Principles of Phase Contrast (Electron) Microscopy

Marin van Heel

LNNano/CNPEM, Brazil

(marin.vanheel@gmail.com)

(© 1980 - 2021)

1. Introduction:

In phase-contrast light and electron microscopy, one exploits the wave properties of photons and electrons respectively. The principles of imaging with waves are the realm of "Fourier Optics". As a very first experiment ("Gedankenexperiment"), let us think back to days when we – as children – would focus the (parallel) light waves of the (far away) sun on a piece of paper in order to set it alight. What lesson did we learn from these early scientific experiments (Fig. 1)?

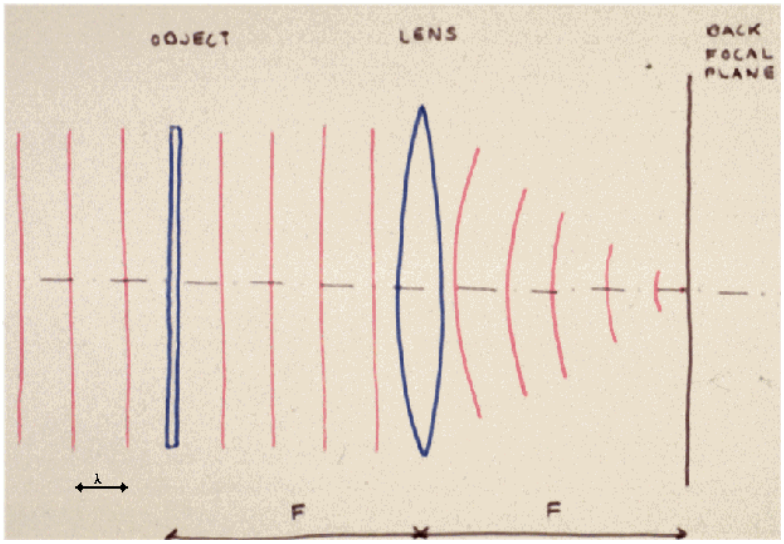
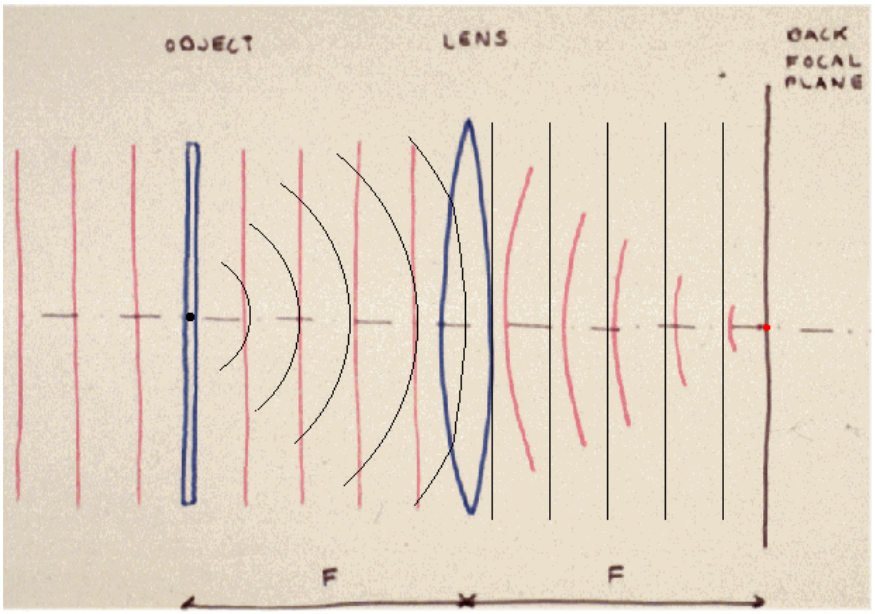


Figure 1: Plane parallel waves are focussed into a single point in the back focal plane of a positive lens (focal distance "F"). The plane waves are one wavelength ("λ") apart.

In Figure 1, the plane waves illuminate an object that is merely a flat sheet of glass and thus the waves exiting the object on right are plane waves indistinguishable from the incident waves. These plane waves are converted into convergent waves which reach a focus in the back focal plane of the lens. This optical system is thus capable of converting a "constant" plane wave in the *front* focal plane into a single point in the *back* focal plane of the lens.



<u>Figure 2</u>: A single point scatterer in the object leads to secondary ("scattered") concentric waves emerging from that point in the object. Since the object is placed in the front focal plane of this lens system, these scattered or "diffracted" secondary waves become plane waves in the back focal plane of the system.

In Fig. 1, the object is a transparent glass plate that essentially does not interact with the incident waves at all. In Fig. 2, a single secondary scatterer is included on the optical axis of the system. The secondary scatterer will become a source radiating concentric waves. Since this point scatterer is in the front focal plane of the lens, parallel waves will emerge from the back of the lens due to the presence of this point scatterer.

These two simple experiments illustrate how a point source in the front focal plane of a simple lens system leads to a plane wave in the back focal plane and vice versa, in the sense that plane waves emerging from the front focal plane, will focus into a single spot in the back focal plane of the system. This special reciprocity relationship between the front – and the back focal plane of a simple lens, is a "Fourier Transform" relationship that will be elaborated on in this document. The Fourier Transform is as fundamental in electron and light microscopy as it is in X-ray crystallography. It is so fundamental in Optics, that all what is discussed in this document falls under the science of "Fourier Optics".

This document provides an overview of the most relevant physical concepts in imaging in the light microscope and the transmission electron microscope. In particular, emphasis is placed on the basic concepts of phase contrast microscopy. Without seriously going into mathematical details, the Phase Contrast Transfer Function ("PhCTF", or short "CTF") is discussed. These concepts are of primary importance for optimising an electron microscope for the imaging of biological macromolecules.

2. Scattering and Diffraction by a periodic object

One of the very fundamental processes in imaging procedures is the interaction between the illuminating waves and the object. It is only after such interaction takes place that the radiation emerging from the object carries – possibly encoded – information about the object. It is the information about the object that we are after we will see that retrieving that information requires a good understanding of the basic physical principles of how the information is coded into the radiation and how to optimise our instrumentation in order to register the information.

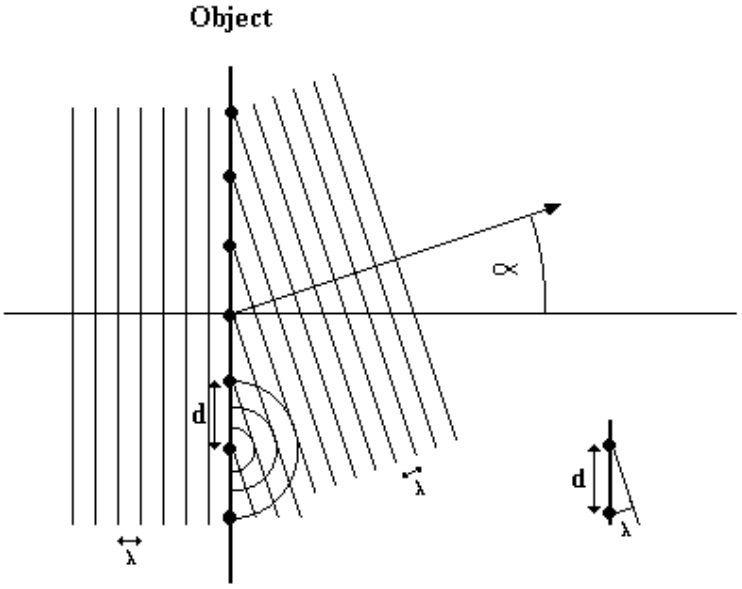


Figure 3: A regular array of single point scatterers in the object plane leads to secondary waves that reinforce each other in specific directions. Drawn in this diagram is the "+1" diffracted beam in which the concentric waves stemming from neighbouring point scatterers in the array are lagging by exactly one wavelength.

Let us, instead of the single point scatterer of Fig. 2, place an array of equidistant point scatterers in the object plane, with each point scatterer placed at a distance "**d**" from its nearest neighbour (Fig. 3). When this array is illuminated from the left with plane waves (wavelength "λ"), each of the point scatterers in the object will start emitting secondary radiation in concentric circles, as drawn in the illustration. In specific directions, the wave fronts from neighbouring point scatterers will be in synch with each others and will constructively interfere. In other directions, the waves emerging from different scatterers arrive at different times (with different "phases") and the radiation in these directions will disappear due to destructive interference. A direction in which there will be constructive interferences in illustrated in Fig. 3 and this "diffraction" direction is given by the formula:

$$\sin(\alpha) = \lambda / d \qquad (1)$$

Note that the smaller "d" is (the distance between the scattering spots in the denominator), that is, the smaller the period of the regular array of point scatterer, ("grating"), the higher the angle the diffracted wave makes with the optical axis of the system. Whereas, close to the object, the various diffracted beams are all intermixed (such as the +1, -1, and the 0 order beam, Fig. 4), at a sufficiently large distance from the object, all the different diffraction directions separate and we can observe its diffraction pattern.

