OPTICS

Sander Konijnenberg, Aurèle J.L. Adam, & H. Paul Urbach
Delft University of Technology

BSc Optics

The LibreTexts libraries are Powered by NICE CXOne and are supported by the Department of Education Open Textbook Pilot Project, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

TABLE OF CONTENTS

Licensing

1: Basic Electromagnetic and Wave Optics

- 1.1: What You Should Know and be able to do After Studying This Chapter
- 1.2: Electromagnetic Theory of Optics and Quantum Optics
- 1.3: The Maxwell Equations in Vacuum
- 1.4: Maxwell Equations in Matter
- 1.5: The Scalar and Vector Wave Equation
- 1.6: Time-Harmonic Solutions of the Wave Equation
- 1.7: Time-Harmonic Maxwell Equations in Matter
- 1.8: Electromagnetic Energy
- 1.9: Time-Averaged Energy
- 1.10: Reflection and Transmission at an Interface
- 1.11: Fiber Optics

2: Geometrical Optics

- 2.1: What You Should Know and be able to do After Studying This Chapter
- 2.2: Introduction
- 2.3: Principle of Fermat
- 2.4: Some Consequences of Fermat's Principle
- 2.5: Perfect Imaging by Conic Sections
- 2.6: Gaussian Geometrical Optics
- 2.7: Beyond Gaussian Geometrical Optics
- 2.8: Aberrations

3: Optical Instruments

- 3.1: What you should know and be able to do after studying this chapter
- 3.2: The Camera Obscura
- 3.3: The Camera
- 3.4: Camera in a Smart Phone
- 3.5: The Human Eye
- 3.6: Magnifying Glasses
- 3.7: Eyepieces
- 3.8: The Compound Microscope
- 3.9: The Telescope

4: Polarization

- 4.1: What you should know and be able to do after studying this chapter
- 4.2: Polarisation States and Jones Vectors
- 4.3: Creating and Manipulating Polarisation States
- 4.4: How to verify whether a Jones Matrix is a Linear Polariser or a Wave Plate?
- 4.5: Decompose Elliptical Polarisation into Linear and Circular States

5: Interference and Coherence

- 5.1: What you should know and be able to do after studying this chapter
- 5.2: Introduction
- 5.3: Interference of Monochromatic Fields of the Same Frequency
- 5.4: Coherence
- 5.5: Temporal Coherence and the Michelson Interferometer
- 5.6: Longitudinally and Transversely Separated Points
- 5.7: Spatial Coherence and Young's Experiment
- 5.8: Increase of Spatial Coherence by Propagation
- 5.9: Stellar Interferometry
- 5.10: Fringe Visibility
- 5.11: Interference and polarisation

6: Scalar diffraction optics

- 6.1: What you should know and be able to do after studying this chapter
- 6.2: Introduction
- 6.3: Angular Spectrum Method
- 6.4: Rayleigh-Sommerfeld Diffraction Integral
- 6.5: Intuition for the Spatial Fourier Transform in Optics
- 6.6: Fresnel and Fraunhofer Approximations
- 6.7: Fraunhofer Diffraction Revisited
- 6.8: Fourier Optics
- 6.9: Super-resolution

7: Lasers

- 7.0: What you should know and be able to do after studying this chapter
- 7.1: Unique Properties of Lasers
- 7.2: Optical Resonator
- 7.3: Amplification
- 7.4: Cavities
- 7.5: Problems with Laser Operation
- 7.6: Types of Lasers

Index

Detailed Licensing

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

CHAPTER OVERVIEW

1: Basic Electromagnetic and Wave Optics

1.1: What You Should Know and be able to do After Studying This Chapter
1.2: Electromagnetic Theory of Optics and Quantum Optics
1.3: The Maxwell Equations in Vacuum
1.4: Maxwell Equations in Matter
1.5: The Scalar and Vector Wave Equation
1.6: Time-Harmonic Solutions of the Wave Equation
1.7: Time-Harmonic Maxwell Equations in Matter
1.8: Electromagnetic Energy
1.9: Time-Averaged Energy
1.10: Reflection and Transmission at an Interface
1.11: Fiber Optics

1.1: What You Should Know and be able to do After Studying This Chapter

This chapter is about Maxwell's equations and is a prerequisite for the rest of the book. It is assumed that the reader is already familiar with electromagnetism at the bachelor level. Therefore the treatments of Maxwell's equations in matter, boundary conditions at interfaces, electromagnetic energy, the field of an electric dipole and the reflection and transmission at an interface are rather concise. After studying this chapter you should

know and be able to

- Derive the scalar wave equation for the electromagnetic field components from Maxwell's equations.
- Work with the complex notation of time harmonic fields.
- Understand time harmonic plane waves, spherical waves and the concept of wave fronts.
- Know the main properties of the field radiated by an electric dipole as the fundamental solution of Maxwell's equations.
- Qualitatively understand the far field radiation pattern of en electric dipole (you do not need to know the formulas).
- Derive long-time averages of products of time-harmonic functions.
- Compute the rate of energy flow using the Poynting vector and its long-time average. (The derivation of the law of conservation of energy is not part of the examination).
- Understand the method of deriving the reflection and transmission of an incident plane wave at an interface by separating in s- and p-polarized states. The formulas for the Fresnel coefficients do not have to be known by hart.
- Understand the Brewster angle, total internal reflection and evanescent waves.
- Understand the principle of the guiding of electromagnetic waves.

1.2: Electromagnetic Theory of Optics and Quantum Optics

Maxwell's equations provide a very complete description of light, which includes diffraction, interference and polarisation. Yet it is strictly speaking not fully accurate, because it allows monochromatic electromagnetic waves to carry any amount of energy, whereas according to quantum optics the energy is quantised. According to quantum optics, light is a flow of massless particles, the photons, which each carry an extremely small quantum of energy:~ ω, where ~ = 6.63 × 10^{-34}/(2π) Js and ν is the frequency, which for visible light is of the order 5 × 10^{14} Hz. Hence ~ ω ≈ 3.3 × 10^{-19} J.

Quantum optics is only important in experiments involving a small number of photons, i.e. at very low light intensities and for specially prepared photons states (e.g. entangled states) for which there is no classical description. In almost all applications of optics the light sources emit so many photons that quantum effects are irrelevant see Figure 1.2.1

Light Source	Number of photons/s.m2
Laserbeam (10m W, He-Ne, focused to 20 µm)	10^{26}
Laserbeam (1 mW, He-Ne)	10^{21}
Bright sunlight on earth	10^{18}
Indoor light level	10^{16}
Twilight	10^{14}
Moonlight on earth	10^{12}
Starlight on earth	10^{10}

Table 1.2.1: The mean photon flflux density for some common sources

The visible part is only a small part of the overall electromagnetic spectrum (see Figure 1.2.1). The results we will derive are however generally valid for electromagnetic waves of any frequency.

1.3: The Maxwell Equations in Vacuum

In a vacuum, light is described by vector fields E(r, t) [Volt/m][1] and B(r, t) [Tesla=Weber/m^2= kg/(C s)], which vary extremely rapidly with position vector r and time t. These vector fields are traditionally called the electric field strength and the magnetic induction, respectively, and together they are referred to as "the electromagnetic field". This terminology is explained by the fact that, because in optics these fields vary with time, the electric and magnetic fields always occur together, i.e. one does not exist without the other. Only when the fields are independent of time, there can be an electric field without a magnetic field and conversely. The first case is called electrostatics, the second magnetostatics.

Time-dependent electromagnetic fields are generated by moving electric charges, the so-called sources. Let the source have charge density ρ(r, t) [C/m^3] and current density J (r, t) [C/(s.m^2)]. Since charge can not be created nor destroyed, the rate of increase of charge inside a volume V must be equal to the flux of charges passing through its surface S from the outside to the inside of V, i.e.:

$$\frac{d}{dt}\int_V \rho\, dV = \int_S J\cdot\hat{n}\, dS$$

where n is the outward-pointing unit normal on S. Using the Gauss divergence theorem, the left-hand side of (1.3.1) can be converted to a volume integral from which follows the differential form of the law of conservation of charge:

$$-\bigtriangledown\cdot J = \frac{\partial\rho}{\partial t}$$

At every point in space and at every time, the field vectors satisfy the Maxwell equations

$$\bigtriangledown\times\varepsilon = -\frac{\partial B}{\partial t}, Faraday's\ Law$$

$$\bigtriangledown\times\frac{B}{\mu_0} = \varepsilon_0\frac{\partial\varepsilon}{\partial t} + J, Maxwell's\ Law$$

$$\bigtriangledown\cdot\varepsilon_0\varepsilon = \rho, Gauss's\ Law$$

$$\bigtriangledown\cdot B = 0, no\ magnetic\ charge$$

where $\varepsilon_0 = 8.8544\times10^{-12}$ C ^{2}N^{-1}m^{-2} is the dielectric permittivity and $\mu_0 = 1.2566\times10^{-6}$m kg C^{-2} is the magnetic permeability of vacuum. The quantity $c = (1/\varepsilon_0\mu_0)^{1/2}$ is the speed of light in vacuum with numerical value of $2.997924562\times10^8\pm1.1$ m/s and $Z = (\mu_0/\varepsilon_0)^{1/2} = 377\Omega = 377$ Vs/C is the impedance of vacuum.

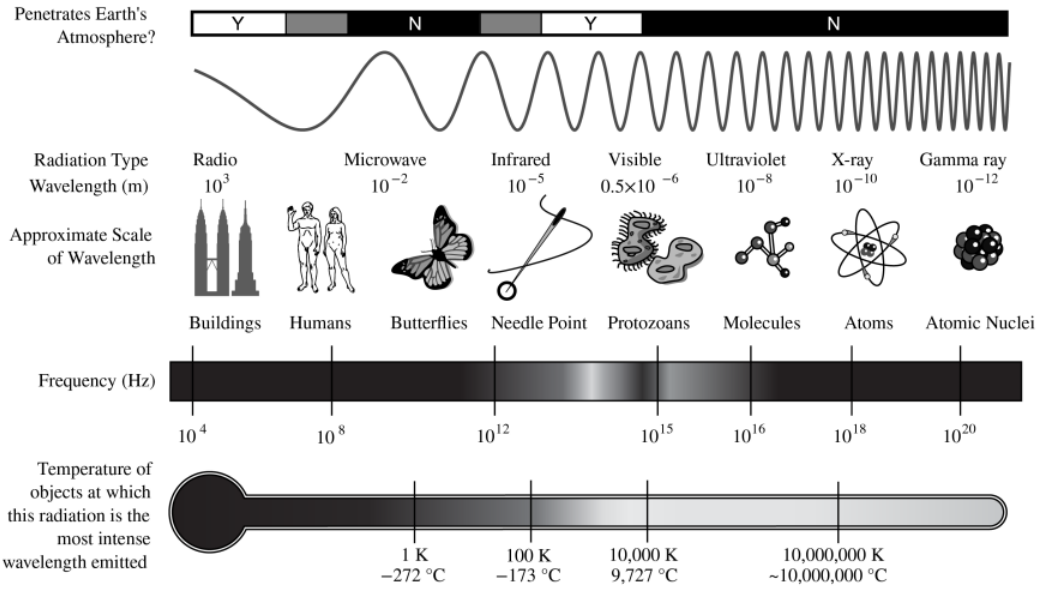


Figure 1.3.1: The electromagnetic spectrum. (from Wikimedia Commons by NASA/ CC BY-SA)

1.4: Maxwell Equations in Matter

Atoms are neutral and consist of a positively charged kernel surrounded by a negatively charged electron cloud. In an electric field, the centres of mass of the positive and negative charges get displaced with respect to each other. Therefore, an atom in an electric field behaves like an electric dipole. In polar molecules, the centres of mass of the positive and negative charges are permanently separated, even without an electric field. But without an electric field, they are randomly orientated and therefore have no net effect, while in the presence of an electric field they line up parallel to the field. Whatever the precise mechanism, an electric field induces a certain net dipole moment density per unit volume $\mathbf{P}(\mathbf{r})$ [C/m^2] i in matter which is proportional to the local electric field $\boldsymbol{\varepsilon}(\mathbf{r})$:

$$P(r,t) = E_0 \chi_e \varepsilon(r,t),$$

where χ_e is a dimensionless quantity, the electric susceptibility of the material. We stress that $\boldsymbol{\varepsilon}$ is the total local field at the position of the dipole, i.e. it contains the contribution of all other dipoles, which are also excited and radiate an electromagnetic field themselves. Only in the case of diluted gasses, the influence of the other dipoles in matter can be neglected and the local electric field is simply given by the field emitted by the external source.

A dipole moment density that changes with time corresponds to a current density $\mathbf{J}_p$ [Ampere/m^2=C/(m^2 s)] and a charge density ρ_p [C/m^3] given by

$$J_p(r,t) = \frac{\partial P(r,t)}{\partial t} = E_0 \chi \frac{\partial \varepsilon(r,t)}{\partial t},$$

$$\rho_p(r,t) = - \bigtriangledown \cdot P(r,t) = - \bigtriangledown \cdot (E_0 \chi \varepsilon),$$

All materials conduct electrons to a certain extent, although the conductivity σ [Ampere/(Volt m)=C/(Volt s] differs greatly between dielectrics, semi-conductors and metals (the conductivity of copper is 10^7 times that of a good conductor such as sea water and 10^{19} times that of glass). The current density $\mathbf{J}_c$ and the charge density corresponding to the conduction electrons satisfy:

$$J_c = \rho \varepsilon,$$

$$\frac{\partial \rho_c}{\partial t} = - \bigtriangledown \cdot J_c = - \bigtriangledown \cdot (\rho \varepsilon),$$

where (1.4.4) is Ohm's Law. The total current density on the right-hand side of Maxwell's Law (1.2.4) is the sum of $\mathbf{J}_p$, $\mathbf{J}_c$ and an external current density $\mathbf{J}_{ext}$, which we assume to be known. Similarly, the total charge density at the right of (1.2.5) is the sum of ρ_p, ρ_c and a given external charge density ρ_{ext}. The latter is linked to the external current density by the law of conservation of charge (1.2.2). Hence, (1.2.4) and (1.2.5) become

$$\bigtriangledown \times \frac{B}{\mu_0} = E_0 E_0 \frac{\partial \varepsilon}{\partial t} + J_p + J_c + J_{ext} = E_0 (1+\chi) \frac{\partial \varepsilon}{\partial t} + \sigma \varepsilon + J_{ext}$$

$$\bigtriangledown \cdot E_0 \varepsilon = \rho_p + \rho_c + \rho_{ext} = - \bigtriangledown \cdot (E_0 \chi \varepsilon) + \rho_c + \rho_{ext}\,.$$

We define the permittivity E by

$$E = E_0 (1 + \chi_e).$$

Then (1.4.6) and (1.4.7) can be written as

$$\bigtriangledown \times \frac{B}{\mu_0} = E \frac{\partial \varepsilon}{\partial t} + \sigma \varepsilon + J_{ext}$$

$$\bigtriangledown \cdot (E \varepsilon) = \rho_c + \rho_{ext}\,.$$

It is verified in Problem 1 that in a conductor any accumulation of charge is extremely quickly reduced to zero. Therefore we may assume that

$$\rho_c = 0.$$

If the material is magnetic, the magnetic permeability is different from vacuum and is written as $\mu = \mu_0(1 + \chi_m)$, where χ_m is the magnetic susceptibility. In the Maxwell equations, one should then replace μ_0 by μ. However, at optical frequencies magnetic

effects are negligible (except in ferromagnetic materials, which are rare). **We will therefore always assume that the magnetic permeability is that of vacuum**: $\mu = \mu_0$.

It is customary to define the magnetic field by $\mathbf{H} = \mathbf{B}/\mu_0$ [Ampere/m=C/(ms)]. By using the magnetic field **H** instead of the magnetic induction **B**, Maxwell's equations become more symmetric:

$$\bigtriangledown \times \varepsilon = -\mu_0 \frac{\partial H}{\partial t}, Faraday's\ Law$$

$$\bigtriangledown \times H = E\frac{\partial \varepsilon}{\partial t} + \sigma\varepsilon + J_{ext}, Maxwell's\ Law$$

$$\bigtriangledown \cdot E\varepsilon = \rho_{ext}, Gauss's\ Law$$

$$\bigtriangledown \cdot H = 0. no\ magnetic\ charge.$$

This is the form in which we will be using the Maxwell equations in matter in this book. It is seen that the Maxwell equations in matter are identical to those in vacuum, with E substituted for E_0.

1.5: The Scalar and Vector Wave Equation

We consider a homogeneous insulator (i.e. E is independent of position and σ=0) in which there are no external sources:

$$J_{ext}\,,\rho_{ext}=0.$$

In optics the external source, e.g. a laser, is normally spatially separated from objects of interest with which the light interacts. Hence the assumption that the external source vanishes in the region of interest is often justified. Take the curl of (1.12) and the time derivative of (1.13) and add the equations obtained. This gives

$$\bigtriangledown\times\bigtriangledown\times\varepsilon+E\mu_0\frac{\partial^2\varepsilon}{\partial t^2}=0.$$

Now for any vector field **A** there holds:

$$\bigtriangledown\times\bigtriangledown\times A=-\bigtriangledown^2 A+\bigtriangledown\bigtriangledown\cdot A.$$

where ∇^2A is the vector:

$$\bigtriangledown^2 A=\bigtriangledown^2 A_x\hat{x}+\bigtriangledown^2 A_y\hat{y}+\bigtriangledown^2 A_z\hat{z},$$

with

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

Because Gauss's law (1.3.14) with ρ_{ext} = 0 and E constant implies that $\nabla \cdot$ E = 0, (1.5.3) applied to E yields

$$\bigtriangledown\times\bigtriangledown\times\varepsilon=-\bigtriangledown^2\varepsilon.$$

Hence, (1.5.2) becomes

$$\bigtriangledown^2\varepsilon-E\mu_0\frac{\partial^2\varepsilon}{\partial t^2}=0.$$

By a similar derivation it is found that also **H** satisfies (1.5.7). Hence in a homogeneous dielectric without external sources, every component of the electromagnetic field satisfies the scalar wave equation:

$$\bigtriangledown^2 U-E\mu_0\frac{\partial^2 U}{\partial t^2}=0.$$

The **refractive index** is the dimensionless quantity defined by

$$n=(\frac{E}{E_0})^{1/2}.$$

The scalar wave equation can then be written as

$$\bigtriangledown^2 U-n^2E_0\mu_0\frac{\partial^2 U}{\partial t^2}=0.$$

The speed of light in matter is

$$\frac{c}{n}=\frac{1}{(E\mu_0)^{1/2}}.$$

1.6: Time-Harmonic Solutions of the Wave Equation

The fact that, in the frequently occurring circumstance in which light interacts with a homogeneous dielectric, all components of the electromagnetic field satisfy the scalar wave equation, justifies the study of solutions of this equation. Since in most cases in optics monochromatic fields are considered, we will focus our attention on time-harmonic solutions of the wave equation.

1.5.1 Time-Harmonic Plane Waves

Time-harmonic solutions depend on time by a cosine or a sine. One can easily verify by substitution that

$$U(r,t) = Acos(kx - \omega t + \phi),$$

where A > 0 and ϕ are constants, is a solution of (1.4.10), provided that

$$k = \omega(E\mu_0)^{1/2} = \omega n(E_0\mu_0)^{1/2} = nk_0,$$

where $k_0 = \omega(E_0\mu_0)^{1/2}$ is the wave number in vacuum. The frequency ω > 0 can be chosen arbitrarily. The wave number k in the material is then determined by (1.6.2). We define T = 2π/ω and λ = 2π/k as the period and the wavelength in the material, respectively. Furthermore, $\lambda_0 = 2\pi/k_0$ is the wavelength in vacuum.

Remark. When we speak of "the wavelength", we always mean the wavelength in vacuum.

We can write (1.6.1) in the form

$$U(x,t) = Acos[k(x - \frac{c}{n}t) + \phi],$$

where $c/n = 1/(E\mu_0)^{1/2}$ is the speed of light in the material. A is the amplitude and the argument under the cosine: k(x-ct/n)+ϕ is called the phase at position x and at time t. A wavefront is a set of space-time points where the phase is constant:

$$x - \frac{c}{n}t = constant.$$

At any fixed time t the wave fronts are planes (in this case perpendicular to the x-axis), and therefore the wave is called a plane wave. As time proceeds, the wavefronts move with velocity c/n in the positive x-direction.

A time-harmonic plane wave propagating in an arbitrary direction is given by

$$U(r,t) = Acos(k \cdot r - \omega t + \phi),$$

where A and ϕ are again constants and $\mathbf{k} = k_x\mathbf{x} + k_y\mathbf{y} + k_z\mathbf{z}$ is the wave vector. The wavefronts are given by the set of all space-time points (**r**, t) for which the phase $\mathbf{k} \cdot \mathbf{r} - \omega t + \phi$ is constant, i.e. for which

$$k \cdot r - \omega t = constant.$$

At fixed times the wavefronts are planes perpendicular to the direction of k as shown in Fig.1.6.1 Eq. (1.6.5) is a solution of (1.4.10) provided that

$$k_x^2 + k_y^2 + k_z^2 = \omega^2 E\mu_0 = \omega^2 n^2 E_0\mu_0 = k_0^2 n^2.$$

The direction of the wave vector can be chosen arbitrarily, but its length is determined by the frequency ω.

1.5.2 Complex Notation for Time-Harmonic Functions

We consider a general time-harmonic solution of the wave equation (1.4.8):

$$U(r,t) = A(r)cos(\phi(r) - \omega t),$$

where the amplitude A(r) > 0 and the phase ϕ(r) are functions of position r. The wavefronts consist of space-time points (r, t) where the phase is constant:

$$\phi(r) - \omega t = constant.$$

At fixed time t, the sets of constant phase: ϕ(r) = ωt + constant are surfaces which in general are not planes, hence the solution in general is not a plane wave. Eq. (1.6.9) could for example be a wave with spherical wavefronts, as discussed below.

Remark. A plane wave is infinitely extended and carries and transports an infinite amount of electromagnetic energy. A plane plane can therefore not exist in reality, but it is nevertheless a usual idealisation because, as will be demonstrated in in Section 7.1, *every time-harmonic solution of the wave equation can always* be expanded in terms of plane waves of the form (1.6.5).

For time-harmonic solutions it is often convenient to use **complex notation**. Define the **complex amplitude** by:

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

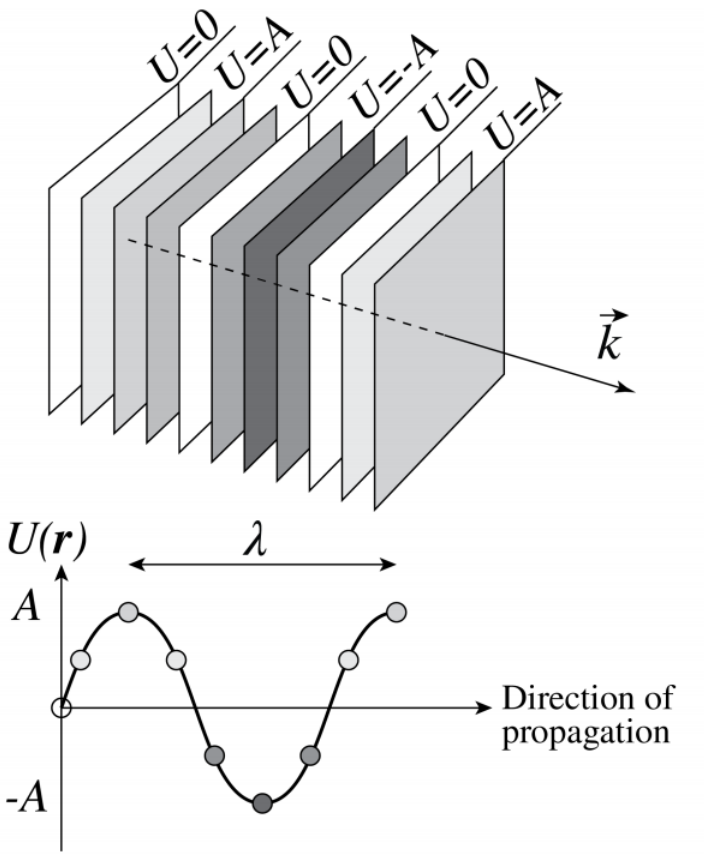


Figure 1.6.1: Planes of constant phase.

i.e. the modulus of the complex number U(**r**) is the amplitude A(**r**) and the argument of U(**r**) is the phase ϕ(**r**) at t = 0. The time-dependent part of the phase: −ωt is thus separated from the space-dependent part of the phase. Then (1.6.8) can be written as

$$U(r,t) = Re[U(r)e^{-i\omega t}].$$

Hence U(**r**, t) is the real part of the complex time-harmonic function

$$U(r)e^{-i\omega t}.$$

Remark. The complex amplitude U(**r**) is also called the complex field. In the case of vector fields such as **E** and **H** we speak of complex vector fields, or simply complex fields. Complex amplitudes and complex (vector) fields are only functions of position **r**; the time dependent factor exp(−iωt) is omitted. To get the physical meaningful real quantity, the complex amplitude or complex field first has to be multiplied by exp(−iωt) and then the real part must be taken.

The following convention is used throughout this book:

Real-valued physical quantities (whether they are time-harmonic or have more general time dependence) are denoted by a calligraphic letter, e.g. U, $\mathcal{E}_x$, or $\mathcal{H}_x$. The symbols are bold when we are dealing with a vector, e.g. $\boldsymbol{\mathcal{E}}$ or $\boldsymbol{\mathcal{H}}$. The complex amplitude of a time-harmonic function is linked to the real physical quantity by (1.6.11) and is written as an ordinary letter such as U and E.

It is easier to calculate with complex amplitudes (complex fields) than with trigonometric functions (cosine and sine). As long as all the operations carried out on the functions are linear, the operations can be carried out on the complex quantities. To get the real-

valued physical quantity of the result (i.e. the physical meaningful result), multiply the finally obtained complex amplitude by exp(−iωt) and take the real part. The reason that this works is that taking the real part commutes with linear operations, i.e. taking first the real part to get the real-valued physical quantity and then operating on this real physical quantity gives the same result as operating on the complex scalar and taking the real part at the end.

By substituting (1.6.12) into the wave equation (1.4.10) we get

$$\begin{aligned}\bigtriangledown^2 U(r,t) - n^2 E_0\mu_0 \frac{\partial^2 U(r,t)}{\partial t^2} &= Re[\bigtriangledown^2 U(r) e^{-i\omega t}] - n^2 E_0\mu_0 Re[U(r)\frac{\partial^2 e^{-i\omega t}}{\partial^2}] \\ &= Re[\bigtriangledown^2 U(r) + \omega^2 n^2 E_0\mu_0 U(r)] e^{-i\omega t}.\end{aligned}$$

Since this must vanish for all times t, it follows that the complex expression between the brackets {.} must vanish. To see this, consider for example the two instances t = 0 and t = π/(2ω. Hence we conclude that the complex amplitude satisfies

$$\bigtriangledown^2 U(r) + k_0^2 n^2 U(r) = 0, Helmholtz\ Equation,$$

where $k_0 = \omega(E_0\mu_0)^{1/2}$ is the wave number in vacuum.

