Chapter 2
Electron Waves and Wave Propagation

Electron diffraction and imaging rely on the wave properties of electrons. A basic understanding of wave and wave properties is thus required for the interpretation of electron diffraction and electron imaging. Wave theory is also required for understanding electron probe formation using electron lenses. This chapter introduces waves, wave-related properties, and wave equations. While the basic concepts described here can be found in introductory text to electron microscopy, the discussions on wave propagation and wave coherence are conducted at a level equivalent to graduate courses in physics. For these sections, the readers are referred to books on quantum mechanics by Griffiths (2004) and on physical optics by Born and Wolf (1999) and Goodman (2004).

2.1 Wave Functions and the Wave Equation

There are two distinct, universal, properties associated with waves; one is the propagation of waves, and the other is local disturbance. If we think of a floating bouy anchored to the ocean floor, its vertical motion as waves pass under it maps out the wave amplitude as a function of time at one point. By contrast, a snapshot photograph of the ocean surface provides a map of the wave amplitude at one time as a function of space. Consider now a one-dimensional wave which propagates at a speed of v along the x direction. An observer moving with the wave, in the wave coordinate x', only sees a local disturbance, and the wave function is then simply given by $\phi = f(x')$. For a stationary observer, the origin of the x' coordinate moves by a distance of vt at time t. The relation between the two coordinates is given by $x' = x - vt$; thus, the 1D wave function seen by the stationary observer has the general form of

$$\phi = f(x - vt) \tag{2.1}$$

© Springer Science+Business Media New York 2017

J.M. Zuo and J.C.H. Spence, *Advanced Transmission Electron Microscopy*,
DOI 10.1007/978-1-4939-6607-3_2

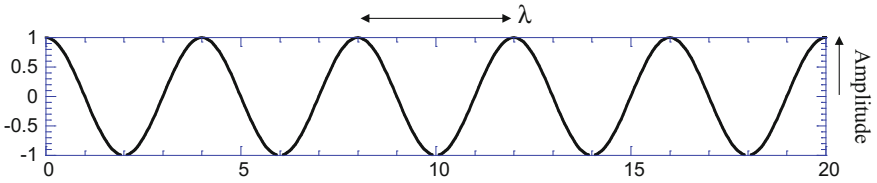

Fig. 2.1 Wave function of Eq. (2.3) plotted as function of x for $t = 0$ and $A = 1$

Double differentiating Eq. (2.1) on both sides by x and t gives the homogeneous wave equation of

$$\frac{\partial^2}{\partial x^2}\phi = \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\phi. \tag{2.2}$$

The simplest wave is a sinusoidal wave in the form of

$$\phi = A\cos\left[\frac{2\pi x}{\lambda} - \omega t\right], \tag{2.3}$$

or

$$\phi = A\sin\left[\frac{2\pi x}{\lambda} - \omega t\right]. \tag{2.4}$$

Here, λ is the wavelength, ω is the frequency, and A is the wave amplitude. The wave velocity is then given by (Fig. 2.1)

$$v = \omega\lambda/2\pi. \tag{2.5}$$

Both the sine and cosine forms of wave function are the solutions of Eq. (2.2), so a general solution is a combination of these two, each with its own amplitude. By choosing an appropriate phase, the combination of sine and cosine functions can be expressed using a single cosine function so that

$$\phi = A\cos\left[\frac{2\pi x}{\lambda} - \omega t + \delta\right], \tag{2.6}$$

Here, δ is called the initial phase of the wave. The solution in Eq. (2.6) can be alternatively written as follows:

$$\phi = \mathrm{Re}\left\langle A\exp\left[i\left(\frac{2\pi x}{\lambda} - wt + \delta\right)\right]\right\rangle$$

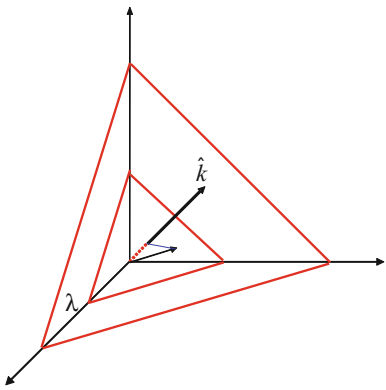

Fig. 2.2 Plane wave in 3D

or

$$\phi = \phi_o \exp\left[i\left(\frac{2\pi x}{\lambda} - wt\right)\right] = \phi_o \exp(2\pi ikx)\exp(-i\omega t) \tag{2.7}$$

by taking the real part afterward. The wave amplitude ϕ_o is complex in general.

So far, we have only discussed the 1D sinusoidal wave that is propagating along the x direction. In 3D, the sinusoidal wave can travel in any direction. Its direction is specified by a unit vector $\hat{k}$ (Fig. 2.2). An important point is that a 3D sinusoidal wave is no different from the 1D sinusoidal wave; the difference is that we instead choose to look at it at a different angle. A property of the wave of Eq. (2.6) is that the phase only changes in the direction of wave propagation (with x only). For this reason, a sinusoidal wave is also called plane wave because of its constant phase in the 2D plane normal to the wave propagation direction. The 3D equivalent is that the phase only depends on the distance along $\hat{k}$, which mathematically is given by $\hat{k} \cdot \vec{r}$. Putting this into Eq. (2.7), we have a general description of a sinusoidal wave in 3D of the form

$$\phi(\vec{r}, t) = \phi_o \exp\left[2\pi i\left(\frac{\hat{k} \cdot \vec{r}}{\lambda}\right)\right] \exp(-i\omega t) = \phi_o \exp\left[2\pi i\vec{k} \cdot \vec{r}\right] \exp(-i\omega t). \tag{2.8}$$

We call the vector $\vec{k} = \hat{k}/\lambda$ the wave vector and $\left|\vec{k}\right| = \frac{1}{\lambda} = \sqrt{k_x^2 + k_y^2 + k_z^2}$ the wave number. Correspondingly, the homogeneous wave equation in 3D is given by

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\phi = \nabla^2\phi = \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\phi \tag{2.9}$$

The plane wave described in Eq. (2.9) is only one of many solutions to the three-dimensional wave equation. The other special solution is the spherical wave, which is emitted from a point source. The wave function of a spherical sinusoidal wave has the form

$$\phi = \frac{\phi_o}{r}\exp(2\pi ikr)\exp(-i\omega t) \tag{2.10}$$

It can be shown that this is a solution of the homogeneous wave equation by using the spherical representation of Eq. (2.9)

$$\nabla^2\phi = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)\phi = \frac{1}{r}\frac{\partial^2}{\partial^2 r}(r\phi) = \frac{1}{v^2 r}\frac{\partial^2}{\partial t^2}(r\phi) \tag{2.11}$$

The spherical sinusoidal wave of Eq. (2.10) represents a diverging wave from the point of origin. By replacing k with $-k$, another solution representing a converging spherical wave is obtained. For plane waves, changing the sign of wave vector reverses the wave propagation direction.

By summing waves of different wave vectors together, three distinct important types of wave can be formed—the running waves treated above, the standing wave, and the pulse or wave packet. It is also possible to sum over a range of wave vectors and frequencies—in that case, the relationship between wave vector and frequency is called a dispersion relation, and this depends on the properties (the dielectric function in optics, or the electromagnetic potential for electron diffraction) of the medium in which the wave is traveling. We discuss all these effects in the following sections. The wave packet will turn out to be particularly important for our later treatment of STEM or electron nanodiffraction using a coherent probe.

2.2 Quantum Mechanical Wave of Electrons and Schrődinger Equation

Waves due to mechanical vibrations (such as the vibration of a guitar string) or electromagnetic oscillations (such as low-frequency radio waves observed directly on an oscilloscope) can be directly observed; e.g., we can measure both the amplitude and phase of these waves in principle. The quantum mechanical wave, including that of an electron, cannot be measured the same way. The mere fact that electrons are diffracted by crystals attests to the wavelike behavior of electrons. We also see the particle-like behavior of electrons inside electron microscopes; electrons are bent by the presence of electric or magnetic fields in a trajectory predicted by the theory of classical mechanics. In fact, we prefer to describe electrons as particles moving along optical paths (trajectories) on most of their trajectory through the electron microscope. Electrons are also detected by collision with another electron or atom. For example, when a photographic film is exposed to electrons, electrons are

"captured" as speckles on the film after chemical processing—we can say that electrons travel as waves, but arrive as particles. (The likelihood of the electron arriving at a particular point is then proportional to the square of the electron wave function, described below.) Anyone who underexposes a lattice image or records low-dose electron images can testify to this, since the electron image is seen to be built-up from many individual electron arrivals, like raindrops. A discernible image or pattern only emerges when there are a sufficient number of these electrons. Thus, whenever we try to measure electrons, we only see particles. Summarizing this situation in the language of quantum mechanics, we have following properties, which are not limited to the electrons:

