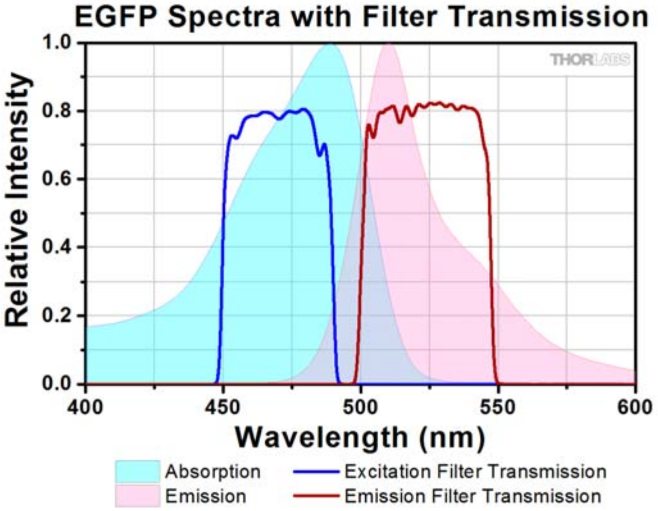


Figure 12: EGFP spectra and transmission data for the Excitation and Emission filters.
(Source: McNamara-Boswell spectra compilation and Chroma Technologies)

There is one last item regarding filters: for convenience, it is often better to do the illumination and the light collection from the same side of the sample. The main reason for this is that often one has patch-clamping apparatus, or flow-cell tubes, etc., attached to the sample from above, and so there is no room for a lens on top of the sample. Another reason for this geometry is that the filters are not perfect, and some light leaks through in wavelength bands where it is not supposed to. In the transmitted light geometry we drew earlier (page 9-7), 100% of the excitation light hits the (excitation-blocking) emission filter. If we instead illuminated from the opposite side, then the excitation would be going *away* from the detector, and all the emission filter would have to block is what little light was reflected off the coverslip and sample – usually ~ 1%.

Illuminating from the same side one collects the light from is called *Epi-illumination*, and the light path looks like this:

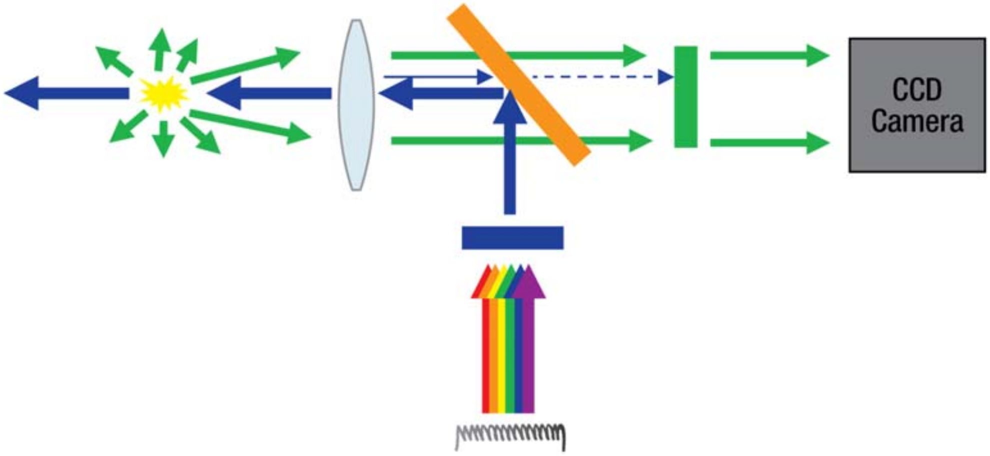


Figure 13: Epi-Fluorescence Geometry

The orange filter in Figure 13 must be another sort of interference filter – it must work at 45° instead of 90° incident light angle, and it must reflect the excitation while transmitting the emitted fluorescence. This filter is called a *dichroic* (*di* meaning two, and *chroic* meaning color) filter. In that sense the other filters are dichroics too, but nobody uses the term that way – usually only the 45° mirror shown above is called a dichroic.

Although it would seem you could get away with only one filter in this manner, the filters are always leaky enough that you actually use three – the excitation and emission filters above, and the dichroic.

Question: What transmission spectrum would you choose for a dichroic to do epi-fluorescence with EGFP and the filters chosen above? Try sketching it before reading further.

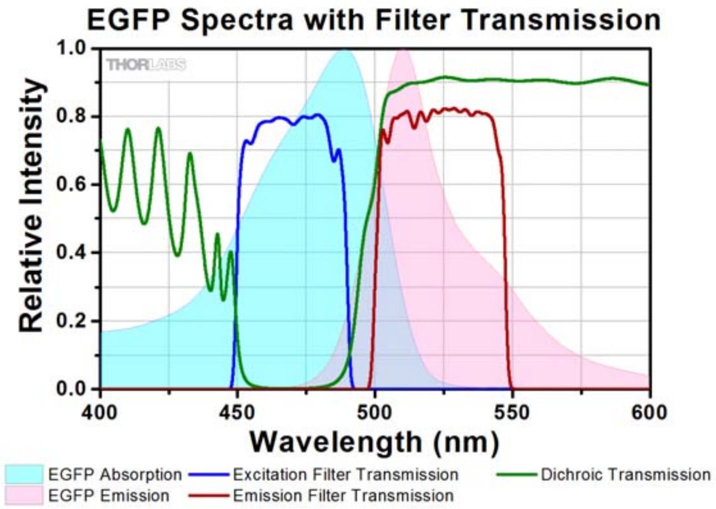


Figure 14: EGFP spectra with dichroic transmission spectrum added.

If you drew something vaguely like what is above, you did very well. If not, go back and think about why this is the right curve. (Note: the oscillations in the transmission curve below 450 nm are due to limitations in the filter design; ignore them. The filter could be reflecting (0% Transmission) below 450 nm and still work fine, so if you drew your filter that way, do not worry.) Note that the emission filter reflection band (reflection = 0% transmission) nicely overlaps the excitation filter transmission band, which is of course no accident.

We will cover the details of optical spectra further in lecture; different lamps and illumination sources have different spectral properties that are very important in terms of their use in fluorescence microscopy. Next lab we will also go into substantial detail on how to choose filters – one can often, with a little effort, do better than the standard recommended filters by a factor of two or more, if one knows how to determine which filters are best for a specific application.

Although we will not do this in class, one can use fluorescence imaging of tiny (~100 nm diameter, << λ) beads to see what the resolution of the microscope is, and quantify it. This method of looking at things (called the PSF, for Point Spread Function) is directly related to the MTF we have just covered and links us back to Abbe theory in a particularly elegant way.

The Optical Spectrum

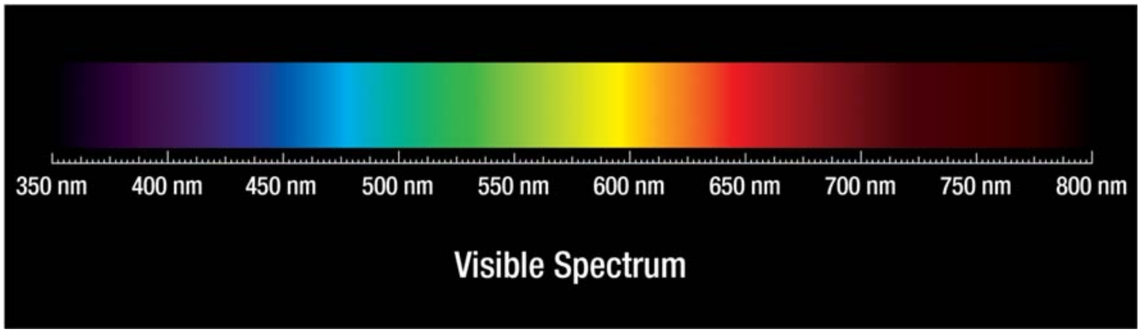


Figure 15. The visible spectrum.

Memorize These:

Color	Wavelength Interval	Notes
Red	~ 700 - 635 nm	Red goes from 635 nm out as far as you can see.
Orange	~ 635 - 590 nm	
Yellow	~ 590 - 560 nm	Yellow is VERY narrow compared to other colors.
Green	~ 560 - 490 nm	Human eye's peak bright light sensitivity is ~550 nm.
Blue	~ 490 - 450 nm	
Violet	~ 450 - 400 nm	Violet goes from as low as you can see up to 450 nm.

THORLABS

Lab 10 Notes: Spectra and Filters

Optical Microscopy Course

Lab 10 Course Notes:
Spectra and Filters

Overview

Most people never learn to choose filters, but rely on the filter companies to tell them what to get. Unfortunately, this often results in fewer good choices than they could have had, especially if they are using an odd fluorophore, or setting up their system in an unusual way. Furthermore, few people know how to estimate how many photons they should expect in a measurement (called a "photon budget"), and so have no way of deciding whether a given experiment or instrument set-up is workable.

Since those two issues are linked, we will cover them both, starting with the "photon budget" and then moving to filter selection.

Do not worry about understanding every last bit of these notes! **The most important task this lab is to understand filter selection so you can work the problem set (and thus be able to choose your own filters)**. If you get bogged down in any of the math below (especially under the "excitation" or "*etendue*" sections), then skim to get the main points, and be sure to pick up carefully again at "Choosing Filters." These notes are intended to be a reference for you in the future, as well as help you do the problem set. You do *not* need to memorize everything.

You do not have to turn in answers to any of the "Questions" unless told to do so by your instructor.

References

All links can be found in the *Reference Links* tab at www.thorlabs.com/OMC.

Handbooks:

- Handbook of Optical Filters for Fluorescence Microscopy
- The Molecular Probes Handbook

Spectra Viewer Tools:

- Fluorescence Spectra Viewer I and II

Sources of Complete Dye and Light Source Spectra:

- University of Arizona Spectra Database
- PhotochemCAD Chemical Spectra
- Fluorescent Proteins (Excel File)

Info on *Etendue* and Light Sources:

- Light Collection and Systems Throughput
- Olympus Noncoherent Sources
- Spectral Irradiance and Lamps
- Microscope Light Sources
- Electronic Imaging Detectors

Data Extraction from Plots

- Data Thief

The Photon Budget

Most biological fluorescence imaging is conducted in cases where there is not as much light as one would like. Understanding why this is, and what steps can be taken to mitigate the problem, requires that we understand what we have to do to get the light in the first place, and subsequently to detect it. This brings us to the topic of the "photon budget."

From the Lab 9 Course Notes, we know that when a fluorescent molecule absorbs a photon it is likely to later emit one at a longer wavelength: one photon in, results in (no more than) one photon out. The "no more than" is due to the possibility that the molecule de-excites in some non-radiative way, or does so on such a long time scale, and/or at such a shifted wavelength (e.g. via phosphorescence) that one never sees the emission. That possibility is captured in the concept of the "Quantum Yield" (QY), which is just the ratio of photons you get out to the number of excitations the molecule undergoes. This can be as high as 0.8 for FITC (fluorescein isothiocyanate, a very popular fluorophore) or 0.92 (for Alexa 488, a proprietary fluorophore from Invitrogen often substituted for FITC), but can be as low as 0.2 (for Acridine Orange) or even lower for "worse" fluorophores. Worse because for every 100 ***absorbed*** excitation photons, one gets 80 emitted fluorescent photons out for FITC, but only 20 for Acridine Orange.

Since molecular excitations and emissions are linked in this simple way by the QY, we can separate them for the purposes of analysis. We will start by looking at the emission side; light collection and detection issues are extremely general – the same analysis also holds for many processes other than fluorescence, including chemiluminescence (which uses chemical reactions, rather than light, for excitation), photons emitted due to radiative decay of radioisotopes, and others.

Let's look at the collection side of our optics:

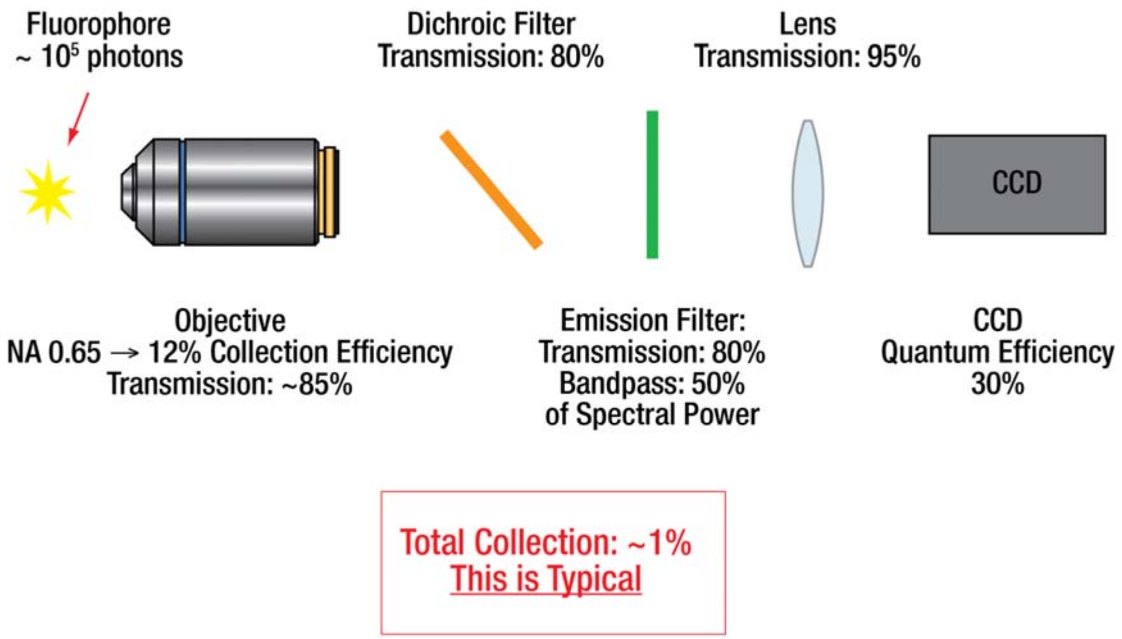


Figure 1: Typical collection and detection efficiency.

There is nothing complicated about this math – we just multiply the efficiencies with which we collect the photons, or with which they make it through a given filter, or with which the detector catches them. In slightly more detail:

1. We collect photons from the sample with efficiency CE = $\sin^2(\theta/2)$, ~ $NA^2/4$. (See Appendix A, *Collection Efficiency* for more details.)

2. The objective has a lot of lenses inside it, and even with antireflection coatings you lose ~1% of light at each surface. 85% transmission is not unusual.

3. The filters typically transmit ~80% in their "pass band," and 0% in their "stop band." (Note: you are not using a dichroic on your rail right now, but one is typically used in epifluorescence so we have included it here).

4. The filter transmission band does not cover all the wavelengths at which the fluorescent molecule may emit – in fact, it may only overlap wavelengths covering half (or less) of the total emitted light (examples in Figure 2).

5. The camera is not equally effective at capturing photons at all wavelengths; for cheap (< $5k) cameras, the fraction of photons detected (called the "Quantum Efficiency" [QE]) tends to peak around 50%, and it often has an average around 30% through the wavelength band of interest (example in Figure 3).

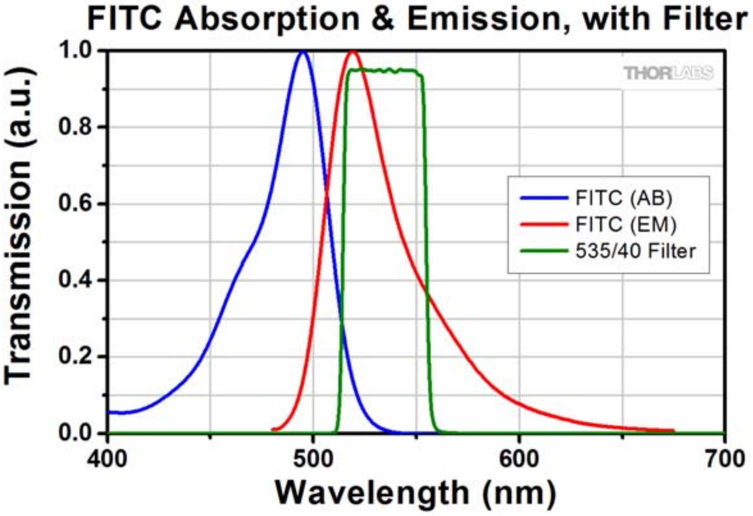


Figure 2: FITC absorption (blue) and emission (red), with a typical 535/40 emission filter transmission band overlaid. Note that 1) the max filter transmission is ~95% and 2) only ~50% of the total area under the emission spectrum falls in the passband, so only approximately 95% * 50% = 48% of the emitted light will make it through this filter.[1] For an extra $50 one can buy a filter that has ~98% transmission, in many cases a very good investment. A slightly different choice of filter transmission band would allow even more light through.

[1] The exact emission fraction under the filter bandpass will depend a bit on excitation λ, since all emitted photons must have a longer wavelength than the excitation. This changes the normalization a little – imagine if one excited at 550 nm in the above graph; any emission would have to be >550 nm, so the fraction of total emission >550 nm would have to be = 1. Practically, absorption and emission bands tend not to overlap too badly, so people disregard this issue for simplicity, even though that results in some minor error.

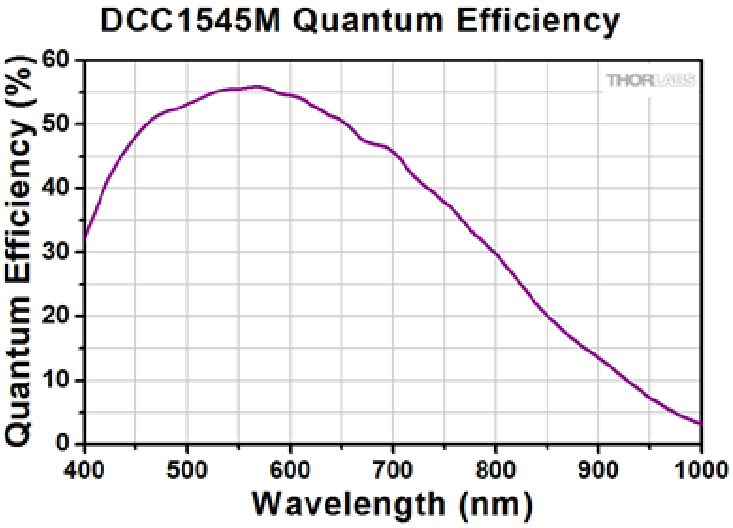


Figure 3: QE for a Thorlabs DCC1545M camera (i.e., the fraction of incident photons at any given λ which are detected by the camera). For an inexpensive camera, this one is pretty good in the green. QE falls off toward zero as λ increases toward 1.1 µm, since the energy of a photon at that wavelength is below the 1.12 eV band-gap energy of the doped silicon. When the photon cannot excite an electron into an accessible state, it cannot be absorbed (and hence detected). For the same reason, silicon is transparent at λ > 1.1 µm.

Let's say we are using the dye FITC, with QY = 0.8, and the numbers given above as a first approximation. Multiplying it all out, we collect only ~1% of the emitted photons. **That means that for every 100 times a FITC molecule gets excited, we will detect 0.8 photons! Not very many.**

It gets worse: few of us can afford a camera that has a noise floor so low one can even hope to distinguish a single photon (those cameras are $35k - $50k, and pretty large), and remember that we are not dumping all our photons on a single pixel, since we need to maintain our resolution by sampling at above the Nyquist frequency for any spatial frequencies in the sample, so our pixels have to be spaced by at least λ/4·NA (see Appendix A, *Nyquist Pixel Spacing* for more details). Let's discuss the noise floor issue first.

Your camera (Thorlabs Item # DCC1545M) probably has a noise floor for a ½-second exposure of ~8 photoelectrons, and probably requires ~ 7500 photoelectrons to read a full 255 counts (i.e., saturate the pixel output). This implies we would need > 8000 excitations of the FITC molecule just to have a reasonable (signal/noise = 1) chance of telling whether it emitted light, and ~ 1 million (i.e., 10^6) excitations to get a fully bright pixel.

This analysis only holds if all the light falls on one pixel, which it should not. We know we need to have our pixels spaced closely enough together that we have at least two pixels per cycle of the smallest spatial frequency we have in the image. Since the spacing of the smallest spatial frequency in the *image* from an objective with numerical aperture NA is M·λ/2·NA (i.e., the same as λ/2·NA at the sample)[2], that means we need pixels spaced no farther than M·λ/4·NA, which is the Nyquist pixel spacing listed on the

[2] This assumes incoherent illumination. It turns out that fluorescence emission is in fact incoherent; usually even more incoherent than "incoherent" illumination, so this is the correct formula to use. It is worth remembering that even if you excite the fluorescence with a *coherent* source (like a laser), the *emission* from the fluorescent molecules is still incoherent, so the incoherent formulae continue to apply even in that case. Not obvious, but true, is that for fluorescence (or any "self-luminous" sample, i.e., a sample producing its own light), you should use $NA_{condenser} = NA_{objective}$ in the formulas.

“Equations to Memorize” sheet. It turns out that this corresponds to having ~5 pixels across the diameter of the Airy disk (i.e., across the diffraction-spread-out image of a point source).

Question 1: Convince yourself that the required pixel spacing is ~ 5 pixels across the image of a point source.