Remark. The complex quantity of which the real part has to be taken is: U exp(−iωt). It is not necessary to drag the time-dependent factor exp(−iωt) along in the computations: it suffices to calculate only with the complex amplitude U, then multiply by exp(−iωt) and then take the real part. However, when a derivative with respect to time is to be taken: ∂/∂t the complex field much be multiplied by −iω. This is also done in the time-harmonic Maxwell's equations in Section 1.6 below.

1.5.3 Time-Harmonic Spherical Waves

A spherical wave depends on position only by the distance to a fixed point. For simplicity we choose the origin of our coordinate system at this point. We thus seek a solution of the form U(r, t) with r = (x2 + y2 + z2)1/2. For spherical symmetric functions we have

$$\bigtriangledown^2 U(r,t) = \frac{1}{r}\frac{\partial^2}{\partial r^2}[rU(r,t)]$$

It is easy to see that outside of the origin

$$U(r,t) = \frac{f(\pm r - ct/n)}{r},$$

satisfies (1.47) for any choice for the function f, where, as before, c = 1/(E0μ0)1/2 is the speed of light and n = (E/E0)1/2. Of particular interest are time-harmonic spherical waves:

$$U(r,t) = \frac{A}{r} cos[k(\pm r - \frac{c}{n}t) + \phi] = \frac{A}{r} cos(\pm kr - \omega t + \phi)$$

where **A** is a constant

$$k = n\omega/c.$$

and ±kr − ωt + ϕ is the phase at **r** and at time t. The wavefronts are space-time points (**r**, t) where the phase is constant:

$$\pm kr - \omega t = constant,$$

which are spheres which move with the speed of light in the radial direction. When the + sign is chosen, the wave propagates outwards, i.e. away from the origin. The wave is then radiated by a **source** at the origin. Indeed, if the + sign holds in (1.6.17), then if time t increases, (1.6.19) implies that a surface of constant phase moves outwards. Similarly, if the − sign holds, the wave propagates towards the origin which then acts as a **sink**. The amplitude of the wave **A**/r is proportional to the inverse distance to the source of sink. Since the local flux of energy is proportional to the square $\mathbf{A}^2/r^2$, the total flux through the surface of any sphere with centre the origin is independent of the radius of the sphere. Since there is a source or a sink at the origin, (1.6.17) satisfies (1.6.15) only outside of the origin. There is a δ-function as source density on the right-hand side:

$$E\mu_0 \frac{\partial^2}{\partial t^2} U(r,t) - \bigtriangledown^2 U(r,t) = 4\pi A\delta(r),$$

where the right-hand side corresponds to either a source or sink at the origin, depending on the sign chosen in the phase.

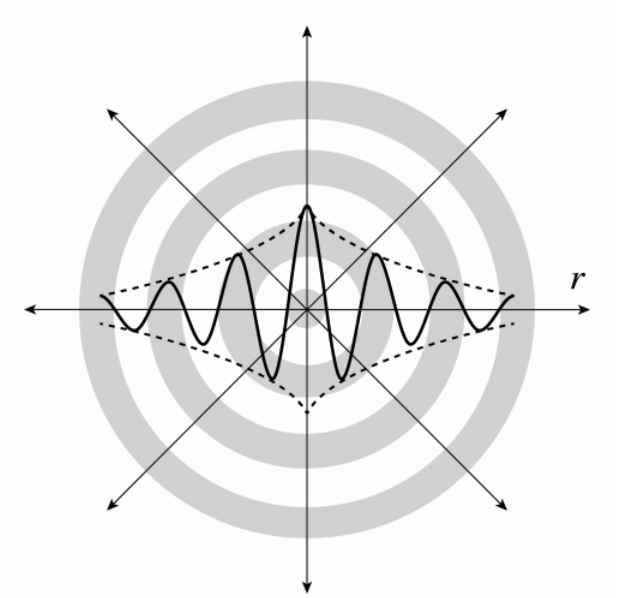


Figure 1.6.2: Spherical wavefronts with amplitude decreasing with distance.

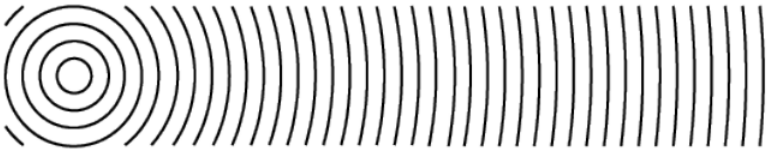


Figure 1.6.3: Planes of constant phase in cross-section. For an observer at large distance to the source the spherical wave looks similar to a plane wave.

Using complex notation we have for the outwards propagating wave:

$$U(r,t) = Re[U(r)e^{-i\omega t}] = Re[\frac{A}{r}e^{i(kr-i\omega t)}]$$

with U(**r**) = A exp(ikr)/r and A = **A** exp(iϕ), where ϕ is the argument and **A** the modulus of the complex number A.

In Figure 1.6.2 and Figure 1.6.3 spherical wavefronts are shown. For an observer who is at a large distance from the source, the spherical wave looks like a plane wave which propagates from the source towards the observer (or in the opposite direction, if there is a sink).

1.7: Time-Harmonic Maxwell Equations in Matter

We now return to the Maxwell equations and consider time-harmonic electromagnetic fields, because these are by far the most important fields in optics. Using complex notation we have

$$E(r,t) = Re[E(r)e^{-i\omega t}],$$

with $E_x(\mathbf{r}) = |E_x(\mathbf{r})|e^{i\phi x(\mathbf{r})}$, $E_y(\mathbf{r}) = |E_y(\mathbf{r})|e^{i\phi y(\mathbf{r})}$, $E_z(\mathbf{r}) = |E_z(\mathbf{r})|e^{i\phi z(\mathbf{r})}$, where $\phi_x(\mathbf{r})$ is the argument of the complex number $E_x(\mathbf{r})$ etc. With similar notations for the magnetic field, we obtain by substitution into Maxwell's equations (1.3.12), (1.3.13), (1.3.14) and (1.3.15), the time-harmonic Maxwell equations for the **complex fields**:

$$\nabla \times E = i\omega\mu_0 H, Faraday's\ Law$$

$$\nabla \times H = -i\omega\varepsilon E + \sigma E + J_{ext}, Maxwell's\ Law$$

$$\nabla \cdot \varepsilon E = \rho_{ext}, Gauss's\ Law$$

$$\nabla \cdot H = 0, no\ magnetic\ charge$$

where the time derivative has been replaced by multiplication of the complex fields by $-i\omega$.

In the time-harmonic Maxwell equations, the conductivity is sometimes included in the imaginary part of the permittivity:

$$E = E_0[1 + \chi_e + i\frac{\sigma}{\omega}].$$

Although it is convenient to do this in Maxwell's Law (1.7.3), one should remember that in Gauss's Law (1.7.4), the original permittivity: $E = 1 + \chi_e$ should still be used. When there are no external sources: $\rho_{ext} = 0$ and the material is homogeneous (i.e. χ_e and σ are independent of position), then (1.7.4) is equivalent to

$$\nabla \cdot E = 0.$$

Hence in this (important) special case, definition (1.7.6) for the permittivity can safely be used without the risk of confusion.

We see that when we use definition (1.7.6), the conductivity makes the permittivity complex and depending on frequency. But actually, also for insulators ($\sigma = 0$), the permittivity E depends in general on frequency and is complex with a positive imaginary part. The positive imaginary part of E is a measure of the absorption of the light by the material. The property that the permittivity depends on the frequency is called **dispersion**. Except close to a resonance frequency, the imaginary part of E (ω) is small and the real part is a slowly increasing function of frequency. This is called normal dispersion. This is illustrated with the refractive index of different glass shown in Figure 1.7.1

Near a resonance, the real part is rapidly changing and *decreases* with ω (this behaviour is called anomalous dispersion), while the imaginary part has a maximum at the resonance frequency of the material, corresponding to maximum absorption at a resonance as seen in Figure 1.7.2 At optical frequencies, mostly normal dispersion occurs and for small-frequency bands such as in laser light, it is often sufficiently accurate to use the value of the permittivity and the conductivity at the centre frequency of the band.

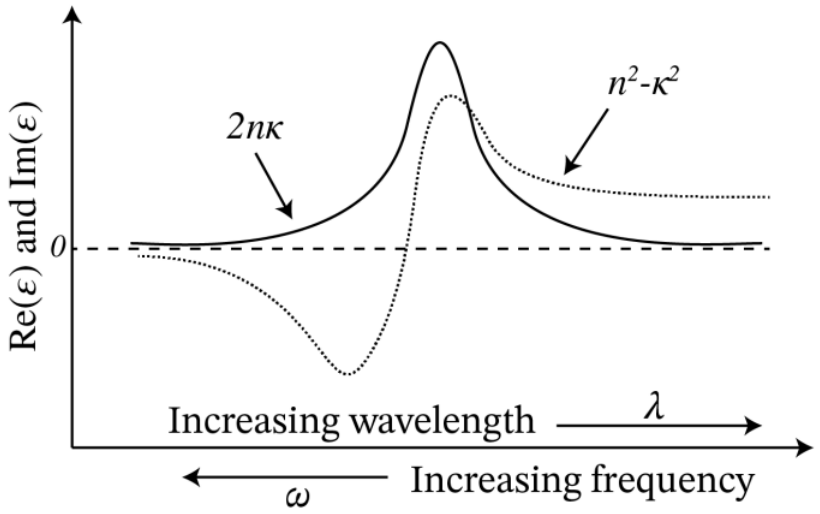


Figure 1.7.1: Real part $n^2 - \kappa^2$ and imaginary part $2n\kappa$ of the permittivity $E = (n + i\kappa)^2$, as function of wavelength and of frequency near a resonance.

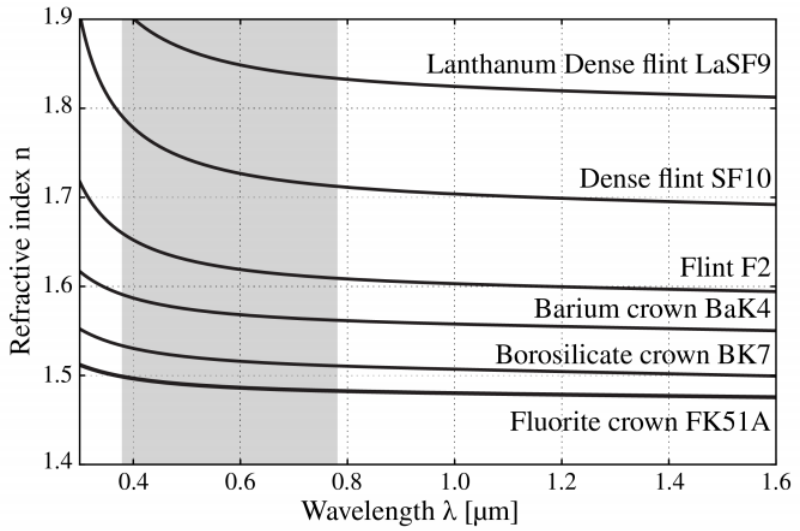


Figure 1.7.2: Refractive index as function of wavelength for several types of glass (from Wikimedia Commons by Geek3 / CC BY-SA).

In many books the following notation is used: $E = (n + i\kappa)^2$, where n and κ ("kappa", not to be confused with the wave number k) are both real and positive, with n the refractive index and κ a measure of the absorption. We then have $\mathrm{Re}(E) = n^2 - \kappa^2$ and $\mathrm{Im}(E) = 2n\kappa$ (see Figure 1.7.1). Note that although n and κ are both positive, Re(E) can be negative for some frequencies. This happens for metals in the visible part of the spectrum.

Remark. When E depends on frequency, Maxwell's equations (1.3.13) and (1.3.14) for fields that are not time-harmonic can strictly speaking not be valid, because it is not clear which value of E corresponding to which frequency should be chosen. In fact, in the case of strong dispersion, the products $E\boldsymbol{\varepsilon}$ should be replaced by convolutions in the time domain. Since we will almost always consider fields with a narrow-frequency band, we shall not elaborate on this issue further.

1.6.1 Time-Harmonic Electromagnetic Plane Waves

In this section we assume that the material in which the wave propagates has conductivity which vanishes: $\sigma = 0$, does not absorb the light and is homogeneous, i.e. that the permittivity E is a real constant. Furthermore, we assume that in the region of space of interest there are no sources. These assumptions imply in particular that (1.7.7) holds. The electric field of a time-harmonic plane wave is given by

$$E(r,t) = Re[E(r)e^{-i\omega t}],$$

with

$$E(r) = Ae^{ik\cdot r},$$

where **A** is a constant complex vector (i.e. it is independent of position and time):

$$A = A_x\hat{x} + A_y\hat{y} + A_z\hat{z},$$

with Ax = |Ax|$e^{i\phi x}$ etc.. The wave vector **k** satisfies (1.5.7). Substitution of (1.7.9) into (1.7.7) implies that

$$E(r)\cdot k = 0$$

for all **r** and hence (1.7.8) implies that also the physical real electric field is in every point **r** perpendicular to the wave vector: **ε**(**r**, t) · **k** = 0. For simplicity we now choose the wave vector in the direction of the z-axis and we assume that the electric field vector is parallel to the x-axis. This case is called a x-polarised electromagnetic wave. The complex field is then written as

$$E(x) = Ae^{ikz}\hat{x},$$

where k = $\omega(E\mu_0)^{1/2}$ and A = |A| exp(iϕ). It follows from Faraday's Law (1.7.2) that

$$H(z) = \frac{k}{\omega\mu_0}\hat{z}\times\hat{x}Ae^{ikz} = (\frac{E}{\mu_0})^{1/2}Ae^{ikz}\hat{y}.$$

The real electromagnetic field is thus:

$$\varepsilon(z,t) = Re[E(z)e^{-i\omega t}] = |A|cos(kz-\omega t+\phi)\hat{x},$$

$$H(z,t) = Re[H(z)e^{-i\omega t}] = (\frac{E}{\mu_0})^{1/2}|A|cos(kz-\omega t+\phi)\hat{y}.$$

We conclude that **in a lossless medium, the electric and magnetic field of a plane wave are in phase and at every point and at every instant perpendicular to the wave vector and to each other**. As illustrated in Figure 1.7.3, at any given point both the electric and the magnetic field achieve their maximum and minimum values at the same time.

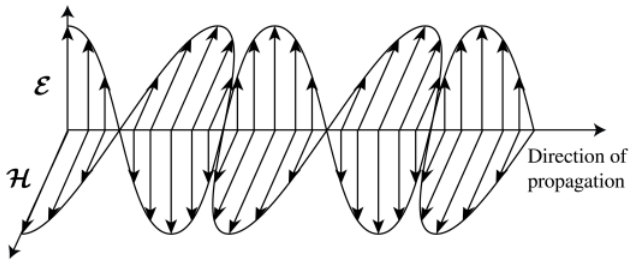


Figure 1.7.3: The time-harmonic vectors **ε** and **H** of a plane polarised wave are perpendicular to each other and to the direction of the wave vector which is also the direction of **ε** × **H**.

1.6.2 Field of an Electric Dipole

An other important solution of Maxwell's equation is the field radiated by a time-harmonic electric dipole, i.e. two opposite charges with equal strength that move time-harmonically around their total centre of mass. In this section the medium is homogeneous, but it may absorb part of the light, i.e. the permittivity may have a nonzero imaginary part. An electric dipole is the classical electromagnetic model for an atom or molecule. Because the optical wavelength is much larger than an atom of molecule, these charges may be considered to be concentrated both in the same point $\mathbf{r}_0$. The charge and current densities of such an elementary dipole are

$$\rho = -p\cdot\nabla\delta(r-r_0),$$

$$J = -i\omega p\delta(r-r_0),$$

with **p** the dipole vector, defined by

$$p = qa,$$

where q > 0 is the positive charge and **a** is the position vector of the positive with respect to the negative charge.

The field radiated by an electric dipole is very important. It is the fundamental solution of Maxwell's equations, in the sense that the field radiated by an arbitrary distribution of sources can always be written as a superposition of the fields of elementary electric dipoles. This follows from the fact that Maxwell's equations are linear and any current distribution can be written as a superposition of elementary dipole currents.

The field radiated by an elementary dipole in $\mathbf{r}_0$ in homogeneous matter can be computed analytically and is given by

$$E(r) = [k2\hat{R} \times (p \times \hat{R}) + (3\hat{R} \cdot p\hat{R} - p)(\frac{1}{R^2} - \frac{ik}{R})]\frac{e^{ikR}}{4\pi ER},$$

$$H(r) = \frac{k^2 c}{n}(1 + \frac{ik}{R})\hat{R} \times p\frac{e^{ikR}}{4\pi R},$$

where k = k_0, n = $(E/E_0)^{1/2}$, with k_0 the wave number in vacuum and with **R** = **r** − $\mathbf{r_0}$. It is seen that the complex electric and magnetic fields are proportional to the complex spherical wave: e^{ikR}/R discussed in Section 1.5.3, but that these fields contain additional position dependent factors. In particular, at large distance to the dipole:

$$H(r) \approx \frac{k^2 c}{n}\hat{R} \times p\frac{e^{ikR}}{4\pi R},$$

\[E(r) ≈k^2 \hat{R} × (p × \hat{R})\dfrac{e^{ikR}}{4πER}=-(\dfrac{μ_{0}}{E})\hat{R} × H(r).

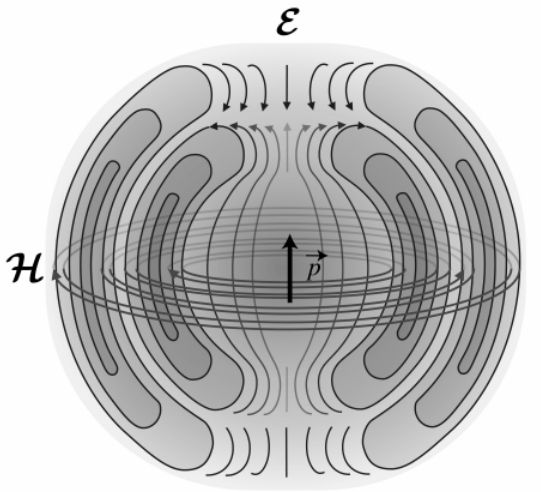


Figure 1.7.4: Electric and magnetic field lines created by a radiating dipole (from Wikimedia Commons, original JPG due to Averse, SVG by Maschen. / CC0).

In Figure 1.7.4 are drawn the electric and magnetic field lines of a radiating dipole. For an observer at large distance from the dipole, the electric and magnetic fields are perpendicular to each other and perpendicular to the direction of the line of sight **R** from the dipole to the observer. Furthermore, the electric field is in the plane through the dipole vector **p** and the vector **R**, while the magnetic field is perpendicular to this plane. So, for a distant observer the dipole field is similar to that of a plane wave which propagates from the dipole towards the observer and has an electric field parallel to the plane through the dipole and the line of sight **R** and perpendicular to **R**. Furthermore, the amplitudes of the electric and magnetic fields depend on the direction of the line of sight, with the field vanishing when the line of sight **R** is parallel to the dipole vector **p** and with maximum amplitude when **R** is in the plane perpendicular to the dipole vector. This result agrees with the well-known radiation pattern of an antenna when the current of the dipole is in the same direction as that of the antenna.

1.8: Electromagnetic Energy

The total energy stored in the electromagnetic field per unit of volume at a point **r** is equal to the sum of the electric and the magnetic energy densities. We postulate that the results for the energy densities derived in electrostatics and magnetostatics are also valid for the fast-oscillating fields in optics; hence we assume that the total electromagnetic energy density is given by:

$$U_{em}(r,t) = \frac{E}{2}\varepsilon(r,t)\cdot\varepsilon(r,t) + \frac{\mu_0}{2}H(r,t)\cdot H(r,t).$$

It is to be noticed that we assume in this section that the permittivity is real, i.e. there is no absorption and the permittivity does not include the conductivity.

Time dependent electromagnetic fields propagate energy. The flow of electromagnetic energy at a certain position r and time t is given by the Poynting vector, which is defined by

$$S(r,t) = \varepsilon(r,t)\times H(r,t).$$

More precisely, the flow of electromagnetic energy through a small surface dS with normal **n** at point **r** is given by

$$S(r,t)\cdot\hat{n}dS.$$

If this scalar product is positive, the energy flow is in the direction of **n**, otherwise it is in the direction of −**n**. Hence the direction of the vector **S**(**r**, t) is the direction of the flow of energy at point r and the length |**S**(**r**, t)| is the amount of the flow of energy, per unit of time and per unit of area perpendicular to the direction of **S**. This quantity has unit J/(sm^2). That the Poynting vector gives the flow of energy can be seen in a dielectric for which dispersion may be neglected by the following derivation. We consider the change with time of the total electromagnetic energy in a volume V :

$$\frac{d}{dt}\iiint_V U_{em}(r,t)\,dV = \iiint_V E\frac{\partial\varepsilon(r,t)}{\partial t}\cdot\varepsilon(r,t) + \mu_0\frac{\partial H(r,t)}{\partial t}\cdot H(r,t)\,dV$$

By substituting (1.3.12), (1.3.13) and using

$$-A\cdot\nabla\times B + B\cdot\nabla\times A = \nabla\cdot(A\times B),$$

which holds for any two vector fields, we find

$$\iiint_V E\frac{\partial\varepsilon(r,t)}{\partial t}\cdot\varepsilon(r,t) + \mu_0\frac{\partial H(r,t)}{\partial t}\cdot H(r,t)\,dV = \iiint_V \varepsilon(r,t)\cdot\nabla\times H(r,t) - H(r,t)\cdot\nabla\times\varepsilon(r,t)\,dV - \iiint_V \sigma$$

$$\varepsilon(r,t)\cdot\varepsilon(r,t)\,dV - \iiint_V \varepsilon(r,t)\cdot J_{ext}(r,t)\,dV = -\iiint_V \nabla\cdot(\varepsilon\times H)\,dV - \iiint_V \sigma\varepsilon(r,t)\cdot\varepsilon(r,t)\,dV - \iiint_V \varepsilon(r,t)$$

$$\cdot J_{ext}(r,t)\,dV = -\iint_S (\varepsilon\times H)\cdot\hat{n}\,dS - \iiint_V \sigma\varepsilon(r,t)\cdot\varepsilon(r,t)\,dV - \iiint_V \varepsilon(r,t)\cdot J_{ext}(r,t)\,dV,$$

where S is the surface bounding volume V and **n** is the unit normal on S pointing out of V. Hence,

$$\frac{d}{dt}\iiint_V U_{em}(r,t)\,dV + \iiint_V \sigma\varepsilon(r,t)\cdot\varepsilon(r,t)\,dV + \iiint_V \varepsilon(r,t)\cdot J(r,t)\,dV = -\iint_S S(r,t)\cdot\hat{n}\,dS.$$

This equation says that the rate of change with time of the electromagnetic energy in a volume V plus the work done by the field on the conduction and external currents inside V is equal to the influx of electromagnetic energy through the boundary of V.

Remark. The energy flux **S** and the energy density U_{em} depend quadratically on the field. For Uem the quadratic dependence on the electric and magnetic fields is clear. To see that the Poynting vector is also quadratic in the electromagnetic field, one should realise that the electric and magnetic fields are inseparable: they together form the electromagnetic field. Stated differently: if the amplitude of the electric field is doubled, then also that of the magnetic field is doubled and hence the Poynting vector is increased by the factor 4. Therefore, when computing the Poynting vector or the electromagnetic energy density of a time-harmonic electromagnetic field, the real-valued vector fields should be used, i.e. the complex fields should **NOT** be used.An exception is the

calculation of the long-time average of the Poynting vector or the energy density. As we will show in the next section, the **time averages** of the energy flux and energy density of time-harmonic fields can actually be expressed quite conveniently in terms of the complex field amplitudes.

If we subsitute the real fields (1.6.12), (1.6.13) of the plane wave in the Poynting vector and the electromagnetic energy density we get:

$$S(z,t) = \varepsilon(z,t) \times H(z,t) = (\frac{E}{\mu_0})^{1/2} |A|^2 cos^2(kz - \omega t + \phi)\hat{z},$$

$$U_{em}(z,t) = E|A|^2 cos^2(kz - \omega t + \phi).$$

We see that the energy flow of a plane wave is in the direction of the wave vector, which is also the direction of the phase velocity. Furthermore, it changes with time at frequency 2ω.

1.9: Time-Averaged Energy

Optical frequencies are in the range of 5×10^{14} Hz and the fastest detectors working at optical frequencies have integration times larger than 10^{-10} s. Hence there is no detector which can measure the time fluctuations of the electromagnetic fields at optical frequencies and any detector always measures an average value, taken over an interval of time that is very large compared to the period $2\pi/\omega$ of the light wave, typically at least a factor 10^5 longer. We therefore compute averages over such time intervals of the Poynting vector and of the electromagnetic energy. Because the Poynting vector and energy density depend nonlinearly (quadratically) on the field amplitudes, we can not perform the computations using the complex amplitudes and take the real part afterwards, but have instead to start from the real quantities. Nevertheless, it turns out that the final result can be conveniently expressed in terms of the complex field amplitudes.

Consider two time-harmonic functions:

$$A(t) = Re[Ae^{-i\omega t}] = |A|cos(\phi_A - \omega t)$$

$$B(t) = Re[Be^{-i\omega t}] = |B|cos(\phi_B - \omega t)$$

with $A = |A| \exp(i\phi A)$ and $B = |B| \exp(i\phi B)$ the complex amplitudes. For a general function of time $f(t)$ we define the time average over an interval T at a certain time t, by

$$\frac{1}{T}\int_{t-T/2}^{t+T/2} f(t')\,dt'.$$

where T is much larger (say a factor of 10^5) than the period of the light. It is obvious that for time-harmonic fields the average does not depend on the precise time t at which it is computed. and we therefore take t = 0 and write

$$\langle f(t)\rangle = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} f(t)\,dt.$$

With A(t) = Re [$Ae^{-i\omega t}$]=1/2[$Ae^{-i\omega t}$+A*$e^{i\omega t}$],where $A*$ is the complex conjugate of A; and with a similar expression for B(t), it follows that

$$\lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} A(t)B(t)\,dt = \lim_{T\to\infty}\frac{1}{4T}\int_{-T/2}^{T/2}[AB* + A*B + ABe^{-2i\omega t} + A*B*e^{2i\omega t}]\,dt =$$

$$\lim_{T\to\infty}\frac{1}{4T}[AB* + A*B + AB\frac{e^{i\omega t}-e^{-i\omega t}}{2iT\omega} + A*B*\frac{e^{i\omega t}-e^{-i\omega t}}{2iT\omega}] = \frac{1}{2}Re[AB*],$$

This important result will be used over and over again. In words:

The average of the product of two time-harmonic quantities over a long time interval compared with the period, is half the real part of the product of the complex amplitude of one quantity and the complex conjugate of the other.

