Introduction to Photonics

Frank Cichos

2025-04-10

Table of contents

Chapter 1

Photonics

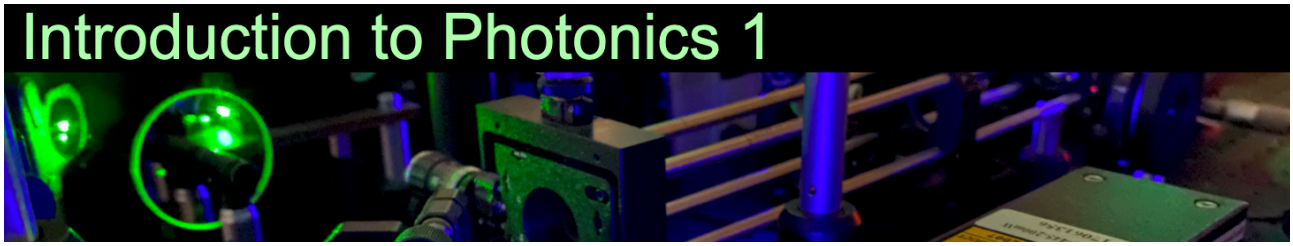


Figure 1.1: Photonics Logo

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

Theories for Light

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.

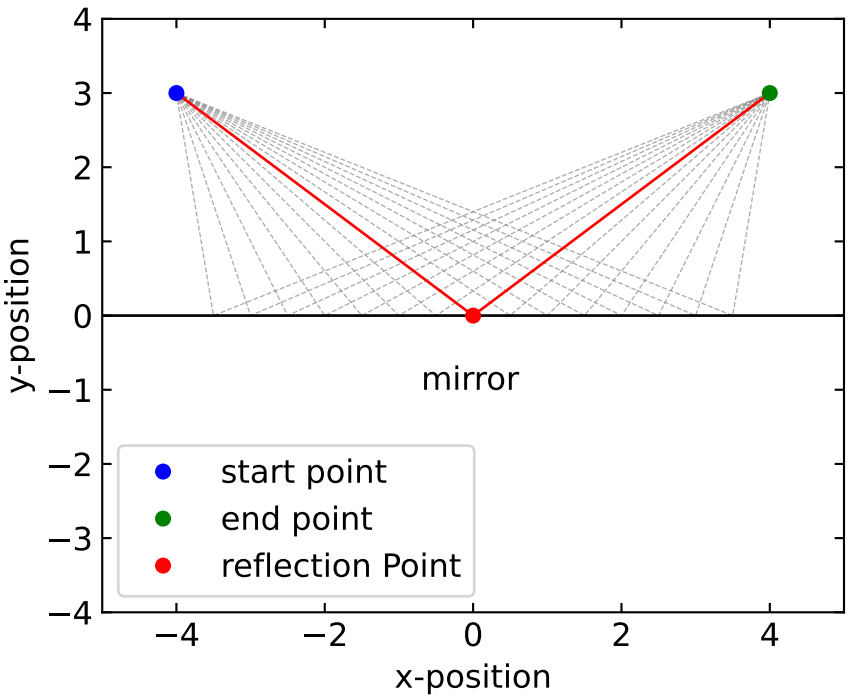


Figure 2.1: Fermat's principle for reflection of light at an interface

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.

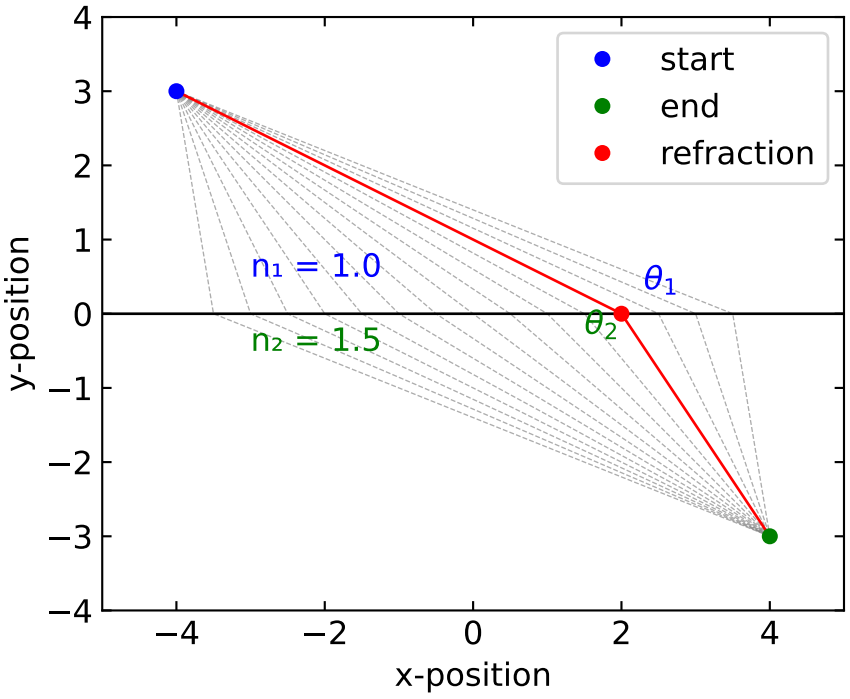


Figure 2.2: Snell's Law from Fermat's Principle

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:

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

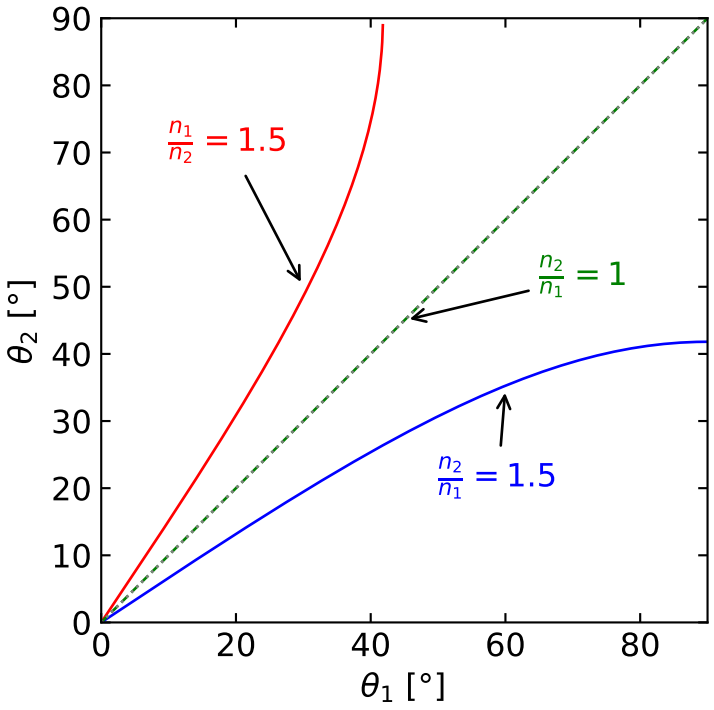


Figure 2.3: Snell's law for different combinations of refractive indices. The plots show the relationship between incident angle (θ_1) and refracted angle (θ_2) for three scenarios: (a) light passing from air to glass, (b) light passing from glass to air, and (c) a comparison of both cases. Note how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

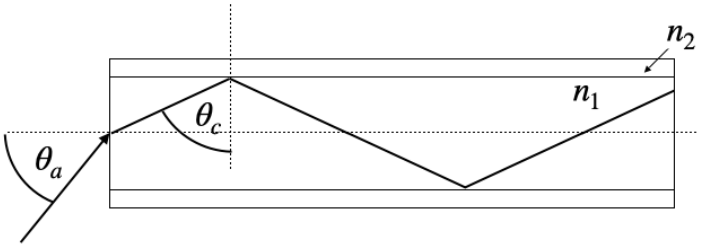


Figure 2.4: Total internal reflection in an optical fiber.

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:

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

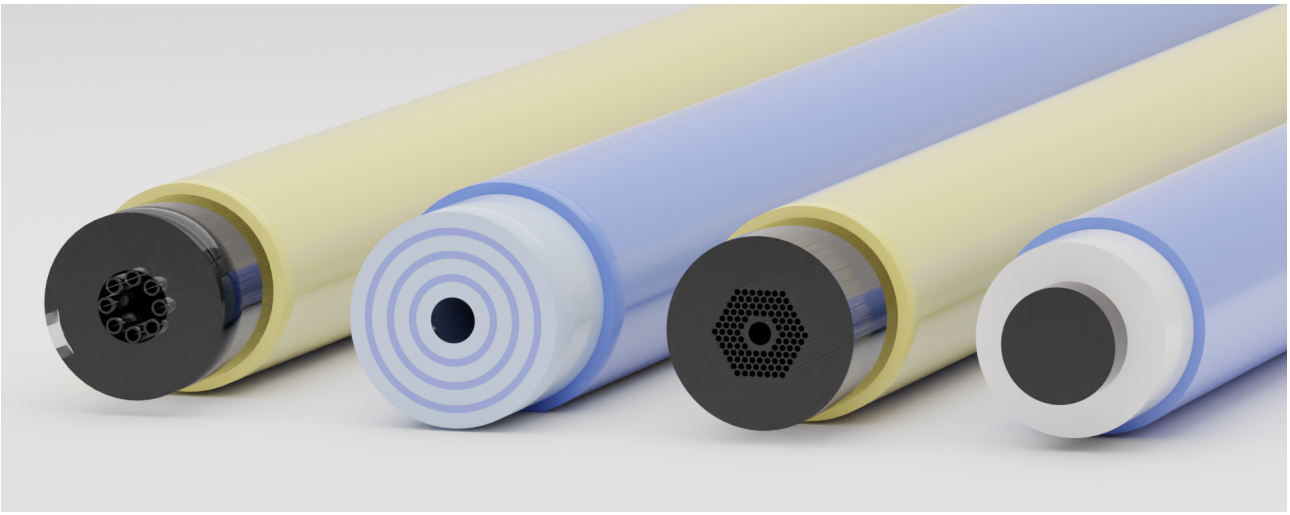


Figure 2.5: Rendering of different optical fibers types (from left to right): Hollow core optical fiber, hollow core bragg fiber, photonic crystal fiber, conventional fiber

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.

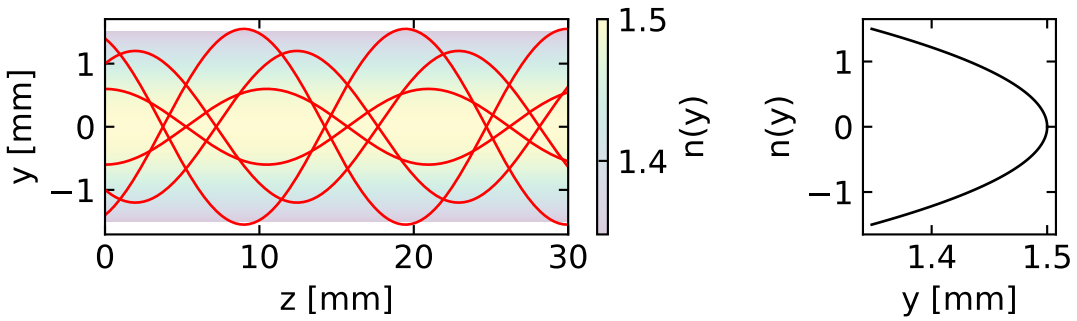


Figure 2.6: Ray-path inside a SELFOC gradient index optical fiber.

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:

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

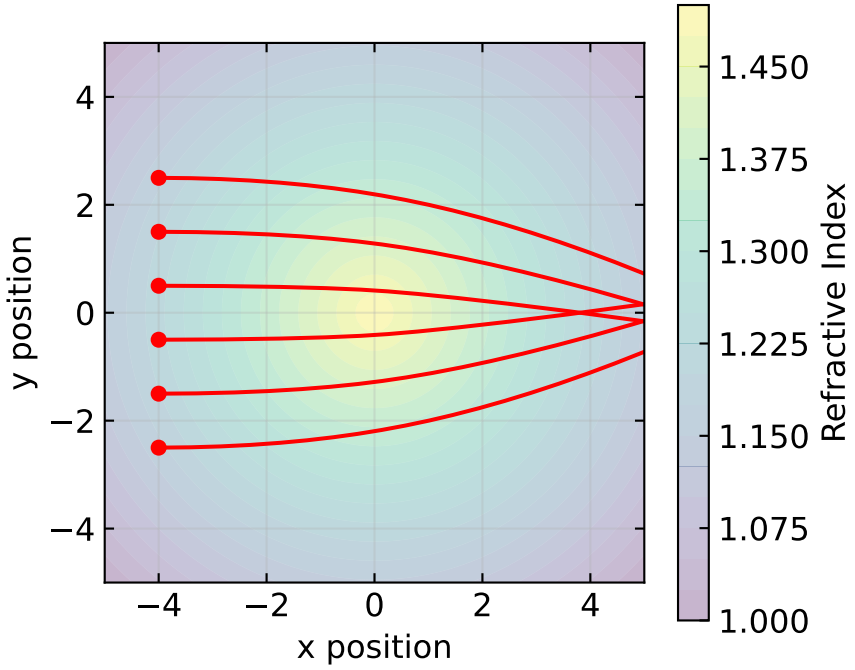


Figure 2.7: F=ma optics - Light rays (red) following paths toward regions of higher refractive index

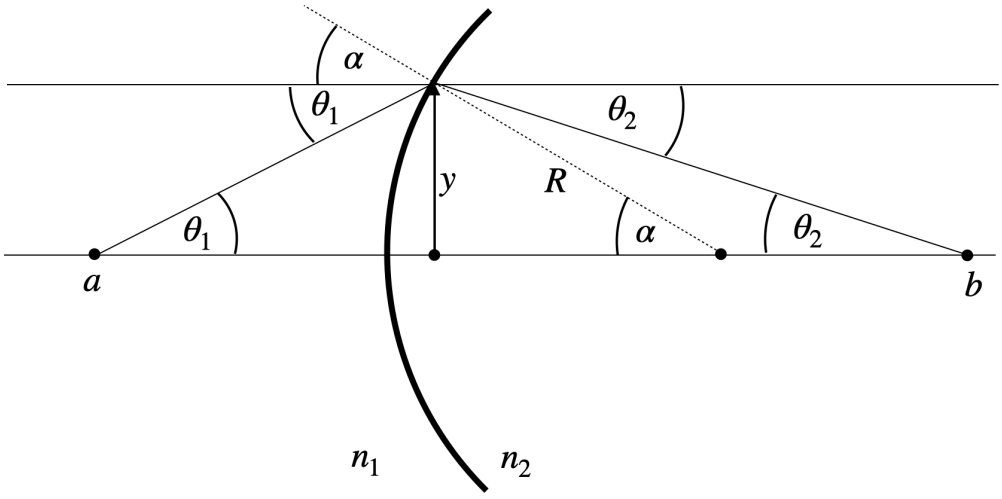


Figure 2.8: Refraction at a curved surface.

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.

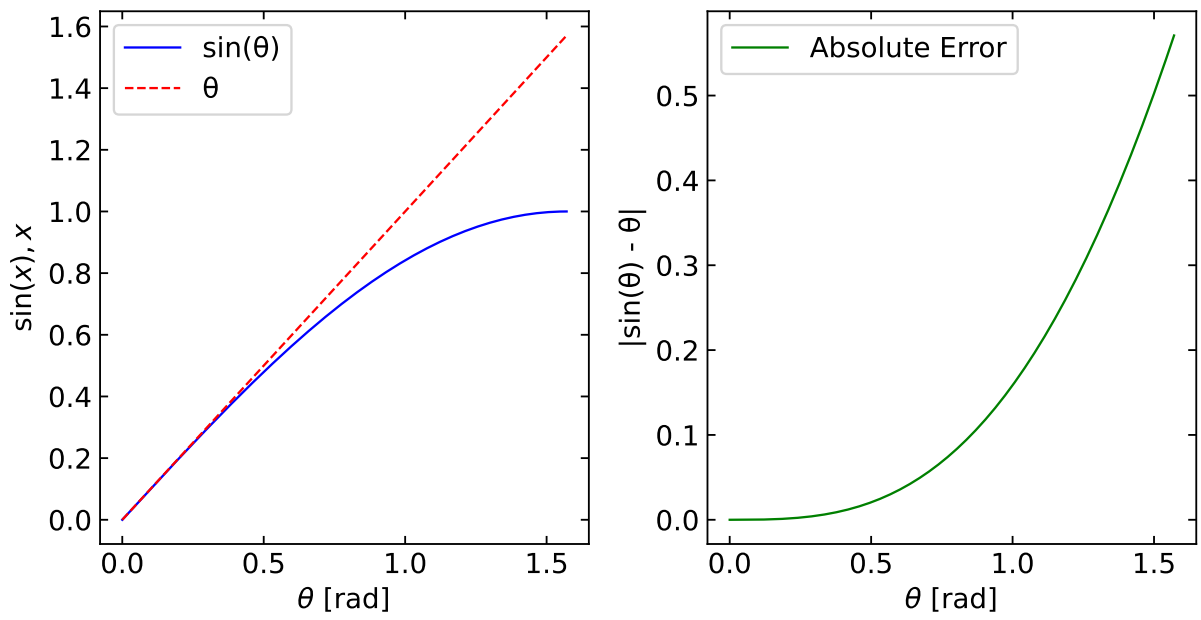


Figure 2.9: Visualization of the paraxial approximation plotting the $\sin(\theta)$ and the linear approximation θ (dashed line) for angles ranging from 0 to $\pi/2$ radians.

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:

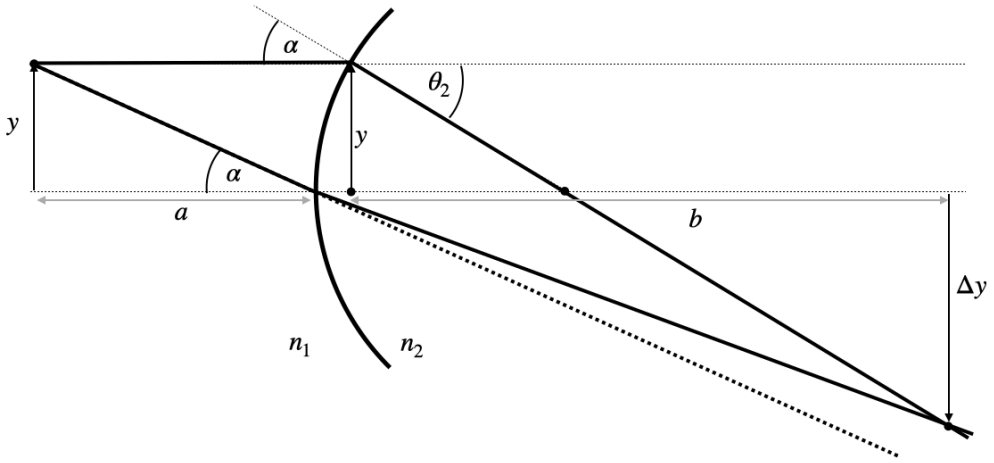


Figure 2.10: Image formation at a curved surface.