(1) An electron can be wavelike or particle-like;
(2) Its property is described by the electron wave function. Wave function itself is not measurable. What is measurable is the distribution of electrons, given by the square intensity of the wave:

$$I = |\phi|^2 = \phi \cdot \phi^*. \tag{2.12}$$

Here, ϕ^* is the complex conjugate of the wave function ϕ. The wave function must be normalized so the overall probability of finding the electron is 1, which is achieved by following integral

$$\int_{-\infty}^{\infty} \phi(\vec{r})\phi^*(\vec{r})d^3\vec{r} = 1. \tag{2.13}$$

(3) The distribution can be any one of the measurable properties of the electrons. Each property is obtained by applying an appropriate operator ($\hat{A}$) to the wave function. For example, the operators of $\hat{A} = x$ and $\hat{A} = -ih\partial/\partial x$ give the position and momentum along the x direction, while $\hat{A} = t$ and $ih\partial/\partial t$ give the time and energy. What is measured is the expectation value, according to

$$\langle \hat{A} \rangle = \langle \phi^* | \hat{A} | \phi \rangle = \int \phi^*(\vec{r})\hat{A}\phi(\vec{r})d^3\vec{r}. \tag{2.14}$$

For electrons inside a potential field, $V(\vec{r})$, the wave function satisfies the Schrődinger equation in the form

$$-\frac{h^2}{8\pi^2 m}\nabla^2\phi - eV(\vec{r})\phi = i\frac{h}{2\pi}\frac{\partial}{\partial t}\phi. \tag{2.15}$$

The time period of electron wave oscillation, $2\pi/\omega$ (4×10^{-20} s for electrons of 100 keV), is however too short to detect for high-energy electrons—it is the differences in this frequency which appear as energy losses when multiplied by Planck's constant; the fundamental frequency is not believed by most physicists to

be an observable. In a typical experiment performed inside an electron microscope, the measurement is over a much longer time period than this frequency. The time period is also much shorter than that of atomic vibrations in the order of 10^{-12} s (picoseconds). Thus, the potential can be assumed to be time independent. Then, the complex wave function can be separated into two parts, the time dependent and time-independent parts, such as $\phi(\vec{r},t) = \phi(\vec{r})\exp(-i\omega t)$. The time-dependent part is same to all electrons of the same energy, which can be taken out of the equation. For the rest of this book, we therefore use the so-called time-independent Schrödinger equation:

$$-\frac{h^2}{8\pi^2 m}\nabla^2\phi - eV(\vec{r})\phi = E\phi, \tag{2.16}$$

where $E = \hbar\omega$ is the electron energy. For an electron traveling in vacuum, the potential is zero. Then, Eq. (2.16) reduces to

$$-\frac{1}{4\pi^2}\nabla^2\phi = \frac{2mE}{h^2}\phi = k^2\phi, \tag{2.17}$$

where $k = \sqrt{2mE/h^2}$. Equation (2.17) is then the same wave equation as in Eq. (2.9) for a single frequency. This equivalence enables us to apply our accumulated knowledge about waves from other fields, such as optics, to understand electron imaging and diffraction, while the specific quantum-mechanical wave properties of electrons are discussed in reference to electron-specimen interaction.

2.3 The Principle of Wave Superposition

The wave equation, in both the classical and quantum mechanical forms (Eqs. 2.9 and 2.15), is linear. By this, we mean that the equation only involves ϕ, and there are no higher order terms including products of ϕ in the wave equation. A general property of linear equations is that the sum of any two solutions is also a solution. Thus, if two waves of ϕ_1 and ϕ_2 are solutions of the wave equation, $\phi = \phi_1 + \phi_2$ is also a solution of the wave equation. Obviously, this extends to any number of such waves, and in general,

$$\phi = \sum_{i=1}^{N}\phi_i \tag{2.18}$$

is a solution, as are the individual solutions.

The property stated in Eq. (2.18) is referred as the principle of superposition. This enables us to write any wave function as a sum of solutions, such as the sinusoidal waves and the spherical waves from individual point sources.

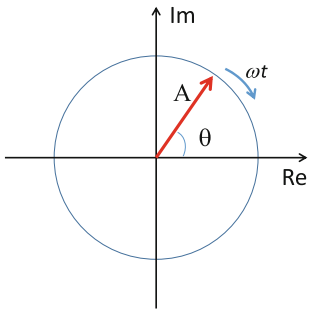

Fig. 2.3 Amplitude and phase diagram of a complex wave function

2.4 Amplitude and Phase Diagrams

The complex wave of Eq. (2.8) can be separated into three components: the amplitude, a position-dependent phase, and a time-dependent phase:

$$\phi = A\exp(i\theta)\exp(-i\omega t) \tag{2.19}$$

where $\theta = 2\pi\vec{k}\cdot\vec{r} + \delta$, and δ is the initial phase. The complex number of Eq. (2.19) can be plotted in the so-called complex plane, with the x-axis for the real and the y-axis for the imaginary parts of the number (see Fig. 2.3). The length of the vector is the amplitude A, and its angle with the x-axis is θ. The vector rotates as time progresses with frequency ω.

In the case that we have two waves with the same ω:

$$\varphi = [A_1\exp(i\theta_1) + A_2\exp(i\theta_2)]\exp(-i\omega t) \tag{2.20}$$

The resultant wave is a linear sum in the way that these two are added together as the addition of two vectors.

2.5 Coherence and Interference

The superposition of two waves of the same frequency results a new wave of the same frequency. The superposition of two waves of different frequencies, however, varies with time with the amplitude from $|A_1 + A_2|$ to $|A_1 - A_2|$. Since in a typical experiment, the electron intensity is measured over a certain limited period of time T, the question of practical importance is how the two types of superposition affect the intensity of waves, and the question arises as to whether waves of different frequency can interfere (the answer, as we now show, is "briefly").

Certain properties of wave superposition give rise to the concept of wave coherence. To examine the results of wave superposition, we start with Eq. (2.18), and for simplification, we look at the superposition of two one-dimensional waves:

$$\begin{aligned}\phi(x,t) &= \phi_1(x,t) + \phi_2(x,t) \\ &= A_1 \exp[2\pi i(k_1 x - v_1 t) + i\delta_1] + A_2 \exp[2\pi i(k_2 x - v_2 t) + i\delta_2]\end{aligned} \tag{2.21}$$

The intensity of the superimposed wave is given by:

$$\begin{aligned}I(x,t) = \phi(x,t)\phi^*(x,t) &= A_1^2 + A_2^2 \\ &+ 2A_1A_2\cos[2\pi(k_1 - k_2)x - (\omega_1 - \omega_2)t + \delta_1 - \delta_2]\end{aligned} \tag{2.22}$$

For an experiment carried out over an extended time T (a typical exposure time T in electron microscopes is in the order of seconds), we observe the average intensity

$$I_{\text{obs}}(x) = A_1^2 + A_2^2 + 2A_1A_2\langle\cos[2\pi(k_1 - k_2)x - (\omega_1 - \omega_2)t + \delta_1 - \delta_2]\rangle_T \tag{2.23}$$

Here, the observed intensity has three terms, the intensities of waves 1 and 2, respectively, plus an interference term in the form of cosine function. This term comes from the interference between the two waves. For this last term, we have several possibilities:

Case 1: $\omega_1 \neq \omega_2$ and $T \gg 2\pi/|\omega_1 - \omega_2|$, where we might call $2\pi/|\omega_1 - \omega_2|$ the "beat period," this being the wavering period we hear, for example, from a slightly out-of-tune piano whose two strings (for the same note) differ in frequency by this amount. (It is interesting to see how this condition, when multiplied by Planck's constant, takes on the form of the energy and time uncertainty principle, as further discussed below.) In this case, the positive and negative contributions of the cosine function cancel each other out, and the overall result is given by

$$I_{\text{obs}}(x) = A_1^2 + A_2^2 \tag{2.20}$$

Case 2: $\omega_1 = \omega_2$ and both δ_1 and δ_2 are constant; in this case, we have

$$I_{\text{obs}}(x) = A_1^2 + A_2^2 + 2A_1A_2\cos[2\pi(k_1 - k_2)x + \delta_1 - \delta_2] \tag{2.21}$$

and the intensity varies with x; the period of variation is determined by the difference between k_1 and k_2.