If we apply this to Poynting's vector of a general time-harmonic electromagnetic field:**ε**(**r**, t) = Re [**E**(**r**)$e^{-i\omega t}$], **H**(**r**, t) = Re [**H**(**r**)$e^{-i\omega t}$],then we find that the time-averaged energy flow denoted by **S**(**r**) is given by

$$S(r) = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} S(r,t)\,dt = \frac{1}{2}Re[E\times H*].$$

Similarly, the time-averaged electromagnetic energy density is:

$$\langle U_{en}(r)\rangle = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} U_{en}(r,t)\,dt = \frac{1}{2}E\varepsilon(r)\cdot\varepsilon(r)* + \frac{\mu_0}{2}H(r)\cdot H(r)* = \frac{1}{2}E|\varepsilon(r)|^2 + \frac{\mu_0}{2}|H(r)|^2.$$

For the special case of plane wave (1.6.12), (1.6.13) in a medium without absorption, we get:

$$S = \frac{1}{2}(\frac{E}{\mu_0})^{1/2} Re[AA*]\hat{z} = \frac{1}{2}(\frac{E}{\mu_0})^{1/2}|A|^2\hat{z}.$$

The length of vector (1.9.8) is the time-averaged flow of energy per unit of area in the direction of the plane wave and is commonly called the **intensity** of the wave. For the time-averaged electromagnetic energy density of the plane wave, we get:

$$\langle U_{en} \rangle = \frac{1}{2}E|A|^2 + \frac{1}{2\mu_0}\mu_0 E|A|^2 = E|A|^2.$$

For a plane wave both the time-averaged energy flux and the time-averaged energy density are proportional to the modulus squared of the complex electric field.

1.10: Reflection and Transmission at an Interface

When an electromagnetic field is incident on an interface between different media, the field is partially reflected and partially transmitted. An important special case is that of a monochromatic plane wave which is incident on a planar interface as in Figure 1.10.2

Let the interface be the plane $z = 0$ between materials in $z < 0$ and $z > 0$ with permittivities E_i and E_t, respectively. We first assume that the materials are lossless, i.e. that the permittivities are real. The plane wave is incident from medium $z < 0$ and the incident electromagnetic field is given by:

$$\varepsilon^i(r,t) = Re[\varepsilon^i(r)e^{-i\omega t}] \tag{1.10.1}$$

$$= Re\left[A^i e^{i(k^i \cdot r - \omega t)}\right] \tag{1.10.2}$$

and

$$H^i(r,t) = Re[H^i(r)e^{-i\omega t}] \tag{1.10.3}$$

$$= Re\left[\frac{k^i}{\omega\mu_0} \times A^i e^{i(k^i \cdot r - \omega t)}\right] \tag{1.10.4}$$

where $\mathbf{k}^i = \mathbf{k}^i_x\mathbf{x} + \mathbf{k}^i_y\mathbf{y} + \mathbf{k}^i_z\mathbf{z}$, with

$$k_z^i = (k_0^2 E_i - (k_x^i)^2 - (k_y^i)^2)^{1/2}.$$

Because the time dependence is given by $\exp(-i\omega t)$ with $\omega > 0$ and the incident wave propagates in the positive z-direction, the positive square root is chosen for k^i_z . Part of the incident field is reflected into $z < 0$ and part is transmitted into $z > 0$. The reflected field is written as

$$\varepsilon^r(r,t) = Re[\varepsilon^r(r)e^{-i\omega t}] = Re[A^r e^{i(k^r \cdot r - \omega t)}],$$

$$H^r(r,t) = Re[H^r(r)e^{-i\omega t}] = Re[\frac{k^r}{\omega\mu_0} \times A^r e^{i(k^r \cdot r - \omega t)}],$$

where $\mathbf{k}^r = \mathbf{k}^r_x\mathbf{x} + \mathbf{k}^r_y\mathbf{y} + \mathbf{k}^r_z\mathbf{z}$, with

$$k_z^r = (k_0^2 E_i - (k_x^r)^2 - (k_y^r)^2)^{1/2},$$

where the minus sign is chosen because the reflected wave propagates in the negative z-direction. The transmitted field is for $z > 0$

$$\varepsilon^t(r,t) = Re[\varepsilon^t(r)e^{-i\omega t}] = Re[A^t e^{i(k^t \cdot r - \omega t)}],$$

$$H^t(r,t) = Re[H^t(r)e^{-i\omega t}] = Re[\frac{k^t}{\omega\mu_0} \times A^t e^{i(k^t \cdot r - \omega t)}],$$

where $\mathbf{k}^t = \mathbf{k}^t_x\mathbf{x} + \mathbf{k}^t_y\mathbf{y} + \mathbf{k}^t_z\mathbf{z}$, with

$$k_z^t = (k_0^2 E_i - (k_x^t)^2 - (k_y^t)^2)^{1/2}.$$

Our aim is to determine $\mathbf{A}^r$ and $\mathbf{A}^t$ for given $\mathbf{A}^i$.

1.9.1 Boundary Conditions at an Interface

There exist conditions for the continuity of the tangential and the normal components of both the electric and magnetic fields at an interface between different media. The boundary conditions for the tangential components follow from the Maxwell equations that contain the curl-operator, i.e. (1.6.2) and (1.6.3). There holds for the interface z = 0 with the incident, reflected and transmitted plane waves introduced above:

$$\hat{z} \times (E^i + E^r) = \hat{z} \times E^t,$$

$$\hat{z} \times (H^i + H^r) = \hat{z} \times H^t,$$

where **z** is the unit normal on the interface. This means that the tangential components of the total electric and total magnetic field are continuous across the interface, or explicitly:

$$E_x^i(x,y,0)+E_x^r(x,y,0)=E_x^t(x,y,0),$$
$$E_y^i(x,y,0)+E_y^r(x,y,0)=E_y^t(x,y,0),$$

and similarly for the magnetic field.

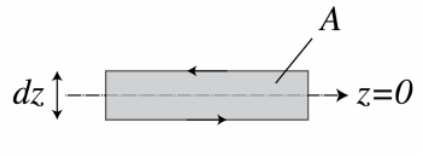


Figure 1.10.1: Closed loop in the (x, z)-plane enclosing the area A and surrounding part of the interface z = 0, as used in Stokes' Law to derive the continuity of the electric and magnetic components which are tangential to the interface and parallel to the plane through the loop.

We will only demonstrate the continuity of the tangential components for the electric field. By choosing a closed loop in the (x, z)-plane which is intersected by the interface z = 0 as shown in Figure 1.10.1, and integrating the y-component of Faraday's Law (1.3.12) for the total electromagnetic field over the area A bounded by the loop L, we obtain:

$$-\mu_0 \frac{d}{dt}\iint_A \hat{y} H\, dA = \iint_A \hat{y}\nabla\times\varepsilon\, dA = \oint_L \varepsilon\, dl$$

where in the last step we used Stokes' theorem with the direction of integration over the loop given by that of the direction of rotation of screw driver when it moves in the direction of the normal **y**. In words: the rate of change of the magnetic flux through the surface A is equal to the integral of the tangential electric field over the bounding closed loop L.

By taking the limit dz → 0, the surface integral and the integrals over the vertical parts of the loop vanish and there remain only the integrals of the tangential electric field over the horizontal parts of the loop on both sides of the interface z = 0. Since these integrals are traversed in opposite directions and the lengths of these parts are the same and arbitrary, we conclude for the loop as shown in Figure 1.10.1 that

$$lim_{z\uparrow 0}\varepsilon_x(x,y,z,t)=lim_{z\downarrow 0}\varepsilon_x(x,y,z,t),$$

where **ε** is the total electric field, i.e. it is equal to the sum of the incident and the reflected field for z < 0, and equal to the transmitted field in z > 0. By choosing the closed loop in the (y, z)-plane instead of the (x, z)-plane one finds similarly that the y-component of the electric field is continuous. The continuity of the tangential components of the magnetic field are derived in a similar manner.

Our derivation holds for electromagnetic fields of arbitrary time dependence. Furthermore, the derivation used above for the planar interface z = 0 can easily be generalized for curved surfaces. Therefore we conclude:

The tangential electric and magnetic field components are continuous across any interface.

By integrating Maxwell's equations that contain the div-operator (1.3.14), (1.3.15) over a pill box with height dz and top and bottom surfaces on either side and parallel to the interface, and considering the limit dz → 0, we find continuity relations for the normal components of the fields:

$$lim_{z\uparrow 0}E_i\hat{z}\varepsilon_x(x,y,z,t)=lim_{z\downarrow 0}E_t\hat{z}\varepsilon x(x,y,z,t),$$
$$lim_{z\uparrow 0}H_x(x,y,z,t)=lim_{z\downarrow 0}H_x(x,y,z,t),$$

The normal components of E**ε** and **H** are continuous across an interface.

Remark

Since all derived boundary conditions hold for all times t, it follows that for timeharmonic fields they also hold for the complex fields. Hence (1.10.10) and (1.10.11) hold and similarly we find that the normal components of E**ε** and **H** are continuous.

1.9.2 Snell's Law

By substituting the complex electric fields derived from (1.10.1), (1.10.4) and (1.10.7) into equation (1.10.10), we get

$$\hat{z} \times [A^i e^{i(k_x^i \cdot x + k_y^i \cdot y)} + A^r e^{i(k_x^r \cdot x + k_y^r \cdot y)}] = \hat{z} \times A^t e^{i(k_x^t \cdot x + k_y^t \cdot y)},$$

Since this equation must be satisfied for all points (x, y), it follows that

$$k_x^i = k_x^r = k_x^t,$$

$$k_y^i = k_y^r = k_y^t,$$

Hence, the tangential components of the wave vectors of the incident, reflected and transmitted waves are identical. In fact, if (1.10.19) would not hold, then by keeping y fixed, the exponential functions in (1.10.18) would not all have the same periodicity as functions of x and then (1.10.18) could never be satisfied for all x. The same argument with x kept fixed leads to the conclusion (1.10.20).

Without restricting the generality, we will from now on assume that the coordinate system is chosen such that

$$k_y^i = k_y^r = k_y^t = 0.$$

The plane through the incident wave vector and the normal to the interface is called the plane of incidence. Hence in the case of (1.10.21) the plane of incidence is the (x, z)-plane.

Since the length of the wave vectors ki and kr is $k_0 n_i$, with k_0 the wave number in vacuum and $n_i = (E_i/E_0)^{1/2}$ the refractive index, and since the length of kt is $k_0 n_t$, with $n_t = (E_t/E_0)^{1/2}$, it follows from (1.10.19)

$$sin\theta_i = \frac{k_x^i}{k_0 n_i} = \frac{k_x^r}{k_0 n_r} = sin\theta_r,$$

and

$$n_i sin\theta_i = \frac{k_x^i}{k_0} = \frac{k_x^t}{k_0} = n_t sin\theta_t,$$

where the angles are as in Figure 1.10.2 Hence,

$$\theta_i = \theta_r, angle\ of\ reflection = angle\ of\ incidence,$$

$$n_i sin\theta_i = n_t sin\theta_t, Snell's\ Law.$$

Snell's Law implies that when the angle of incidence θ_i increases, the angle of transmission increases as well. If the medium in z < 0 is air with refractive index $n_i = 1$ and the other medium is glass with refractive index $n_t = 1.5$, then the maximum angle of transmission occurs when $\theta_i = 90^o$ with

$$\theta_{t,max} = arcsin(n_i/n_t) = 41.8^o.$$

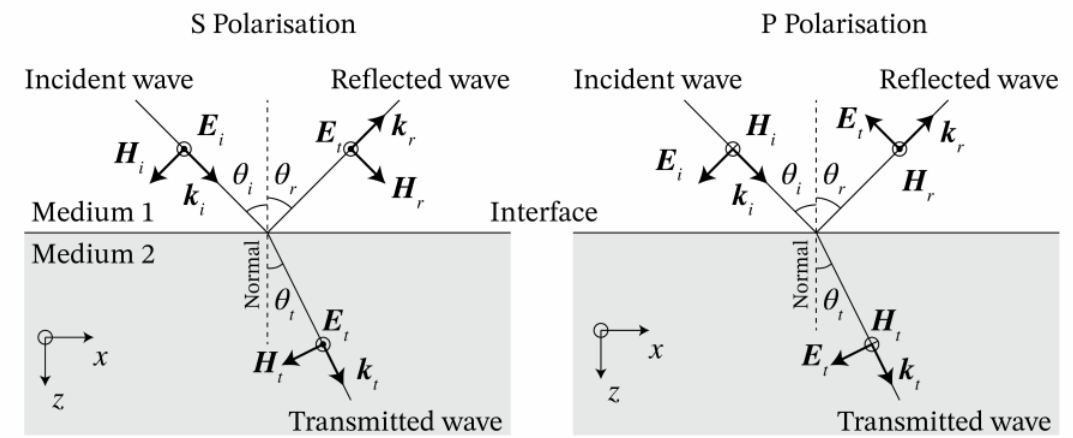


Figure 1.10.2: The incident, reflected, and transmitted wave vectors with the electric and magnetic vectors for s- and p-polarisation. For s-polarisation the electric field points out of the plane at the instant shown while for p-polarisation the magnetic field points out of the plane at the instant shown.

In case the light is incident from glass, i.e. $n_i = 1.5$ and $n_t = 1.0$, the angle of incidence θ_i cannot be larger than 41.8^o because otherwise there is no real solution for θ_t . It turns out that when $\theta_i > 41.8^o$, the wave is totally reflected and there is no propagating transmitted wave in air. As explained in Section 1.9.5, this does however not mean that there is no field in $z > 0$. In fact there is a non-propagating so-called evanescent wave in $z > 0$. The angle $\theta_{i,crit} = 41.8^o$ is called the **critical angle of total internal reflection**. It exists only if a wave is incident from a medium with larger refractive index on a medium with lower refractive index ($n_t < n_i$). The critical angle is independent of the polarisation of the incident wave.

1.9.3 Fresnel Coefficients

Because of (1.10.19) and (1.10.21), we write $k_x = k^i_x = k^r_x = k^t_x$ and therefore $k^i_z = (k^2_0 E_i - k^2_x)^{1/2} = -k^r_z$ and $k^t_z = (k^2_0 E_t - k^2_x)^{1/2}$. Hence,

$$k^i = k_x\hat{x} + k^i_z\hat{z}, k^r = k_x\hat{x} - k^i_z\hat{z},$$

and

$$k^t = k_x\hat{x} - k^t_z\hat{z},$$

According to (1.6.11), for the incident, reflected and transmitted plane waves there must hold:

$$A^i \cdot k^i = A^r \cdot k^r = A^t \cdot k^t = 0.$$

We choose an orthonormal basis perpendicular to k^i with unit vectors:

$$\hat{s} = \hat{y}, \hat{p}^i = \frac{1}{|k^i|}(-k^i_z\hat{x} + k_x\hat{z}),$$

where

$$|k^i| = (k^i \cdot (k^i)^*)^{1/2} = (k^2_x \cdot |k^i_z|^2)^{1/2},$$

and where in writing the complex conjugate we anticipate the case the k^i_z is complex, which may happen for example when E_i is complex (a case that has been excluded so far but which later will be considered) or in the case of evanescent waves discussed in Section 1.9.5. Note that when k^i_z is real, $|k_i| = (k^2_x + (k^i_z)^2)^{1/2} = k0ni$. It is easy to see that the basis (1.10.30) is orthonormal in the space of two-dimensional complex vectors and that $\mathbf{s} \cdot \mathbf{k}^i = \mathbf{p}^i \cdot \mathbf{k}^i = 0$. The vector $\mathbf{s}$ is perpendicular to the plane of incidence, therefore the electric field component in this direction is polarised perpendicular to the plane of incidence and is called s-polarised ("Senkrecht" in German). The other basis vector $\mathbf{p}^i$ is (for real $\mathbf{k}^i$) parallel to the plane of incidence and when the electric component in this direction is called p-polarised. The complex vector $\mathbf{A}^i$ can be expanded on this basis:

$$A^i = A^i_s\hat{y} + A^i_p\hat{p}^i.$$

Since

$$k^i \times \hat{y} = |k^i|\hat{p}^i, k^i \times \hat{p}^i = -\frac{k_0^2 E_i}{|k^i|}\hat{y},$$

it follows that the electric and magnetic field of the incident plane wave can be written as

$$E^i(r) = (A_s^i \hat{y} + A_p^i \hat{p}^i)e^{ik^i \cdot r},$$

$$H^i(r) = (\frac{|k^i|}{\omega\mu_0} A_s^i \hat{p}^i - \frac{\omega E_0 E_i}{|k^i|} A_p^i \hat{y})e^{ik^i \cdot r}.$$

The reflected field is expanded on the basis **y** and $\mathbf{p}^r$ with

$$\hat{p}^r = -\frac{1}{|k^i|}(k_z^i \hat{x} + k_x \hat{z}).$$

The sign in front of the unit vector $\mathbf{p}^r$ is chosen such that that its x-component is the same as that of $\mathbf{p}^i$. Since

$$k^r \times \hat{y} = -|k^i|\hat{p}^r, k^r \times \hat{p}^r = \frac{k_0^2 E_i}{|k^i|}\hat{y},$$

it follows that

$$E^r(r) = (A_s^r \hat{y} + A_p^r \hat{p}^r)e^{ik^r \cdot r},$$

$$H^r(r) = (-\frac{|k^i|}{\omega\mu_0} A_s^r \hat{p}^r + \frac{\omega E_0 E_i}{|k^i|} A_p^r \hat{y})e^{ik^r \cdot r},$$

where we used that $k^r \cdot k^r = k_0^2 n_i^2$ and $|k^r| = (k_x^2 + |k_z^r|^2)^{1/2} = (k_x^2 + |k_z^i|^2)^{1/2} = |k^i|$. For the transmitted plane wave we use the basis **y** and $\mathbf{p}^t$ with

$$\hat{p}^t = \frac{1}{|k^t|}(-k_z^t \hat{x} + k_x \hat{z}),$$

where $\mathbf{p}^t$ is chosen such that the x-component of $\mathbf{p}^t$ has the same sign as the x-component of $\mathbf{p}^i$. Since

$$k^t \times \hat{y} = |k^t|\hat{p}^t, k^t \times \hat{p}^t = \frac{k_0^2 E_t}{|k^t|}\hat{y},$$

we get

$$E^t(r) = (A_s^t \hat{y} + A_p^t \hat{p}^t)e^{ik^t \cdot r},$$

$$H^t(r) = (\frac{|k^t|}{\omega\mu_0} A_s^t \hat{p}^t - \frac{\omega E_0 E_i}{|k^t|} A_p^t \hat{y})e^{ik^t \cdot r},$$

We now consider an s-polarised incident plane wave, i.e. $A_p^i = 0$. We will show that all boundary conditions can be satisfied by $A_p^r = A_p^t = 0$ and by appropriately expressing A_s^r and A_s^t in terms of A_s^i . This implies that if the incident plane wave is s-polarised, the reflected and transmitted waves are s-polarised as well. For s-polarisation, the electric field has only a y-component and this component is tangential to the interface z = 0. This leads to the condition

$$A_s^i + A_s^r = A_s^t.$$

The only tangential component of the magnetic field is the x-component and requiring it to be continuous for z = 0 leads to

$$-k_z^i A_s^i + k_z^i A_s^r = -k_z^t A_s^t.$$

Solving (1.10.44), (1.10.45) for A_s^r and A_s^t gives the following formula for the reflection and transmission coefficients:

$$r_s = \frac{A_s^r}{A_s^i} = \frac{k_z^i - k_z^t}{k_z^i + A_z^t},$$

$$t_s = \frac{A_s^t}{A_s^i} = \frac{2k_z^i}{k_z^i + A_z^t}.$$

Only the magnetic field has a z-component and it easy to verify that $H_z^i + H_z^r = H_z$ for z = 0.

By looking at the case of a p-polarised incident wave: $A_s^i = 0$, we see that the expression for the magnetic field in the p-polarised case become similar (except for the chosen signs) to that of the electric field for s-polarisation and conversely. Enforcing the continuity of the tangential components at z = 0 gives for p-polarisation:

$$r_p = \frac{A_p^r}{A_p^i} = -\frac{\frac{k_p^i}{E_i} - \frac{k_p^t}{E_t}}{\frac{k_p^i}{E_i} + \frac{k_p^t}{E_t}},$$

$$t_p = \frac{A_p^t}{A_p^i} = \frac{2\frac{k_p^i}{E_i}}{\frac{k_p^i}{E_i} + \frac{k_p^t}{E_t}}.$$

It is easy to verify that E_z is the only normal component and that $e_i(E_z^i + E_z^r) = e_t E_z^t$ for z = 0.

The reflection and transmission coefficients r_s, r_p, t_s and tp are called **Fresnel coefficients**. As follows from the derivation, there is no cross talk between s- and p-polarised plane waves incident on a planar interface. A generally polarised incident plane wave can always be written as a linear combination of s- and a p-polarised incident plane waves. Because in general $r_s \neq r_p$ and $t_s \neq t_p$, it follows that the reflected and transmitted fields are also linear combinations of s- and p-polarised fields, but with different coefficients (weights) of these two fundamental polarisation states than for the incident wave.

Remarks.

1. In the derivation of the Fresnel coefficients the continuity of the normal field components was not used and was automatically satisfied. The reason is that the electromagnetic fields of the plane waves where chosen to be perpendicular to the wave vectors. This implies that the divergence of E**ε** and of **H** vanishes which in turns implies that the normal components are automatically continuous across the the interface.

2. When kiz and ktz are both real, we have $|k^i| = k_0 n_i$ and $|k^t| = k_0 n_t$ and the Fresnel coefficients can be expressed in the angles θ_i, θ_r and θ_t and the refractive indices $n_i = E_i^{1/2}/E_0$ and $n_t = (E_t/E_0)^{1/2}$. Because $k_z^i = k_0 n_i \cos\theta_i$ and $k_z^t = k_0 n_t \cos\theta_t$, we find

$$r_s = \frac{n_i cos\theta_i - n_t cos\theta_t}{n_i cos\theta_i + n_t cos\theta_t} = -\frac{sin(\theta_i - \theta_t)}{sin(\theta_i + \theta_t)},$$

$$t_s = \frac{2n_i cos\theta_i}{n_i cos\theta_i + n_t cos\theta_t} = \frac{2cos\theta_i sin\theta_t}{sin(\theta_i + \theta_t)},$$

and

$$r_p = -\frac{\frac{cos\theta_i}{n_i} - \frac{cos\theta_t}{n_t}}{\frac{cos\theta_i}{n_i} - \frac{cos\theta_t}{n_t}} = -\frac{tan(\theta_i - \theta_t)}{tan(\theta_i + \theta_t)},$$

$$t_p = \frac{\frac{2cos\theta_i}{n_i}}{\frac{cos\theta_i}{n_i} + \frac{cos\theta_t}{n_t}} = -\frac{2cos\theta_i sin\theta_t}{tan(\theta_i + \theta_t)}.$$

To obtain the expressions at the far right in (1.10.50), (1.10.51), (1.10.52) and (1.10.53) Snell's Law has been used.

3. The advantage of the expressions (1.10.46), (1.10.47), (1.10.48), (1.10.49) in terms of the wave vector components kiz and ktz is, that they also apply when k_z^i and/or k_z^t are complex. This happens for example when there is absorption in z < 0 or in z > 0 or both, or when $E_i > E_t$ and the incident angle is above the critical angle, due to which k_z^t is purely imaginary (see Section 1.9.5).

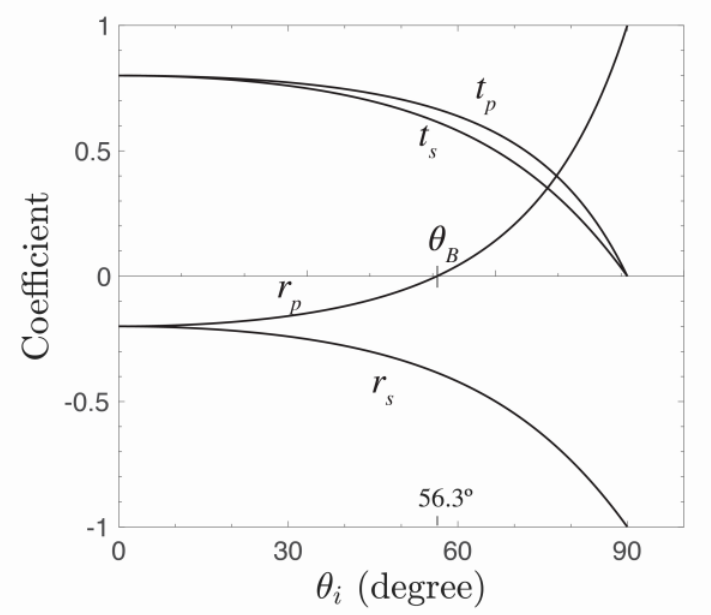


Figure 1.10.3: Reflection and transmission coefficients as function of the angle of incidence of s- and p-polarised waves incident from air to glass. The Brewster angle θ_B is indicated.

1.9.4 Properties of the Fresnel Coefficients

For normal incidence: $\theta_i = 0$, Snell's Law implies: $\theta_t = 0$. Hence, (1.10.50), (1.10.52) give:

$$r_s(\theta_i = 0) = r_p(\theta_i = 0) = \frac{n_i - n_t}{n_i + n_t},$$

Is is seen that for normal incidence $r_p = r_s$, as expected. Note however that if we would not have defined $\mathbf{p}^r$ such that its tangential component is the same as that of $\mathbf{p}^i$, the two reflection coefficients would have the opposite sign for normal incidence (as is the case in some books). If the incident medium is air and the other medium is glass ($n_i = 1.0$, $n_t = 1.5$), we get

$$r_s(\theta_i = 0) = r_p(\theta_i = 0) = -0.2,$$

and since the flow of energy is proportional to the square of the field, it follows that 4% of normal incident light is reflected by the glass. Hence a lens of glass without anti-reflection coating reflects approximately 4% of the light at normal incidence. The transmission coefficient for normal incidence is:

$$t_s(\theta_i = 0) = r_p(\theta_i = 0) = \frac{2n_i}{n_i + n_t},$$

which for air-glass becomes 0.8.

Remark. Energy conservation requires that the normal component < Sz > of the time-averaged energy flux through the interface is continuous. By using the formula for the time-averaged Poynting vector of a plane wave (1.8.8), it can be verified that the Fresnel coefficients are such that the energy flux is indeed continuous.

It follows from Snell's Law (1.10.25) that $\sin\theta_t = (n_i/n_t)\sin\theta_i$. Hence θ_t monotonically increases with θ_i and therefore there exists some θ_i such that

$$\theta_i + \theta_t = 90^o.$$

For this particular angle of incidence, the denominator of (1.10.52) is infinite and hence $r_p = 0$, i.e. the p-polarised wave is not reflected at all. This angle of incidence is called the **Brewster angle** θ_B . It is easy to see from (1.10.50) that the reflection is never zero for s-polarisation.