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:

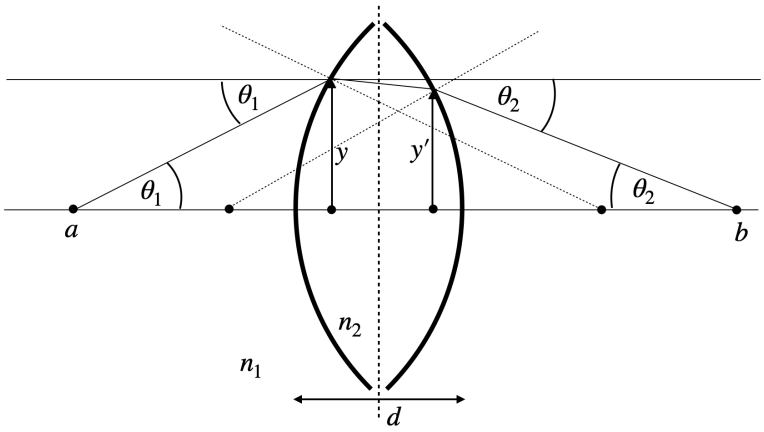


Figure 2.11: Refraction on two spherical surfaces.

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:

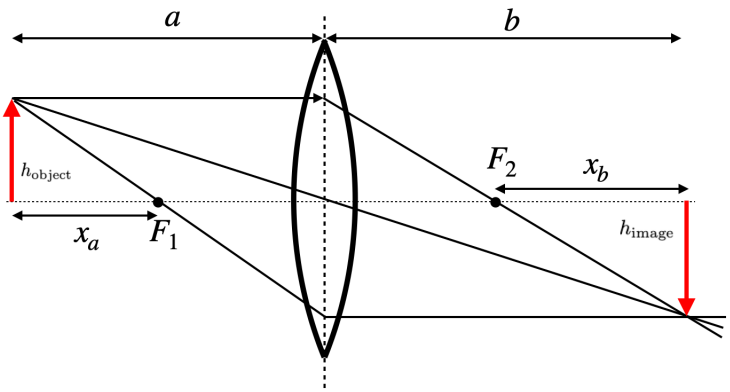


Figure 2.12: Image construction on a thin lens.

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.

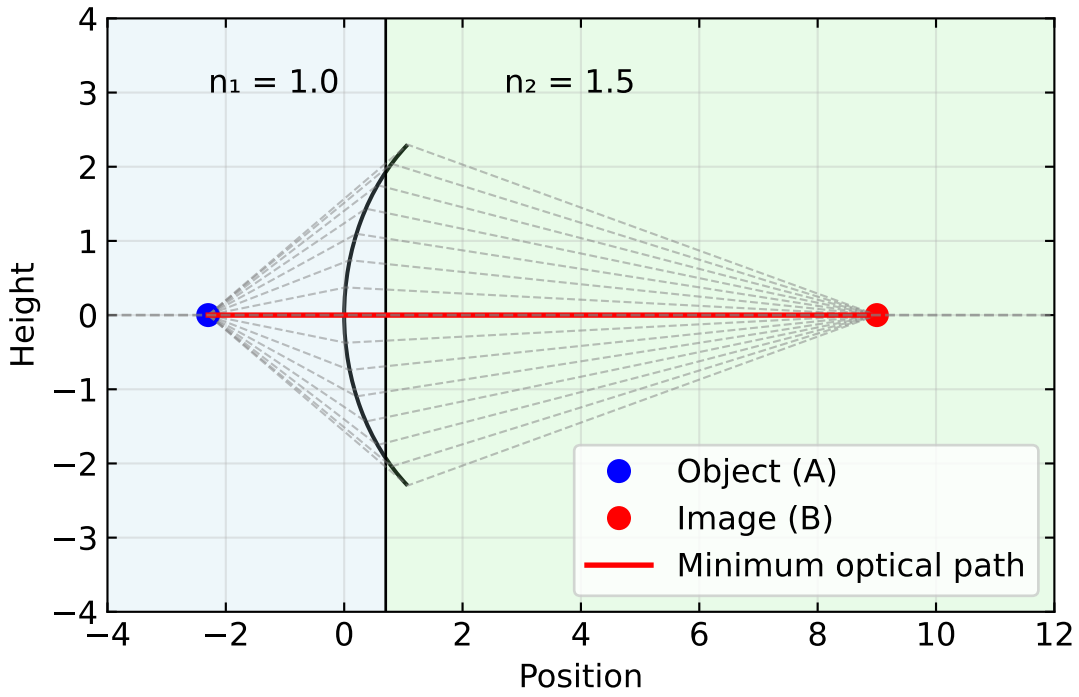


Figure 2.13: Fermat's principle applied to a spherical refracting surface

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.

Chapter 3

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.

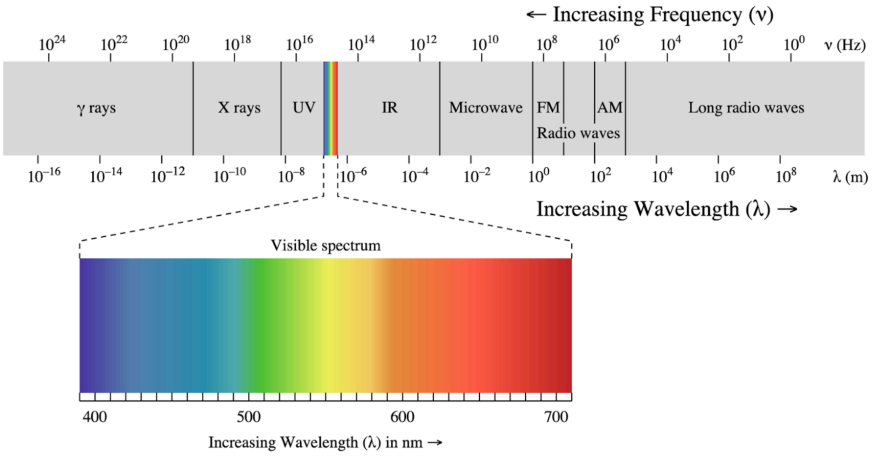


Figure 3.1: Electromagnetic Spectrum with its different regions

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.