Case 3: $\omega_1 = \omega_2$ and both δ_1 and δ_2 vary with time randomly; in this case, the positive and negative contributions of the cosine function again cancel each other out, which gives rise to the following result

$$I_{\text{obs}}(x) = A_1^2 + A_2^2$$

Case 4: $\omega_1 \neq \omega_2$ but $T \ll 2\pi/|\omega_1 - \omega_2|$; in this case, the detection time is shorter than the "beat period" within which waves of different frequency can interfere, so we get the same results as cases 2 or 3. (This case is the basis for experiments in which laser light from two different, independent lasers, operating at

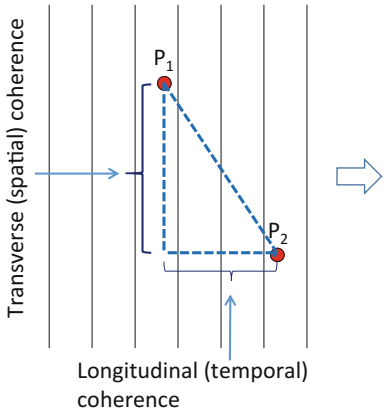

Fig. 2.4 Spatial coherence and temporal coherence between two wave-front points. The temporal coherence and spatial coherence are measured along and normal to the wave propagating direction, respectively

slightly different frequencies, can be shown to interfere if the detection time is short enough.)

Among the 4 cases above, cases 1 and 3, where the intensity of the total wave is simply the sum of the individual waves, are said to be incoherent. Case 2 is said to be coherent. Here, the intensity of the sum of two waves is the sum of individual wave intensities plus their interference effect, which depends on the relative phase difference between the two waves (An example of electron interference is shown in Fig. 2.11.) This interference effect is missing in the incoherent cases.

The above discussion was concerned with two opposite cases, one fully incoherent and the other fully coherent. In practice, wave interference often lies in between these two, which we call partially coherent. The degree of interference can vary continuously, and thus, it is useful to define a measure for the degree of partial coherence. To do so, we consider the general case of interference between two points on a propagating wave front, P_1 and P_2 as shown in Fig. 2.4, which are separated by a vector $\vec{r}$. The waves travel between these two source points and the detector through two different paths. Assuming the first wave arrives at time t and the second one arrives at $t + \tau$, the intensity averaged over a period of time is given by

$$\begin{aligned}\langle I\rangle &= \left\langle |\phi_1(t)+\phi_2(t+\tau)|^2 \right\rangle \\ &= \left\langle |\phi_1(t)|^2 \right\rangle + \left\langle |\phi_2(t+\tau)|^2 \right\rangle + 2\mathrm{Re}\left\{\left\langle \phi_1(t)\phi_2^*(t+\tau)\right\rangle\right\} \\ &= I_1 + I_2 + 2\mathrm{Re}\{\Gamma_{12}(\tau)\}\end{aligned} \tag{2.24}$$

where

$$\Gamma_{12}(\tau) = \left\langle \phi_1(t)\phi_2^*(t+\tau)\right\rangle, \tag{2.25}$$

which is known as the correlation function of the two waves. A correlation function is simply a function which measures the degree of similarity between two functions. When the two waves are the same, for example, $\Gamma_{11}(\tau) = \langle \phi_1(t)\phi_1^*(t+\tau)\rangle$, the function then defines the self-correlation (also known as autocorrelation function). Using these definitions, we rewrite Eq. (2.24) as

$$\langle I\rangle = I_1 + I_2 + 2\sqrt{I_1 I_2}\,\mathrm{Re}\{\gamma_{12}(\tau)\} \tag{2.26}$$

where

$$\gamma_{12}(\tau) = \frac{\Gamma_{12}(\tau)}{\sqrt{\Gamma_{11}(0)}\sqrt{\Gamma_{22}(0)}}. \tag{2.27}$$

The degree of coherence between the two waves (i.e., the extent to which the waves can interfere) can be measured from the maximum intensity and minimum intensity recorded in an interference pattern (such as Young's pinhole experiment), formed when these two waves are allowed to interfere. This is called the "visibility"

$$\upsilon = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}.$$

The intensity of minimum ($I_{\min}$) and maximum ($I_{\max}$) is obtained with $\gamma_{12}(\tau) = -|\gamma_{12}(\tau)|$ and $\gamma_{12}(\tau) = |\gamma_{12}(\tau)|$, respectively. In case $I_1 = I_2$, then $\upsilon = |\gamma_{12}(\tau)|$. For this reason, $\gamma_{12}(\tau)$ is called the complex degree of partial coherence. The normalization against the autocorrelation function in the definition of $\gamma_{12}(\tau)$ ensures that its maximum will not exceed 1 (Born and Wolf 1999). The fully incoherent and coherent cases correspond to $|\gamma_{12}(\tau)| = 0$ and 1, respectively.

Experimentally, the longitudinal and transverse components of $\gamma_{12}(\tau)$ are measured as shown in Fig. 2.4 using interferometry. For the longitudinal component, the source points are selected to be on the same axis along the propagation direction using a Michelson interferometry (Steel 1985). They are selected by the time delay set by the path difference in the interferometer. The value of $\gamma_{12}(\tau)$ obtained in this way gives a measurement of longitudinal coherence, which is also known as temporal coherence. The transverse coherence can be measured by the Young's slit experiment, where the wave arrives at the detector from the two slits simultaneously with $\tau = 0$. This will be further discussed in Sect. 2.9.

2.6 Wave Packets and the Uncertainty Principle

So far, we have only discussed sinusoidal waves, which are continuous, infinite, and monochromatic with a single frequency. The sinusoidal wave can be only an approximation of the actual wave function of the free electrons for two reasons. First, it cannot be normalized since for the sinusoidal wave, we have

$$\int_{-\infty}^{\infty} \phi(\vec{r})\phi^*(\vec{r})d^3\vec{r} = A^2 \int_{-\infty}^{\infty} d^3\vec{r} = A^2(\infty).$$

Secondly, the emitted electrons have a finite energy spread (ΔE) as well as a finite angular distribution. Thus, instead of a single momentum k, there is a range of electron momenta (or $\vec{k}$). Since the sinusoidal wave of Eq. (2.8) is a solution of the wave equation, the wave function of a free electron can be expressed as a superposition of sinusoidal waves according to the principle of wave superposition

$$\phi(\vec{r},t) = \int \phi\left(\vec{k}\right) \exp\left[2\pi i\vec{k}\cdot\vec{r}\right] \exp\left(-\pi i\frac{hk^2}{m}t\right) d^3\vec{k} \tag{2.28}$$

Here, $\phi\left(\vec{k}\right)$ is the complex amplitude of the sinusoidal wave of wave vector $\vec{k}$. The resulting wave function is called a wave packet. This wave function can be normalized in the following way (the mathematics employed here is described in Appendix E)

$$\begin{aligned}\int \phi(\vec{r},0)\phi^*(\vec{r},0)d^3\vec{r} &= \int\int \phi\left(\vec{k}\right)\phi^*\left(\vec{k}'\right)d^3\vec{k}d^3\vec{k}' \int \exp\left[2\pi i\left(\vec{k}-\vec{k}'\right)\cdot\vec{r}\right]d^3\vec{r} \\ &= \int \phi\left(\vec{k}\right)\phi^*\left(\vec{k}'\right)\delta\left(\vec{k}-\vec{k}'\right)d^3\vec{k}d^3\vec{k}' \\ &= \int \phi\left(\vec{k}\right)\phi^*\left(\vec{k}\right)d^3\vec{k} = 1\end{aligned} \tag{2.29}$$

Both $\phi\left(\vec{k}\right)$ and $\phi(\vec{r},t)$ describe the same electron wave function. The only difference is the representation, one in the real space and one in momentum or k space.

At $t = 0$, we have the wave function in integral form, known as a Fourier transform:

$$\phi(\vec{r},0) = \int \phi\left(\vec{k}\right) \exp\left[2\pi i\vec{k}\cdot\vec{r}\right] d^3\vec{k}. \tag{2.30}$$

Inside an electron microscope, the emitted electrons emerge from a source, which can be real or virtual. (For example, the source is virtual when it is placed closer to a convex lens than the focal distance.) If we set $t = 0$ at the source position, then by applying an inverse Fourier transform (Appendix E), we obtain

$$\phi\left(\vec{k}\right) = \int \phi(\vec{r},0)\exp\left[-2\pi i\vec{k}\cdot\vec{r}\right]d^3\vec{r} \tag{2.31}$$

The emitted electrons in the TEM has a small divergence angle of few to tens of milliradians (mrad) and an energy spread of $\sim$0.3 to $\sim$2 eV. A useful approximation is as follows:

$$\phi\left(\vec{k}\right) = A\left(k_x,k_y\right)\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-(k_z-k_\mathrm{o})^2/2\sigma_k^2}, \tag{2.32}$$

where $A\left(k_x,k_y\right) = 1/\pi k_{\max}^2$ for $\sqrt{k_x^2+k_y^2} \le k_{\max}$ and 0 otherwise. This model assumes that momentum is uniformly distributed within a disk along the x direction and y direction and has a Gaussian distribution along the z direction around the mean value of $k_\mathrm{o} = 1/\lambda$.

