Introduction to Photonics

Frank Cichos

2025-04-10

Table of contents

Chapter 1

Photonics

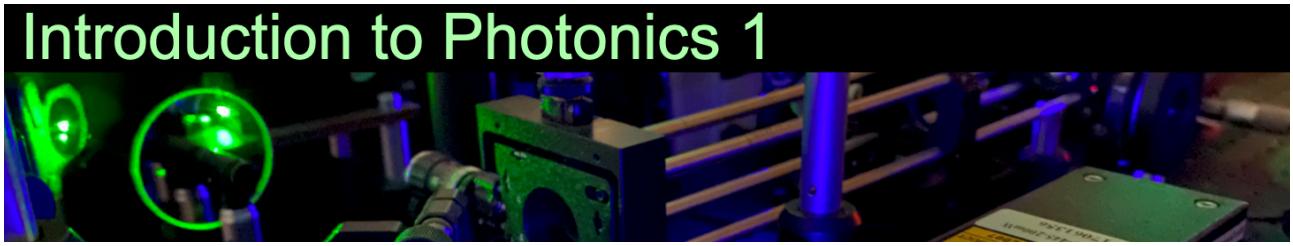


Figure 1.1: Photonics Logo

Photonics is a field of science that is manipulating the flow of light. It contains many facets of research involving light propgation from fundamentals involving light matter interaction to applications involving photonic computing with disordered media or single light quanta to adaptive superresolution microscopy. It is one of the fastest growing fields.

In this course we will introduce into the field of optics and photonics. We will start with simple but powerful descriptions of light propagation using ray optics to more advanced physics using electromagnetic waves. We will explore Fourier optics, anisotropic media and non-linear optics to lay the foundation to more complex topics in advanced lecture series.

Part I

Lecture 1

Chapter 2

Theories for light

Light has been described through increasingly sophisticated theoretical frameworks throughout the history of physics. The simplest framework is *Ray Optics* or *Geometrical Optics*, which treats light as rays traveling along straight paths and applies geometrical principles to describe interactions with optical elements like lenses and mirrors. Moving beyond this approximation, *Wave Optics* introduces the wave nature of light, explaining phenomena such as interference and diffraction that ray optics cannot address. *Electromagnetic Optics* further refines our understanding by treating light as electromagnetic waves governed by Maxwell's equations, providing a complete classical description of light-matter interactions. For intense light sources, *Nonlinear Optics* becomes essential, describing how materials respond nonlinearly to strong electromagnetic fields, giving rise to frequency conversion and other novel effects. Finally, at the most fundamental level, *Quantum Optics* treats light as consisting of photons—quantum mechanical particles exhibiting both wave and particle properties—essential for understanding phenomena like spontaneous emission, entanglement, and the quantum nature of light-matter interactions. This course will progressively build your understanding through these increasingly sophisticated frameworks.

Ray Optics

Ray optics, or geometric optics, provides a powerful framework for understanding light propagation when the wavelength is much **smaller than the dimensions of optical elements** involved. In this approach, light travels along straight lines called rays in homogeneous media, with well-defined paths that can be mathematically traced. This description serves as the foundation for analyzing many optical systems, from simple mirrors to complex microscopes and telescopes.

Fermat's Principle: Integral and Differential Forms

Fermat's Principle forms one of the foundations of ray optics, stating that light travels along the route that takes the total optical path length between any two points to an extremum (commonly a minimum). This optical path length, expressed mathematically as $\int_C n(s)ds$, represents the effective distance light traverses through media of varying refractive indices. When this quantity is divided by the vacuum speed of light c_0, it yields the total travel time required for light to journey between those points.

In its integral form:

$$\delta \int_C n(s)ds = 0$$

where $n(s)$ is the refractive index along path C and ds is the differential path length.

The same principle can be expressed as a differential equation that describes how light bends in media with varying refractive indices:

$$\frac{d}{ds}\left(n\frac{d\mathbf{r}}{ds}\right) = \nabla n$$

This equation shows that rays bend toward regions of higher refractive index. In homogeneous media ($\nabla n = 0$), it simplifies to $\frac{d^2\mathbf{r}}{ds^2} = 0$, confirming that light follows straight lines.

Optical Laws Derived from Fermat's Principle

Reflection: At a planar interface, Fermat's Principle directly yields the law of reflection:

$$\theta_i = \theta_r$$

where θ_i is the angle of incidence and θ_r is the angle of reflection, both measured from the normal to the surface.

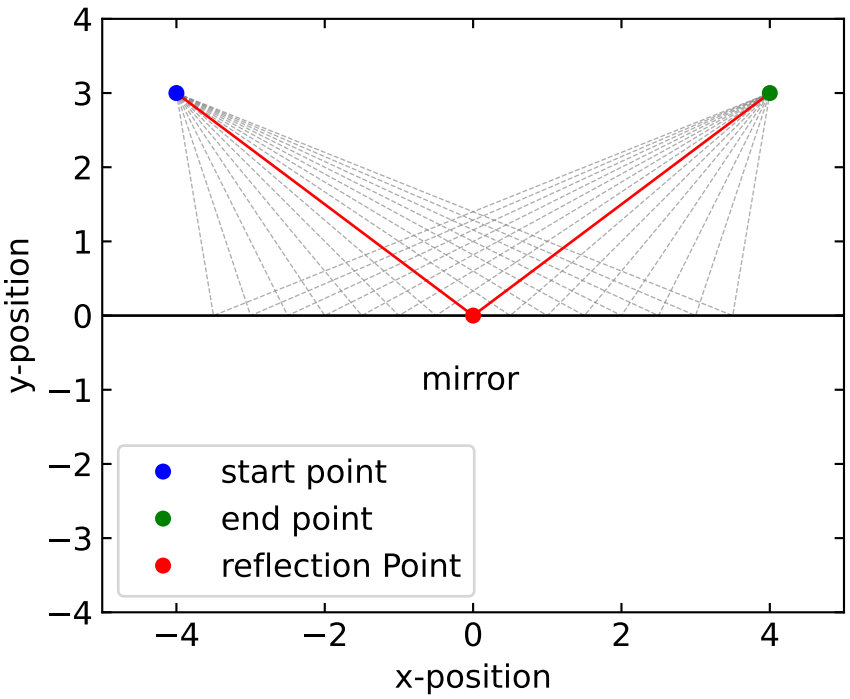


Figure 2.1: Fermat's principle for reflection of light at an interface

i Reflection Law Derivation

For reflection at a planar interface, we consider a ray traveling from point A to point B via reflection at point P on a mirror, as illustrated in Fig. 2.1. The total path length is $L = |AP| + |PB|$.
Let's establish a coordinate system where the mirror lies along the x-axis at y = 0. If point A is at coordinates $(-a, h_1)$ and point B is at (b, h_2), with the reflection point P at $(x, 0)$, the total path length is:

$$L(x) = \sqrt{(x+a)^2 + h_1^2} + \sqrt{(b-x)^2 + h_2^2}$$

According to Fermat's Principle, the actual path minimizes L, so we differentiate with respect to x and set it equal to zero:

$$\frac{dL}{dx} = \frac{x+a}{\sqrt{(x+a)^2 + h_1^2}} - \frac{b-x}{\sqrt{(b-x)^2 + h_2^2}} = 0$$

Rearranging this equation:

$$\frac{x+a}{\sqrt{(x+a)^2+h_1^2}} = \frac{b-x}{\sqrt{(b-x)^2+h_2^2}}$$

Now, let's interpret this geometrically. The angle of incidence θ_i is the angle between the incident ray AP and the normal to the mirror (y-axis). Similarly, the angle of reflection θ_r is the angle between the reflected ray PB and the normal.
From trigonometry:

- $\sin(\theta_i) = \frac{x+a}{\sqrt{(x+a)^2+h_1^2}}$
- $\sin(\theta_r) = \frac{b-x}{\sqrt{(b-x)^2+h_2^2}}$

Therefore, our minimization condition directly yields:

$$\sin(\theta_i) = \sin(\theta_r)$$

Since both angles are measured in the same quadrant (from the normal to the mirror), this equality implies:

$$\theta_i = \theta_r$$

This is the law of reflection: the angle of incidence equals the angle of reflection.
Law of Reflection: The angle of incidence equals the angle of reflection.

$$\theta_i = \theta_r$$

Refraction: Between media with different refractive indices, Fermat's Principle yields Snell's law:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

where θ_1 and θ_2 are the angles of incidence and refraction, respectively.

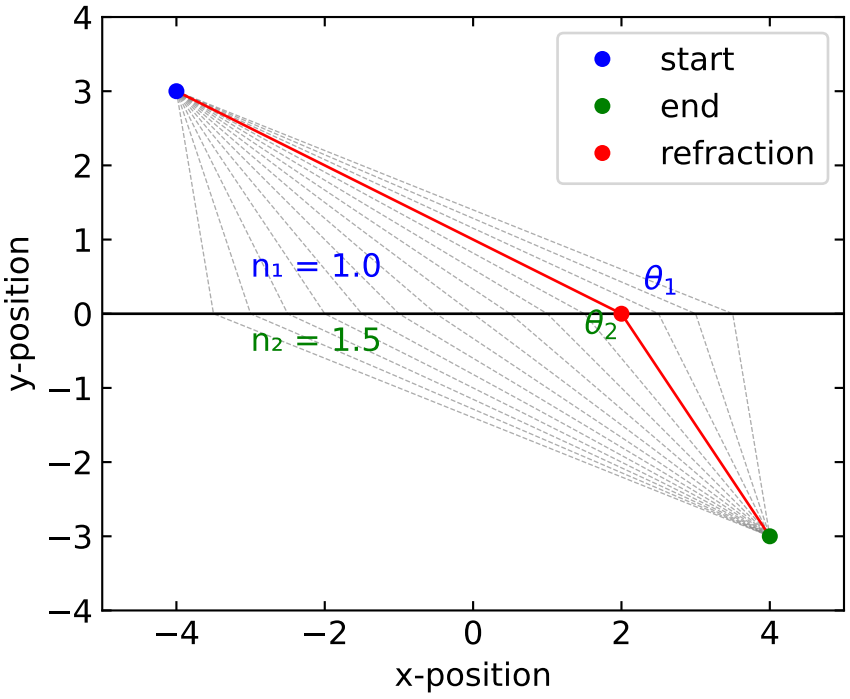


Figure 2.2: Snell's Law from Fermat's Principle

Refraction Law Derivation

For refraction between two media with different refractive indices, we apply Fermat's principle to find the path that minimizes the total optical path length. Consider a ray traveling from point A in medium 1 to point B in medium 2, with refraction occurring at point P on the interface, as illustrated in Fig. 2.2. The total optical path length is:

$$L = n_1|AP| + n_2|PB|$$

To determine the exact refraction point P that minimizes this path, we establish a coordinate system with the interface along the x-axis at y = 0. If point A is at coordinates (x_A, y_A) where $y_A > 0$, and point B is at (x_B, y_B) where $y_B < 0$, with the refraction point P at $(x, 0)$, the total optical path length is:

$$L(x) = n_1\sqrt{(x - x_A)^2 + y_A^2} + n_2\sqrt{(x_B - x)^2 + y_B^2}$$

According to Fermat's Principle, we minimize L by differentiating with respect to x and setting it equal to zero:

$$\frac{dL}{dx} = n_1\frac{x - x_A}{\sqrt{(x - x_A)^2 + y_A^2}} - n_2\frac{x_B - x}{\sqrt{(x_B - x)^2 + y_B^2}} = 0$$

Rearranging this equation:

$$\frac{n_1(x - x_A)}{\sqrt{(x - x_A)^2 + y_A^2}} = \frac{n_2(x_B - x)}{\sqrt{(x_B - x)^2 + y_B^2}}$$

From geometry, we can identify the sine of the angles of incidence and refraction:

- $\sin(\theta_1) = \frac{|x-x_A|}{|AP|} = \frac{|x-x_A|}{\sqrt{(x-x_A)^2+y_A^2}}$
- $\sin(\theta_2) = \frac{|x_B-x|}{|PB|} = \frac{|x_B-x|}{\sqrt{(x_B-x)^2+y_B^2}}$

Taking the sign into account based on our coordinate system, our minimization condition becomes:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$$

This is Snell's law, stating that the ratio of the sines of the angles of incidence and refraction equals the ratio of the refractive indices of the two media.

Snell's Law: The ratio of the sines of the angles of incidence and refraction equals the reciprocal of the ratio of the refractive indices.

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

Total Internal Reflection

When light travels from a medium with a higher refractive index (n_1) to one with a lower refractive index (n_2), a fascinating phenomenon can occur. As the angle of incidence increases, the refracted ray bends away from the normal until, at a critical angle, it travels along the boundary between the two media. Beyond this critical angle, light can no longer pass into the second medium and is instead completely reflected back into the first medium. This phenomenon is known as **total internal reflection** (TIR).

From Snell's law, the critical angle θ_c occurs when the refracted angle $\theta_2 = 90°$:

$$n_1 \sin\theta_c = n_2 \sin(90°) = n_2$$

Therefore:

$$\theta_c = \arcsin\left(\frac{n_2}{n_1}\right)$$

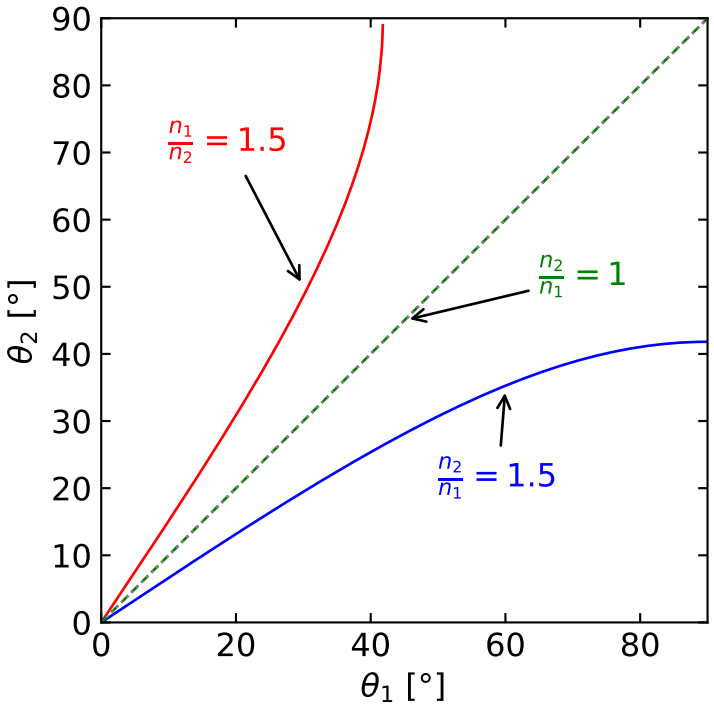


Figure 2.3: Snell's law for different combinations of refractive indices. The plots show the relationship between incident angle (θ_1) and refracted angle (θ_2) for three scenarios: (a) light passing from air to glass, (b) light passing from glass to air, and (c) a comparison of both cases. Note how the curves differ when light moves into a medium with higher refractive index versus a lower refractive index.

For total internal reflection to occur, two conditions must be satisfied:

1. Light must travel from a higher to a lower refractive index medium ($n_1 > n_2$)
2. The angle of incidence must exceed the critical angle ($\theta_1 > \theta_c$)

From Fermat's principle perspective, total internal reflection represents a scenario where no physical path through the second medium can satisfy the minimum optical path length requirement. Instead, the path of least time becomes the reflected path within the original medium. This phenomenon has numerous practical applications, including:

- **Fiber optic communication**: Light signals travel long distances through optical fibers via successive total internal reflections with minimal loss
- **Prisms and reflectors**: Total internal reflection in prisms provides perfect reflection without needing reflective coatings
- **Gemstones**: The brilliance of diamonds results from light being trapped through multiple internal reflections
- **Optical instruments**: Binoculars, periscopes, and endoscopes use prisms with TIR to redirect light

Total internal reflection demonstrates how Fermat's principle enforces an absolute constraint on light's behavior—when no path through the second medium can minimize the optical path length, light must remain in the first medium, following the path of least time.

Optical Fibers and Total Internal Reflection

Total internal reflection plays a crucial role in modern telecommunications, particularly in optical fibers, which are also part of many experimental setups. These fibers are essentially ultra-thin glass wires, ranging in diameter from a few micrometers to several hundred micrometers, designed to transport light over long distances with minimal loss.

The structure of an optical fiber is key to its function:

1. Core: A central glass core with a refractive index n_1
2. Cladding: A surrounding layer with a slightly lower refractive index n_2

This difference in refractive indices is what allows total internal reflection to occur within the fiber.

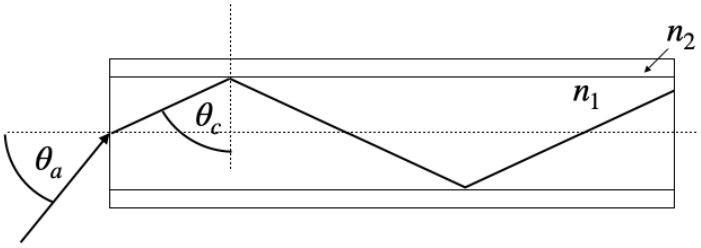


Figure 2.4: Total internal reflection in an optical fiber.

For light to propagate effectively through the fiber, it must enter at an angle that ensures total internal reflection at the core-cladding interface. This leads to the concept of the acceptance angle, θ_a, which is the maximum angle at which light can enter the fiber and still undergo total internal reflection.

To characterize this acceptance angle, optical engineers use a parameter called the **Numerical Aperture (NA)**.

Numerical Aperture

The Numerical Aperture of a fiber is defined as the sine of the maximum acceptance angle:

$$NA = \sin(\theta_a) = \sqrt{n_1^2 - n_2^2} \tag{2.1}$$

This equation relates the NA directly to the refractive indices of the core and cladding. The derivation of this formula involves applying Snell's law at the air-fiber interface and at the core-cladding interface, then using the condition for total internal reflection.

In practice, typical values for the refractive indices might be $n_1 = 1.475$ for the core and $n_2 = 1.46$ for the cladding. Plugging these into our equation:

$$NA = \sqrt{1.475^2 - 1.46^2} \approx 0.2 \tag{2.2}$$

This means that light entering the fiber within a cone of about 11.5° (arcsin(0.2)) from the fiber's axis will be transmitted through the fiber via total internal reflection.

The NA is an important parameter in fiber optic design:

1. It determines the light-gathering ability of the fiber.
2. It affects the fiber's bandwidth and its susceptibility to certain types of signal distortion.
3. It influences how easily the fiber can be coupled to light sources and other fibers.

Optical fibers come in various types, each optimized for different applications. Some fibers are designed to transmit light over long distances with minimal loss, while others are engineered for specific wavelengths or to guide light in unusual ways. The figure below shows a few examples of optical fiber types.

Differential Form of Fermat's Law

To derive the differential ray equation from Fermat's integral principle, we apply the calculus of variations. Starting with the optical path length functional:

$$L = \int_C n(s)ds = \int_{t_1}^{t_2} n(\mathbf{r}(t)) \left|\frac{d\mathbf{r}}{dt}\right| dt$$

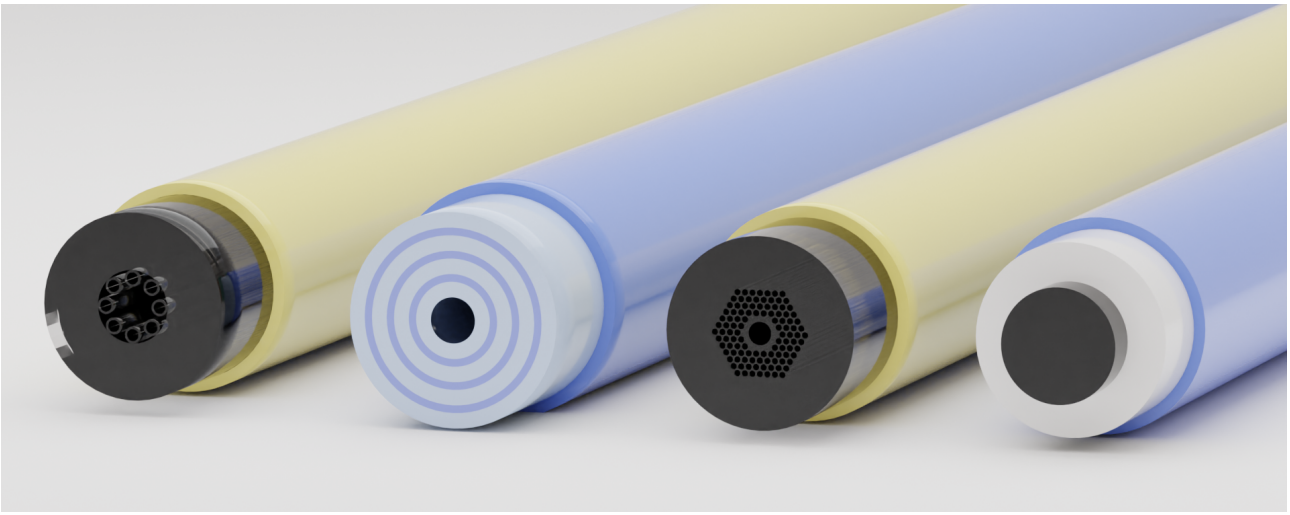


Figure 2.5: Rendering of different optical fibers types (from left to right): Hollow core optical fiber, hollow core bragg fiber, photonic crystal fiber, conventional fiber

Where $\mathbf{r}(t)$ parametrizes the path. The term $\left|\frac{d\mathbf{r}}{dt}\right|$ represents the differential element of arc length ds along the path, so $ds = \left|\frac{d\mathbf{r}}{dt}\right| dt$. This parametrization allows us to convert the path integral over the curve C into a definite integral over the parameter t. According to Fermat's principle, the true path makes this integral stationary (L = 0).

Consider a small variation in the path: $\mathbf{r}(t) \to \mathbf{r}(t) + \epsilon\,(t)$ where $(t_1) = (t_2) = 0$ (fixed endpoints). Expanding the variation of the integral to first order in :

$$\delta L = \frac{d}{d\epsilon}\bigg|_{\epsilon=0} \int_{t_1}^{t_2} n(\mathbf{r}(t) + \epsilon\,(t)) \left|\frac{d}{dt}(\mathbf{r}(t) + \epsilon\,(t))\right| dt$$

Using the chain rule and reparametrizing with arc length s (where $\frac{d\mathbf{r}}{ds}$ is a unit vector), the stationarity condition leads to:

$$\int_C \left[\nabla n \cdot \quad - \frac{d}{ds}\left(n\frac{d\mathbf{r}}{ds}\right) \cdot \right] ds = 0$$

Since this must hold for any variation , we obtain the Euler-Lagrange equation:

$$\frac{d}{ds}\left(n\frac{d\mathbf{r}}{ds}\right) = \nabla n$$

This shows that rays bend toward regions of higher refractive index, directly analogous to how a mechanical particle's trajectory is affected by a potential field in classical mechanics.

SELFOC Gradient Index Lens

SELFOC (SELf-FOCusing) gradient-index fibers are interesting optical elements that guide light through a continuous refraction process rather than discrete refractions at interfaces. Let me demonstrate how Fermat's principle can be used to determine the ray paths in these fibers. A SELFOC fiber has a radially varying refractive index, typically following a parabolic profile:

$$n(r) = n_0\left(1 - \frac{1}{2}\alpha^2 r^2\right)$$

where: - n_0 is the refractive index at the central axis - r is the radial distance from the axis - α is the

gradient constant that determines how quickly the index decreases with radius

Fermat's Principle in Gradient-Index Media

For a medium with a spatially varying refractive index, Fermat's principle states that light follows the path that minimizes the optical path length:

$$\delta \int_C n(r) ds = 0$$

This yields the differential equation:

$$\frac{d}{ds}\left(n\frac{d\mathbf{r}}{ds}\right) = \nabla n$$

Deriving the Ray Path Equation

For our parabolic index profile, the gradient of the refractive index is:

$$\nabla n = \frac{\partial n}{\partial r}\hat{\mathbf{r}} = -n_0 \alpha^2 r \hat{\mathbf{r}}$$

Using cylindrical coordinates with z along the fiber axis, and assuming the paraxial approximation (rays make small angles with the z-axis), we can simplify the ray equation to:

$$\frac{d^2 r}{dz^2} + \alpha^2 r = 0$$

This is the equation for a harmonic oscillator, which has the solution:

$$r(z) = r_0 \cos(\alpha z) + \frac{\theta_0}{\alpha}\sin(\alpha z)$$

where r_0 is the initial radial position and θ_0 is the initial angle of the ray with respect to the fiber axis.

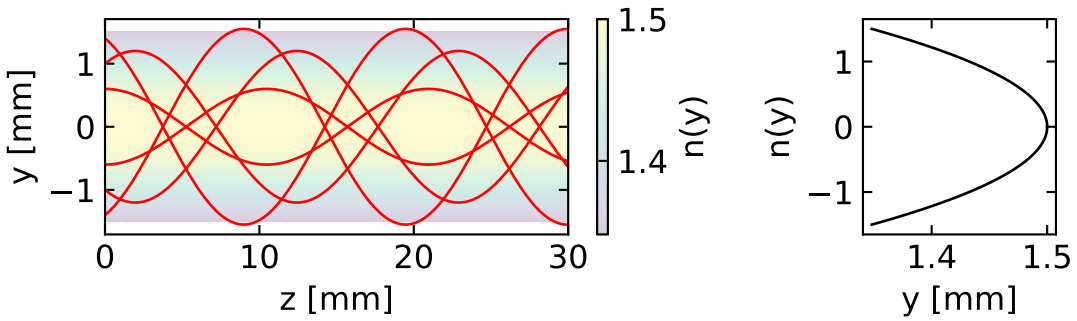


Figure 2.6: Ray-path inside a SELFOC gradient index optical fiber.

Fermat's Principle and the "F=ma" Analogy in Optics

The differential form of Fermat's principle:

$$\frac{d}{ds}\left(n\frac{d\mathbf{r}}{ds}\right) = \nabla n$$

reveals a profound analogy with Newton's Second Law of motion:

$$\mathbf{F} = m\mathbf{a} = m\frac{d^2\mathbf{r}}{dt^2}$$

This comparison, sometimes called "F=ma optics," illustrates how light rays follow trajectories mathematically similar to those of mechanical particles. To see this connection more clearly, we can expand the ray equation as:

$$n\frac{d^2\mathbf{r}}{ds^2} + \frac{d\mathbf{r}}{ds}\frac{dn}{ds} = \nabla n$$

Using the chain rule, $\frac{dn}{ds} = \nabla n \cdot \frac{d\mathbf{r}}{ds}$, and denoting $\mathbf{t} = \frac{d\mathbf{r}}{ds}$ as the unit tangent vector along the ray:

$$n\frac{d^2\mathbf{r}}{ds^2} + (\nabla n \cdot \mathbf{t})\mathbf{t} = \nabla n$$

Rearranging to isolate the ray curvature term:

$$n\frac{d^2\mathbf{r}}{ds^2} = \nabla n - (\nabla n \cdot \mathbf{t})\mathbf{t}$$

The right side represents the component of ∇n perpendicular to the ray direction, which we can denote as $(\nabla n)_\perp$. Therefore:

$$\frac{d^2\mathbf{r}}{ds^2} = \frac{1}{n}(\nabla n)_\perp$$

This equation reveals that the ray curvature is proportional to the perpendicular component of the refractive index gradient and inversely proportional to the refractive index itself. Crucially, this shows that light rays bend **toward** regions of higher refractive index, not away from them.

This behavior explains many optical phenomena:

- Light bending toward the normal when entering a medium with higher refractive index
- Light guiding in optical fibers where light remains confined in the higher-index core
- Formation of mirages where light curves toward the denser air near the ground
- Focusing in gradient-index (GRIN) lenses where the refractive index decreases radially from the center

While the mathematical form resembles Newton's equation for particle motion, the analogy must be carefully interpreted: unlike particles that accelerate toward lower potential energy, light rays curve toward regions of higher refractive index.

Lenses

Lenses are among the most fundamental optical elements in photonics, using curved surfaces (typically spherical) to manipulate light paths. Understanding how lenses work requires analyzing refraction at spherical surfaces and applying this to the thin lens model.