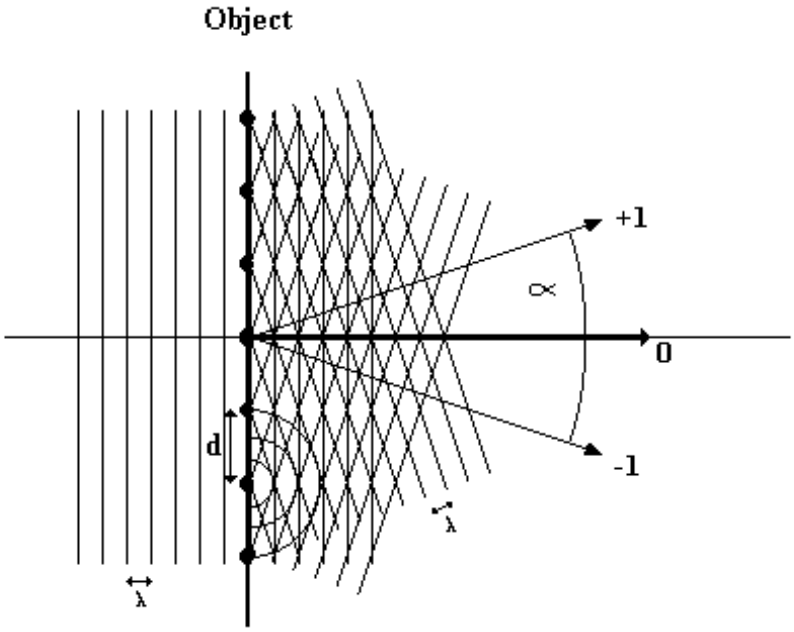
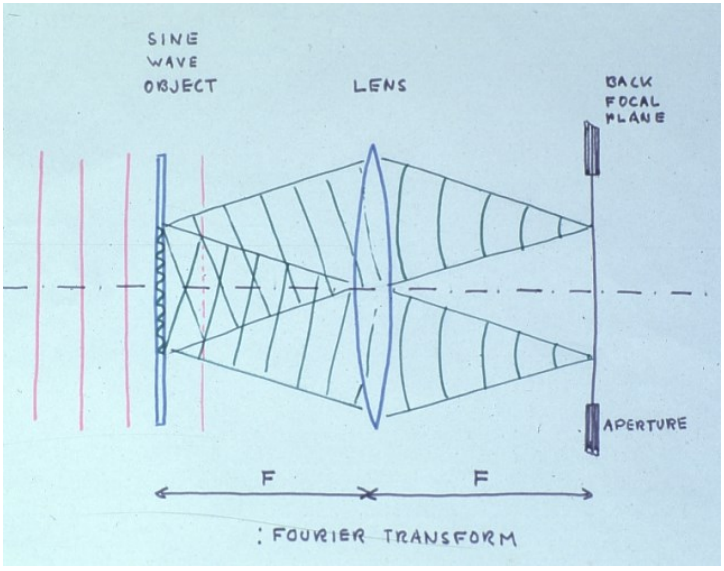


Figure 4: A regular array of single point scatterers in the object plane leads to secondary waves that reinforce each other in specific directions. Drawn in this diagram are the "+1" , the corresponding "-1" and the "0" order diffracted beams. Close behind the object all these waves are intermingled, but they separate with the increasing distance from the object. Eventually, a diffraction pattern of the object is obtained.

Thus, if we look in a plane, that is placed far enough away from the grating, we will see two spots (at least) due to the light being diffracted by each periodic grating, in full analogy to the diffraction of X-ray waves by a 3D protein crystal. Every "spatial frequency" (= periodic) component in the object thus corresponds to a certain radiation (diffraction) direction. If we replace the grating by a finer grating, the light will be diffracted at a higher angle. If we have objects containing more than a single spatial frequency, we can obtain highly complicated diffraction patterns. In particular, periodic objects such as 2D or 3D crystals lead to diffraction patterns that consist of an intricate raster of diffraction peaks. Special slides with periodic patterns, can also lead to spectacular diffraction patterns when illuminated by a laser. Such diffraction by a periodic object is well known from X-ray crystallography. Indeed, the formula for constructive interference (1) is essentially identical to Bragg's Law in X-ray crystallography (Sir William Henry Bragg, and his son Sir William Lawrence Bragg, shared the 1915 Physics Nobel prize, http://www.nobel.se/).

3. Imaging Diffraction patterns by a simple lens

The diffraction experiments described in the previous paragraphs rely on a sufficiently large ("infinite") distance behind the object for separation for separating the various plane waves exiting the object in different directions. Since it may be difficult to look at a plane that is placed very far away from the object (at infinity: the "Fraunhofer" diffraction plane) we can use the trick explained in the Introduction (Fig. 1) to bring "infinity" within reach. The trick is to place a lens behind the object and to look at the diffraction plane in the back focal plane of the lens (Fig. 5).



<u>Figure 5</u>: A periodic in the object plane leads to secondary waves that reinforce each other in specific directions. Drawn in this diagram are the "+1" and the corresponding "-1" order diffracted beams. (If the object is a pure sine wave, these are the only diffraction spots apart from the zero order beam; if the object consists of a regular array of point scatterers, higher order diffraction peaks may appear). Close behind the object all these waves are intermingled, but they separate with the increasing distance from the object. A lens placed at distance "F" from the object will focus plane waves in different points in the back focal plane, depending on the angle α that the directions of the plane waves make with respect to the optical axis.

In the setup of Fig. 5, every direction is focused in a point in the back focal plane of the lens. In other words, each spatial frequency in the object corresponds to a point in the back focal plane. Actually, each spatial frequency corresponds to *two* points in the back focal plane because there will be two diffraction maxima in the set up of Fig. 5 (like there were two maxima in Fig. 4). The simple lens system decomposes the wave front exiting the object in its sine/cosine-wave components and actually performs an exact "Fourier Transform" of the wave front. This Fourier Transform property of a simple lens system is presented here mainly in the context of explaining the principles of image formation; this property is actually exploited routinely in electron microscopy to judge the quality of electron micrographs. Electron micrographs are routinely placed in the "optical diffractometer" (Figure 6).

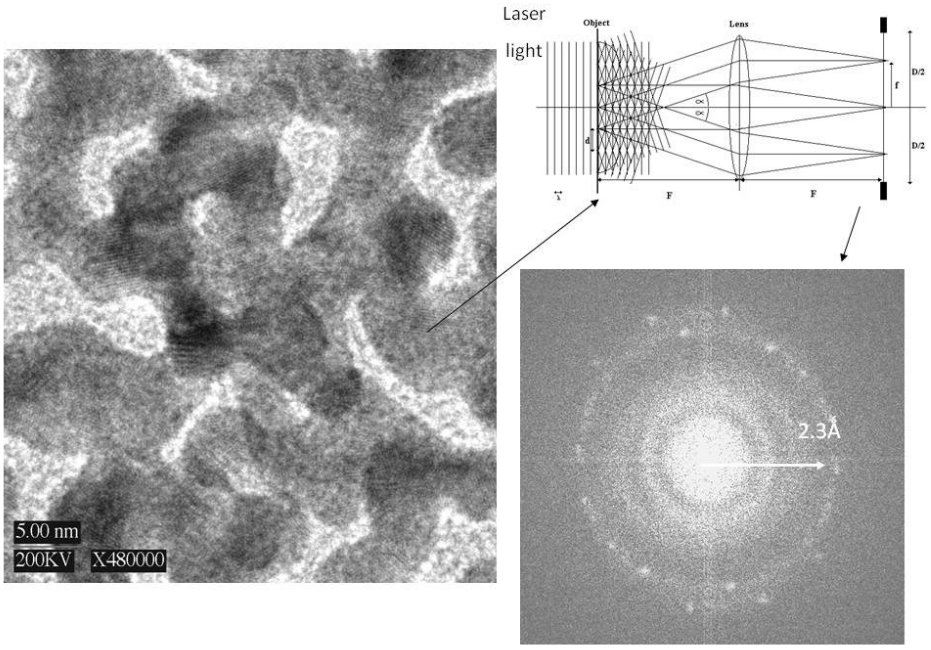


Figure 6: Use of the optical diffractometer (OD) to assess the quality of an electron micrograph. The micrograph (a negative) is placed in front of the diffraction lens and its diffraction pattern is observed in the back focal plane of the lens. The further out the diffraction pattern extends, the higher the level of resolution in the micrograph. This sample consists of small gold crystals on a carbon support film; its diffractogram indicates the presence of 2.3Å information. Inorganic samples can often tolerate high electron exposures: this image was obtained using a dose of ~10000el./Å^2. Biological samples, in contrast, may already show significant damage at exposure levels of ~10el./Å^2 even when cooled to liquid Nitrogen temperatures. Analogue optical diffractometers are no longer in use, now that the data collection is almost always performed digitally and now that digital Fourier Transforms can be performed in real time given the speed of modern computing equipment.