The Fourier transform in the integral of Eq. (2.31) has the property that a broadly distributed function gives rise to a narrowly distributed Fourier spectrum or vice versa in a reciprocal relationship. For example, consider the Fourier transform of a Gaussian function in the form

$$\mathrm{FT}\left(\mathrm{e}^{-ax^2}\right) = \int_{-\infty}^{\infty}\mathrm{e}^{-ax^2}\mathrm{e}^{-2\pi ikx}\mathrm{d}x = \sqrt{\frac{\pi}{a}}\mathrm{e}^{-\pi^2k^2/a}. \tag{2.33}$$

The standard deviation of the Gaussian function of e^{-ax^2} is $\sigma_x = 1/\sqrt{2a}$. Thus, $\sigma_k = \sqrt{a}/\sqrt{2}\pi$ and $\sigma_x\sigma_k = 1/2\pi$. In general, a wave packet of short duration in size has a broad range of electron momenta, while a sinusoidal wave extending over all space has only a single value of momentum. Such reciprocal relationships are broadly defined in the Heisenberg's uncertainty principles of quantum mechanics, which state that for two incompatible observables, the product of the measurement uncertainties (σ) is greater than the Planck's constant:

$$\begin{matrix}\sigma_x\sigma_p \ge \hbar \\ \sigma_t\sigma_E \ge \hbar\end{matrix}. \tag{2.34}$$

The same relationship also applies to position and momentum along y and z directions, as well as the angles and angular momentums. We note that Eq. (2.34) in the so-called modern representation of the uncertainty principle has $\hbar$ divided by 2. Heisenberg's original formulation had $\hbar$ instead of $\hbar/2$ on the right side of the inequality. Since the uncertainty principle only sets a lower limit, the two formulations only differ in the estimate of this limit. For a Gaussian function, we have $\sigma_x\sigma_k = 1/2\pi$. Since $p = hk$, $\sigma_x\sigma_p = \hbar$ for the Gaussian wave function.

2.7 The Gaussian Wave Packet and Its Propagation

It is instructive to examine the propagation of a Gaussian wave packet as described in Eq. (2.32). At $t = 0$, by converting the 3D integral in k space to cylindrical form, we have

$$\begin{aligned}\phi(\vec{r},0) &= \frac{1}{\sigma_k\sqrt{2\pi}}\int A(k_x,k_y)\mathrm{e}^{-(k_z-k_\mathrm{o})^2/2\sigma_k^2}\exp\left[2\pi i\vec{k}\cdot\vec{r}\right]d^3\vec{k} \\ &= \frac{1}{\sigma_k\sqrt{2\pi}}\frac{1}{\pi k_{\max}^2}\int_0^{k_{\max}}\int_0^{2\pi}\mathrm{e}^{2\pi i k_\parallel\rho\cos\theta}k_\parallel \mathrm{d}k_\parallel \mathrm{d}\theta\left\{\mathrm{e}^{2\pi i k_\mathrm{o} z}\int_{-\infty}^{\infty}\mathrm{e}^{-k^2/2\sigma_k^2}\mathrm{e}^{2\pi i k_z z}\mathrm{d}k_z\right\}\end{aligned} \tag{2.35}$$

where $\rho = \sqrt{x^2+y^2}$ and θ is the angle between $\vec{k}$ and $\vec{\rho}$. Using the result in Eq. (2.33), we have for the integral over k_z

$$\frac{1}{\sigma_k\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{e}^{-k^2/2\sigma_k^2}\mathrm{e}^{2\pi i k_z z}\mathrm{d}k_z = \mathrm{e}^{-2(\pi\sigma_k z)^2}$$

The integral over $k_\parallel$ is carried out first by integrating over θ, which gives the well-known zero-order Bessel function

$$\begin{aligned}\int_0^{k_{\max}}\int_0^{2\pi}\mathrm{e}^{2\pi i k_\parallel\rho\cos\theta}k_\parallel\mathrm{d}k_\parallel\mathrm{d}\theta &= \int_0^{k_{\max}}k_\parallel\mathrm{d}k_\parallel\left[2\pi J_\mathrm{o}(2\pi k_\parallel\rho)\right] \\ &= \frac{1}{2\pi\rho^2}\int_0^{2\pi k_{\max}\rho}xJ_\mathrm{o}(x)\mathrm{d}x\end{aligned}$$

where $x = 2\pi k_\parallel\rho$. Using the integral identity $\int_0^{x_o}xJ_\mathrm{o}(x)\mathrm{d}x = x_\mathrm{o}J_1(x_\mathrm{o})$ and putting the above results together, we obtain the following wave function at $t = 0$

$$\phi(\vec{r},0) = \mathrm{e}^{2\pi i k_\mathrm{o} z}\mathrm{e}^{-2(\pi\sigma_k z)^2}2\left[\frac{J_1(2\pi k_{\max}\rho)}{2\pi k_{\max}\rho}\right] \tag{2.36}$$

Thus, the wave function amplitude falls off away from $z = 0$ according to a Gaussian distribution, while normal to z, the function gives rise to the well-known Airy disk function $[2J_1(x)/x]^2$, which has a maximum at $\rho = 0$ and falls to its first zero at $x = 3.8317$ or $\rho = 0.6/k_{\max}$ and oscillates as ρ increases.

At $t \neq 0$, the electron wave function $\phi(\vec{r},t)$ can be obtained using the propagation of individual plane waves that make up the wave packet, according to the integral in (2.28). This integral can also be separated into two parts similar to

Eq. (2.35). Here, we focus on the propagation along the z direction, for which we have

$$\begin{aligned}
&\frac{1}{\sigma_k\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{e}^{-k_z^2/2\sigma_k^2}\mathrm{e}^{2\pi i k_z z}\exp\left(-\pi i\frac{h(k_z+k_\mathrm{o})^2}{m}t\right)\mathrm{d}k_z \\
&\quad=\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left(-\pi i\frac{hk_\mathrm{o}^2}{m}t\right)\int_{-\infty}^{\infty}\mathrm{e}^{-ak_z^2}\mathrm{e}^{2\pi i k_z z'}\mathrm{d}k_z \\
&\quad=\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left(-\pi i\frac{hk_\mathrm{o}^2}{m}t\right)\sqrt{\frac{\pi}{a}}\mathrm{e}^{-\pi^2 z'^2/a}
\end{aligned}$$

where

$$z'=z-\frac{hk_\mathrm{o}}{m}t$$

and

$$a=\frac{1}{2\sigma_k^2}\left(1+i\frac{2\pi\sigma_k^2 ht}{m}\right)=\frac{1}{2\sigma_k^2}\left(\frac{1+\Gamma^2t^2}{1-i\Gamma t}\right) \tag{2.37}$$

With $\Gamma=\frac{2\pi\sigma_k^2 h}{m}$.

Putting the above results together, we have along the z-axis for the Gaussian wave packet

$$\phi(z,t)=C\sqrt{\frac{1-i\Gamma t}{1+\Gamma^2t^2}}\exp\left[-2\pi^2\sigma_k^2\left(z-\frac{hk_\mathrm{o}}{m}t\right)^2/\left(1+\Gamma^2t^2\right)\right]\mathrm{e}^{2\pi i k_\mathrm{o}z} \tag{2.38}$$

Here, C contains two phase terms, one is simply the phase oscillation with time according to the average frequency, and the other comes from the complex term in Eq. (2.37).

The wave packet is thus again described by a Gaussian function, and its center moves with the so-called group velocity hk_o/m. Its width increases with time according to $\sqrt{2}\pi\sigma_k/\sqrt{1+\Gamma^2t^2}$. The rate of wave packet broadening is given by Γ.

As we will see in later chapters, the relationship between the frequency and wave vector $\omega\left(\vec{k}\right)$ is modified by electron interaction with the potential, and it becomes a general function of $\vec{k}$ in a complex form. In such cases, as long as the frequency spectrum is narrow, $\omega\left(\vec{k}\right)$ can be approximately expanded around the mean in a Taylor series of

$$\omega(k) = \omega(k_o) + (k - k_o)\omega'(k_o) + \frac{1}{2}(k - k_o)^2\omega''(k_o) + \cdots$$
$$= 2\pi\left[k_o v_p + (k - k_o)v_g + \frac{1}{2}(k - k_o)^2\Gamma + \cdots\right]$$

where $v_p = \omega(k)/2\pi k$ is the phase velocity, $v_g = (d\omega(k)/dk)/2\pi$ is the group velocity, and $\Gamma = (d^2\omega(k)/dk^2)/2\pi$ gives the dispersion.

In the case of a free electron, $v_p = E/hk = hk/2m$ and $v_g = (dE/dk)/h = hk/m$. Thus, the quantum mechanical phase velocity is half the speed of classical particles, while the group velocity is the same as the classical speed.

2.8 Temporal Coherence

A direct consequence of having an electron wave packet of finite length is the limited temporal coherence $\gamma_{12}(\tau)$, as defined in Eq. (2.27). To calculate $\gamma_{12}(\tau)$ for an electron wave packet, we approximate its wave function using the model of the quasi-monochromatic wave

$$\phi(z,t) = \phi_o \exp[2\pi ikz]\exp(-i\omega t)\exp(-i\delta(t)) \tag{2.41}$$

Here, the phase $\delta(t)$ is taken as constant within a coherence time τ_o, e.g., $\delta(t) = \Delta_n$ for the time period of $n\tau_o \le t < (n+1)\tau_o$. However, from one coherent period to next, the value of Δ_n fluctuates randomly in the so-called random phase approximation.