Consequently, the light will fall on approximately 25 (5 x 5) pixels, with the exact distribution depending on where the center of the Airy disk is with respect to the pixel pattern:

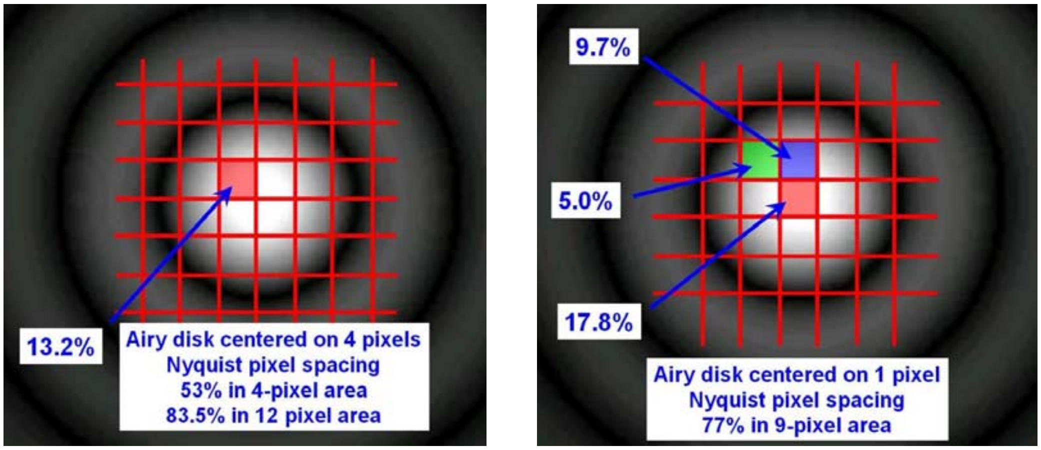


Figure 4: Light fraction falling on different pixels as a function of whether the Airy disk is centered at the vertex between 4 pixels (left image) or at the center of one pixel (right image). Roughly speaking, 80% of the total light falls in a 9 - 12 pixel region depending on the registration between the Airy disk and the pixel pattern, so a decent estimate is that the brighter pixels will get ~ 10% of the signal.

The result of all this is that we need ~10 million excitations of the dye molecule to get a single decent image on our camera and that is for a really good dye molecule! This is not idle calculation – the super-resolution microscopy technique PALM (Nobel Prize, 2014) involves calculations almost exactly like this. While we are at it, remember that the average dye molecule will bleach (photodestruct) after only ~1 million excitations, so it would take 10 dye molecules just to get us this single image.

Let’s pause and recap what we have so far:

- It takes ~10 million dye excitations to get one fully-exposed Airy disk image, using reasonable equipment.

- That is a lot. One implication is that much imaging is done with fewer collected photons, which means that the S/N (signal to noise) ratio is lower – remember, using our numbers here, the 255 counts at the middle of the image is only 32 times greater than the 8-photoelectron noise level. Already, that level is so high we will barely see the 5%-level pixels in Figure 4, so our Airy disk images will be blobs, not nice ringed shapes. If we have fewer excitations, we will have even lower S/N, and our image will be that much worse (below S/N ~ 10, you start to really notice). This is one reason people spend $35k+ on very low-noise cameras – if you cannot have more signal photons, then you have to suppress the noise level if you want better S/N.

- If we do not have many dye molecules, bleaching will definitely be an issue – in fact, for one dye molecule we can only get *one* really good image (using expensive equipment and allowing for bad S/N, one can push this to 30 or so).

- More efficiency would sure help. How could we do this?
    - Better camera with 90% QE and noise level 10X lower: ~3X improvement. Cost: 100X as much ($35k instead of $350).
    - Higher NA objective: 0.95 NA air (or 1.4 NA oil) = CE of 34%: ~3X improvement. Cost: $3k or so, depending.
    - Better filters: 95% transmission and 75% (instead of 50%) of spectral energy: ~2x improvement. Cost: ~ $300-$1k.
        - **Notice: filters are *by far* the cheapest way to improve performance!**

Filter Leakage

Light throughput is only half the story for filters; their main purpose is to block the excitation light from getting to the detector. It is worth doing a quick, back-of-the-envelope calculation to get a sense of how good we need these filters to be at blocking excitation light, since the more we can get away with, the more emitted light we can get to our detector and the less we will probably need to pay for fancy filters.

Ideally, one wants uniform illumination at the sample (we are ignoring laser scanning techniques), so that the image is evenly lit across the whole field of view. The Nyquist pixel sampling distance (at the sample) is λ/4·NA, so the amount of *excitation* hitting a given pixel – i.e., the background light in our image – is given by:

Equation 1: Illumination

$$\frac{\boldsymbol{photons}}{\boldsymbol{sec}}\,\boldsymbol{on\ a\ pixel} = \frac{\boldsymbol{photons/sec}}{\boldsymbol{m^2}}\ \boldsymbol{at\ the\ sample} * \left(\frac{\boldsymbol{\lambda}}{\boldsymbol{4\ NA}}\right)^{\boldsymbol{2}} * \boldsymbol{filter\ blocking\ ratio}$$

From our earlier analysis, we know that the light we collect from the fluorescent molecules will be something like:

Equation 2: Signal

$$\frac{\boldsymbol{photons}}{\boldsymbol{sec}}\,\boldsymbol{on\ a\ pixel} \sim \boldsymbol{\#\ dye\ molecules\ at\ the\ sample} * \frac{\frac{\boldsymbol{excitations}}{\boldsymbol{sec}}}{\boldsymbol{dye\ molecule}} * \boldsymbol{CE} * \boldsymbol{1\%}$$

Where we have pulled the collection efficiency (CE) of the objective out and made it explicit.[3]

Conveniently, the number of fluorescence excitations is proportional to the illumination intensity at the sample. The easiest way to think about this is to imagine the dye molecules have a small cross-sectional area, and if a photon hits that area it gets absorbed and the molecule is excited. This is not quite accurate – the photon is an electromagnetic wave which interacts with the electron wave function of the dye molecule, which is spread out over space, but one can mathematically reduce this complexity to an equivalent cross-sectional area, denoted σ. What is σ? Just the area of the molecule photons can "hit" – if you think of photons as like raindrops, this is equivalent to the area of your head or a roof; the roof has a larger area σ, and more rain hits the roof than your head, which is smaller (smaller σ). The "cross section" σ has units of area, and represents the area over which photons will be absorbed, so if we know how

[3] To avoid confusion, note that the 1% number seems unchanged because we took out the CE (12% in the original calculation) and inserted another (numerically similar) factor of 10% since we have only got ~ 10% of the light on a given pixel if we sample the Airy disk at Nyquist. As a result, the 1% number did not change.

many photons we have per unit area, per unit time, then we know how many excitations our dye molecule will have:

Equation 3:

$$\frac{\boldsymbol{excitations/sec}}{\boldsymbol{dye\ molecule}} = \boldsymbol{\sigma} * \frac{\boldsymbol{illumination\ photons/sec}}{\boldsymbol{m^2}}$$

Inserting Eq. 3 into Eq. 2 gives:

Equation 4:

$$\frac{\boldsymbol{signal\ photons}}{\boldsymbol{sec}}\boldsymbol{on\ a\ pixel} \sim \boldsymbol{\#\ dye\ molecules\ at\ the\ sample} * \boldsymbol{\sigma} * \frac{\boldsymbol{illumination\ photons/sec}}{\boldsymbol{m^2}} * (\boldsymbol{1\%} * \boldsymbol{CE})$$

Dividing this by Eq. 1 results in what is called the "signal-to-background ratio," S/B, which gives an idea of how well we will be able to see the sample:

Equation 5:

$$\frac{\boldsymbol{signal\ photons}}{\boldsymbol{illumination\ photons}}\boldsymbol{on\ a\ pixel} \sim \frac{[\boldsymbol{\#dye\ molecules\ at\ the\ sample} * \boldsymbol{\sigma} * \boldsymbol{1\%} * \boldsymbol{CE}]}{[(\frac{\boldsymbol{\lambda}}{\boldsymbol{4\ NA}})^2 * \boldsymbol{filter\ blocking\ ratio}]}$$

We would like to keep the S/B high, say ~ 10, so the background shot noise is low (shot noise = statistical photon noise = √(# photons), so even if we subtract off the *average* background we would be left with a lot of extra noise in our image). We will also substitute in CE ~ $NA^2/4$, and rearrange Eq. 5 to get:

Equation 6:

$$\boldsymbol{Required\ filter\ blocking\ ratio} \sim \frac{\boldsymbol{4\%} * \boldsymbol{\#\ dye\ molecules} * \frac{\boldsymbol{\sigma}}{\boldsymbol{\lambda^2}} * \boldsymbol{NA^4}}{\boldsymbol{desired\ S/B}}$$

There is no need to memorize that, but it is in red because it is quite important. Before discussing it, let's see what the ratio is for some reasonable numbers, say NA = 0.65, λ = 0.5 μm, S/B ~ 10, 1 dye molecule, and σ ~ 2 $Å^2$. Plugging these values in, we find **the required blocking ratio is ~ 5.7 x 10^{-11}, or about OD 10.**

That is really quite high – usually a good filter only gives ~ OD 6. This is the reason people use epi-illumination for serious fluorescence work: the excitation light reflects off the dichroic and goes through the sample *away from* the detector:

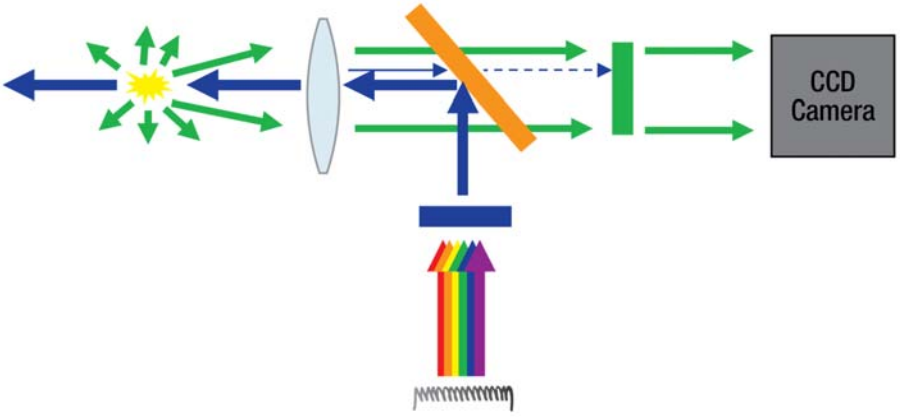


Figure 5: Epi-illumination geometry (also called epifluorescence). Note that the only excitation going toward the camera is what had reflected off the sample (~ 1%, equivalent to OD 2), and this light must also then go through the dichroic (which has ~ OD 1+ at the excitation wavelengths), for an effective additional blocking of the excitation by OD 3.

The result of that is that only the *reflected* illumination must be rejected and for an oil-immersion objective (typical for fluorescence imaging, since one usually wants to use high NA) the main reflection is only from the glass-water interface, where R ~ $[(1.52 – 1.33)/(1.51+1.33)]^2$ ~ 0.5%. Practically, one also gets some from imperfect antireflection coatings in the objective, but the result is still a few percent, a huge benefit compared to using transmitted light. In addition, the dichroic mirror, which initially reflected the excitation toward the sample, also serves to deflect any returning illumination back toward the lamp. Filters work much less effectively at 45°, so the effective blocking here is only ~ OD 1, but that still results in a combined decrease of OD 3 in illumination incident on the final emission filter. If that filter has OD 6, then the result is OD 9, very close to what we calculated we need for single-fluorophore imaging, previously. In practice, increasing NA, getting better filters, and using a better camera do allow you to get away with a lower blocking ratio, so it is only rarely that people really need to double-up on the emission filters to get the necessary blocking.

While remembering that making the blocking small (i.e. high OD; see Equation 6) is hard to do, we nonetheless want the S/B ratio as large as possible because it will make the image better. Looking at Equation 6, notice that:

- The 4% represents the emission light that got through the filters, lenses, and was detected. Improving this (better filter transmission, higher camera QE, etc.) helps in a linear way.

- More dye molecules in a given sample area also helps in a linear way, which makes sense – essentially more sample!

- A larger absorption cross-section σ helps, since there are more excitations for a given illumination level. The important ratio is σ / λ^2. The smaller the λ, the smaller the Airy disk the signal's concentrated in, and the larger σ the more excitations (and thus emissions) one gets.

- There is extremely strong scaling with NA! Higher NA increases the signal capture, and also reduces the size of the Airy disk. **Required filter blocking is vastly reduced at higher NA!**

- The lower we want our background, compared to the signal, the better filter blocking we need.

- Illumination intensity does not actually show up anywhere in the formula!

Question 2: In the case of the class set-ups, we have OD 6 blocking since we are not doing epi-illumination, and we have only 0.4 NA objectives, how many dye molecules would we need in one tiny sub-resolution spot (so that all the light contributes to the same Airy disk in the image) if we want a S/B ~ 10?

Question 3: Typically, one uses dye-impregnated polystyrene microspheres to test fluorescence set-ups. A good guess for the typical dye concentration in the plastic is ~100 µM. What radius bead would have the required number of dye molecules to satisfy your result from Question #2?

Excitation

So far we have only dealt with the question of how much the filters must block, not what they must let through. Before we can get to that, however, we need to get a sense for the illumination / excitation photon budget. There are two main parts to this: the dye and the source. The dye is simplest, so we will start there.

We mentioned above that the dye's absorption can be characterized in terms of a cross-section σ. Usually it is hard to find this number (actually, it is often hard to find many of the numbers that matter for these calculations). What is usually cited is the "extinction coefficient," ε, which has units of Mol^{-1} cm^{-1}. When light traverses an absorbing medium its intensity decays exponentially. This should make sense – if it decays 10% in the first millimeter it travels, there is only 90% left, so in the second millimeter of travel, 10% of 90% = 9% more (of the original) will be lost. Notice that the change (derivative) is proportional to the intensity at any point, which is characteristic of an exponential. This is known as the Beer-Lambert law, and is written:

Equation 7: $I = I_0 \, 10^{-\varepsilon\, x\, c}$ **, where ε = extinction coeff., x is the distance, and c is the concentration**

It is obviously not that tricky to measure ε – just shine some light and measure what you get with and without a known thickness of material in the path. Many labs have specialized equipment just for this purpose, and so those values (ε) often get stated in papers.

Conveniently, ε can be related to the cross section σ (which is more useful for our calculations – we want to know what area can absorb photons, not how far they can travel in some material) by some simple manipulations: if a light beam of area A is propagating through a medium with a concentration *n* of particles per unit volume (*n* is just = c·N$_a$, where N$_a$ = Avogadro's number and c is the Molar concentration), each of cross-sectional area σ, then the fractional area blocked in a distance dx is given by (number of particles) * (area per particle) / (total area), which is to say [(*n*·A dx) · σ] / A. The A cancels, and substituting *n* = c·N$_a$ gives the fractional area as c·N$_a$·σ dx. Since the loss of intensity across dx is given by the (intensity) * (fraction of clear area), we can write dI = – I [n·σ dx], with the solution:

Equation 8: $I = I_0 \, e^{-N_a\, \sigma\, x\, c}$ **, where N$_a$ = Avogadro's #, σ = cross section, and x, c as above.**

Setting Equations 7 and 8 equal to each other yields:

Equation 9: **σ = 3.82 x 10^{-25} ε, with σ in m^2 and ε in M^{-1} cm^{-1}; this comes from σ = ε / [10 N$_a$ log$_{10}$(e)], with the factor of 10 converting from ε in M^{-1} cm^{-1} to σ in m^2.**

Typical ***peak*** values of ε are 10,000 - 100,000 M^{-1} cm^{-1}, so typical values of σ are ~ (0.5 Å)^2 to (2 Å)^2, not far from what one might expect for things of molecular dimensions. That word "peak" is crucial – ε and σ are wavelength dependent; if the optical frequency (∝ photon energy) is not on resonance with a molecular orbital transition, then absorption is unlikely and σ is commensurately smaller. The dye absorption spectra are just tabulations of ε, and so are proportional to σ; one can just read the relative values right off the graph (e.g. Figure 2). Often the maximum value is given, and the graph is simply scaled so the maximum is 1.0 and then (in a pathological practice) the units are left off the y-axis. If you encounter this, keep looking to find the peak value for ε, and convert using Eq. 9.

A lot of math there, and for what? Quite simple: with σ, we can rapidly find the amount of light we are going to get from a sample using our illumination intensity. As an example, consider a 1 mW, 532 nm laser with a 1 mm² beam cross section. What is the illumination intensity? Obviously 1 mW / mm², but one must *BEWARE* here: one photon absorbed leads to one photon emitted (ignoring the issue of quantum yield) but 1 mW *absorbed IS NOT* 1 mW *emitted*, because the energy of photons changes with wavelength. This should also make sense from the perspective that some energy is lost to thermalization in the molecule before the fluorescence is emitted so the **emission has less energy than was absorbed, but the same number of photons.**

You really need to be careful here, in part because there is terrible, pernicious confusion in the way much of the data you will use for these sorts of calculations is presented. For instance, illumination data is almost always in intensity (e.g. W/m²), rather than in photons per unit area and, even worse, even the intensity data is often given normalized for the sensitivity of the human eye (units of lumens or candela), and it can be hard to undo this to get the right information. An example: a 1 mW, 500 nm laser has 2/3 as many photons/sec as a 1 mW, 750 nm laser, since photon energy scales as 1 / λ:

Equation 10: $\boldsymbol{Photon\ Energy = h\,v = \frac{hc}{\lambda}}$**, where h = 6.626 x 10^{-34} kg m²/s, Planck's constant.**

However, the human eye is more sensitive at 500 nm than 750 nm, so the 500 nm laser might have 10x as high a lumen rating, so be very careful about the units on things. **We usually convert everything to photons,** for a few reasons: dye absorption is proportional to σ, camera QE is a relative photon detection sensitivity, and dye emission is (usually, when they bother to label the axis at all) reported in "counts," which is again usually a photon-based measure. Here "photon-based" means that one detected photon is proportional to one count – this may seem obvious, but it is not necessarily: if a photon excites an electron in a semiconductor, then one will measure an electron – it does not matter whether the electron was highly excited by a blue photon or barely excited by a red one – you still just measure one electron (at least for visible light; x-rays can excite multiple electrons). Compare that to a thermal detector that absorbs light and reports how much it is heated up; in that case a blue photon will look twice as "hot" as a red one at twice the wavelength.

Back to our example, the 1 mW, 1 mm², 532 nm laser. (Lasers are thankfully always reported in mW, never in lumens, etc.) How many photons/sec/mm² is this? From Eq. 10, we have
photon/sec = 1 mW / E = 1 mW * λ / (h c) = 2.7e15 photons/sec/mm². That seems like quite a lot, right? And it is – a 1 mW green laser pointer looks *very* bright if you look directly at it (do ***not*** try that!).

But how many excitations of a single FITC dye molecule would that be? (Note: actually, 532 nm would excite Rhodamine 6G better, but aside from that R6G has similar properties to FITC in terms of σ and QY.) We can write this as

Equation 11:

$$\frac{\boldsymbol{excitations}}{\boldsymbol{sec}} = \frac{\boldsymbol{illumination\ power/unit\ area}}{\frac{\boldsymbol{hc}}{\boldsymbol{\lambda}}} * \boldsymbol{\sigma} = \frac{\boldsymbol{\#\ photon}}{\mathbf{sec} * \boldsymbol{area}} * \boldsymbol{\sigma}$$

From previously, we have σ ~ (2 Å)^2, so the number of excitations of our dye molecule is ~100 / sec, not very many. Remember from above that we need ~10 million excitations to image the sample well just once. Ignoring bleaching, with a 1 mW laser source getting that many excitations would take 10^5 sec, or about 28 hours. Since most cameras have dark current (due to thermal agitation exciting electrons into the conduction band), the pixels would fill up with dark signal long before that unless the camera was cooled with liquid nitrogen (or liquid helium; astronomers do this regularly).

So what can we do? We could turn up the power, except that the very brightest arc lamps will only give us ~20 mW/mm^2. Lasers are pretty expensive ($10k+) and only give you one wavelength, plus if you turn them up too far you will cook your sample. A halogen lamp is very poor by comparison: using a good collector lens it would give us ~0.1 mW/mm^2; ours probably provides ~0.02 mW/mm^2.

Let's actually use that example: we have halogen lamps, and when you looked at the lens paper with the pink marker on it (pink is sort of like Rhodamine 6G) you used ~0.5 sec exposure. Using the numbers from above, we need ~10 million photons to image well, and all that in 0.5 sec, so 20 million photons/sec. Since the lamp is ~ 2% as powerful as our laser example (if that), we expect ~2 excitations/sec/molecule; thus we need ~2.5 million molecules to image well. We also need them to be in an area small enough that the light is not getting spread out much past an Airy disk diameter, so let's say for simplicity they need to be in an area the diameter of the Airy disk itself – a circle of radius 0.61 λ / NA, ~ 2 µm^2. So we probably needed ~ 1 x 10^{18} molecules/m^2 to see the lens paper without using a lot of gain.

How reasonable is that estimate? We used ~2 µl of ink on each slide, covering roughly a 0.5 cm x 1 cm stripe; furthermore, ~0.5 mM ink concentration is probably not a bad guess, which results in ~ 1 x 10^{19} molecules/m^2. Given that we used a lower NA objective than in my example above, and that the dye is probably not as good as FITC (QY maybe only ~0.3, ε maybe ~ 3 x 10^5 M^{-1} cm^{-1}), these numbers are reasonably consistent. Photon budgets work!