Refraction at Spherical Surfaces

When light encounters a spherical boundary between two media, we can analyze its path using Snell's law and geometric considerations as shown below:

To determine how an image forms, we need to find where rays originating from a point at distance a from the surface will converge after refraction. Using Snell's law for a ray hitting the surface at angle $\alpha + \theta_1$:

$$n_1 \sin(\alpha + \theta_1) = n_2 \sin(\alpha - \theta_2)$$

Where:

$$\sin(\alpha) = \frac{y}{R}, \quad \tan(\theta_1) = \frac{y}{a}, \quad \tan(\theta_2) = \frac{y}{b}$$

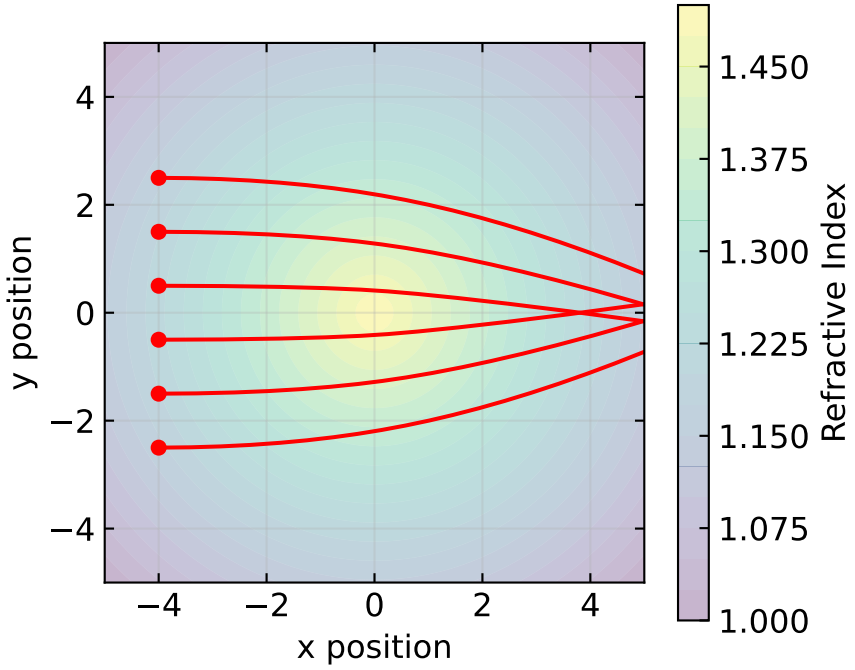


Figure 2.7: F=ma optics - Light rays (red) following paths toward regions of higher refractive index

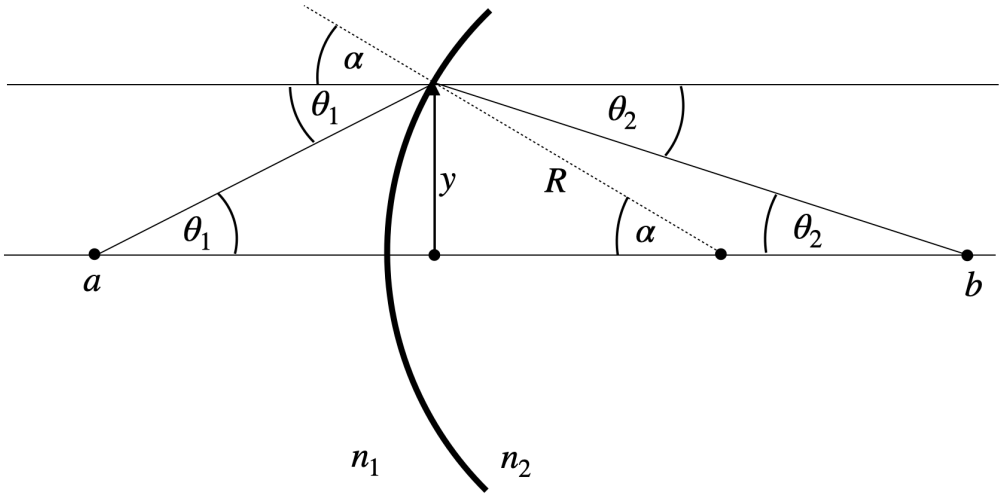


Figure 2.8: Refraction at a curved surface.

For practical optical systems, we employ the **paraxial approximation**, where all angles are assumed small enough that:

$$\sin(\theta) \approx \theta + O(\theta^3), \quad \tan(\theta) \approx \theta + O(\theta^3), \quad \cos(\theta) \approx 1 + O(\theta^2)$$

This simplifies Snell's law to:

$$n_1(\alpha + \theta_1) = n_2(\alpha - \theta_2)$$

After appropriate transformations (detailed in the online lecture), we obtain:

$$\theta_2 = \frac{n_2 - n_1}{n_2 R} y - \frac{n_1}{n_2}\theta_1$$

This linear relationship between input (y, θ_1) and output (θ_2) parameters is a hallmark of paraxial optics.

i Paraxial Approximation

The paraxial approximation is a fundamental simplification in optics that assumes all angles are small. This allows us to use linear approximations for trigonometric functions, significantly simplifying calculations while maintaining accuracy for most practical scenarios involving lenses.
To visualize the validity of this approximation, let's examine two plots:

1. The first plot compares sin( ) (blue line) with its linear approximation (red dashed line) for angles ranging from 0 to /2 radians.
2. The second plot shows the absolute error between sin( ) and .

These plots demonstrate that:

1. For small angles (roughly up to 0.5 radians or about 30 degrees), the approximation is very close to the actual sine function.
2. The error increases rapidly for larger angles, indicating the limitations of the paraxial approximation.

In most optical systems, especially those involving lenses, the angles of incident and refracted rays are typically small enough for this approximation to be valid. However, it's important to be aware of its limitations when dealing with wide-angle optical systems or scenarios where precision is critical.

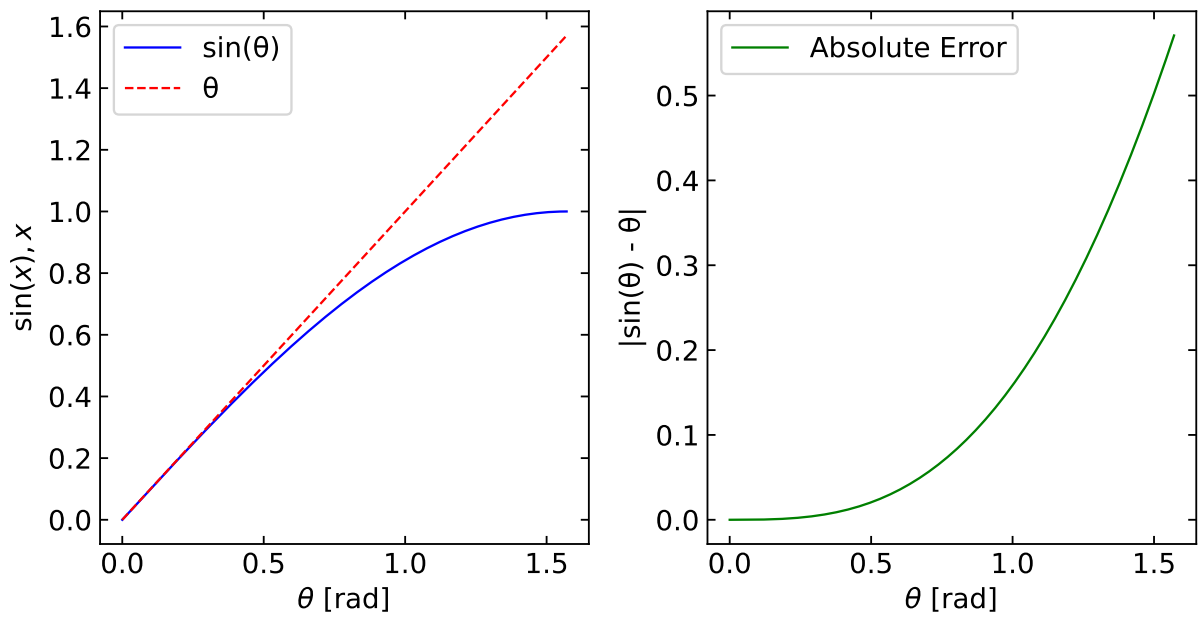


Figure 2.9: Visualization of the paraxial approximation plotting the $\sin(\theta)$ and the linear approximation θ (dashed line) for angles ranging from 0 to $\pi/2$ radians.

To derive the imaging equation, we analyze how light from a point object forms an image after refraction. Consider two special rays from an off-axis point:

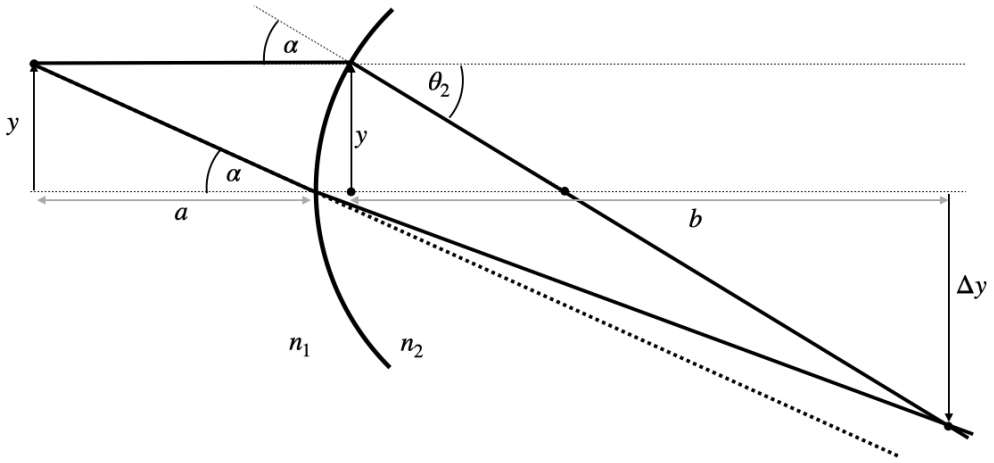


Figure 2.10: Image formation at a curved surface.

For a ray parallel to the optical axis ($\theta_1 = 0$):

$$\theta_2 = \frac{n_2 - n_1}{n_2}\frac{y}{R} = \frac{y + \Delta y}{b}$$

For a ray through the center of curvature ($y = 0$):

$$n_2 \frac{\Delta y}{b} = n_1 \frac{y}{a}$$

Combining these equations yields the fundamental imaging equation for a spherical surface:

$$\frac{n_1}{a} + \frac{n_2}{b} = \frac{n_2 - n_1}{R}$$

From this, we define the **focal length** of the spherical surface:

$$f = \frac{n_2}{n_2 - n_1} R$$

Imaging Equation for Spherical Refracting Surface

The sum of the inverse object and image distances equals the inverse focal length of the spherical refracting surface:

$$\frac{n_1}{a} + \frac{n_2}{b} \approx \frac{n_2}{f}$$

where the focal length of the refracting surface is given by:

$$f = \frac{n_2}{n_2 - n_1} R$$

in the paraxial approximation.

Thin Lens

A lens consists of two spherical surfaces in close proximity. To analyze how a lens forms images, we consider refraction at both surfaces:

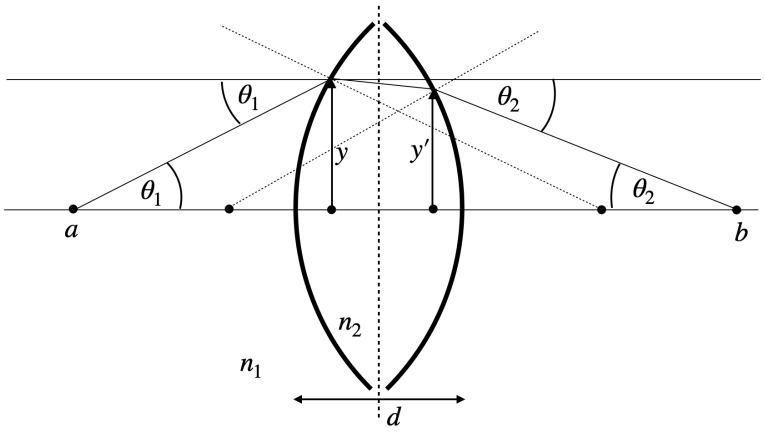


Figure 2.11: Refraction on two spherical surfaces.

When the lens thickness d is much smaller than the radii of curvature ($d \ll R_1, R_2$), we can apply the **thin lens approximation**. This assumes: 1. The ray height at both surfaces is approximately equal ($y \approx y'$) 2. All refraction effectively occurs at a single plane (the **principal plane**) 3. The change in angle is additive from both surfaces

This approximation, combined with the sign convention for radii (positive for convex surfaces facing incoming light, negative for concave), leads to the thin lens formula:

Imaging Equation for Thin Lens

The sum of the inverse object and image distances equals the inverse focal length of the thin lens:

$$\frac{1}{a} + \frac{1}{b} = \frac{1}{f}$$

where:

$$\frac{1}{f} = \frac{n_2 - n_1}{n_1}\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

This can be rearranged to give the **lensmaker equation**:

Lensmaker Equation

The focal length of a thin lens is calculated by:

$$f = \frac{n_1}{n_2 - n_1}\left(\frac{R_1 R_2}{R_2 - R_1}\right)$$

in the paraxial approximation.

Image Construction and Magnification

To construct the image formed by a lens, we typically trace two or three special rays: 1. A ray parallel to the optical axis, which passes through the far focal point after refraction 2. A ray through the center of the lens, which passes undeflected 3. A ray through the near focal point, which emerges parallel to the optical axis

The intersection of these rays locates the image position:

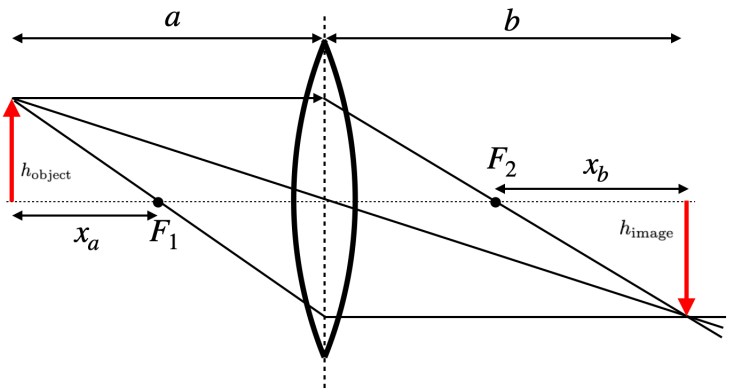


Figure 2.12: Image construction on a thin lens.

The ratio of image height to object height defines the **magnification**:

Magnification of a Lens

The magnification is given by:

$$M = \frac{h_{\text{image}}}{h_{\text{object}}} = -\frac{b}{a} = \frac{f}{f - a}$$

where the negative sign indicates image inversion for real images.

The image characteristics depend on the object distance relative to the focal length:

Object Position	Image Characteristics	Magnification (M)	Image Type
$a < f$	Upright and magnified	$M > 0$	Virtual
$f < a < 2f$	Inverted and magnified	$M < -1$	Real
$a = 2f$	Inverted, same size	$M = -1$	Real
$a > 2f$	Inverted and reduced	$-1 < M < 0$	Real
$a = f$	Image at infinity	$M = \infty$	-

The diagram below illustrates these various imaging scenarios for a biconvex lens:

Fig.: Image construction on a biconvex lens with a parallel and a central ray for different object distances.

Matrix Optics

The above derived equations for a single spherical surface yield a linear relation between the input variables y_1 and θ_1 and the output variables y_2 and θ_2. The linear relation yields a great opportunity to express optical elements in terms of linear transformations (matrices). This is the basis of **matrix optics**. The matrix representation of a lens is given by

$$\begin{pmatrix} y_2 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \theta_1 \end{pmatrix}$$

where the matrix is called the **ABCD matrix** of the lens. Due to the linearization of Snells law w can write down more generally

$$\begin{pmatrix} y_2 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} y_1 \\ \theta_1 \end{pmatrix}$$

and one can obtain a Matrix for all types of optical elements such as free space of dustance d.

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}$$

Here are some useful matrices for optical elements:

$$\mathbf{M} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \quad \text{(Free space)}$$

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{n_1}{n_2} \end{bmatrix} \quad \text{(Planar interface)}$$

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ -\frac{(n_2 - n_1)}{n_2 R} & \frac{n_1}{n_2} \end{bmatrix} \quad \text{(Spherical Boundary)}$$

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix} \quad \text{(Tin Lens)}$$

If we have now a system of optical elements, we can multiply the matrices of the individual elements to obtain the matrix of the whole system.

$$\rightarrow \mathrm{M}_1 \rightarrow \mathrm{M}_2 \rightarrow \mathrm{M}_N \rightarrow \mathrm{M} = \mathbf{M}_N \ldots \mathrm{M}_2 \mathbf{M}_1.$$

This is a very powerful tool to analyze optical systems.

2.1 Fermat's Principle for Spherical Surfaces

The power of Fermat's principle becomes particularly evident when applied to spherical refracting surfaces. Consider a spherical boundary of radius R between two media with refractive indices n_1 and n_2. According to Fermat's principle, light will follow the path that minimizes the total optical path length.

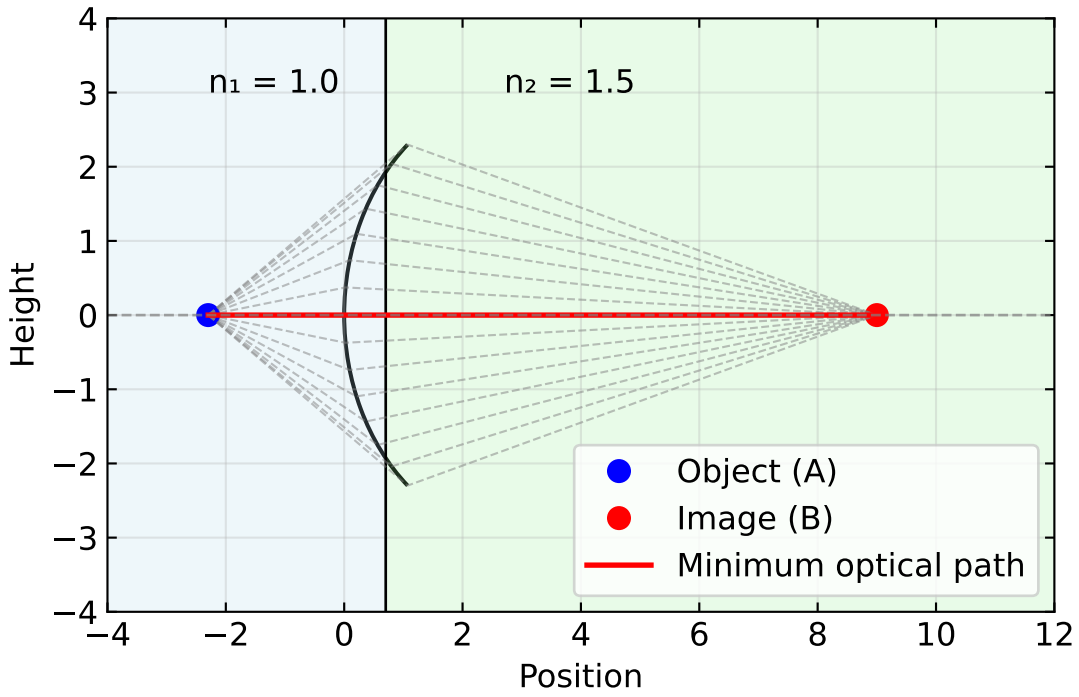


Figure 2.13: Fermat's principle applied to a spherical refracting surface

When we apply Fermat's principle to a spherical surface, we can derive the laws of refraction. Consider a spherical boundary between two media with refractive indices n_1 and n_2. We'll place our coordinate system so that the spherical surface intersects the x-axis at x=0, with radius R and its center at position (R,0) to the right.

For a point P on the spherical surface at height y from the optical axis, the total optical path length from object point A at (-a,0) to image point B at (b,0) is:

$$L = n_1|AP| + n_2|PB|$$

where:

$$|AP| = \sqrt{a^2 + y^2}$$
$$|PB| = \sqrt{b^2 + y^2}$$

According to Fermat's principle, light follows the path where this length is stationary:

$$\frac{dL}{dy} = n_1 \frac{d|AP|}{dy} + n_2 \frac{d|PB|}{dy} = 0$$

Computing these derivatives:

$$\frac{d|AP|}{dy} = \frac{y}{|AP|}$$

$$\frac{d|PB|}{dy} = \frac{y}{|PB|}$$

Substituting into our condition:

$$n_1 \frac{y}{|AP|} + n_2 \frac{y}{|PB|} = 0$$

This equation is incorrect. The right-hand side should not be zero because we need to account for the geometry of the spherical surface. The correct form includes the effect of the surface normal:

$$n_1 \frac{y}{|AP|} + n_2 \frac{y}{|PB|} = \frac{(n_2 - n_1)y}{R}$$

This correction comes from the fact that at point P, the normal to the spherical surface makes an angle with the optical axis, where sin( ) y/R in the paraxial approximation.

Dividing by y (assuming y 0):

$$\frac{n_1}{|AP|} + \frac{n_2}{|PB|} = \frac{n_2 - n_1}{R}$$

In the paraxial approximation, we can use |AP| a and |PB| b, yielding:

$$\frac{n_1}{a} + \frac{n_2}{b} = \frac{n_2 - n_1}{R}$$

This is the correct imaging equation for a spherical refracting surface.

The elegance of Fermat's principle is preserved, as it still naturally produces the same result as our geometric derivation, once we properly account for the geometry of the refracting surface.

Deriving the Thin Lens Equation from Fermat's Principle

To derive the thin lens equation, we apply Fermat's principle to the two spherical surfaces that make up a lens. Consider a lens with refractive index n_2 in a medium of index n_1, with surface radii R_1 and R_2. The total optical path for a ray passing through the lens at height y from the optical axis is: - Path from object to first surface: $n_1 s_1$ - Path through the lens: $n_2 s_2$ - Path from second surface to image: $n_1 s_3$ For a thin lens, the optical path length simplifies to:

$$L(y) = n_1\sqrt{a^2 + y^2} + n_2 d(y) + n_1\sqrt{b^2 + y^2}$$

Where $d(y)$ is the thickness of the lens at height y, which can be approximated as:

$$d(y) \approx d_0 + \frac{y^2}{2}\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

Applying Fermat's principle ($\frac{dL}{dy} = 0$) and using the paraxial approximation:

$$\frac{n_1 y}{\sqrt{a^2 + y^2}} + n_2 y\left(\frac{1}{R_1} - \frac{1}{R_2}\right) + \frac{n_1 y}{\sqrt{b^2 + y^2}} = 0$$

In the paraxial limit ($y \ll a, y \ll b$), this becomes:

$$\frac{n_1 y}{a} + n_2 y\left(\frac{1}{R_1} - \frac{1}{R_2}\right) + \frac{n_1 y}{b} = 0$$

Dividing by y and rearranging:

$$\frac{1}{a} + \frac{1}{b} = \frac{n_2 - n_1}{n_1}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = \frac{1}{f}$$

This is the thin lens equation with the focal length given by the lensmaker's equation:

$$f = \frac{n_1}{n_2 - n_1}\left(\frac{R_1 R_2}{R_2 - R_1}\right)$$

Thus, both the imaging equation and the lensmaker equation emerge naturally from Fermat's principle applied to the geometry of a thin lens, showing that light follows paths of equal optical length from object to image when passing through any part of the lens.
From a wave perspective, what makes a lens focus light to a point is that all paths from object to image through any part of the lens have equal optical path lengths (to first order in the paraxial approximation), ensuring constructive interference at the image point.

Part II

Lecture 2

Chapter 3

Theories for light

Refraction at Spherical Surfaces

When light encounters a spherical boundary between two media, we can analyze its path using Snell's law and geometric considerations as shown below:

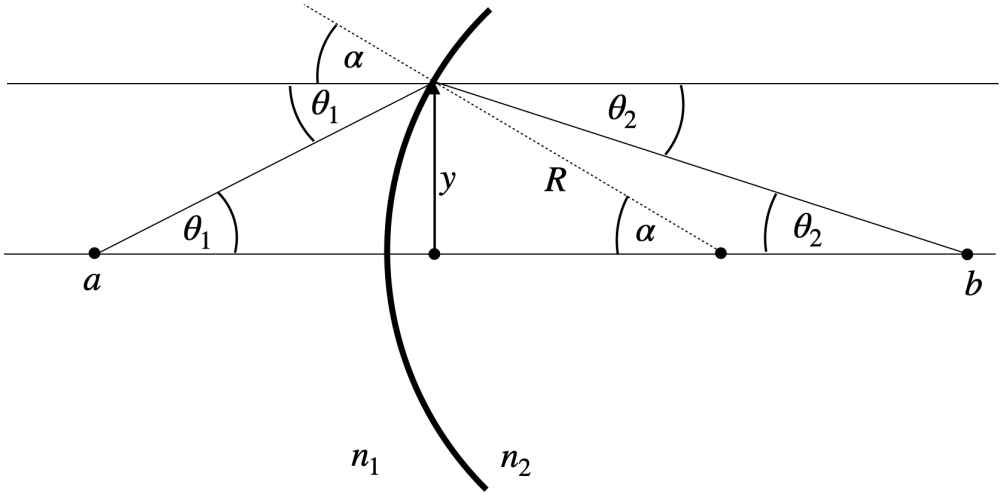


Figure 3.1: Refraction at a curved surface.

To determine how an image forms, we need to find where rays originating from a point at distance a from the surface will converge after refraction. Using Snell's law for a ray hitting the surface at angle $\alpha + \theta_1$:

$$n_1 \sin(\alpha + \theta_1) = n_2 \sin(\alpha - \theta_2)$$

Where:

$$\sin(\alpha) = \frac{y}{R}, \quad \tan(\theta_1) = \frac{y}{a}, \quad \tan(\theta_2) = \frac{y}{b}$$

For practical optical systems, we employ the **paraxial approximation**, where all angles are assumed small enough that:

$$\sin(\theta) \approx \theta + O(\theta^3), \quad \tan(\theta) \approx \theta + O(\theta^3), \quad \cos(\theta) \approx 1 + O(\theta^2)$$

This simplifies Snell's law to:

$$n_1(\alpha + \theta_1) = n_2(\alpha - \theta_2)$$

After appropriate transformations (detailed in the online lecture), we obtain:

$$\theta_2 = \frac{n_2 - n_1}{n_2 R} y - \frac{n_1}{n_2}\theta_1$$

and

$$y = y_1 = y_2$$

This linear relationship between input (y, θ_1) and output (y,θ_2) parameters is a hallmark of paraxial optics and a result of the linearization of Snells law.

Matrix Optics

The linear relation between input and output parameters allows us to express optical elements as linear transformations (matrices). This approach forms the foundation of **matrix optics**. For a lens, the matrix representation is:

$$\begin{pmatrix} y_2 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \theta_1 \end{pmatrix}$$

This 2×2 matrix is called the **ABCD matrix** of the lens. Thanks to the linearization of Snell's law, we can generalize this to any optical element:

$$\begin{pmatrix} y_2 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} y_1 \\ \theta_1 \end{pmatrix}$$

Each element in the ABCD matrix has a specific physical meaning:

Matrix Element	Physical Meaning
A	Magnification - relates output position to input position
B	Position-to-angle conversion - relates output position to input angle
C	Focusing power - relates output angle to input position
D	Angular magnification - relates output angle to input angle

Every optical element can be characterized by these parameters. For example, a lens has C = -1/f (focusing power), while free space has B = d (position-dependent angle change). An important property is that the determinant of the matrix equals the ratio of refractive indices: det(M) = n /n , which equals 1 in a single medium.

Here are the ABCD matrices for common optical elements:

$$\mathbf{M} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \quad \text{(Free space)}$$

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{n_1}{n_2} \end{bmatrix} \quad \text{(Planar interface)}$$

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ -\frac{(n_2-n_1)}{n_2 R} & \frac{n_1}{n_2} \end{bmatrix} \quad \text{(Spherical Boundary)}$$

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix} \quad \text{(Thin Lens)}$$

For a system containing multiple optical elements, we simply multiply their matrices in the order that light passes through them:

$$\mathrm{M} = \mathbf{M}_N \cdots \mathbf{M}_2 \mathbf{M}_1$$

A ray entering the first optical element at a height y_1 at an angle θ_1 is transformed according to the matrix $\mathbf{M}$ by the whole system. This elegant approach provides a powerful tool for analyzing complex optical systems efficiently.

i Example: Optical Cloaking with Lens Systems

Optical cloaking refers to making objects "invisible" by guiding light rays around them such that to an observer, it appears as if the rays traveled through free space without encountering any object. Using matrix optics, we can design such a system.