If unpolarised light is incident at the Brewster angle, the reflected light will be purely s-polarised.

Since at the Brewster angle s-polarised light is only partially reflected and the rest is transmitted, the transmitted light at the Brewster angle is a mixture of s- and p-polarisation. We have $\theta_t = 90^o - \theta_i$, hence $\sin\theta_t = \cos\theta_i$ and by Snell's Law (writing θ_i =

θ_B):

$$tan(\theta_B) = \frac{n_t}{n_i}.$$

We see that there is always a solution, independent of whether the wave is incident from the material with the smallest or largest refractive index. For the air-glass interface we have $\theta_B = 56.3^o$ and $\theta_t = 33.7^o$. By (1.10.50):

$$r_s(\theta_B = 56.3^o) = -0.38$$

so that $(0.38)^2/2 = 0.07$, or 7 % of the unpolarised light is reflected as purely s-polarised light at the air glass interface. For a wave incident from glass, $\theta_B = 33.7^o$.

In Figure 1.10.3 the reflection and transmission coefficients of s- and p-polarised waves are shown as a function of the angle of incidence for the case of incidence from air to glass. There is no critical angle of total reflection in this case. The Brewster angle is indicated. It is seen that the reflection coefficients decrease from the values −0.2 for $\theta_i = 0^o$ to -1 for $\theta_i = 90^o$. The transmission coefficients monotonically decrease to 0 at $\theta_i = 90^o$.

Figure 1.10.3 shows the Fresnel coefficients when the wave is incident from glass to air. The critical angle is $\theta_{i,crit} = 41.8^o$ as derived earlier. At the angle of total internal reflection the absolute values of the reflection coefficients are identical to 1. There is again an angle where the reflection of p-polarised light is zero $\theta_B = 33.7^o$.

Depending on the refractive indices and the angle of incidence, the reflection coefficients can be negative. The reflected electric field then has an additional π phase shift compared to the incident wave. In contrast, (provided that the materials are lossless), the transmitted field is always in phase with the incident field, i.e. the transmission coefficients are always positive.

1.9.5 Total Internal Reflection and Evanescent Waves

We return to the case of a wave incident from glass to air, i.e. $n_i = 1.5$ and $n_t = 1$. As has been explained, there is then a critical angle, given by $\sin\theta_{i,crit}=n_t/n_i$.

This is equivalent to

$$k_x^t = k_0 n_i sin\theta_{i,crit} = k_0 n_t$$

The wave vector $k^t = k_x^t\mathbf{x} + k_z^t\mathbf{z}$ in z > 0 satisfies:

$$(k_x^t)^2 + (k_z^t)^2 = k_0^2 n_t^2.$$

Because of (1.10.60), we have at the critical angle

$$k_z^t = 0.$$

For angles of incidence above the critical angle we have: $k_x^t > k_0 n_t$ and it follows from (1.10.61) that $(k_z^t)^2 = k_0^2 n_t^2 - (k_x^t)^2 < 0$, hence k_z^t is imaginary:

$$k_z^t = \pm(k_0^2 n_t^2 - (k_x^t)^2)^{1/2} = \pm i((k_x^t)^2 - k_0^2)^{1/2},$$

where the last square root is a positive real number. It can be shown that above the critical angle the reflection coefficients are **complex** numbers with modulus 1: $|r_s| = |r_p| = 1$. This implies that the reflected intensity is identical to the incident intensity, while at the same time the transmission coefficients are not zero! For example, for s-polarisation we have according to (1.10.46), (1.10.47):

$$t_s = 1 + r_s \neq 0,$$

because $r_s \neq -1$ (although $|r_s| = 1$). Therefore there is an electric field in z > 0, given by

$$E(x,z)e^{-i\omega t} = t_s e^{ik_x^t x + ik_z^t z - i\omega t}\,\hat{y} = t_s e^{i(k_x^t x - \omega t)}\, e^{-z((k_x^t)^2 - k_0^2 n_t^2)^{1/2}}\,\hat{y}, z > 0,$$

where we have chosen the + sign in (1.10.63) to prevent the field from blowing up for $z \to \infty$. Since k_x^t is real, the wave propagates in the x-direction. In the z-direction, however, the wave is **not** propagating. Its amplitude decreases exponentially as a function of distance z to the interface and therefore the wave is confined to a thin layer adjacent to the interface. Such a wave is called an evanescent wave. One can compute the Poynting vector of the **evanescent wave** and find that this vector is parallel to the

interface. Hence, **the flow of energy of an evanescent wave propagates parallel to the interface namely in the direction in which k^t_x is positive**

Hence no energy is transported away from the interface into the air region. We shall return to evanescent waves in the chapter on diffraction theory.

External sources in recommended order

1. Youtube video - 8.03 - Lect 18 - Index of Refraction, Reflection, Fresnel Equations, Brewster Angle - Lecture by Walter Lewin
2. MIT OCW - Reflection at The Air-glass Boundary: demonstration of reflection of polarised light and the Brewster angle.

1.11: Fiber Optics

By using the phenomenon of total internal reflection, light can be transported over long distances without reduction of the energy density due to divergence of the beam. The principle has been known for a long time, but the topic was greatly boosted by the invention of the laser.

Consider a straight glass cylinder of refractive index n_i , surrounded by air with refractive index $n_t = 1$. The core of the cylinder has a cross section approximately the size of a human hair and hence, although quite small, it is still many optical wavelengths thick. This implies that when light strikes the cylindrical surface, we can locally consider the cylinder as a flat surface. By focusing a laser beam at the entrance plane of the fiber, light can be coupled into the fiber. The part of the light inside the fiber that strikes the cylinder surface at an angle with the normal that is larger than the critical angle of total reflection will be totally reflected. As it hits the opposite side of the cylinder surface, it will again be totally reflected and so on (Figure 1.11.2left).

Since visible light has such high frequencies (order 10^{15} Hz), roughly a hundred thousand times more information can be carried through a fiber than at microwave frequencies. Today fibers with very low losses are fabricated so that signals can be sent around the earth with hardly any attenuation. Abraham van Heel, professor of optics at Delft University of Technology, showed for the first time in a paper published in Nature in 1954 that by packing thousands of fibers into a cable, images can be transferred, even if the bundle is bent (Figure 1.11.2right).

Fibers can also be used to transmit and combine the light of many Diode laser in order to provide locally large amount of energy This can be used for cutting metals for example.

External sources in recommended order

1. MIT OCW - Single Mode Fiber: Demonstration of a single-mode fiber.
2. MIT OCW - Multi-mode Fiber: Demonstration of a multimode fiber.

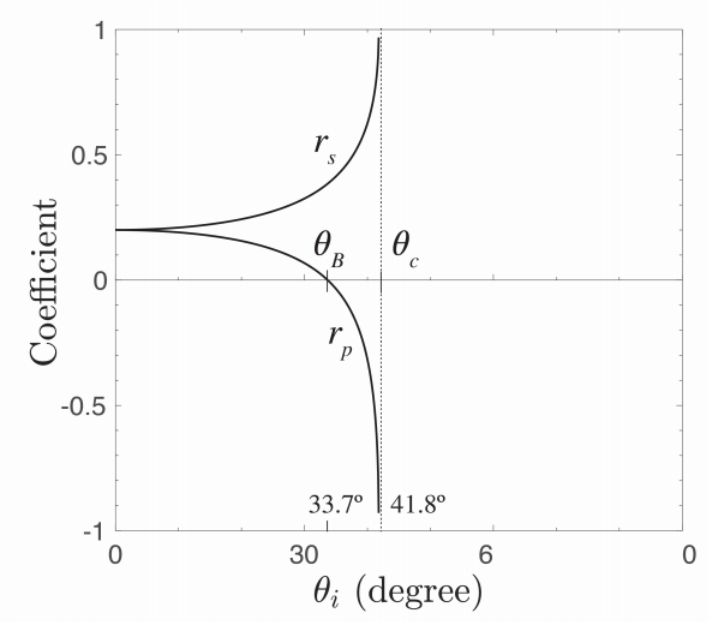


Figure 1.11.1: Reflection and transmission coefficients as function of the angle of incidence of s- and p-polarised waves incident from glass to air. Furthermore, the absorption coefficients are the squares of the coefficients shown.

Figure 1.11.2: Light guided within a piece of glass. (from Wikimedia Commons by Keerthi - CC BY).

CHAPTER OVERVIEW

2: Geometrical Optics

2.1: What You Should Know and be able to do After Studying This Chapter
2.2: Introduction
2.3: Principle of Fermat
2.4: Some Consequences of Fermat's Principle
2.5: Perfect Imaging by Conic Sections
2.6: Gaussian Geometrical Optics
2.7: Beyond Gaussian Geometrical Optics
2.8: Aberrations

2.1: What You Should Know and be able to do After Studying This Chapter

- Principle of Fermat.
- Understand the approximation made in Gaussian geometrical optics.
- Know how to work with the sign convention of the Lens Maker's Formula (not the derivation of the formula).
- Understand how the Lens Maker's Formula of a single lens follows from the formula for a single interface.
- Understand how the image of two and more lenses is derived from that of a single lens by construction and by computing the intermediate images. You do not need to know the imaging equation and the formulae for the focal distances of two thin lenses.
- Understand the matrix method (you do not need to know the matrices by hart).
- Understand the modification of the lens model to incorporate a thick lens.
- Understand the limitations of geometrical optics, in particular when diffraction optics is needed.

Nice software for practicing geometrical optics: https://www.geogebra.org/m/X8RuneVy

2.2: Introduction

Geometrical optics is an old subject but it is still very essential to understand and design optical instruments such as camera's, microscopes, telescopes etc.. Geometrical optics started long before light was described as a wave as is done in wave optics, and long before it was discovered that light is an electromagnetic wave and that optics is included in electromagnetism.

In this chapter we go back in history and treat geometrical optics. That may seem strange now that we have a much more accurate and better theory at our disposal. However, the predictions of geometrical optics are under quite common circumstances very useful and also very accurate. In fact, for many optical systems and practical instruments there is no good alternative for geometrical optics because more accurate theories are much too complicated to use.

When a material is illuminated, the molecules start to radiate spherical waves (more precisely, they radiate like tiny electric dipoles) and the total wave scattered by the material is the sum of all these spherical waves. A time-harmonic wave has at every point in space and at every instant of time a well defined phase. A **wavefront** is a set of space-time points where the phase has the same value. At any fixed time, the wavefront is called a surface of constant phase. This surface moves with the phase velocity in the direction of its local normal. For plane waves we have shown that the surfaces of constant phase are planes and that the normal to these surfaces is in the direction of the wave vector which also coincides with the direction of the phase velocity as well as with the direction of the flow of energy (the direction of the Poynting vector). For more general waves, the local direction of energy flow is given by the direction of the Poynting vector. Provided that curvature of the surfaces much smaller than the wavelength, the normal to the surfaces of constant phase may still be considered to be in the direction of the local flow of energy. Such waves behave locally as plane waves and their effect can be accurately described by the methods of geometrical optics.

Geometrical optics is based on the intuitive idea that light consists of a bunch of rays.

A ray is an oriented curve which is everywhere perpendicular to the surfaces of constant phase and points in the direction of the flow of energy.

Consider a point source at some distance before an opaque screen with an aperture. According to the ray picture, the light distribution on a second screen further away from the source and parallel to the first screen is simply an enlarged copy of the aperture (see Figure 2.2.1). The copy is enlarged due to the fanning out of the rays. However, this description is only accurate when the wavelength of the light is very small compared to the diameter of the aperture. If the aperture is only ten times the wavelength, the pattern is much broader due to the bending of the rays around the edge of the aperture. This phenomenon is called **diffraction**. It can not be explained by geometrical optics and will be studied in Chapter 6. Although it is possible to incorporate polarisation into geometrical optics, this is not standard theory and we will not consider polarisation effects in this chapter.

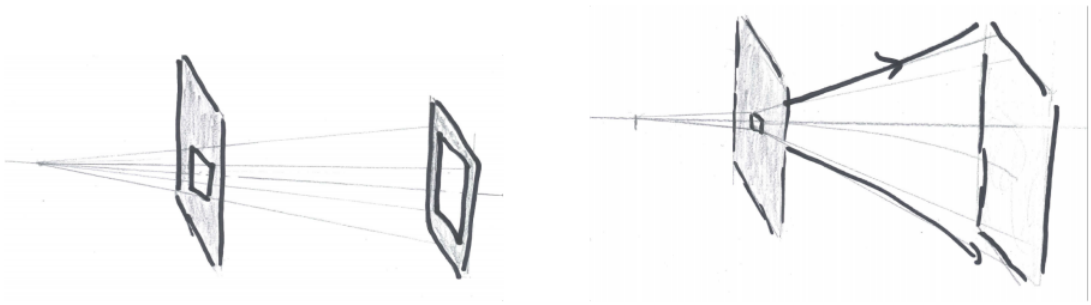


Figure 2.2.1: Light distribution on a screen due to a rectangular aperture. Left: for a large aperture, we get an enlarged copy of the aperture. Right: for an aperture that is of the order of the wavelength there is strong bending (diffraction) of the light.

Geometrical optics is accurate when the size of the objects in the system are large compared to the wavelength. It is possible to derive geometrical optics from Maxwell's equations by expanding the electromagnetic field in a Taylor series in the wavelength and retaining only the first term of this expansion.

2.3: Principle of Fermat

The starting point of the treatment of geometrical optics is the very powerful.

Principle of Fermat (1657)

The path followed by a light ray between two points is the one that takes the least amount of time.

The speed of light in a material with refractive index n, is c/n, where c = 3 × 10^8 m/s is the speed of light in vacuum. At the time of Fermat, the conviction was that the speed of light must be finite, but nobody could suspect how incredibly large it actually is. In 1676 the Danish astronomer Ole Römer computed the speed from inspecting the eclipses of a moon of Jupiter and arrived at an estimate that was only 30% too low.

Let $\mathbf{r}(s)$, be a ray with s the length parameter. The ray links two points S and P. Suppose that the refractive index varies with position: $n(\mathbf{r})$. Over the infinitesimal distance from s to $s+ds$, the speed of the light is

$$\frac{c}{n(\mathbf{r}(s))}.$$

Hence the time it takes for light to go from $\mathbf{r}(s)$ to $\mathbf{r}(s+ds)$ is:

$$dt = \frac{n(\mathbf{r}(s))}{c} ds,$$

and the total total time to go from S to P is:

$$t_{S\to P} = \int_0^{s_P} \frac{n(\mathbf{r}(s))}{c}\, ds,$$

where s_P is the distance along the ray from S to P. The **optical path length** [m] of the ray between S and P is defined by:

$$OPL = \int_0^{s_P} n(r(s))\, ds,$$

So the OPL is the distance weighted by the refractive index.

Fermat's principle is thus equivalent to the statement that a ray follows the path with shortest OPL.

Remark

Actually, Fermat's principle as formulated above is not complete. There are circumstances that a ray can take two paths between two points that have different travel times. Each of these paths then corresponds to a minimum travel time compared to nearby paths, so the travel time is in general a *local minimum*. An example is the reflection by a mirror discussed in the following section.

2.4: Some Consequences of Fermat's Principle

Homogeneous matter

In homogenous matter, the refractive index is constant and therefore paths of shortest OPL are straight lines. Hence in homogeneous matter rays are straight lines.

Inhomogeneous matter

When the refractive index is a function of position such as air with a temperature gradient, the rays bend towards regions of higher refractive index, leading several well-known effects (see Figure 2.4.1).

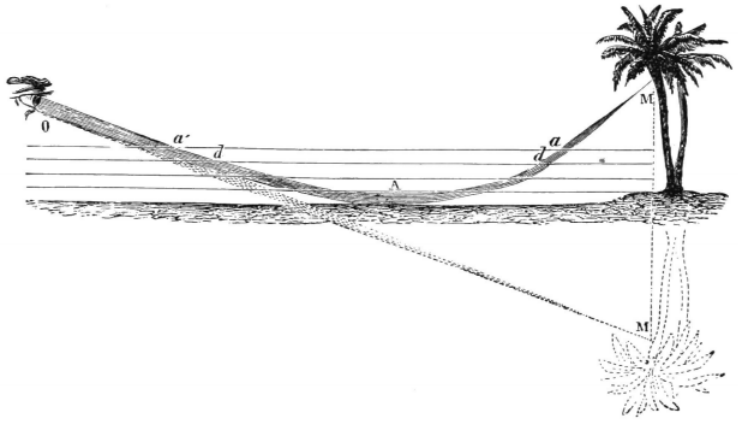


Figure 2.4.1: Because the temperature close to the ground is higher, the refractive index is lower there. Therefore the rays bend upwards, creating a mirror image of the tree below the ground. (From Popular Science Monthly Volume 5, Public Domain).

Law of reflection

Consider the mirror shown in Figure 2.4.2 A ray from point P can end up in Q in two ways: by going straight form P to Q or alternatively via the mirror. Both possibilities have different path lengths and hence different travel times and hence both are local minima mentioned at the end of the previous section. We consider here the path by means of reflection by the mirror. Let the x-axis be the intersection of the mirror and the plane through the points P and Q and perpendicular to the mirror. Let the y-axis be normal to the mirror. Let (x_P , y_P) and (x_Q, y_Q) be the coordinates of P and Q, respectively. If (x, 0) is the point where a ray from P to Q hits the mirror, the travel time of that ray is

$$\frac{n}{c}d_1(x)+\frac{n}{c}d_2(x)=\frac{n}{c}((x-x_P)^2+y_P^2)^{1/2}+\frac{n}{c}((x_Q-x)^2+y_Q^2)^{1/2},$$

where n is the refractive index of the medium in y > 0. According to Fermat's Principle, the point (x, 0) should be such that the travel time is minimum , i.e.

$$\frac{d}{dx}[d_1(x)+d_2(x)]=\frac{x-x_P}{d_1(x)}-\frac{x_Q-x}{d_2(x)}=0.$$

Hence

$$\sin\theta_i=\sin\theta_r,$$

or

$$\theta_r=\theta_i.$$

where θ_i and θ_r are the angles of incidence and reflection as shown in Figure 2.4.2.

Snell's law of refraction

Next we consider refraction at an interface. Let y = 0 be the interface between a medium with refractive index n_i in y > 0 and a medium with refractive index n_t in y < 0. Let (x_P , y_P) and (x_Q, y_Q) with y_P > 0 and y_Q < 0 be the coordinates of two points P and

Q are shown in Figure 2.4.3.

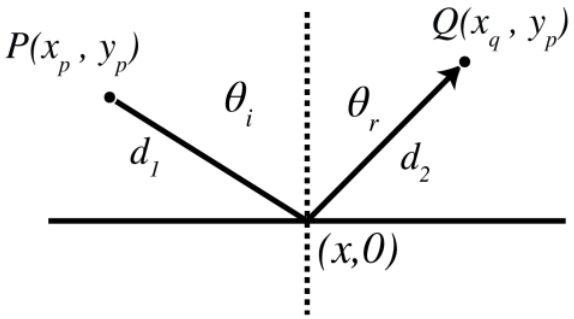


Figure 2.4.2: Ray from P to Q via the mirror.

What path will a ray follow that goes from P to Q? Since the refractive index is constant in both half spaces, the ray is a straight line in both media. Let (x, 0) be the coordinate of the intersection point of the ray with the interface. Then the travel time is

$$\frac{n_i}{c}d_1(x)+\frac{n_t}{c}d_2(x)=\frac{n_i}{c}((x-x_P)^2+y_P^2)^{1/2}+\frac{n_t}{c}((x_Q-x)^2+y_Q^2)^{1/2}.$$

The travel time must be minimum, hence there must hold

$$\frac{d}{dx}[n_i d_1(x)+n_t d_2(x)]=n_i\frac{x-x_P}{d_1(x)}-n_t\frac{x_Q-x}{d_2(x)}=0.$$

where the travel time has been multiplied by the speed of light in vacuum. Eq. (2.4.6) implies

$$n_i\sin\theta_i=n_t\sin\theta_t,$$

where θ_i and θ_t are the angles between the ray and the normal to the surface in the upper half space and the lower half space, respectively (2.4.3).

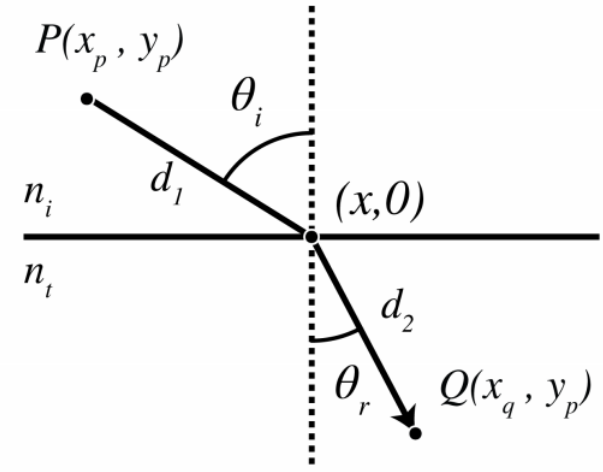


Figure 2.4.3: Ray from P to Q, refracted by an interface.

Hence we have derived the law of reflection and Snell's law from Fermat's principle. In Chapter 1 the reflection law and Snell's law have been derived by a different method, namely from the continuity conditions for the electromagnetic field components at the interface.

2.5: Perfect Imaging by Conic Sections

In this section the conic sections ellipse, hyperbole and parabola are important. In Figure 2.5.1 their definitions are shown as a quick reminder.

We start with explaining what in geometrical optics is meant by **perfect imaging**. Let S be a point source. The rays perpendicular to the spherical wave fronts emitted by S radially fan out from S. Due to objects such as lenses etc. the spherical wave fronts are deformed and the direction of the rays start to deviate from the radial propagation direction. When there is a cone of rays coming from point S and all rays in that cone intersect in the same point P, then by Fermat's principle, all these rays have traversed paths of minimum travel time. In particular, their travel times are equal and therefore they **all add up in phase** when they arrive in P. Hence at P there is a high light intensity. If there is a cone of rays from point S which all intersect in a point P, point P is called the **perfect image** of S. By reversing the direction of the rays, S is similarly a perfect image of P. The optical system in which this happens is called **stigmatic for the two points S and P**.

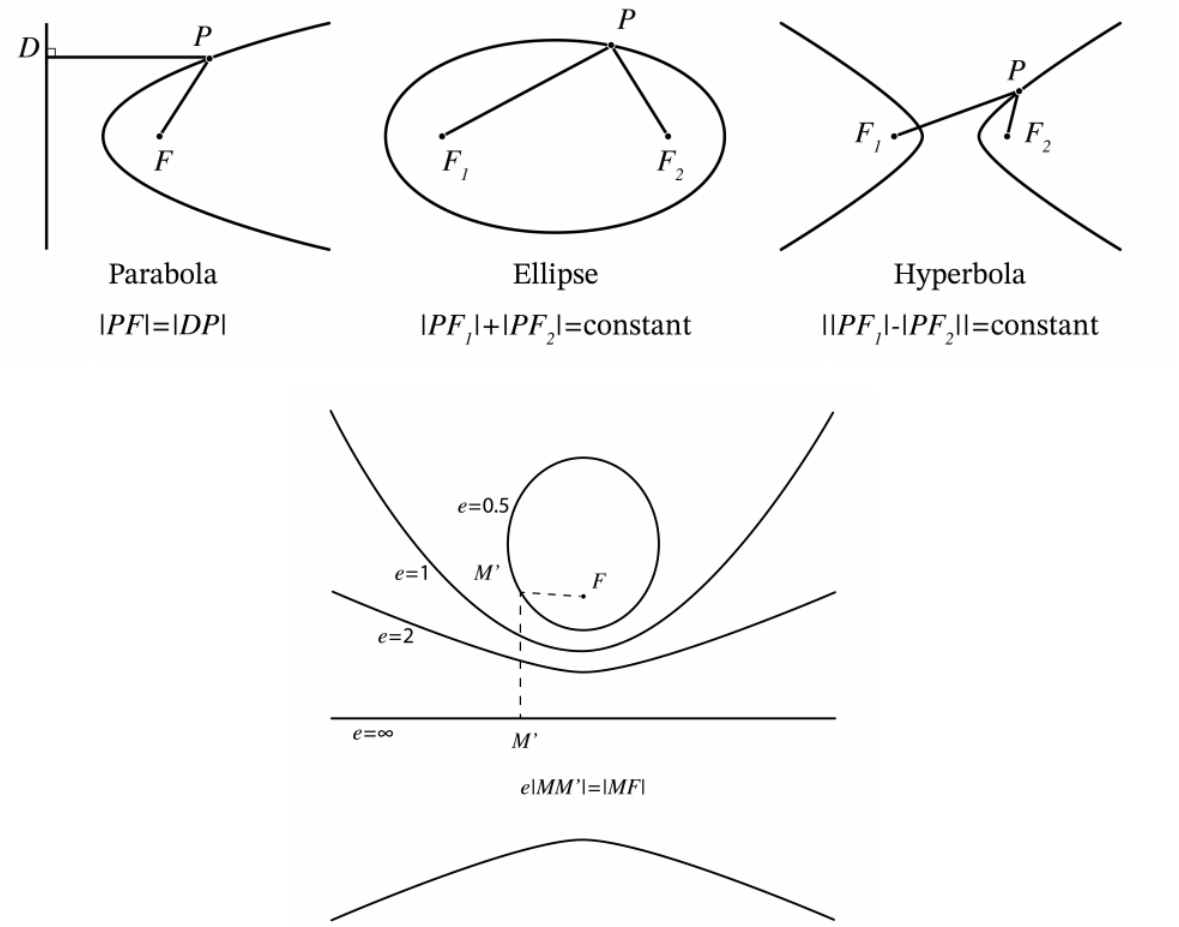


Figure 2.5.1: Overview of conic sections. The lower figure shows a definition that unifies the three definitions in the figure above by introducing a parameter called the eccentricity e. The point F is the focus and the line e = ∞ is the directrix of the conic sections.

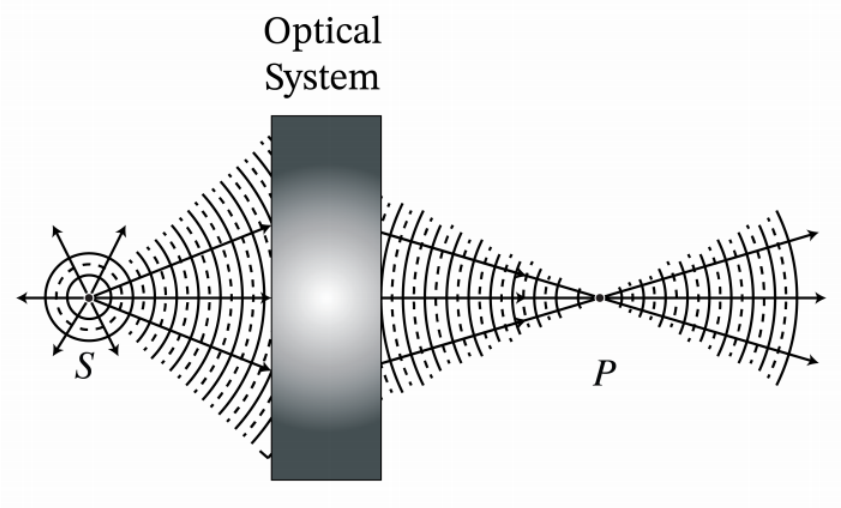


Figure 2.5.2: Perfect imaging: a cone of rays which diverge from S and all intersect in point P. The rays continue after P.