Going the other direction, let's consider trying to image labeled actin moving back from an advancing lamellipodium as a cell crawls; see fluorescence image Fig. 4 in Course Notes 1, and the (phase contrast) video *Crawling Keratocyte* under the *Reference Links* tab at www.thorlabs.com/OMC. We have a new microscope capable of collecting and detecting about 25% of the emitted light from a sample, and we can illuminate that sample with at least 20 mW/mm^2 of laser light before sample damage begins; however, such a movie requires ~33 ms exposure times. How will that work out? That illumination intensity will gives us ~2000 excitations / sec, or 60 excitations per 33 ms of which we will *detect* 14 photoelectrons. These are spread out onto several pixels, so maybe ~ 1 - 2 photoelectrons/pixel. We have an amazing camera, so that might work out to a S/N ~ 0.5, but it is certainly not enough to do real-time tracking unless we turn up the laser (which we could easily do), undersample the image so all the light gets dumped into one pixel (this can be done by electronically "binning" the pixels before readout), or multiply label the actin so that there is more than one fluor in a subresolution area, or some combination of those things. This should give you an idea of why that movie was so impressive, why the noise levels looked high in it, and what sort of research tasks merit the expenditure of so much money on cameras, etc. In the case of rapid dynamics with very few fluors, the details really matter and figuring out the details starts with doing photon calculations ("budgets") like the calculations previously.

Practically speaking, the difference between those two cases consists of 25% vs. ~ 1% collection/detection efficiency, ~ 1,000 times more intense illumination, and a better S/N for the case where we imaged the lens paper. The place it is most easy to improve those numbers is the filters, which we will come to shortly. However, it is worth a final word on illumination before moving on.

Etendue

Note: *Etendue* can be rather confusing; while it is important enough to merit a good bit of discussion here, the main take-homes from this section are the limit on usable light available from incoherent sources and the fact that it is brightness that matters, not total power.

If this section is confusing to you, please skip down to "Choosing Filters," and use your time to better understand the other parts of these notes.

It may seem counterintuitive that one cannot just buy a bigger arc lamp and focus the light down more tightly to get more intense illumination. After all, if you want a more powerful searchlight, you just get a bigger arc source; so why not for microscopy?

The answer to that question lies in something called "*Etendue*," which essentially means "extent." The phenomenon is also known as the Lagrange or optical invariant, and is a very powerful concept. In short, the optical invariant states that the product of the area of an image and the angular distribution of light at the image remain constant:

Equation 12: $$n_1{}^2 A_1 \Omega_1 = n_2{}^2 A_2 \Omega_2$$
where n = refractive index, A = area and Ω = solid angle.

Taking the square root of both sides, gives (in the small angle approximation),

Equation 13: $$R_1 NA_1 = R_2 NA_2, \text{ since } \Omega \sim \frac{NA^2}{4n^2}, \text{ and } A \sim R^2.$$

See Appendix A, *Optical Invariant* for more details.

An example can be useful. Let's say we have a 1 kW halogen lamp, with a filament temperature of 2800 K and a filament size of 10 mm^2. We also have a 100 W halogen lamp with a 2800 K, 1 mm^2 filament. What is the benefit of the bigger lamp?

The amount of light we collect from any part of the filament is (naturally) related to the collector NA, and scales like NA^2/4. Higher NA is obviously better, since we collect more light from each point on the filament.

However, the magnification, M, is given by M = NA$_{image}$ / NA$_{object}$ and NA can never be bigger than 1 (or, for an immersion objective, bigger than the index n). Let's say I use the highest NA objective I have, 1.4 NA oil immersion. As we increase the collector NA, I collect more light proportional to the $NA^2_{collector}$, but at the same time, the magnification M is increasing like $1/NA^2_{collector}$. These two terms cancel – as the collector NA goes up, we ***do*** collect more light from the filament, but we spread it over more area at the sample and so we gain nothing. For this reason, it will not matter whether we use the 100 W or the 1 kW lamp; since both filaments are the same temperature, they are just as bright in terms of emitted power / unit area, and the extra area cannot be focused down onto the sample without reducing the collection efficiency of the collector lens by a roughly equal amount. This is a manifestation of an "*etendue* limit;" it does not happen for the searchlight because the arc is not being focused down at all, so the NA is low on the image side and there is plenty of flexibility to adjust it to focus a bigger source over the huge area desired. This stops being useful once one wants to focus the source down to a spot of the same size as the source or smaller.

As a result of the *etendue* limit, what matters most in a source is how "bright" it is. Brightness is carefully defined as the emitted power, per unit area, per solid angle. Incoherent sources emit in all directions, so the solid angle is already fixed at 4π (or 2π if one considers only one side of the bulb, and uses no reflector). In that case, **what matters is emitted power *per unit area*, not total power**. The most popular Hg arc lamps are "HBO 100" lamps, which are 100 W bulbs. One can buy more powerful lamps, but the HBO 100s are called "short arc" bulbs and have a very small (0.25 mm x 0.25 mm) emitting area, where all the power is concentrated. It turns out these bulbs emit more power per unit area than any other arc lamps. You can get a bigger one, but it will not help. Similarly, the power emitted from a piece of hot metal is related (via the Stefan-Boltzmann law) to the temperature. Once your filament is bigger than the largest sample you are going to examine, it will not help to make it any bigger; however, making it *hotter* will increase the power per unit area and help a lot. Unfortunately, making it hotter also dramatically decreases the life of the bulb (a 2800 K filament might last 1500 hours, while a 3100 K filament only lasts around 250 hours).

A last comment on *etendue*: one can use it to quickly figure out how much light one can use from a source. As an example, consider a 1.4 x 1.4 mm (i.e., 2 mm^2), high power, 350 mW LED that emits into 180° (= 2π sr; solid angle is measured in "steradians," abbreviated sr; there are 4π sr in a full sphere). If we want to illuminate a 1 mm diameter field of view ($A = \pi / 4$ mm^2) through a 0.4 NA objective, what is the most power we can get to the sample?

1. The *etendue* at the LED is $A*\Omega = 2 \text{ mm}^2 * 2\pi = 4\pi \text{ mm}^2$, and

2. The *etendue* at the sample is $A*\Omega = \pi / 4 \text{ mm}^2 * (4\pi * NA^2 / 4) \approx 0.04\, \pi^2 \text{ mm}^2$

since the solid angle is 4π * the collection efficiency. Note: the LED only emits in one direction; hence the 2π solid angle.

The initial *etendue* is 4π mm^2, but any lens we use (with NA < 1) will reduce this. However, until we reduce the collector lens NA to the point where the *etendue* on that side is <0.16π mm^2, it will not matter: it will still be limited by the *etendue* on the sample side.

Question 4: At what collector NA will it begin to be *etendue* limited on the collector side?

One way of thinking about this is that I can drop the collector NA until I have (de-)magnified the LED by a factor of M = 1 / 1.4 and it is exactly the same size of my sample field of view; then the resulting collector NA determines how much useable power I can collect from the LED (and thus effectively use).

The ratio of the initial and final *etendue*s (4π mm^2 and 0.04 π^2 mm^2) also gives a measure of the system throughput: in this case ~3% (= 0.04 π / 4). Although the LED power was 350 mW to start with, we will get only ~11 mW to the sample. This is very handy, and worth remembering:

Equation 14: $$\boldsymbol{System\ Throughput = \frac{n_2{}^2 A_2 \Omega_2}{n_1{}^2 A_1 \Omega_1}}$$

You can convince yourself this is true by calculating the collection efficiency and the magnification for some 1-lens geometries, and see how it comes out in terms of the percentage of original power delivered to the sample. In the case above, since the initial Ω is the entire half-space, and the sample field of view (A) is fixed, the only things we could do to improve the situation are to find an LED with a smaller die size but the same power (i.e., a "brighter" LED), or to use a higher NA objective, which would allow us to focus the LED down further.

As a benchmark for the relative brightness of things, consider this table:

Source	Input Power (W)	Emitter Size (mm)	Approx. Brightness (Radiance), 450 - 490 nm EGFP Band, (mW/mm^2/sr)	Source *Etendue*, n^{2}AΩ, (mm^2 sr)	Max Intensity at Focus, (mW/mm^2)	Max Power at Sample, in EGFP Band (mW)
3100 K Quartz Tungst. Hal.	100	4.2 x 2.3	10.7	30.3	22	6.8
XBO 75 W Xenon Arc	75	0.25 x 0.5	79	0.4	164	50
470 nm Blue LED, 760 mW Output	5	1.0 x 1.0	242	7.0	227	70
488 nm Laser, 30 mW	0.5	1.5 (diameter)	1.3 x 10^8	2.4 x 10^{-7}	7.0 x 10^7	21

Table 1: Relative input powers and maximum obtainable excitation intensities at the sample for various light sources. Not shown is the HBO 100 W mercury arc lamp, which can deliver significantly higher intensities than Xenon when used at wavelengths where it has strong emission lines (the eGFP band used above happens not to be one of those, hence its omission). Xenon arcs are less bright, but have a much flatter spectrum through the visible range. The laser, of course, trumps everything in terms of brightness and maximum intensity at the sample. The intensity listed assumes the laser has been focused to its smallest diameter; expanding the beam would drop the intensity to 68 mW/mm^2 across the sample. It's easy to get a more powerful laser, but lasers are most often used focused down, as part of laser-scanning systems, so the number listed above is more usually relevant. Note: Calculations assume a 40X, 0.85 NA dry Zeiss objective (*etendue* 0.91); 70% transmission through the microscope; the LED has an acrylic (n = 1.49) cap on it; and that sources other than the laser are Lambertian emitters. Also, these numbers are an upper limit – there are many ways to deliver less power to the sample (e.g. using less efficient optics). The laser intensity is based on optimal beam focusing (power / beam area).

Because lasers do not emit their light incoherently in all directions, but are (coherently) directional, they typically have very small *etendue* values ($A\ \Omega = \lambda^2$ for a Gaussian laser beam; the optical invariant for a Gaussian beam is radius * $\theta = \lambda / \pi$). Compare that to the ~ π mm^2 values for the LED (only π, because it is a Lambertian emitter; radiation evenly into the full half-space would of course be 2π). This small *etendue* is related to the ability to focus lasers down to very small spots (at which point they have a large angular spread) or collimate them such that they have a larger area but very small angular spread. As a result of their versatility (embodied in the small *etendue* value, which allows a large range of A vs Ω tradeoffs), their power can be brought to bear on a sample much more effectively than is true for an incoherent source with a larger *etendue* value.[4]

More discussion of all this can be found in the Newport/Oriel Application Notes on *Light Collection and Systems Throughput* and *Spectral Irradiance and Lamps*, as well as the Zeiss article on *Microscope Light Sources*; see the references listed on page 10-1.

Choosing Filters

Since we often cannot get more *usable* power out of the light sources (either due to *etendue* limits or problems with sample damage from the high intensity light), it behooves us to make the best use of what we can get, which brings us back to the subject of filter choice. A good place to begin is in revisiting what we want filters to do; this usually boils down to letting as much effective excitation through as possible, while simultaneously collecting the most emission possible. Since the excitation and emission spectra of the dyes tend to overlap a bit, there are tradeoffs. In addition, depending on the experiment, one may want to emphasize certain things. For instance:

- Short exposure times: If one is observing by eye, or with a camera that does not allow long exposure times, then maximizing total brightness of the image matters most. Brightness will be the product of the total effective excitation and total detected emission, so one may compromise excitation if that results in an even greater benefit in collected emission, or vice-versa.

- Autofluorescence: If one is concerned about "autofluorescence" (i.e. fluorescence from a molecule one is not interested in; hemoglobin is a great offender in this regard) then one may choose to sacrifice other things to exclude it, either by limiting excitation in the band where the autofluorescence is excited, or cutting off the emission bandpass where the autofluorescent emission will be.

- Bleaching: Sometimes the limiting factor is the bleaching of the very few fluorescent molecules one has (especially an issue in single-molecule work, where there may be only 1-5 fluorophores per molecule). In that case, one would maximize collected emission, at the expense of limiting both the bandwidth available for the excitation and where the excitation occurs with respect to the effective excitation curve (which we will discuss shortly).

Spectra

Filters simply modify the spectra that get transmitted and reflected in our microscope, so in order to make intelligent choices regarding them, we need to make sure we are aware of all the spectra in the system. For instance, if there is no light at 400 nm, who cares if we filter at that wavelength?

In addition to the dye absorption and emission spectra, there are two more cases of great importance: the spectrum of our excitation source and the sensitivity of our detector (eye or camera) at different wavelengths. You have already seen an example of the latter – when you inserted a visible-light blocking filter in your microscopes, you could still image in the remaining infrared light, but your eye could not see the IR. Since many commercial filter sets are optimized for people viewing the fluorescence by eye, you can see how it might matter a lot if you are using a (IR-sensitive) camera instead! This is one reason it helps to choose your own filters.

[4] Though it is beyond the scope of this class, if you are noticing that the *etendue* for a laser (which is a source of fully coherent light) is much smaller than that for an incoherent source and are wondering if coherence and *etendue* are related, they are, in a rather profound way involving the van Cittert-Zernike theorem (this is the same Zernike that invented phase contrast).

Detector Spectra

Below are spectra for several common types of detectors, including your eye:

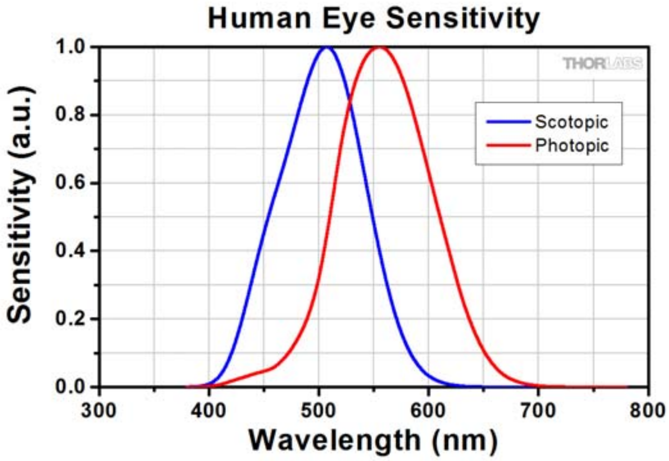


Figure 6: QE of the human eye, (red line) under dim conditions (called Photopic response) and (blue line) brightly lit conditions (called Scotopic response). X-axis is wavelength in nm; y-axis is the QE scaled to 1 at maximum. The difference in efficiencies is due to the increasing reliance on the retinal rod cells in dim conditions vs. the color-sensing cone cells in brighter conditions. Note that all this is unrelated to why red is used at night (and that much of that is based on misinformation – night vision is a separate, complex subject). For more articles on the Sensitivity of the Human Eye, please see the *Reference Links* tab at www.thorlabs.com/OMC.

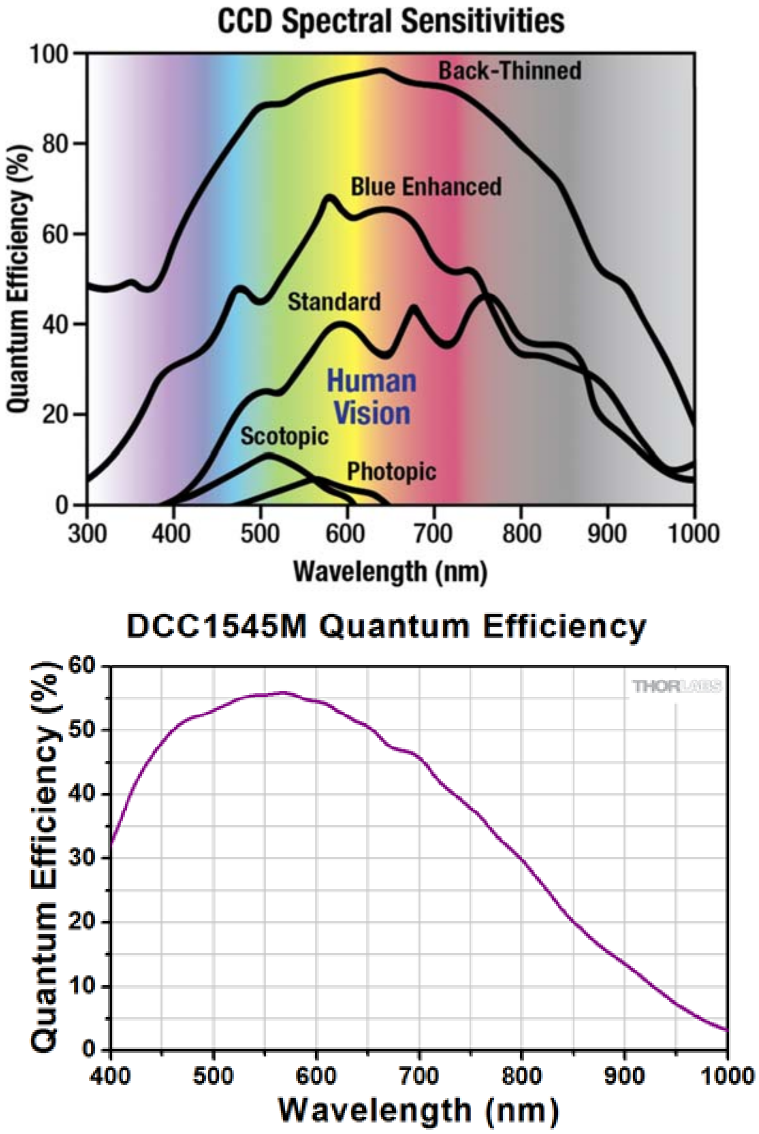


Figure 7: QE of some silicon detectors. (Top): A comparison of several CCD types, along with human eye response. The higher QE curves correspond to more expensive cameras. (Bottom): The CMOS sensor in the Thorlabs DCC1545M, which you are using in your rigs.

Light Source Spectra

This subject is broad, and we will touch on it only briefly. For more information, see *Microscope Light Sources* under the *Reference Links* tab at http://www.thorlabs.com/OMC.

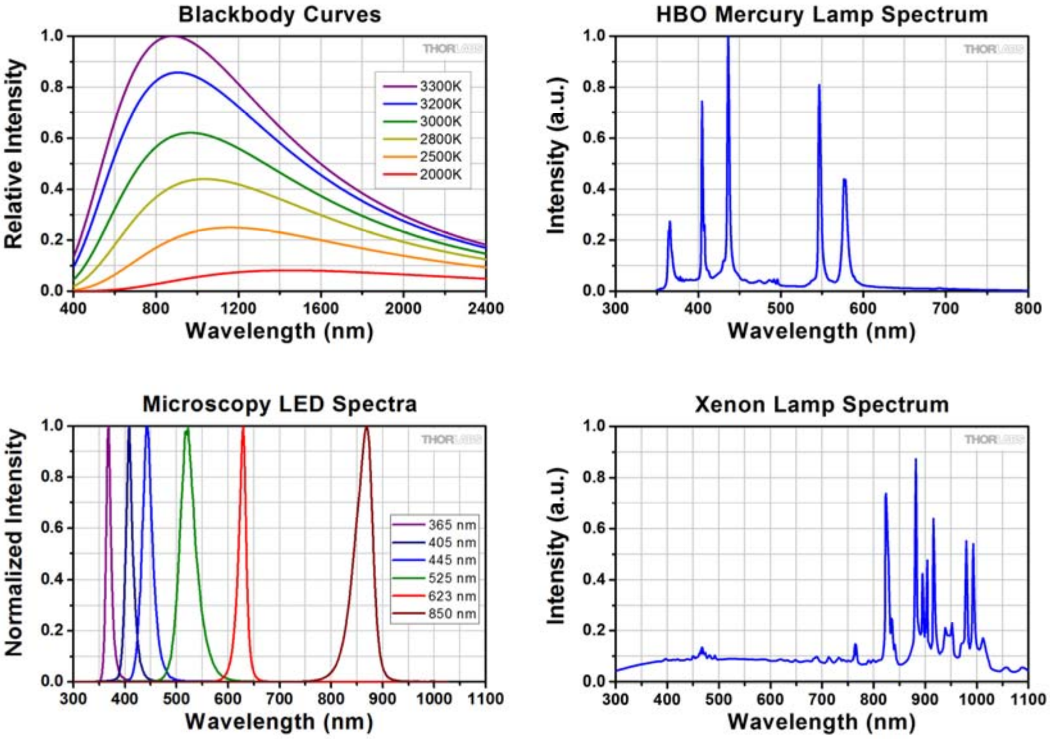


Figure 8. Spectra for various common light sources. Top Left: Thermal source (tungsten-halogen); Top Right: Hg arc lamp; Bottom Left: High power LEDs; Bottom Right: Xe arc lamp. We have left off lasers, since for CW lasers (continuous wave, the most common type) the spectrum is for our purposes essentially a spike at a single wavelength.

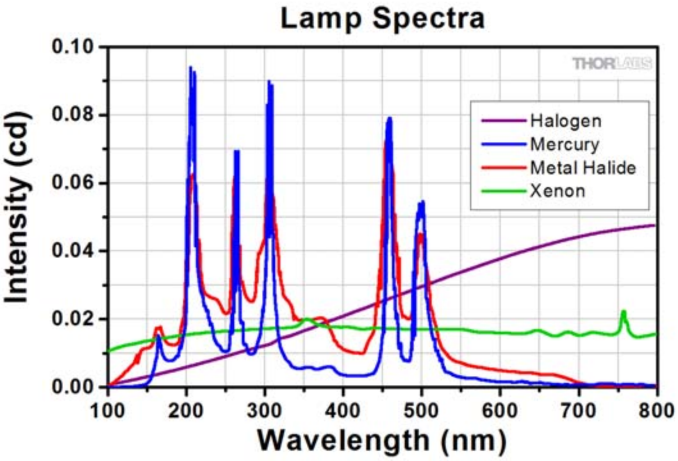


Figure 9: Comparison of various light sources.

Comments on Figure 9:

1. The Halogen (tungsten-halogen) lamp output is multiplied by 10X, and is still small. There is very little power in it, which is why nobody does fluorescence with thermal sources.

2. The Mercury arc gives the most power in narrow bands, but has little power near 470 - 500 nm, where GFP and FITC (very popular fluorophores) excite. The Metal Halide lamp fixes this a bit, at the expense of less power at the peaks.