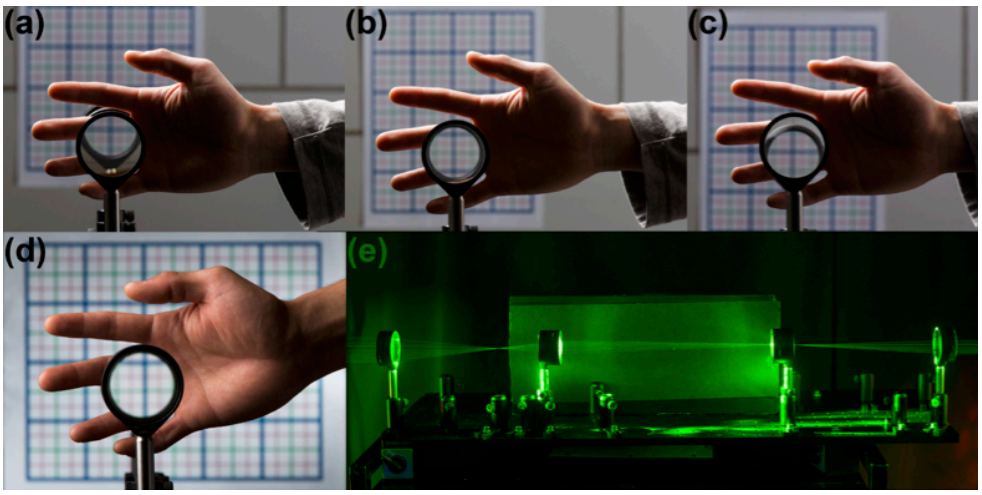


Figure 3.2: Example of a practical paraxial cloak. (a)–(c) A hand is cloaked for varying directions, while the background image is transmitted properly.(d) On-axis view of the ray optics cloaking device. (e) Setup using practical, easy to obtain optics, for demonstrating paraxial cloaking principles. (Photos by J. Adam Fenster, videos by Matthew Mann / University of Rochester) Source

For perfect optical cloaking, the ABCD matrix of our system must be equivalent to that of free space:

$$\mathbf{M}_{cloaking} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}$$

Where d is the total effective optical path length. Let's explore why we need exactly 4 lenses to achieve this.

Analysis of Different Lens Configurations

1. Single Lens Configuration
For a single lens with focal length f, the ABCD matrix is:

$$\mathbf{M}_{single} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix}$$

This clearly cannot match the free space matrix due to the non-zero C element.

2. Two-Lens Configuration

For two lenses with focal lengths f_1 and f_2 separated by distance d_{12}:

$$\mathbf{M}_{two} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f_2} & 1 \end{bmatrix} \begin{bmatrix} 1 & d_{12} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{f_1} & 1 \end{bmatrix}$$

When $d_{12} = f_1 + f_2$ (telescopic arrangement), this simplifies to:

$$\mathbf{M}_{two} = \begin{bmatrix} -\frac{f_1}{f_2} & 0 \\ 0 & -\frac{f_2}{f_1} \end{bmatrix}$$

Since both magnification (A) and angular magnification (D) cannot simultaneously equal 1 while maintaining $\det(\mathbf{M}) = 1$, two lenses are insufficient.

3. Three-Lens Configuration

With three lenses, we have more parameters but still need to determine if we can satisfy all constraints simultaneously. For a three-lens system with focal lengths f_1, f_2, and f_3, separated by distances d_{12} and d_{23}, the system matrix would be:

$$\mathbf{M}_{three} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f_3} & 1 \end{bmatrix} \begin{bmatrix} 1 & d_{23} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{f_2} & 1 \end{bmatrix} \begin{bmatrix} 1 & d_{12} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{f_1} & 1 \end{bmatrix}$$

When multiplied out, the focusing power (C element) of the system is:

$$C = -\frac{1}{f_3} - \frac{1}{f_2} - \frac{1}{f_1} + \frac{d_{12}}{f_1 f_2} + \frac{d_{23}}{f_2 f_3} + \frac{d_{12} d_{23}}{f_1 f_2 f_3}$$

The magnification (A element) of the system is:

$$A = 1 - \frac{d_{12}}{f_1} - \frac{d_{23}}{f_2} + \frac{d_{12} d_{23}}{f_1 f_2}$$

For perfect cloaking, we need both C = 0 and A = 1. From A = 1, we can derive:

$$\frac{d_{12}}{f_1} + \frac{d_{23}}{f_2} - \frac{d_{12} d_{23}}{f_1 f_2} = 0$$

Solving for d_{23}, we get:

$$d_{23} = \frac{d_{12}/f_1}{1/f_2 - d_{12}/(f_1 f_2)}$$

Substituting this into the condition for C = 0, we obtain a complex expression that places constraints on the possible values of f_1, f_2, and f_3. For typical lens configurations, this results in values that are difficult to realize physically, as it often requires either negative separations or negative focal lengths.

While the three-lens system provides more parameters to work with than the two-lens system, the mathematical constraints of simultaneously achieving zero focusing power (C = 0) and unit magnification (A = 1) still make perfect cloaking challenging with conventional optical elements.

4. Four-Lens Configuration: The Solution

We can arrange four lenses in two pairs:

- First pair (lenses 1 and 2): A beam compressor
- Second pair (lenses 3 and 4): A beam expander

For the beam compressor, with lenses at their combined focal length apart:

$$\mathbf{M}_{comp} = \begin{bmatrix} -\frac{f_1}{f_2} & 0 \\ 0 & -\frac{f_2}{f_1} \end{bmatrix}$$

Similarly, for the beam expander:

$$\mathbf{M}_{exp} = \begin{bmatrix} -\frac{f_3}{f_4} & 0 \\ 0 & -\frac{f_4}{f_3} \end{bmatrix}$$

The combined system matrix is:

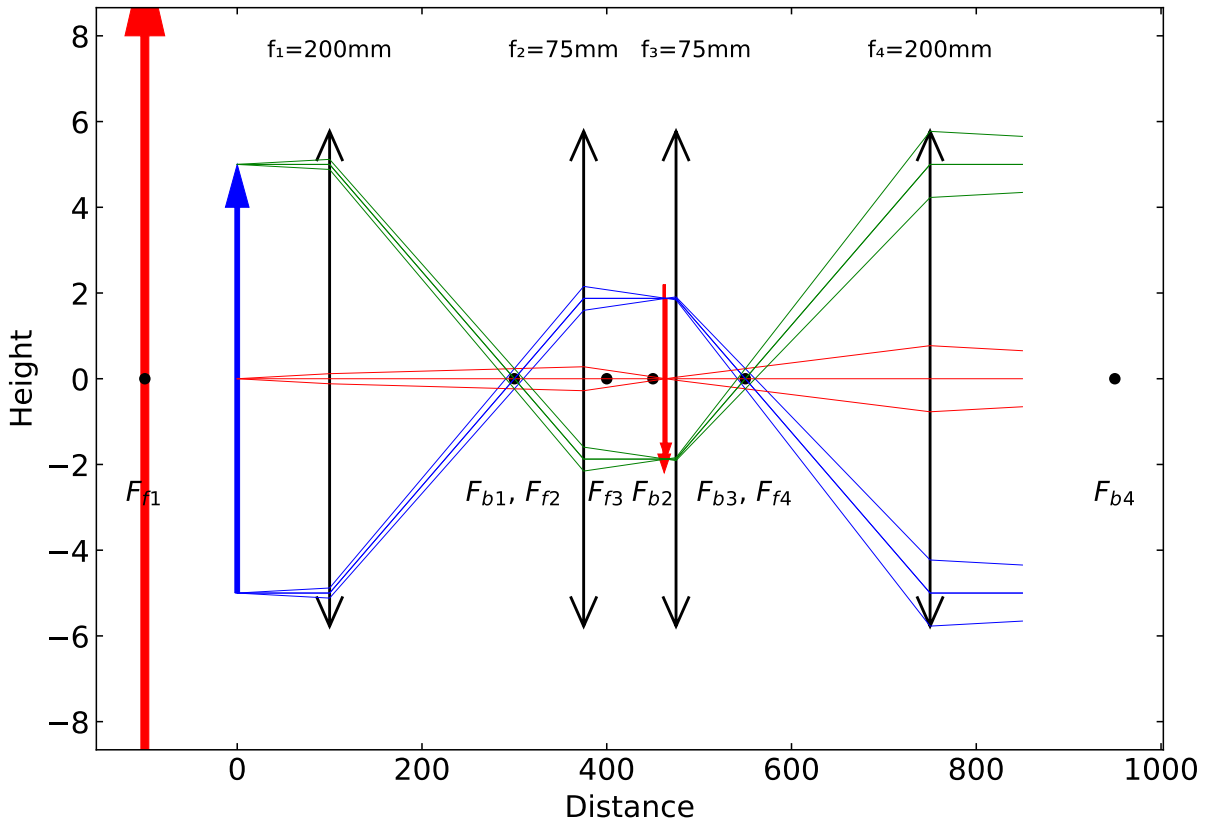
$$\mathbf{M}_{total} = \mathbf{M}_{exp} \times \mathbf{M}_{comp} = \begin{bmatrix} \frac{f_1}{f_2} \times \frac{f_3}{f_4} & 0 \\ 0 & \frac{f_2}{f_1} \times \frac{f_4}{f_3} \end{bmatrix}$$

For perfect cloaking, we need: - $\frac{f_1}{f_2} \times \frac{f_3}{f_4} = 1$ - $\frac{f_2}{f_1} \times \frac{f_4}{f_3} = 1$
This is satisfied when $f_1 = f_4$ and $f_2 = f_3$.
With the lenses properly spaced and an additional free space distance d_0 between the two pairs, the complete system matrix becomes:

$$\mathbf{M}_{cloaking} = \begin{bmatrix} 1 & d_0 \\ 0 & 1 \end{bmatrix}$$

Which perfectly mimics free space propagation, creating the optical cloaking effect.

```
BeginnerHint [in figure.py]: Infinite field of view: cannot use limitObjectToFieldOfView=True. The ob
BeginnerHint [in imagingpath.py]: Field of view is infinite. You can pass useObject=True to use the f
BeginnerHint [in figure.py]: No aperture stop in the system: cannot use onlyPrincipalAndAxialRays=Tru
```



Multifocal Imaging

We would like to explore the concept of multifocal imaging to obtain 3D resolution from a single image. In a microscope, an objective lens together with a tube lens is used to focus light from a sample onto a detector. Either the tube lens is no modified to shift the focal plane of the objective lens or one uses multiple tube lenses with different focal lengths with a single objective lens.

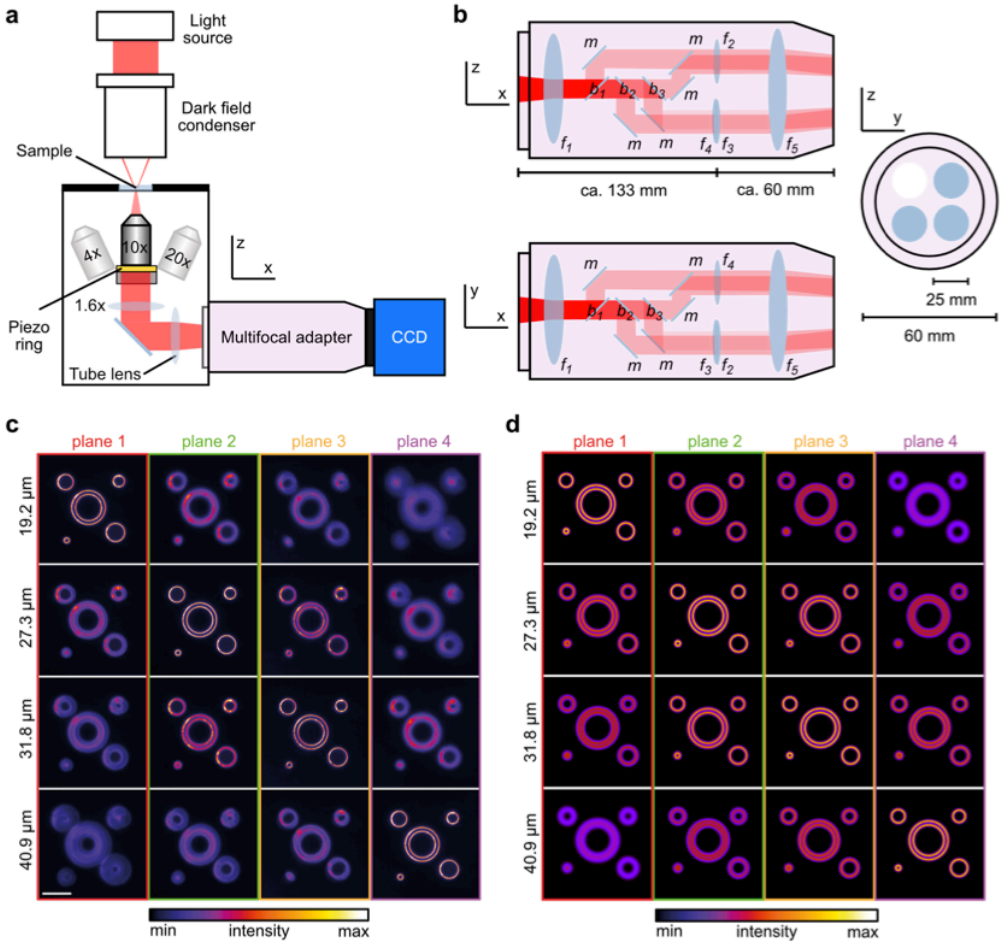


Figure 3.3: The multifocal adapter. (a) Composition of the multifocal imaging setup. The multifocal adapter (MFA) is inserted into the light path between a standard dark-field microscope and a camera. (b) Composition of the multifocal adapter. Incoming light from the microscope is infinity projected by the lens f1 and split by three consecutive beam-splitters (b1 = 25/75; b2 = 33/67; b3 = 50/50). Five mirrors (m) guide the light into four separate optical paths. The lens f5 projects the four images onto the camera chip. The inter-plane distance is set by lenses of different focal length (f1 - f4). (c, d) Experimentally acquired (c) and simulated (d) (see Methods) multifocal dark-field images (planes 1 - 4, 20x objective) of a calibration grid at the four focal positions (indicated on the left side), where one of the four focal planes maps the grid sharply. Scale bar represents 20 µm. Experiment was performed six times with similar results. Source

In the simplest configuration of an objective lens and a tube lens, the light first travels from the object to the objective lens in free space for a distance s, which is represented by

$$M_1 = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix}$$

Then it is passing the objective lens with

$$M_2 = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f_1} & 1 \end{bmatrix}$$

traveling further through free space with distance d

$$M_3 = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}$$

and then hitting the tube lens with

$$M_4 = \begin{bmatrix} 1 & 0 \\ -\frac{1}{\Delta_{ft}+f_t} & 1 \end{bmatrix}$$

where $\Delta_f t$ is a parameter modifying the original tube lens focal distance f_t. Finally, the light will propagate to the detector at distance f_t from the tube lens

$$M_s = \begin{bmatrix} 1 & f_t \\ 0 & 1 \end{bmatrix}$$

gicing for the whole system:

$$M_t = M_4 M_3 M_2 M_1$$

The final matrix has again 4 parameters A, B, C and D. For sharp imaging the parameter B of the system must be equal to zero, as the outgoing light should be independent of the incoming angle. This is achieved

$$\delta s = \frac{f_1 \left(\Delta_{ft} d + \Delta_{ft} f_t + f_t^2\right)}{\Delta_{ft} d - \Delta_{ft} f_1 + \Delta_{ft} f_t + f_t^2} - s_0$$

i The B Parameter

The **B parameter** is the most direct indicator of a sharp image in matrix optics. For an optical system with ABCD matrix:

$$\begin{pmatrix} y_2 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} y_1 \\ \theta_1 \end{pmatrix}$$

The output position is given by:

$$y_2 = A y_1 + B\theta_1$$

For an image to be sharp, all rays originating from the same object point must converge to a single image point, regardless of their initial angles. This means that for a given y_1 (object position), the final position y_2 must be independent of the initial angle θ_1.
This condition is satisfied when $\mathbf{B = 0}$.
When B = 0: - The final position depends only on the initial position ($y_2 = A y_1$) - All rays from a point source converge to a single point in the image plane - The imaging is “stigmatic” (point-to-point mapping)

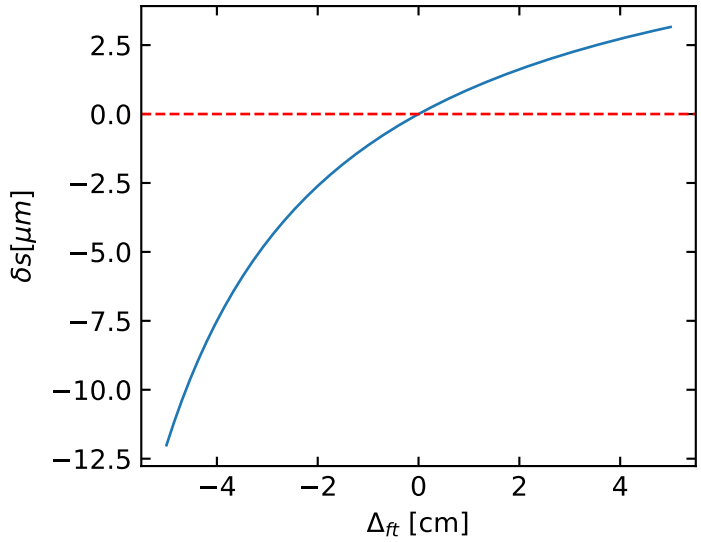


Figure 3.4: Plot of the solution with specific parameter values

In our lens example, the system matrix was:

$$\mathbf{M}_{system} = \begin{bmatrix} 1 - \frac{d}{f} & d \\ -\frac{1}{f} & 1 \end{bmatrix}$$

For parallel input rays (different y_1 but all $\theta_1 = 0$), we needed A = 0 to make them all converge to a single point, which happened when d = f.
But for general imaging of points (not just parallel rays), the B = 0 condition is what determines whether the image is sharp. This is why, in optical design, finding conjugate planes (where B = 0) is essential for sharp imaging.

Chapter 4

Theories for light

Wave Optics

Wave optics extends our understanding beyond the limitations of geometric optics by treating light as a wave phenomenon. This approach explains effects that cannot be accounted for by ray tracing alone, such as:

- Interference (the combination of waves)
- Diffraction (the bending of waves around obstacles or through apertures)
- Color (the wavelength-dependent nature of light)

Light is part of the electromagnetic spectrum, which spans an enormous range of frequencies. The visible region, extending approximately from 400 nm (violet) to 700 nm (red), represents only a small fraction of this spectrum. This wave description is essential for understanding many optical phenomena that geometric optics cannot explain, particularly when dealing with structures comparable in size to the wavelength of light.

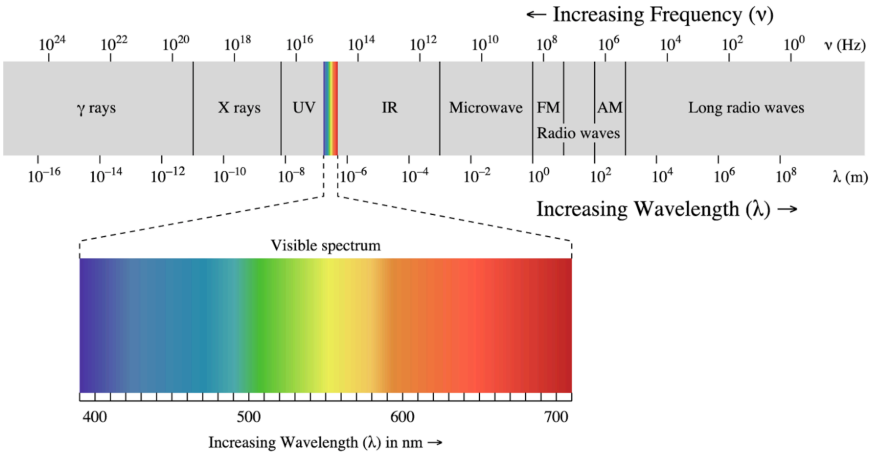


Figure 4.1: Electromagnetic Spectrum with its different regions

In the following, we would like to introduce wave by discarding the fact, that light is related to electric and magnetic fields. This is useful as the vectorial nature of the electric and magnetic field further complicates the calculations, but we do not need those yet. Accordingly we also do not understand how light really interacts with matter and we therefore have to introduce some postulates as well.

4.1 Postulates of Wave Optics

> **Wave**
>
> A wave corresponds to a physical quantity which oscillates in space and time. Its energy current density is related to the square magnitude of the amplitude. A wave satisfies the wave equation.

Wave equation

$$\nabla^2 u - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = 0$$

where the Laplace operator ∇^2 is defined as:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

The wave equation is a linear differential equation, which implies that the superposition principle holds. Specifically, if $u_1(\mathbf{r}, t)$ and $u_2(\mathbf{r}, t)$ are solutions of the wave equation, then any linear combination:

$$u(\mathbf{r}, t) = a_1 u_1(\mathbf{r}, t) + a_2 u_2(\mathbf{r}, t)$$

is also a solution, where a_1 and a_2 are arbitrary constants.

Monochromatic Wave

A monochromatic wave consists of a single frequency ω. By definition, such a wave must be infinite in time and free from phase disturbances (such as sudden jumps). The mathematical expression for a monochromatic wave is:

$$u(\mathbf{r}, t) = a(\mathbf{r}) \cos(\omega t + \phi(\mathbf{r}))$$

where:

- $a(\mathbf{r})$ represents the amplitude
- $\phi(\mathbf{r})$ represents the spatial phase
- ω represents the angular frequency

Complex Amplitude

The wave can be represented in complex form as:

$$U(\mathbf{r}, t) = a(\mathbf{r}) e^{i\phi(\mathbf{r})} e^{i\omega t}$$

This is known as the complex wavefunction.

> **Note**
>
> A phasor displays the complex amplitude with magnitude and phase as a vector in the complex plane.

The relationship between the complex and real wavefunctions is:

$$u(\mathbf{r}, t) = \mathrm{Re}\{U(\mathbf{r}, t)\} = \frac{1}{2}[U(\mathbf{r}, t) + U^*(\mathbf{r}, t)]$$

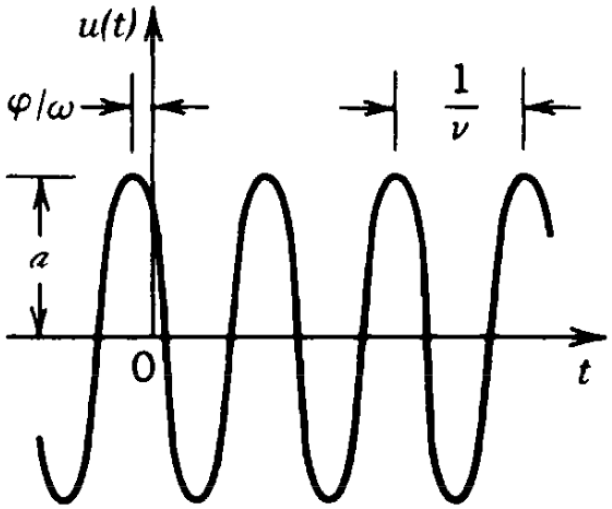


Figure 4.2: Representation of a wavefunction over time (constant position) denoting the phase ϕ and the period $T = 1/\nu$

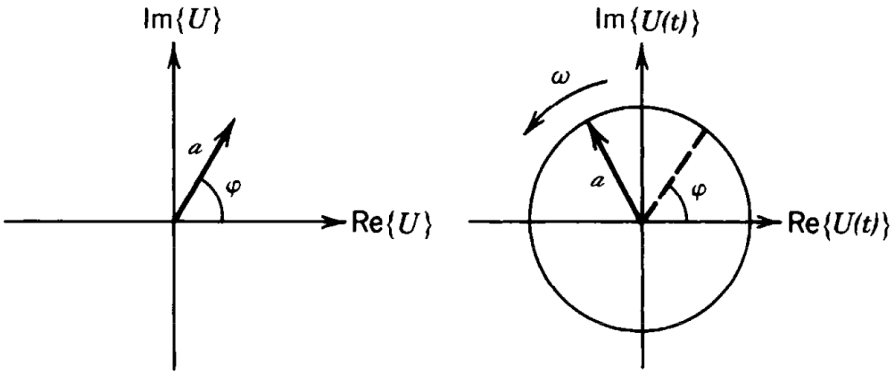


Figure 4.3: Phasor diagram of the complex amplitude $U(\mathbf{r})$ (left) and $U(t)$ (right)

The complex wavefunction satisfies the same wave equation:

$$\nabla^2 U - \frac{1}{c^2}\frac{\partial^2 U}{\partial t^2} = 0$$

We can separate the complex wavefunction into spatial and temporal components:

$$U(\mathbf{r}, t) = U(\mathbf{r})e^{i\omega t}$$

where

$$U(\mathbf{r}) = a(\mathbf{r})e^{i\phi(\mathbf{r})}$$

Here, ϕ represents the spatial phase of the wavefunction. Substituting this into the wave equation and noting that the time derivatives bring down factors of $i\omega$:

$$\nabla^2[U(\mathbf{r})e^{i\omega t}] - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}[U(\mathbf{r})e^{i\omega t}] = 0$$

$$\nabla^2 U(\mathbf{r})e^{i\omega t} + \frac{\omega^2}{c^2}U(\mathbf{r})e^{i\omega t} = 0$$

The time dependence $e^{i\omega t}$ factors out, leaving us with **the Helmholtz equation**:

$$\nabla^2 U(\mathbf{r}) + k^2 U(\mathbf{r}) = 0$$

where $k = \omega/c$ is the wave number. This equation describes the spatial behavior of monochromatic waves.

Intensity of Waves

The intensity of a wave at position $\mathbf{r}$ and time t is defined as:

$$I(\mathbf{r}, t) = 2\langle u^2(\mathbf{r}, t)\rangle$$

where I is measured in units of $\left[\frac{W}{m^2}\right]$. The angle brackets $\langle ...\rangle$ represent a time average over one oscillation cycle of u. For visible light, this averaging occurs over an extremely brief period - for example, light with a wavelength of 600 nm has a cycle duration of just 2 femtoseconds.

The optical power P of a wave can be calculated by integrating the intensity over a surface area A:

$$P = \int_A I(\mathbf{r}, t)\, dA$$

Inserting the seperation of the complex wavefunction into spatial and temporal components leads to the following expression for the intensity:

$$I(\mathbf{r}) = |U(\mathbf{r})|^2$$

Thus the physical quantity forming the spatial and temporal oscillation of the wavefunction is also providing the intensity of the wave when its magnitude is squared. This is a fundamental property of wavefunctions and for example not the case when temperature oscillates in space and time in a medium.

Wavefronts

Wavefronts are surfaces in space where the phase is constant:

$$\phi(\mathbf{r}) = \text{const}$$

Typically, this constant is chosen to represent points of maximum spatial amplitude, such that:

$$\phi(\mathbf{r}) = 2\pi q$$

where q is an integer.

The direction normal to these wavefronts can be described by the gradient vector:

$$\mathbf{n} = \nabla\phi = \left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}, \frac{\partial\phi}{\partial z}\right)$$

This vector $\mathbf{n}$ is always perpendicular to the wavefront surface and points in the direction of wave propagation. The evolution of these wavefronts in time provides important information about the wave's propagation characteristics.

4.2 Plane Waves

A plane wave represents a fundamental solution of the homogeneous wave equation. In its complex form, it is expressed as:

$$U(\mathbf{r}, t) = Ae^{-i\mathbf{k}\cdot\mathbf{r}}e^{i\omega t} \tag{4.1}$$

where:

- The first exponential term contains the spatial phase
- The second exponential term contains the temporal phase
- A is the (potentially complex) amplitude of the plane wave

The wavefront of a plane wave is defined by:

$$\mathbf{k}\cdot\mathbf{r} = 2\pi q + \arg(A)$$

where 1 is an integer. It just means that the projection of the position vector $\mathbf{r}$ onto the wavevector $\mathbf{k}$ is a multiple of 2π. This equation describes a plane perpendicular to the wavevector $\mathbf{k}$. Adjacent wavefronts are separated by the wavelength $\lambda = 2\pi/k$, where k represents the spatial frequency of the wave oscillation.

The spatial component of the plane wave is given by:

$$U(\mathbf{r}) = Ae^{-i\mathbf{k}\cdot\mathbf{r}} \tag{4.2}$$

In vacuum, the wavevector $\mathbf{k} = \mathbf{k}_0$ is real-valued and can be written as:

$$\mathbf{k}_0 = \begin{pmatrix} k_{0x} \\ k_{0y} \\ k_{0z} \end{pmatrix} \tag{4.3}$$

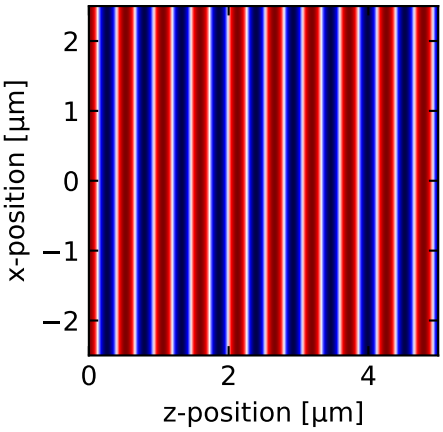


Figure 4.4: Plane wave propagating along the z-direction

4.3 Dispersion Relation

Using the plane wave solution

$$U(\mathbf{r}, t) = Ae^{-i\mathbf{k}\cdot\mathbf{r}}e^{i\omega t} \tag{4.4}$$

we can write down the sum of the spatial and temporal phase as

$$\phi(r, t) = \omega t - \mathbf{k} \cdot \mathbf{r}$$

If we select a point on the wavefront $\mathbf{r}_m$, and follow that over time, the phase $\phi(t) = \text{const}$. Taking the time derivative results in

$$\mathbf{k} \cdot \frac{d\mathbf{r}_m}{dt} = \omega$$

If we choose the direction of the wavevector for measuring the propagation speed, i.e. $\mathbf{r}_m = r_m\mathbf{e}_k$ then we find for the propagation speed

$$\frac{dr_m}{dt} = \frac{\omega}{k}$$

or in vacuum

$$c_0 = \frac{\omega}{k_0} \tag{4.5}$$

This fundamental relationship connects:

- The momentum (k),
- The energy (ω)

and is called a dispersion relation despite the fact, that we do not really understand why those quantities are related to energy and momentum.