4. Imaging systems and their Transfer Functions

In X-ray crystallography, one collects data in the diffraction plane, that is, in Fourier space. In microscopy, in contrast, the data is collected in the image plane, that is, in real space. Irrespective of the actual implementation of the microscopical instrument, its imaging properties can be described formally using a double Fourier Transform device as depicted in Fig. 7. The system drawn here is a 1:1 magnification imaging system. In fact, by changing the ratio of the focal lengths of the two Fourier transform lenses (**F'/F**) the same formalism can be used to describe the imaging properties of instruments ranging from microscopes (**F'/F** >> 1), photographic cameras (**F'/F** ≈ 1), to astronomical telescopes (**F'/F** << 1).

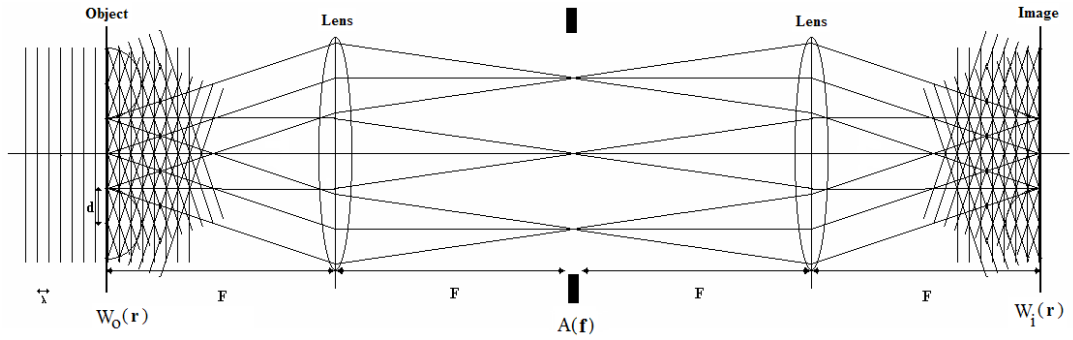


Figure 7. Imaging devices perform a "double" Fourier transform. This leaves space for manipulation of the data in Fourier space by modifications of the diffracted waves in the back-focal plane.

The great advantage of describing imaging systems by this formalism is that all of the properties of the "linear" imaging instrument can be dealt using a straightforward mathematical formalism. All of the properties of the imaging system can be described by a simple multiplication of the Fourier transform of the signal, in the back focal plane of the imaging device, by a "transfer function" characterising the instrument. The signal on the output side of the microscope can then be calculated by Fourier transforming that product back into real space. The formalism of using "Transfer Functions" in Fourier space to describe the effects of a component of an information chain is a standard formalism in Signal Processing that can be applied to all "Linear Systems". The formalism is used in 2D or 3D imaging but also for describing, say, the characteristics of a (1D) loudspeaker system in an audio amplification chain.

As an example of how this works, let us look at the resolution limits achievable by a conventional light microscope. In Fig 5,7-8, a limiting aperture is included in the back focal plane of the first lens. This aperture in a light-microscope objective lens may be a real aperture, but it may also be defined by the diameter of the front lens facing the object. In the Electron Microscope (EM) it normally is a physical aperture inserted into the electron beam in a aperture holder. For completeness, in the astronomical telescope it is given by the diameter of the telescope lens. We have seen above that the smaller the periodicity of the object "d", the higher the angle α the diffracted beam makes with the optical axis (the direction of the incoming waves).

5. Resolution limits of the (light) microscope

In Fig. 8, a limiting aperture is included in the back focal plane of the first lens. The wave falling onto the aperture is the Fourier Transform of the wave emerging from the object. It is subsequently "masked" by the opaque aperture that removes all information for which **f >> D/2.** Note from Fig. 8, that **$f \approx F \sin(\alpha)$**. We have seen previously that the general diffraction direction α, is related to the distance **d** by formula (1). To find the highest resolution possible with this imaging system, we simply need to look at what the maximum α_{max} angle would be that is not yet blocked by the aperture. The smallest periodic detail d_{min} that can still be transmitted by this optical system is:

$$d_{min} = \lambda / \sin(\alpha_{max}) \qquad (2)$$

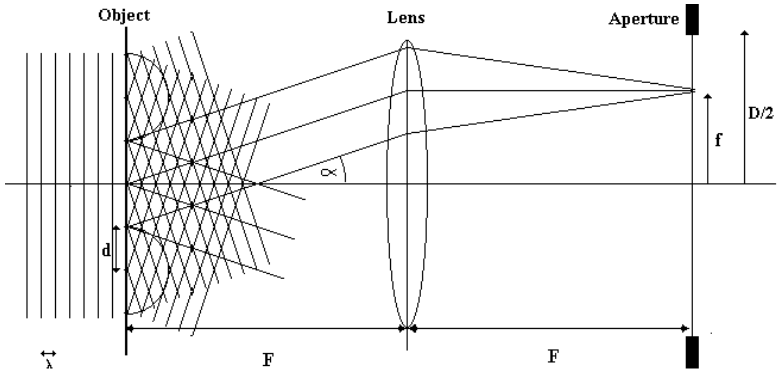


Figure 8. The limiting aperture (with diameter "D") in the back focal plane of the imaging instrument prevents the very fine detail image information to reach the image (overall set-up: Fig. 7) and it thus limits the resolution achievable by the imaging system.

This formula is generally known as the "Raleigh" resolution criterion in a slightly different form:

$$d_{min} = 0.61\,\lambda / n \sin(\alpha_{max}) = 0.61\,\lambda / NA \qquad (3)$$

"NA" stands for the Numerical Aperture of the lens (defined as **n sin (α_{max})).** The "**n**" in this formula stands for the refractive index of the medium; for example, in oil-immersion microscopy one can reach a higher resolution than in air. The factor 0.61 in (3) stems from the definition of the Raleigh resolution criterion which was derived for just two point scatterers rather than a periodic array of point scatterers, and for a different type of illumination than used here (Lord John William Strutt Raleigh: 1842-1919; Nobel Laureate 1904: http://www.nobel.se/). Note that this criterion is (again) closely related to Bragg's law.

Too high spatial frequencies (f > D/2) are blocked by the aperture and will thus no longer contribute to the image. The "transfer function" for intensity information (ITF: Intensity Transfer Function) will thus be:

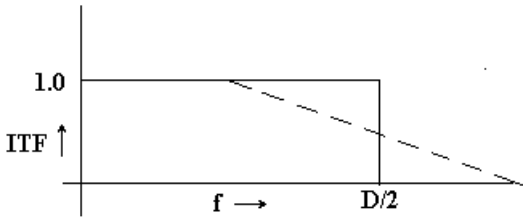


Figure 9. The transfer function of the imaging system of Fig. 7-8, assuming coherent illumination. (The dashed line indicates a transfer function that could be obtained by using partially coherent illumination.)

There can be variants of this ITF depending on the type of illumination. We have hitherto been discussing (coherent) parallel plane wave illumination of the sample. For incoherent or partially coherent illumination [Van Heel, 1978], the high-frequency cut-off of this ITF at D/2 is not sharp but only gradually drops to zero (dashed line). Before we go more into details on how different types of objects – in particular phase

objects – are imaged in the in the microscope, we need some further understanding of what the objects do to the incident plane waves.

Note that in electron microscopy the wavelength of the electron radiation can be calculated from the approximate formula: $\lambda \approx \sqrt{(150/V)}$ in which V is the acceleration voltage (relativistically corrected). Thus, for 300kV electrons we have a wavelength of: ~0.022 Å.

6. Interaction between the incident waves and the specimen

To be able to image an object, we must first have to illuminate the object. We may see the illumination we apply to the object as a vector (Fig. 11a). The length of the vector gives the amplitude of the incoming wave; its direction represents the phase of the incoming wave. The phase of the incoming wave changes continuously and very rapidly with time, but this is not the phase we want to focus on. We are only interested in changes in the phase of the incoming wave due to the interactions with the object ("scattered wave") relative to the phase of the unscattered wave. (The rapid changes in the phase of the unscattered wave with time will also be present in the scattered wave and these will thus cancel as soon as we look at their phase difference.) We thus define the phase of the illumination vector in Fig. 11a to be zero and pointing to the right. The object can influence the incoming wave in two essentially different ways: it can absorb part of the waves or it can change its phase.