The quasi-monochromatic wave applies to the experimental case when the electron momenta spread along the x and y directions are very small, and the emitted electrons are far from each other in time. Thus, the probability of having two or more electrons emitted within the coherence time is very small.

Assuming the amplitudes of two waves are the same, we obtain from the above wave function

$$\gamma_{12}(\tau) = \frac{\Gamma_{12}(\tau)}{\sqrt{\Gamma_{11}(0)}\sqrt{\Gamma_{22}(0)}} = \langle \exp(-i\omega\tau)\exp(i[\delta(t) - \delta(t+\tau)])\rangle$$
$$= \exp(-i\omega\tau)\lim_{T\to\infty}\frac{1}{T}\int_0^T \exp(i[\delta(t) - \delta(t+\tau)])\mathrm{d}t \tag{2.39}$$

To evaluate the integral in Eq. (2.39), we first consider the case of $\tau > \tau_o$; e.g., the time delay exceeds the coherence time. The phase difference is then given by the difference between two random phases, which gives another random phase, so that $\delta(t) - \delta(t+\tau) = \Delta$. The integral over random phases averages to zero, and thus, $\gamma_{12}(\tau) = 0$ for $\tau > \tau_o$.

For $0<\tau<\tau_o$, we have the following two scenarios

$$\delta(t)-\delta(t+\tau)=\begin{cases}0, & 0<t<\tau_o-\tau\\ \Delta_n-\Delta_{n+1}, & \tau_o-\tau<t<\tau_o\end{cases}$$

This applies to every coherent period in the quasi-monochromatic wave. By summing up all coherent periods, we obtain the integral of Eq. (2.39) in the form

$$\begin{aligned}\gamma_{12}(\tau)&=\exp(-i\omega\tau)\lim_{N\to\infty}\frac{1}{N\tau_o}\left\{\sum_{n=0}^{N}\int_0^{\tau_o-\tau}\mathrm{d}t+\int_{\tau_o-\tau}^{\tau_o}\exp[i(\Delta_n-\Delta_{n+1})]\mathrm{d}t\right\}\\&=\exp(-i\omega\tau)\frac{1}{\tau_o}\int_0^{\tau_o-\tau}\mathrm{d}t\\&=\left(1-\frac{\tau}{\tau_o}\right)\exp(-i\omega\tau)\end{aligned}\tag{2.40}$$

for $0<\tau<\tau_o$. Here, the second sum over the random phases averages to zero.

Together, the above results show that the visibility of interference fringes $|\gamma_{12}(\tau)|$ decreases linearly with the delay time and disappears beyond the coherence time. The path difference between the two waves thus must be smaller than $L=v\tau_o$ in order to observe the interference between the two waves.

ForThe longitudinal coherence between two waves of slightly different wavelengths ($\lambda_1=\lambda$ and $\lambda_2=\lambda-\Delta\lambda$) is defined as the length over which the two waves become completely out of phase with each other (e.g., by 180°) as shown in Fig. 2.5. According to this definition

$$2L_C=N\lambda=(N+1)(\lambda-\Delta\lambda)$$

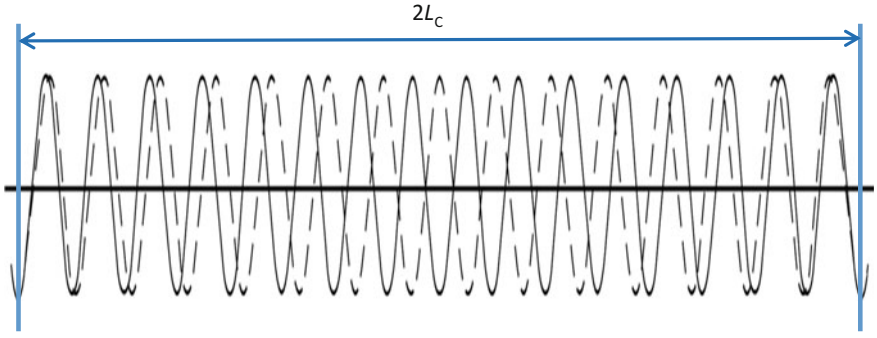

Fig. 2.5 Longitudinal coherence between two waves of different wavelengths (After D. Attwood, University of California, Berkeley)

where N is the number of periods where the two waves become in-phase again. From this, we obtain

$$N+1 \approx N = \frac{\lambda}{\Delta\lambda}$$

and

$$L_C = \frac{\lambda^2}{2\Delta\lambda} \tag{2.41}$$

For the electron wave packet, it can be shown the same form is obtained by taking

$$L_C = \frac{1}{2} v\Delta t = \frac{\lambda^2}{2\Delta\lambda} = \frac{\lambda E}{2\Delta E}$$

where Δt is the uncertainty time obtained from the uncertainty principle. Thus, if we consider the coherence time is same as the uncertainty time ($\tau_o = \Delta t$), at L_C, we expect 50 % contrast in the interference. This expression for the temporal coherence of electron beams was first investigated by Mollenstedt and Ducker (1955).

2.9 Spatial Coherence

In defining the spatial, or transverse, coherence length (or width), we first consider two waves of the same wavelength, originating from two separate source points in space as shown in Fig. 2.6. Each gives rise to a set of wave fronts. Using the same criterion for

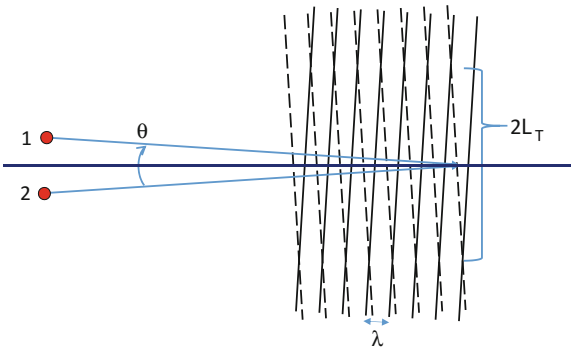

Fig. 2.6 Transverse coherence

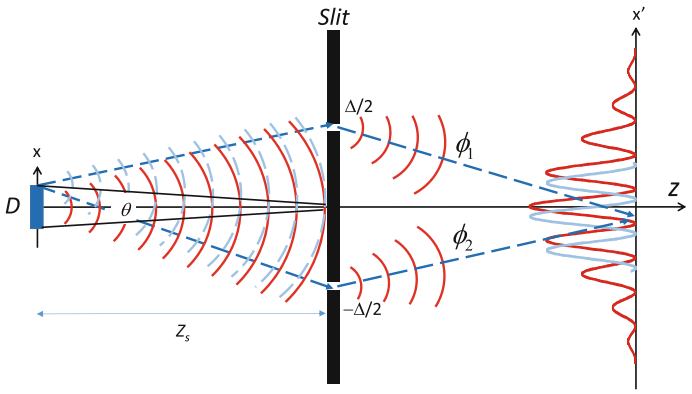

Fig. 2.7 Young's double-slit interference experiment for measurement of transverse coherence length

the definition of temporal coherence length, the transverse coherence length is defined as the lateral distance along a wave front over which there is a complete dephasing between the two waves. For small θ, according to Fig. 2.6, we have

$$2L_T \approx \frac{\lambda}{\theta}$$

Thus, the transverse coherence length is inversely proportional to the angle sustained by the two source points.