> **Note**
>
> Light in free space exhibits a linear dispersion relation, i.e. the frequency of light changes linearly with the wavevector magnitude.

Note that if we choose a different propagation direction $\mathbf{e}$ than the one along the wavevector $\mathbf{e}_k$, we can write the phase velocity as

$$\mathbf{k} \cdot \mathbf{e} \frac{dr}{dt} = k \cos(\measuredangle \mathbf{k}, \mathbf{e}) \frac{dr}{dt} = \omega$$

or

$$\frac{dr}{dt} = \frac{\omega}{k \cos(\measuredangle \mathbf{k}, \mathbf{e})}$$

which means that if you observe the wavepropagation not in the direction of the wavevector, the phase velocity is actually bigger than the speed of light and even tends to infinity if the angle between the wavevector and the observation direction tends to 90°.

4.4 Propagation in a Medium

When a wave propagates through a medium:

1. The frequency ω remains constant (determined by the source)
2. The wave speed changes according to:
$$c = \frac{c_0}{n}$$
where n is the refractive index of the medium

This leads to changes in:

- the wavelength, which becomes shorter in the medium
$$\lambda = \frac{\lambda_0}{n}$$
- the length of the wavevector, which increases in the medium
$$k = nk_0$$

4.5 Snells Law

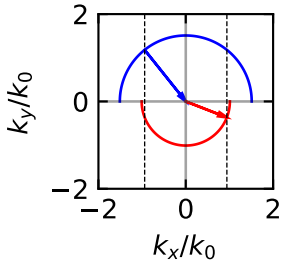
The change in the length of the wavevector has some simple consequence for Snells law. We can write Snells law as

$$n_1 k_0 \sin(\theta_1) = n_2 k_0 \sin(\theta_2)$$

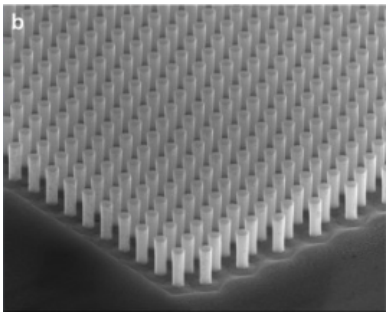
where k_0 is the wavevector length in vacuum. As the $n_1 k_0$ is the magnitude of the wavevector in medium 1, and $n_2 k_0$ is the magnitude of the wavevector in medium 2, we can rewrite Snells law as

$$k_1 \sin(\theta_1) = k_2 \sin(\theta_2)$$

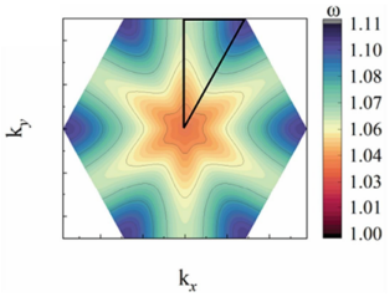
which means that the component of the wavevector parallel to the interface is conserved. If the wavevector has constant length then the wavevector incident at different angles is between a point on a circle and the origin in the diagram below. The circle corresponds to an isofrequency surface.



(a) Snells law construction using the conservation of the wavevector component parallel to the interface. The vertical dashed lines indicate the parallal component of the wavevector in the two media.



(a) Electron microscopy image of a 2D photonic crystal



(a) Isofrequency surfaces of a photonic crystal

Isofrequency surfaces can have non-spherical shape. In anisotropic media, they can be ellipsoids. In photonic crystals, i.e. crystals with a periodic structure on the scale of the wavelength, they can have a more complex shape.

4.6 Spherical Waves

A spherical wave, like a plane wave, consists of spatial and temporal components, but with wavefronts forming spherical surfaces. For spherical waves, $|\mathbf{k}||\mathbf{r}| = kr = \text{const}$. Given a source at position $\mathbf{r}_0$, the spherical wave can be expressed as:

$$U = \frac{A}{|\mathbf{r} - \mathbf{r}_0|} e^{-ik|\mathbf{r}-\mathbf{r}_0|} e^{i\omega t} \tag{4.6}$$

> **!** Important
>
> The $1/|\mathbf{r}-\mathbf{r}_0|$ factor in the amplitude is necessary for energy conservation - ensuring that the total energy flux through any spherical surface centered on the source remains constant.

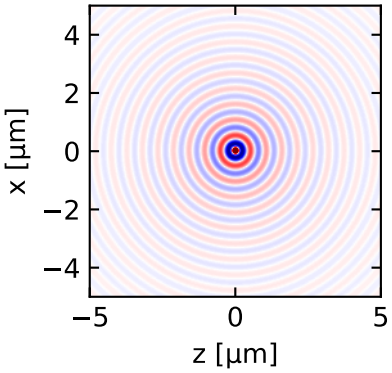


Figure 4.8: Spherical wave propagation. The wave is emitted from the origin and propagates in the positive z-direction. The wavefronts are spherical surfaces. The wave is visualized in the xz-plane.

Note: The direction of wave propagation can be reversed by changing the sign of the wavenumber k.

Part III

Lecture 3

Chapter 5

Interference in space and time

Interference is a fundamental physical phenomenon that demonstrates the superposition principle for linear systems. This principle, which states that the net response to multiple stimuli is the sum of the individual responses, is central to our understanding of wave physics. Interference appears across many domains of physics: in optics where it enables high-precision measurements and holography, in quantum mechanics where it reveals the wave nature of matter, and in acoustics where it forms the basis for noise cancellation technology. The ability of waves to interfere constructively (amplifying each other) or destructively (canceling each other) has profound practical applications, from the anti-reflective coatings on optical elements to the operational principles of interferometric gravitational wave detectors like LIGO. Understanding interference is therefore not just of theoretical interest but crucial for modern technology and experimental physics.

When two wave solutions $U_1(\mathbf{r})$ and $U_2(\mathbf{r})$ combine, their superposition gives:

$$U(\mathbf{r}) = U_1(\mathbf{r}) + U_2(\mathbf{r})$$

The resulting intensity is:

$$\begin{aligned} I &= |U|^2 && (5.1)\\ &= |U_1 + U_2|^2 && (5.2)\\ &= |U_1|^2 + |U_2|^2 + U_1^* U_2 + U_1 U_2^* && (5.3) \end{aligned}$$

The individual wave intensities are given by $I_1 = |U_1|^2$ and $I_2 = |U_2|^2$. Using this, we can express each complex wave amplitude in polar form, separating its magnitude (related to intensity) and phase:

$$U_1 = \sqrt{I_1} e^{i\phi_1}$$
$$U_2 = \sqrt{I_2} e^{i\phi_2}$$

Substituting these expressions back into our interference equation and performing the algebra, the total intensity becomes:

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\Delta\phi)$$

where $\Delta\phi = \phi_2 - \phi_1$ is the phase difference between the waves. This equation is known as the interference formula and contains three terms:

- I_1 and I_2: the individual intensities
- $2\sqrt{I_1 I_2}\cos(\Delta\phi)$: the interference term that can be positive or negative

A particularly important special case occurs when the interfering waves have equal intensities ($I_1 = I_2 = I_0$). The equation then simplifies to:

$$I = 2I_0(1 + \cos(\Delta\phi)) = 4I_0 \cos^2\left(\frac{\Delta\phi}{2}\right)$$

This last form clearly shows that:

- Maximum intensity ($4I_0$) occurs when $\Delta\phi = 2\pi n$ (constructive interference)
- Zero intensity occurs when $\Delta\phi = (2n + 1)\pi$ (destructive interference)
- The intensity varies sinusoidally with the phase difference

Constructive Interference

Occurs when $\Delta\phi = 2\pi m$ (where m is an integer), resulting in $I = 4I_0$

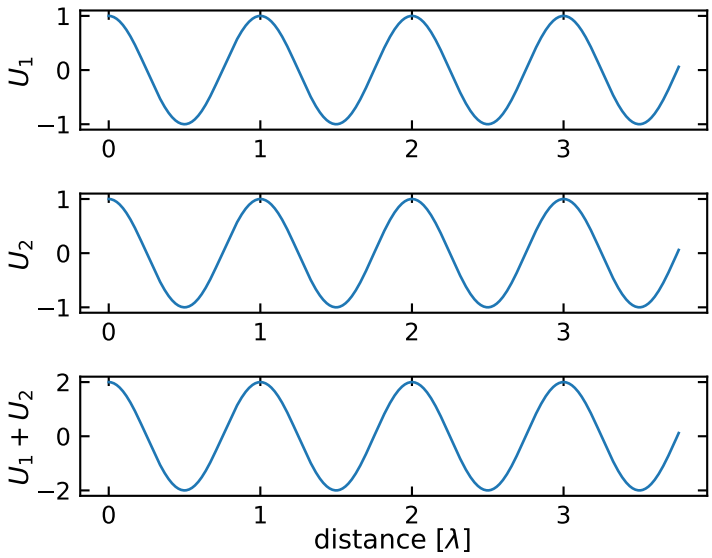


Figure 5.1: Constructive interference of two waves (top, middle) and the sum of the two wave amplitudes (bottom)

Destructive Interference

Occurs when $\Delta\phi = (2m - 1)\pi$ (where m is an integer), resulting in $I = 0$

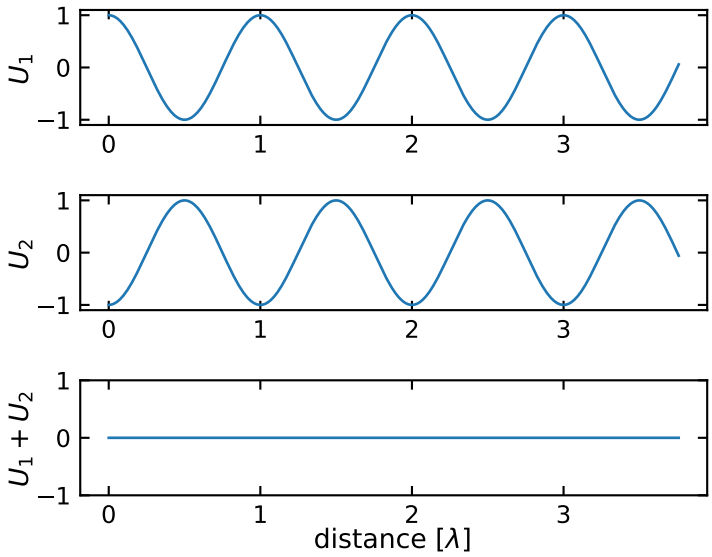


Figure 5.2: Destructive interference of two waves (top, middle) and the sum of the two wave amplitudes (bottom)

Phase and Path Difference

The phase difference $\Delta\phi$ can be related to the path difference Δs between the two waves. For two waves with the same frequency ω, we can write their complete phase expressions as:

$$\phi_1(\mathbf{r}, t) = \mathbf{k}_1 \cdot \mathbf{r} - \omega t + \phi_{01}$$

$$\phi_2(\mathbf{r}, t) = \mathbf{k}_2 \cdot \mathbf{r} - \omega t + \phi_{02}$$

where:

- $\mathbf{k}_i$ are the wave vectors
- $\mathbf{r}$ is the position vector
- ω is the angular frequency
- ϕ_{0i} are initial phase constants

The instantaneous phase difference is then:

$$\Delta\phi(\mathbf{r}, t) = \phi_2(\mathbf{r}, t) - \phi_1(\mathbf{r}, t) = (\mathbf{k}_2 - \mathbf{k}_1) \cdot \mathbf{r} + (\phi_{02} - \phi_{01})$$

For stationary interference patterns, we typically observe the time-independent phase difference. When the waves travel along similar paths (same direction), this reduces to:

$$\Delta\phi = k\Delta s + \Delta\phi_0$$

where Δs is the path difference and $\Delta\phi_0$ is any initial phase difference between the sources.

! Phase Difference and Path Difference

A path difference Δs corresponds to a phase difference $k\Delta s = 2\pi\Delta s/\lambda$. Path differences of integer multiples of λ result in phase differences of integer multiples of 2π.

Interference of Waves in Space

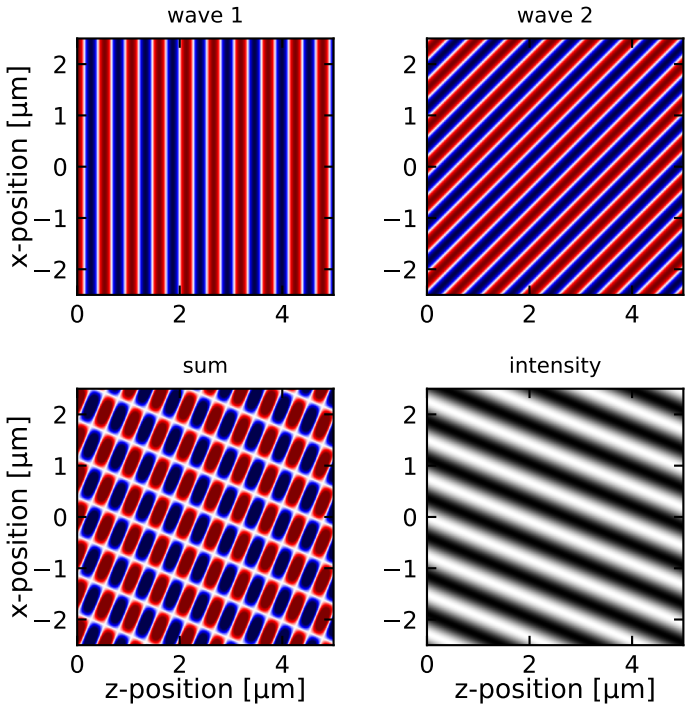


Figure 5.3: Interference of two plane waves propagating under an angle of 45°. The two left graphs show the original waves. The two right show the total amplitude and the intensity pattern.

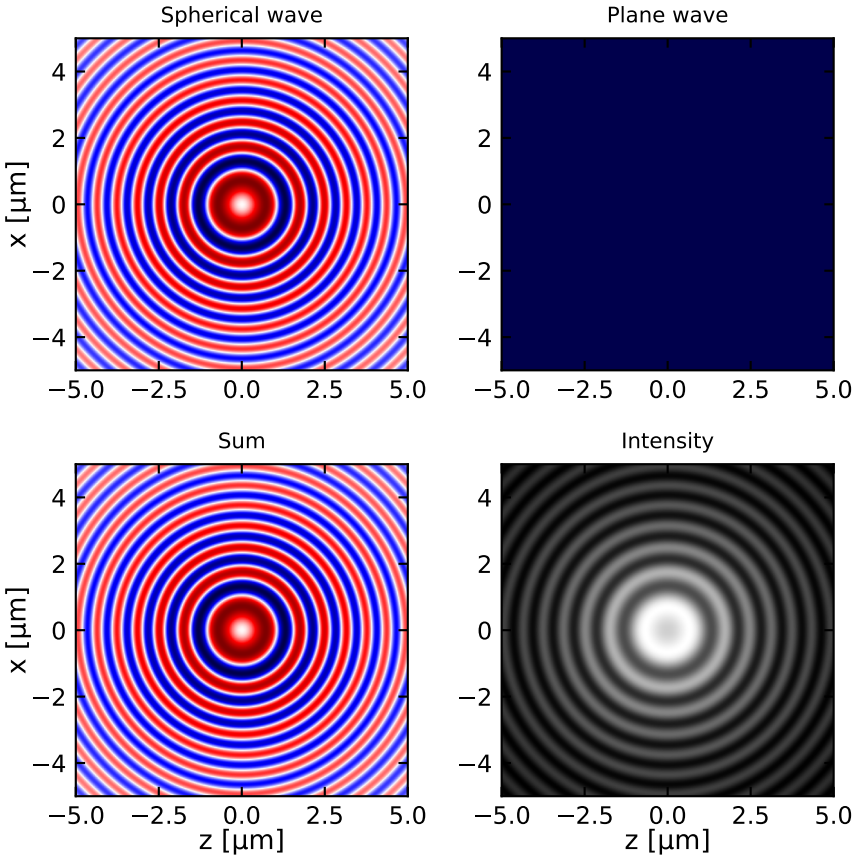


Figure 5.4: Interference of a spherical wave and a plane wave. The top graphs show the original waves. The two bottom show the total amplitude and the intensity pattern.

The interference of the spherical and the plane wave (also the one of the two plane waves) give also an interesting result. The intensity resembles to be a snapshot of the shape of the wavefronts of the spherical wave. We can therefore measure the wavefronts of the spherical wave by interfering it with a plane wave. This is also the basic principle behind holography. There we use a reference wave to interfere with the wave that we want to measure. The interference pattern is recorded and can be used to reconstruct the wavefronts of the wave.

A super nice website to try out interference interactively is here.

Coherence

In the earlier consideration we obtained a general description for the phase difference between two waves. TIt is given by and contains the pathlength difference Δs and some intrinsic phase $\Delta\phi_0$ that could be part of the wave generation process.

$$\Delta\phi = k\Delta s + \Delta\phi_0$$

To observe stationary interference, it is important that these two quantities are also stationary, i.e. the phase relation between the two waves is stationary. This relation between the phase of two waves is called coherence and was assumed in all the examples before.

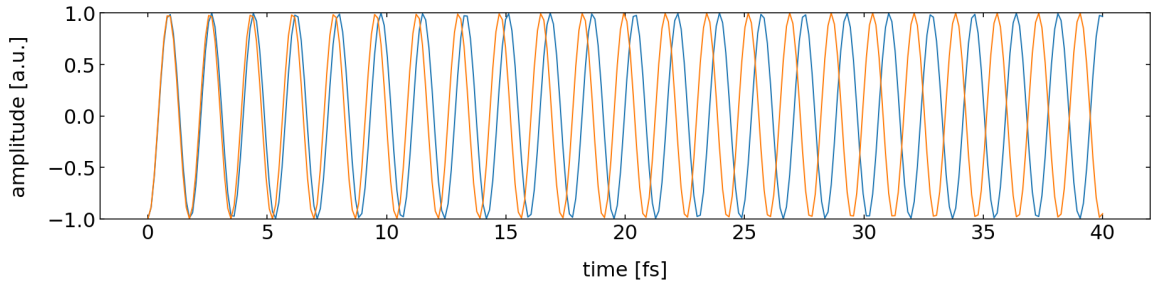


Figure 5.5: Two waves of different frequency over time.

The above image shows the timetrace of the amplitude of two wave with slightly different frequency. Due to the frequency, the waves run out of phase and have acquired a phase different of π after 40 fs.

The temporal coherence of two waves is now defined by the time it takes for the two waves to obtain a phase difference of 2π. The phase difference between two wave of frequency ν_1 and ν_2 is given by

$$\Delta\phi = 2\pi(\nu_2 - \nu_1)(t - t_0)$$

Here t_0 refers to the time, when thw two waves were perfectly in sync. Lets assume that the two frequencies are seperarated from a central frequency ν_0 such that

$$\nu_1 = \nu_0 - \Delta\nu/2$$
$$\nu_2 = \nu_0 + \Delta\nu/2$$

Inserting this into the first equation yields

$$\Delta\phi = 2\pi\Delta\nu\Delta t$$

with $\Delta t = t - t_0$. We can now define the coherence time as the time interval over which the phase shift $\Delta\phi$ grows to 2π, i.e. $\Delta\phi = 2\pi$. The coherence time is thus

$$\tau_c = \Delta t = \frac{1}{\Delta\nu}$$

Thus the temporal coherence and the frequency distribution of the light are intrisincly connected. Monochromatic light has $\Delta nu = 0$ and thus the coherence time is infinitely long. Light with a wide spectrum (white light for example) therefore has and extremly short coherence time.

The coherence time is also connected to a coherence length. The coherence length L_c is given by the distance light travels within the coherence time τ_c, i.e.

$$L_c = c\tau_c$$

Coherence

Two waves are called coherent, if they exihibit a fixed phase relation in space or time relation over time. It measures their ability to interfer. The main types of coherence are

Temporal Coherence

- Measures phase correlation of a wave with itself at different times
- Characterized by coherence time τ_c and coherence length $L_c = c\tau_c$

- Related to spectral width: $\tau_c = 1/\Delta\nu$
- Perfect for monochromatic waves (single frequency)
- Limited for broad spectrum sources (like thermal light)

Spatial Coherence

- Measures phase correlation between different points in space
- Important for interference from extended sources
- Determines ability to form interference patterns
- Related to source size and geometry

Coherence is a property of the light source and is connected to the frequency distribution of the light. Sources can be:

- **Fully coherent**: ideal laser
- **Partially coherent**: real laser
- **Incoherent**: thermal light

i More General Description of Coherence

While the above definition provides an intuitive picture based on frequency spread, we can describe coherence more rigorously using correlation functions. These functions measure how well a wave maintains its phase relationships:
In real physical systems, perfect coherence (constant phase relationship) between waves is rare. Partial coherence describes the degree to which waves maintain a consistent phase relationship over time and space. We can characterize this using correlation functions:

1. **Temporal Coherence** The complex degree of temporal coherence is given by:

$$g^{(1)}(\tau) = \frac{\langle U(t)U^*(t+\tau)\rangle}{\sqrt{\langle|U(t)|^2\rangle\langle|U(t+\tau)|^2\rangle}}$$

where:

- τ is the time delay
- $U(t)$ is the electric field
- $\langle ...\rangle$ denotes time averaging

2. **Spatial Coherence** Similarly, spatial coherence between two points is characterized by:

$$g^{(1)}(\mathbf{r}_1, \mathbf{r}_2) = \frac{\langle U(\mathbf{r}_1)U^*(\mathbf{r}_2)\rangle}{\sqrt{\langle|U(\mathbf{r}_1)|^2\rangle\langle|U(\mathbf{r}_2)|^2\rangle}}$$

The obtained correlation functions can be used to calculate the coherence time and length and have the following properties:

- $|g^{(1)}| = 1$ indicates perfect coherence
- $|g^{(1)}| = 0$ indicates complete incoherence
- $0 < |g^{(1)}| < 1$ indicates partial coherence

A finite coherence time and length is leads to partial coherence affects interference visibility through:

- Reduced contrast in interference patterns
- Limited coherence length/area
- Spectral broadening

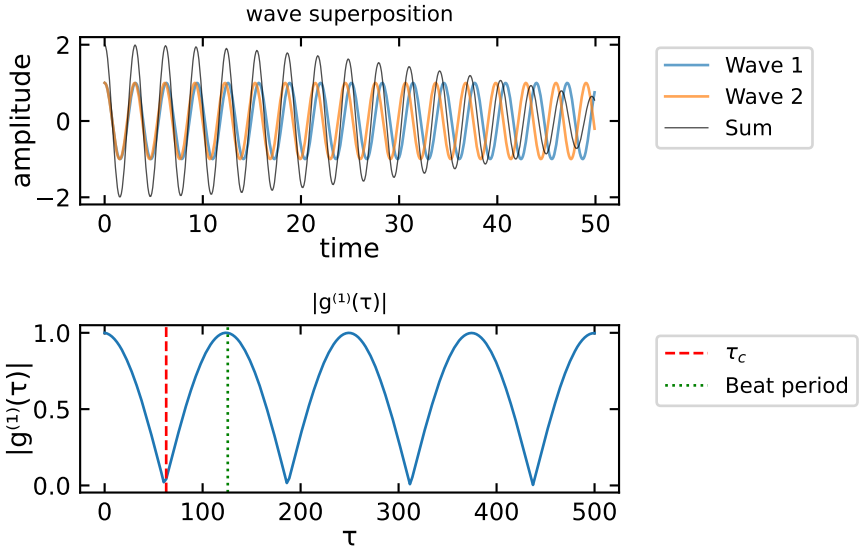


Figure 5.6: Temporal correlation for two waves with slightly different frequencies. The vertical line indicates the coherence time $c = /\Delta$.

Besides different frequencies the coherence time can also be affected by phase jumps. The following example shows two waves with the same frequency but multiple phase jumps. The temporal correlation function shows the decoherence due to the phase jumps.

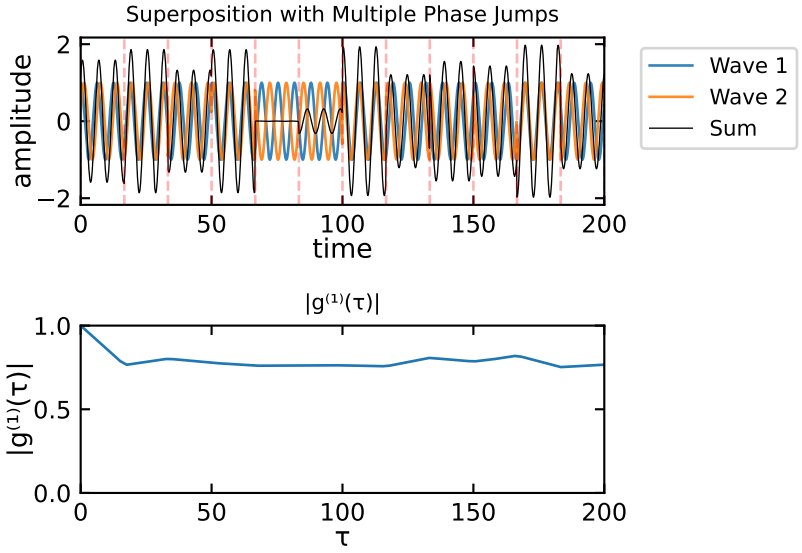


Figure 5.7: Temporal correlation for two waves of same frequency showing decoherence due to multiple phase jumps. Vertical lines indicate positions of phase jumps.

Multiple Wave Interference

So far we looked at the interference of two waves, which was a simplification as I mentioned already earlier. Commonly there will be a multitude of partial waves contribute to the oberved intereference. This is what we would like to have a look at now. We will do that in a quite general fashion, as the resulting formulas will appear several times again for different problems.

Nevertheless we will make a difference between

- multiwave interference of waves with the constant amplitude
- multiwave interference of waves with decreasing amplitude

Especially the latter is often occuring, if we have multiple reflections and each reflection is only a fraction of the incident amplitude.

Multiple Wave Interference with Constant Amplitude

In the case of constant amplitude (for example realized by a grating, which we talk about later), the total wave amplitude is given according to the picture below by

$$U = U_1 + U_2 + U_1 + U_3 + ... + U_M$$

where we sum the amplitude over M partial waves. Between the neighboring waves (e.g. U_1 and U_2), we will assume a phase difference (because of a path length difference for example), which we denote as $\Delta\phi$.

The amplitude of the p-th wave is then given by

$$U_p = \sqrt{I_0} e^{i(p-1)\Delta\phi}$$

with the index p being an interger $p = 1, 2, ..., M$, $h = e^{i\Delta\phi}$ and $\sqrt{I_0}$ as the amplitude of each individual wave. The total amplitude U can be then expressed as

$$U = \sqrt{I_0}\,(1 + h + h^2 + ... + h^{M-1})$$

which is a geometric sum. We can apply the sum formula for geometric sums to obtain

$$U = \sqrt{I_0}\frac{1 - h^M}{1 - h} = \sqrt{I_0}\frac{1 - e^{iM\Delta\phi}}{1 - e^{i\Delta\phi}}$$

We now have to calculate the intensity of the total amplitude

$$I = |U|^2 = I_0 \left| \frac{e^{-iM\Delta\phi/2} - e^{iM\Delta\phi/2}}{e^{-i\Delta\phi/2} - e^{i\Delta\phi/2}} \right|^2$$

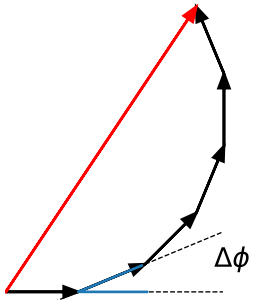
which we can further simplify to give

$$I = I_0 \frac{\sin^2(M\Delta\phi/2)}{\sin^2(\Delta\phi/2)}$$

The result is therefore an oscillating function. The numerator $\sin^2(M\Delta\phi/2)$ shows and oscillation frequency, which is by a factor of M higher than the one in the denominator $\sin^2(\Delta\phi/2)$. Therefore the intensity pattern is oscillating rapidly and creating a first minimum at

$$\Delta\phi = \frac{2\pi}{M}$$

This is an important result, since it shows that the number of sources M determines the position of the first minimum and the interference peak gets narrower with increasing M. Since the phase difference $\Delta\phi$ between neighboring sources is the same as for the double slit experiment, i.e. $\Delta\phi = 2\pi d/\lambda \sin(\theta)$, we can also determine the angular position of the first minimum. This is given by



(a) Multiple wave interference of $M = 6$ waves with a phase difference of $\phi = \pi/8$. The black arrows represent the individual waves, the red arrow the sum of all waves.