3.1 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:

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

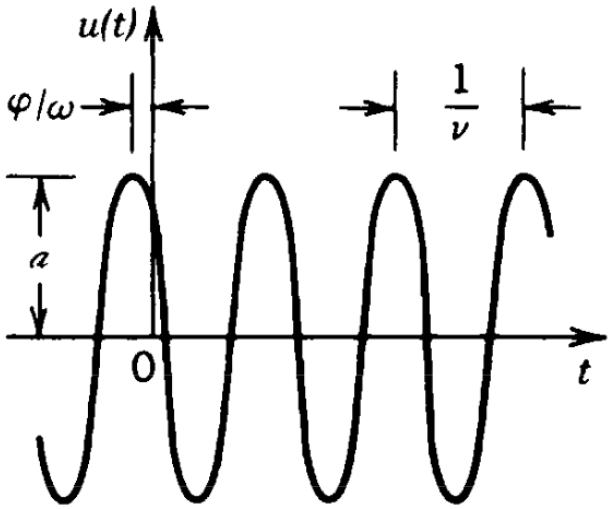


Figure 3.2: Representation of a wavefunction over time (constant position) denoting the phase ϕ and the period $T = 1/\nu$

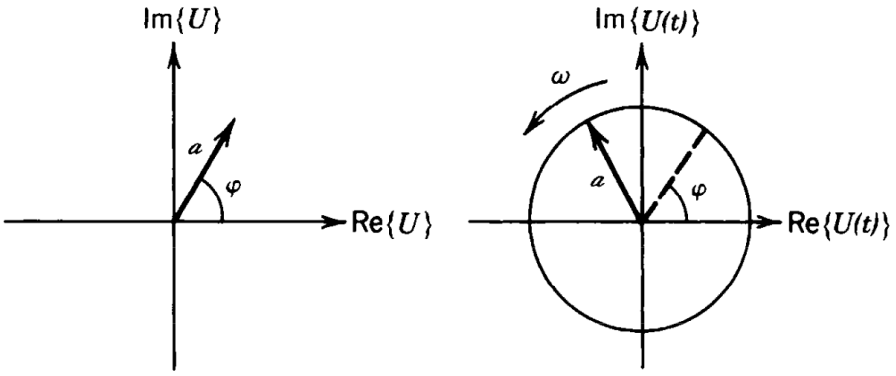


Figure 3.3: Phasor diagram of the complex amplitude $U(\mathbf{r})$ (left) and $U(t)$ (right)

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.

3.2 Plane Waves

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

$$U(\mathbf{r}, t) = Ae^{-i\mathbf{k}\cdot\mathbf{r}}e^{i\omega t} \tag{3.1}$$

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:

$$U(\mathbf{r}) = Ae^{-i\mathbf{k}\cdot\mathbf{r}} \tag{3.2}$$

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

$$\mathbf{k}_0 = \begin{pmatrix} k_{0x} \\ k_{0y} \\ k_{0z} \end{pmatrix} \tag{3.3}$$

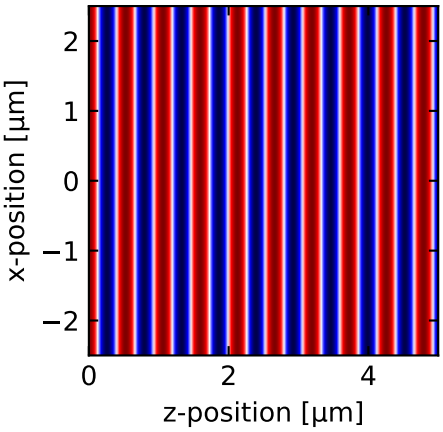


Figure 3.4: Plane wave propagating along the z-direction

3.3 Dispersion Relation

Using the plane wave solution

$$U(\mathbf{r}, t) = Ae^{-i\mathbf{k}\cdot\mathbf{r}}e^{i\omega t} \tag{3.4}$$

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

$$c_0 = \frac{\omega}{k_0} \tag{3.5}$$

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°.

3.4 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
$$k = nk_0$$

3.5 Snells Law

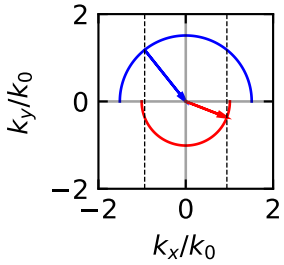
The change in the length of the wavevector has some simple consequence for Snells law. We can write Snells law as

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

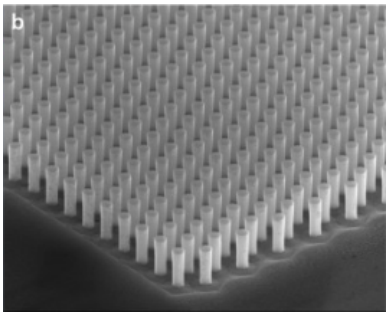
where k_0 is the wavevector length in vacuum. As the $n_1 k_0$ is the magnitude of the wavevector in medium 1, and $n_2 k_0$ is the magnitude of the wavevector in medium 2, we can rewrite Snells law as

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

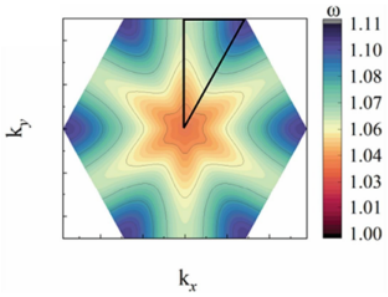
which means that the component of the wavevector parallel to the interface is conserved. If the wavevector has constant length then the wavevector incident at different angles is between a point on a circle and the origin in the diagram below. The circle corresponds to an isofrequency surface.