The **amplitude** object may absorb a part of the incoming radiation. The outgoing wave will have a smaller amplitude than the incoming wave (Fig. 10a, 11b). This type of object is called an "amplitude" or "intensity" object. The absorption in the object is different from place to place and we can describe the amplitude object by its amplitude function $\psi_A(\mathbf{r})$, where **r** describes the position in the object plane.

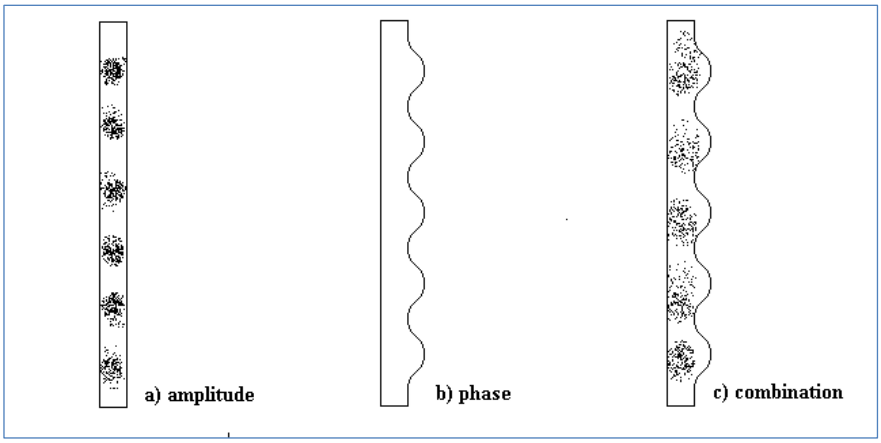


Figure 10. An amplitude object (a) will absorb some of the incoming wave and will thus transmit only part of the radiation. A phase object (b) has a varying thickness which changes the phase of the incident waves without affecting their intensity. Combinations of both (c) are also possible.

A **phase object** does not absorb radiation, but rather delays or advances the incoming wave, leading to an object wave that locally has a phase different from than in the rest of the object plane (Fig. 10b, 11c). We can also see such a phase object as a "glass plate" that locally has a different thickness than it has in the rest of the object plane. It need not really be physically thicker: a locally different refractive index has the same effect as a local thickness variation. Note that the full wave goes through the object and - as a whole - suffers phase retardation due to the average optical thickness of the object. This global phase delay is, again, not interesting since it affects all parts of the object plane. After passage through the object, we can define the average of the object wave to be zero. The phase object locally rotates the wave-vector over an angle ψ_{Ph} (r) relative to the average phase. For small phase shifts, this angle equals the distance ψ_{Ph} (r) as is drawn in Fig. 11c.

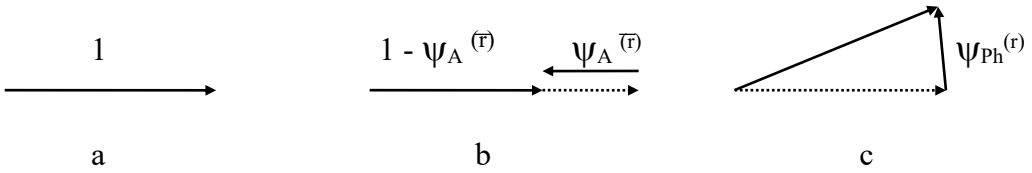


Figure 11. Interactions of the incident beam with different types of specimens. A) The incident plane waves are described by a unit vector of length "1". B) An amplitude object will absorb some of the incoming energy and will thus transmit only part of the radiation. C) In contrast, a phase object is fully transparent and does not absorb any energy. A phase object has a varying thickness which locally changes (rotates) the relative phase of the incident wave.

To describe the imaging of amplitude and phase objects in the microscope, or to explain the "phase problem" (see below), we need the mathematics of complex numbers (See: Appendix 1). The incident wave we have called "1", above (Fig. 11a). The (complex) wave just behind the object can be described by ("1" times) the object wave function:

$$O(\mathbf{r}) = \psi_A(\mathbf{r}) \cdot e^{\{-i \cdot \psi_{Ph}(\mathbf{r})\}} \qquad (4)$$

In this equation, $\psi_A(\mathbf{r})$ describes the amplitude variations in the object (as function of position "**r**" in the object) and $\psi_{Ph}(\mathbf{r})$ describes the phase variations in the object. This formalism is thus suited to describe all three different types of objects depicted in Fig. 10. In the "ideal" microscope, the wave function in image plane would be identical (apart from the magnification) to the modulated wave just behind the object. When we record the image, we record the intensities in the image, that is, the square of the wave functions (see Appendix 1, and below):

$$I(\mathbf{r}) = O(\mathbf{r})^2 = \psi^2_A(\mathbf{r}) \qquad (5)$$

Thus, although the amplitude information is recorded in the image intensity in the ideal microscope, the phase information is not. Our main objects in electron microscopy – and certainly all cryo-EM specimens – are phase objects (Fig. 10a). Thus, the image we register will not show anything of the phases $\psi_{Ph}(\mathbf{r})$ we are most interested in, but rather just a homogeneous grey image.

7. The Phase problem in X-ray vs. EM

Both in X-ray crystallography and in Electron Microscopy we face a "phase problem". Although the "phase problem" means different things in X-ray crystallography and in Electron Microscopy, both are based on the fact that when recording a diffraction pattern or and image, we only register the *square* of the complex wave rather than the wave itself. As we just have seen in formula (5), the phase of the complex wave disappears when the intensities (square of the Amplitude) are registered.

The phase problem in X-ray crystallography occurs when registering the diffraction pattern (Fourier space). The phases of the diffraction pattern peaks (*relative to the phase of the zero-order beam*) contain the most important information in terms of the structure of the molecules in the unit cell of the crystals. Much of the effort in X-ray crystallography goes into finding the phases of the diffraction spots.

In Electron Microscopy, we register the square of the complex wave in the image plane. Again, the phases of that wave are lost in the process. However, in phase contrast microscopy – and electron microscopy is a form of phase contrast microscopy – one optimises the instrument to convert the *phase* variations in the object plane into *amplitude* variations in the image plane (see below). Thus, although strictly speaking the phases in the image are lost, the amplitudes in the image which one registers contain the information about the phases in the object, which is exactly what we are interest in when studying the structure of biological macromolecules by cryo-EM. Thus, there is not really a phase problem in EM.

But, even if it is not directly relevant in electron microscopy, why is it that we loose the phases of the wave functions (both in EM and in X-ray crystallography) when we register the wave functions in the first place? What is the fundament problem we face here? The problem is associated with the time scale of the events. Suppose on stand in the water at the beach and observes the ocean waves going to the shore. The distance between the tops of two consecutive waves defines the wavelength to the waves, the height of the waves defines their amplitudes, and we can even observe the phase of the arriving waves. After all, we can synchronise our actions to the phase of the arriving waves and jump over the wave tops assuming their amplitudes will allow us to do that. Thus, we can easily measure the amplitudes of the incoming waves as well as their phases. Why does that not also work for electron and photons?

Suppose we expose a photographic film to light using a very short exposure time, say $1/1000^{th}$ of a second. During that exposure the light waves will have travelled ~ 300,000,000 m/sec x 0.001 sec = 300,000 m. Thus, the equivalent of ~ 10^{10} wavelengths has passed through the film during the exposure time. The story for X-ray photons or 300kV electrons is similar. Although determining the phases of ocean waves or radio waves is quite straightforward, the detectors we use in electron microscopy and in X-ray crystallography are simply too slow for determining the phases of the incoming waves directly. There is also not much hope of developing phase-sensitive detectors since there are limitations to what one can do with all known

materials. The phase sensitive detectors can be built if we create a reference wave with which we can generate an interference pattern with the wave we are interested in. This is the principle of Holography, invented by the Imperial College scientist Dennis Gabor (Nobel prize for Physics 1971; http://www.nobel.se/).

7. The Phase Contrast Microscope

As was mentioned above (equation 5), in an ideal microscope, a pure phase object will lead to homogeneously grey image from which all phase information has disappeared. This means that when we are, say, investigating live bacteria in water, nothing is seen in the image. It had long since been common knowledge that one could observe such objects by somewhat defocusing the light microscope and we will get to that below (Zernike, 1942). However, it required a stroke of genius to fundamentally understand the matter and to then design the phase contrast microscope exploiting this basic understanding. Fritz Zernike (1888-1966) invented the phase-contrast microscope in 1930 at the University of Groningen, The Netherlands. The first phase-contrast microscopes were not built until 1941. He received the Nobel Prize for physics 1956 (http://www.nobel.se/). To avoid losing the phase information he proposed the use of a "phase plate" in the back focal plane of the microscope.