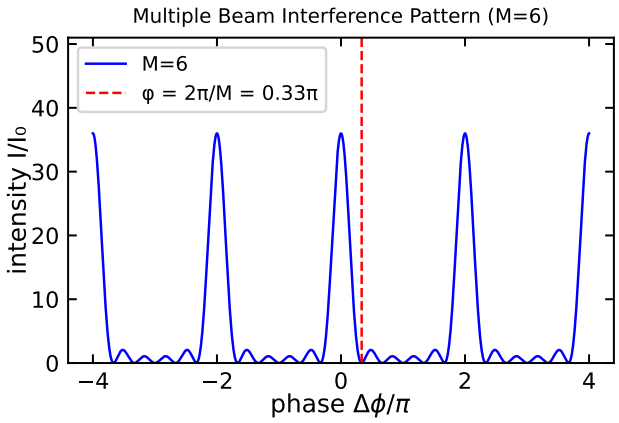


Figure 5.9: Multiple beam interference pattern for M=6 beams. The intensity distribution is shown as a function of the phase shift ϕ. The first minimum is at $\phi = 2\pi/M$. The intensity distribution is symmetric around $\phi = 0$.

$$\sin(\theta_{\min}) = \frac{1}{M}\frac{\lambda}{d}$$

This again has the common feature that it scales as λ/d. A special situation occurs, whenever the numerator and the denominator become zero. This will happen whenever

$$\Delta\phi = m2\pi$$

where m is an integer and denotes the interference order, i.e. the number of wavelength that neighboring partial waves have as path length difference. In this case, the intensity distributiion will give us

$$I = I_0\frac{0}{0}$$

and we have to determine the limit with the help of l'Hospitals rule. The outcome of this calculation is, that

$$I(\Delta\phi = m2\Delta\pi) = M^2 I_0$$

which can be also realized when using the small angle approximation for the sine functions.

Wavevector Representation

We would like to introduce a different representation of the multiple wave interference of the grating, which is quite insightful. The first order ($m = 1$) constructive interference condition is given by

$$\frac{1}{\lambda}\sin\theta = \frac{1}{d}$$

which also means that

$$\frac{2\pi}{\lambda}\sin\theta = \frac{2\pi}{d}$$

This can be written as

$$k \sin\theta = K$$

where k is the magnitude of the wavevector of the light and K is the wavevector magnitude that corresponds to the grating period d. As the magnitude of the wavevector of the light is conserved, the wavevectors of the incident light and the light traveling along the direction of the first interence peak form the sides of an equilateral triangle. This is shown in the following figure.

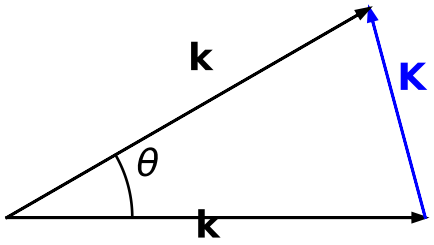


Figure 5.10: Wavevector summation for the diffraction grating. The wavevector of the incident light k and the wavevector of the light traveling along the direction of the first interference peak K form an equilateral triangle.

This means that the diffraction grating is providing a wavevector K to alter the direction of the incident light. This is again a common feature reappearing in many situations as for example in the X-ray diffraction of crystals.

i Multiple Wave Interference with Decreasing Amplitude

We will turn our attention now to a slight modification of the previous multiwave interference. We will introduce a decreasing amplitude of the individual waves. The first wave shall have an amplitude $U_1 = \sqrt{I_0}$. The next wave, however, will not only be phase shifted but also have a smaller amplitude.

$$U_2 = hU_1$$

where $h = re^{i\phi}$ with $|h| = r < 1$. r can be regarded as a reflection coefficient, which deminishes the amplitude of the incident wave. According to that the intensity is reduced by

$$I_2 = |U_2|^2 = |hU_1|^2 = r^2 I_1$$

The intensity of the incident wave is multiplied by a factor r^2, while the amplitude is multiplied by r. Note that the phase factor $e^{i\Delta\phi}$ is removed when taking the square of this complex number.

Intensity at Boundaries

The amplitude of the reflected wave is diminished by a factor $r \leq 1$, which is called the reflection coefficient. The intensity is diminished by a factor $R = |r|^2 \leq 1$, which is the **reflectance**.

In the absence of absorption, reflectance R and **transmittance** T add to one due to energy conservation.

$$R + T = 1$$

Consequently, the third wave would be now $U_3 = hU_2 = h^2 U_1$. The total amplitude is thus

$$U = U_1 + U_2 + U_3 + ... + U_M = \sqrt{I_0}(1 + h + h^2 + ...)$$

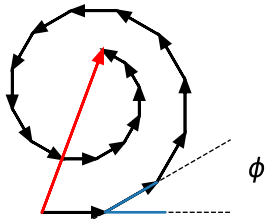


Figure 5.11: Phase construction of a multiwave intereference with M waves with decreasing amplitude due to a reflection coefficient $r = 0.95$.

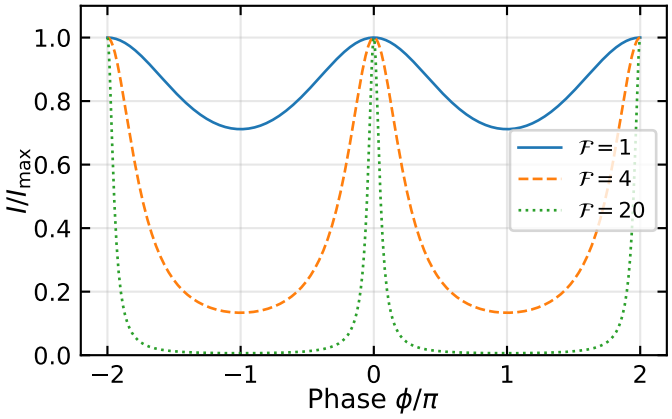


Figure 5.12: Multiple wave interference with decreasing amplitude. The graph shows the intensity distribution over the phase angle ϕ for different values of the Finesse $\mathcal{F}$.

This yields again

$$U = \sqrt{I_0}\frac{(1-h^M)}{1-h} = \frac{\sqrt{I_0}}{1-re^{i\Delta\phi}}$$

Calculating the intensity of the waves is giving

$$I = |U|^2 = \frac{I_0}{|1-re^{i\Delta\phi}|^2} = \frac{I_0}{(1-r)^2 + 4r\sin^2(\Delta\phi/2)}$$

which is also known as the **Airy function**. This function can be further simplified by the following abbrevations

$$I_{\max} = \frac{I_0}{(1-r)^2}$$

and

$$\mathcal{F} = \frac{\pi\sqrt{r}}{1-r}$$

where the latter is called the *Finesse*. With those abbrevations, we obtain

$$I = \frac{I_{\max}}{1 + 4\left(\frac{\mathcal{F}}{\pi}\right)^2 \sin^2(\Delta\phi/2)}$$

for the interference of multiple waves with decreasing amplitude.

This intensity distribution has a different shape than the one we obtained for multiple waves with the same amplitude.
We clearly observe that with increasing Finesse the intensity maxima, which occur at multiples fo π get much narrower. In addition the regions between the maxima show better contrast and fopr higher Finesse we get complete destructive interference.

Light beating

Beating of two waves

Let us consider now interference in the time domain. We introduce two monochromatic waves of frequencies ν_1 and ν_2. We will denote their amplitudes by $\sqrt{I_1}$ and $\sqrt{I_2}$.

The total amplitude is thus

$$U = U_1 + U_2 = \sqrt{I_1}\exp(i2\pi\nu_1 t) + \sqrt{I_2}\exp(i2\pi\nu_2 t)$$

such that we obtain an Intensity

$$I = |U|^2 = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(2\pi(\nu_1 - \nu_2)t)$$

The intensity is thus time dependent and oscillates at a frequency $\nu_1 - \nu_2$, which is the so-called beating frequency. Similar schemes are used in **optical heterodyne detection** but also in acoustics when tuning your guitar.

Multiple wave beating and pulse generation

Consider now a whole set of $M = 2L + 1$ each with an amplitude $\sqrt{I_0}$. The frequencies of the waves are given by $\nu_q = \nu_0 + q\Delta\nu$ with $q = -L, \dots, L$ with ν_0 beeing the center frequency of the spectrum and $\Delta\nu$ the frequency spacing. We will assume that $\Delta nu << \nu_0$ such that the total amplitude of the waves is given by

$$U = \sum_{q=-L}^{L} \sqrt{I_0}\exp(i2\pi(\nu_0 + q\Delta\nu)t)$$

The total intensity can then be calculated in the same way as for the multiple source in space before. Using $\phi = 2\pi\Delta\nu t$ we obtained

$$I(t) = I_0 \frac{\sin^2(M\pi t/T)}{\sin^2(\pi t/T)}$$

with $T = 1/\Delta\nu$ and a maximum intensity of $I_{\max} = M^2 I_0$.

Frequency Combs: Phase-Coherent Temporal Interference

The pulse generation we just examined leads us to an important concept in modern optics: frequency combs. A frequency comb is a spectrum consisting of a series of discrete, equally spaced frequency lines that results from a train of phase-coherent pulses in the time domain.

From Pulse Trains to Frequency Combs

Let's extend our understanding of multiple wave beating to include phase coherence. When we have a set of equidistant frequency components that maintain a fixed phase relationship, the resulting time-domain signal is a periodic train of pulses. Conversely, a periodic train of pulses in the time domain corresponds to a frequency comb in the spectral domain.

The relationship between these domains is described by the Fourier transform. For a pulse train with repetition rate $f_{rep} = \Delta\nu$ (the spacing between frequency components), the frequency spectrum consists of lines at:

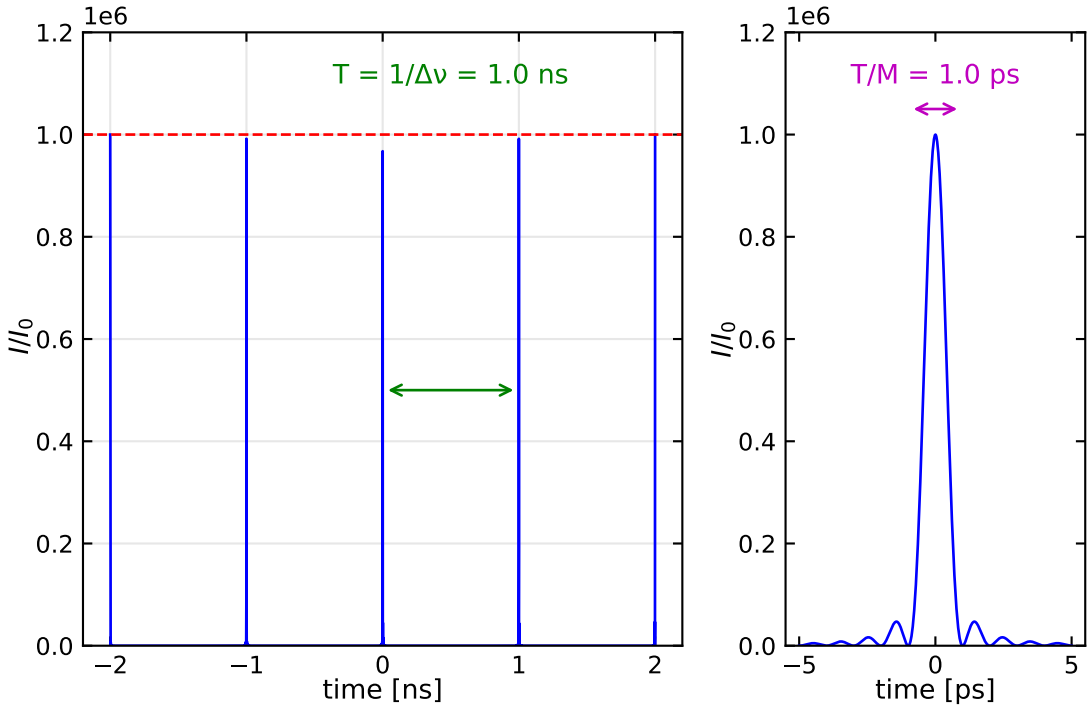


Figure 5.13: Multiple wave beating with M=1000 monochromatic waves separated by Δ =1 GHz. The intensity oscillates with period T=1/Δ =1 ns. Each pulse has a width of approximately T/M=1 ps with maximum intensity I_max=M²I .

$$f_n = f_0 + n \cdot f_{rep}$$

where f_0 is the carrier-envelope offset frequency and n is an integer.

Mathematical Description

Consider a train of pulses described by the electric field:

$$E(t) = \sum_{m=-\infty}^{\infty} A(t - mT_{rep}) e^{i[\omega_c(t-mT_{rep})+\phi_{CE}\cdot m+\phi_0]}$$

where: - $A(t)$ is the pulse envelope - $T_{rep} = 1/f_{rep}$ is the pulse repetition period - ω_c is the carrier frequency - ϕ_{CE} is the carrier-envelope phase slip (the phase shift from pulse to pulse) - ϕ_0 is a constant phase

The Fourier transform of this pulse train gives us a frequency comb with:

$$f_0 = \frac{\phi_{CE}}{2\pi} \cdot f_{rep}$$

This carrier-envelope offset frequency (f_0) is crucial for determining the absolute positions of the comb lines.

Applications of Frequency Combs

Frequency combs have revolutionized precision measurements in physics and enabled numerous applications:

1. **Optical Clocks**: Frequency combs provide a “gear mechanism” to count optical frequencies, enabling optical atomic clocks that are orders of magnitude more precise than conventional atomic clocks.

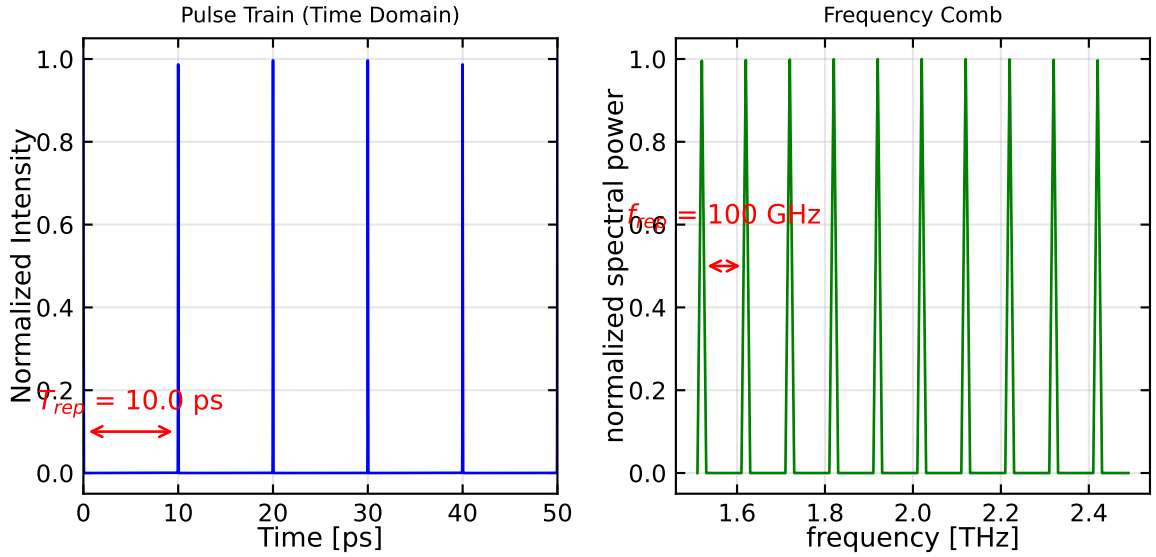


Figure 5.14: Demonstration of a frequency comb. (Left) Time domain representation showing a train of phase-coherent pulses. (Right) Frequency domain representation showing equally spaced frequency lines forming a comb structure.

2. **Precision Spectroscopy**: The precise and stable frequency references allow for high-resolution molecular spectroscopy.
3. **Calibration of Astronomical Spectrographs**: Frequency combs enable the detection of Earth-like exoplanets by providing precise wavelength calibration for astronomical instruments.
4. **Distance Measurements**: They enable precise absolute distance measurements used in applications from gravitational wave detectors to satellite formation flying.
5. **Telecommunications**: Frequency combs can be used for wavelength-division multiplexing in optical communications.

Connection to Mode-Locked Lasers

In practice, frequency combs are often generated using mode-locked lasers. In such lasers, multiple longitudinal modes of the laser cavity oscillate with a fixed phase relationship, resulting in short pulses. The Fourier transform of these regularly spaced pulses is precisely the frequency comb.

The mode-locking can be achieved through various mechanisms:

- **Active mode-locking**: Using an external modulator driven at the cavity round-trip frequency
- **Passive mode-locking**: Using saturable absorbers or Kerr-lens mode-locking

The 2005 Nobel Prize in Physics was awarded to Theodor W. Hänsch and John L. Hall for their contributions to the development of laser-based precision spectroscopy, including the optical frequency comb technique.

The Fundamental Link Between Time and Frequency Domains

Frequency combs beautifully illustrate the duality between time and frequency domains in physics. A perfectly periodic sequence of events in time (the pulse train) corresponds to discrete, equally spaced frequencies. The more precise and stable the temporal pattern, the more precise and stable the frequency components.

This duality is fundamental to many areas of physics and engineering, from quantum mechanics to signal processing, and frequency combs represent one of its most elegant and useful manifestations in optics.

Part IV

Lecture 4

Chapter 6

Introduction to Gaussian Beams

In optics and laser physics, Gaussian beams represent one of the most fundamental and important mathematical descriptions of laser light propagation. They are particularly relevant for understanding laser resonators, optical systems, and coherent light behavior. This section introduces Gaussian beams from first principles and explores their mathematical description.

Derivation from the Helmholtz Equation

We begin with the Helmholtz equation, which describes monochromatic electromagnetic waves in a homogeneous medium:

$$\nabla^2 U + k^2 U = 0$$

where U represents the electric field component, $k = 2\pi/\lambda$ is the wave number, and λ is the wavelength of light. For a wave predominantly traveling along the z-axis, we can express the electric field as:

$$U(x, y, z) = u(x, y, z)e^{-ikz}$$

Here, $u(x, y, z)$ is a complex amplitude function that varies slowly with z compared to the wavelength. Substituting this into the Helmholtz equation and expanding the Laplacian operator yields:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)(ue^{-ikz}) + k^2(ue^{-ikz}) = 0$$

Computing the derivatives and simplifying:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} - 2ik\frac{\partial u}{\partial z} = 0$$

The Paraxial Approximation

The paraxial approximation applies when the beam's angular spread is small, meaning the wavefronts are nearly perpendicular to the propagation axis. Mathematically, this means that the amplitude u varies slowly along the propagation direction compared to transverse directions:

$$\left|\frac{\partial^2 u}{\partial z^2}\right| \ll \left|2k\frac{\partial u}{\partial z}\right|$$

Under this approximation, the Helmholtz equation simplifies to the paraxial Helmholtz equation:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - 2ik\frac{\partial u}{\partial z} = 0$$

To solve this equation, we propose the ansatz:

$$u(x, y, z) = A(z) \exp\left[-\frac{k}{2q(z)}(x^2 + y^2)\right]$$

where $A(z)$ and $q(z)$ are complex functions to be determined. Substituting this into the paraxial equation and solving the resulting differential equations:

$$\frac{dq}{dz} = 1 \quad \text{and} \quad \frac{dA}{dz} = -\frac{A}{q}$$

These yield solutions $q(z) = q_0 + z$ and $A(z) = \frac{A_0}{q(z)}$, where q_0 and A_0 are constants.

The complex beam parameter $q(z)$ relates to physical parameters through:

$$\frac{1}{q(z)} = \frac{1}{R(z)} - i\frac{\lambda}{\pi w^2(z)}$$

where $R(z)$ is the radius of curvature of the wavefront and $w(z)$ is the beam radius at which the intensity falls to $1/e^2$ of its axial value.

Setting $q_0 = iz_0$ where z_0 is the Rayleigh range, we can express these parameters as:

$$w(z) = w_0\sqrt{1 + \left(\frac{z}{z_0}\right)^2}$$

$$R(z) = z\left[1 + \left(\frac{z_0}{z}\right)^2\right]$$

where $w_0 = \sqrt{\frac{\lambda z_0}{\pi}}$ is the beam waist (minimum beam radius).

The complete Gaussian beam solution is:

$$U(x, y, z) = U_0\frac{w_0}{w(z)} \exp\left[-\frac{x^2 + y^2}{w^2(z)}\right] \exp\left[-ikz - ik\frac{x^2 + y^2}{2R(z)} + i\phi(z)\right]$$

where $\phi(z) = \arctan(z/z_0)$ is the Gouy phase shift, representing an additional phase beyond that of a plane wave.

In scalar wave theory, the intensity of the Gaussian beam is proportional to the square of the amplitude. It can be calculated as:

$$I(x, y, z) = |U(x, y, z)|^2 = I_0\frac{w_0^2}{w^2(z)} \exp\left[-\frac{2(x^2 + y^2)}{w^2(z)}\right]$$

where $I_0 = |U_0|^2$ is the peak intensity at the beam waist. This expression shows that the intensity has a Gaussian profile in any transverse plane, with its peak on the beam axis. The intensity falls to $1/e^2$ of its axial value at a radial distance $r = w(z)$ from the axis, which defines the beam radius. The total power carried by the beam is conserved during propagation, but the peak intensity decreases as $w(z)$ increases with distance from the waist.

Gaussian Beam Propagation in the x-z Plane

We can visualize how a Gaussian beam's intensity varies across both the propagation direction (z-axis) and transverse direction (x-axis) simultaneously using a 2D contour plot.

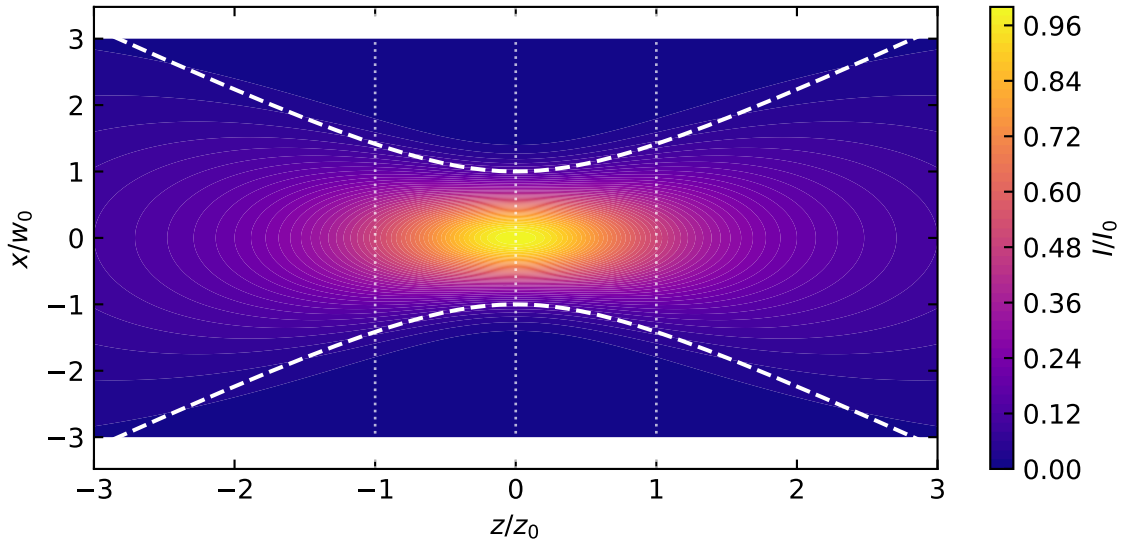


Figure 6.1: Gaussian Beam Intensity Distribution in the x-z Plane for a wavelength of 632.8 nm and a beam waist of $w_0 = 0.1\,\text{mm}$

This contour plot illustrates how the Gaussian beam intensity distribution evolves as it propagates. The horizontal axis represents the normalized propagation distance (z/z), while the vertical axis shows the normalized transverse distance (x/w). The color gradient indicates intensity values, with brighter colors representing higher intensities.
The white dashed lines trace the beam width w(z), where the intensity falls to $1/e^2$ (approximately 13.5%) of its value on the beam axis. Note how the beam width reaches its minimum at the beam waist (z=0) and expands as the beam propagates away from the focus.
The plot clearly shows that the highest intensity occurs at the beam waist, with the intensity decreasing both as we move away from the center axis and as the beam propagates away from the focal point.

Key Gaussian Beam Parameters

The following table summarizes the important parameters that characterize a Gaussian beam:

Parameter	Expression	Description
Beam waist (w_0)		Minimum beam radius at focus ($z = 0$)
Beam width ($w(z)$)	$w(z) = w_0\sqrt{1 + \left(\frac{z}{z_0}\right)^2}$	Beam radius at position z
Rayleigh length (z_0)	$z_0 = \frac{\pi w_0^2}{\lambda}$	Distance over which beam area doubles
Radius of curvature ($R(z)$)	$R(z) = z\left[1 + \left(\frac{z_0}{z}\right)^2\right]$	Radius of wavefront curvature
Divergence angle (θ)	$\theta = \frac{\lambda}{\pi w_0}$	Far-field half-angle of beam spread
Gouy phase ($\phi(z)$)	$\phi(z) = \arctan\left(\frac{z}{z_0}\right)$	Additional phase beyond plane wave
Complex beam parameter ($q(z)$)	$q(z) = z + iz_0$	Combined parameter for beam properties

These parameters are interrelated, forming a complete description of how a Gaussian beam propagates. The Rayleigh length z_0 is particularly important as it defines the transition between the near field (where the beam is approximately collimated) and the far field (where the beam diverges linearly). At a distance of one Rayleigh length from the waist, the beam width increases by a factor of $\sqrt{2}$ and the intensity drops to half its maximum value.

Gaussian Beam Intensity Profiles

To better understand the spatial distribution of intensity in a Gaussian beam, it's helpful to visualize how the intensity varies along different directions. Here we explore two fundamental cross-sections: the axial intensity along the beam propagation path, and the transverse intensity profile at the beam waist.

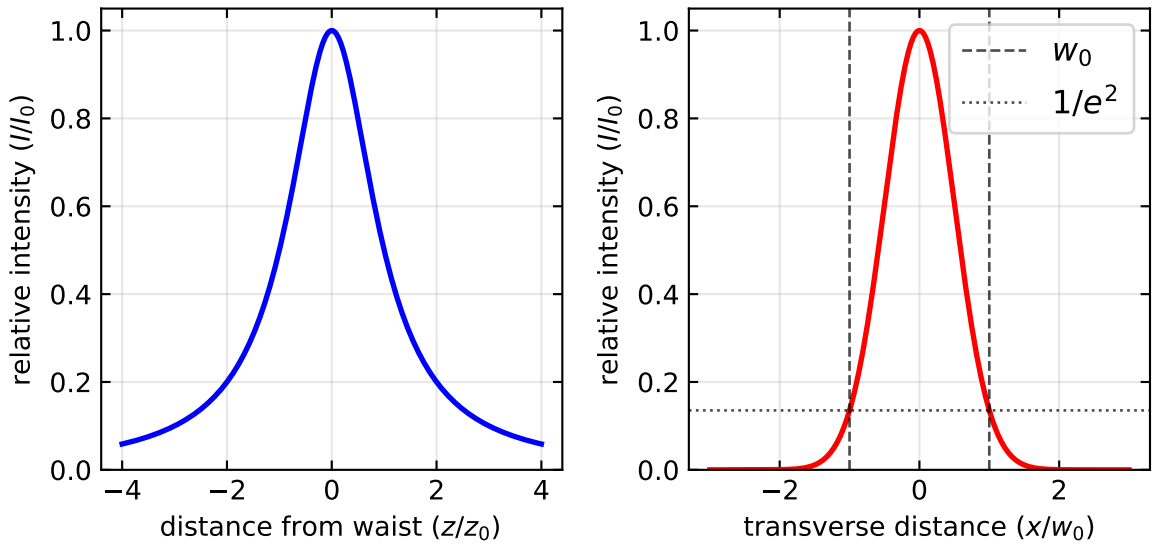


Figure 6.2: Gaussian Beam Intensity Profiles for a wavelenght of 632.8 nm and a beam waist of $w_0 = 0.1\,m$

Gaussian Beam Propagation

To better understand the spatial evolution of a Gaussian beam as it propagates, we can visualize how two key parameters change with distance: the beam width $w(z)$ and the wavefront radius of curvature $R(z)$.