(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.

3.6 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:

$$U = \frac{A}{|\mathbf{r} - \mathbf{r}_0|} e^{-ik|\mathbf{r}-\mathbf{r}_0|} e^{i\omega t} \tag{3.6}$$

> **!** 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.

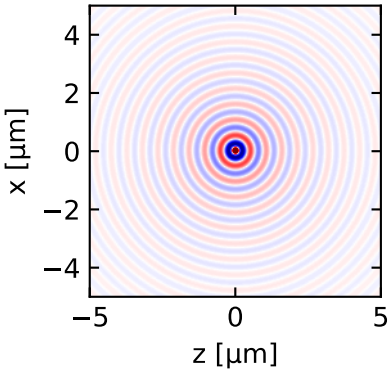


Figure 3.8: Spherical wave propagation. The wave is emitted from the origin and propagates in the positive z-direction. The wavefronts are spherical surfaces. The wave is visualized in the xz-plane.

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

Chapter 4

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:

$$\begin{aligned} I &= |U|^2 && (4.1)\\ &= |U_1 + U_2|^2 && (4.2)\\ &= |U_1|^2 + |U_2|^2 + U_1^* U_2 + U_1 U_2^* && (4.3) \end{aligned}$$

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 $I = 4I_0$

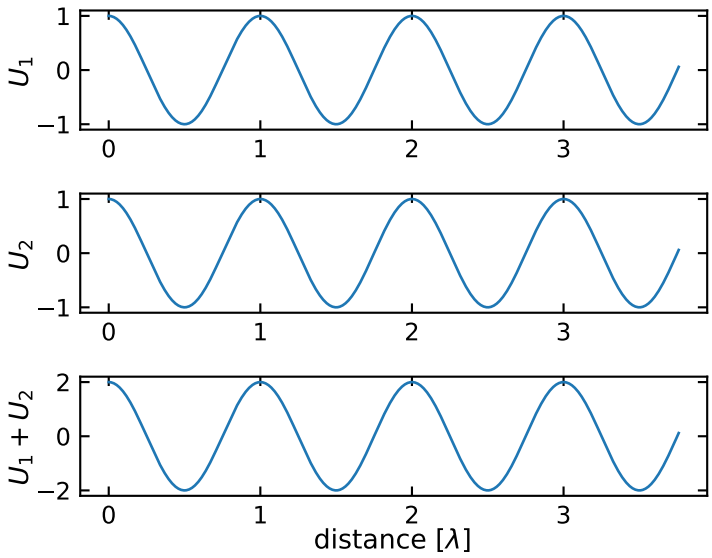


Figure 4.1: Constructive interference of two waves (top, middle) and the sum of the two wave amplitudes (bottom)

Destructive Interference

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

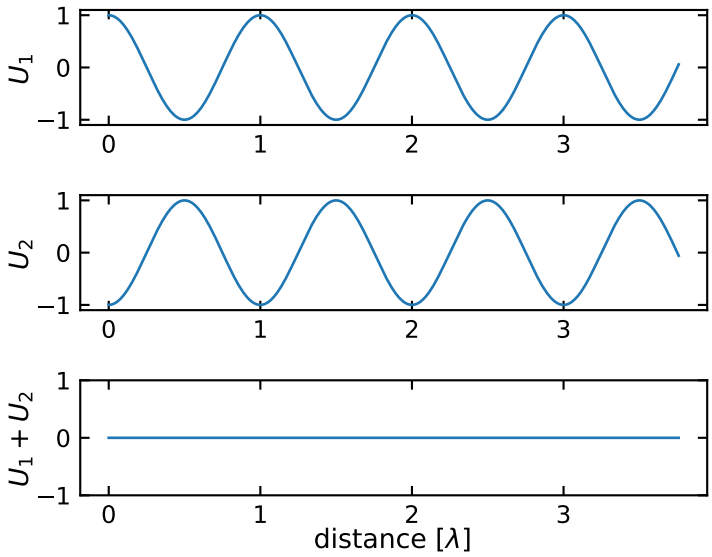


Figure 4.2: Destructive interference of two waves (top, middle) and the sum of the two wave amplitudes (bottom)

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

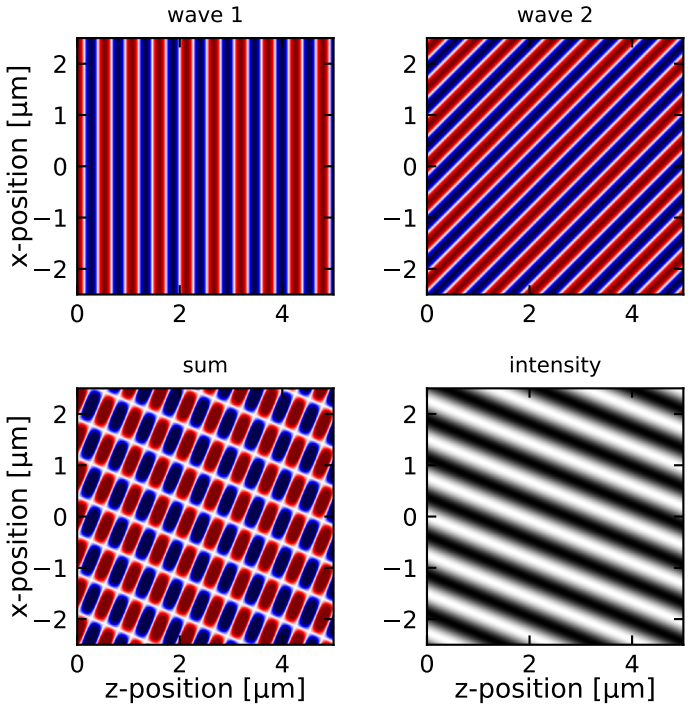


Figure 4.3: Interference of two plane waves propagating under an angle of 45°. The two left graphs show the original waves. The two right show the total amplitude and the intensity pattern.