Remark. The concept of a perfect image point exists only in geometrical optics. In reality finite apertures of lenses and other imaging systems cause diffraction due to which image points are never perfect but blurred.

We summarise the main examples of stigmatic systems.

1. **Perfect focusing and imaging by refraction.** A parallel bundle of rays propagating in a medium with refractive index n_2 can be focused into a point F in a medium n_1. If $n_2 > n_1$, the interface between the media should be a hyperbole with focus F, whereas if $n_2 < n_1$ the interface should be an ellipse with focus F (see Figs. 2.5.3) and 2.5.4. By reversing the rays we obtain perfect collimation. Therefore, a point S in air can be perfectly imaged onto a point F in air by inserting a piece of glass in between them with hyperbolic surfaces as shown in Figure 2.5.4.
2. **Perfect focusing of parallel rays by a mirror.** A bundle of parallel rays in air can be focused into a point F by a mirror of parabolic shape with F as focus (see Figure 2.5.5. By reversing the arrows, we get (within geometrical optics) a perfectly parallel beam. Parabolic mirrors are used everywhere, from automobile headlights to radio telescopes.

The remainder of this section can be skipped since the following demonstrations that the ellipse, hyperbole and parabole are stigmatic for certain points are not essential for the sequel. First suppose that there are two media with refractive indices $n_1 > n_2$ and that point S is at infinity in the medium with refractive index n_2. We will construct a surface (interface) between the two media such that all rays from S are focused into the same point F (see Figure 2.5.3a). Because S is at very large distance, the rays entering from the right are parallel. Since all parallel rays have travelled the same distance when they hit the surface DD' perpendicular to the rays, all parallel rays have the same phase at their intersection points with the plane DD' . If point A is on the interface sought for, the travel time for a ray from D to F via A must be minimum and hence must be the same for all points A. Hence

$$\frac{n_2}{c}|DA| + \frac{n_1}{c}|AF| = constant,$$

where “constant” means the same value for all rays, hence for all points A on the interface. By moving the plane DD' parallel to itself, we can achieve that for this new plane DD' we get:

$$e|DA| + |AF| = 0,$$

where $e = n_2/n_1 < 1$. Hence the set of points A defines an **ellipsoid**.

In contrast, when $n_2 < n_1$, as shown at the right of Figure 2.5.3, then $e < 1$ and we have an ellipsoid with F as one of its focal points.

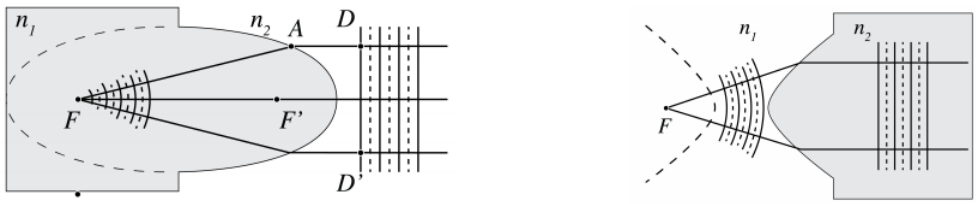


Figure 2.5.3: (a) Ellipsoid ($n_2 < n_1$) and (b) hyperboloid to perfectly focus a parallel beam incident from the medium with refractive index n_2 into a point in a medium with refractive index n_1.

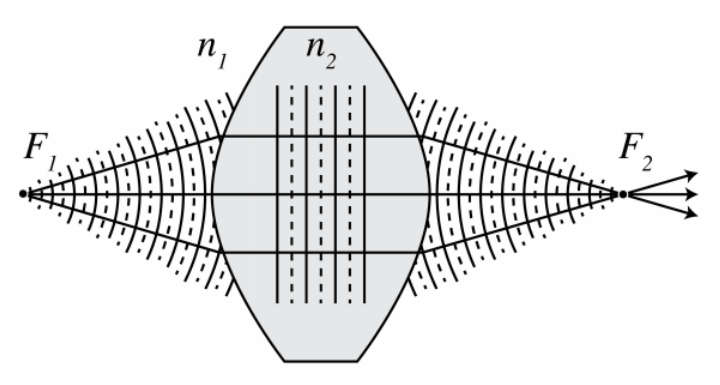


Figure 2.5.4: Lens with hyperboloid surfaces for perfect imaging of a pair of points.

The direction of the rays in Figure 2.5.3 can obviously also be reversed in which case the rays from point F are all perfectly collimated (i.e. parallel). If medium 2 consists of glass and medium 1 of air, we conclude that by joining two hyperboloids as shown in Figure 2.5.4, point S in air is perfectly imaged to point P, also in air.

Next we consider perfect focusing of parallel rays by a mirror. Let there be a parallel bundle of rays in air (n = 1) and suppose we want to focus all rays in point F. We draw a plane Σ_1 perpendicular to the rays as shown in Figure 2.5.5. The rays that hit Σ_1 have traversed the same optical path length. We draw a second surface Σ_2 parallel to Σ_1. Consider rays hitting the mirror in A_1 and A_2. The OPL from W_j via A_j to F must be the same for all rays:

$$OPL = |W_1A_1| + |A_1F| = |W_2A_2| + |A_2F|.$$

Since Σ_2 is parallel to Σ_1:

$$|W_1A_1| + |A_1D_1| = |W_2A_2| + |A_2D_2|$$

Hence (2.5.3) will be satisfied for points A for which |AF| = |AD|, i.e. for which the distance to F is the same as to Σ_2. This is a paraboloid with F as focus and Σ_2 as directrix.

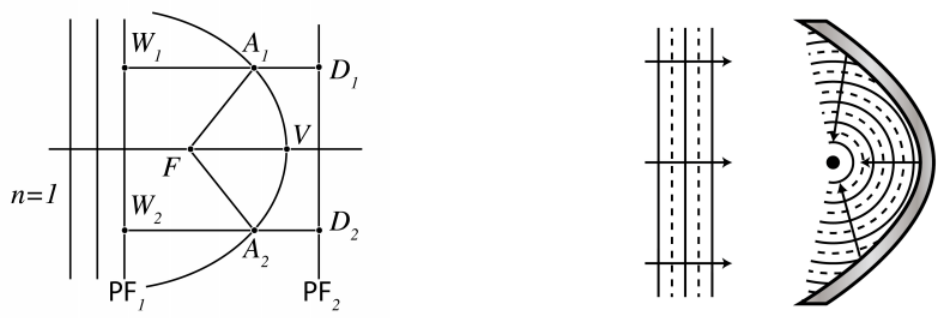


Figure 2.5.5: A paraboloid mirror.

Remark.

Although we found that conic surfaces give perfect imaging for a certain pair of points, other points do *not* have perfect images in the sense that for a certain cone of rays, all rays are refracted (or reflected) to the same point.

External sources in recommended order

- KhanAcademy - Geometrical Optics: Playlist on elementary geometrical optics.
- Yale Courses - 16. Ray or Geometrical Optics I - Lecture by Ramamurti Shankar
- Yale Courses - 17. Ray or Geometrical Optics II - Lecture by Ramamurti Shankar

2.6: Gaussian Geometrical Optics

We have seen above that by using lenses or mirrors which have surfaces that are conic sections we can perfectly image a certain pair of points, but for other points the image is in general not perfect. The imperfections are caused by rays that make larger angles with the **optical axis**, i.e. with the symmetry axis of the system. Rays for which these angles are small are called paraxial rays. Because for **paraxial rays** the angles of incidence and transmission at the surfaces of the lenses are small, the sine of the angles in Snell's Law are replaced by the angles themselves:

$$n_i\theta_i = n_t\theta_t. \qquad \text{(paraxial space rays space only)}$$

This approximation greatly simplifies the calculations. When only paraxial rays are considered, one may replace any refracting surface by a sphere with the same curvature at its vertex. For paraxial rays, errors caused by replacing the general surface by a sphere are of second order and hence insignificant. Spherical surfaces are not only more simple in the derivations but they are also much easier to manufacture. Hence in the optical industry spherical surfaces are used a lot. To reduce imaging errors caused by non-paraxial rays one applies two strategies:

1. adding more spherical surfaces
2. replacing one of the spherical surfaces (typically the last before image space) by a non-sphere.

In Gaussian geometrical optics only paraxial rays and spherical surfaces are considered. In Gaussian geometrical optics every point has a perfect image.

2.5.1 Gaussian Imaging by a Single Spherical Surface

We will first show that within Gaussian optics a single spherical surface between two media with refractive indices $n_1 < n_2$ images all points perfectly (Figure 2.6.1). The sphere has radius R and centre C which is inside medium 2. We consider a point object S to the left of the surface. We draw a ray from S perpendicular to the surface. The point of intersection is V. Since for this ray the angle of incidence with the local normal on the surface vanishes, the ray continues into the second medium without refraction and passes through the centre C of the sphere. Next we draw a ray that hits the spherical surface in some point A and draw the refracted ray in medium 2 using Snell's law in the paraxial form (2.6.1) (note that the angles of incidence and transmission must be measured with respect to the local normal at A, i.e. with respect to CA). We assume that this ray intersects the first ray in point P. We will show that within the approximation of Gaussian geometrical optics, all rays from S pass through P. Furthermore, with respect to a coordinate system (y, z) with origin at V , the z-axis pointing from V to C and the y-axis positive upwards as shown in Figure 2.6.1, we have:

$$-\frac{n_1}{s_0} + \frac{n_2}{s_i} = \frac{n_2 - n_1}{R}.$$

where s_0 and s_i are the z-**coordinates** of S and P, respectively (hence $s_0 < 0$ and $s_i > 0$ in Figure 2.6.1).

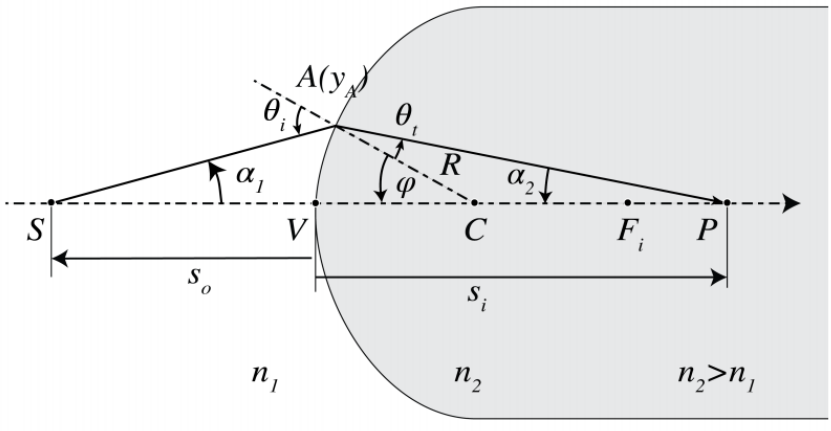


Figure 2.6.1: Imaging by a spherical interface between two media with refractive indices $n_2 > n_1$.

Proof.

(Note: the proof is not part of the exam). It suffices to show that P is independent of the ray, i.e. of A. We will do this by expressing s_i into s_o and showing that the result is independent of A. Let α_1 and α_2 be the angles of the rays SA and AP with the z-axis as shown in Figure 2.6.1. Let θ_i be the angle of incidence of ray SA with the local normal CA on the surface and θ_t be the angle of refraction. By considering the angles in Δ SCA we find

$$\theta_i = \alpha_1 + \phi.$$

Similarly, from Δ CPA we find

$$\theta_t = -\alpha_2 + \phi.$$

By substitution into the paraxial version of Snell's Law (2.6.1), we obtain

$$n_1\alpha_1 + n_2\alpha_2 = (n_2 - n_1)\phi.$$

Let y_A and z_A be the coordinates of point A. Since $s_o < 0$ and $s_i > 0$ we have

$$\alpha_1 \approx tan(\alpha_1) = \frac{y_A}{z_A - s_o}, \alpha_2 \approx tan(\alpha_2) = \frac{y_A}{s_i - z_A}.$$

Furthermore,

$$\phi \approx \sin\phi \approx \frac{y_A}{R}.$$

which is small for paraxial rays. Hence,

$$z_A = R - R\cos\phi = R - R(1 - \frac{\phi^2}{2}) = \frac{R}{2}\phi^2 \approx 0,$$

because it is second order in y_A and therefore is neglected in the paraxial approximation. Then, (2.6.6 becomes

$$\alpha_1 = -\frac{y_A}{s_o}, \alpha_2 = \frac{y_A}{s_i}.$$

By substituting (2.6.9 and (2.6.7 into (2.6.5 we find $-(n_1/s_o)y_A + (n_2/z_i)y_A = [(n_2 - n_1)/R]y_A$, or $-n_1/s_o + n_2/z_i = (n_2 - n_1)/R$, which is Eq. (2.6.2. It implies that s_i, and hence P, is independent of y_A, i.e. of the ray chosen. Therefore, P is a perfect image within the approximation of Gaussian geometrical optics.

When $s_i \to \infty$, the ray after refraction is parallel to the z-axis and we get $s_o \to -n_1R/(n_2 - n_1)$. The object point for which the rays in the medium 2 are parallel to the z-axis is called the **first focal point** or **object focal point** F_o, Its z-coordinate is:

$$f_o = -\frac{n_1 R}{n_2 - n_1}.$$

In spite of the fact that this is a coordinate and hence has a sign, it is also called the **front focal length** or **object focal length**.

When $s_o \to -\infty$, the incident ray is parallel to the z-axis in medium 1 and the corresponding image point F_i is called the second focal point or image focal point. Its z-coordinate is given by:

$$f_i = \frac{n_2 R}{n_2 - n_1},$$

and it is also referred to as the **second focal length** or **image focal length**. With (2.6.11) and (2.6.10), (2.6.2) can be rewritten as:

$$-\frac{n_1}{s_o} + \frac{n_2}{s_i} = \frac{n_2}{f_i} = -\frac{n_1}{f_o}.$$

By adopting the sign convention listed in Table 2.6.1 below, it turns out that (2.6.2) holds generally. For example, when point S is between the front focal point and the vertex V so that $f_o < s_o < 0$, the rays from S are so strongly diverging that the refraction is insufficient to obtain an image point in medium 2. Instead there is a diverging ray bundle in medium 2 which for an observer in medium 2 seems to come from a point P in medium 1 with z-coordinate si , hence si < 0, in agreement with the fact that P is now to the left of V. Point P is called a **virtual image** because it does not correspond to an actual concentration of light energy in space. When $s_i > 0$ there is a concentration of light energy in P which therefore is then called a **real image**.

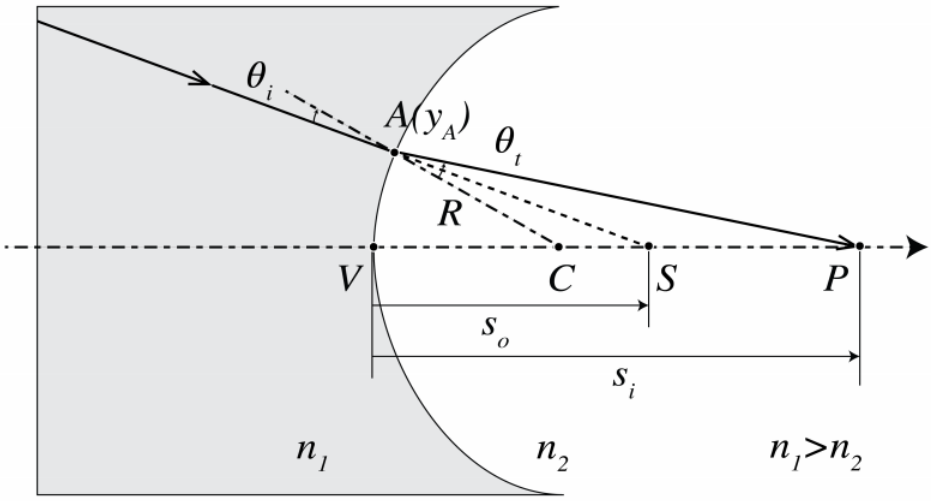


Figure 2.6.2: Imaging of a virtual object S by a spherical interface between two media with refractive indices $n_1 > n_2$.

Table 2.6.1: Sign convention for spherical surfaces and thin lenses. The convention for s_o, f_o, s_i, f_i follows from the fact that these are z-coordinates with the origin at vertex V and the positive z-axis pointing to the right.

quantity	positive	negative
s_o, f_o, s_i, f_i	corresponding point is to the right of vertex	corresponding point is to left of vertex
R	centre of curvature right of vertex	centre of curvature left of vertex
α (astute ray angle)	if counter clockwise rotation over α makes ray parallel to z-axis	if clockwise rotation over α makes ray parallel to z-axis
Refr. index n ambient medium of mirror	before reflection	after reflection

If $n_1 > n_2$ or $R < 0$ (i.e. the surface is concave when seen from the left of the vertex), the right-hand side of (2.6.2) is negative:

$$\frac{n_2 - n_1}{R} < 0.$$

Light rays incident from the left are then refracted away from the z-axis and incident rays that are parallel to the z-axis are refracted such that they never intersect the z-axis in medium 2. Instead, they seem to be emitted by a point in medium 1. As illustrated in Figure 2.6.2, the second focal point is thus virtual with negative focal length given by:

$$f_i = \frac{n_2 R}{n_2 - n_1} < 0.$$

Furthermore, $f_o > 0$ and an incident ray that after refraction is parallel to the z-axis in medium 2 seems for an observer in medium 1 to converge to a point in medium 2. Hence the first focal point is virtual as well: $f_o > 0$. When (2.6.12) holds, for all object points S in front of the lens ($s_o < 0$), the image point is always virtual. In fact, si as given by (2.6.2) is then always negative.

2.5.2 Ray Vectors and Ray Matrices

In geometrical optics it is convenient to use ray vectors and ray matrices. In any plane perpendicular to the z-axis, a ray is determined by the y-coordinate of the point of intersection of the ray with the plane and the acute angle α of the ray with the z-axis. Here $y > 0$ when the intersection point is above the z-axis and $y < 0$ otherwise. We define the ray vector

$$\begin{pmatrix} n\alpha \\ y \end{pmatrix},$$

where n is the local refractive index. The definition with the refractive index as factor in the first element of the ray vector turns out to be convenient. The acute angle α has sign according to the convention in Table 2.6.1.

The ray vectors of a ray in any two planes z = z_1, z = z_2, with $z_2 > z_1$, are related by a so-called ray matrix:

$$\begin{pmatrix} n_2\alpha_2 \\ y_2 \end{pmatrix} = M \begin{pmatrix} n_1\alpha_1 \\ y_1 \end{pmatrix}.$$

where

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

The elements of matrix M depend on the optical components and materials between the planes z = z_1 and z = z_2.

As an example consider the ray matrix that relates a ray vector in the plane immediately before the spherical surface in Figure 2.6.2 to the corresponding ray vector in the plane immediately behind that surface. Using (2.6.5) and (2.6.7) it follows

$$n_1\alpha_1 - n_2\alpha_2 = \frac{(n_2 - n_1)y_1}{R},$$

where we have replaced α_2 by $-\alpha_2$ in (2.6.5), because according to the sign convention, the angle α_2 in Figure 2.6.1 should be taken negative. Because furthermore $y_2 = y_1$, we conclude

$$\begin{pmatrix} n_2\alpha_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} n_1\alpha_1 - \dfrac{(n_2 - n_1)y_1}{R} \\ y_1 \end{pmatrix} = \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ y_1 \end{pmatrix}, \; spherical\ surface,$$

where

$$P = \frac{n_2 - n_1}{R},$$

is called the **power** of the surface.

For a spherical mirror with radius of curvature R, we see that

$$\alpha_1 = \theta_i - \phi,$$

$$\alpha_2 = -\theta_r - \phi,$$

where we take the sign convention for the angles into account. Because $\theta_r = \theta_i$ we find

$$\alpha_2 = -\alpha_1 + 2\phi = -\alpha_1 + 2\frac{y_1}{R},$$

where we used (2.6.7) with $y_A = y_1$. As mentioned in Table 2.6.1, the refractive index after reflection is to be chosen negative. This means that for rays that propagate from right to left, the refractive index in the ray vector should be chosen negative. Hence,

$$n_2\alpha_2 = -n_1\alpha_2 = n_1\alpha_1 - 2n_1\frac{y_1}{R}.$$

Then

$$\begin{pmatrix} n_2\alpha_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ y_1 \end{pmatrix}, \; spherical\ reflector,$$

where

$$P = \frac{2n_1}{R}$$

is the power of the mirror. We conclude that the ray matrix for reflection by a spherical mirror is the same as that for refraction by a spherical surface, provided that n_2 is replaced by $-n_1$. When a ray is propagating from the right to the left, the refractive index of all media and interfaces through which the ray propagates, and at which it is being refracted, should have negative refractive index. After the ray has been reflected a second time, due to which it propagates again from left to right, all refractive indices should again be positive. In this way the matrix method can be applied to optical systems consisting of both refracting elements, such as lenses, and reflecting elements.

Finally we consider the case that between the two planes there is homogeneous material with refractive index n. In that case $\alpha_2 = \alpha_1$ and $y_2 = y_1 + \alpha_1(z_2 - z_1)$, hence

$$M = \begin{pmatrix} 1 & 0 \\ \frac{z_2 - z_1}{n} & 1 \end{pmatrix}, homogeneous\ space.$$

For two planes between which there are a number of optical components, possibly separated by regions with homogeneous material (e.g. air), the ray matrix can be obtained by multiplying the matrices of the individual components and of the homogeneous regions. The order of the multiplication of the matrices is such that the **right-most matrix corresponds to the first component that is encountered while propagating**, and so on.

It should be remarked that the rays considered in the ray matrix approach stay in the same plane, namely the plane through the ray and the z-axis. These rays are called **meridional rays**. By considering only meridional rays, the imaging by optical systems is restricted to two dimensions. Non-meridional rays are called **skew rays**. Skew rays do not pass through the optical axis and they are not considered in the paraxial theory.

2.5.3 The Lens Matrix

We apply ray matrices to a lens. Figure 2.6.4 shows a lens with two spherical surfaces. The refractive index of the lens is n_l and that of the ambient medium is n_m and the distance between the vertices is d. We will first derive the matrix which maps the ray vector in the plane **immediately in front** of the lens to that in the plane **immediately behind** the lens. Let

$$\begin{pmatrix} n_m\alpha_1 \\ y_1 \end{pmatrix} and \begin{pmatrix} n_m\alpha_2 \\ y_2 \end{pmatrix}$$

be two vectors in the two planes which correspond to the same ray. The ray is first refracted by the spherical surface with radius R_1 and centre C_1. Using (2.6.25) and (2.6.20) it follows that the matrix between the ray vectors just before and just behind the spherical surface with radius R_1 and centre C_1 is given by

$$M_1 = \begin{pmatrix} 1 & -\frac{n_l - n_m}{R_1} \\ 0 & 1 \end{pmatrix}$$

The ray propagates then over the distance d through the material of which the lens is made. The matrix that maps ray vectors from the plane inside the lens immediately behind the left spherical surface to a ray vector in the plane immediately before the right spherical surface follows from (2.6.27):

$$M = \begin{pmatrix} 1 & 0 \\ \frac{d}{n_l} & 1 \end{pmatrix}.$$

Finally, the matrix that maps ray vectors from the plane in the lens immediately before the second spherical surface to vectors in the plane immediately behind it is

$$M_1 = \begin{pmatrix} 1 & -\frac{n_m - n_l}{R_2} \\ 0 & 1 \end{pmatrix}.$$

Hence the matrix that maps ray vectors in the plane immediately before the lens to ray vectors in the plane immediately behind the lens is given by the matrix product:

$$M = M_3 M_2 M_1 = \begin{pmatrix} 1 - \frac{d}{n_l}P_2 & -P_1 - P_2 + \frac{d}{n_l}P_1P_2 \\ \frac{d}{n_l} & 1 - \frac{d}{n_l}P_1 \end{pmatrix}, lens,$$

where

$$P_1 = \frac{n_l - n_m}{R_1}, P_2 = \frac{n_m - n_l}{R_2}.$$

The quantity

$$P = P_1 + P_2 - \frac{d}{n_l} P_1 P_2$$

is called the **power** of the lens. It has dimension 1/length and is given in diopter (D), where 1 D = m^{-1}. The power can be positive and negative. The z-axis is the axis of symmetry and is called the **optical axis**. The space to the left of the lens is called the **object space** and that to the right of the lens is called the **image space**.

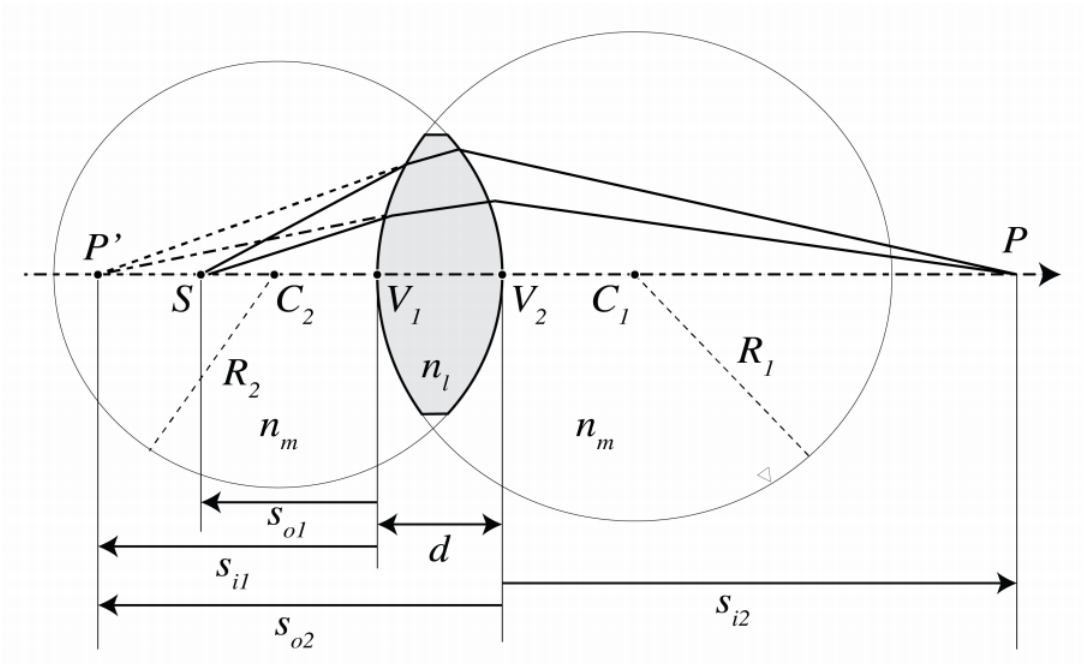


Figure 2.6.3: A spherical lens made of glass of index n_l in a medium of index n_m. The point S is imaged in P.