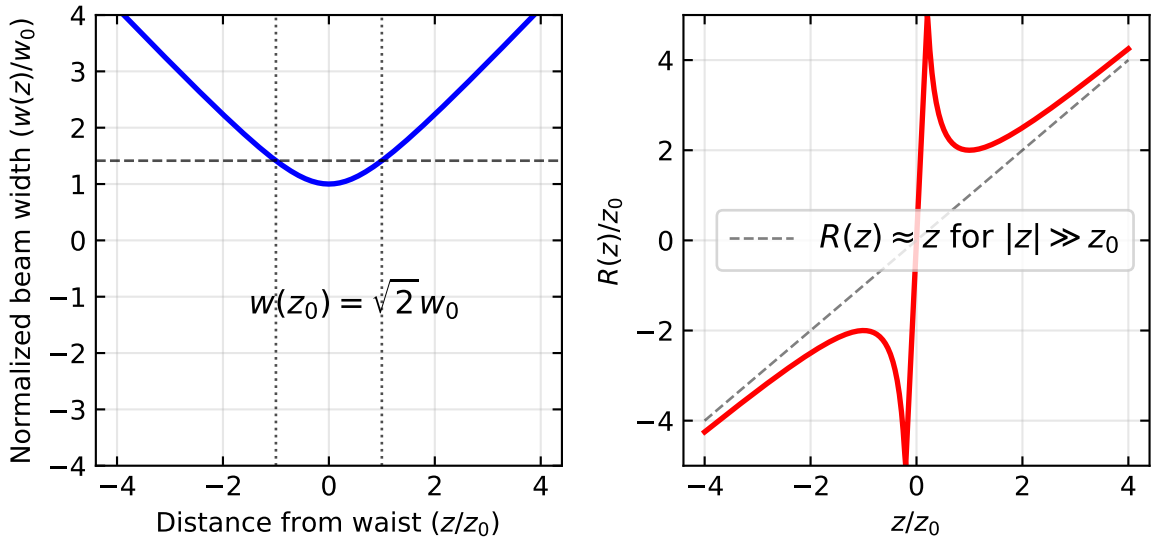


Figure 6.3: Gaussian Beam Width and Wavefront Curvature for a wavelength of 632.8 nm and a beam waist of $w_0 = 0.1\,\text{mm}$

The left plot shows how the beam width $w(z)$ evolves with distance from the beam waist. At $z = 0$, the beam is at its narrowest point w_0. At the Rayleigh range ($z = \pm z_0$), the width increases to $\sqrt{2}w_0$. For $|z| \gg z_0$, the beam width increases approximately linearly with distance, corresponding to a constant far-field divergence angle $\theta = \lambda/(\pi w_0)$.
The right plot illustrates the wavefront radius of curvature $R(z)$. At the beam waist, the wavefronts are flat ($R = \infty$). The curvature reaches its minimum absolute value of $2z_0$ at $z = \pm z_0$. For $z > 0$, $R(z)$ is positive (converging wavefronts), while for $z < 0$, $R(z)$ is negative (diverging wavefronts). As $|z|$ increases, $R(z)$ approaches the asymptotic behavior of a spherical wave, where $R(z) \approx z$.
These parameters together provide a complete description of how the Gaussian beam transforms from a tightly focused wave near the waist to an approximately spherical wave in the far field.

Gaussian Beam Transformation Through Optical Systems

The ABCD Matrix Formalism

The propagation of Gaussian beams through optical systems can be elegantly described using the ABCD matrix formalism from ray optics. While ray optics typically tracks the position and angle of rays, for Gaussian beams we track the transformation of the complex beam parameter $q(z)$.

When a Gaussian beam passes through an optical system characterized by an ABCD matrix, the complex beam parameter transforms according to:

$$q_2 = \frac{Aq_1 + B}{Cq_1 + D}$$

where q_1 is the initial complex beam parameter and q_2 is the transformed parameter. This remarkable result, known as the ABCD law for Gaussian beams, allows us to determine how the beam waist and wavefront curvature change through arbitrary optical systems.

Common Optical Elements

Different optical elements transform Gaussian beams in characteristic ways:

1. **Free-space propagation** over distance d is represented by:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix}$$

This matrix describes how the beam naturally diverges as it propagates.

2. **Thin lens** with focal length f:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix}$$

A lens modifies the wavefront curvature without changing the beam diameter at the lens location.

3. **Curved interface** between media with refractive indices n_1 and n_2 and radius of curvature R:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\frac{n_2-n_1}{n_2 R} & \frac{n_1}{n_2} \end{pmatrix}$$

Multiple optical elements can be analyzed by multiplying their respective ABCD matrices in the order encountered by the beam.

Focusing of Gaussian Beams

A particularly important case is the focusing of a Gaussian beam by a lens. If a collimated Gaussian beam with waist w_0 is incident on a lens with focal length f, the focused beam will have a new waist:

$$w_0' = \frac{\lambda f}{\pi w_0}$$

located approximately at the focal point. This equation highlights the fundamental diffraction limit of focusing: smaller focal spots require larger input beam diameters relative to the wavelength.

When a Gaussian beam is focused by a lens, the Rayleigh range of the focused beam also changes. For a collimated input beam, the new Rayleigh range after focusing is:

$$z_0' = \frac{\pi {w_0'}^2}{\lambda} = \frac{\lambda f^2}{\pi w_0^2}$$

This means that tightly focused beams (small w_0') have correspondingly shorter Rayleigh ranges, resulting in more rapid divergence beyond the focal point. This inverse relationship between spot size and Rayleigh range represents a fundamental trade-off in beam focusing: achieving a smaller spot size necessarily results in a beam that diverges more quickly after the focus.

The divergence angle of the focused beam is also affected, increasing as the spot size decreases:

$$\theta' = \frac{\lambda}{\pi w_0'} = \frac{w_0}{f}$$

This relationship shows that the focused beam's divergence is inversely proportional to the input beam width. A wider input beam produces a more tightly focused spot with greater divergence, while a narrower input beam creates a larger focal spot with less divergence.

The transformation matrices enable us to design optical systems that reshape Gaussian beams to desired specifications—expanding, collimating, or focusing them for specific applications. This matrix approach bridges ray optics and wave optics, providing a powerful tool for optical system design with coherent light sources.

Higher-Order Gaussian Modes

Hermite-Gaussian Beams

Hermite-Gaussian modes form a complete set of solutions to the paraxial wave equation in Cartesian coordinates. They can be expressed as:

$$U_{nm}(x,y,z) = U_0 \frac{w_0}{w(z)} H_n\left(\frac{\sqrt{2}x}{w(z)}\right) H_m\left(\frac{\sqrt{2}y}{w(z)}\right) \exp\left[-\frac{x^2+y^2}{w^2(z)}\right]$$
$$\times \exp\left[-ikz - ik\frac{x^2+y^2}{2R(z)} + i(n+m+1)\phi(z)\right]$$

where H_n and H_m are Hermite polynomials of orders n and m. The indices $n, m = 0, 1, 2, ...$ determine the number of nodes in the intensity pattern along x and y directions. The fundamental Gaussian beam corresponds to $n = m = 0$.

These modes naturally arise in laser resonators with rectangular symmetry and maintain their intensity pattern during propagation, though they scale in size. Each higher-order mode experiences an additional Gouy phase shift, causing different modes to accumulate phase at different rates during propagation.

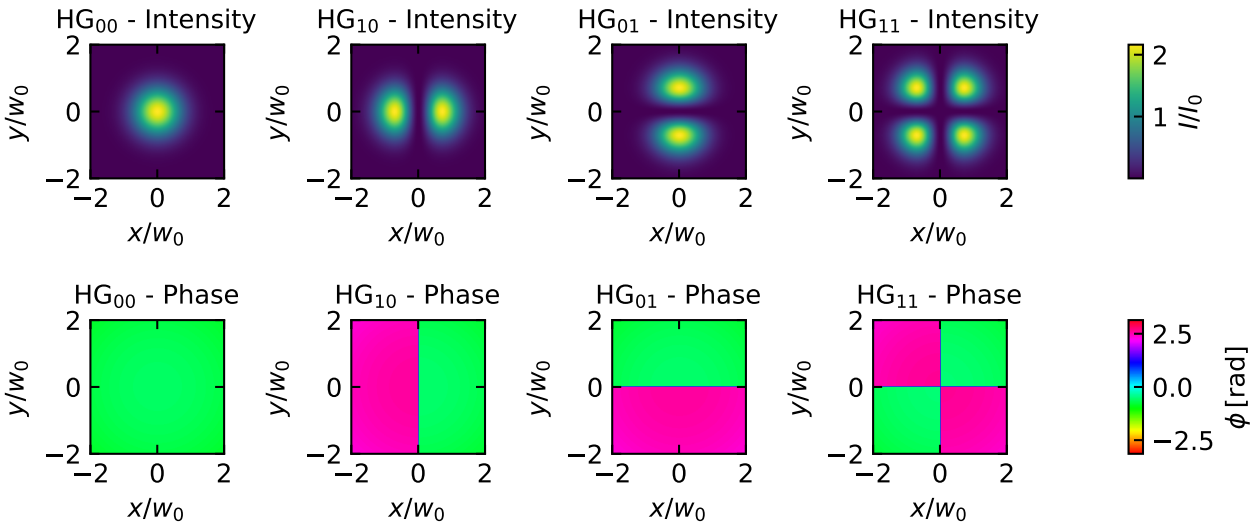


Figure 6.4: Intensity (top row) and phase (bottom row) distributions of the first four Hermite-Gaussian modes in the xy-plane at z=1z : (a) HG , (b) HG , (c) HG , and (d) HG . Higher-order modes clearly show multiple intensity peaks.

Laguerre-Gaussian Beams

In systems with cylindrical symmetry, Laguerre-Gaussian modes provide a more natural description. In cylindrical coordinates (r, θ, z), they are given by:

$$U_{pl}(r,\theta,z) = U_0 \frac{w_0}{w(z)} \left(\frac{\sqrt{2}r}{w(z)}\right)^{|l|} L_p^{|l|}\left(\frac{2r^2}{w^2(z)}\right) \exp\left[-\frac{r^2}{w^2(z)}\right]$$
$$\times \exp\left[-ikz - ik\frac{r^2}{2R(z)} + i(2p+|l|+1)\phi(z) + il\theta\right]$$

where $L_p^{|l|}$ are the associated Laguerre polynomials, $p \geq 0$ is the radial index determining the number of radial nodes, and l is the azimuthal index that determines the helical structure of the wavefront.

A remarkable property of Laguerre-Gaussian modes with $l \neq 0$ is that they carry orbital angular momentum (OAM) of $l\hbar$ per photon. This OAM arises from the helical phase structure represented by the term $\exp(il\theta)$, which creates a twisted wavefront resembling a spiral staircase. The intensity distribution forms a ring-like pattern with a dark center for $l \neq 0$ due to the phase singularity along the beam axis. As p increases, additional concentric rings appear in the intensity pattern.

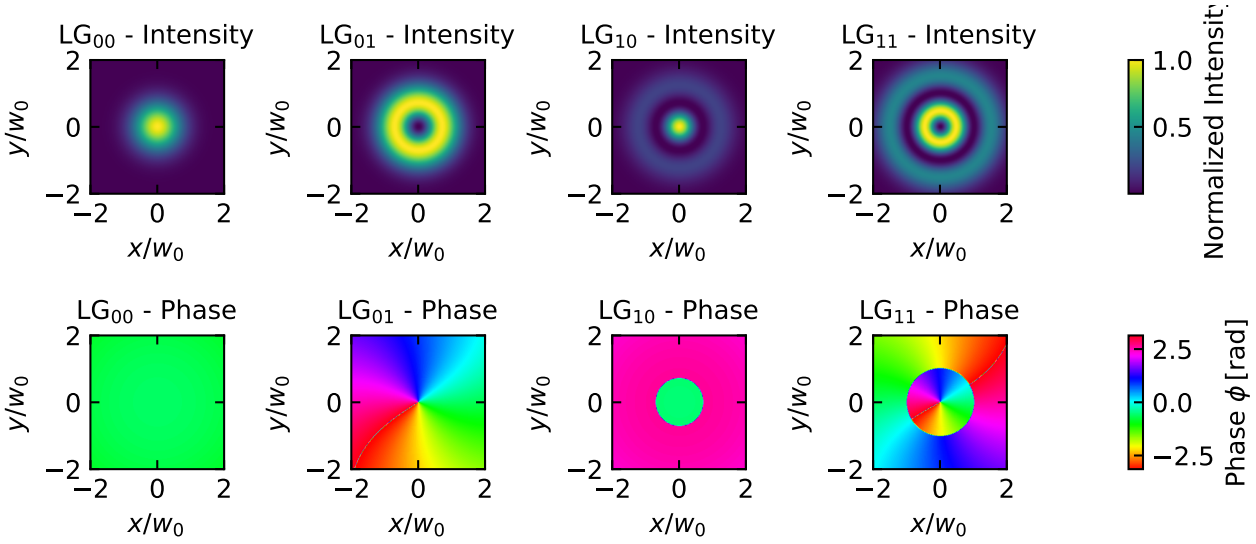


Figure 6.5: Intensity (top row) and phase (bottom row) distributions of the first four Laguerre-Gaussian modes in the xy-plane at z=0.1z : (a) LG , (b) LG , (c) LG , and (d) LG . Note the ring-like intensity patterns and spiral phase structures in modes with non-zero azimuthal index l.

The orbital angular momentum of light is distinct from spin angular momentum (SAM), which is associated with circular polarization ($\pm\hbar$ per photon). While SAM relates to the polarization state of light, OAM relates to the spatial structure of the wavefront. Importantly, these two forms of angular momentum can interact through spin-orbit coupling in certain optical systems, particularly in anisotropic or inhomogeneous media, at interfaces, or when light experiences strong focusing. Such spin-orbit coupling enables novel phenomena like spin-to-orbital angular momentum conversion, where the polarization state can influence the spatial structure of the beam and vice versa. This coupling mechanism has found specific applications in:

1. **Optical tweezers** - Spin-orbit coupling allows precise control of trapped particles by converting polarization changes into rotational motion, enabling manipulation of microscopic objects with unprecedented precision.
2. **Quantum cryptography** - The coupling between SAM and OAM creates additional degrees of freedom for encoding quantum information, enhancing the security and information capacity of quantum key distribution protocols.
3. **Optical vortex metrology** - Using the phase singularities created by spin-orbit interactions to detect nanoscale surface imperfections with superior sensitivity compared to conventional techniques.
4. **Chiral spectroscopy** - The interaction between polarization and spatial modes enables enhanced detection of chiral molecules by amplifying the difference in light-matter interactions between enantiomers.
5. **Structured light microscopy** - Coupling between SAM and OAM generates complex field patterns that improve resolution beyond the diffraction limit in specific imaging configurations.

Both families of higher-order modes are important in modern optics applications, including optical manipulation, quantum information processing, and mode-division multiplexing in optical communications. They represent different orthogonal bases of the same solution space and can be transformed into each other through appropriate optical systems.

Part V

Lecture 5

Chapter 7

Introduction to Fourier Optics

Fourier optics offers a robust analytical approach to understanding light propagation through optical systems by employing Fourier analysis techniques on optical fields. This framework elegantly connects image formation and optical resolution to the transmission of spatial information via light waves. Our exploration begins with examining complex transmittance functions, which give us fundamental insights into how various samples shape optical wavefronts. From this foundation, we will progress to the essential principles of Fourier optics and the associated diffraction integrals.

Transmission

When light interacts with an optical component or object, its amplitude and phase can be modified. Following Saleh and Teich's formalism, we can characterize this interaction using the complex transmission factor $t(x, y)$, which is defined as the ratio of the output field amplitude to the input field amplitude at each point (x, y) in a plane:

$$t(x, y) = \frac{U_{\text{out}}(x, y)}{U_{\text{in}}(x, y)}$$

This transmission factor is generally complex-valued, with its magnitude representing amplitude modulation and its phase representing phase modulation of the incident light.

For a thin lens, the primary effect is phase modulation. To derive the transmission function for a thin lens, we need to consider the optical path length through the lens at each point. Consider a planoconvex lens with one flat surface and one spherical surface of radius R. The thickness of the lens varies with position according to:

$$d(x, y) = d_0 - \frac{(x^2 + y^2)}{2R}$$

where d_0 is the thickness at the center. As light passes through the lens, it experiences a phase delay proportional to the optical path length, which is the product of the refractive index n and the physical path length:

$$\phi(x, y) = k \cdot n \cdot d(x, y) - k \cdot d(x, y)_{\text{air}}$$

where $k = 2\pi/\lambda$ is the wavenumber. Simplifying:

$$\phi(x, y) = k(n - 1)d(x, y) = k(n - 1)\left(d_0 - \frac{(x^2 + y^2)}{2R}\right)$$

The first term represents a constant phase shift that we can ignore, and the second term gives us the position-dependent phase modulation. For a lens with focal length f, the relationship between R and f is given by the lensmaker's formula, which for a planoconvex lens simplifies to:

$$(n-1)\frac{1}{R} = \frac{1}{f}$$

Substituting this into our phase equation:

$$\phi(x,y) = -k(n-1)\frac{(x^2+y^2)}{2R} = -k\frac{(x^2+y^2)}{2f}$$

The complex transmission factor is then:

$$t(x,y) = \exp[j\phi(x,y)] = \exp\left[-j\frac{k}{2f}(x^2+y^2)\right]$$

This quadratic phase factor represents the position-dependent phase delay introduced by the lens, with greater delays at the thicker portions of the lens.

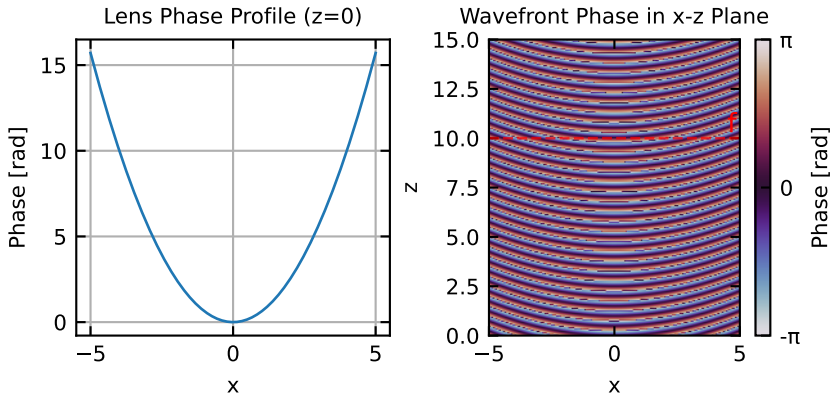


Figure 7.1: Phase modulation effect of a thin lens on an incident plane wave. (a) The quadratic phase profile introduced by the lens at z=0. (b) The wavefront shape in the x-z plane after passing through the lens, showing how the initially flat wavefront is transformed into a converging spherical wavefront.

This transmission function is crucial in Fourier optics as it allows us to mathematically model how a lens transforms an incident field. When placed in the path of a light wave, the lens modifies the wavefront according to this transmission factor, effectively performing a spatial Fourier transform of the input field at its focal plane.

Generalization to Arbitrary Thickness Objects

For arbitrary thickness objects, we can extend our treatment beyond the thin-element approximation. When light propagates through a medium of varying thickness and refractive index, the transmission function becomes:

$$t(x,y) = A(x,y)e^{i\phi(x,y)}$$

where $A(x,y)$ represents amplitude modulation (absorption or gain) and $\phi(x,y)$ represents phase modulation. For a thick object, the phase shift is given by the path integral through the object:

$$\phi(x,y) = k\int_{\text{path}} [n(x,y,z) - n_0]dz$$

where $n(x,y,z)$ is the spatially varying refractive index within the object, n_0 is the refractive index of the surrounding medium, and the integration is performed along the light path through the object.

This formulation accounts for complex three-dimensional objects where both the thickness and the refractive index may vary with position. For inhomogeneous media, we can express the transmission function as:

$$t(x,y) = \exp\left[-\frac{1}{2}\alpha(x,y) + ik\int_0^{d(x,y)} n(x,y,z)dz\right]$$

where $\alpha(x,y)$ is the absorption coefficient integrated along the path, and $d(x,y)$ is the thickness at position (x,y).

For many practical applications, this can be approximated by considering the effective phase and amplitude changes, leading to the more manageable form:

$$t(x,y) = \tau(x,y)e^{ik(n-n_0)d(x,y)}$$

where $\tau(x,y)$ is the amplitude transmission coefficient accounting for reflection and absorption losses.

This mathematical framework will become crucially important later when we describe image formation from waves that have propagated through an object. The transmission function directly encodes how an object modifies both the amplitude and phase of the incident light field, which determines how the object appears in an imaging system. Different imaging modalities (such as bright-field, phase-contrast, or differential interference contrast microscopy) essentially measure different aspects of this complex transmission function, revealing different properties of the object being imaged.

Wave Propagation Through Objects

When a plane wave propagating along the z-axis encounters an object, its wavefronts are modified according to the object's transmission function. This section explores how different types of objects transform incident wavefronts, which is fundamental to understanding phenomena from simple refraction to complex wavefront shaping.

The wavefront visualizations above illustrate how different optical elements transform an incident plane wave:

1. **Free Space Propagation**: In the absence of any optical element, a plane wave maintains flat wavefronts perpendicular to the propagation direction.
2. **Lens Effect**: A converging lens introduces a quadratic phase modulation, transforming plane wavefronts into converging spherical wavefronts that focus at the focal point.
3. **Prism Effect**: A prism applies a linear phase gradient across the wavefront, tilting the wavefronts and changing the propagation direction according to Snell's law.
4. **Arbitrary Phase Objects**: More complex phase profiles create correspondingly complex wavefront shapes, which can be designed for specific applications like wavefront correction or beam shaping.

Understanding these wavefront transformations is essential in optical system design, as the shape of the wavefront directly determines how light propagates through subsequent optical elements and ultimately forms images or interference patterns.

Spatial Frequencies and Angular Spectrum

Building on our analysis of wave propagation through various optical elements, we now explore a fundamental concept in Fourier optics that connects spatial patterns to wave propagation directions. This relationship between spatial structure and angular distribution is a direct extension of how different optical elements transform wavefronts, as visualized in the previous section. Just as a lens converts a plane wave into a converging spherical wave and a prism tilts the wavefront to change the propagation direction, complex spatial patterns decompose into multiple propagation directions—a relationship that will become essential when we discuss optical imaging systems, diffraction limits, and the resolution capabilities of microscopes and telescopes.

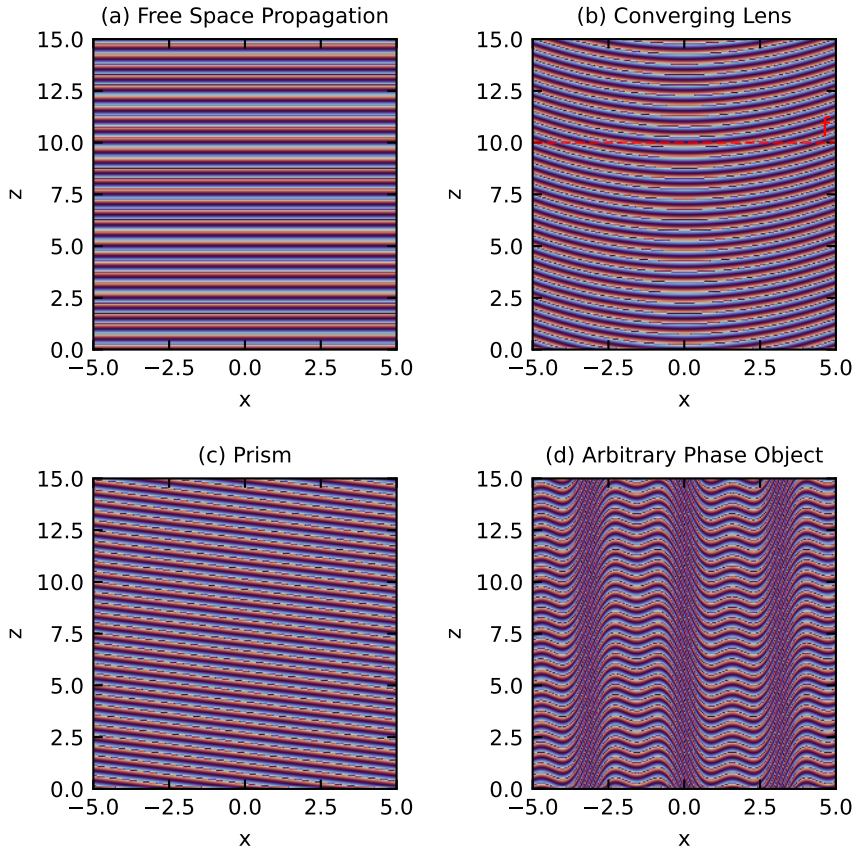


Figure 7.2: Wavefront propagation after transmission through different optical elements. (a) A plane wave passing through free space maintains flat wavefronts. (b) After passing through a converging lens, the wavefronts become spherical, converging toward the focal point. (c) Transmission through a prism tilts the wavefronts, changing the propagation direction. (d) A phase plate with arbitrary phase profile creates custom-shaped wavefronts.

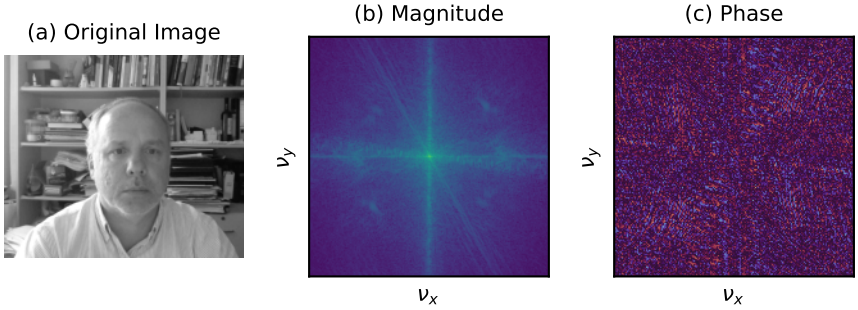


Figure 7.3: Spatial frequency analysis of an image. (a) Original grayscale image. (b) Magnitude of the 2D Fourier transform, showing the distribution of spatial frequencies. (c) Phase of the Fourier transform.

The Concept of Spatial Frequencies

Just as a temporal signal can be decomposed into frequency components through Fourier analysis, a spatial pattern or object can be represented as a superposition of spatial harmonic functions with different spatial frequencies. The spatial frequency ν represents how rapidly the intensity or phase of an optical field changes with distance.

For a two-dimensional complex spatial harmonic function:

$$f(x, y) = Ae^{i2\pi(\nu_x x + \nu_y y)}$$

where:

- ν_x and ν_y are the spatial frequencies in the x and y directions (in cycles per unit length)
- A is the complex amplitude

Higher spatial frequencies correspond to finer details in an object, while lower spatial frequencies represent coarser features.

Overall, the function $f(x, y)$ can be expressed as the Fourier transform of its spatial frequency components:

$$f(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(\nu_x, \nu_y) e^{i2\pi(\nu_x x + \nu_y y)} d\nu_x d\nu_y$$

where $F(\nu_x, \nu_y)$ is the spatial frequency spectrum of $f(x, y)$.

Fourier Transform Review

Basic Definitions

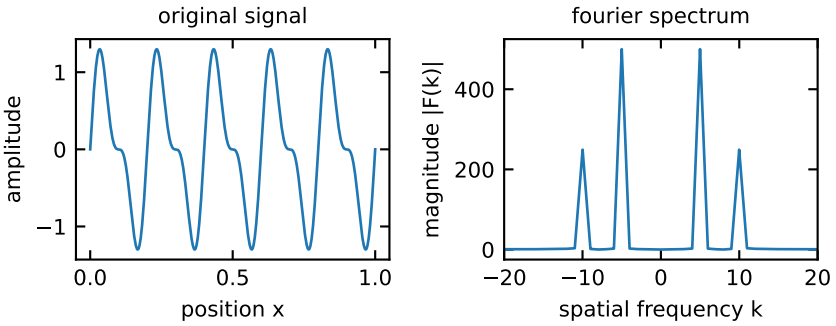
The Fourier transform decomposes a function into its constituent frequencies. For a function $f(x)$, its Fourier transform $F(k)$ is defined as:

$$F(k) = \int_{-\infty}^{\infty} f(x) e^{-ikx} dx$$

The inverse Fourier transform reconstructs the original function:

$$f(x) = \int_{-\infty}^{\infty} F(k) e^{ikx} dk$$

In optics, x typically represents spatial coordinates and k represents spatial frequencies. When working with discrete data, as you will in your computational exercises, you'll use the Discrete Fourier Transform (DFT), which is efficiently computed using the Fast Fourier Transform (FFT) algorithm:



Important Properties

1. **Linearity**: $\mathcal{F}\{af(x) + bg(x)\} = aF(k) + bG(k)$

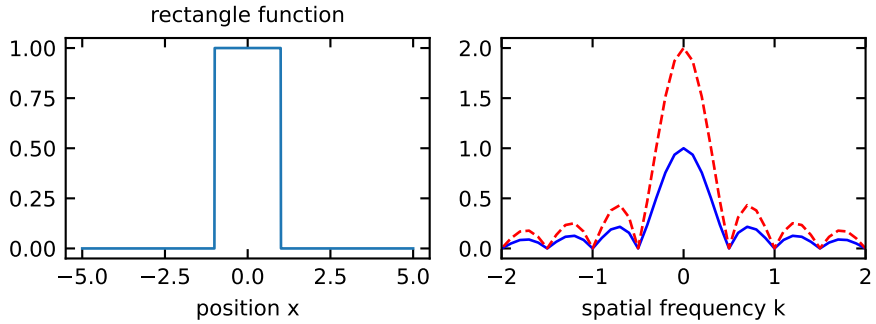
This means the Fourier transform of a sum is the sum of the Fourier transforms, allowing us to analyze complex signals by breaking them into simpler components.