To generalize the above discussion for two discrete source points to a source of finite area, we consider the Young's double-slit interference experiment for the measurement of degree of transverse coherence. For simplicity, we will consider the two-dimensional case first, as shown in Fig. 2.7. Two narrow slits of the same width separated by a distance Δ are placed symmetrically relative to the source. The slits are illuminated by a one-dimensional source of finite width with intensity distribution $I(x)$. Interference is observed at the detector between the two waves selected by the two slits. Considering that electrons are typically emitted from areas about the size of an atom, and the source dimension is much larger than an atom, most electron sources can be considered to be ideally incoherent, that is, to consist of a statistically independent close-packed array of emitters. For such an incoherent extended source, each atomic point-source point generates an independent interference pattern at the detector. What is recorded then is the sum of the intensities of the interference patterns generated by each independent source point. To put this in mathematical form, we take a source point at x, so that the two waves at the slits can be written in the form

$$\begin{aligned}\phi_1 &= A(x)\exp(2\pi ikz)\exp(-2\pi ik_x x')\exp(-i\delta_1)\\ \phi_2 &= A(x)\exp(2\pi ikz)\exp(2\pi ik_x x')\exp(-i\delta_2).\end{aligned} \tag{2.42}$$

where δ_1 and δ_2 are the phases of the two waves at the slits, and their difference is determined by the path difference from the source point to the two slits, which is simply $\delta_2 - \delta_1 = (2\pi/\lambda)x\Delta/Z_S$. For what follows, we will consider only the interference recorded at the center of the detector ($x' = 0$). In this case, the arrival time to the detector point is same from each slit, and thus, $\tau = 0$. According to Eq. (2.26), the intensity contribution from the source point at x is given by

$$\langle I(x)\rangle \mathrm{d}x = 2I(x)\left[1 + \mathrm{Re}\left\{\gamma'_{12}(x)\right\}\right]\mathrm{d}x$$

where

$$\gamma'_{12}(x) = \exp(2\pi ix\Delta/\lambda Z_S).$$

The overall complex degree of partial coherence is obtained by integrating over all source points, in the form

$$\gamma_{12}(\Delta, 0) = \int_{-\infty}^{\infty} I(x)\exp(2\pi ix\Delta/\lambda Z_S)\mathrm{d}x \tag{2.43}$$

Extending this to a two-dimensional source, we have

$$\gamma_{12}\left(\Delta_x, \Delta_y, 0\right) = \frac{1}{I_o}\int_{-\infty}^{\infty} I(x,y)\exp\left[2\pi i\left(x\Delta_x + y\Delta_y\right)/\lambda Z_S\right]\mathrm{d}x\mathrm{d}y \tag{2.44}$$

where $I_o = \int_{-\infty}^{\infty} I(x,y)\mathrm{d}x\mathrm{d}y$ is the integrated source intensity. The above result is known as Van Cittert–Zernike theorem in optics (Born and Wolf 1999). According to this theory, the wave front from a small incoherent source will appear mostly coherent at a large source distance (Z_s). An intuitive explanation of this phenomenon can be provided based on the uncertainty principle. Considering first the range of electron momenta being measured, since the electrons are confined within the source before their emission, their directions are thus determined by the source angle, sustained by the source over the detector point. The range of electron momenta is proportional to the source angle, and thus, it is large for a detector placed close to the source. Intensity recorded by the detector is dominated by the contribution coming from the closest source point, which can be determined with a high degree of accuracy according to the uncertainty principle. For a detector placed far from the source, the momentum distribution is small, and our measurement will

no longer be able to distinguish the contributions from specific source points. Consequently, all source points appear to be same, and they contribute almost equally at a large source distance.

For a one-dimensional source of size D, $I(x) = 1$ for $|x| \leq D/2$ and zero elsewhere. The Fourier transform of Eq. (2.43) gives the following result

$$\gamma_{12}(\Delta, 0) = \frac{1}{D} \int_{-D/2}^{D/2} \exp(2\pi i x \Delta / \lambda Z_S) \mathrm{d}x = \frac{\sin(\pi D \Delta / \lambda Z_S)}{\pi D \Delta / \lambda Z_S},$$

and the first zero of $\gamma_{12}(\Delta, 0)$ occurs at $D\Delta/\lambda Z_S = 1$ or

$$\Delta = \lambda Z_S / D = \lambda / (D/Z_S) = \lambda / \theta$$

At the length of $L_\mathrm{T} = \lambda/2\theta$, we have

$$\gamma_{12}(\Delta/2, 0) = \frac{\sin(\pi/2)}{\pi/2} = 0.64.$$

To summarize, the coherence width at the TEM sample (the largest distance between points across the beam at which waves will interfere) is about $\lambda/2\theta$. Highly coherent conditions therefore require that the (ideally incoherent) source subtends a small angle θ at the sample (i.e., a well-collimated beam). This can be achieved by placing a large source (such as a star) at a very large distance, with the result that the detected intensity, although coherent, becomes weak. The coherence width from sunlight is in fact about a tenth of a millimeter. This trade-off between spatial coherence and intensity is of great importance in designing radiation sources, from synchrotrons to electron microscopes and neutron facilities, and is quantified by the quantity of emittance, which is the product of source area and the solid angle subtended by the source at the sample. Low emittance, coupled with high brightness, is a major goal for accelerator physicists and electron-optical designers.

2.10 Electron Refraction and the Refractive Index

Another wave property is refraction. Although we will see in later chapters that the effects of electron multiple scattering by a crystal can lead to multiple electron wave vectors inside the crystal, here we will ignore these effects and consider only the mean "refractive index" effect of the average electrostatic potential of the sample. The electron speeds up as it enters the sample, being attracted to the positive atomic nuclei, thereby gaining kinetic energy and a longer wave vector. We define V_o as the mean sample potential, in volts. Then, the magnitude K of the mean wave vector inside the sample is given by

$$K^2 = k_o^2 + 2meV_o/h^2 = k_o^2 + U_o$$

so that

$$K \approx k_o + U_o/2k_o \tag{2.45}$$

Here, U_o has the dimension of length^{-2}. Values of V_o are tabulated for various crystals [Appendix F and Okeeffe and Spence (1994); Kim et al. (1998); Kruse et al. (2006)].

The change in wavelength as the wave travels across the interface gives the well-known refraction effect, similar to the application of Snell's law in optics. By analogy with optics, an electron refractive index n can be defined using the approximation of Eq. (2.45) and Eqs. A.6 and A.7 in Appendix A:

$$n = K/k_o \approx 1 + U_o/2k_o^2 = 1 + U_o\lambda^2/2 = 1 + \frac{V_o}{\Phi}\frac{1 + e\Phi/m_ec^2}{1 + e\Phi/2m_ec^2}, \tag{2.46}$$

For $\Phi = 200$ kV and $V_o = 20$ V, n = 1.0000997. Thus, for high-energy electrons with small wavelength, the refractive index is close to 1. Since the electron refractive index is larger than unity (as for light, but unlike X-rays), an electron beam entering a surface at a low angle is bent toward the normal. The refraction effect is most noticeable when the electron beam is at a glancing angle to the sample surface. In that case, the diffraction condition depends strongly on U_o, but for most transmission microscopy on untilted slab samples, we will see that it is the components of the wave vectors K and k_o normal to sample surface which determine the diffraction condition, and these are approximately equal.

2.11 Wave Propagation

2.11.1 Huygens–Fresnel Principle

We start by asking how electron waves propagate from one plane to another in space. Or, if we know the wave amplitude and phase in plane A, how to calculate the wave in the plane A', which is at a distance of Z downstream, as shown in Fig. 2.8, illustrates.

Mathematically, the above problem can be solved using either a Green's function in quantum mechanics, or Kirchhoff's formulation of the solution of the wave equation. We will not repeat this approach here, but instead will follow the more intuitive approach based on Huygens' principle of wave propagation. Huygens proposed that propagation of waves in space involves the generation of spherical waves at every point on the wave front; these secondary waves propagate in the forward direction and form an envelope that becomes a new wave front. This

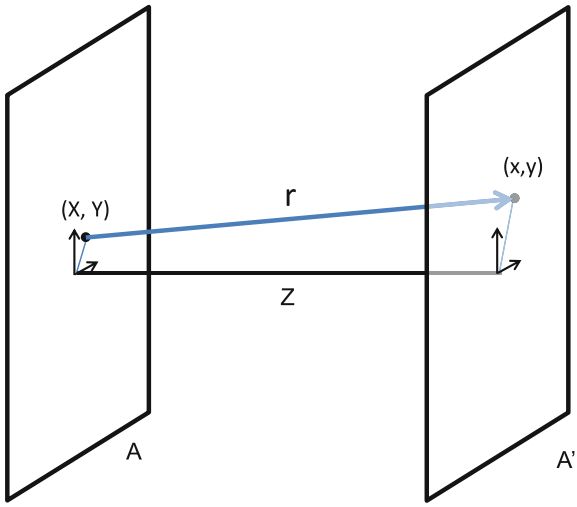

Fig. 2.8 Wave propagation from plane A to plane A', separated by distance Z. The (X, Y) and (x, y) are coordinates on A and A', respectively

principle was further developed by Fresnel. By combining Huygens' principle with the principle of interference, Fresnel was able to explain both the rectilinear propagation of light and also diffraction effects. To achieve this, Fresnel made a number of assumptions about the wave amplitude and phase. Together, the Huygens–Fresnel principle states:

(1) the wave front can be divided into small, finite sized, elements with each element acts the source of a secondary spherical wave;
(2) The secondary waves interfere with each other at point P of a distance Z away according to the principle of superposition;
(3) Each element contributes an amount of wave proportional to its wave amplitude ϕ_o and area dS;
(4) The contribution at point P is inversely proportional to the distance times the wavelength;
(5) There is an obliquity factor of $(1 + \cos\theta)/2$ to the contribution, which is 1 in the forward direction ($\theta = 0$) and zero in reverse direction ($\theta = \pi$); and
(6) The secondary waves vibrate at a quarter of the wavelength behind the primary disturbance.