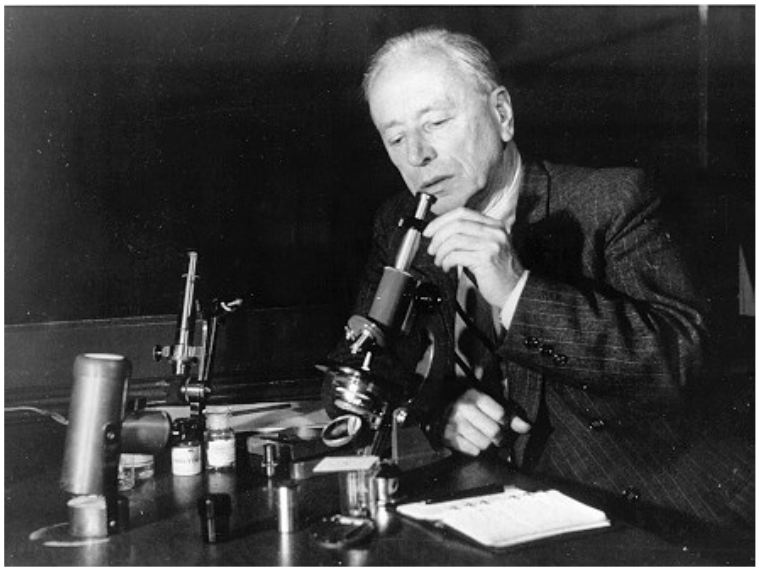


Figure 11. Fritz Zernike (1888-1966) with one of first phase-contrast microscopes were not built until 1941. (Nobel Prize for physics, 1956 (http://www.nobel.se/).

The phase plate (Fig. 12) changes the relative length of the optical path of the diffracted beams with respect to the zero-order beam (such that an extra phase difference of 90º (or 270º) is introduced between the two (Fig. 13). The influence of this extra phase shift in the resulting image is dramatic. We discussed above (Fig. 11) why phase objects are not visible in the conventional microscope: the phase fluctuations in the image (due to the presence of the phase object in the object plane) are "perpendicular" to the plane waves of the un-diffracted beam (Fig. 11c).

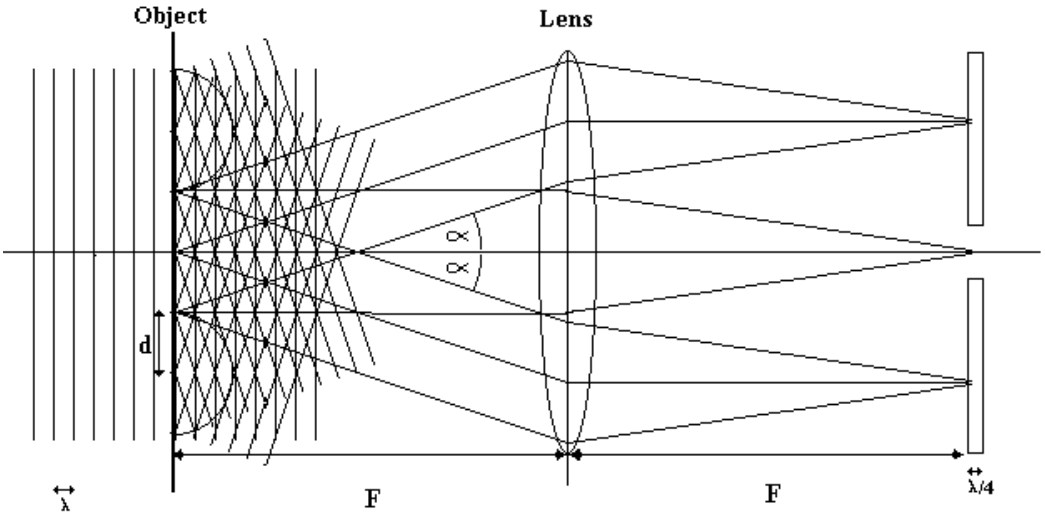


Figure 12. A phase object does not modulate the intensity of the incoming wave, but rather changes the phase of the incident waves. The brilliant trick of phase-contrast microscopy is to manipulate the phases of the diffracted beam, relative to the phase of the 0-order beam (on the optical axis in this example).

The phase plate will change the world around: that phase information that originally led to fluctuations perpendicular to the background, now leads to fluctuation along the background light vector. At the same time, the amplitude information fluctuations that was along the direction of the background light (Fig. 11b) now becomes perpendicular to the background vector and is thus invisible in the image.

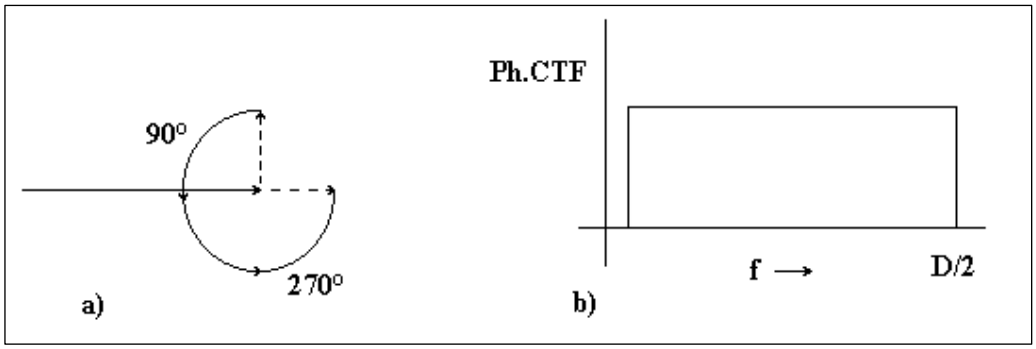
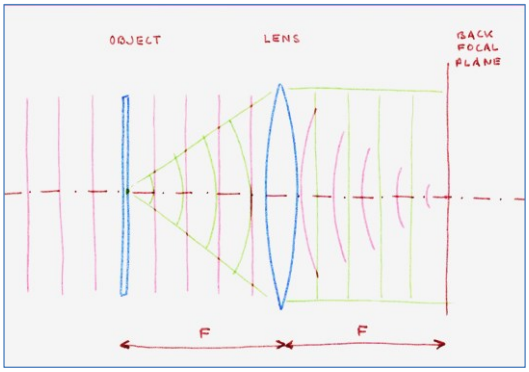


Figure 13. The Zernike phase plate in the back focal plane of the objective lens manipulates the phases of the diffracted beams with respect to the zero order beam such that they will interfere constructively in the image plane. The phase rotation (a) can be either ~90º or ~270º. The difference between the resulting images (90º versus 270º), is a complete reversal in contrast.

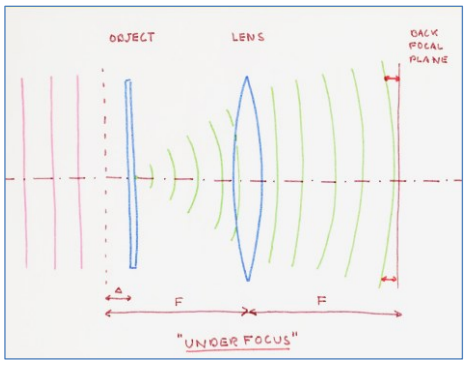
In terms of the transfer of the phase information through the microscope we must realise that the Phase Contrast Transfer Function (PhCTF) is somewhat different from the transfer function we discussed earlier (Fig. 9). The phase contrast microscope translates phase fluctuation in the object into amplitude modulations in the image. The PhCTF (short: CTF) illustrates who well the instrument does this. Since PhCTF thus compares different entities is its sometimes called a "pseudo" transfer function.

8. The Aberrated Microscope

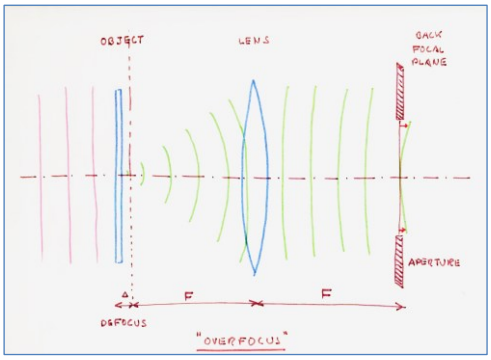
A simple experiment: If we frame a piece of polyethylene foil in a slide holder, we do not see anything if the slide projector is focused. However, if we turn the lens out of focus, thickness variations in the foil due to its manufacturing process will become visible. The aberrated microscope can thus transfer phase information and that effect has puzzled many microscopists already more than a century ago. The *modus operandi* of the "defocused" microscope is pictured in Fig. 14.