2. **Shift Theorem**: $\mathcal{F}\{f(x-a)\} = e^{-ika}F(k)$
   A shift in the spatial domain corresponds to a phase change in the frequency domain, critical for understanding how optical elements that cause phase shifts affect the spectrum.
3. **Convolution Theorem**: $\mathcal{F}\{f * g\} = F(k) \cdot G(k)$
   Convolution in the spatial domain becomes multiplication in the frequency domain. This is particularly useful in optics, where the effect of a lens or aperture can be modeled as a convolution operation.
4. **Parseval's Theorem**: $\int |f(x)|^2 dx = \int |F(k)|^2 dk$
   This theorem establishes energy conservation between domains, showing that the total energy in a signal is preserved in its Fourier transform.

Common Fourier Transform Pairs

Function	Fourier Transform
$\delta(x)$ (Delta function)	1 (constant)
1 (constant)	$\delta(k)$ (Delta function)
$\text{rect}(x)$ (Rectangle function)	$\text{sinc}(k)$ (Sinc function)
$e^{-\pi x^2}$ (Gaussian)	$e^{-\pi k^2}$ (Gaussian)
$\cos(2\pi a x)$	$\frac{1}{2}[\delta(k-a) + \delta(k+a)]$

Understanding these transform pairs is essential for optical analysis. For example, a rectangular aperture produces a sinc-function diffraction pattern, and a Gaussian beam maintains its Gaussian profile under propagation.



Correspondence to Plane Wave Angular Components

One of the most profound insights in Fourier optics is the relationship between spatial frequencies and the angular spectrum of plane waves. To understand this relationship, consider a plane wave $U(x, y, z)$ with wavevector $\mathbf{k}$ and wavelength λ incident on the plane $z = 0$. The wavevector can be written as:

$$\mathbf{k} = k_x\hat{\mathbf{x}} + k_y\hat{\mathbf{y}} + k_z\hat{\mathbf{z}}$$

where $|\mathbf{k}| = 2\pi/\lambda$. The components of this wavevector can be expressed in terms of the propagation angles θ_x and θ_y (with respect to the z-axis):

$$k_x = \frac{2\pi}{\lambda}\sin\theta_x$$

$$k_y = \frac{2\pi}{\lambda}\sin\theta_y$$

$$k_z = \frac{2\pi}{\lambda}\cos\theta_z$$

where $\cos\theta_z = \sqrt{1 - \sin^2\theta_x - \sin^2\theta_y}$ from the constraint that $|\mathbf{k}| = 2\pi/\lambda$.

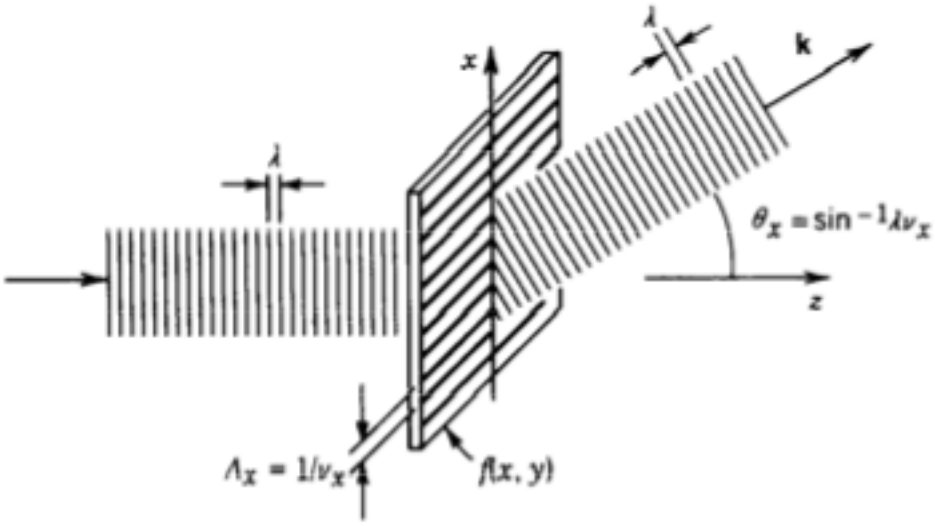


Figure 7.4: Principle of plane wave angular decomposition. (Image taken from Saleh/Teich "Principles of Photonics")

At the plane $z = 0$, this plane wave can be represented as:

$$U(x, y, 0) = U_0 e^{j(k_x x + k_y y)}$$

where $k_x/2\pi$ and $k_x/2\pi$ are the spatial frequencies of the plane wave along the x- and y direction. This equation shows that a plane wave propagating at angles θ_x and θ_y manifests as a spatial harmonic function at the $z = 0$ plane, with spatial frequencies directly related to the propagation angles:

$$\frac{k_x}{2\pi} = \frac{1}{\lambda}\sin\theta_x$$
$$\frac{k_y}{2\pi} = \frac{1}{\lambda}\sin\theta_y$$

We can match now the spatial frequencies of the object $f(x, y)$ to the plane wave $U(x, y, 0)$ by adjusting the wavevector angles θ_x, θ_y to yield the same periodicity.

The means that

$$U(x, y, 0) = f(x, y)$$

or concerning the frequencies

$$\nu_x = \frac{k_x}{2\pi} = \frac{1}{\lambda}\sin\theta_x$$
$$\nu_y = \frac{k_y}{2\pi} = \frac{1}{\lambda}\sin\theta_y$$

This means each spatial frequency of the sample $f(x, y)$ is diffracting the incident plane wave $U(x, y, z)$ into a certain angle, when the frequencies are matched. Behind the sample, the plane wave is propagating further without any change with the additional phase factor $e^{-ik_z z}$ such that

$$U(x, y, z) = U(x, y, 0)e^{-ik_z z} = U_0 e^{i(k_x x + k_y y)} e^{-ik_z z}$$

where the wavevector component k_z is given by:

$$k_z = \sqrt{k^2 - k_x^2 - k_y^2} = \frac{2\pi}{\lambda}\sqrt{1 - \lambda^2(\nu_x^2 + \nu_y^2)}$$

This expression for k_z shows how the propagation along the z-direction depends on the spatial frequencies in the x and y directions. This relationship provides a direct connection between the spatial structure of an object and the directions in which light propagates after interacting with it.

Spatial Frequency and Propagation Angles of a Grating

We saw this principle in action when analyzing diffraction gratings, where we decomposed the grating's periodic structure into angular components using the grating vector. For a grating with period d, the spatial frequency is $\nu_x = 1/d$, and the directions of diffracted orders are given by:

$$\sin\theta_m = m\lambda/d = m\lambda\nu_x$$

where m is the diffraction order. This shows how the grating's spatial frequency determines the angles of diffracted light, which is a specific application of the more general Fourier relationship between spatial frequencies and propagation angles.

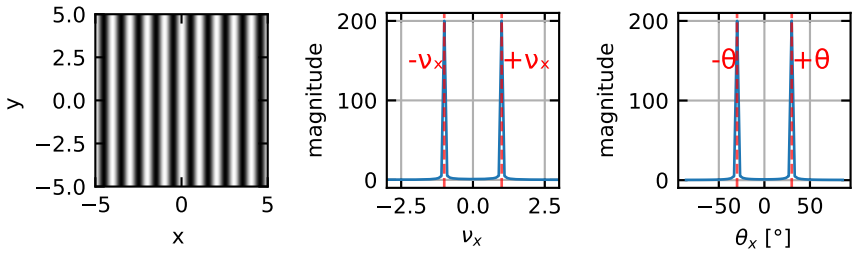


Figure 7.5: Visualization of a 2D object containing a single spatial frequency in the x-direction. (a) The object pattern showing sinusoidal variation along x with frequency . (b) The Fourier transform magnitude of the object, showing two symmetric points corresponding to ± . (c) The corresponding angular spectrum representation, where the spatial frequency maps to specific diffraction angles ± according to sin( ) = .

Part VI

Lecture 6

Chapter 8

Spatial Spectral Analysis

Corresponding to our previous analysis, the angular spectrum representation can be formalized using Fourier analysis. The complex amplitude transmittance $f(x, y)$ can be written as a Fourier transform

$$f(x,y) = \iint_{-\infty}^{\infty} F(\nu_x, \nu_y) e^{i2\pi(\nu_x x + \nu_y y)} d\nu_x d\nu_y$$

where $F(\nu_x, \nu_y)$ give the amplitudes of the frequency components of the transmittance. With our previous expression, then the field at any plane z can be obtained by:

$$U(x,y,z) = \iint_{-\infty}^{\infty} F(\nu_x, \nu_y) e^{i2\pi(\nu_x x + \nu_y y)} e^{-ik_z z} d\nu_x d\nu_y$$

where $k_z = 2\pi\sqrt{(1/\lambda)^2 - \nu_x^2 - \nu_y^2}$ is the z-component of the wavevector.

This formulation shows that the field at any distance z can be calculated by multiplying each spatial frequency component by the appropriate phase factor $e^{-ik_z z}$ and then performing an inverse Fourier transform. This approach provides an elegant and computationally efficient method for modeling wave propagation, particularly in homogeneous media.

For propagating waves, where $\nu_x^2 + \nu_y^2 < (1/\lambda)^2$, the factor $e^{-ik_z z}$ represents a phase shift. For evanescent waves, where $\nu_x^2 + \nu_y^2 > (1/\lambda)^2$, k_z becomes imaginary, resulting in exponential decay with distance.

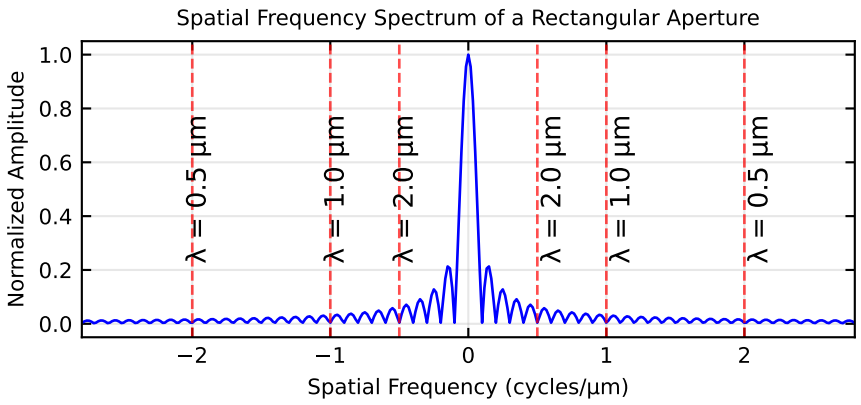


Figure 8.1: Spatial frequency spectrum of a rectangular aperture with width 10 m. The dashed lines indicate the cutoff frequencies at which light with wavelength can propagate.

Transfer Function of Free space

We now examine the propagation of a monochromatic optical wave of wavelength λ and complex amplitude $U(x, y, z)$ in the free space between the planes $z = 0$ and $z = d$, called the input and output planes, respectively. Given the complex amplitude of the wave at the input plane, $f(x, y) = U(x, y, 0)$, we want to determine the complex amplitude at the output plane, $g(x, y) = U(x, y, d)$.

The input field $f(x, y)$ propagates through free space to form the output field $g(x, y)$. Using the angular spectrum representation, we can express the relationship between input and output as:

$$g(x,y) = \iint_{-\infty}^{\infty} F(\nu_x, \nu_y) e^{i2\pi(\nu_x x + \nu_y y)} e^{-ik_z d} d\nu_x d\nu_y$$

where $F(\nu_x, \nu_y)$ is the Fourier transform of $f(x, y)$, and $k_z = 2\pi\sqrt{(1/\lambda)^2 - \nu_x^2 - \nu_y^2}$ is the z-component of the wavevector.

The transfer function of free space, denoted as $H(\nu_x, \nu_y)$, is defined as the ratio of the output spectrum to the input spectrum:

$$H(\nu_x, \nu_y) = e^{-ik_z d} = e^{-i2\pi d\sqrt{(1/\lambda)^2 - \nu_x^2 - \nu_y^2}}$$

This transfer function has two distinct regimes based on the values of ν_x and ν_y:

1. **Propagating waves**: When $\nu_x^2 + \nu_y^2 < (1/\lambda)^2$, k_z is real, and $H(\nu_x, \nu_y) = e^{-ik_z d}$ represents a pure phase shift. These are propagating waves that carry energy from the input to the output plane.
2. **Evanescent waves**: When $\nu_x^2 + \nu_y^2 > (1/\lambda)^2$, k_z becomes imaginary, and $H(\nu_x, \nu_y) = e^{-|k_z|d}$ represents an exponential decay. These evanescent waves decay exponentially with distance and do not propagate energy to the far field. For spatial frequencies slightly beyond the propagating limit, where $\nu_x^2 + \nu_y^2 \approx (1/\lambda)^2 + \Delta$, the decay constant can be approximated as $|k_z| \approx 2\pi\sqrt{\Delta} \approx \pi\lambda/(2d^2)$, where d is the characteristic distance from the object. This means that features with spatial frequencies significantly above $1/\lambda$ become effectively undetectable at distances greater than a few wavelengths.

A simplification of the transfer function of free space maybe obtained when considering only spatial frequencies that are much smaller than the cut-off frequency. This simplificaltion is called the Fresnel approximation and leads to

$$H\left(\nu_x, \nu_y\right) \approx H_0 \exp\left[i\pi\lambda d\left(\nu_x^2 + \nu_y^2\right)\right]$$

Its inverse Fourier transform is the impulse response function $h(x, y)$, which is given by

$$h(x,y) \approx h_0 \exp\left[-ik\frac{x^2 + y^2}{2d}\right]$$

with $h_0 = (i/\lambda d)\exp(-ikd)$.

Fresnel Approximation

The expression for the transfer function may be simplified if the input function $f(x, y)$ contains only spatial frequencies that are much smaller than the cutoff frequency $1/\lambda$, so that $\nu_x^2 + \nu_y^2 \ll 1/\lambda^2$. The plane-wave components of the propagating light then make small angles $\theta_x \approx \lambda\nu_x$ and $\theta_y \approx \lambda\nu_y$ corresponding to paraxial rays.
Denoting $\theta^2 = \theta_x^2 + \theta_y^2 \approx \lambda^2\left(\nu_x^2 + \nu_y^2\right)$, where θ is the angle with the optical axis, the phase factor in the transfer function is

$$2\pi\left(\frac{1}{\lambda^2}-\nu_x^2-\nu_y^2\right)^{1/2} d = 2\pi\frac{d}{\lambda}\left(1-\theta^2\right)^{1/2}$$
$$= 2\pi\frac{d}{\lambda}\left(1-\frac{\theta^2}{2}+\frac{\theta^4}{8}-\cdots\right)$$

Neglecting the third and higher terms of this expansion, the transfer function may be approximated by

$$H\left(\nu_x,\nu_y\right) \approx H_0 \exp\left[i\pi\lambda d\left(\nu_x^2+\nu_y^2\right)\right]$$

where $H_0 = \exp(-ikd)$. In this approximation, the phase is a quadratic function of ν_x and ν_y. This approximation is known as the Fresnel approximation.
The condition of validity of the Fresnel approximation is that the third term in the expansion is much smaller than π for all θ. This is equivalent to

$$\frac{\theta^4 d}{4\lambda} \ll 1$$

If a is the largest radial distance in the output plane, the largest angle $\theta_m \approx a/d$, and this condition may be written in the form

$$\frac{N_{\mathrm{F}}\theta_m^2}{4} \ll 1,$$

where $N_{\mathrm{F}} = a^2/\lambda d$ is the Fresnel number. For example, if $a = 1$ cm, $d = 100$ cm, and $\lambda = 0.5\mu$ m, then $\theta_m = 10^{-2}$ radian, $N_{\mathrm{F}} = 200$, and $N_{\mathrm{F}}\theta^2/4 = 5\times10^{-3}$. In this case the Fresnel approximation is applicable.

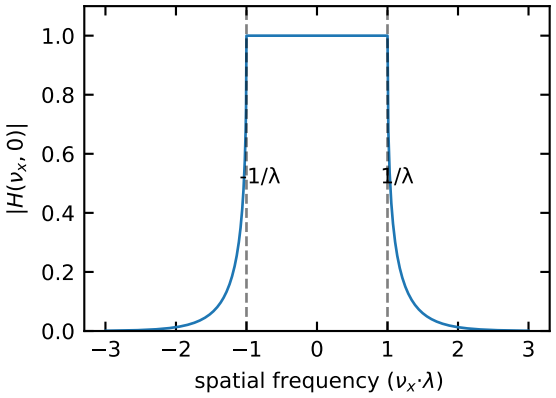


Figure 8.2: The magnitude of the free-space transfer function. For propagating waves ($\nu_x^2+\nu_y^2 < 1/\lambda^2$), the transfer function has magnitude 1, representing pure phase delay. For evanescent waves ($\nu_x^2+\nu_y^2 > 1/\lambda^2$), the magnitude decays exponentially with distance from the origin.

i Total Internal Reflection: A Spatial Frequency Interpretation

In Fourier optics, we analyze wave propagation in terms of spatial frequencies. For a monochromatic wave with vacuum wavelength λ_0:

1. **Wavevector components** in a medium with refractive index n:

$$|\mathbf{k}| = \frac{2\pi n}{\lambda_0}$$

And its components must satisfy:

$$k_x^2 + k_y^2 + k_z^2 = \left(\frac{2\pi n}{\lambda_0}\right)^2$$

2. **Spatial frequency components** are related to the wavevector:

$$\nu_x = \frac{k_x}{2\pi}, \nu_y = \frac{k_y}{2\pi}, \nu_z = \frac{k_z}{2\pi}$$

Thus, the dispersion relation in terms of spatial frequencies:

$$\nu_x^2 + \nu_y^2 + \nu_z^2 = \left(\frac{n}{\lambda_0}\right)^2$$

Total Internal Reflection Condition

When light travels from medium 1 (n_1) to medium 2 (n_2), where $n_1 > n_2$:

1. **Tangential continuity**: The tangential component of the wavevector (k_x) must be conserved across the boundary. For an incident angle θ_i:

$$k_x = \frac{2\pi n_1}{\lambda_0} \sin\theta_i$$

Or in terms of spatial frequency:

$$\nu_x = \frac{n_1}{\lambda_0} \sin\theta_i$$

2. **Critical angle condition**: At the critical angle θ_c:

$$\sin\theta_c = \frac{n_2}{n_1}$$

In spatial frequency terms, this corresponds to:

$$\nu_x^{critical} = \frac{n_2}{\lambda_0}$$

3. **Total internal reflection regime**: When $\theta_i > \theta_c$, or equivalently, when $\nu_x > n_2/\lambda_0$:

$$\nu_x > \frac{n_2}{\lambda_0}$$

The normal component of the wavevector in medium 2 becomes:

$$k_z^2 = \left(\frac{2\pi n_2}{\lambda_0}\right)^2 - k_x^2 < 0$$

Or in spatial frequency terms:

$$\nu_z^2 = \left(\frac{n_2}{\lambda_0}\right)^2 - \nu_x^2 < 0$$

4. **Evanescent wave solution**: Since $\nu_z^2 < 0$, we have $\nu_z = \pm i\gamma$ where γ is real. Choosing the physically meaningful solution:

$$\nu_z = i\gamma = i\sqrt{\nu_x^2 - \left(\frac{n_2}{\lambda_0}\right)^2}$$

The field in medium 2 takes the form:

$$E(x,z) = E_0 e^{i2\pi\nu_x x} e^{-2\pi\gamma z}$$

This represents an evanescent wave that propagates along the interface but decays exponentially perpendicular to it.

5. **Penetration depth**: The amplitude of the evanescent wave decreases by a factor of 1/e at a distance:

$$d = \frac{1}{2\pi\gamma} = \frac{1}{2\pi\sqrt{\nu_x^2 - \left(\frac{n_2}{\lambda_0}\right)^2}}$$

This can be rewritten in terms of the incident angle:

$$d = \frac{\lambda_0}{2\pi\sqrt{n_1^2 \sin^2\theta_i - n_2^2}}$$

Spatial Frequency Filtering Interpretation

Total internal reflection can be understood as a spatial frequency filtering phenomenon:

1. The medium with refractive index n_2 has a maximum spatial frequency cutoff at $\nu_{max} = n_2/\lambda_0$ for propagating waves.
2. When the incident wave has a tangential spatial frequency component $\nu_x > \nu_{max}$, the second medium cannot support propagating waves at this spatial frequency.
3. The resulting evanescent wave can be viewed as a "frustrated" attempt to propagate high spatial frequencies that exceed the medium's capability.
4. Only spatial frequencies that satisfy $\nu_x \leq n_2/\lambda_0$ can propagate in medium 2, constituting a low-pass spatial filter.

This mathematical framework demonstrates why optical systems with lower numerical apertures (effectively lower refractive indices) cannot resolve features with spatial frequencies beyond their cutoff frequency - the same principle that underlies the diffraction limit in microscopy.

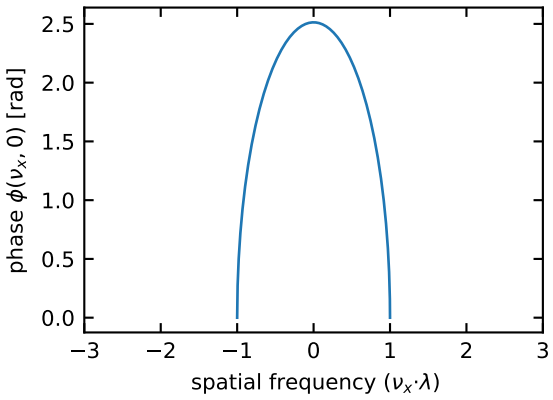


Figure 8.3: The phase of the free-space transfer function. For propagating waves (_x² + _y² < 1/ ²), the transfer function introduces a phase delay that increases with spatial frequency. This phase represents the wavefront curvature during propagation.

The figure below now visualizes the effect of free space propagation on the phase of the transfer function when light of certain wavelength is used to illuminate and rectangular aperture of 10 µm width.

This description of free space propagation provides insight into important phenomena such as:

Diffraction limits: Spatial frequencies beyond $1/\lambda$ correspond to evanescent waves that decay exponentially with distance, explaining why sub-wavelength features cannot be observed in the far field. This would mean that light should not propagate through subwavelength holes. This is what you would expect for the grid in front of your microwave. Yet, light can penetrate through subwavelength holes not only as evanescent fields. Bethe used in 1944 an idealized model where the film was infinitely thin and the metal was a perfect conductor. Under these assumptions, he derived a straightforward expression for the transmission efficiency η_B (normalized

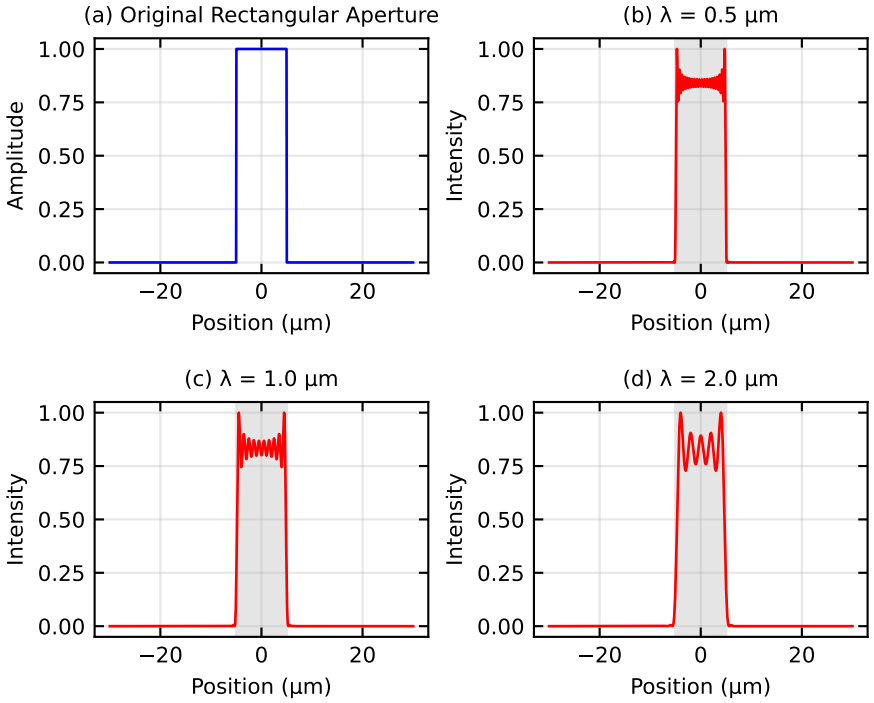


Figure 8.4: Effect of spatial frequency cutoff on image reconstruction. (a) Original rectangular aperture. (b-d) Reconstructed intensity after applying different wavelength cutoffs. As the cutoff wavelength increases, more high-frequency components are lost, resulting in blurring and loss of edge sharpness.

to the aperture area):

$$\eta_B = \frac{64(kr)^4}{27\pi^2}$$

where $k = 2\pi/\lambda$ represents the wavevector magnitude of the incoming light with wavelength λ, and r is the hole radius. This equation clearly shows that η_B scales as $(r/\lambda)^4$, indicating that the optical transmission would decrease rapidly as λ becomes larger than r.

However, real apertures with finite depth exhibit waveguide properties. Light transmission through these waveguides differs fundamentally from free-space propagation. The confined geometry modifies the field's dispersion relation, with the aperture's lateral dimensions determining the cutoff wavelength λ_c beyond which propagation ceases. When incident wavelength $\lambda > \lambda_c$, transmission decays exponentially, indicating the non-propagating regime (Fig. 2). In real metals, the transition from propagative to evanescent regimes occurs gradually rather than at a sharply defined λ_c.

The transmission of light through subwavelength apertures provides exciting new tools for spectroscopy and imaging applications. These tools enable the manipulation of light at the nanoscale, leading to advancements in areas such as microscopy, sensing, and communication.

Resolution limits: An optical system with a maximum acceptance angle θ_{max} can only capture spatial frequencies up to $\sin\theta_{max}/\lambda$, limiting the finest details that can be resolved.

Spatial filtering: Optical components like apertures and lenses act as spatial filters, selectively transmitting or modifying certain spatial frequency components.

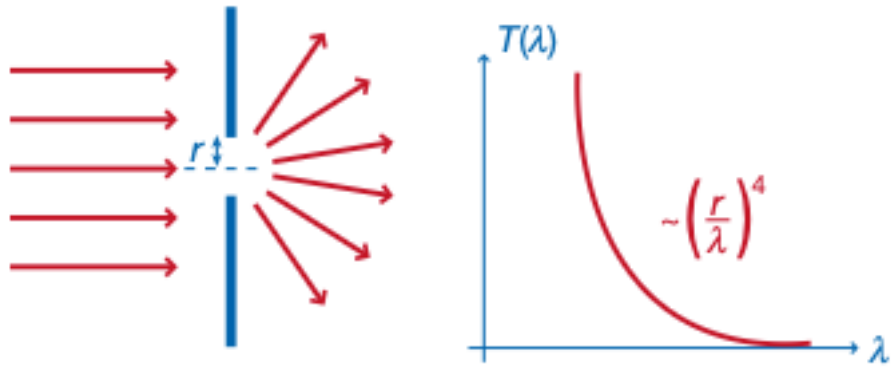


Figure 8.5: Diffraction and typical transmission spectrum of visible light through a subwavelength hole in an infinitely thin perfect metal film.