3. The Xe arc is wonderfully flat through the visible, and has more power at the FITC/GFP excitation than anything else. These are the reasons it is so popular.

4. The y-axis is unfortunately in candela, a unit which involves weighting by the human eye response (Figure 5), and is also based on power rather than photon count. To be useful for a photon budget, one would need to divide these curves by the scotopic response, then convert to photons by dividing by the photon energy at each wavelength.

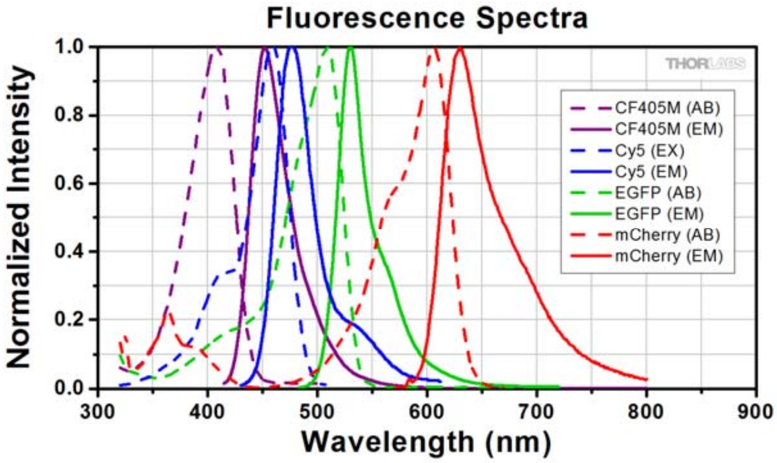


Figure 10: Spectra of some fluorescent dyes: Purple: CF405M (dye); Blue: Cy5 (cyanine dye); Green: EGFP (green fluorescent protein); Red: mCherry (red fluorescent protein). Sources: MacNamara/Boswell and the (excellent) Spectra database at the University of Arizona (Cy5), Tsien lab (EGFP and mCherry fluorescent protein spectra), Biotium Inc. (CF405M).

One can select the dye based on the available excitation light source, or for its overlap (or lack thereof) with another dye – e.g. by using CF405M and mCherry, one could do two-color imaging by changing filters between images. A historic advantage of GFP was that it excited using the (strong) Argon-ion laser line at 488 nm. This convenient light source – now replaced by solid-state lasers developed to duplicate it – drove use and development of a number of dyes exciting at that wavelength (e.g. FITC, GFP, etc.), which remain in widespread use.

Filter Spectra

Now we will begin to put it all together: look at the superposed lamp, dye, and filter spectra below:

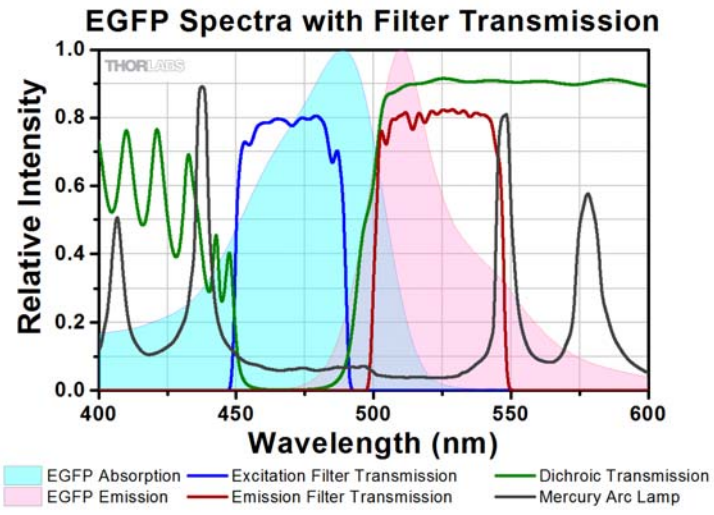


Figure 11: Spectra of an Hg arc lamp (gray), the EGFP fluorescent protein (light blue: absorption; pink: emission) and a filter set designed for EGFP (blue: excitation; green: dichroic; red: emission. Notice the general lack of lamp power in the excitation band – one reason Hg lamps are not so good for EGFP. Also note the strong lamp band at 546 nm; if one was really worried about filter leakage, one would trade off light collection to narrow the emission filter to cut that line out. The blue (~400 nm) transmission of the dichroic does not matter, since the excitation filter will have already removed that energy from the light beam. Just in case, though, the emission filter blocks at those wavelengths.

Here you can begin to see how the spectra combine to affect how much light you detect or excitation you get, or both. Roughly what we want to know is how much of the gray curve (lamp) and light blue curve (EGFP excitation) lie under the excitation filter passband, and how much of the pink emission curve falls under the emission filter passband. The product of those two things will be proportional to the total brightness of the light you get at the detector.

A problem with graphing things separately, as in Figure 11, is that it is hard to sum up (by eye) all of the blue excitation curve under the excitation filter passband, multiplied by the lamp intensity at each wavelength. Of course, ***it is*** easy to do this in software like Excel.

Effective Excitation and Emission

In terms of choosing filters by eye, however, it helps to consider what really matters, and to plot things accordingly. For instance, in terms of excitation, what matters is the product of the lamp photon flux at each wavelength with the dye excitation at the same wavelength. This may not be the same as the dye spectrum alone – if the light source has most of its power off in one region of the dye spectrum, that region will be the important area in terms of excitation efficiency. Similarly, since we are detecting the emitted photons, what will matter is not the dye emission spectrum alone, but the product of the dye emission and the efficiency of the detector we are using (our eye, or the camera, etc) at those wavelengths. This is most easily understood in pictures (which is why you should always plot it!):

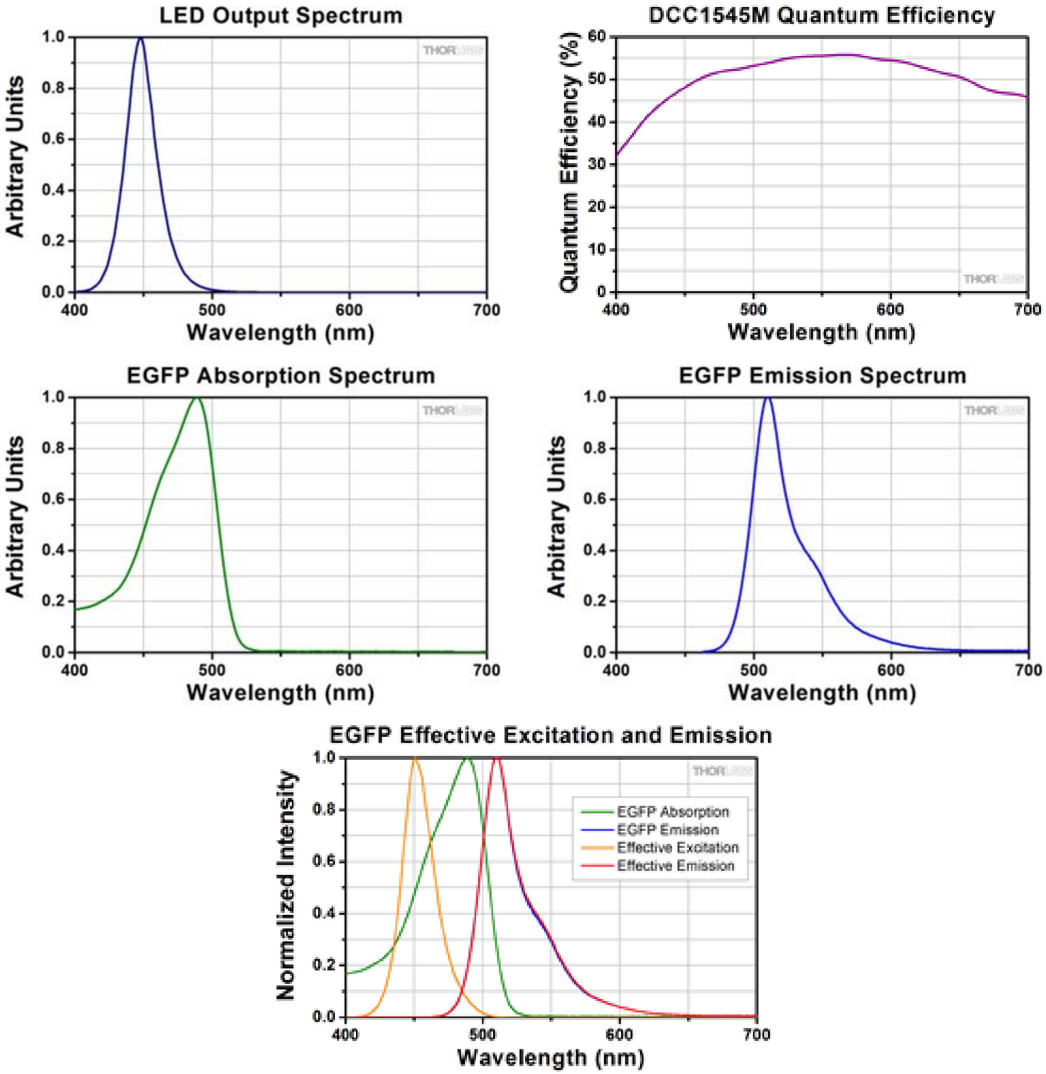


Figure 12: The *effective* excitation (orange line) and emission (red line) spectra are respectively the point-by-point products of the dye absorption and light source photon flux, and the dye emission and detector quantum efficiency. (Bottom): Note that the excitation (an LED, emitting in a narrow range around 450 nm) is *not* centered on the absorption maximum, and so the effective excitation is shifted substantially. The emission falls in a relatively flat area of the camera QE curve, and so it is not changed noticeably.

Plotting effective excitation and emission is very simple in Excel – just multiply the columns for dye absorption and light source output, and for dye emission and detector QE, and plot them.

The reason it is so handy to plot things in terms of the effective excitations and emissions is that it makes it easy to judge by eye which will be the correct filters to use. For instance, the filters one would choose based on the raw dye spectra would naturally overlap those spectra; the filters one would choose based on the *effective* excitation and emission spectra will reflect any shifts those have with respect to the raw dye data. One would expect that a relatively flat light source spectrum would result in little spectral shift of the effective excitation and emission with regard to the data raw dye. In this case, the stock filters (upper left graph in Figure 13 below) would be a good choice. However, if the effective excitation and emission spectra are shifted (due to a light source or detector response with peaks at certain wavelengths), then the filters that overlap best will be different:

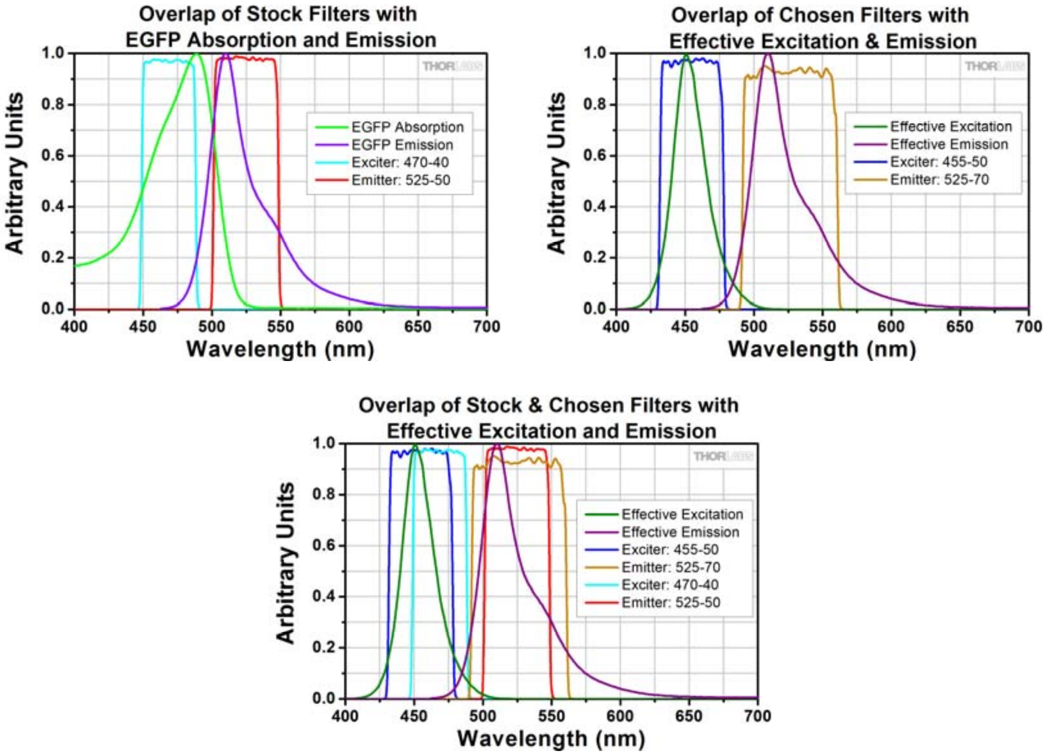


Figure 13: Filter overlap with raw EGFP spectra and with effective excitation and emission spectra (spectra weighted by the light source and detector QE spectra). Note in the bottom graph that the stock filters (red, light blue lines) do not match well with the effective excitation and emission, and are hence not the best choice.

A crucial point is that the units for the various columns in Excel must all be proportional to photons – not to power.

The reason for this is that the detectors we use (CCDs or CMOS silicon cameras) are photon detectors- they will count a blue photon the same as a red one, even though the energies of the photons are different. This would not be true for a bolometric detector (one that detects heat, sometimes called a thermopile); in that case, one would need to adjust properly for the difference in photon energies. In the usual case, however, the detector is based on photon count. That is convenient, because the number of photons a fluorophore emits is proportional to the number of photons it absorbs; the proportionality constant is the quantum yield (QY), and for a good dye it is nearly 1 (0.92 for Alexa 488, for instance). If there are numerous ways for the molecule to de-excite without emitting a photon, then the QY will be lower – poor dyes can have QY < 0.1. Regardless, fluorophores convert photons to photons with a roughly direct proportionality. One must be careful in interpreting dye spectra from various sources: absorption is a direct measure of ε, the molar extinction coefficient, and hence of σ, the cross section. However, emission may be measured either in power or in something proportional to photon counts; if in doubt, the answer is most likely photon counts, since the detectors in most spectrometers used for fluorescence measurements are silicon band-gap devices like the cameras. Still, read the figure axes carefully.

Light sources are a different matter – except for lasers, most sources are reported in units of power (or, equivalently for our purposes, power/unit area/unit wavelength/unit solid angle). Worse, often they are reported in units of lumens (lm) or candela (cd), which are power units weighted by the response of the human eye. This is handy if one wants to know how brightly lit a room will appear when using that light source, but pernicious if one wants to figure out how best to use it for fluorescence – to get to the number of photons emitted, one must divide by the scotopic eye response, and then divide by the photon energy at each wavelength. This is all a mess – ideally one can get the data in units of power, at which point one need only divide by the photon energy at each λ:

$$\textbf{Equation 15: Intensity (proportional to \# of photons)} = \frac{\boldsymbol{power}}{\frac{\boldsymbol{hc}}{\boldsymbol{\lambda}}} = \frac{\boldsymbol{power*\lambda}}{\boldsymbol{hc}}$$

where h = Planck's constant, $6.626\ 10^{-34}$ J·s and c is the speed of light, 3e8 m/s. Doing this calculation at every λ is easy to do in Excel (as is dividing by the scotopic response, if necessary to convert from lumens to power units).

Assuming one has gotten the various data into photon-based units, things become quite simple; on a column-by-column basis (with wavelength in the first column), one calculates:

1. Effective excitation = dye absorption * light source spectrum
2. Effective emission = dye emission * detector QE spectrum
3. Applied excitation = effective excitation * excitation filter spectrum
4. Obtained emission = effective emission * emission filter spectrum
5. Total excitation = sum of applied excitation column
6. Total emission = sum of obtained emission column
7. Total brightness = total excitation * total emission
8. Total leakage = sum (light source * excitation filter * emission filter * detector QE)

With this, one can produce the "figures of merit" on which to judge choice of filters:

1. Total brightness; the higher this is, the shorter exposure times you can use, or the brighter things will look by eye.

2. Total emission; the higher this is, the smaller any bleaching problem will be.

3. Total leakage; the lower this is, the lower the background in your image. Since background produces noise (shot noise = √# background photons), and in any case it reduces your camera dynamic range, lower is better.

4. S/B: Often one is concerned with signal to background (S/B) and signal to noise (S/N). Relative S/B is just total brightness / total leakage.

5. S/N: Since the noise depends on the square root of the absolute # of photons, it cannot be well judged in relative terms. If it really matters, one must carry the whole photon budget through exactly (all units absolute, not relative). This is definitely harder to do, and usually only accurate to a factor of ~ 10, and so must be measured experimentally (usually people do not calculate it ahead of time unless forced to). Metrics 1 - 4 are easy to do, though.

When comparing two (or more) filter combinations, one computes the above and compares them, as shown in the *Filter Selection Example* spreadsheet (under the *Reference Links* tab at www.thorlabs.com/OMC). For the sets shown in Figure 13, the stock set is 59% as bright as the set we chose, but has 57% of the leakage, so its S/B is 4% higher. The stock set collects 15% less of the total emission, though. In general, we would choose the brighter set, since one accumulates more noise during longer exposures, likely offsetting any nominal gain due to slightly lower leakage in the stock set.

Notably, just by eyeballing the plot in Fig. 12, we were able to get nearly half again the brightness, and 15% more of the total emission. Not bad, for < $1k. Full optimization (which we will not lay out in these notes, but is not hard once you understand the basics of the above) would probably give another 10 - 20% benefit.

There is one last thing of major importance in choosing filters: leakage. While we laid out the calculation of this above, the easiest way to check it is by eye, with the filter curves plotted on a log scale:

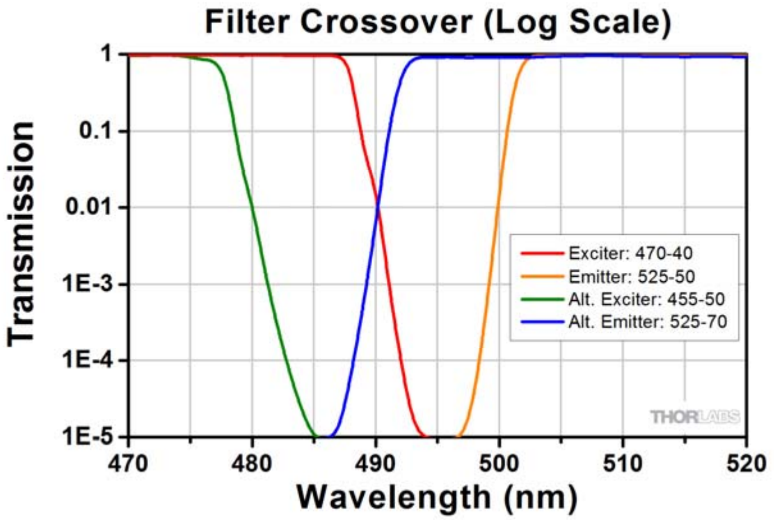


Figure 14: Filter crossover, log scale. Note that either set of filters crosses at OD > 5, for a combined OD > 10 at crossover (since transmissions multiply, log(T) – i.e., ODs – add). This is excessive; in the emission filter transmission band, OD is ~5 from the excitation filter, and vice-versa, so there is little point in having higher OD over a few nm of bandwidth near crossover. Having the filters cross at OD 3+ gives OD 6+ blocking through the entire band, and allows the filter transitions to be brought closer together so that you can catch more of the emission light. Typically ~ 10 - 15 nm between the excitation filter cut-off to the emission filter cut-on is fine.

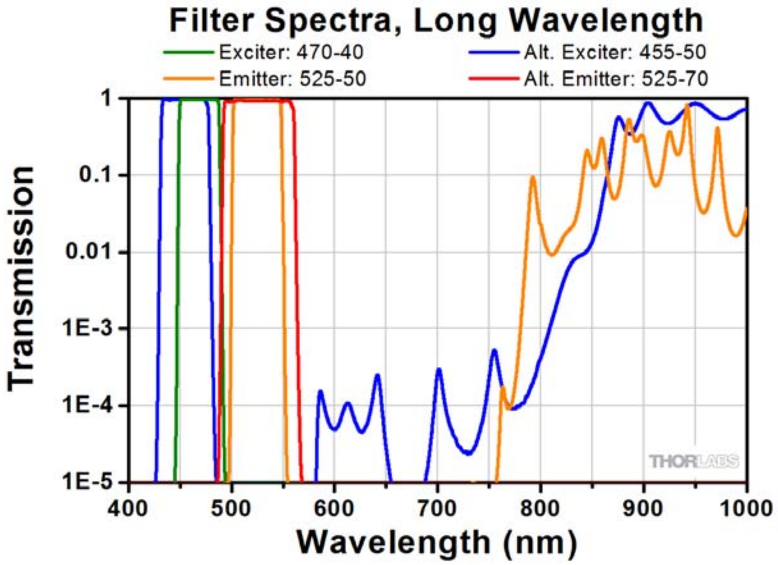


Figure 15: Log plot of the entire spectrum of filters is important; here notice that the emitter for the stock set (orange line) has poor IR blocking, so it is important to make sure the excitation filter is good out there. Conversely, the excitation filter I chose (blue line; actually, an emission filter re-purposed) has a lot of IR transmission, so it is crucial my emission filter blocks well there. In general, emission filters often have IR transmission, while excitation filters usually do not, but beware and always plot the full spectra so you can check!

Often checking both the IR filter behavior and OD at crossover by eye is sufficient to make a pretty good filter choice.