2.5.4 Focusing with a Thin Lens

For a thin lens the vertices V_1 and V_2 coincide and d = 0, hence (2.6.34) becomes

$$M = \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix}, thin\ lens,$$

where $P = P_1 + P_2$, with P_1 and P_2 given by (2.6.35). The origin of the coordinate system is chosen in the common vertex $V_1 = V_2$. For a ray emerging in image space at height y_2 and parallel to the optical axis: $\alpha_2 = 0$, we have $y_1 = y_2$ and

$$n_m \alpha_1 = P_y.$$

If the power is positive: P > 0, the angle α1 has the same sign as y_1, which implies that the ray in object space has intersected the optical axis in a point F_o with z-coordinate: $z = f_o$ satisfying

$$\frac{1}{f_o} = \frac{\alpha_1}{y_1} = \frac{n_m}{P} = \frac{n_l - n_m}{n_m}\left(\frac{1}{R_1} - \frac{1}{R_2}\right).$$

The point Fo is called the **first focal point** or **object focal point**.

Similarly, by considering a ray in medium 1 which is parallel to the optical axis ($\alpha_1 = 0$) and at height y_1, we get $n_m\alpha_2 = -P\ y_1$ and $y_2 = y_1$. Hence, when P > 0, the angle α_2 of the ray has sign opposite to y_2 and therefore the ray in image space is bent back to the optical axis, yielding a **second focal point** or **image focal point** F_i. Its z-coordinate f_i satisfies:

$$\frac{1}{f_i} = -\frac{\alpha_2}{y_2} = -\frac{n_m}{P} = -\frac{1}{f_o}.$$

When the power P of the lens is positive, $f_i = -f_o > 0$, which means that the first and second focal points are in the object and image space, respectively. A lens with positive power is called **convergent** or **positive**.

A lens with negative power is called divergent and has $f_i = -f_o < 0$. Then incident rays parallel to the optical axis are refracted away from the optical axis and seem to come from a point in front of the lens with z-coordinate $f_i < 0$. Hence the second focal point does not correspond to a location where there is actual concentration of light intensity and hence it is virtual. The first focal point is a virtual object point, because only for a bundle of incident rays that are converging to a certain point behind the lens, the negative refraction can give a bundle of rays that are all parallel to the optical axis.

With the results obtained for the focal coordinates we can rewrite the lens matrix of a thin lens alternatively as

$$M = \begin{pmatrix} 1 & -\frac{n_m}{f_i} \\ 0 & 1 \end{pmatrix}, thin\ lens.$$

2.5.5 Imaging with a Thin Lens

We first consider the general ray matrix (2.6.16), (2.6.17) between two planes $z = z_1$ and $z = z_2$ and ask the following question: what are the properties of the ray matrix such that the two planes are images of each other, or (as this is also called) are each other's conjugate? Clearly for these planes to be each other's image, we should have that for every point coordinate y_1 in the plane $z = z_1$, there is a point with coordinate y_2 in the plane $z = z_2$ such that any ray through (y_1, z_1) (within some cone of rays) will pass through point (y_2, z_2). Hence for any angle α_1 (in some interval of angles) there is an angle α_2 such that (2.6.16) is valid. This means that for any y_1 there is a y_2 such that for all angles α_1:

$$y_2 = Cn_1\alpha_1 + Dy_1,$$

This requires that

$$C = 0, condition\ for\ imaging.$$

The ratio of y_2 and y_1 gives the magnification. Hence

$$\frac{y_2}{y_1} = D,$$

is the **magnification** of the image (this quantity has sign).

To determine the image of a point by a thin lens we first derive the ray matrix between the planes $z = z_1 < 0$ and $z = z_2 > 0$ with a thin lens in between with vertex at the origin. This matrix is the product of the matrix for propagating from $z = z_1$ to the plane immediately in front of the lens, the matrix of the thin lens and the matrix for propagation from the plane immediately behind the lens to the plane $z = z_2$:

$$M = \begin{pmatrix} 1 & 0 \\ \frac{z_2}{n_m} & 1 \end{pmatrix} \begin{pmatrix} 1 & -\frac{n_m}{f_i} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\frac{z_1}{n_m} & 1 \end{pmatrix} = \begin{pmatrix} 1+\frac{z_1}{f_i} & -\frac{n_m}{f_i} \\ -\frac{z_1}{n_m} + \frac{z_2}{n_m} + \frac{z_1 z_2}{n_m f_i} & 1-\frac{z_2}{f_i} \end{pmatrix}$$

The imaging condition (2.6.41) implies:

$$-\frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{f_i}, Lensmaker's\ Formula.$$

where we have written $s_o = z_1$ and $s_i = z_2$ for the z-coordinates of the object and the image. Because for the thin lens matrix (2.6.43): $D = 1 - z_2/f_i$, it follows by using (2.6.44) that the magnification (2.6.42) is given by

$$M = \frac{y_i}{y_o} = 1 - \frac{s_i}{f_i} = \frac{s_i}{s_o},$$

where we have written now y_o and y_i instead of y_1 and y_2, respectively.

For a positive lens: $f_i > 0$ and hence (2.6.44) implies that $s_i > 0$ provided $|s_o| < f_i = |f_o|$, which means that the image by a convergent lens is real if the object is further from the lens than the first focal point F_o. The case $s_o > 0$ corresponds to a virtual object, i.e. to the case of a converging bundle of incident rays, which for an observer in object space seems to converge to a point at distance so behind the lens. A convergent lens ($f_i > 0$) will then make an image between the lens and the second focal point. In contrast, a diverging lens ($f_i < 0$) can turn the incident-converging bundle into a real image only if the virtual object point is between the lens and the focal point. If the virtual object point has larger distance to the lens, the convergence of the incident bundle is too weak and

the diverging lens then refracts this bundle into a diverging bundle of rays with vertex at the virtual image point in front of the lens ($s_i < 0$).

One can also construct the image with a ruler. Consider imaging a finite object S_1S_2 as shown in Figure 2.6.4. Let y_o be the y-coordinate of S_2. We have $y_o > 0$ when the object is above the optical axis. Draw the ray through the focal point F_o in object space and the ray through the centre V of the lens. The first ray becomes parallel in image space. The latter intersects both surfaces of the lens almost in their (almost coinciding) vertices and therefore the refraction is opposite at both surfaces and the ray exits the lens parallel to its direction of incidence. Furthermore, its lateral displacement can be neglected because the lens is thin. Hence, **the ray through the centre of the lens is not refracted**. The intersection in image space of the two rays gives the location of the image point P_2 of S_2. The image is real if the intersection occurs in image space and is virtual otherwise. For the case of a convergent lens with a real object with $y_o > 0$ as shown in Figure 2.6.4, it follows from the similar triangles Δ AVF$_i$ and Δ P$_2$P$_1$F$_i$ that

$$\frac{y_o}{|y_i|} = \frac{f_i}{s_i - f_i},$$

where we used $|f_o| = f_i$. From the similar triangles Δ S$_2$S$_1$F$_o$ and Δ BVF$_o$:

$$\frac{|y_i|}{y_o} = \frac{f_i}{f_o - s_o}.$$

(the absolute value of y_i is taken because according to our sign convention y_i in Figure 2.6.4 is negative whereas (2.6.44) is a ratio of lengths). By multiplying these two equations we get the **Newtonian form** of the lens equation:

$$x_o x_i = -f_i^2 = -f_o^2,$$

where x_o and x_i are the z-coordinates of the object and image relative to those of the first and second focal point, respectively:

$$x_o = s_o - f_o, x_i = s_i - f_i.$$

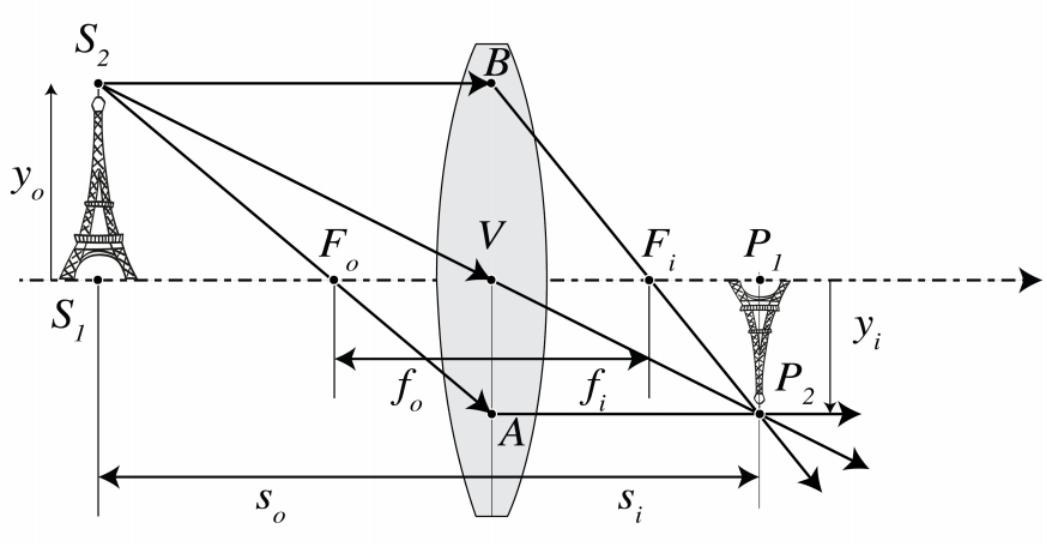


Figure 2.6.4: Object and image for a thin lens.

Hence x_o is negative if the object is t$_o$ the left of F_o and x_i is positive if the image is to the right of F_i.

The **transverse magnification** is

$$M = \frac{y_i}{y_o} = \frac{s_i}{s_o} = -\frac{x_i}{f_i},$$

where the second identity follows from considering the similar triangles ΔP$_2$P$_1$F$_i$ and ΔAVF$_i$ in Figure 2.6.4. A positive M means that the image is erect, a negative M means that the image is inverted (as is always the case for a single lens).

All equations are also valid for a thin negative lenses and for virtual objects and images. Some examples of real and virtual object and image points for a positive and a negative lens are shown in Figs. 2.6.5 and 2.6.6.

2.5.6 Two Thin Lenses

The imaging by two thin lenses L_1 and L_2 can easily be obtained by construction. We simply construct the image obtained by the first lens as if the second lens were not present and use this image as (possibly virtual) object for the second lens. In Figure 2.6.7 an example is shown where the distance between the lenses is larger than the sum of their focal lengths. First the image P' of S is constructed as obtained by L_1 as if L_2 were not present. We construct the intermediate image P' due to lens L_1 using ray 2 and 3. P_1 is a real image of lens L_1 which is also real object for lens L_2. Ray 3 is parallel to the optical axis between the two lenses and is thus refracted by lens L2 through its back focal point F_{2i}. Ray 4 is the ray from P1 through the centre of lens L_2. The image point P is the intersection of ray 3 and 4.

In the case of Figure 2.6.8 the distance d between the two positive lenses is smaller than their focal lengths. The intermediate image P' is a real image for L_1 obtained as the intersection of rays 2 and 3 passing through the object and image focal points F_{1o} and F_{1i} of lens L_1. P' is now a virtual object point for lens L_2. To find its image by L_2, draw ray 4 from P' through the centre of lens L_2 back to S (this ray is refracted by lens L_1 but not by L_2) and draw ray 3 as refracted by lens L_2. Since ray 3 is parallel to the optical axis between the lenses, it passes through the back focal point F_{2i} of lens L_2. The intersection point of ray 3 and 4 is the final image point P.

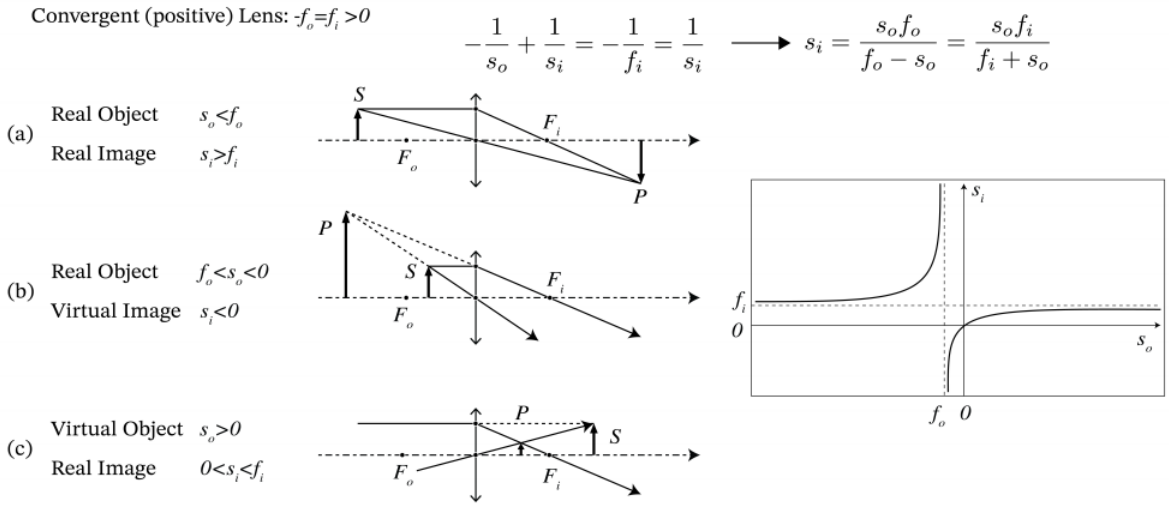


Figure 2.6.5: Real and virtual objects and images for a convergent thin lens, i.e. $-f_o = f_i > 0$. In (a) the object is real with $s_o < f_o$ and the image is real as well ($s_i > 0$). In (b) the object is between the focal point in front of the lens and the lens: $f_o < s_o < 0$. Then the rays from the object are too divergent for the lens to make them convergent in image space and the image is virtual: $s_i < 0$. In (c) there is a cone of converging rays incident on the lens from the left which, in the absence of the lens, would converge to point S behind the lens. Hence S is a virtual object ($s_0 > 0$). The image is real and can be constructed with the two rays shown. In (d) si is shown as function of so for a convergent lens (see Eq. (2.6.44)).

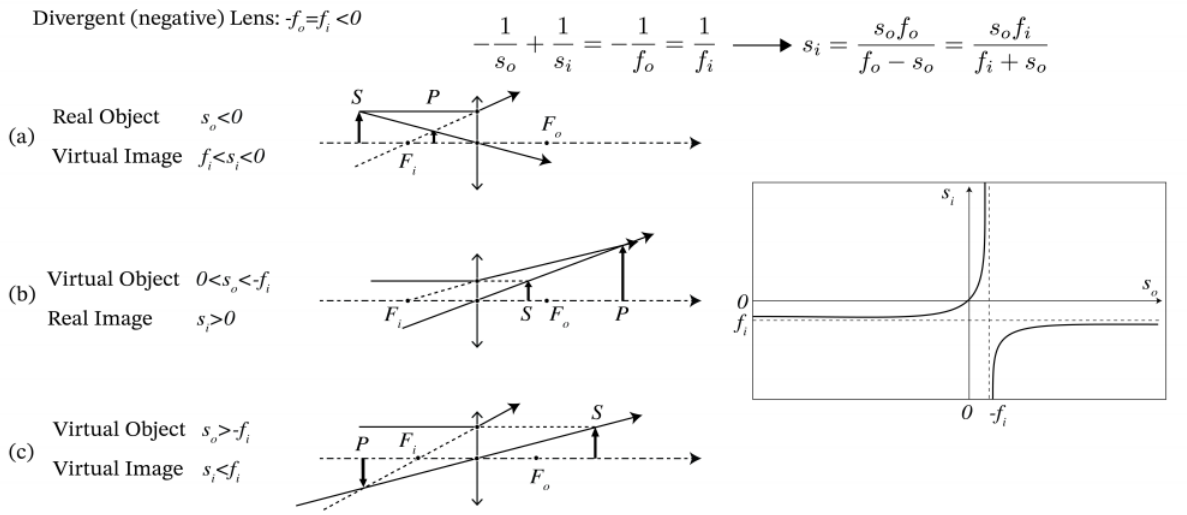


Figure 2.6.6: Real and virtual objects and images for a divergent thin lens, i.e. $-f_o = f_i < 0$. In (a) the object is real, i.e. $s_o < 0$. The diverging lens makes the cone of rays from the object more divergent so that the image is virtual: $s_i < 0$. When the object is virtual, there is a cone of converging rays incident from the left which after extension to the right of the lens (as if the lens is not present) intersect in the virtual object S ($s_o > 0$). It depends on how strong the convergence is whether the diverging lens turns this cone into converging rays or whether the rays keep diverging. In (b) $0 < s_o < -f_i$, and the image is real. In c) $s_o > -f_i$ and the image is real ($s_i > 0$). In (d) si is shown as function of so for a divergent lens ($f_i < 0$ (see Eq.‘(2.6.44)).

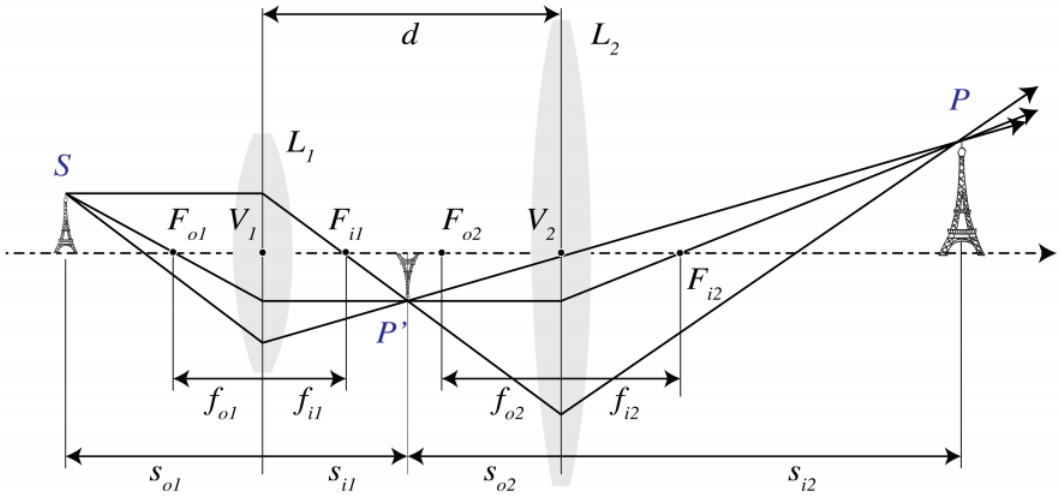


Figure 2.6.7: Two thin lenses separated by a distance that is larger than the sum of their focal lengths.

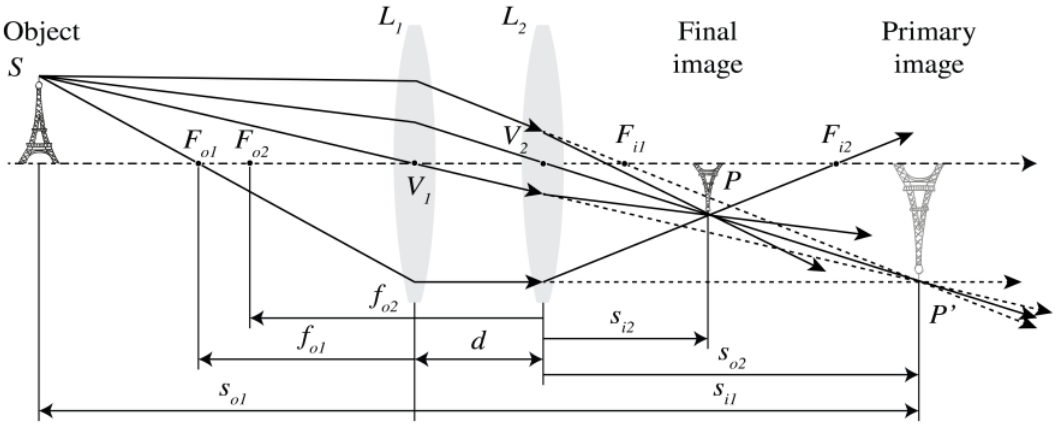


Figure 2.6.8: Two thin lenses at a distance smaller than their focal lengths.

It is easy to express the z-coordinate s_i with respect to the coordinate system with origin at the vertex of L_2 of the final image point, in the z-component s_o with respect to the origin at the vertex of lens L_1 of the object point. We use the Lensmaker's Formula for each lens while taking care that the proper local coordinate system is used in each case. The *intermediate* image P' due to lens L_1 has z-coordinate s_{1i} with respect to the coordinate system with origin at the vertex V_1, which satisfies:

$$\frac{1}{s_o}+\frac{1}{s_{1i}}=\frac{1}{f_{1i}}.$$

P' is object for lens L_2 with z-coordinate with respect to the coordinate system with origin at V_2 given by: $s_{2o} = s_{1i} - d$, where d is the distance between the lenses. Hence, with $s_i = s_{2i}$ the Lensmaker's Formula for lens L_2 implies:

$$-\frac{1}{s_{1i}-d}+\frac{1}{s_i}=\frac{1}{f_{2i}}.$$

By solving $(2.6.51)$ for s_{1i} and substituting the result into $(2.6.52)$, we find

$$s_i=\frac{-df_{1i}f_{2i}+f_{2i}(f_{1i}-d)s_o}{f_{1i}(f_{2i}-d)+(f_{1i}+f_{2i}-d)s_o}, two\ thin\ lenses.$$

By taking the limit $s_o \to -\infty$, we obtain the z-coordinate f_i of the image focal point of the two lenses, while $s_i \to \infty$ gives the z-coordinate f_o of the object focal point:

$$f_i=\frac{(f_{1i}-d)f_{2i}}{f_{1i}+f_{2i}-d},$$

$$f_o=-\frac{(f_{2i}-d)f_{1i}}{f_{1i}+f_{2i}-d},$$

Except when the refractive indices of the media before and after the lens are different, the object and image focal lengths of a thin lens are the same. However, as follows from the derived formula for an optical system with two lenses, the object and image focal lengths are in general different when there are several lenses.

By construction using the intermediate image, it is clear that the magnification of the two-lens system is the product of the magnifications of the two lenses:

$$M=M_1M_2.$$

Remarks

1. When $f_{1i} + f_{2i} = d$ the focal points are at infinity. Such a system is called **telecentric**.

2. In the limit where the lenses are very close together: $d \to 0$, $(2.6.53)$ becomes

$$-\frac{1}{s_o}+\frac{1}{s_i}=\frac{1}{f_{1i}}+\frac{1}{f_{2i}}.$$

The focal length fi of the system of two lenses in contact thus satisfies:

$$\frac{1}{f_i}=\frac{1}{f_{1i}}+\frac{1}{f_{2i}}.$$

Two positive lenses in close contact enforce each other, i.e. the second positive lens makes the convergence of the first lens stronger. Similarly, two negative lenses in contact make a more strongly negative system. The same applies for more than two lenses in close contact.

3. Although for two lenses the image coordinate can still be expressed relatively easily in the object distance, for systems with more lenses finding the overall ray matrix and then using the image condition $(2.6.41)$ is a much better strategy.

2.5.7 The Thick Lens

At the left of Figure $2.6.9$ a thick lens is shown. The first focal point is defined as the point whose rays are refracted such that the emerging rays are parallel to the optical axis. By extending the incident and emerging rays by straight segments, the points of intersection are found to be on a curved surface, which close to the optical axis, that is in the paraxial approximation, is in good approximation a plane perpendicular to the optical axis. This plane is called the **primary principal plane** and its intersection with

the optical axis is called the primary principal point H_1. By considering incident rays which are parallel to the optical axis and therefore focused in the back focal point, the **secondary principal plane** and secondary principal point H_2 are defined in a similar way (see the drawing at the right in Figure 2.6.9). The principal planes need not be inside the lens. In particular for meniscus lenses, this is not the case. It can be seen from Figure 2.18 that the principal planes are images of each other, with unit magnification. Hence, if an object is placed in the primary principal plane (hypothetically if this plane is inside the lens), its image is in the secondary principal plane. The image is erect and has unit magnification.

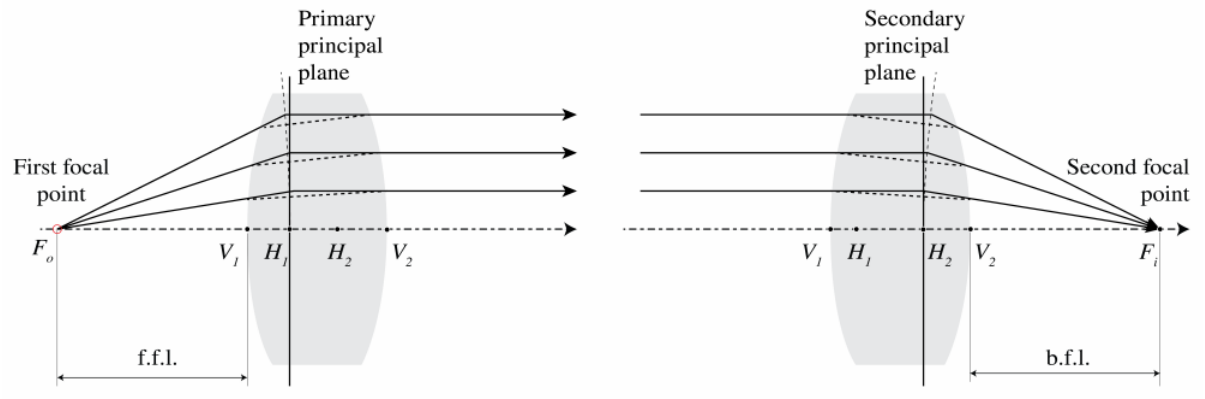


Figure 2.6.9: Principal planes of a thick lens, with front and back focal lengths: f.f.l and b.f.l.

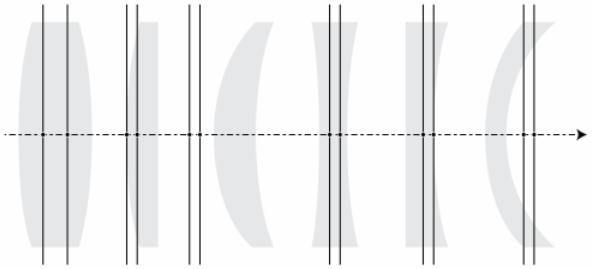


Figure 2.6.10: Position of the principal planes for several lenses.