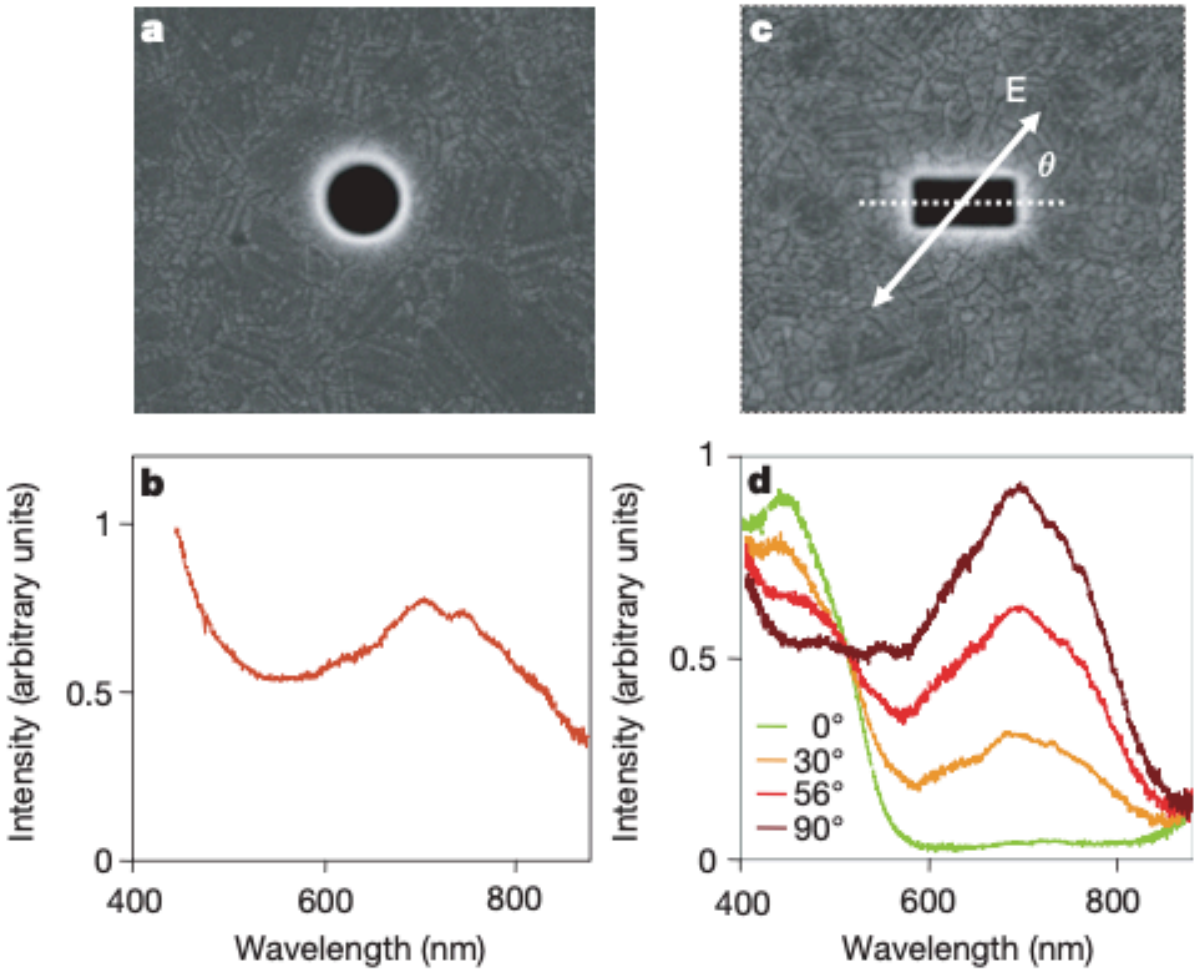


Figure 8.6: Optical transmission properties of single holes in metal films. The holes were milled in suspended optically thick Ag films illuminated with white light. a, A circular aperture and b, its transmission spectrum for a 270 nm diameter in a 200-nm-thick film. c, A rectangular aperture and d, its transmission spectrum as a function of the polarization angle h for the following geometrical parameters: 210 nm 3 310 nm, film thickness 700 nm. (Degiron, A., Lezec, H. J., Yamamoto, N. & Ebbesen, T. W. Optical transmission properties of a single subwavelength aperture in a real metal. Opt. Commun. 239, 61–66 (2004).)

8.1 Amplitude Modulation

Let's examine how spatial amplitude modulation affects the angular propagation of light. Consider a transparency with complex amplitude transmittance $f_0(x, y)$. If its Fourier transform $F_0(\nu_x, \nu_y)$ extends over spatial frequency ranges $\pm\Delta\nu_x$ and $\pm\Delta\nu_y$ in the x and y directions, the transparency will deflect an incident plane wave by angles θ_x and θ_y within the ranges:

$$\pm \sin^{-1}(\lambda\Delta\nu_x)$$

and

$$\pm \sin^{-1}(\lambda\Delta\nu_y)$$

respectively.

Now consider a second transparency with complex amplitude transmittance:

$$f(x, y) = f_0(x, y)e^{-i2\pi(\nu_{x0}x+\nu_{y0}y)}$$

where $f_0(x, y)$ varies slowly compared to the exponential carrier term, meaning $\Delta\nu_x \ll \nu_{x0}$ and $\Delta\nu_y \ll \nu_{y0}$. This represents an amplitude-modulated function with spatial carrier frequencies ν_{x0} and ν_{y0} and modulation function $f_0(x, y)$. According to the shift property of the Fourier transform, the transform of $f(x, y)$ is:

$$F_0(\nu_x - \nu_{x0}, \nu_y - \nu_{y0})$$

The transparency will deflect a plane wave in directions centered around the angles:

$$\theta_{x0} = \sin^{-1}(\lambda\nu_{x0})$$

and

$$\theta_{y0} = \sin^{-1}(\lambda\nu_{y0})$$

This behavior can be understood by viewing $f(x, y)$ as a combination of the base transmittance $f_0(x, y)$ with a phase grating having transmittance $e^{-i2\pi(\nu_{x0}x+\nu_{y0}y)}$ that provides the angular deflection.

This principle enables spatial-frequency multiplexing, where two images $f_1(x, y)$ and $f_2(x, y)$ can be recorded on the same transparency using the encoding:

$$f(x, y) = f_1(x, y)e^{-i2\pi(\nu_{x1}x+\nu_{y1}y)} + f_2(x, y)e^{-i2\pi(\nu_{x2}x+\nu_{y2}y)}$$

By illuminating this combined transparency with a plane wave, the two images are deflected at different angles determined by their carrier frequencies, allowing them to be spatially separated. This technique is particularly valuable in holography, where separating different image components recorded on the same medium is often necessary.

Structured Illumination Microscopy (SIM)

The amplitude modulation concepts presented here form the theoretical foundation for Structured Illumination Microscopy (SIM), a super-resolution imaging technique. In SIM, a sample is illuminated with a known spatially structured pattern, typically a sinusoidal grid. This can be mathematically represented as an illumination intensity pattern:

$$I_{\text{illum}}(x, y) = I_0[1 + m\cos(2\pi\nu_0 x + \phi)]$$

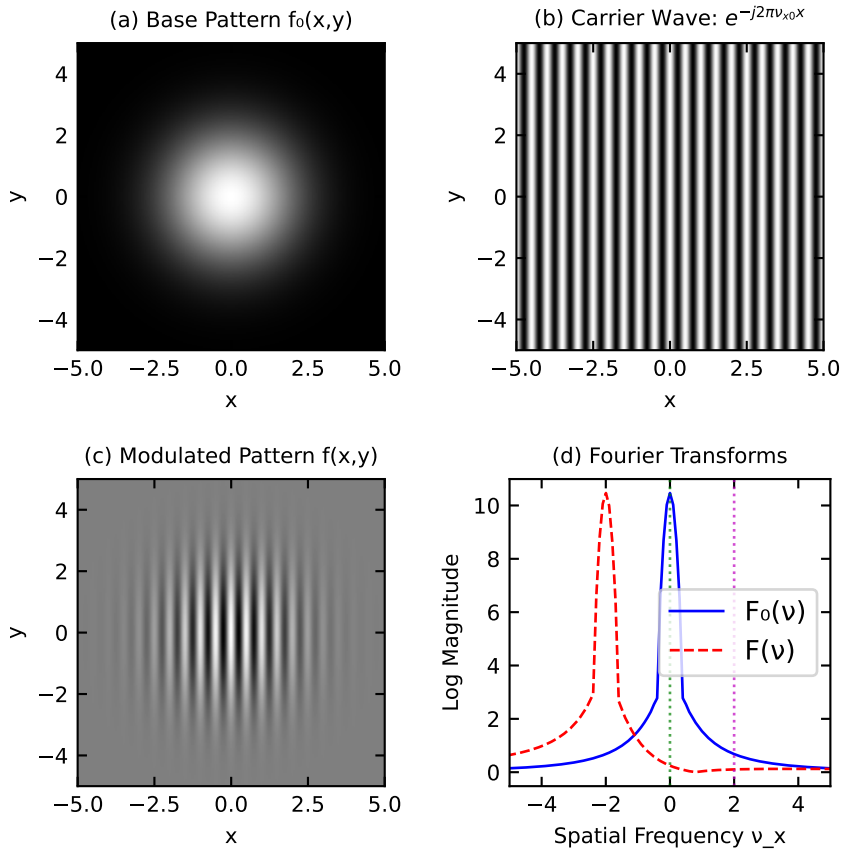


Figure 8.7: Visualization of amplitude modulation and the corresponding angular deflection. (a) The base pattern f (x,y) - a Gaussian envelope. (b) A carrier wave with spatial frequency . (c) The amplitude-modulated pattern f(x,y)=f (x,y)e^(-j2 x). (d) Fourier transforms showing how the spectrum shifts with modulation, corresponding to angular deflection.

where I_0 is the average intensity, m is the modulation depth, ν_0 is the spatial frequency of the illumination pattern, and ϕ is the phase.
When this structured pattern illuminates a sample with spatial structure $S(x, y)$, the resulting observed image is simply the product:

$$D(x, y) = S(x, y) \cdot I_{\text{illum}}(x, y)$$

Substituting the illumination pattern:

$$D(x, y) = S(x, y) \cdot I_0[1 + m\cos(2\pi\nu_0 x + \phi)]$$
$$D(x, y) = I_0 \cdot S(x, y) + I_0 \cdot m \cdot S(x, y)\cos(2\pi\nu_0 x + \phi)$$

Using Euler's formula, we can rewrite the cosine term:

$$D(x, y) = I_0 \cdot S(x, y) + \frac{I_0 \cdot m}{2} \cdot S(x, y)[e^{j(2\pi\nu_0 x+\phi)} + e^{-j(2\pi\nu_0 x+\phi)}]$$

In the frequency domain, this becomes:

$$\tilde{D}(\nu_x, \nu_y) = I_0\tilde{S}(\nu_x, \nu_y) + \frac{I_0 \cdot m}{2}[\tilde{S}(\nu_x - \nu_0, \nu_y)e^{j\phi} + \tilde{S}(\nu_x + \nu_0, \nu_y)e^{-j\phi}]$$

where $\tilde{S}$ is the Fourier transform of the sample structure, and $\tilde{D}$ is the Fourier transform of the detected image.
This equation reveals how structured illumination enables access to high spatial frequencies beyond the conventional diffraction limit. In standard microscopy, the optical system acts as a low-pass filter due to the diffraction limit, restricting detectable spatial frequencies to $|\nu| \leq \nu_{\max} = \frac{NA}{\lambda}$, where NA is the numerical aperture and is the wavelength.
The key insight is that structured illumination creates a "moiré effect" between the illumination pattern and the sample structure. Consider a sample with high spatial frequency components that exceed $\nu_{\max}$ and would normally be undetectable. When this sample is illuminated with the structured pattern of frequency ν_0, these high-frequency components interact with the illumination pattern to produce difference frequencies that fall within the detectable range.

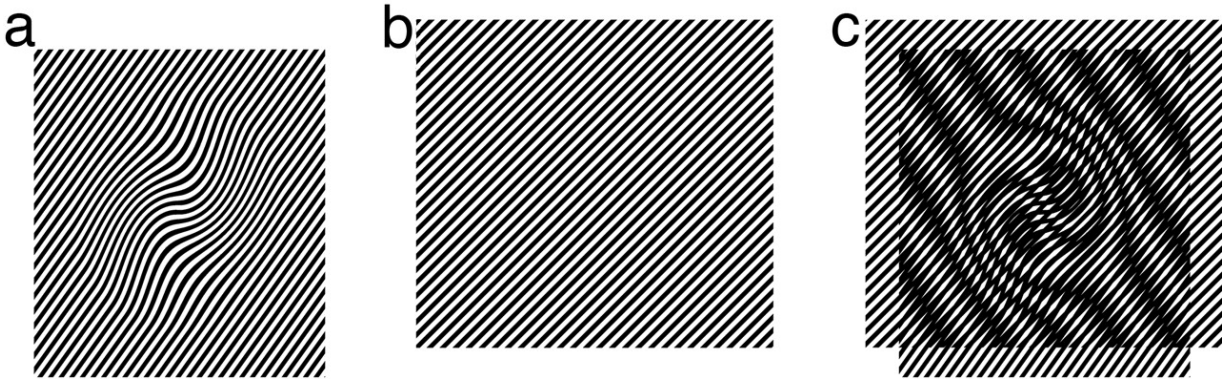


Figure 8.8: The moiré effect in SIM. When a sample with high spatial frequency features is illuminated with a structured pattern, the interference creates moiré fringes at lower frequencies that can be detected by the microscope. This allows information about sub-diffraction structures to be encoded in observable signals. (see Gustafsson, M. G. L. Nonlinear structured-illumination microscopy: Wide-field fluorescence imaging with theoretically unlimited resolution. Proc. Natl. Acad. Sci. 102, 13081–13086 (2005))

Specifically, sample features with spatial frequency $\nu_s > \nu_{\max}$ combine with the illumination frequency ν_0 to produce components at $\nu_s - \nu_0$ and $\nu_s + \nu_0$. If $\nu_s - \nu_0 < \nu_{\max}$, then this difference frequency becomes detectable by the optical system, effectively bringing previously inaccessible high-frequency information into the observable range.

For example, if a sample contains structures with spatial frequency $\nu_s = 1.7\nu_{\max}$ and we apply illumination with $\nu_0 = 0.8\nu_{\max}$, the difference frequency becomes $\nu_s - \nu_0 = 0.9\nu_{\max}$, which falls within the detectable range. This allows us to extract information about sample features that would be invisible under conventional illumination.
To separate and reconstruct these frequency-shifted components, we need multiple images with different phases of the illumination pattern. Typically, three images with phases $\phi = 0°, 120°, 240°$ are acquired, allowing us to solve the following system of equations:

$$\begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix} = \begin{pmatrix} 1 & e^{j\phi_1} & e^{-j\phi_1} \\ 1 & e^{j\phi_2} & e^{-j\phi_2} \\ 1 & e^{j\phi_3} & e^{-j\phi_3} \end{pmatrix} \begin{pmatrix} I_0\tilde{S}(\nu) \\ \frac{I_0 \cdot m}{2}\tilde{S}(\nu - \nu_0) \\ \frac{I_0 \cdot m}{2}\tilde{S}(\nu + \nu_0) \end{pmatrix}$$

By extracting these frequency-shifted components and computationally restoring them to their original positions in frequency space, we can reconstruct spatial frequencies beyond the conventional diffraction limit, typically achieving a resolution improvement factor of 2 in each dimension, or a factor of 2 beyond what would be possible with the same wavelength and numerical aperture in conventional microscopy.

8.2 Frequency Modulation

We now examine the transmission of a plane wave through a transparency made of a "collage" of several regions, the transmittance of each of which is a harmonic function of some spatial frequency, as illustrated below. If the dimensions of each region are much greater than the period, each region acts as a grating or a prism that deflects the wave in some direction, so that different portions of the incident wavefront are deflected into different directions. This principle may be used to create maps of optical interconnections, which may be used in optical computing applications

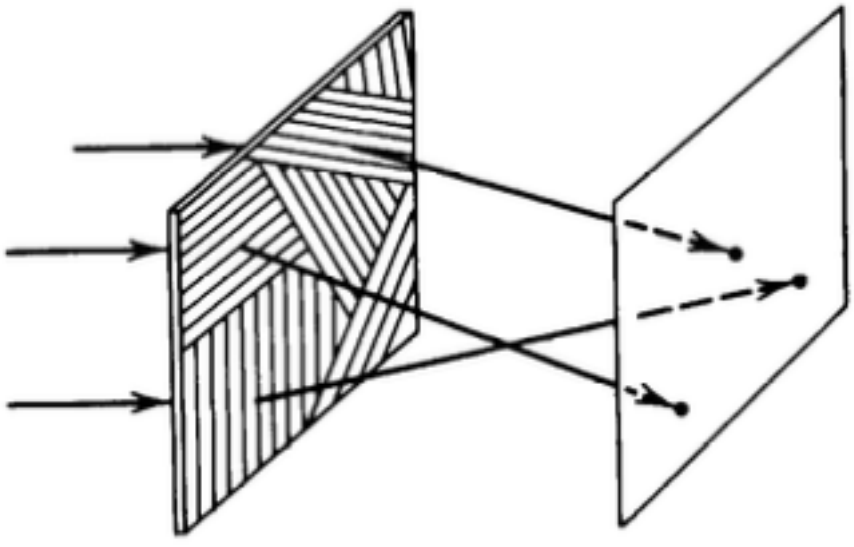


Figure 8.9: Deflection of light by a transparency made of several harmonic functions (phase gratings) of different spatial frequencies. (source Saleh/Teich Principles of Photonics).

This concept of directed light deflection through controlled spatial frequency modulation stands in stark contrast to what happens when spatial frequencies are distributed randomly, as we'll see next with speckle patterns.

8.3 Speckle: Random Frequency Modulation

When coherent light interacts with a rough surface or passes through a scattering medium, the resulting intensity pattern exhibits a characteristic granular appearance known as speckle. This phenomenon represents a natural example of random spatial frequency modulation that can be elegantly described using the Fourier optics framework we've developed.

While the optical interconnect example above demonstrates how carefully designed spatial frequency distributions can create predictable, useful light paths, speckle represents the opposite case - where randomly distributed spatial frequencies create complex interference patterns. In essence, speckle is what happens when nature creates its own "random optical interconnect map."

Speckle Formation

Speckle arises when coherent wavefronts undergo spatially varying phase shifts that cause complex interference. From our spatial frequency perspective, we can model a rough surface as applying a random phase modulation to the incident wavefront:

$$E_{\text{scattered}}(x, y) = E_0 e^{i\phi(x,y)}$$

where $\phi(x, y)$ is a random phase function corresponding to the surface height variations. When this scattered field propagates and interferes with itself, it produces the characteristic speckle pattern. Unlike our controlled interconnect example, where each region deliberately directs light in a specific direction, here each microscopic region of the rough surface randomly deflects light, creating a complex superposition of wavefronts with random phases and directions.

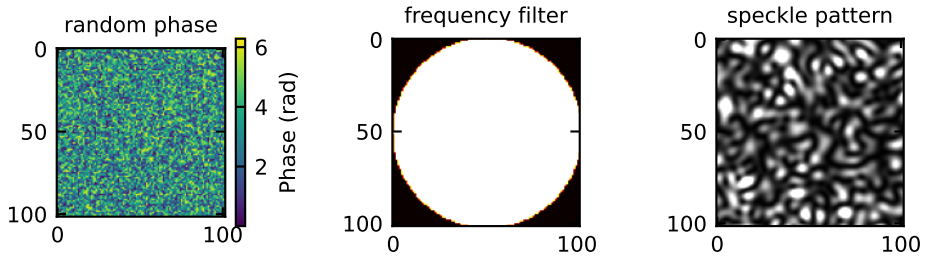


Figure 8.10: Speckle formation process. (a) Coherent light incident on a rough surface acquires random phase shifts. (b) The resulting scattered field creates a speckle pattern in the observation plane. (c) The characteristic granular appearance of a fully developed speckle pattern.

The random yet deterministic nature of speckle patterns shares remarkable similarities with the weight matrices of trained neural networks, offering an intriguing conceptual framework that connects Fourier optics with modern computational systems.

> **i** Conceptual Connections to Neuronal Networks
>
> 1. **Wave Superposition vs. Neuronal Contributions**: Just as each speckle grain represents the constructive interference of many wave components, each connection in a neural network can be viewed as the "superposition" of many training examples that collectively shaped that weight.
> 2. **Information Encoding**:
>   - Speckle patterns encode information about the scattering medium in a distributed, holographic manner
>   - Neural networks encode learned features in a distributed pattern across weight matrices
> 3. **Statistical Properties**:
>   - Speckle intensity follows a negative exponential distribution for fully developed speckle
>   - Neural network weights often approximate Gaussian distributions after training

4. **Fourier Domain Representation**:
    - Speckle can be analyzed in the Fourier domain to reveal the spatial frequency content of the scattering medium
    - Neural network weights can be analyzed in the frequency domain to reveal the spectrum of features they've learned to detect

The mathematical connection becomes even more apparent when considering both systems as complex-valued functions:

- A speckle field at a plane can be expressed as: $E(x,y) = \sum_k A_k e^{j\phi_k} e^{j(k_x x + k_y y)}$
- A neural network layer output can be expressed as: $y_i = \sigma\left(\sum_j w_{ij} x_j + b_i\right)$

Both involve a weighted sum of inputs, and both transform information from one domain to another through operations that can be represented as spatial filtering operations.

These connections between speckle field and information processing give rise to research field of imaging and computing with disordered materials.

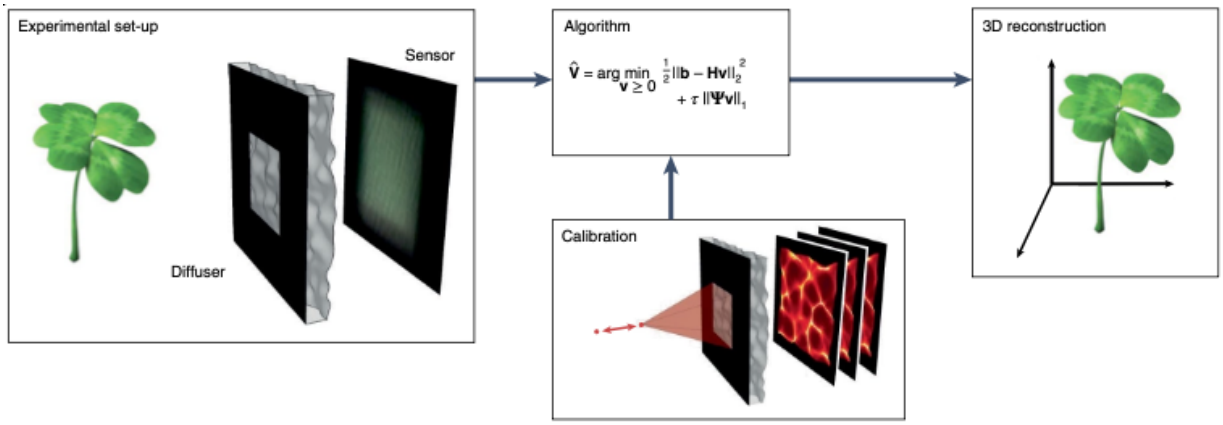


Figure 8.11: DiffuserCam: A compressive framework allows 3D imaging using a surface diffuser and a camera, using prior calibration of the diffuser and a minimization algorithm. Image taken from the perspective article Gigan, S. Imaging and computing with disorder. Nat. Phys. 18, 980–985 (2022).

8.4 Frequency Modulation with Continuously Varying Spatial Frequencies

A transparency may also have a harmonic transmittance with a spatial frequency that varies gradually across its surface, similar to how a musical note changes pitch in an FM radio signal. This is easier to visualize than the fixed-frequency patterns we discussed earlier.

Let's look at a transparency with this phase-altering property:

$$f(x,y) = \exp[-i2\pi\phi(x,y)]$$

where $\phi(x,y)$ is a smooth function that changes slowly compared to the wavelength of light (λ).

If we zoom in around any point (x_0, y_0), we can approximate $\phi(x,y)$ using the first few terms of a Taylor series (which you've seen in calculus):

$$\phi(x,y) \approx \phi(x_0,y_0) + (x-x_0)\nu_x + (y-y_0)\nu_y$$

where ν_x and ν_y are just the partial derivatives of ϕ with respect to x and y at that point:

$$\nu_x = \frac{\partial \phi}{\partial x} \quad \text{and} \quad \nu_y = \frac{\partial \phi}{\partial y}$$

Near this point, our transparency function behaves like:

$$\exp[-i2\pi(\nu_x x + \nu_y y)]$$

which we recognize as a harmonic function with local spatial frequencies ν_x and ν_y.

Since these derivatives change as we move across the transparency, the spatial frequencies also vary with position. As a result, different parts of the incoming light wave get deflected by different angles:

$$\theta_x = \sin^{-1}\left(\lambda \frac{\partial \phi}{\partial x}\right) \quad \text{and} \quad \theta_y = \sin^{-1}\left(\lambda \frac{\partial \phi}{\partial y}\right)$$

This is how optical elements like lenses can bend light in position-dependent ways to focus or shape wavefronts.

Imaging

When a thin transparency has a complex amplitude transmittance of

$$f(x, y) = \exp\left(j\pi x^2/\lambda f\right)$$

,

it introduces a phase shift of $2\pi\phi(x, y)$ where $\phi(x, y) = -x^2/2\lambda f$.

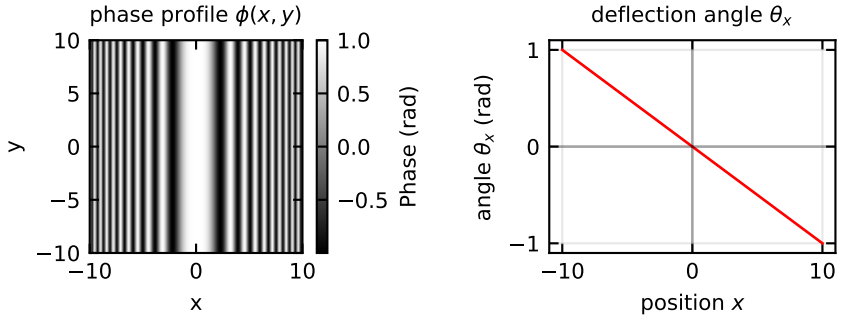


Figure 8.12: Phase function of a cylindrical Fourier lens. (left) The 2D phase profile showing how phase varies quadratically with x. (right) The resulting deflection angle as a function of position, showing the linear relationship that causes focusing.

This phase profile causes the wave at position (x, y) to be deflected by angles $\theta_x = \sin^{-1}(\lambda \partial\phi/\partial x) = \sin^{-1}(-x/f)$ and $\theta_y = 0$. For small values where $|x/f| \ll 1$, the deflection angle simplifies to $\theta_x \approx -x/f$, creating a linear relationship between deflection angle and position. When a plane wave illuminates this transparency, each point of the wavefront experiences a position-dependent deflection, transforming the overall wavefront shape. At each position x, the local wavevector is redirected by an angle $-x/f$, causing all light rays to converge at a focal point located at distance f from the transparency along the optical axis, as illustrated below.

This transparency operates exactly like a cylindrical lens with focal length f. Extending this concept, a transparency with transmittance $f(x, y) = \exp\left[j\pi\left(x^2 + y^2\right)/\lambda f\right]$ acts as a spherical lens with focal length f. This elegant mathematical expression perfectly captures the phase transformation property of a thin lens.

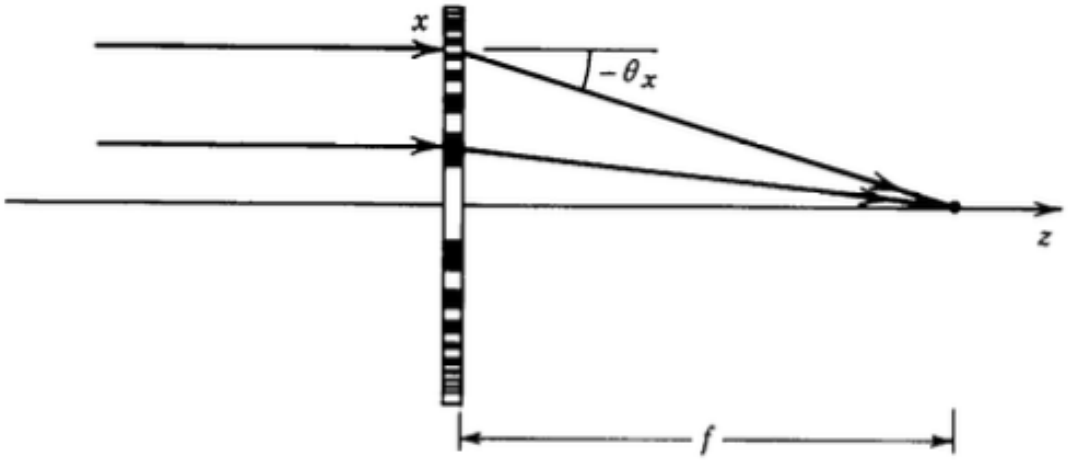


Figure 8.13: A transparency with transmittance $f(x, y) = \exp(j\pi x^2/\lambda f)$ bends the wave at position x by an angle $\theta_x \approx -x/f$, functioning as a cylindrical lens with focal length f.