With this, you should be able to choose filters pretty effectively; what follow are a few advanced notes which are less critical for right now but useful if/when you actually choose your own later on.

You do not have to read this part now:

1. We assumed a direct-illumination path (as in your current rigs) in all the discussion previously. Usually one does epi-illumination, which uses a dichroic filter.

    a. The dichroic must *reflect* the excitation, so one just multiplies the excitation filter spectrum by (R = 1 – dichroic transmission) – easy to do in Excel.

    b. The emission must *pass through* the dichroic, so one must also multiply the emission filter spectrum by the dichroic transmission spectrum.

    c. After doing that, everything else is the same.

2. Often one worries about autofluorescence, at which point filter choice gets trickier. By plotting the spectrum of the things one expects to autofluoresce (e.g. plastic in a sample holder, or hemoglobin in tissue, etc.), one can purposefully choose filters that block light that would excite autofluorescence or that block autofluorescence emission from getting to the camera.

    a. Often, however, one does not know the exact details of the autofluorescing stuff. The trick then is to avoid large excitation power at wavelengths where your desired dye does not excite efficiently. In this case, you might decide to forgo a big peak in the *effective* excitation that occurs on the edge of the raw dye absorption spectrum, since such a peak implies lots of light power there but little absorption by your dye. So you get more possible autofluorescence for less possible extra dye excitation – a bad tradeoff.

    b. Another trick is to use bandpass (rather than long-pass) emission filters that cut off when the effective emission drops < 10%. Again, the reason for this is that you are getting little extra light from your dye at this point, but still picking up who knows how much background – a bad ratio.

3. As an advanced note, there is an oversimplification in our formulae before; it generally makes little difference, but should you want to be completely precise, we discuss it here:

    a. If the excitation wavelength overlaps the emission spectrum, then since any emission is generally at *longer* wavelengths than the excitation, the *effective* emission spectrum is enhanced at the wavelengths longer than that particular excitation. As an example, if we excite EGFP at 400 nm, then our emission could fall practically anywhere in the emission spectrum, while if we excite exactly the same number of times at 550 nm then any emission must be >550 nm (see Fig. 13, purple trace). One can think of the emission spectrum as a probability: the photon comes out at *some* wavelength, so the total area under the curve must = 1. For the 550 nm excitation, the area under the curve >550 nm must still be equal to 1 so the effective emission curve is effectively higher for all those points, and given that particular excitation wavelength, essentially zero at wavelengths <550 nm. The practical effect of this is that for a very broad excitation (meaning wide bandpass excitation filter and light source with a broad spectrum), the effective emission may shift a little bit to the red. However, taking this into account quantitatively is beyond the scope of this class.

THORLABS

Appendix A:
Equations to Memorize

Optical Microscopy
Course

Appendix A: Equations to Memorize

Please memorize the formulas on this page. They will show up on quizzes throughout the semester. The following pages discuss the equations in some additional (but brief) detail to help clarify things, but you need not memorize all that content. You will cover it all more thoroughly during the course.

Name	Paraxial Approx.	Full Version
Numerical Aperture	$NA \cong \frac{r}{s_{Object}} \cong \frac{r}{f}$	$NA = n\,sin(\theta)$
Magnification	$M = -\frac{S_{Image}}{S_{Object}}$	$\lvert M \rvert = \frac{NA_{Object}}{NA_{Image}}$
Optical Invariant	$R_{image}\,\theta_{image} = R_{object}\,\theta_{object}$	$R_{image}\,NA_{image} = R_{object}\,NA_{object}$
Depth of Field	$DoF = \frac{n_{sample}\,\lambda}{NA^2}$	$DoF = \frac{\lambda}{4\,n\,sin^2\left(\frac{\theta}{2}\right)}$
Collection Efficiency	$CE \cong \frac{NA^2}{4\,n_{sample}^2}$	$CE = sin^2\left(\frac{\theta}{2}\right)$

Name	Self-luminous Sample	Transilluminated Sample
Rayleigh Criterion	$\delta = \frac{0.61\,\lambda}{NA}$	$\delta = \frac{1.22\,\lambda}{NA_{Objective} + NA_{Condenser}}$
Max. Spatial Frequency	$k = \frac{2\,NA}{\lambda}$	$k = \frac{(NA_{Objective} + NA_{Condenser})}{\lambda}$
Nyquist Pixel Spacing	$d \leq \frac{M\,\lambda}{4\,NA}$	$d \leq \frac{M\,\lambda}{2\,(NA_{Objective} + NA_{Condenser})}$

Name	Equation
Snell's Law	$n_1\,sin(\theta_1) = n_2\,sin(\theta_2)$
Back Focal Plane Diameter	$\text{BFP dia.} = 2\,\text{f}\,\text{NA}$
Wavelength in a material	$\lambda = \frac{\lambda_{Vacuum}}{n}$
Thins Lens Formula	$\frac{1}{f} = \frac{1}{S_{Object}} + \frac{1}{S_{Image}}$
Reflection Coefficient	$R = \left[\frac{n_2 - n_1}{n_2 + n_1}\right]^2$

- Field of view, microscope objective: $\text{FoV} = \frac{\text{FN}}{M}$; FN = Field of view number, and M = magnification. Typically the FN is about 20 - 25 mm, and M is printed on the objective barrel.
- Note: In the Depth of Field equations n is the index of the immersion medium, not the lens glass.

Comments

In general, it is much better to understand where something (e.g. an equation, or a name) comes from than to simply memorize it – if merely memorized, once you forget it, it is gone. If you know where it came from, you can re-derive it when you need it, or at least get back to the general relationship even if you forget the precise numerical factors.

During this course, we will cover all of the equations discussed below, and either derive them, discuss intuitive ways of getting the most important parts, or both. However, doing optics (like all science and engineering) requires having at least some facts quick at hand. This is critical: being able to quickly analyze a system and decide if it will do what you want (or what someone else claims) is what distinguishes good scientists and engineers from average ones – decisions are most frequently made in group discussions, and it is usually only a few of the more critical points that are slated for further investigation after the main direction is decided. Being able to hone in on the right direction *during the discussion* is very powerful – it will make you valuable to the team, make people respect your technical ability, and allow you to spend the precious analysis time on the most important (and interesting) parts of the problem.

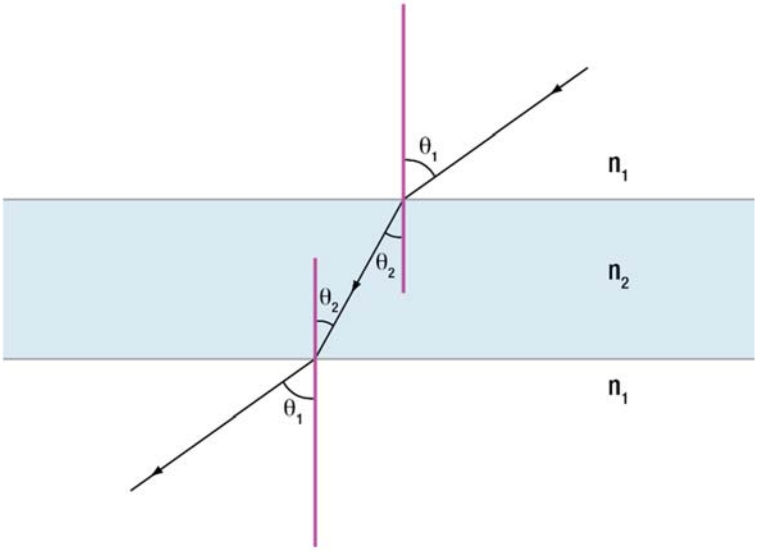
Snell's Law

$$n_1\, sin(\theta_1) \;=\; n_2\, sin(\theta_2)$$

Snell's law is the relationship determining the bending of a wave at an interface between two media. It is not specific to light – acoustic waves bend (and can be focused) the same way (e.g. in medical ultrasound).

It is useful to remember that *the angle is always higher on the low-index side* of an interface; equivalently, ***light bends toward the surface normal when crossing into a higher index* material.**

(Note: the "surface normal" just means a line perpendicular to the surface at that point.)



Obviously, Snell's law underlies our ability to make lenses the way we do, as well as how light deviates when it goes from one medium into another. Interestingly, Snell's law can be derived just from the fact that light is a wave[1], and an identical relation holds for any wave incident on an interface, such as water waves, sound waves, radio waves, etc.

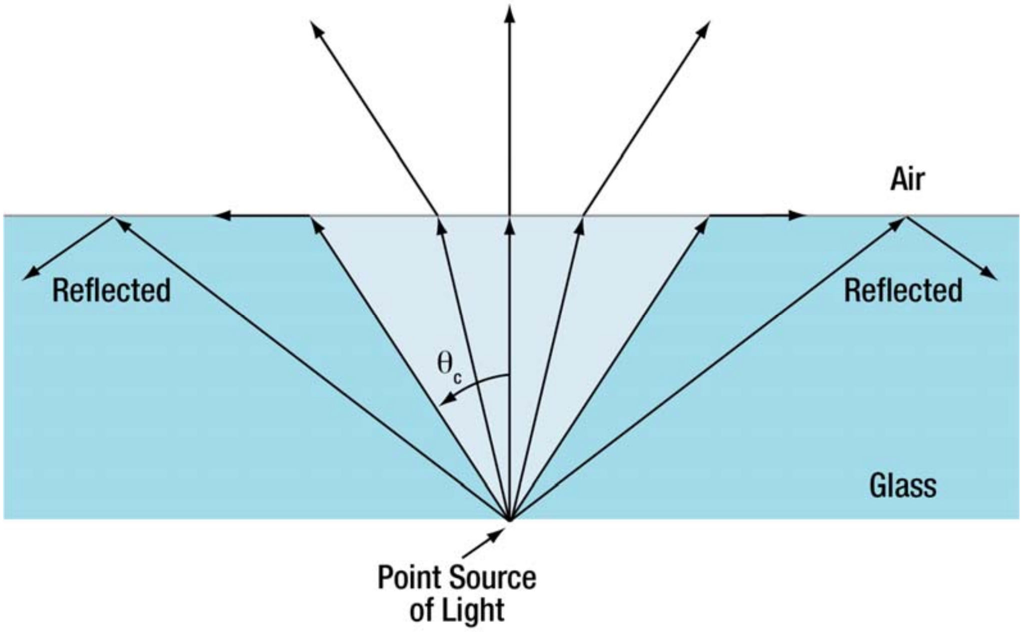
[1] See the Lab 1 Course Notes for further discussion.

Critical Angle

What if the angle with which the wave is incident is such that Snell's Law *cannot be satisfied at all*? Since by the nature of the sine function sin(θ_2) ≤ 1, this will happen for any angle θ_1 for which Snell's Law would give n_1 $sin(\theta_1) > n_2$. The angle where this first happens is called the critical angle, $\theta_{critical}$:

$$sin(\theta_{critical}) = \frac{n_2}{n_1}$$

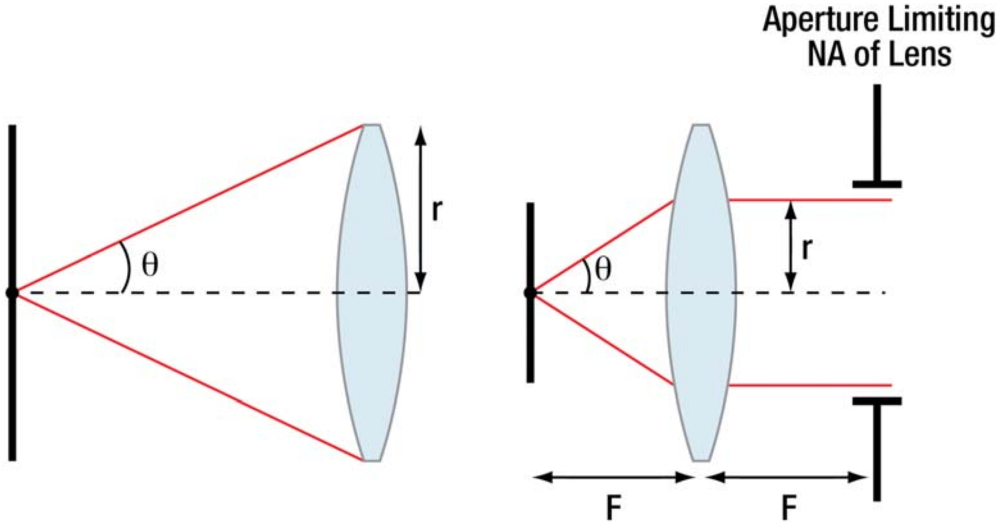
There will only be a critical angle when light is going from a higher index n_1 into a lower index n_2. For situations where $n_2 > n_1$ Snell's Law can always be satisfied and light will be able to transmit for any angle θ_1 (though reflection does get higher at glancing angles, for reasons beyond the scope of this course). However, for $n_2 < n_1$ – i.e., when the light is incident on a lower index material (e.g. going from water into air) – beyond $\theta_{critical}$ the wave cannot transmit into the lower-index material, and is totally reflected. This effect is known as TIR, for Total Internal Reflection. This is of special interest in microscopy because commonly the sample of interest is in an aqueous buffer (n ~ 1.38), or fixed in an oil-like substance with an index similar to glass (n ~ 1.51). If there is air between the sample and the lens, then the higher-angle rays will be totally internally reflected and thus be unusable for imaging or light collection purposes.



To avoid this problem (and for reasons related to resolution, discussed later) one will often use an "immersion lens," also known as an "immersion objective," where the lens sits directly in the water bath the cells are in, or has a layer of water or oil between it and the cells or glass coverslip (the oil is chosen to have the same index as glass and/or cell mounting medium). As a result, the rays do not bend crossing the boundary (or bend inwards rather than outwards), so there is no reduction in the angles of light collected from the sample due to TIR loss. TIR is the reason that the maximum NA of a lens is set by the *lowest* index between it and the sample.

Numerical Aperture

$$NA = n\, sin(\theta)$$



Numerical aperture is related to the maximum angle of light that can be collected by a lens. Note that the usual formula for it $\left(NA = \frac{r}{f}\right)$ assumes that the entire lens is available to collect light, as in the left image above; obviously, if something restricts the rays of light that are collected by a lens, the NA will be reduced, as shown in the right image above. Be sure to understand that; many people do not.

A common mistake is to have a narrow laser beam going through a large diameter lens, and to use the lens diameter (not the beam diameter) to calculate the NA and hence the minimum spot size the laser will be focused to.

An under-filled lens has a lower NA than a fully filled lens.

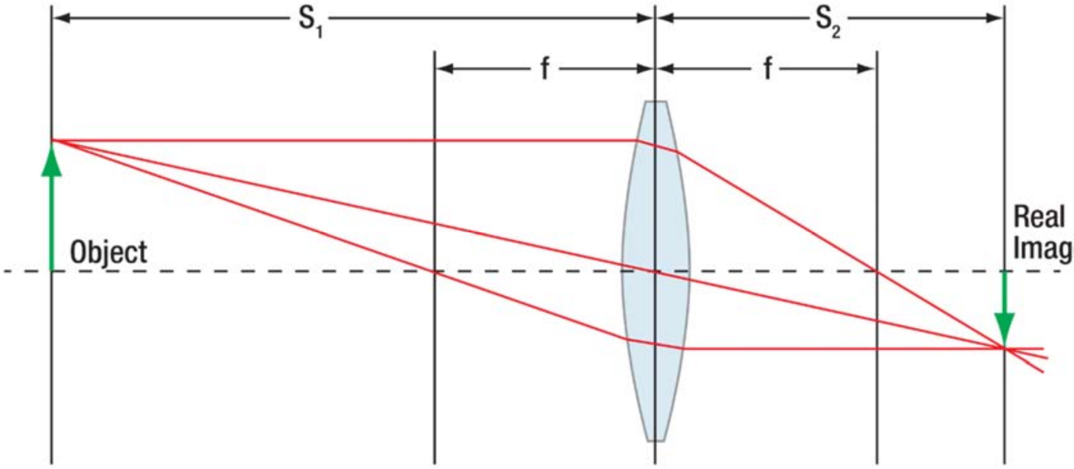
Notably Snell's Law can be re-written $NA_1 = NA_2$, which is to say the NA of light does not change as it goes through a planar surface. This is quite useful, and is part of the reason people often speak of the NA of light rays rather than the angle.

In the paraxial (small-angle) approximation, the NA can be written $\cong \frac{r}{f}$, where r is the radius of the lens used, f is the focal length, and n = 1 is assumed. The assumption made here is that the sample is placed a focal length away from the lens; if the sample is farther away, the NA should be approximated $NA \cong \frac{r}{S_{object}}$.

➔ **It is a common mistake to use f instead of S_{object} (or, if interested in the image-side, S_{image})**

Thin Lens Formula

$$\frac{1}{f} = \frac{1}{S_{object}} + \frac{1}{S_{image}}$$



Note: The is image is inverted (upside down); hence the "−" sign in the formula for Magnification

Magnification

$$|\mathrm{M}| = \frac{NA_{object}}{NA_{image}}$$

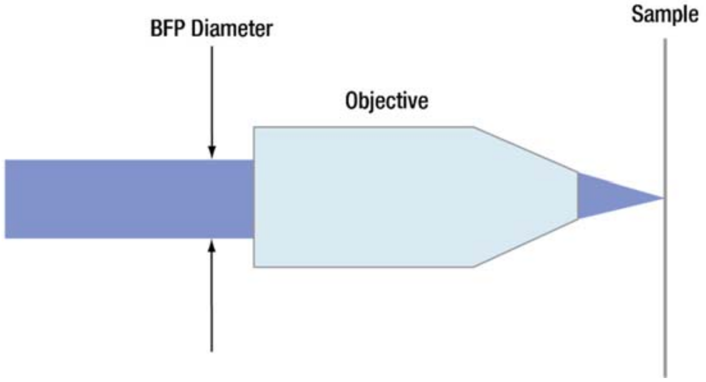
This is true for all aberration-free imaging; it boils down to the Abbe sine condition (which we will cover midway through the course). Because the ray going through the middle of a thin lens is undeviated, the angle it makes on each side is equal and opposite. Thus (see the image above) geometry tells us that $h_{object} = S_{object}\tan(\theta)$ and $h_{image} = S_{image}\tan(\theta)$ which can be rearranged to give

$$M = -\frac{h_{image}}{h_{object}} = \frac{S_{image}}{S_{object}}$$

For a single thin lens and small angles (small enough that $tan\theta \cong sin\theta$), this formula for M becomes equal to the NA-based version. For higher angles, the equivalency is lost, unless a lot of design effort is put into the lens system, as in the case of microscope objectives, where these two formulae continue to be equivalent even at very high NAs.

Back Focal Plane Diameter

$$BFP\ dia. = \ 2\ f\ NA$$



This is the diameter of the collimated beam from a single point source at the focus of a lens. It is not intuitive; geometrically one would be correct in thinking that the relationship should involve the tangent of the collection angle, not the sine. However, good imaging (via the Abbe imaging criterion) requires that the relationship be that of the sine, and microscope objective designers go to a lot of trouble to make their lenses obey that relationship instead of the seemingly more natural tan(θ) one. For those who wish for more detail: treating the objective as a thick lens, the design effort modifies the first principle plane of the objective to be a curved (spherical, with radius f centered on the sample point) principle surface, with the result that the height at which a ray hits the surface is proportional to sin(θ). However, it is much easier to simply remember that the BFP *diameter* = 2 f NA.

Collection Efficiency

$$CE = sin^2\left(\frac{\theta}{2}\right)$$

This is the fraction of light collected from a point source emitting equally in all directions ("isotropically"). Note that while the solid angle – the formula given above – does not depend on the index, n, the actual collection efficiency may if more than one index is involved. To see why, consider an oil-immersion (n = 1.5) objective imaging a sample fixed in n = 1.5 media. In this case, Snell's law says the light rays will not bend at all in going from the sample to the objective entrance, and so the objective will collect the full range of angles expected, and the formula above will be correct.

It is best to use $CE = sin^2\left(\frac{\theta}{2}\right)$ for oil- and water-immersion objectives, especially high-NA ones

At a smaller NA, this formula would seem to reduce to $CE = sin^2\left(\frac{\theta}{2}\right) \cong \left(\frac{\theta}{2}\right)^2 \cong \frac{\theta^2}{4} \cong \frac{sin^2(\theta)}{4} \cong \frac{NA^2}{4} \ldots$ However, imagine the (fairly common) case of an air objective (n = 1) imaging a sample in biological buffer (n ~ 1.4). The rays coming from the sample will bend *away* from the surface normal when exiting the sample into the air, and so the cone of angles collected from inside the sample will be smaller than the cone defined by the objective NA; one can find the new angle by using Snell's law. Specifically, for $n_{immersion}$ = 1(for air), NA = n sin(θ) = sin(θ), and Snell's law then says the half-angle collected from *inside* the sample will be $sin(\theta_{sample}) = \frac{NA_{objective}}{n_{sample}}$. As a result, for most low- and moderate-NA objectives, the formula becomes:

$$CE \cong \frac{NA^2}{4\, n_{sample}^2}$$

It is best to use $CE \cong \frac{NA^2}{4\, n_{sample}^2}$ for air objectives imaging into higher index media

This correction for air objectives typically results in ~2X lower light collection (i.e., 1.4^2 or 1.5^2) than would otherwise be expected. In a similarly important way, use of the full (trigonometric) CE equation for high-NA immersion objectives gives about 50% higher light collection than the $\sin(\theta) \sim \theta$ approximation would suggest.

What is most useful to remember is that the collection efficiency scales as CE ~ NA^2.