Putting these points together mathematically, an area dS on the wave front contributes to the wave function at P by the amount

$$\mathrm{d}\phi_P = -i\frac{\phi_o}{r\lambda}\left(\frac{1+\cos\theta}{2}\right)\mathrm{e}^{2\pi ikr}\mathrm{d}S \tag{2.47}$$

and the total contribution at point P is an integration over the surface area of the wave front:

$$\phi_P = -i \iint_S \frac{\phi_o}{r\lambda} \left(\frac{1 + \cos\theta}{2} \right) e^{2\pi i k r} dS \tag{2.48}$$

The assumptions that lead to Eq. (2.48) emerge automatically in Kirchhoff's diffraction formula, in the form of an integral solution to the wave equation. A formal derivation of Eq. (2.48) based on Kirchhoff's integral can be found in Longhurst (1986) and Cowley (1995). For a comprehensive treatment and comparison of these results with the first Born approximation and reconciliation of their superficially different dependence on wavelength and the ninety degree scattering phase (factor i in Eq. 2.51), see the 7th edition of Born and Wolf (1999). For electron waves propagating in the forward direction inside an electron microscope, the small-angle approximation, as used in the paraxial equation of Chap. 6, is assumed, and the obliquity factor is taken to be approximately unity.

2.11.2 Propagation of Plane Wave and Fresnel Zones

We now apply the Huygens–Fresnel principle to the propagation of plane waves. The secondary waves are expected to produce another planar wavefront; thus, the result is known. By going through this exercise, we will introduce the concept of Fresnel zones in wave propagation. This concept is often employed in the calculation of electron diffraction from crystals containing defects. It is fundamental for X-ray focusing using zone plates.

A plane wave is represented by

$$\phi = e^{2\pi i \vec{k} \cdot \vec{r}} \tag{2.49}$$

Here, the wave vector $\vec{k}$ is taken to be along Z as shown in Fig. 2.9. According to (2.47), each point in the planar wave front acts as a secondary source. The contribution from this point depends on the distance of r and $\cos\theta$; both are rotationally symmetrical around the Z-axis. Thus, the integral of (2.48) can be carried out in a circular area of

$$d\sigma = 2\pi\rho d\rho$$

For reasons that will later become clear, we divide the planar wave front into circular zones numbered from 1 to infinity. Inside each zone, the change in the angle is small. Then, the contribution from this zone is approximately given by

$$\phi_n \approx -i \left(\frac{1 + \cos\theta_n}{2} \right) \int_{r_{n-1}}^{r_n} \frac{1}{r\lambda} e^{2\pi i k r} 2\pi\rho d\rho \tag{2.50}$$

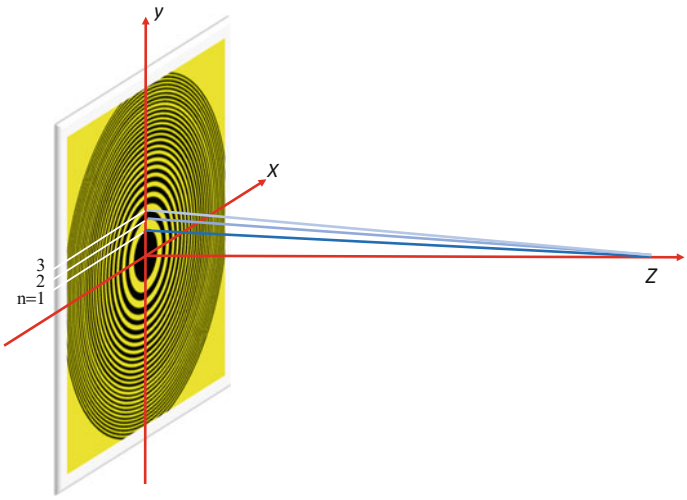

Fig. 2.9 The definition of Fresnel zones

To evaluate Eq. (2.50), we use the relationship

$$r^2 = z^2 + \rho^2$$

and

$$2r\mathrm{d}r = 2\rho\mathrm{d}\rho \tag{2.51}$$

Substituting (2.51) into (2.50), we obtain

$$\phi_n = -\left(\frac{1+\cos\theta_n}{2}\right)\left(\mathrm{e}^{2\pi ikr_n} - \mathrm{e}^{2\pi ikr_{n-1}}\right) \tag{2.52}$$

Next, we define the zone radius in such a way that the radius of two neighboring zones differs only by half a wavelength such that (also see Fig. 2.9)

$$r_n = z + n\lambda/2 \tag{2.53}$$

The zones defined this way are called Fresnel zones. Putting the above radius into (2.52), we have

$$\phi_n = (1+\cos\theta_n)(-1)^{n+1}\mathrm{e}^{2\pi ikz}$$

The contribution from the nth zone, ϕ_n, is positive or negative depending on whether n is odd or even. The total wave obtained at the point p from N Fresnel zones is the sum of contributions from all included zones:

$$\phi_P = |\phi_1| + (|\phi_3| - |\phi_2|) + \cdots + (|\phi_{N-1}| - |\phi_{N-2}|) - |\phi_N|$$

Here, N is an even number. When $Z \gg \lambda$, to a very good approximation, the terms inside each bracket are approximately equal and they cancel each other. Then, one may write

$$\phi_P = |\phi_1| - |\phi_N|$$

When N is taken to be a very large number, the angle θ approaches 90°, and we have $\phi_1 = 2e^{2\pi ikz}$, $\phi_N = -e^{2\pi ikz}$, and

$$\phi_P = \frac{\phi_1}{2} = e^{2\pi ikz} \tag{2.54}$$

Thus, using Huygens–Fresnel principle we have successfully demonstrated that the secondary waves in a plane wave front give rise to another plane wave further away, as we expected. Equation (2.54) also shows that the wave function at the point P is half of the contribution from the first Fresnel zone, while contributions from the rest of zones cancel the other half.

2.11.3 *Fresnel Diffraction—The Near-Field Small-Angle Approximation*

For wavelengths much smaller than the size of the object (L), the range of diffraction angles, which can be estimated by the uncertainty principle, $\Delta k/k \sim 1/kL$, is small. Thus, the distance between the source point (X, Y) on plane A and the detection point (x, y) on plane A' can be approximated by

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

If we assume that the wave distribution on plane A is $\phi_e(X, Y)$, the wave function at a distance z away is then approximately given by:

$$\phi(x, y) = -i\frac{e^{2\pi ikz}}{z\lambda} \iint \phi_e(X, Y) e^{\frac{\pi i}{\lambda z}\left[(x-X)^2 + (y-Y)^2\right]} dX dY, \tag{2.55}$$

which is called the Fresnel propagation equation. This equation can be used to explain a class of electron diffraction effects since the electron wavelengths employed in TEM are smaller than an atom. Fresnel diffraction can be observed directly in a TEM around the edge of a sample or an aperture in imaging mode, while Fraunhofer diffraction, to be discussed later, is observed at large distances or in the back focal plane of the objective lens of a TEM.

Next, we examine Fresnel diffraction from the edge of an aperture and the resulting Fresnel integral. These fringes are often used to correct for astigmatism in the electron microscope. To proceed, we consider the case of an opaque aperture with a straight edge. Suppose that this aperture covers half of the space in the x direction and the incident wave is a plane wave propagating along the z direction with $\phi_e(X, Y) = 1$ for $X > 0$, where the aperture is absent. Substituting this into (2.55), we obtain the following wave function:

$$\phi(x, y) = -i\frac{e^{2\pi ikz}}{z\lambda}\int_0^{\infty}\int_{-\infty}^{\infty} e^{\frac{\pi i}{\lambda z}\left[(x-X)^2+(y-Y)^2\right]}dXdY \tag{2.56}$$

The integral in (2.56) is known as the Fresnel integral, which has the general form of (Fig. 2.10)

$$\int_0^s e^{\frac{\pi i}{2}s^2}ds = \int_0^s \cos\left(\frac{\pi}{2}s^2\right)ds + i\int_0^s \sin\left(\frac{\pi}{2}s^2\right)ds = \overline{X} + i\overline{Y}$$

The complex value of the Fresnel integral can be plotted in 2D with the y-axis representing the imaginary part and the x-axis for the real part, and the result defines a curve known as Cornu's spiral. A plot of the Cornu's spiral is shown in Fig. 2.10. The s value is marked on the top part of the curve. The two spirals have an inversion symmetry with the center of two spirals at the s limit of positive and negative infinite. At these two limits, both $\bar{X}$ and $\bar{Y}$ approach the value of ±½. Using this for (2.56), we have