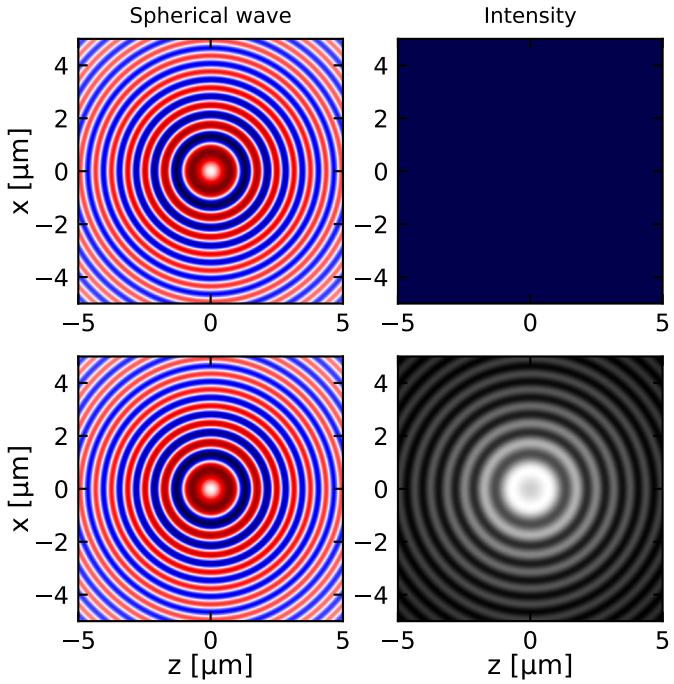


Figure 4.4: Interference of a spherical wave and a plane wave. The top graphs show the original waves. The two bottom show the total amplitude and the intensity pattern.

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.

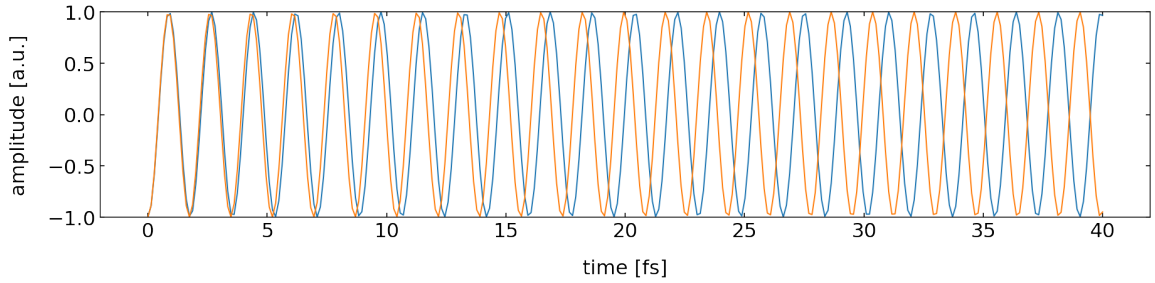


Figure 4.5: Two waves of different frequency over time.

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

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

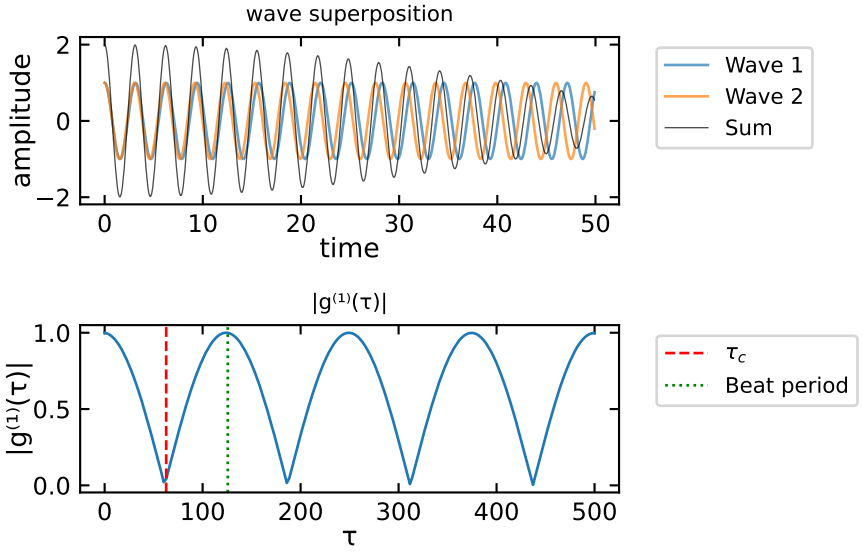


Figure 4.6: Temporal correlation for two waves with slightly different frequencies. The vertical line indicates the coherence time c = /Δ .

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.

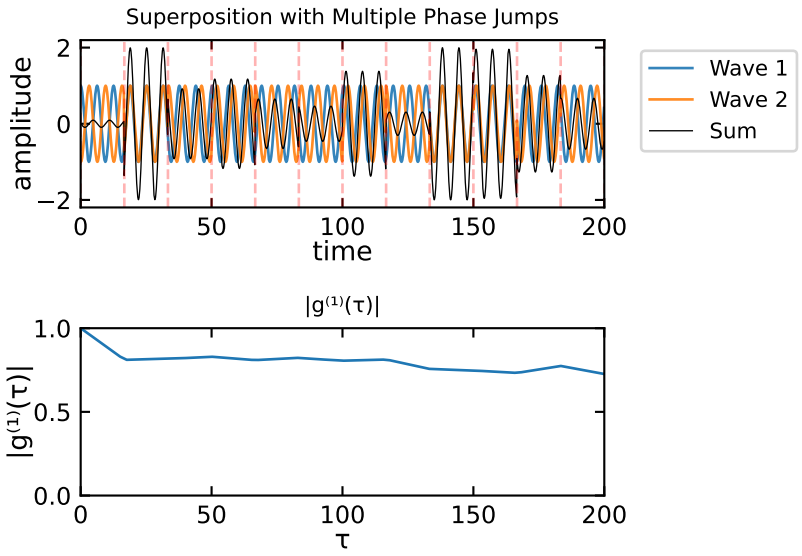


Figure 4.7: Temporal correlation for two waves of same frequency showing decoherence due to multiple phase jumps. Vertical lines indicate positions of phase jumps.

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

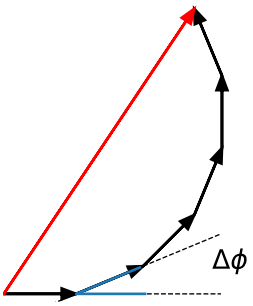
$$U = \sqrt{I_0}\frac{1-h^M}{1-h} = \sqrt{I_0}\frac{1-e^{iM\Delta\phi}}{1-e^{i\Delta\phi}}$$

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

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



(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.