Rayleigh Resolution

$$\delta = \frac{0.61\, \lambda}{NA}$$

This actually only holds for two point objects which are both self-luminous and incoherent – individual fluorescent molecules are a good example, while two tiny pinholes illuminated by a plane wave are not (in the latter case the light emitted by the two pinholes will be *coherent* and the light from them will interfere in a consistent way. With *incoherent* sources the light (on average) displays no interference effects between the two sources.)

For samples illuminated by incoherent light (what you get when the condenser aperture is open more than a tiny pinhole), the angles of the light illuminating the sample also matter, as can be seen by the presence of the condenser NA in the resolution equations:

$$\delta = \frac{1.22\lambda}{NA_{objective} + NA_{condenser}}$$

Maximum Spatial Frequency

$$k = \frac{2\, NA}{\lambda}$$

The derivation of this formula is discussed at length in the Lab 6 Course Notes on the Abbe Theory of Image Formation.

This formula only applies to self-luminous objects; more generally for trans-illuminated samples the condenser NA will also matter, just as it does for resolution:

$$k = \frac{NA_{objective} + NA_{condenser}}{\lambda}$$

You should note the reciprocal relationship between these formulas and the resolution formulas above – that is no accident.

Nyquist Pixel Spacing

$$d \leq \frac{M\,\lambda}{4\,NA}$$

The Nyquist criterion requires that to avoid aliasing (which can cause undesirable image artifacts) **the digital sample rate must be at least *twice* the highest spatial frequency**. The resolution of objectives varies, as does the spacing of the pixels in cameras; the way one matches these so that sampling is sufficient is by choosing an appropriate magnification – hence the appearance of M in the equation above. This should not be a big surprise, given the previous formula linking magnification to the ratio of the image- and object-NAs, and the fact that the maximum spatial frequencies present depend on these NAs.

The most general way of writing the requirement is in terms of the pixel spacing d, spatial frequency k, and magnification M. Since 1/d is the sampling frequency, we require:

$$\frac{1}{d} \leq 2\,k_{image}$$

But the magnification changes the spatial frequency at the image compared to the object; since M is usually defined from object to image, $k_{image} = \frac{k_{object}}{M}$, and combining these two equations with the relation between k and NA yields

$$d \leq \frac{M}{2\,k_{object}} = \frac{M\lambda}{2(NA_{objective} + NA_{condenser})} \approx \frac{M\lambda}{4NA_{object}}$$

From this and the formulas for maximum spatial frequency you can see that the first equation above is a worst-case estimate (if you use it, you will be safe for sure). The equality only holds for self-luminous objects or cases where the $NA_{condenser} = NA_{objective}$. In cases where the condenser NA is lower than the objective NA, the maximum spatial frequency in the image will be lower and one can plug in the appropriate NA values to find the pixel spacing necessary to avoid aliasing. In practice it can be useful to oversample if one needs to fully reconstruct the image; the price paid for oversampling is a smaller amount of the object gets imaged, since more pixels are being used per unit distance and the camera has a limited number of pixels. For that reason, many mobile phone and consumer digital cameras actually *under* sample, gaining a larger field of view with a smaller number of pixels, at the cost of resolution (and some aliasing).

Depth of Field

$$DoF = \frac{\lambda}{4\,n\,sin^2\left(\frac{\theta}{2}\right)}$$

The depth of field is the distance one can move the sample along the optical axis before it starts to look blurry. The definition listed here is the total focal depth – from where the sample is blurry on one side of the focus to the point where it is blurry again on the other side. Be careful – many papers and books define the DoF as from the plane of perfect focus to where it gets blurry on one side, different from our definition by a factor of two. The formula above is also the correct formula for high-NA systems; again, be careful – it is stated incorrectly in a number of places. It is worth noting that the DoF is a very strong function of NA – at low NA it can be quite long, whereas at high NA it can be very short such that only very thin sections of a sample are in focus at any given time. The DoF is essentially the axial resolution; notice that the axial resolution scales differently with NA than the lateral (Rayleigh) resolution. At low NA (paraxial approximation) the formula becomes:

$$DoF = \frac{n\,\lambda}{NA^2}$$

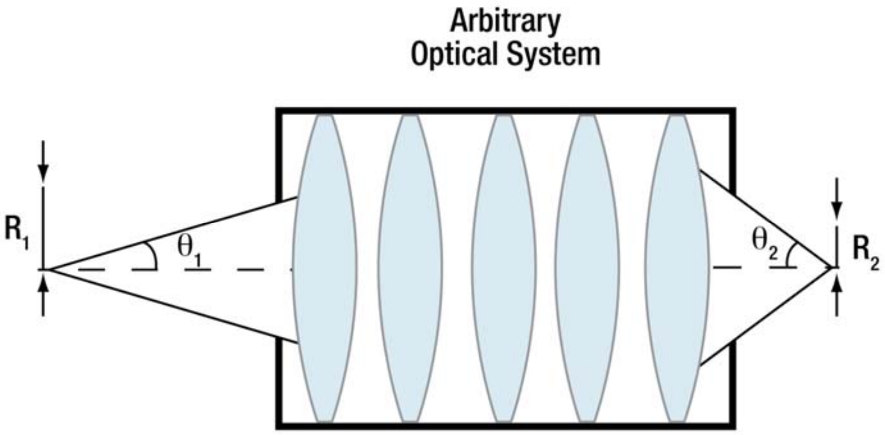
The important thing is to remember that DoF scales roughly like 1/NA².

Two additional notes: first, the index n in this equation is the index of the medium the sample is in (or of the region in which the light is focusing), not that of the glass lens. Second, this formula assumes diffraction-limited imaging, with sufficient sampling of the image to tell when the image starts to blur. There are other equations for DoF, used for situations where no imaging is done and all the light is put onto a single detector (e.g. a big photodiode, essentially a single pixel). Those non-imaging applications do not concern us here.

Optical Invariant

$$R_{image} \cdot NA_{image} = R_{object} \cdot NA_{object}$$

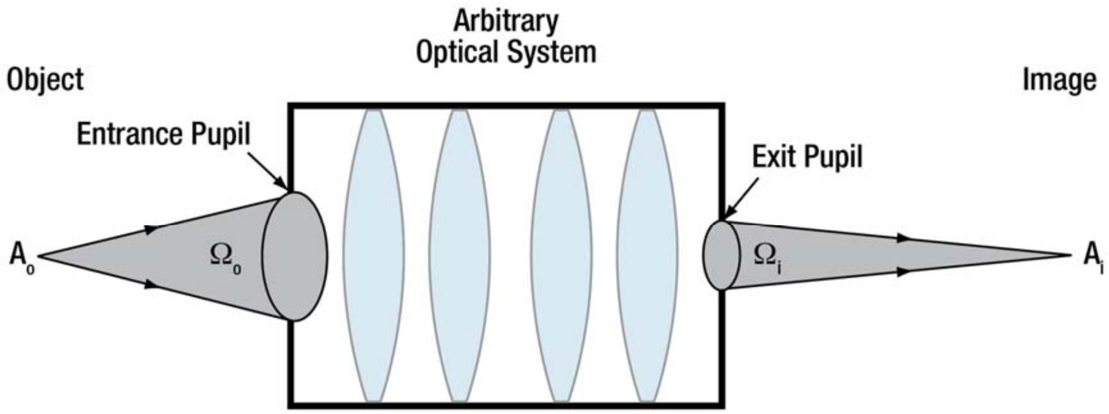
The optical invariant goes by a number of names, including Étendue, Lagrange Invariant, and Optical Extent. In short, the product of the angular distribution of light and the emitting area must be constant at any image plane in a lossless optical system.



Since the area of an image is proportional to the square of its radius, and the solid angle Ω is proportional to the square of the NA (at least in the paraxial approximation, where it is easy to prove this relation), this can also be written in a 2-D form:

$$n_{image}^2 A_{image} \Omega_{image} = n_{object}^2 A_{object} \Omega_{object}$$

Note the factors of *n* (contained in the NA) and *n*² respectively in the 1- and 2-dimensional versions of the formula. Often the index at the image and object planes is the same, but if not then one must take the index into account.



These relationships can be used, e.g., to derive the magnification as the ratio of NA's, and also to put an upper limit on how much one can focus a lamp down onto a sample – one cannot make the lamp spot infinitely bright by demagnifying it to make it smaller. This is counterintuitive, but a lower-power lamp with a small emitting area is better for high-intensity microscope illumination than a high-power lamp that has a proportionally larger filament (or arc). Light intensity is power/area, and the larger power cannot compensate for the inability to focus the larger arc down to a sufficiently small size due to limitations based on the *etendue* formula.

Separately, combining the Optical Invariant with the Depth of Focus reveals that the focal depth at the image is M^2 times the focal depth at the object. As an example of how this works, imagine using a 20X, 0.4 NA objective to image a sample. The objective has a depth of focus of ~ 3 µm; however, at the image you could move the camera up to 20^2 * 3 µm = 1.2 mm before the image is out of focus.

Wavelength of Light

$$\lambda = \frac{\lambda_{vacuum}}{n}$$

The wavelength of light becomes shorter in matter; this allows for better resolution since the diffraction limit (Rayleigh resolution criterion) depends on the wavelength. This is part of the reason people use oil-immersion objectives (oil has an index of n = 1.5, so the possible resolution becomes 1.5 X higher than it would be in air; notice that the NA in the Rayleigh resolution formula contains the index, n). Throughout this document, and in general, when you see λ it means the wavelength of light of that frequency *in vacuum* – while the frequency does not change when a wave goes from one medium to another, its wavelength does. Hence, people reference the wavelength to that in a specific medium (usually a vacuum).

Reflection Coefficient

$$R = \left[\frac{n_2 - n_1}{n_2 + n_1}\right]^2$$

The reflection coefficient[2] given here is for "normal incidence," i.e., for light hitting a surface perpendicularly. That is generally a good approximation, but at higher angles (≳40°) the reflection losses can increase dramatically from this formula. A good number to remember is that a single air-glass

[2] This is also the intensity reflection coefficient, which is the square of the coefficient for the field amplitude.

interface (n = 1.0 into n = 1.5) will give 4% reflection, so one loses 8% just going through a single glass plate (two surfaces). For objectives with as many as 14 lens surfaces (and often high angles of incidence too), it should be apparent why antireflection coatings are critical.

Shot Noise

$$\sigma \propto \sqrt{signal}\,,\ \therefore\ \mathrm{S/N} \propto \sqrt{signal}$$

Photons arrive in a fundamentally random way governed by Poisson statistics. A characteristic of Poisson statistics is that the variance of the distribution is equal to the mean. Since the standard deviation σ is the square root of the variance, we get the formulas above.

Be careful to use the actual photon (or photoelectron) count to figure the noise level – using some arbitrary units (say, the digital level coming out of a camera) will give the wrong answer (the constant relating the number of photons to the number of digital counts then shows up in the equation too; you can actually use that fact to calibrate the camera). The main point to note is that, at light levels often encountered in microscopy, the signal-to-noise ratio of an image can be significantly limited by shot noise, and that noise can only be reduced by collecting more light. As an example, say a typical fluorescent molecule emits ~10^5 photons before it photodestructs (“bleaches”). You might collect and detect about 1% of those, resulting in 10^3 photons. The S/N ratio for that number of photons is only ~30, and there is nothing you can do to improve that other than to try to keep the molecule from photodestructing as fast, so that you can get more photons out. More generally, even when you do not know the calibration in terms of photons per digital number (say, the numerical value from some pixel in a camera), one can still assess how much *more* light you might need in order to reduce the noise level by a given amount, just by using the proportionality: $\mathrm{S/N} \propto \sqrt{signal}$.

Field of View

$$FoV = \frac{FN}{M}$$

Microscope objectives are designed to image accurately (have small aberrations) over a specific field of view. This is given by the “field of view number,” which is usually buried in the specifications for the objective and hard to find. The field of view number is understood to be in millimeters at the image plane of the objective and its tube lens (if any), so the actual field of view *at the sample* is smaller by the amount of magnification. For example, a 100X widefield objective with FN 26.5 (hence the “widefield” – a regular field would be FN = 18 to 25) will have a field of view of 265 µm (= 26.5 mm / 100) at the sample.

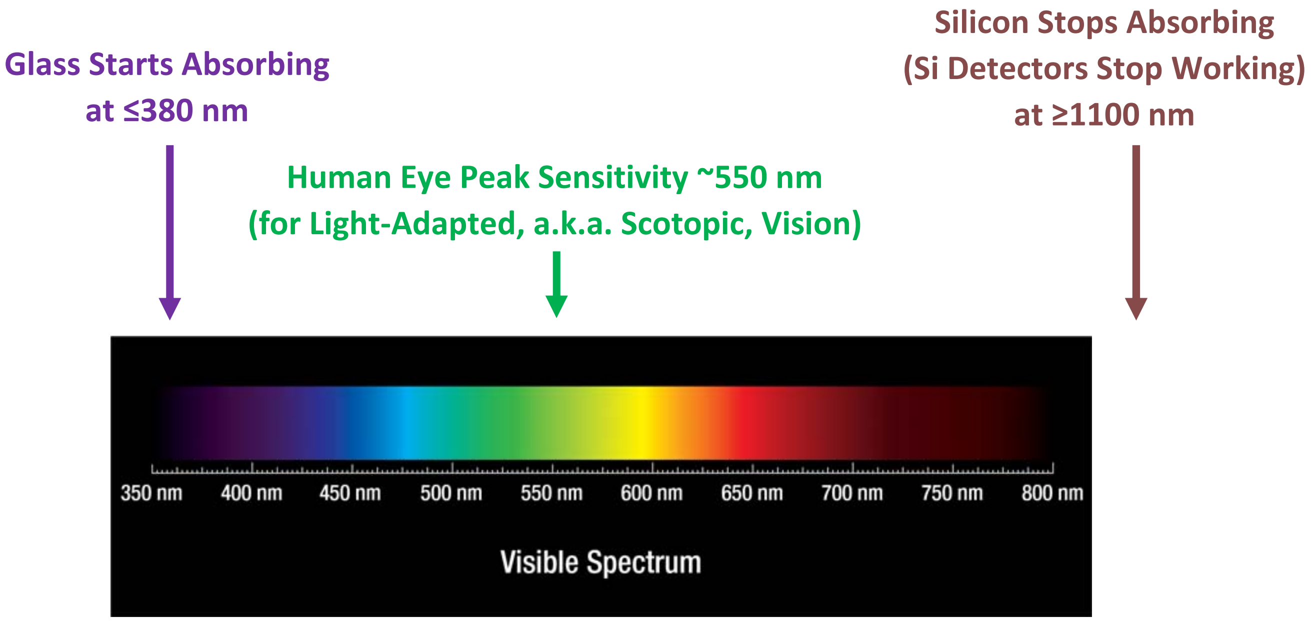
Optical Spectrum and Colors

Since optics depends on the movement and detection of light, it is critical to be able to think quantitatively in terms of the spectrum (color) of light. The usual units for the wavelength of light are nanometers (nm, 10^{-9} meters) or Angstroms (Å, 10^{-10} m), but it is not unusual for people to use microns (µm, 10^{-6} m). It is helpful to be able to think in these terms, at least to some extent; one way to do this is to consider everyday objects. The wavelength of green light (near the peak of the human eye visual response) is ~0.5 µm = 500 nm = 5,000 Å.

In everyday terms, 500 nm is:

- 1/50,000th of an inch, which is 1/50th of the best cutting tolerance of a good metalworking mill or lathe (which can usually hold a bit better than one thousandth of an inch (25 µm) tolerance)
- about 1/75th the thickness of average aluminum foil, or an average human hair (both about 1.5 thousands of an inch, or 40 µm thick)
- 1/40th the size of a typical mammalian cell (20 µm diameter; a mammalian cell is roughly as thick as plastic wrap)
- about 1/25th the thickness of kitchen plastic wrap (12.5 µm, or 0.0005 inches)
- 1/2 the size of an E. Coli bacillus
- 120X the size of a (small-ish) protein (GFP, green fluorescent protein).
- 5,000 times the size (diameter) of an atom

While you do not need to know all those dimensions, you will need to know the following table. We will expect you to be able to state the wavelength or range (roughly) corresponding to a given color, as tabulated below:



Glass Starts Absorbing at ≤380 nm

Human Eye Peak Sensitivity ~550 nm (for Light-Adapted, a.k.a. Scotopic, Vision)

Silicon Stops Absorbing (Si Detectors Stop Working) at ≥1100 nm

Color	Wavelength Interval
Red	~ 700 - 635 nm
Orange	~ 635 - 590 nm
Yellow	~ 590 - 560 nm
Green	~ 560 - 490 nm
Blue	~ 490 - 450 nm
Violet	~ 450 - 400 nm

THORLABS

Appendix B: Zernike's Phase Contrast

Optical Microscopy Course

Appendix B: Zernike's Phase Contrast

Introduction

Many biological specimens – cells, small organisms – are primarily water and thus primarily transparent, and consequently hard to see. One can dye ("stain") the cells in order to see them, but the staining process typically kills the cells, and thus is not practical if one wants to examine living materials (e.g. to study the life-cycle of cells or organisms).

Darkfield and, more importantly, phase contrast allow observation of transparent materials *without* any staining steps, and for this reason phase contrast (and to a lesser extent darkfield) is used in nearly every biology research lab in the world, and many medical labs. Both make transparent samples visible, but phase contrast (as opposed to darkfield contrast) presents an image where the degree of shading is linearly proportional to the optical thickness of the sample. As a result, phase images look closer to what one might expect from a stained sample, making them easier to interpret.

For this reason, the Dutch physicist Frits Zernike received the 1953 Nobel Prize for developing phase contrast. Zernike is a name worth knowing in optics: he also developed the Zernike polynomials, a set of functions that allow easier analysis of optical systems[1], and helped develop the theory of optical coherence (related to our discussions of coherent and incoherent illumination) with the famous van Cittert-Zernike theorem.[2] [For readability, notes for this Appendix are placed at the end of the document.]

Zernike has a beautiful essay explaining phase and phase contrast, and the famous microscopist Shinya Inoué also did a nice presentation of Zernike's work; see the references at the end of this document for additional (and fun) reading. The discussion below owes much to both of their presentations, which are paraphrased below and placed in the context of this class.

Zernike's Explanation of Phase Contrast

Zernike notes that the wave nature of light is important to take into account in microscopy because 1) the sample features being observed are typically near the size of a wavelength of light, and 2) the transmitted light does not change much on passing through the (typically thin, and nearly clear) sample. The combination of these two things makes the effects of diffraction in a microscope particularly prominent.

To see how this affects contrast, consider a set of experiments you can do with your microscope. In all cases, close down the aperture stop as small as it will easily go. The result is plane-wave illumination along the axis, and the light will then be collected by the objective and brought to a focus at a point at the center of the back focal plane of the objective. We can use this fact to selectively block only the light undeviated by the sample. Let's clarify this with figures; to start with, imagine having no sample present:

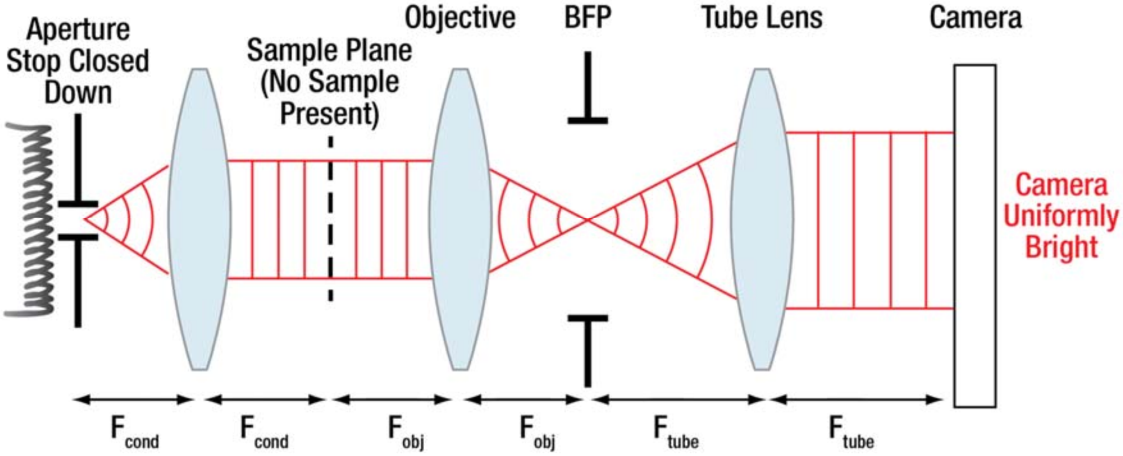


Figure 1: With no sample present, plane wave illumination at the sample is collected by the objective and focused down to a point in the objective back focal plane, and then converted back into a plane wave by the tube lens, creating uniform illumination – brightness – again at the camera.

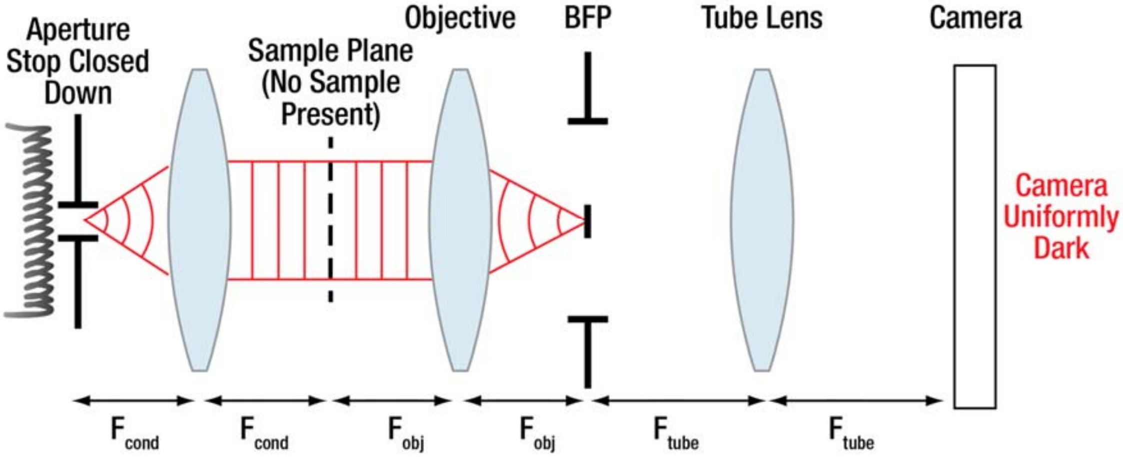


Figure 2: Introducing a small light-block at the center of the objective back focal plane. Illumination is fully blocked, resulting in darkness – no image at all – at the camera.