We recall the result (2.6.32) for the ray matrix between the planes through the front and back vertices V_1, V_2 of a thick lens with refractive index n_l and thickness d:

$$M_{V_1V_2} = \begin{pmatrix} 1-\frac{d}{n_l}P_2 & -P \\ \frac{d}{n_l} & 1-\frac{d}{n_l}P_1 \end{pmatrix}, thick\ lens,$$

where

$$P_1 = \frac{n_l - n_m}{R_1}, P_2 = \frac{n_m - n_l}{R_2},$$

and n_m is the refractive index of the ambient medium, and

$$P = P_1 + P_2 - \frac{d}{n_l}P_1P_2.$$

If h_1 is the z-coordinate of the first principal point H_1 with respect to the coordinate system with origin vertex V_1, we have according to (2.6.27) for the ray matrix between the primary principal plane and the plane through vertex V_1

$$M_1 = \begin{pmatrix} 1 & 0 \\ \frac{h_1}{n_m} & 1 \end{pmatrix}.$$

Similarly, if h_2 is the coordinate of the secondary principal point H_2 with respect to the coordinate system with V_2 as origin, the ray matrix between the plane through vertex V_2 and the secondary principal plane is

$$M_2 = \begin{pmatrix} 1 & 0 \\ \dfrac{h_2}{n_m} & 1 \end{pmatrix}.$$

The ray matrix between the two principle planes is then

$$M_{H_1H_2} = M_2 M_{V_1V_2} M_1.$$

The coordinates h1 and h2 can be found by imposing to the resulting matrix the imaging condition (2.6.41): C = 0 and the condition that the magnification should be unity: D = 1, which follows from (2.6.42). We omit the details and only give the resulting expressions here:

$$h_1 = \frac{n_m}{n_l}\frac{P_2}{P}d,$$

$$h_2 = -\frac{n_m}{n_l}\frac{P_1}{P}d.$$

Furthermore, (2.6.64) becomes

$$M_{H_1H_2} = \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix}.$$

We see that **the ray matrix between the principal planes is identical to the ray matrix of a thin lens** (2.6.35). We therefore conclude that if the coordinates in object space are chosen with respect to the origin in the primary principal point H_1, and the coordinates in image space are chosen with respect to the origin in the secondary principal point H_2, the expressions for the first and second focal points and for the coordinates of the image point in terms of that of the object point are identical to that for a thin lens. An example of imaging by a thick lens is shown in Figure 2.6.11.

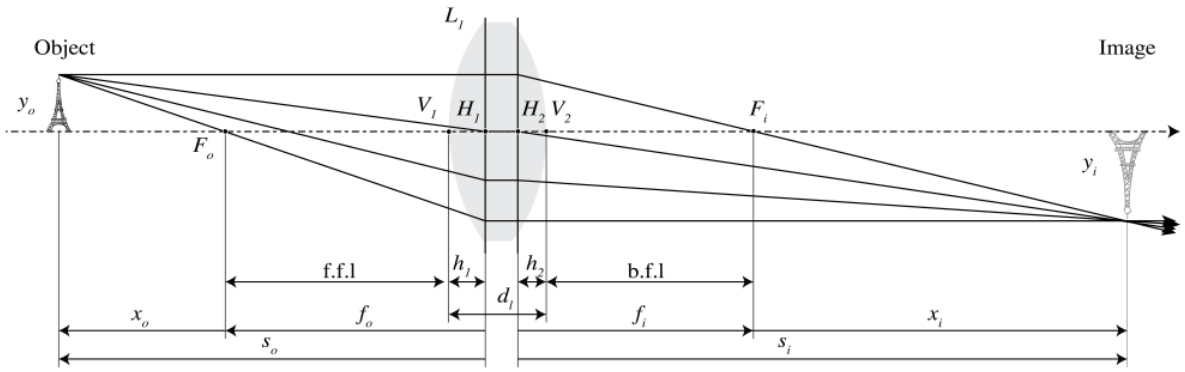


Figure 2.6.11: Thick-lens geometry. There holds $f_i = f_o$ if the ambient medium left of the lens is the same as to the right of the lens.

2.5.8 Stops

An element such as the rim of a lens or a diaphragm which determines the set of rays that can contribute to the image, is called the **aperture stop**. An ordinary camera has a variable diaphragm.

The **entrance pupil** is the image of the aperture stop by all elements to the left of the aperture stop. If there are no lenses between object and aperture stop, the aperture stop itself is the entrance pupil. Similarly the **exit pupil** is the image of the aperture stop by all elements to the right of it. The entrance pupil determines for a given object the cone of rays that enters the optical system, while the cone of rays leaving and taking part in the image formation is determined by the exit pupil (see Figure 2.6.12). Note that in constructing the entrance pupil as the image of the aperture stop by the lenses to the left of it, are propagating from the right to the left. Hence the aperture stop is a real object in this construction, while the entrance pupil can be a real or a virtual image. The rays used in constructing the exit pupil as the image of the aperture stop by the lenses following the stop are propagating from the left to the right. Hence also in this case the aperture stop is a real object while the exit pupil can be a real or a virtual image of the aperture stop.

For any object point, the **chief ray** is the ray in the cone that passes through the centre of the entrance pupil, and hence also through the centres of the aperture stop and the exit pupil. A marginal ray is the ray that for an object point on the optical axis passes

through the rim of the entrance pupil (and hence also through the rims of the aperture stop and the exit pupil).

For a fixed diameter D of the exit pupil and given x_o, the magnification of the system is according to (2.6.50) and (2.6.48) given by $M = -x_i/f_i = f_i/x_o$. It follows that when fi is increased, the magnification increases. A larger magnification means a lower energy density, hence a longer exposure time, i.e. **the speed of the lens is reduced**. Camera lenses are usually specified by two numbers: the focal length f_i and the diameter D of the exit pupil. The **f-number** is the ratio of the focal length to this diameter:

$$f-number = \frac{f}{D}.$$

For example, f-number= 2 means f = 2D. Since the exposure time is proportional to the square of the f-number, a lens with f-number 1.4 is twice as fast as a lens with f-number 2.

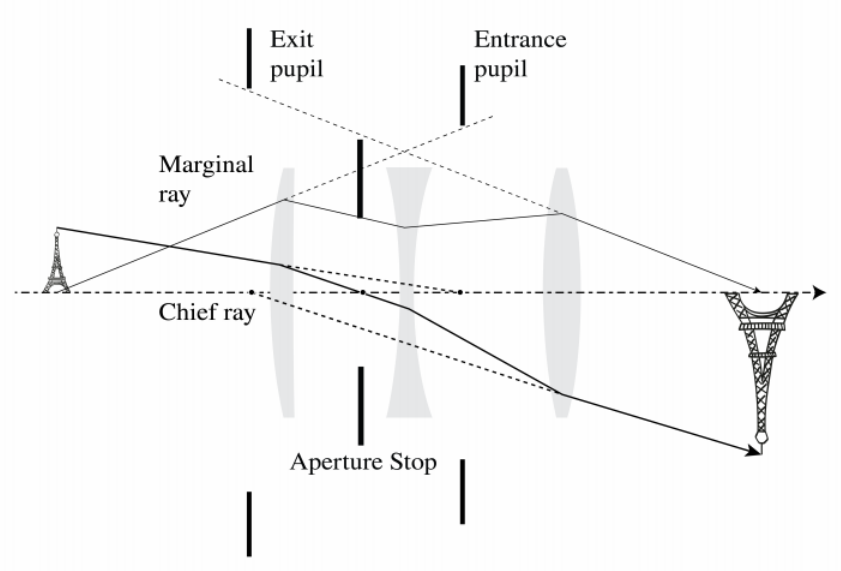


Figure 2.6.1: Aperture stop (A.S.) between the second and third lens, with entrance pupil and exit pupil (in this case these pupils are virtual images of the aperture stop). Also shown are the chief ray and the marginal ray.

Welcome to the Physics Library. This Living Library is a principal hub of the LibreTexts project, which is a multi-institutional collaborative venture to develop the next generation of open-access texts to improve postsecondary education at all levels of higher learning. The LibreTexts approach is highly collaborative where an Open Access textbook environment is under constant revision by students, faculty, and outside experts to supplant conventional paper-based books.

2.8: Aberrations

For designing advanced optical systems Gaussian geometrical optics is not sufficient. Instead nonparaxial rays, and among them also non-meridional rays, must be traced using software based on Snell's Law with the sine of the angles of incidence and refraction. Often many thousands of rays are traced to evaluate the quality of an image. It is then found that in general the non-paraxial rays do not intersect at the ideal Gaussian image point. Instead of a single spot, a spot diagram is found which is more or less confined. The deviation from an ideal point image is quantified in terms of **aberrations**. One distinguishes between monochromatic and chromatic aberrations. The latter are caused by the fact that the refractive index depends on wavelength. Recall that in paraxial geometrical optics Snell's Law (2.3.7) is replaced by: $n_i\theta_i = n_t\theta_t$, i.e. sin θ_i and sin θ_t are replaced by the linear terms. If instead one retains the first two terms of the Taylor series of the sine, the errors in the image can be quantified by five monochromatic aberrations, the so-called **primary** or **Seidel aberrations**. The best known is **spherical aberration**, which is caused by the fact that for a convergent spherical lens, the rays that makes a large angle with the optical axis are focused closer to the lens than the paraxial rays (see Figure 2.8.1). **Distortion** is one of the five primary aberrations. It causes deformation of images due to the fact that the magnification depends on the distance of the object point to the optical axis.

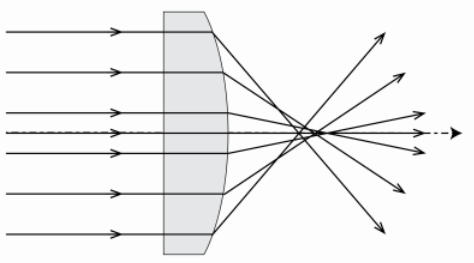


Figure 2.8.1: Spherical aberration of a planar-convex lens.

For high-quality imaging the aberrations have to be reduced by adding more lenses and optimising the curvatures of the surfaces, the thicknesses of the lenses and the distances between them. For high quality systems, a lens with an aspherical surface is sometimes used. Systems with very small aberrations are extremely expensive, in particular if the field of view is large, as is the case in lithographic imaging systems used in the manufacturing of integrated circuits.

2.7.1 Diffraction Optics

Aberrations can be quantified by analysing the spot diagram or, alternatively, by considering the shape of the actual wavefront in image space, which in Gaussian geemotrical optics would converge to the Gaussian image point. Aberrations cause the wavefront to deviate from a perfect sphere. According to a generally accepted criterion formulated first by Rayleigh, aberrations start to deteriorate images considerably if the wavefront aberrations cause path length differences of more than a quarter of the wavelength. When the aberrations are less than this, the system is called **diffraction limited**. A very comprehensive treatment of aberration theory can be found in the book by Braat et al. Even if the wave transmitted by the exit pupil would be perfectly spherical, the wave front consists of only a circular section of a sphere since the field is limited by the aperture. An aperture causes **diffraction**, i.e. bending and spreading of the light. When one images a point object on the optical axis, diffraction causes the light distribution called the Airy spot, as shown in Figure 2.8.3. The Airy spot has full-width at half maximum:

$$FWHM = 0.6\frac{\lambda}{NA},$$

with NA= arcsin(a/si) is the numerical aperture (i.e. 0<NA<1) with a the radius of the exit pupil and si the image distance as predicted by Gaussian geometrical optics. Diffraction depends on the wavelength and hence it cannot be described by geometrical optics, which applies in the limit of vanishing wavelength. We will treat diffraction by apertures in Chapter 6.

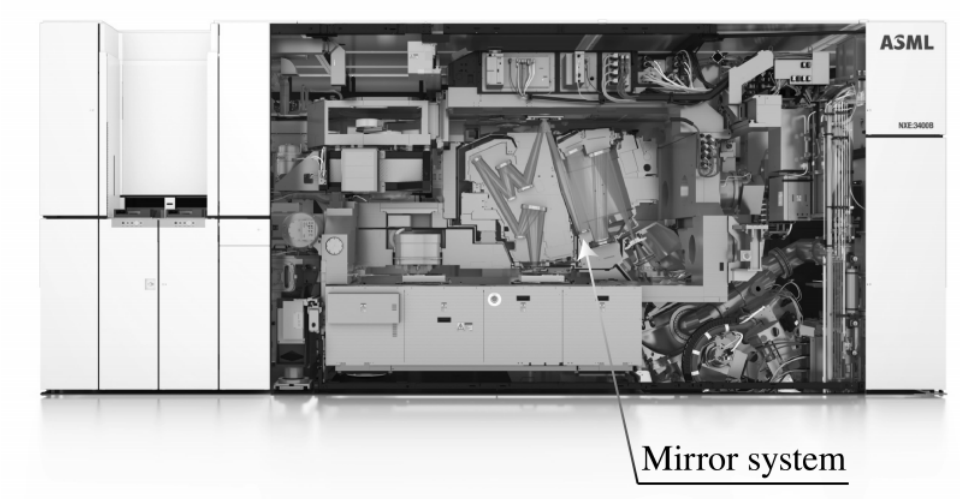


Figure 2.8.2: The EUV stepper TWINSCAN NXE:3400B.Lithographic lens system for DUV (192 nm), costing more than € 500.000. Ray paths are shown in purple. The optical system consists of mirrors because there are no suitable lenses for this wavelength (Courtesy of ASML).

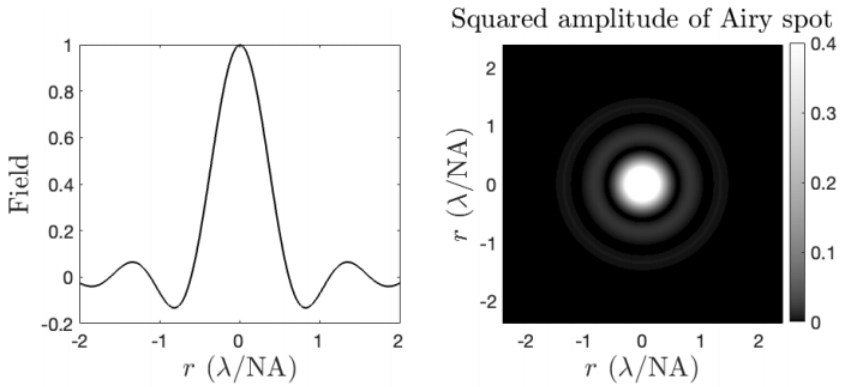


Figure 2.8.3: Left: cross section of the field of the Airy pattern. Right: intensity pattern of the Airy pattern.

CHAPTER OVERVIEW

3: Optical Instruments

3.1: What you should know and be able to do after studying this chapter
3.2: The Camera Obscura
3.3: The Camera
3.4: Camera in a Smart Phone
3.5: The Human Eye
3.6: Magnifying Glasses
3.7: Eyepieces
3.8: The Compound Microscope
3.9: The Telescope

3.1: What you should know and be able to do after studying this chapter

- Understand the working principle of a camera.
- Understand the optics of the eye and its accommodation with the near and far point.
- The working of eye glasses.
- Understand the principle of the magnifier and the eyepiece and its use in the microscope and the telescope.
- Understand the microscope and the telescope concept and the (angular) magnification in both cases.

After the treatment in the preceding chapter of the laws of Gaussian geometrical optics, more complex systems based on lenses and reflectors can now be considered.

3.2: The Camera Obscura

The **camera obscura** or pinhole camera is the simplest image forming system. It consists of a closed box with a pinhole on one side. An inverted image is cast on the opposite side of the box as shown in Figure 3.2.1. If the hole is too large, the image is very blurred. At the cost of less light, the image can be made sharper by reducing the aperture The camera obscura can form images of objects across an extremely wide angular field due to great depth of focus and over a large range of distances (great depth of field) as you can see in the right picture of Figure 3.2.1. If a film would be used to record the image, very long exposure times are however needed because only a small amount of light enters the pinhole, (f-number=500). It is believed that painters such as Johannes Vermeer have used the camera obscura to make paintings of external scenes.

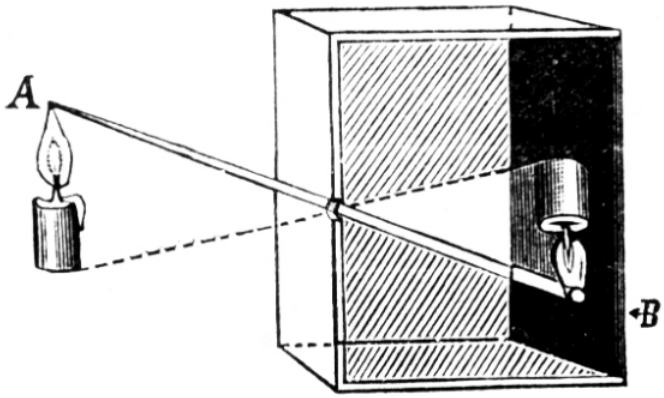


Figure 3.2.1: The principle of the camera obscura (from Wikimedia Commons in Fizyka z. (1910) / Public Domain). Examples of pictures made with a camera obscura can be found here.

3.3: The Camera

In Figure 3.3.1 a single-lens reflex (SLR) camera is shown. The name does not mean that there is only one lens in the optical system, but that the photographer looks through the same lenses that the picture is taken with, instead of looking through a separate parallel optical system as in the twin reflex camera. After traversing the first few lens elements, the light passes through an iris diaphragm with adjustable diameter with which the f-number can be changed. After the lenses the light is reflected by a movable mirror tilted at 45^{o}, passes through a prism and exits the camera through the finder eyepiece. When the shutter is released, the diaphragm closes to a preset value, the mirror swings up and the CCD is exposed. To focus the camera, the entire lens is moved toward or away from the detection plane. The autofocus is based on maximising the contrast of the images. The **angular field of view** (AFOV) is defined for scenes at large distances and is equal to the angle subtended at the lens by the detector when the image distance is the **focal length** f, i.e. the object is at infinity. The AFOV decreases when f increases. A standard SLR has a focal length of around 6 cm and the AFOV is then between 40° and 50°.

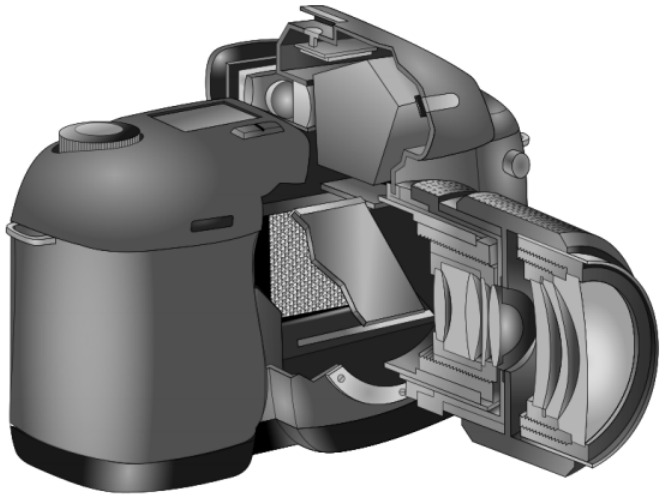


Figure 3.3.2: Digital SLR camera. The pixelated digital sensor is behind a movable mirror (from Wikimedia Commons by Jean François WITZ / CC BY-SA 3.0).

More complex systems can have a variable focal length by changing the distance between the lenses, i.e. they are able to zoom into a scene.

The **depth of focus** is a range of object distances around a given distance for which the images on the sensor are sharp. The depth of focus depends on the diaphragm. When the aperture is wide open, rays forming the image will make larger angles with the optical axis. When these rays come from objects at various distances they will for a large diaphragm cause more blurred images on the sensor (see Figure 3.4.2). When the aperture is reduced, this effect is less and therefore a smaller diaphragm implies a larger depth of focus. The drawback is that less light reaches the sensor, therefore a longer exposure time is needed.

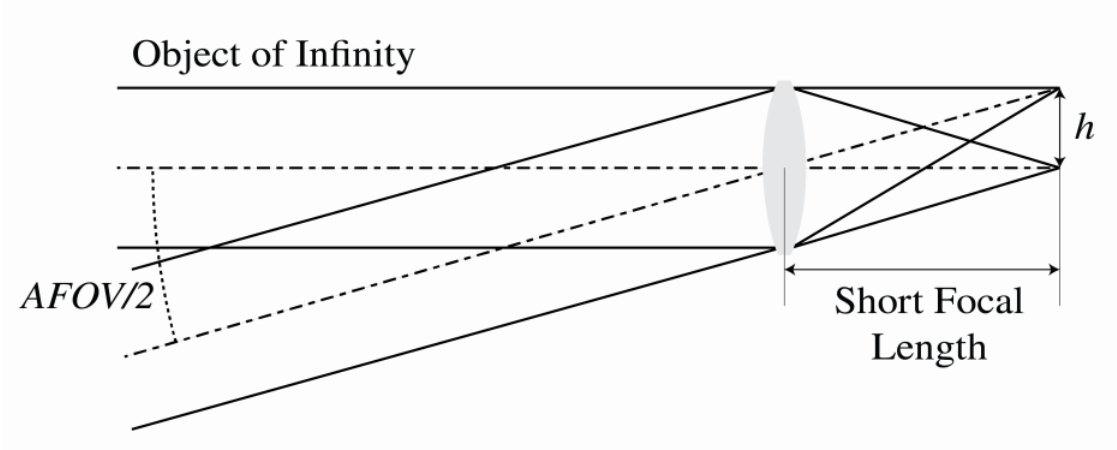


Figure 3.3.1: Angular field of view

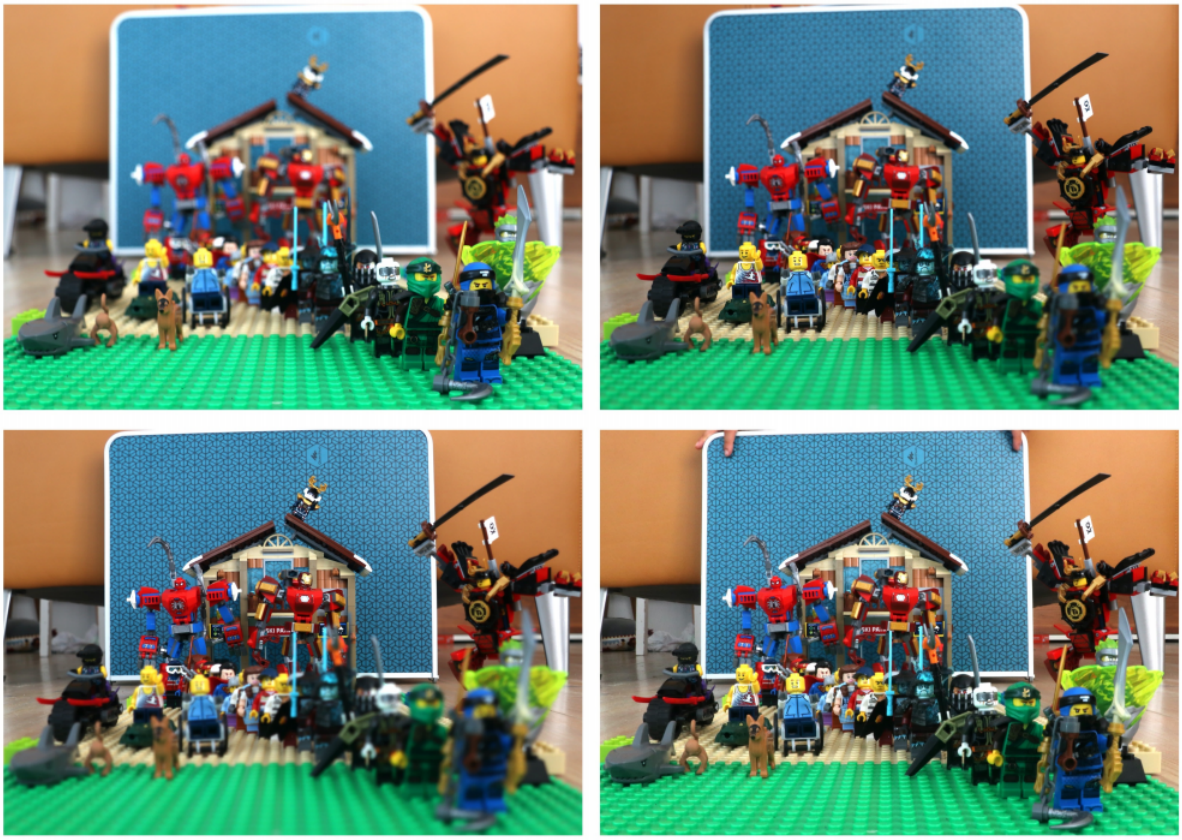


Figure 3.3.2: Four images taken with different diaphragm and different focal plane. The image on bottom right is taken with a small diaphragm and all the image appears clear (photos taken by Aurèle J.L. ADAM / CC BY-SA).

3.4: Camera in a Smart Phone

These can contain standard **double Gauss** or **Cook triplet** lenses and sometimes more advanced aspheres. The image sensor is CMOS. Nowadays smart phones have auto-focus systems where the lens is moved towards or away from the sensor using different criteria. In the standard contrast detection auto-focus the lens is moved until the contrast in the image is largest. This trial-and-error method is relatively slow: it usually takes 1 second to focus. In high-end cameras so-called phase detection auto-focus is used. The relative positions of two spots obtained from the focusing of two small diaphragms on either side of the optical axis is analysed. Depending on the relative positions of the spots, the lens is moved towards or away from the sensor. This auto-focus system is three times faster than the contrast-based system. The third method that is currently used is an active auto-focus system similar to radar. An infrared laser pulse is emitted and the distance of the object of interest is determined from the return time of the reflected pulse. This method is roughly as fast as the phase detection method. It works very well under low light conditions, but is not accurate for objects at distances of more than 5 m. In all smart phone cameras, blurry images are sharpened by post-processing using software.

3.5: The Human Eye

The eye is an adaptive optical device.

3.4.1 Anatomy

The human eye is made of an almost spherical (24 mm long and 22 mm across) gelatinous substance called the **vitreous humor** with refractive index 1.337, surrounded by a white shell, the **sclera** (Figure 3.5.3). At the front, the sclera has an opening with a transparent lens called the **cornea**, with for green light an index of refraction of 1.376. Most of the bending of the rays takes place at the air-cornea interface and this is why it is difficult to see under water (n_{water} = 1.33). After passing the cornea, the rays reach the **aqueous humour** (n ≈1.336) with the **iris** (i.e. pupil). It can expand or contract from a 2 mm (bright sun) to 8 mm (low light) diameter to adapt to the light intensity. The iris gives colour to the eye. After the iris, the rays reach the flexible **crystalline lens** which has the size of a bean (9 mm in diameter, and 4 mm thick in unaccommodated condition). Its index of refraction varies from 1.406 in the centre to 1.386 at the edge.