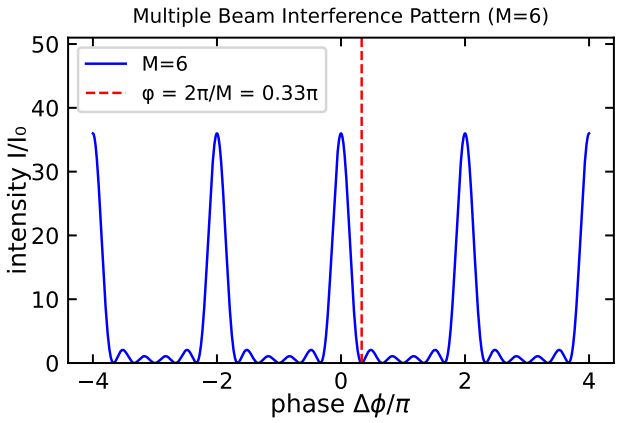


Figure 4.9: Multiple beam interference pattern for M=6 beams. The intensity distribution is shown as a function of the phase shift ϕ. The first minimum is at $\phi = 2\pi/M$. The intensity distribution is symmetric around $\phi = 0$.

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

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

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.

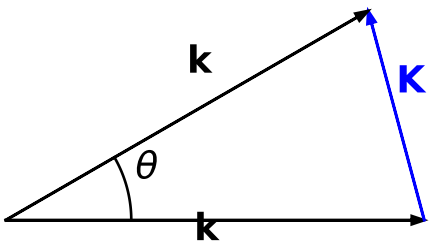


Figure 4.10: Wavevector summation for the diffraction grating. The wavevector of the incident light k and the wavevector of the light traveling along the direction of the first interference peak K form an equilateral triangle.

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.

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

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

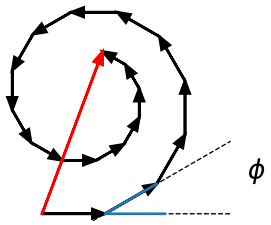


Figure 4.11: Phase construction of a multiwave intereference with M waves with decreasing amplitude due to a reflection coefficient $r = 0.95$.

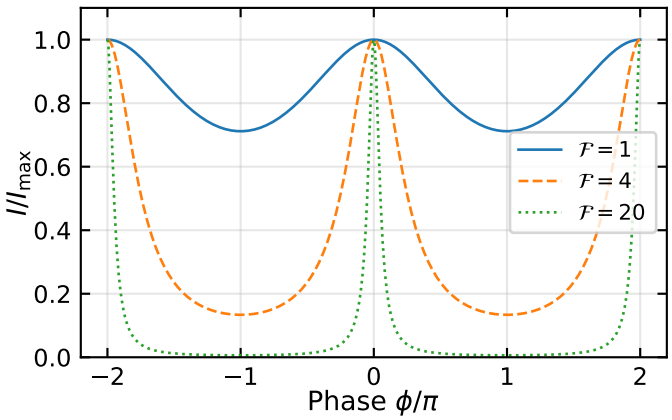


Figure 4.12: Multiple wave interference with decreasing amplitude. The graph shows the intensity distribution over the phase angle ϕ for different values of the Finesse $\mathcal{F}$.

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 $q = -L, \ldots, L$ 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 $I_{\max} = M^2 I_0$.

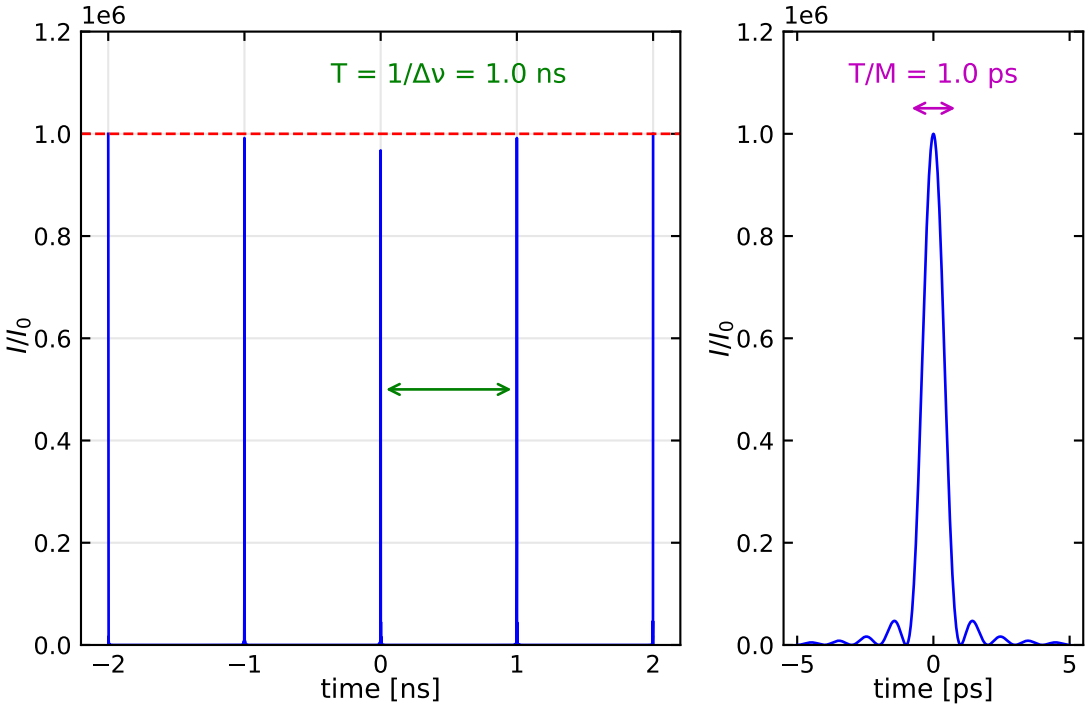


Figure 4.13: Multiple wave beating with M=1000 monochromatic waves separated by Δ =1 GHz. The intensity oscillates with period T=1/Δ =1 ns. Each pulse has a width of approximately T/M=1 ps with maximum intensity I_max=M²I .