With no sample present, all the light will be collected and spread evenly across the camera, resulting in a uniformly bright image (Figure 1). If we then introduce a small light-blocking stop at the center of the objective BFP, we block all of the light, resulting in a camera image that is uniformly dark (Figure 2).

Zernike then describes several sets of experiments using variations on this configuration. For the first set of experiments, he takes a sample of small *light-absorbing* objects – tiny dust particles, or (for instance) cheek epithelial cells stained with absorbing blue dye.

Using this sample, Zernike does three experiments:

a. With the BFP iris closed down to block all but the very center of the BFP. This allows essentially only the illumination through, and so the camera is, as before, **uniformly illuminated** (Figure 3A).

b. Open up the BFP iris, but insert a light-blocking stop in the center. Before (with no sample) we saw that the image was then dark. Now, however, the camera has **bright features against a dark background** (i.e., a darkfield image), with the bright features corresponding to absorbing features in the sample (Figure 3B).

   ➔ The fact that we can see anything at all is proof that some light is being diffracted from the sample.

c. With the BFP iris open, and no stop in the center to block the direct illumination light, then the camera registers an image of the sample where the **features are dark against a bright background** (Figure 3C).

   ➔ The only way the image features could go from being bright when there is no (direct) illumination background present to being dark when the diffracted light combines with the direct (illumination) light is if **the direct and diffracted light interfere destructively.**

It is not that surprising that the interference would turn out to be destructive – one can view the absorption of light by the sample as subtracting a portion from the illumination wavefront. Zernike phrases this particularly well, saying:

> "Each (absorbing) particle obliterates a small part of the wave; it can be said to make a hole in the wave front. The diffraction effect of these holes is found as follows: instead of subtracting a small patch of wave front, each particle may be said to add a negative piece – that is, a wave of opposite phase. In this way the transmitted light is seen to consist of two parts that behave differently: (1) the unchanged incident wave, which will be called the *direct light*; (2) the wavelets starting from the black particles, to be called the *diffracted wave*." [quoted from Zernike, Reference 2.]

Mathematically one can write this interference as the combination of the illumination electric field and the electric field of the light diffracted by the sample. Hence, the field at the camera is:

Equation 1: $$\boldsymbol{E_{Image}(x,y,t) = E_{Direct}(x,y)\cos(\omega t) - E_{Diffracted}(x,y)\cos(\omega t)}$$

Since the direct illumination is uniform, it does not depend on x and y. Also, cos(ωt + 180°) = – cos(ωt), so:

Equation 2: $$\boldsymbol{E_{Image}(x,y,t) = E_{Direct}\cos(\omega t) + E_{Diffracted}(x,y)\cos(\omega t + 180^\circ)}$$

This gives the important result that one can view the absorption of light as equivalent to the generation of diffracted light describing the sample, but with a half-wavelength (i.e., 180°) phase shift compared to the illumination. This allows the diffracted light to always decrease the illumination at the camera, making a dark image against a light background.[3]

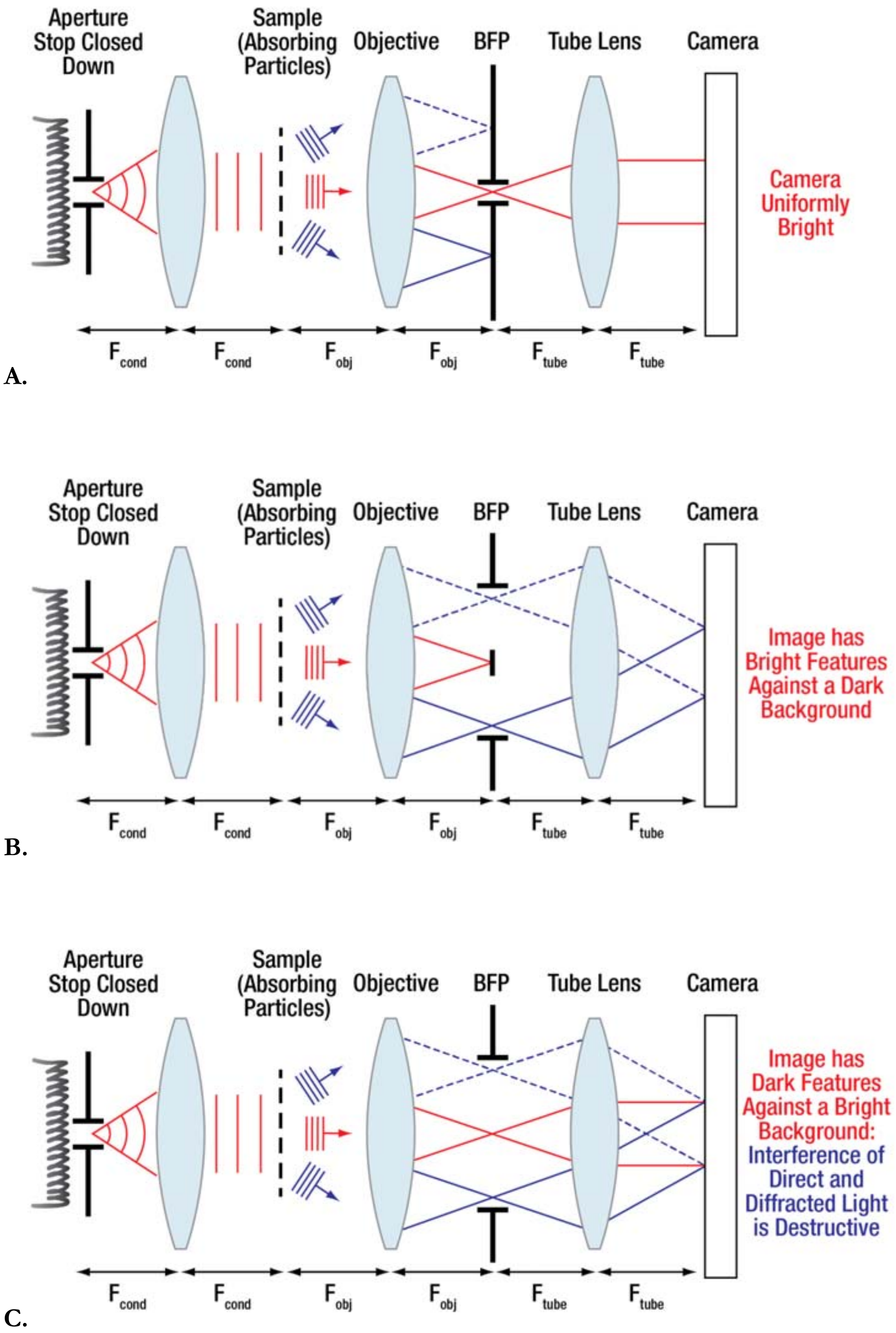

Aperture Stop Closed Down
Sample (Absorbing Particles)
Objective
BFP
Tube Lens
Camera
Camera Uniformly Bright
F_{cond}
F_{cond}
F_{obj}
F_{obj}
F_{tube}
F_{tube}
A.
Aperture Stop Closed Down
Sample (Absorbing Particles)
Objective
BFP
Tube Lens
Camera
Image has Bright Features Against a Dark Background
F_{cond}
F_{cond}
F_{obj}
F_{obj}
F_{tube}
F_{tube}
B.
Aperture Stop Closed Down
Sample (Absorbing Particles)
Objective
BFP
Tube Lens
Camera
Image has Dark Features Against a Bright Background: Interference of Direct and Diffracted Light is Destructive
F_{cond}
F_{cond}
F_{obj}
F_{obj}
F_{tube}
F_{tube}
C.

Figure 3: **A)** Allowing only direct illumination light through results in uniformly bright image. **B)** Allowing diffracted light through but blocking the direct light (illumination) results in bright features against a dark background, indicating that in fact some light is diffracted by the sample at angles other than along the axis. **C)** Allowing *all* the light through results in dark image features against a bright background, indicating that the interference of the diffracted and direct light at the camera is *destructive*.

Now Zernike shifts to consideration of a transparent sample – for example, biological cells in water (or, more accurately, salty aqueous buffer). Using such a sample, he again does the same three experiments, though here we will vary the order:

In his second set of experiments, Zernike:

a. Closes down the BFP iris closed down to block all but the very center of the BFP. This allows essentially only the illumination through, and he finds that the camera is, as before, **uniformly illuminated**.

b. Opens the BFP iris wide. The camera again captures a uniformly bright image – **there are no image features at all**.

➔ This makes sense – after all, the sample is transparent, so one does not expect to see anything.

Aside: if you are trying this yourself, it is critical to actually be in perfect focus – if you are slightly out of focus, then you will start to be able to see features even in a transparent sample. The reason for this is discussed nicely, and very readably, by Zernike in Reference 2.

c. Opens the BFP iris wide, but also inserts a light-blocking stop in the center. As above, the camera shows **bright features against a dark background** (i.e., a darkfield image), with the bright features corresponding to (otherwise invisible!) features in the sample. You can see the cells!

➔ This indicates that *the transparent sample* also diffracts light!

These experiments beg the question of why one cannot see the transparent cells in brightfield, as one can with the absorbing sample. On the one hand, it makes perfect sense – they absorb no light, so why would you expect to see them? On the other hand, they clearly diffract light – which is why you can see them in darkfield – so why *can't* you see them in brightfield?

Zernike hypothesized that the reason for this is that the diffracted light from a transparent sample is 90° – a quarter wavelength – out of phase with the direct (illumination) light. This was an educated guess, based in part on the fact that a phase shift of 90° results in no interference *on average*.[4]

In his discussion, Inoué notes that there is a convenient feature of "the paths taken by the waves diffracted by the transparent specimen and those of the background illumination. The diffracted wave passes the full aperture of the objective lens, while the direct (illumination) wave is focused into only a smaller region, the conjugate of the aperture stop. The two waves do not appreciably overlap, and they occupy distinct parts of the back focal plane of the objective lens. This separation of the two waves in space allows them to be manipulated separately at the objective rear focal planes." [Paraphrased from Inoué, Reference 1.]

Zernike takes advantage of exactly this in order to evaluate the hypothesis about the 90° phase shift for diffracted light. In his last experiment he describes involves placing a small plate (called a phase mask) in the back of the objective; the plate has, in the center, a small transparent dot that shifts the phase of the light going through it by 90°. This plate introduces the phase shift to essentially only the direct light, the light which has not been diffracted by the sample.[5]

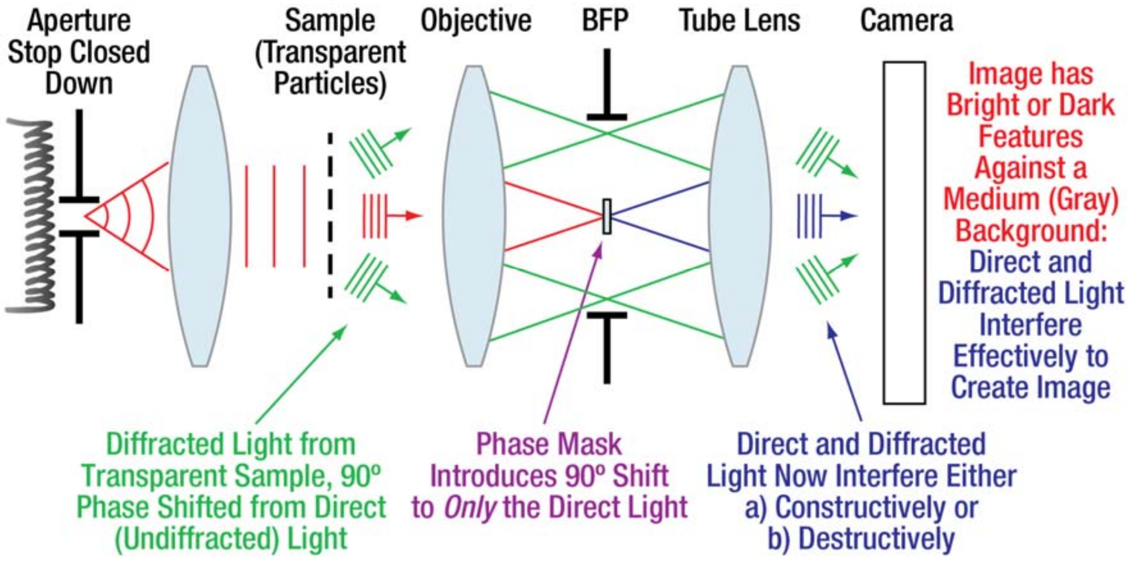

Aperture Stop Closed Down
Sample (Transparent Particles)
Objective
BFP
Tube Lens
Camera
Image has Bright or Dark Features Against a Medium (Gray) Background: Direct and Diffracted Light Interfere Effectively to Create Image
Diffracted Light from Transparent Sample, 90° Phase Shifted from Direct (Undiffracted) Light
Phase Mask Introduces 90° Shift to *Only* the Direct Light
Direct and Diffracted Light Now Interfere Either a) Constructively or b) Destructively

Figure 4: Phase contrast. For a transparent sample, the diffracted light is only 90° out of phase with the direct light, and so they will not interfere destructively – in fact, when averaged over time the direct and diffracted light will not interfere at all, and so no image features will be visible. Placing a 90° phase shifter in the center of the objective back focal plane (where it will affect only the direct light) results in either a net 0° or 180° phase shift between the direct and diffracted light, and thus effective (constructive or destructive) interference, and consequently visible image features.[6]

When one does this, one sees on the camera that the sample no longer appears transparent – the features are either brighter or darker than the overall background illumination, depending on whether the sample was optically thinner (shorter path length, or lower index) in a certain region, or optically thicker (longer path, or higher index). The reason for this is that the illumination light has now been shifted such that it is no longer 90° out of phase with the diffracted light; instead, it will be either 0° or 180° depending on the sign of the phase shift introduced and the index variations in the sample. The critical thing is that both of these give interference on average – either constructive (for 0°) or destructive (for 180°).

Even better, if one works through the math (see Note 4 at the end of this Appendix) one sees that the image intensity variation is *linear* in the variation in the sample thickness[7], making the images fairly natural to interpret (it looks a lot like a brightfield image) and easier to analyze quantitatively (though truly quantitative phase microscopy is done fairly rarely).

The phase plate is essentially just a piece of glass that is usually built into in the objective back focal plane[8] and that is slightly thinner (or thicker) in the region where the phase shift is desired, such that light going through that area is shifted by ±90°. To increase image contrast, the phase mask usually also blocks about 90% of the light, known as apodization (described in the Lab 8 Course Notes). To understand why, remember that in a transparent sample where the index does not vary much – e.g., biological cells in water – very little light is scattered. As a result, the direct illumination is vastly brighter than the small changes that constitute the image. Reducing the intensity of the direct illumination light (*after* it has already passed through the sample; doing so beforehand also reduces the diffracted light by the same amount, for no net gain) results in a larger percent change in brightness – i.e., higher contrast – when the diffracted and direct light interfere at the camera.[9] This light blocking in the phase plate – which is built into the objective – makes phase contrast objectives a very bad choice when trying to do fluorescence microscopy, where one usually has little light from the sample and cannot afford to waste any of it.

As described in the sets of experiments above, the small aperture stop opening required to get plane-wave (on-axis) illumination results in a very small condenser NA. This makes our examples conceptually simpler, but since resolution is given by $\delta x = 1.22\ \lambda\ /\ (NA_{obj} + NA_{cond})$, it would result in a significant loss of resolution as well as creating fringing (coherence effects). To get around this problem, Zernike used an *annulus* for the phase plate instead of a central dot; the illumination then requires a matching *ring* at the condenser aperture stop, such that the ring is imaged onto the phase annulus in the objective back focal plane (see Fig. 13 of the Lab 8 Course Notes). The annulus, due to its larger area, also allows much more light through than a small hole (closed aperture stop) would, and thus helps make weakly diffracting (thin/small) samples more visible. Realistically the ring and phase annulus cannot have zero width, since light needs to get through, and there are also alignment tolerances. This results in some image artifacts (the "halos" typical of phase contrast), since low-spatial-frequency light, which diffracts at small angles, will pass through the phase plate along with the direct illumination and thus there is no contrast for low spatial frequencies (i.e., the low-resolution portion of the image will have no contrast).

The critical elements of phase contrast are thus:

1. An annulus (usually a plate with a clear ring area) placed in the condenser aperture stop;
2. A phase ring producing a ± 90° phase shift, placed in the objective back focal plane, and matching and aligned with the image of the aperture stop, such that all of the direct illumination light has its phase shifted;
3. Limitation to "optically thin" samples: it is worth remembering that the linear relationship between phase shift in the sample (proportional to the thickness; technically speaking the "optical thickness") and the intensity variations in the image only holds for thin (so called "weak") phase samples, where the total phase shift is << 1.

Notes:

1. The Zernike polynomials are a set of complete, orthogonal functions, much like the sines and cosines, Bessel functions, etc. Their completeness allows any other function to be written as a sum of different Zernike polynomials, and their orthogonality makes finding the correct sum (relatively) easy to do – analogous to finding a Fourier series. The advantage of writing a function (say, the distribution of light in an image) in terms of the Zernike functions is that the different Zernike polynomials correspond cleanly to different optical aberrations (like defocus, spherical aberration, coma, etc.). Thus, the Zernike functions are heavily used in optical design and analysis.
2. The van-Cittert-Zernike theorem involves concepts beyond the scope of this course, but for our purposes the principle result is that the area over which illumination is coherent – i.e., will interfere on average with light from nearby – has a radius $\delta r = 0.16 \frac{\lambda}{NA_{illumination}}$. When we close down the condenser aperture stop to get spatially "coherent" illumination, the smallest the iris in the kit goes is a radius of ~ 0.5 mm, and so with focal length 50 mm the $NA_{condenser}$ ~ r/f ~ 0.01. For green light with λ ~ 0.5 μm, the illumination is thus coherent over *diameters* ~ 8 μm at the sample, coherent, but not terribly so; note that the LEDs used in Lab 3 have an effectively much smaller NA (they have maybe ~ 100 μm diameter emitting surface, and you placed them ~ 25 cm away, for an $NA_{illumination}$ ~ 2 x 10^{-4}), and the diffraction fringes in the LED-illuminated images were much more visible.

3. Advanced note: one can take this further by noting that the image the camera detects is due to the average light *intensity*, which is given by $I = \langle |E|^2 \rangle$, where the brackets < > indicate the time average. This would give, from Equation 2,

Equation 3: $$\boldsymbol{I \propto E^2_{Direct} - E_{Direct}E_{Diffracted} + E^2_{Diffracted}}$$

since the time average of $\cos^2(\omega t) = ½$. Usually the diffracted light has much smaller amplitude than the illumination (direct) light, so one can ignore the rightmost term and the image intensity becomes a constant, E^2_{Direct} (essentially the original illumination intensity), minus a term that is linearly proportional to the diffracted light – which is in turn proportional to the sample absorption.

4. There are (at least) two reasons one might suspect the phase shift would be 90°. The first is that light can have different speeds, and thus wavelengths, in different parts of a transparent sample – e.g. if the sample has a higher index of refraction (remember, $\lambda = \lambda_{vacuum} / n$) in different locations. If the index is a function of x and y, i.e. $n = n_0 + \Delta n(x,y)$, then for a sample of thickness Δz a wave will be proportional to

$$\boldsymbol{\cos\left[\frac{2\pi}{\lambda}\Delta z - \omega t\right] = \cos\left[\frac{2\pi n_0}{\lambda_{vac}}\Delta z - \omega t + \frac{2\pi\,\Delta n(x,y)}{\lambda_{vac}}\Delta z\right] = \cos[k_0\Delta z - \omega t + \Delta\phi(x,y)]}$$

where the phase change going through the sample is $\Delta\phi(x,y) = \frac{2\pi\,\Delta n(x,y)}{\lambda_{vac}}\Delta z$. Assuming a thin sample, i.e. one with only small variations in phase, the Taylor series expansion for the cosine gives

Equation 4:

$$\boldsymbol{\cos[k_0\Delta z - \omega t + \Delta\phi(x,y)] \cong \cos[k_0\Delta z - \omega t] - \Delta\phi(x,y) * \sin[k_0\Delta z - \omega t] \cdots}$$

➔ In other words, the differential change in the cosine due to interaction with the sample will be proportional to a sine, which has a relative 90° phase shift from a cosine.

Aside: often the sinusoids are plotted versus time; then it is simpler to have ωt positive, and the equation becomes

Equation 5:

$$\boldsymbol{\cos[\omega t - k_0\Delta z - \Delta\phi(x,y)] \cong \cos[\omega t - k_0\Delta z] + \Delta\phi(x,y) * \sin[\omega t - k_0\Delta z] \cdots}$$

Since sin(ωt) = cos(ωt – 90°), it is said that the diffracted light has a 90° *phase delay*, i.e. a phase shift of –90°.