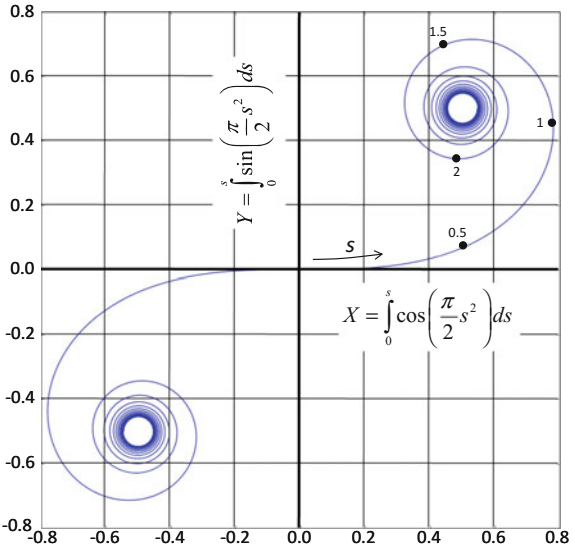

Fig. 2.10 Cornu's spiral. The two spirals converge to ±(1/2, 1/2) as s moves from 0 to ±∞

$$\phi(x,y) = c\int_0^\infty e^{\frac{\pi i}{\lambda z}(x-X)^2}dX = c\int_{s_1}^\infty e^{\frac{\pi i}{2}s^2}ds = c\left(\int_0^\infty e^{\frac{\pi i}{2}s^2}ds - \int_0^{s_1} e^{\frac{\pi i}{2}s^2}ds\right)$$
$$= c\left[\left(\frac{1}{2}+i\frac{1}{2}\right) - (\bar{X}+i\bar{Y})\right] \quad (2.57)$$

Here, $s = \sqrt{2(x-X)^2/\lambda z}$ and $s_1 = \sqrt{2/\lambda z}x$. For $x > 0$, s_1 travels along the bottom left part of the spiral. Since the diffracted wave intensity $I(x,y) = |\phi(x,y)|^2$ is proportional to the square of the vector length from s_1 to (1/2, 1/2), the intensity oscillates as s_1 goes around the spiral and approaches—∞. On the other side, for $x < 0$, s_1 travels along the top right part of the spiral and the length decreases continuously. Figure 2.11 shows the intensity calculated from (2.57), and the oscillation expected for $x > 0$ and the monotonic decrease expected for $x < 0$ are both reflected in this plot.

Fresnel fringes are observed in an out of focus electron image. Figure 2.12 shows an example. It is an out of focus image of a sharp W tip (Beleggia et al. 2014). The tip is placed near a negatively biased electrode (−90 V) at a distance away. Electron wave propagation around the W tip creates the Fresnel fringes. They are deflected by the tip electric fields, giving rise to the shape of deflected wave, and the interference of the deflected waves is also shown in Fig. 2.12. The number of Fresnel fringes observed in an out of focus electron image is ultimately determined by the coherence of the electron source. A 300 kV FEI TEM equipped with a field emission gun was used to record the electron image here. The coherence properties of different electron sources and electron illumination are discussed in Chaps. 8 and 10.

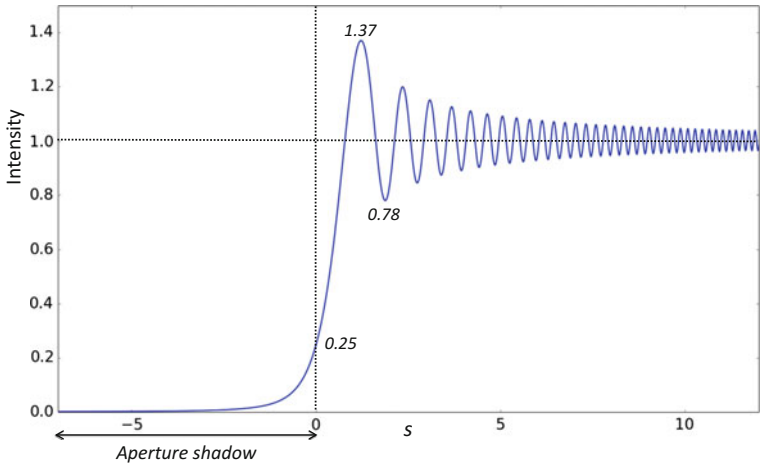

Fig. 2.11 Fresnel diffraction intensity from an opaque aperture with a straight edge at $s = 0$

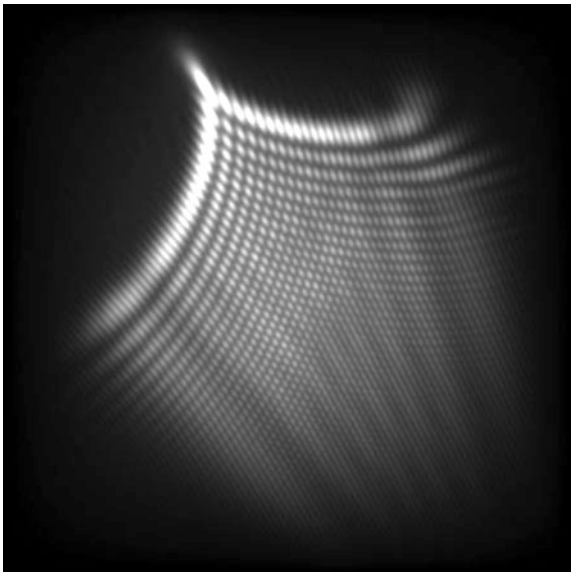

Fig. 2.12 Experimental image of Fresnel fringes and their interference formed by the propagation of coherent electron waves through the electric fields created by a sharp W tip and a negatively biased electrode (Beleggia et al. 2014). The image is approximately 6 mm away from the tip and electrode with a field view of 1 μm. The pattern of interference fringes results from overlap of waves from either side of the tip. (Image provided by Rafal Dunin–Borkowski, Ernst Ruska-Centre for Microscopy and Spectroscopy, Jülich, Germany)

2.11.4 *Fraunhofer Diffraction—Far-Field Forward Diffraction*

With the electron detector placed in the far field, and the detector size much larger than the extent of the object, to a good approximation for electron diffraction, we have

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

and

$$kr = \sqrt{z^2 + (X-x)^2 + (Y-y)^2} \approx kr - \frac{x}{\lambda R}X - \frac{y}{\lambda R}Y$$

Most of electron diffraction occurs in the forward direction with $\cos\theta \approx 1$. Then, the integral of (2.48) can be simplified into the so-called Fraunhofer diffraction equation:

$$\phi_P = \frac{-i\mathrm{e}^{2\pi ikr}}{R\lambda}\iint_S \phi_\mathrm{o}(X,Y)\mathrm{e}^{-2\pi i\left(k_x X + k_y Y\right)}\mathrm{d}X\mathrm{d}Y \tag{2.58}$$

With $k_x = \frac{x}{\lambda R}$ and $k_y = \frac{y}{\lambda R}$.

The Fraunhofer diffraction equation can be compared with the kinematical theory of diffraction in Chap. 4. Both involve the same type of integral known as a

Fourier transform. The difference is that in the kinematic theory of electron diffraction, an assumption is made about the nature of the electron scattering (that there is no multiple scattering), whereas the Fraunhofer diffraction equation simply relates the wave function at the far field to the exit-face wave function across the downstream face of the sample.

References

Beleggia M, Kasama T, Larson DJ, Kelly TF, Dunin-Borkowski RE, Pozzi G (2014) Towards quantitative off-axis electron holographic mapping of the electric field around the tip of a sharp biased metallic needle. J Appl Phys 116:024305

Born M, Wolf E (1999) Principles of optics: electromagnetic theory of propagation, interference and diffraction of light, 7th edn. Cambridge University Press, Cambridge

Cowley JM (1995) Diffaction physics, 3rd edn. Elsevier Science, Amsterdam

Goodman J (2004) Introduction to fourier optics, 3rd edn. Roberts and Company Publishers, Englewood

Griffiths DJ (2004) Introduction to quantum mechanics, 2nd edn. Pearson Prentice Hall, Upper Saddle River

Kim MY, Zuo JM et al (1998) Ab-initio LDA calculations of the mean Coulomb potential V_o in slabs of crystalline Si, Ge and MgO. Phys Status Solidi A 166:445–451

Kruse P, Schowalter M, Lamoen D, Rosenauer A, Gerthsen D (2006) Determination of the mean inner potential in III-V semiconductors, Si and Ge by density functional theory and electron holography. Ultramicroscopy 106:105–113

Longhurst RS (1986) Geometrical and physical optics, 3rd edn. Orient BlackSwan

Mollenstedt G, Ducker H (1955) Fresnelscher Interferenzversuch mit einem Bi-prisma fur Electronenwellen. Naturwissenschaften 42:41

O'Keeffe M, Spence (1994) On the average coulomb potential and constraints on the electron density in crystals. Acta Cryst A50:33–45

Steel WH (1985) Interferometry, 2nd edn. Cambridge University Press, Cambridge

http://www.springer.com/978-1-4939-6605-9

Advanced Transmission Electron Microscopy
Imaging and Diffraction in Nanoscience
Zuo, J.M.; Spence, J.C.H.
2017, XXVI, 729 p. 310 illus., 218 illus. in color.,
Hardcover
ISBN: 978-1-4939-6605-9