Simple lens: "in focus"



"Under-focus"



"Over-focus"

Figure 14. A simple defocusing of a light microscope may help visualising phase differences in the object. The dotted vertical lines in the lower illustrations mark the ideal "focus" position. The departure of the diffracted wave from the ideal focal plane can be interpreted as a phase plate in the back-focal plane and will lead to visualising certain spatial frequencies in the object.

We had already seen that a point in the object plane corresponds to a plane wave in the back focal plane (Fig. 14a). If we shift the point to the right towards the lens – the under-focus situation – the wave-front becomes divergent (Fig. 14b). If we shift that point slightly to the left (over focus direction) the wave in the back focal plane becomes convergent (Fig. 14c). In the under-focus case (Fig. 14b; Fig. 15), the wave in the back-focal plane will lag in phase relative to the wave on the optical axis. This is explained in more detail in the legend of Fig. 15. Interestingly, we could have obtained the same situation by inserting a physical phase plate in the back focal plane that would cause this phase shift. This is why the phase structure of the plastic sheet

becomes visible upon defocusing or why we can see phase information in the normal (but defocused) light microscope. Although we normally don't think of defocusing as being an image aberration, we can generally consider lens aberrations to work like a phase plate in the back focal plane of an optical system. The PhCF for this situation will schematically behave as shown in Fig. 16. Since the PhCF in the defocused microscope is no longer zero everywhere, phase information becomes visible.

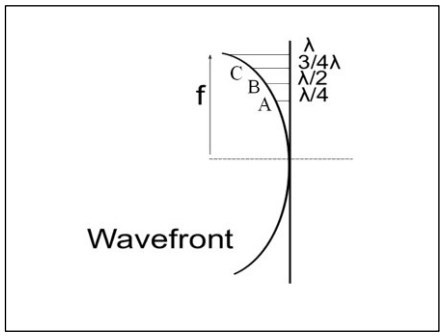


Figure 15. Defocus as a phase plate.

At position A, in the back focal plane the phase is delayed by 90° with respect to the phase at the optical axis: that leads to the imaging of this particular spatial frequency component with ideal (positive) phase contrast. A bit further up (B) the total phase delay is 180°, leading to no imaging of that spatial frequency. At position C, the total phase difference is 270° and we obtain ideal, negative phase contrast etc.

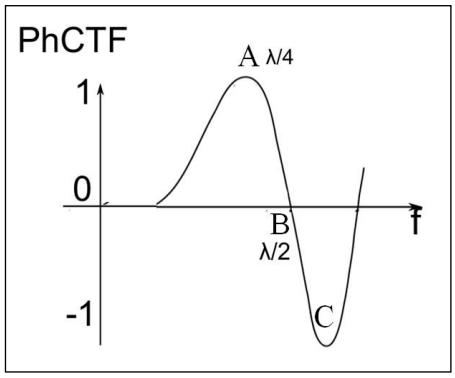
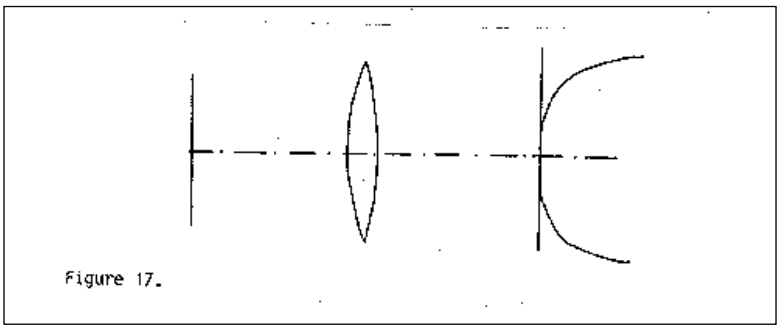


Figure 16. The associated PhCTF.

The information transfer in the defocused microscope is expressed by a Phase Contrast Transfer Function which is the sine of the wave aberration function shown in Figure 15. It has its positive maximum (at "A") where the phase shift reaches 90° (λ/4) and its negative maximum where the phase shift is 270° (3/4 λ).

9. The Electron Microscope as a Phase Contrast Microscope

The electron microscope has an aberration that cannot easily be corrected: the spherical aberration. The spherical aberration coefficient "**Cs**" is measured in mm (see below). The additional phase shifts in the back focal plane due to spherical aberration are schematically drawn in Fig. 17.



For the higher spatial frequencies, the phase shifts grow rapidly, namely with the *fourth* power of the distance from the axis. ,In combination with a certain amount of "underfocus" (Fig. 15) leads to an overall phase shift in the back focal plane that is

exactly λ/4 over a large range of frequencies (Fig. 18). Given the right combination of underfocus and spherical aberration we obtain an "ideal" phase-contrast microscope for frequency ranges from ~20Å to ~4Å ("Scherzer focus"). The PhCF for this situation is shown schematically in (Fig. 19); and computationally in (Fig. 20).

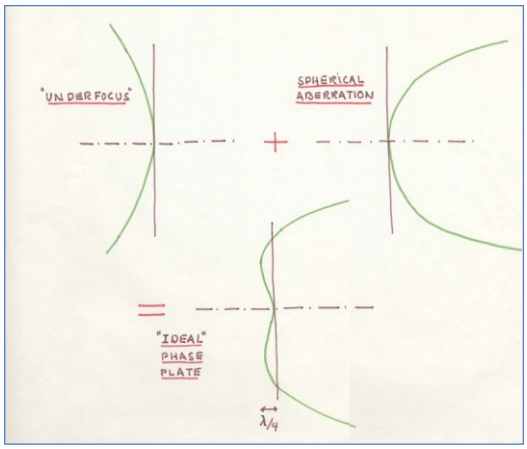


Figure 18

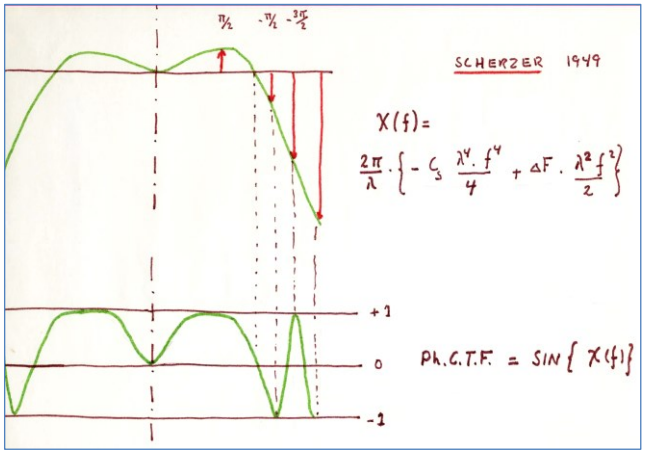


Figure 19

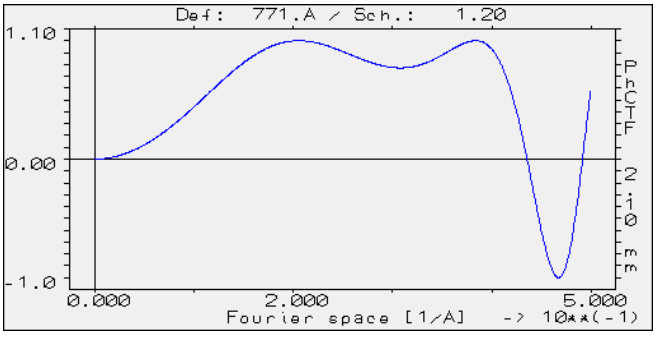


Figure 20

Figures 18-20. The combination of under-focus and spherical aberration in the electron microscope makes for an almost ideal virtual phase plate. When the amount of under-focus optimally matches the spherical aberration of the objective lens, the information transfer in the EM is optimal (Fig. 20). This situation is called "Scherzer focus" in honour of Otto Scherzer.