The second reason one might suspect a 90° phase shift is due to the fact that the intensity of the image in brightfield was uniform – there were no variations in intensity (see experiment (e) above). If in Equation 2 above we instead have

Equation 6:

$$\boldsymbol{E_{Image}(x,y,z,t) = E_{Direct}\cos(\omega t) + E_{Diffracted}(x,y)\cos(\omega - 90°)}$$

Then the time-averaged intensity will then be (see Note 3, above)

Equation 7: $$\boldsymbol{I(x,y,z) \propto E^2_{Direct} + E^2_{Diffracted} \cong I_{Direct},}$$

i.e. the image will show nothing but uniform intensity since the time average $\langle\sin^2(\omega t)\rangle = \langle\cos^2(\omega t)\rangle = ½$, while $\langle\cos(\omega t)\cos(\omega t \pm 90°)\rangle = \langle\cos(\omega t)\sin(\omega t)\rangle = 0$. Note that $I_{Direct} \propto E^2_{Direct}$, and then again we assume that the diffracted light, $E_{Diffracted}$, is much smaller than the direct light such that the squared term can be ignored.

➔ In other words, a 90° phase shift causes the main intensity variation in the interference of the diffracted and direct light to average to zero, giving the lack of image variation that one in fact sees with a (in focus) transparent sample.

5. If you would like to make one of your own, it is not that hard – see instructions in Appendix C of these Course Notes.

6. Advanced note: we have been purposefully vague about whether the shift is to be plus or minus 90°. Depending on which is chosen, positive index variations (i.e., longer optical path lengths) in the sample, which result in –90° phase shift (a *phase delay*) of the diffracted light, will either result in bright features against the background (if the direct light is also shifted by –90°, such that it constructively interferes with the diffracted light) or dark features against the background (if the direct light is shifted by +90°, such that it destructively interferes with the diffracted light). Setting up the phase mask to produce dark features is most common, and is known as "positive" phase contrast since the phase shift given to the direct light is positive; the converse approach is known as "negative" phase contrast.

7. Technically, the intensity variation is linear in the sample phase variation, which is, for small index variations, proportional to the change in the optical path length in the sample. In simple cases (e.g. for a rough glass surface), the optical path length in the sample is proportional to actual thickness in terms of distance.

8. For the pro user, at least one manufacturer (Nikon) makes a microscope where the phase plate is actually at a separate plane that is conjugate to the objective BFP, allowing more versatility in use. E.g., when doing fluorescence one typically does not want a phase objective, because they block a lot of light; this set-up allows for fluorescence microscopy through a normal, unobstructed objective, and then, at some separate wavelength or time, and without changing objectives, one can also do phase contrast via the optical path that goes through the external phase plate.

9. This issue of contrast is discussed in more depth in the Lab 8 Course Notes. Ideally one would like the direct and diffracted light to have approximately equal intensities; in practice it is good to make sure the direct light is always *at least* as intense as the diffracted light (and it is good to have some extra margin just to be sure). This prevents the field from dropping below zero, which – since the intensity is the square of the field – would start to look brighter again, reversing the contrast.

References:

1. Shinya Inoué, Video Microscopy, ISBN 0-306-42120-8 (the 1st edition), pp. 119-122, has an excellent (and excellently phrased) discussion of this same material, based on Zernike's (see reference below).
   a. Note: the 1st edition of Video Microscopy has a lot of content relevant to this course that got dropped from the subsequent 2nd edition for space reasons. The 1st edition is worth tracking down if you can find it, and parts of Chapter 5: Microscope Image Formation are especially relevant.
2. F. Zernike, *The Wave Theory of Microscopic Image Formation*, Appendix K in Strong, Concepts of Classical Optics, WH Freeman, 1958. Now available free online and also reprinted by Dover, ISBN 0486432629. Zernike's exposition is extremely readable and (hardly surprisingly) very well done.

THORLABS

Appendix C: Additional Projects

Optical Microscopy Course

Appendix C: Additional Projects

One of the main goals of this course is to teach you the practical skills necessary to carry out your own projects involving optics and light microscopy. Depending on the format of your course (and whether you are on a semester or quarter system), there may be time for you to undertake some additional exploration.

Some of the projects below require additional equipment; naturally, choose projects for which you have the required hardware. As an example, computational exploration (e.g. using Python or MATLAB) naturally requires access to appropriate programming languages.

To get you thinking, some possible topics are listed below.

➔ ***Many other projects are also possible, so definitely do not feel limited by this list! If you are interested in sharing your own labs or ideas contact techsupport@thorlabs.com .***

Project Ideas

- ***Rheinberg Illumination (Color Contrast)***: A technique for generating color contrast for unstained samples; uses masks of different colors in the aperture stop. Images are often quite beautiful; an excellent project in terms of exploring and understanding Abbe theory.
    - For this project you need to explain where the color filters go, how one determines the appropriate diameters for the central and annular portion of the filter, and what the relation is between this technique and brightfield and between this technique and darkfield.
    - You also need to find appropriate color combinations and samples to show the advantage of the technique, and provide images in your report.
    - Colored plastic sheet (e.g. inexpensive "gel" filters for theater lights) works well.
    - Resources:
        - Optical Microscopy by M.W. Davidson and M. Abramowitz, see the *Reference Links* tab at www.thorlabs.com/OMC.
        - M. Abromowitz, "Rheinberg Illumination," American Laboratory, 1983, v15, #4, p 38.
- ***Polarization Microscopy***: Use crossed polarizers to image different samples; explain theory and find some interesting samples to investigate. Note: this project can also be combined with brightfield epi-illumination (see below).
    - For this project, you need to explain what polarization is, what birefringence is, and how it can affect the polarization of light going through a sample.
    - The kit contains polarizers you can use.
        - A good first step is to figure out how they work:
            - What is the axis of the polarizers (and how could you figure it out, e.g. using Brewster's angle)?
            - Plot total transmission vs. relative angle between the two polarizers. Does it fit the theory? What is the total "extinction ratio" of the crossed polarizers (and how did you determine it)?
    - Samples for polarization microscopy are ***not*** included with the kit. Understanding what samples might be good, and making some, is typically part of the project; hair, synthetic

fibers, paper (e.g. Kimwipe® or lens paper in water between coverslips) are good places to start. Additional options include:

        - Chemical melts: place a few grains of some chemical on a slide, melt them (e.g. using a hotplate), and smear with a coverslip (or put a coverslip on and press down – it is hot; do not use your finger! – to make this sample). Per Andrew Davidson's website, vitamins, moth balls / moth crystals, aspirin and other painkillers, sugar, etc. all can make nice samples this way.
        - One can also obtain thin bone sections (e.g. from Ward's or Carolina Biosciences) or mineral sections (e.g. asbestos or other materials, already fixed on slides from McCrone, www.McCrone.com).
    - Setting up a rotating stage can be done, but we typically just rotate the polarizers instead – less ideal, but easier to set up.
    - Resources:
        - Optics textbooks, e.g. Hecht, Optics, 5th ed., ISBN 0133977226.
        - Essentials of Polarized Light Microscopy, by Delly (of the McCrone Institute). Excellent coverage, with images; see the *Reference Links* tab at www.thorlabs.com/OMC.
        - S. Bradbury and P.J. Evennett, Contrast Techniques in Light Microscopy, by BIOS Scientific Publishers Ltd, 1996, ISBN 1-85996-085-5
        - Polarized Light Microscopy Techniques; see the *Reference Links* tab at www.thorlabs.com/OMC.

- ***Color Imaging and Color Balance:*** Image a color sample using your color camera with white light illumination, and then again using three separate exposures with your mono camera with three different (red, green, blue) wavelength illuminations (e.g. using color filters). Figure out how to combine your three mono images to get a single good color image – this involves some linear algebra – and compare your image taken with the color camera with your own synthesized color image.
    - Doing this simply – multiplying each of your color images by different constants, and then combining them – can be done fairly directly using NIH ImageJ. However, getting good color balance usually requires a more sophisticated calculation using a matrix, since – assuming the illumination color ranges overlap, as usual for cheap plastic filters – the information for different sample absorbances will show up in *each* of the three single-color images, thus resulting in the need for a 3x3 array of coefficients in order to convert your images to a single color RGB image. Manipulating your images in this way will likely require programming (e.g. in Python or MATLAB).
        - Learning to open and manipulate images in Python or MATLAB is an excellent skill to learn.
    - You could explore what happens to your color balance if you remove the IR filter in the illumination path – does the balance change?
- ***Phase Contrast***: Make custom phase masks using soot/wax (method described in section 2, below), and explore phase contrast further, e.g. by duplicating the Zernike example of central phase, and comparing to central darkfield using the zero order mask in the kit. Exploring different phase samples could also be rewarding. This is actually a very nice lab to do – and is actually the original way we did this in the class, before adding the phase contrast objective with annular phase plate.
    - Make a central-phase ("zero order phase") mask, as detailed in this appendix (section 2, below).

    - Image in phase contrast. Compare your images using central phase (with your mask) to images made using the fully-black "zero order mask" used for darkfield in Lab 6. What are the differences?
    - Compare your phase imaging to imaging of a similar (or identical) sample using the phase objective included with the kit.

- ***Epi-Illumination:*** Illuminating through the microscope objective (instead of through the sample) is a standard way to image opaque samples (like metal, computer chips, etc.), and is nearly always used in fluorescence microscopy to reduce background from the illumination.
    - For brightfield this can be done using, e.g., the beamsplitter already included in the kit and putting the lamp and collector on the 90° dogleg rail that normally holds the color (BFP) camera. One can set this up, then explore opaque samples (like a silicon chip of some sort, or a metal surface, etc.)
        - One can further set up polarization with brightfield epi-illumination, which can provide additional contrast for metal and other surfaces. If doing this, using the color camera may make for more aesthetically pleasing images (polarized light images can be quite colorful).
    - (Requires additional equipment) For fluorescence one would normally do this using a dichroic filter (not included in the kit) that reflects the excitation wavelength and transmits the emission wavelengths. Setting up this type of illumination – possibly in conjunction with a high-power LED light source (also not included) – is an excellent project.
        - Even without the additional parts, this can be an excellent theory project. Items to consider include choosing the appropriate collector NA, how to incorporate an adjustable field- and aperture-stop separate from the objective BFP (all within reasonable length limitations for the illumination system), and figuring out how to optimize for maximum illumination intensity at the sample given optical invariant/*etendue* limitations.

- ***High Dynamic Range (HDR) Imaging:*** Explore/understand what this is, and use it for something – possibly acquisition of high-bit-depth images of diatoms (in brightfield, darkfield, or phase contrast), or (harder) a high-bit-depth (HDR) PSF.
    - HDR imaging involves taking multiple images at different exposures, then combining them to create a final image that has a larger total range of grey levels (i.e., larger dynamic range).
    - Doing this requires programming (e.g. in Python or MATLAB)
        - Registering the images (making sure they overlap properly even if the sample got bumped) can be important; many software packages (e.g. MATLAB) have options for registering images prior to doing other calculations.

- ***Aberrations***: Investigate various aberrations using your system – spherical aberration, coma, distortion, field curvature.
    - This may require substantial rebuilding of your system, since the microscope setup is somewhat optimized to minimize these aberrations.
    - Resource: Hecht, Optics, 5th ed., ISBN 0133977226, section on Aberrations – note especially discussion of stop location on distortion.

- ***Point-Spread Function***: Generate a substantial 3-D model of the PSF for your system, based on images you take. This will require careful imaging and post processing of your images. Show PSF's demonstrating good imaging, and spherical aberration (the system already has a good bit of spherical aberration). Explain what happens to the PSF based on a ray diagram showing spherical aberration for a spherical lens.

- It can be helpful to work at lower NA, such that the DoF is large, allowing a good number of separate steps along the z-axis with the sample stage (remember, the micrometer only has 10 µm resolution, so you need to take steps at least that big).
- This lab requires facility with programming (e.g. Python or MATLAB) to register the individual images before stacking them for 3-D viewing (we register in MATLAB, then view as a stack in ImageJ; one could do all of this in MATLAB or Python, however).
- (Requires additional equipment): Use of a small (e.g. 1 µm) pinhole as the sample, and a high-power LED for illumination helps with image acquisition. Neither of these is included with the kit.

- ***Learning to Use a Research Microscope:*** If you have access to a research microscope, it is an excellent project to learn to set it up for Köhler illumination, as well as phase contrast (if the microscope has phase).
    - Absolutely worth doing: follow the instructions in the microscope manual to set up Köhler, and also to properly align the eyepieces – setting the diopter adjustment to accommodate your eyes, and adjusting the eyepiece spacing to best suit the distance between your eyes. Properly doing this results in dramatic improvement of the image, like seeing a movie in 3D instead of in the usual 2D.
    - Learn to set up Köhler illumination. Tip: usually it is much easier to start with a low-magnification (5X or 10X) objective and focus on (find) the sample *before* shifting to a higher magnification objective.
        - Learn to use the Bertrand lens (to see the objective BFP), if your microscope has one. If not, the usual technique is to remove the eyepiece and just look into the tube in order to see the BFP (only do this with permission of the microscope owner).
    - (Only with permission of the microscope owner): learn to properly adjust the phase rings so they are aligned with the phase mask in the objective BFP.
    - ***Advanced:***
        - ***Graded-Field Microscopy***: Also known as "the poor scientist's DIC," one can implement a phase-contrast like technique that gives images very similar to those from DIC (Differential Interference Contrast). The technique is extremely easy to set up – using two business cards to block half the condenser and half the objective aperture. Worth trying if your microscope gives you access to the objective BFP and aperture stop.
            - Resource: J. Mertz, et al., "Graded field microscopy in white light," Optics Express, Vol. 14, No. 12, June 2006.
            - This can also be implemented directly using the kit rail system.
        - *"**Poor Scientist's Darkfield**"*: If the microscope is set up for phase contrast, then use of a regular (non-phase) low-NA objective and a high-NA phase ring (e.g. a 5X objective and a Phase 2 or Phase 3 ring) yields darkfield imaging without having to buy any extra parts. Worth trying if your microscope is set up for phase contrast and also has a lower NA objective you can try this with.
            - Explain how (and why) "poor scientist's darkfield" works.

Home-Made Phase Masks

Phase masks for phase contrast microscopy can be made extremely inexpensively, and as part of student projects. The method detailed below was used for making phase contrast objectives in commercial microscopes in the 1950s, and was developed by Maksymilian Pluta, who literally wrote the (multivolume set of) books on microscope design.

The basic method is this: moving a coverslip through a candle flame (really, a bit above the flame **where the soot is forming) allows formation of a thin** soot **layer on glass which both absorbs and provides a phase delay (e.g. the index for soot from a stearin**, **a type of wax**, **candle is $n = 2.32$ compared to** $n = 1$ in **air). Conveniently, the absorption of the soot and the phase shift are both related (by the thickness of the soot layer), and this has been measured and tabulated (**see **Pluta reference, below). Simply light the candle, pass a coverslip** just above **it (moving horizontally, with the coverslip parallel to the floor)** once or twice**, measure the change in transmission through the soot,** and **repeat until the transmission matches the thickness of soot which provides ¼ λ retardation** (see steps below)**. Then using a razor blade, scrape off all the soot except for a 1.0 to 1.5 mm dot in the center (sufficient to cover the image of the aperture stop iris when that iris is fully closed), and mount the coverslip in a short SM1** lens **tube so it can be screwed onto the back of your homebuilt objective.**

The darkness of the soot provides some apodization in addition to the phase shift due to the wax; it may be worth emphasizing to students that there is no inherent optical link between the phase shift and the apodization (lower transmission) – it is due in this case to the materials used, and will vary for other types of candles (soot), substances (e.g. thin metals layers), etc. It is possible in principle to have absorption without phase shift and phase shift without absorption.

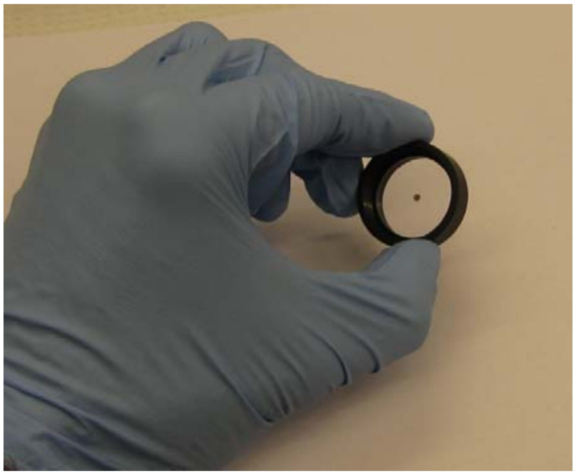


Figure 1: Home-Made Phase Mask: 1.25 mm diameter stearin candle soot spot, with transmission ~10%, on a coverslip mounted in an SM1 tube.

Procedure:

1. Materials:
   a. Some clean 25 mm circular coverslips
   b. A stearin candle (can be ordered from Swedish goods stores like IKEA)
   c. Razor blades
   d. SM1L03 or SM1L05 lens tube to mount the coverslips in later

2. Preparation
    a. Put a coverslip in an SM1 lens tube, screw the tube onto something, and then put a permanent marker against the center of the coverslip and rotate the SM1 lens tube so that you get a perfectly centered dot to use as a template for where to scrape off the soot on the other coverslips.
    b. Set up a microscope rig; after setting up proper Köhler, remove the sample. You will use this to test the transmission of the coverslips.
        i. Turn off the room lights – you need very low background, since the correct transmission of the soot is ~10% (24 counts if 240 is what you get in a clear area of the glass), so very low background helps.
        ii. Hold the clean coverslip in the sample area of the microscope rig, and optimize the camera so a clear area of the coverslip is close to, but not quite saturating (240 counts or so).
    c. Light the candle so it warms up
3. Wave a clean coverslip through the candle, above the main flame where the flame is tapering into soot.
    a. It helps to keep the coverslip normal to the flame axis, and to move your hand at a constant velocity and slowly enough that the wind from your motion does not perturb the candle flame while the coverslip is in it.
    b. Hold the now-sooty coverslip in the sample area of the microscope rig, and see what the transmittance is. It should be approximately 10 – 13% (say, 24 counts if the original was 240; be sure you are subtracting off any background.)
        i. If it is too low, get a new coverslip.
        ii. If it is too high, wave the coverslip just above the flame again, and re-measure.
        iii. If it is just right, move to the next step.
    c. Optional:
        i. Per Pluta (see reference below), moistening the soot on the coverslip with a drop of "absolute alcohol" and then letting it dry before the next steps toughens the soot layer and (after the alcohol has evaporated) makes the soot layer more robust and easier to process. This is an easy step to add if you happen have reagent-grade ethanol handy, as many biology labs do ("absolute alcohol" is just anhydrous alcohol, in this case 99%+ ethanol). Just drip a little on the coverslip, then wait for it to evaporate.
    d. Place the coverslip on a Kimwipe® (lint free paper) and scrape the soot off with a razorblade in a circular motion, leaving a dot in the center.
        i. Compare to your template; if it looks good, go to the next step.
    e. Use a folded-up Kimwipe® to wipe the coverslip down everywhere but the central dark spot (which must remain perfect), to get all remaining soot fragments off.
        i. A good diameter for use with this kit is 1.25 mm; between 1.0 and 1.5 mm should work. Larger dots make for easier alignment (superposition of the aperture stop image with the phase spot) but block more of the low-spatial-frequency light.
            1. Examining the variation in the image as a function of phase spot diameter could be an interesting student project; one would expect it to affect the "halos" one sees around the edges of objects in phase images.

    f. Put the coverslip in an SM1 lens tube and secure it with a retaining ring.

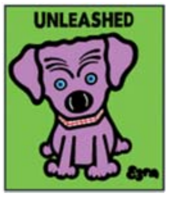
4. You are finished. Test the phase plate in your rig to see if it works for phase contrast.

Phase Contrast:

1. Get a decent phase sample (e.g. human cheek epithelial cells, as described in the Lab 8 Lab Notes).
2. Set up your rig for good imaging, in Köhler illumination, etc.
3. Screw your phase plate onto the back of your objective.
4. Open the BFP iris to several millimeters (say, 5.0 mm)
    a. The reason for only opening the objective BFP to 5 mm is to reduce aberrations in the objective; the better the imaging system the harder it is to see clear samples at the plane of best focus. Aberrations introduce phase shifts which can actually make the sample more visible if it is clear.
5. Close the aperture stop iris down to its minimum.
    a. Check in the BFP camera to see that the AS is closing down so that only the area covered by the soot is illuminated. If not, and there is filament visible around the edge of the soot spot, fix your alignment.
        i. If the AS and spot do not overlap perfectly, especially vertically (for which there are no mechanical adjustments), try rotating the phase mask SM1 tube; the coverslip may not be centered in the SM1 tube.
        ii. If you cannot center the soot spot make another phase plate and try again.
6. Once you close down the AS so that only the soot is illuminated, you should be in phase contrast. Open the Aperture Stop iris to get to (modified) brightfield; close it to go back to phase.

References:

1. "Zernike on phase contrast," from Inoue 1st ed. ISBN 0-306-42120-8, pp. 119-122.
2. Pluta, Maksymilian; "Stray-light problem in phase contrast microscopy or toward highly sensitive phase contrast devices: a review," Optical Engineering v. 32 n. 12, pp. 3199-3214 (December 1993).
    a. Note: the actual data plot appears to show that the correct transmittance of the soot layer for $\Delta\Phi = 0.25\lambda$ is 13% (±0.5% as best we can read it). Pluta does not seem to list the wavelength he uses, but it is surely in the visible (very likely the Fraunhofer/Hg line at 546 nm) and the masks work fine in our experience.
    b. Note: Stearin candles are available (in the US at least) at IKEA.



THORLABS

www.thorlabs.com