The phase shift in the back focal plane due to spherical aberration and defocusing together, are given in (Scherzer 1949):

$$X(f) = \frac{2\pi}{\lambda}\left\{-\frac{C_s\,\lambda^4 \cdot f^4}{4} + \frac{\Delta f\,\lambda^2 \cdot f^2}{2.}\right\} \qquad (6)$$

The PhCTF is then obtained by calculating the SINE of the wave aberration function:

$$PhCTF(f) = \mathrm{Sin}\left\{\frac{2\pi}{\lambda}\left(-\frac{C_s\,\lambda^4 \cdot f^4}{4} + \frac{\Delta f\,\lambda^2 \cdot f^2}{2}\right)\right\} \qquad (7)$$

In this formula, the variables are:

C_s: the spherical aberration coefficient (measured in mm.)
Δf : defocus (positive for underfocus) (-40000Å to + 40000Å)
f : spatial frequency
λ : wavelength of the electrons (at 80kV : 0.043 Å)

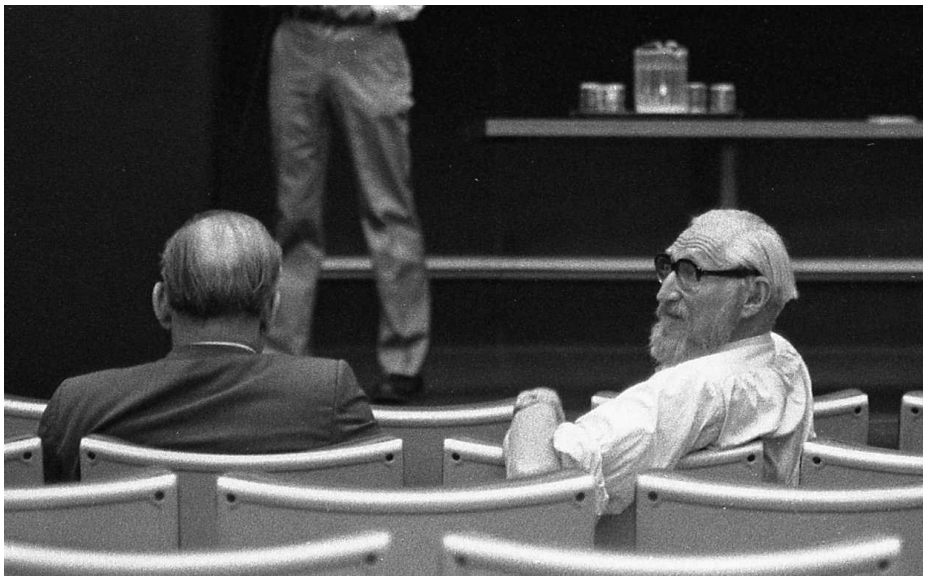


Figure 21. Professor Otto Scherzer (1909–1982) used the phase-contrast formalism to explain the main image contrast in the electron microscope (1949). Taken in 1978 by MvH at the Intern. EM conference in Toronto, Canada.

A final word of warning: this document is primarily about the linear transfer of information. Nowhere has the issue of harvesting noisy information has been mentioned. For a modern idea about that see reference **#12**.

References:

1. E. Abbe, "**Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung**", *Archiv f. Mikroskopischer Anatomie* Bd. **9** (1873) 413-480.

   It is amazing how much someone as brilliant as Abbe can say in words without resorting to the use of even a single formula.

2. F. Zernike, "**Phase contrast, a new method for the microscopic observation of transparent objects, Part I**", *Physica* **9** (1942) 686-693.

   This paper starts with: "Every microscopist knows that transparent objects show light or dark contours in different ways varying with the change of focus and depending on the kind of illumination used. Curiously enough, the wave theory of light has never been explicitly applied to the case of absolutely transparent objects, the details of which differ only in thickness or refractive index."

3. F. Zernike, "**Phase contrast, a new method for the microscopic observation of transparent objects, Part II**", *Physica* 9 (1942) 974-986.

4. P.M. Duffieux, **L' Intégrale de Fourier et ses Aplications à l'Optique**, Facultée des Sciences Besançon (1946).

5. O. Scherzer, *J. Appl. Phys*. **20** (1949) 20-29.

6. M. Born and E. Wolf, "**Principles of Optics**", Pergamon Press (1975).

7. R.N. Bracewell, "**The Fourier Integral and its Applications**", McGraw-Hill (1965).

8. D.C. Champeney, "**Fourier Transforms and their Physical Applications**", Academic Press (1973)

9. J.W. Goodman, "**Introduction to Fourier Optics**", McGraw-Hill (1968).

   Oriented towards theoreticians/ physicists. Still one of the standard books in the field. No mention of electron microscopy at all.

10. M. van Heel, "**On the imaging of relatively strong objects in partially coherent illumination in optics and electron optics**", *Optik* **47** (1978) 389-408.

11. M. van Heel, "**The Phase Contrast Transfer Function is - per definition - zero at the origin."** in preparation.

12. M. van Heel, M. Schatz: **Information: to Harvest, to Have and to Hold** (https://arxiv.org/abs/2009.03223)

Appendix 1: Complex Numbers

Complex numbers, rather than representing a point along a line as do "Real" numbers, represent a point in a two-dimensional plane. A point in a plane can be described by its X- and Y-coordinates; in complex numbers nomenclature, these are called the "Real" and "Imaginary" part. Alternatively, the complex number can be described by its "Amplitude" and "Phase".

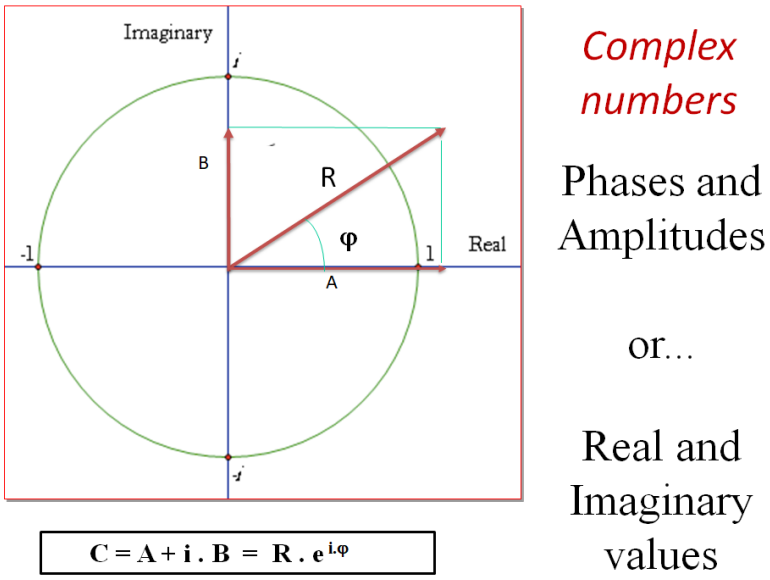


Figure (A1-1). Classical complex number representations: as "*Phases and Amplitudes*" or as "*Real and Imaginary*" parts.

In **Real** and **Imaginary** representation:

$$\mathbf{C} = \mathbf{A} + \mathbf{i} \cdot \mathbf{B} \qquad (1)$$

In **Amplitude** and **Phase** representation:

$$\mathbf{C} = \mathbf{R} \cdot \mathbf{e}^{\mathbf{i}.\varphi} \qquad (2)$$

In the amplitude and phase notation, one uses polar co-ordinates rather than Cartesian (x-y) co-ordinates. In polar co-ordinates, one describes the position of a point in a plane by its distance to the origin (the radius or the amplitude), and by the angle that the line connecting the origin to the point makes with the positive X-axis.
By definition:

$$\mathbf{e}^{\mathbf{i}.\varphi} = \mathbf{Cos}\,(\varphi) + \mathbf{i} \cdot \mathbf{Sin}\,(\varphi) \qquad (3)$$

Thus, putting (1), (2) and (3) together we have:

$$\mathbf{A} = \mathbf{R} \cdot \mathbf{Cos}\,(\varphi) \qquad (4)$$

And:

$$\mathbf{B} \quad = \quad \mathbf{R \,.\, Sin\,(\varphi)} \qquad \mathbf{(5)}$$

These two formulas allow one to convert a complex number written in amplitude and phase notation to a complex number written in real and imaginary part notation.

The square of a complex number "**C**" (the "Intensity") is defined as:

$$\mathbf{C^2 = C \,.\, C^* \quad = (\,A + i.B)\,.\,(A + i.B)^*} \quad =$$

$$\mathbf{(A + i.B)\,.\,(A - i.B) \quad = \; A^2 + B^2} \qquad \mathbf{(6)}$$

(C* is the complex conjugate of C). Alternatively, in the amplitude and phase representation:

$$\mathbf{C^2 = C \,.\, C^* \quad = \; R \,.\, e^{\,i.\varphi} \,.\, R \,.\, e^{\,-i.\varphi} \;= R^2} \qquad \mathbf{(7)}$$

Note that in the squaring operation one uses a special version of the multiplication operation, the "conjugate multiplication". In the conjugate multiplication, one uses the conjugate of the second factor for the multiplication (replace "i" by "-i"). After the squaring operation, the amplitude "**R**" can be retrieved, not the phase **φ** of the complex number.

Appendix 2: The Fourier Transform (in 1D-2D-….)

The Fourier Transform, invented (discovered?) by Jean-Baptiste Fourier (1768-1830), plays a central role in signal and image processing. In order to convey some sense for the behaviour of this important mathematical transform, here are some sketched examples first:

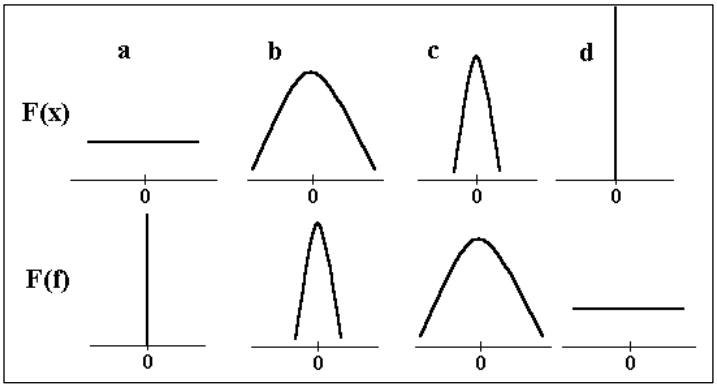


Figure A2-1

In the top row of Fig. 1, a number of functions (**F(x)**) is drawn, while the lower row shows their corresponding Fourier Transforms (**F(f)**). In Fig. **1a**, a constant signal leads to a very strong peak (a "***delta function***") at frequency zero in the Fourier transform. This is not too surprising since the signal **F(x)** contains exclusively one very low frequency (namely, the frequency zero). In going from **Fig**. **1a** to **1d**, the signal **F(x)** becomes narrower and higher, until the signal itself becomes a delta function (**1d**). At the same time, we see that the corresponding Fourier transforms broaden: in **1d** the transform has become a constant. This is the ***reciprocal*** behaviour of a function and its Fourier transform: short distances in real space correspond to large distances in Fourier space and vice-versa. (Keyword: "***reciprocity theorem***")

As was mentioned above, the FT decomposes a signal into sine and cosine waves. If the signal consists of a single cosine wave, the FT of that signal thus takes a simple form (we will also soon understand why these transform pairs provide nice links to the field of optics and X-ray crystallography):

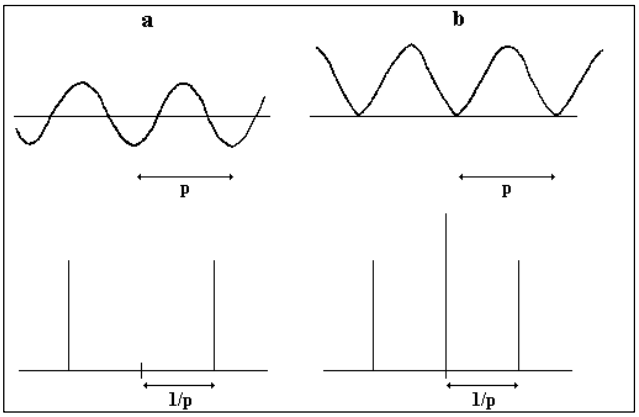


Figure A2-2

The cosine waves presented in Fig. 4a,b are essentially identical, but the one in Fig 4a (top row) has a zero average, whereas the one in Fig. 4b is always positive. We can think of the latter one as being the sum of the first and a constant. From Fig. 3a, we have seen that the FT of a constant is a large peak at the origin. Thus, the FT of the function in Fig. 4b is simply the FT of Fig.4a, plus a large peak at the origin. The peak at the origin is called the zero-order or the unscattered beam in X-ray and Electron crystallography. (Just to test your understanding of the reciprocity principle of the FT: what happens with the transforms, if we change the period of the sine-shaped waves in **F(x)**? That is to say, what happens if we change the period **p** to **2p** or to **p/2**?)

There are a number of definitions of the Fourier transform currently in use. The definition I prefer, since it avoids the presence of "2π" factors in transform pairs, is:

$$\underline{\mathbf{F}}(\mathbf{f}) = \int_{-\infty}^{+\infty} \mathbf{F(x)}\ \exp(-2\pi i\ \mathbf{f}\cdot\mathbf{x})\, d\mathbf{x} \qquad (1)$$

With its inverse:

$$\mathbf{F(x)} = \int_{-\infty}^{+\infty} \underline{\mathbf{F}}(\mathbf{f})\ \exp(+2\pi i\ \mathbf{f}\cdot\mathbf{x})\, d\mathbf{f} \qquad (2)$$

We are not doing anything with these Fourier integrals here: the definitions are only given to facilitate their recognition in the literature. The link to Optics was first made by Duffieux in 1946 in his historical work: "l'Intégrale de Fourier et ses applications a l'Optique" (which, I must admit, I've never read). In Fourier Optics, the Fourier transforms are typically used in their two-dimensional **2D** form:

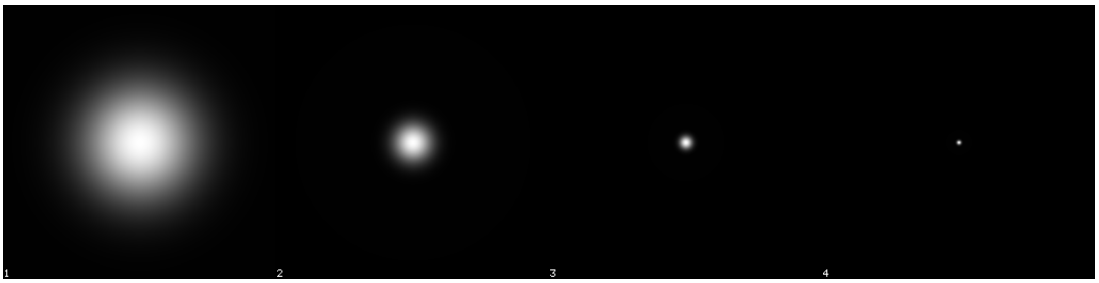
$$\underline{\mathbf{F}}(\mathbf{f}, \mathbf{g}) = \int_{-\infty}^{+\infty} \mathbf{F(x, y)}\ \exp\{-2\pi i\,(\mathbf{f}\cdot\mathbf{x} + \mathbf{g}\cdot\mathbf{y})\}\, d\mathbf{x}\, d\mathbf{y} \qquad (3)$$

With its inverse:

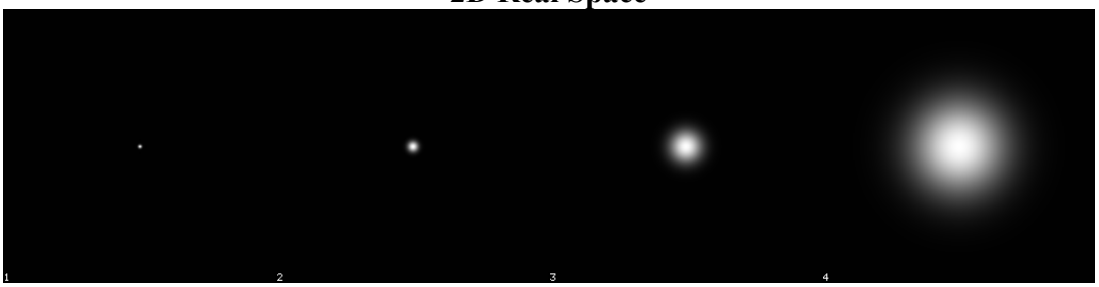
$$\mathbf{F}(\mathbf{x}, \mathbf{y}) = \int_{-\infty}^{+\infty} \underline{\mathbf{F}}(\mathbf{f}, \mathbf{g})\ \exp\{+2\pi i\,(\mathbf{f}\cdot\mathbf{x} + \mathbf{g}\cdot\mathbf{y})\}\, d\mathbf{f}\, d\mathbf{g} \qquad (4)$$

More compact formulation:

$$\underline{\mathbf{F}}(\mathbf{f}, \mathbf{g}) = FT^{+2}\{\mathbf{F}(\mathbf{x},\mathbf{y})\}\ ; \quad \mathbf{F}(\mathbf{x},\mathbf{y}) = FT^{-2}\{\underline{\mathbf{F}}(\mathbf{f},\mathbf{g})\} \qquad (5)$$



2D Real Space



2D Fourier Space

<u>Figure A2-3</u>: Gaussian 2D Fourier transform pairs: In the top row of the figure, a number of 2D Gaussian functions (F(x)**) are depicted, while the lower row shows their corresponding Fourier Transforms (**F(f)**). In the top line, left, a very wide Gaussian object leads to a very strong peak (almost a "***delta function***") around frequency zero in 2D Fourier space (left in second line). In going from left to right in the top row, the 2D signal** F(x) **becomes narrower, until the signal itself almost becomes a delta function. At the same time, we see that the corresponding Fourier transform broadens continuously: from almost a delta function to becoming a very wide Gaussian and finally a constant (again, not depicted). Another illustration of the** *reciprocal* **behaviour of a function in Real-space versus Fourier-space. (Keyword: "***2D reciprocity theorem***")**