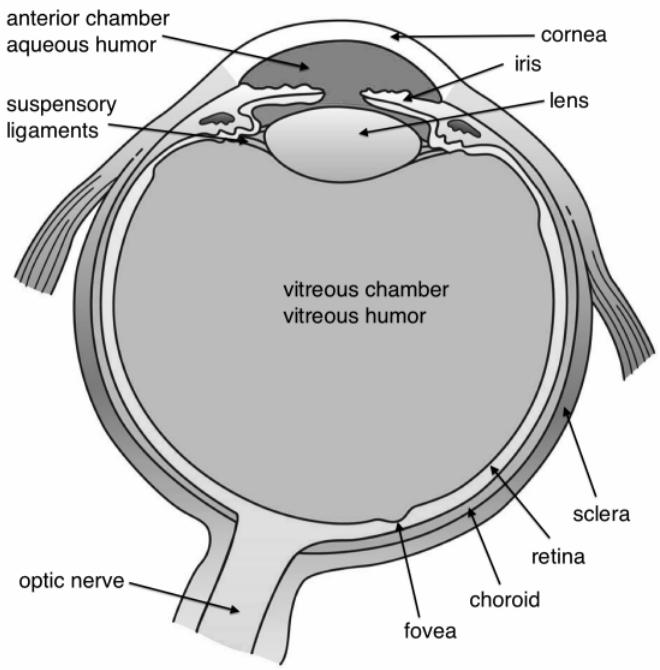


Figure 3.5.3: Cross section of a human eye (from Wikimedia Commons by Holly Fischer / CC BY).

3.4.2 Working of the eye

The entire eye can very accurately be treated as two lenses in contact, of which the crystalline lens can change its focal length. Often the system is approximated by only a single lens. In unaccommodated condition, the first focal distance of the lens system is f_o = 16 mm as measured from the cornea. The second focal distance is equal to the length of the eye: f_i = 24 mm. These focal distances are different, because the refractive indices of the surrounding medium (air and vitreous humour) differ. The power of the intact, unaccommodated eye lens system is, according to (2.5.35), (2.5.39):

$$\mathfrak{D} = \frac{n_{vh}}{f_2} = \frac{1.337}{0.0243} = 55 \text{ Diopter}.$$

In relaxed condition the lens focuses light coming from infinity on the retina. When the object is closer, the eye muscles contract due to which the crystalline lens becomes more convex and the focal length of the system decreases, as seen on the right of Figure 3.5.4. At a certain point, the object will be too close to be focused on the retina. This is called the near point of the eye. Due to the loss of elasticity of the muscle, the **near point** moves from 7 cm for teens to 100 cm for a 60-year-old. Figure 3.5.4 shows the optical rays entering the eyes, for two configurations: an object at infinity and an object nearby. The so-called **normal near point**

is at 25 cm. The **far point** is the furthest object which is imaged on the retina by the unaccommodated eye. For the normal eye the far point is at infinity.

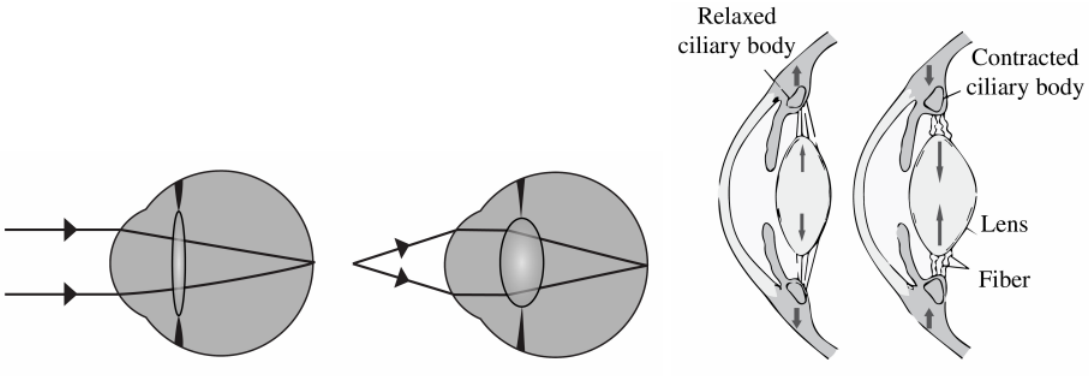


Figure 3.5.4: Left: Optical rays showing how an eye accommodates by changing its focal length. Right: Relaxed and contracted muscle at the crystalline lens needed for this accommodation.(Left: adapted from Wikimedia Commons Erin Silversmith / BY-NC-SA 2.5 Generic. Right: adapted from Sjaastad O.V., Sand O. and Hove K. (2010) Physiology of domestic animals, 2nd edn., Oslo: Scandinavian Veterinary Press).

3.4.3 Retina

The retina is composed of approximately 125 million photoreceptor cells: the **rods** and the **cones**. The rods are highly sensitive black and white (intensity) sensors, while the cones are colour sensitive for the wavelengths 390 nm - 780 nm. UV light is absorbed by the lens (people whose lens is removed because of cataract can "see" UV light). The **fovea centralis** is the most sensitive centre of the retina with a high density of cones. The eyes move continuously to focus the image on this area. The information is transferred by the optical nerve, placed at the back of the eye, where it causes a blind spot.

3.4.4 Dioptric Power of a lens

For a combination of two thin lenses, it is not so easy to derive the resulting focal length from the individual focal length of the lenses. It is easier to consider the dioptric power. For a single lens the dioptric power is defined by:

$$\mathfrak{D} = \frac{1}{f} = (n_i - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

with R_1 and R_2 the radii of the thin lens and n_i the index of refraction of the glass. Because for two lenses in contact, the focal length is given by:

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2},$$

the combined power of the two lenses is the sum of the individual powers:

$$\mathfrak{D} = \mathfrak{D}_1 + \mathfrak{D}_2$$

A lens of focal length $f_1 = 10\,\text{cm}$ has a dioptric power $\mathfrak{D}_1 = 10\text{D}$. If it is in contact with a negative lens of dioptric power $\mathfrak{D}_2 = -10\text{D}$, the resulting power is $\mathfrak{D} = 0$, equivalent to a parallel sheet of glass.

3.4.5 Eyeglasses

The eye can suffer from imperfections as seen in Figure 3.5.5. We discuss the most common imperfections and their solutions.

a. **Myopia or Nearsightedness**. A myopic eye has too short focal distances (has too high power) so that distant objects are focused in front of the retina by the unaccommodated eye. The far point is thus not at infinity, but closer. This can be corrected by a negative lens. Suppose the far point is at 2 m. If the concave lens makes a virtual image of a distant object at distance 2 m in front of the cornea, the unaccommodated eye can see it clearly. The lens Law (2.5.44), with $s_o = -\infty$ implies then $f_i = s_i = -2\text{ m}$. Hence the required power of the lens is:

$$\mathfrak{D} = \frac{1}{f} = -0.5 \text{ Diopter}.$$

The lens is best put in the front focal plane of the unaccomodated eye, i.e. at approximately 16 mm in front of the cornea. (Since the distance of the retina to the eye lens is roughly 22 mm and the refractive index of the vitrous humor is 1.337, this implies a focal distance in air of the unaccomodated eye of $22/1.337 \approx 16$ mm). The reason is that in this case the magnification of the eye and the negative lens together are the same as for the uncorrected eye. To see this, draw a ray from the top of the object through the centre of the negative lens. This will then be made parallel to the optical axis by the eye lens; the distance of this ray to the optical axis is the image size on the retina. This ray will end up at the same point of the retina as when the negative lens is taken out, because it is not refracted by this lens.

Contact lenses are very close to the eye lens and hence the total power of the eye with a contact lens is simply the sum of the power of the eye and the contact lens.

b. **Hyperopia or Farsightedness**. In this case a distant object is imaged by the unaccommodated eye behind the retina, i.e. the back focal distance of the unaccommodated eye is larger than the depth of the eye. Close objects can not be imaged on the retina, hence the near point is relatively far from the cornea. In order to bend the rays more, a positive lens is placed in front of the eye. Suppose that a hyperopic eye has near point at distance 125 cm. For an object at $s_o = -25$ cm to have virtual image at $s_i = -125$ cm, so that it can be seen, the focal length must satisfy

$$\frac{1}{f} = -\frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{0.25} - \frac{1}{1.25} = \frac{1}{0.31},$$

hence the power must be $\mathfrak{D} = 1/f = +3.2$ Diopter.

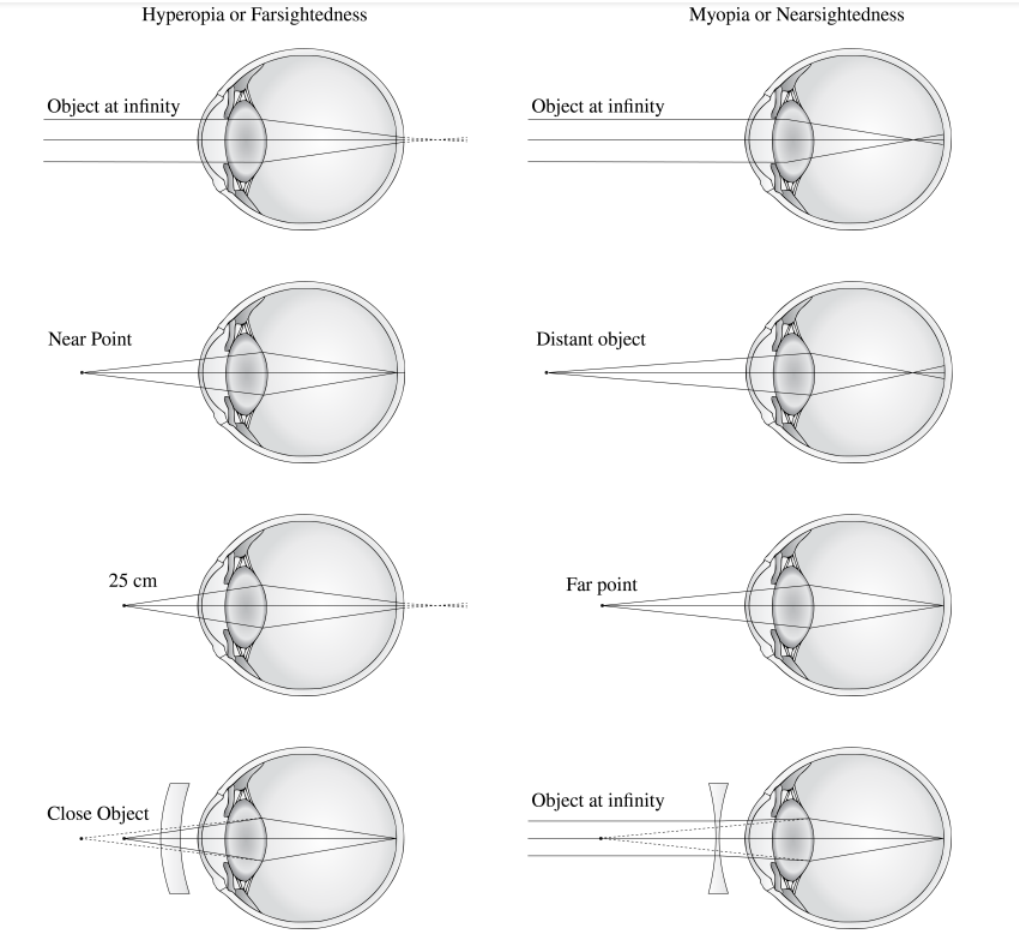


Figure 3.5.5: Correction of farsighted (left) and nearsighted (right) eye (adapted from Wikimedia Commons by Gumenyuk I.S. / CC BY-SA 4.0).

c. **Presbyopia**. This is the lack of accommodation of the eye as is common in people over 40. It results in an increase in the distance between the near point and the retina. This defect affects all images. Presbyopia is usually corrected by glasses with progressive correction, the upper part of glass used for distance vision and the lower part for near vision.

d. **Astigmatism**. In this case the focal distances for two directions perpendicular to the optical axis are different. It is attributed to a lack of symmetry of revolution of the cornea. This is compensated by using glasses which themselves are astigmatic.

3.4.6 New Correction Technique

In recent years, to correct eye defects such as myopia and astigmatism, technology has been developed to change the local curvatures of the surface of the cornea using an excimer laser. The laser is computer-controlled and causes photo-ablation in parts of the cornea.

3.6: Magnifying Glasses

A magnifying glass causes an image on the retina which is larger than without the magnifier. In principle, the image on the retina can be increased by simply bringing the object closer to the eye (reduce $|s_o|$ at fixed s_i). But $|s_o|$ can not be smaller than the near point d_o, which we take here to be 25 cm. It is desirable to use a lens that makes a magnified erect image at a distance to the eye greater than d_o. This can be achieved by a positive lens with the object closer to the lens than the first focal point, thereby producing a magnified virtual image. An example is given in Figure $3.6.1$.

Figure $3.6.1$: Example of a positive lens used as a magnifying glass (picture taken by A.J.L. Adam / CC-BY-SA 4.0).

3.5.1 Magnifying Power

The **magnifying power** MP or angular magnification M_a is defined as the ratio of the size of the retinal image obtained with the instrument and the size of the retinal image as seen by the unaided eye at normal viewing distance d_o. To estimate the size of the retinal image, we compare in both cases where **the chief ray through the top of the object and the centre of the pupil of the eye hits the retina**. Since the distance between the eye lens and the retina is fixed, the ratio of the image size on the retina for the eye with and without magnifying glass is:

$$\mathrm{MP} = \frac{\alpha_a}{\alpha_u},$$

where α_a and α_u are the angles between the optical axis and the chief rays for the aided and the unaided eye, respectively, as shown in Figure $3.6.2$. Working with these angles instead of distances is in particular useful when the virtual image of the magnifying glass is at infinity. Using $\alpha_a \approx y_i/L$ and $\alpha_u \approx y_0/d_0$ with y_i and y_0 positive and L the positive distance from the image to the eye (with as requirement : $L \geq d_o$), we find

$$\mathrm{MP} = \frac{y_i d_0}{y_0 L}.$$

Since $s_i < 0$ and $f_o < 0$ we have,

$$\frac{y_i}{y_o} = \frac{s_i}{s_o} = 1 + \frac{s_i}{f_o},$$

where we used the lens equation for the magnifying glass. We have $s_i = -|s_i| = -(L - \ell)$, where C is the distance between the magnifying glass and the eye. Hence, \((3.6.2)\) becomes:

$$\begin{aligned} \mathrm{MP} &= \frac{d_0}{L}\left[1+\frac{L-\iota}{|f_o|}\right] \\ &= \frac{d_0}{L}[1+\mathfrak{D}(L-\iota)], \end{aligned}$$

where $\mathfrak{D}$ is the power of the magnifying glass.

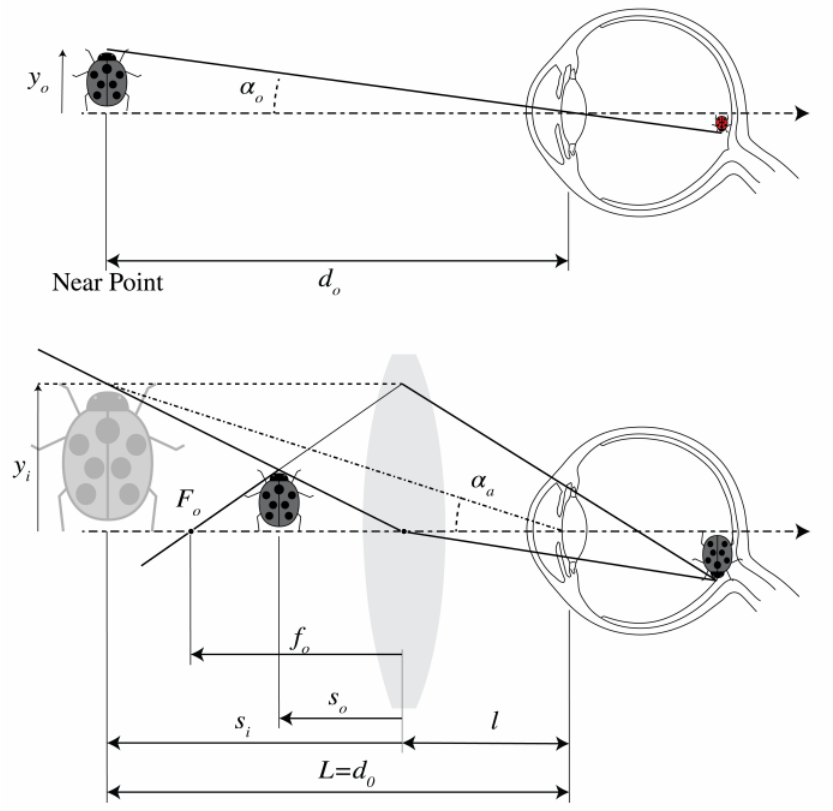


Figure 3.6.2: An unaided view (top) and an aided view using a magnifier.

1. $\ell = |f_o|$: the magnifying power is then $\mathrm{MP} = d_0\mathfrak{D}$.

2. $\ell = 0$: hence $L = d_0$ is smallest while MP is maximum:

$$\mathrm{MP}|_{\ell=0, L=d_0} = d_0\mathfrak{D}+1$$

3. The object is at the focal point of the magnifier $(s_0 = f_o)$, so that the virtual image is at infinity $(L=\infty)$ and hence

$$\mathrm{MP}|_{L=\infty} = d_0\mathfrak{D},$$

for every distance l between the eye and the magnifying glass. The rays are parallel, so that the eye views the object in a relaxed way. This is the most common use of the magnifier.

In practice $d_0\mathfrak{D} = d_o/\,|f_o|$ is much larger than 1 , so that MP is similar in the three cases.

3.5.2 Nomenclature

Normally magnifiers are expressed in terms of the magnifying power when $L=\infty$ (case 3 above). For example, a magnifier with a power of 10 Diopter has a MP equal to 2.5 or $2.5\times$. In other words, the image is 2.5 times larger than it would be if the object would be at the near point of the unaided eye.

3.7: Eyepieces

An **eyepiece** or **ocular** is a magnifier used before the eye at the end of an other optical instrument such as a microscope and a telescope. The eye looks into the ocular and the ocular "looks" into the optical instrument. The ocular provides a magnified virtual image of the image produced by the optical instrument. Similar to the magnifying glass, the virtual image should preferably be at or near infinity to be viewed by a relaxed eye. Several types of eye pieces exist and most of them are made out of two lenses: 1 . the field lens, which is the first lens in the ocular; 2 . the eye-lens, which is closest to the eye at a fixed distance called the **eye relief**. The aperture of the eyepiece is controlled by a field stop. An example is given in Figure 3.7.1.

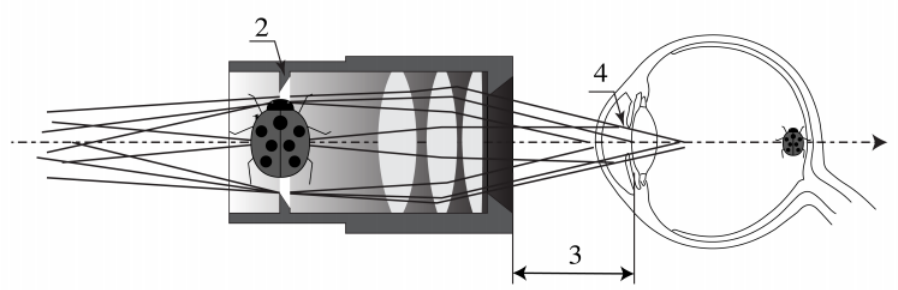


Figure 3.7.1: Example of an eyepiece; 1) real image 2) field diaphragm 3) eye relief 4) eye pupil (adapted from Wikimedia Commons by Tamas-flex/CC BY-SA 3.0).

3.8: The Compound Microscope

A magnifier alone can provide very high magnification only at the cost of intolerable aberrations. The **compound microscope** is a magnifier of close objects with a high angular magnification, generally more than $30\times$. It was invented by Zacharias Janssen in Middelburg in 1590 (this claim is disputed). The first element of the compound microscope is an objective (in Figure 3.8.1 a simple positive lens) which makes a real, inverted and magnified image of the object in the front focal plane of an eyepiece (where there is also the field stop). The eyepiece will make a virtual image at infinity, as explained above. The magnifying power of the entire system is the product of the transverse linear magnification of the objective M_T and the angular magnification of the eyepiece M_{Ae} :

$$\mathrm{MP} = M_T M_{Ae}.$$

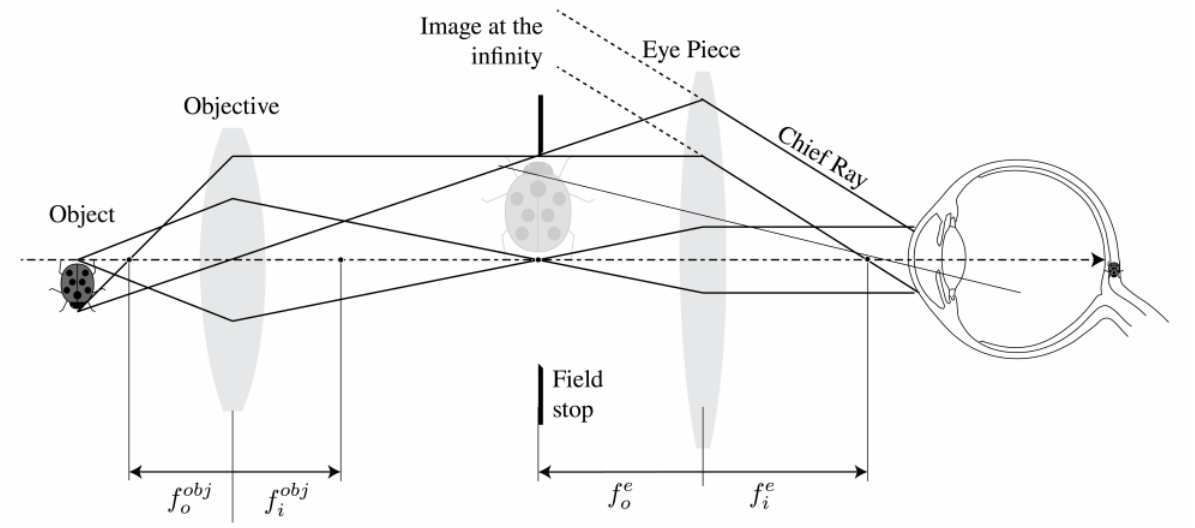


Figure 3.8.1: Simple compound microscope. The objective forms a real image of a nearby object. The eyepiece enlarges this intermediate image. The final image can be bigger than the barrel of the device, since it is virtual.

According to (2.5.51): $M_T = -x_i/f_i^{obj}$, where x_i is the distance of the image made by the objective to its back focal plane with focal distance f_i^{obj} . We have $x_i = L$ which is the tube length, i.e. the distance between the second focal point of the objective and the first focal point of the eyepiece. The tube length is standardised at $16\ \mathrm{cm}$. Furthermore, according to (3.5.5), the angular magnification is for a virtual image at infinity $M_{Ae} = d_o/f_i^e$. Hence, we obtain:

$$\mathrm{MP} = \frac{-x_i}{f_i^{obj}} \frac{d_o}{f_i^e} = \frac{-16}{f_0} \frac{25}{f_e},$$

with the standard near-point $d_o = 25\ \mathrm{cm}$. As an example, an Amici objective gives $40\times$ and combined with a $10\times$ eye piece one gets $MP = 400$.

The **Numerical aperture** of a microscope is a measure of the capability to gather light from the object. It is defined by:

$$\mathrm{NA} = n_{im} \sin\theta_{\max}$$

with n_{im} the refractive index of the immersing medium, usually air, but it could be water or oil, and $\theta_{\max}$, the half-angle of the maximum cone of light accepted by the lens. The numerical aperture is the second number etched in the barrel of the objective. It ranges from 0.07 (lowpower objectives) to 1.4 for high-power objectives. Note that it depends on the object distance. In Chapter 7 it will be explained that NA is for the a given object distance proportional to the resolving power, which is the minimum transverse distance between two object points that can be resolved in the image.

3.9: The Telescope

A telescope enlarges the retinal image of a distant object. Like a compound microscope, it is also composed of an objective and an eyepiece as seen in Figure 3.9.1. The object in this figure is at a large but finite distance; therefore, an image is formed by the objective just after its second focal point. The eyepiece makes a virtual magnified image, to be viewed with a relaxed eye. Therefore, the intermediary image of the objective must be within the focal length f_i^e from the eyepiece. The final image is inverted.

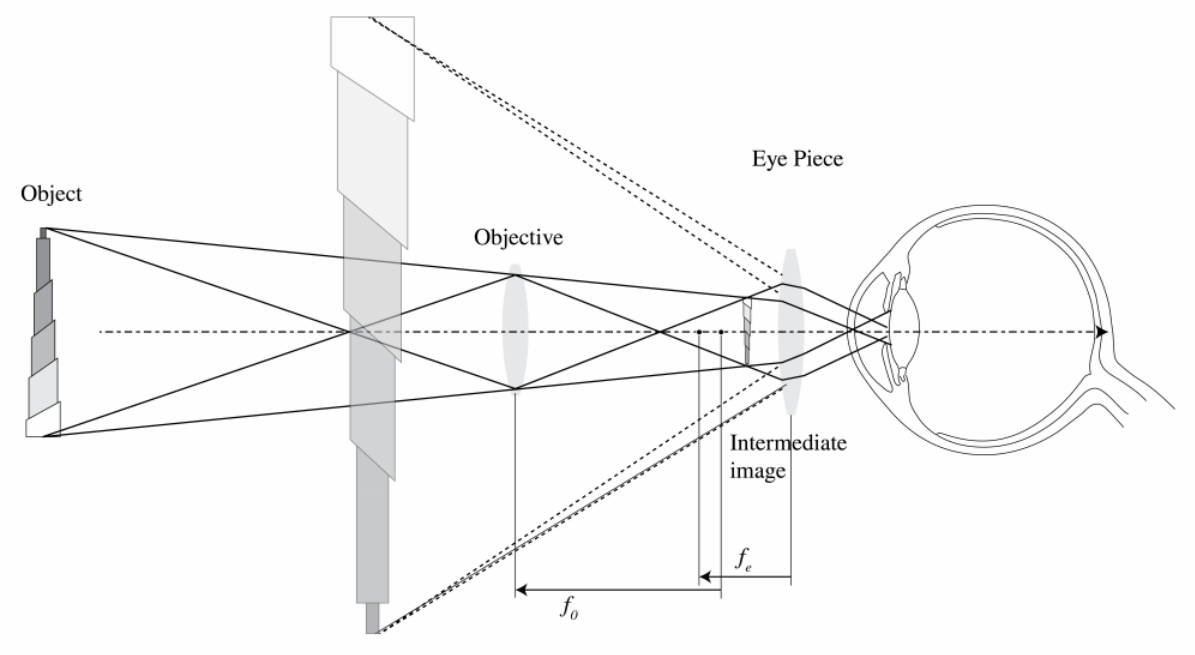


Figure 3.9.1: Keplerian astronomical telescope.

As seen earlier, the angular magnification is: $\mathrm{MP} = \alpha_a / \alpha_u$ where α_u is the half angle of the cone of light that would be collected without telescope and α_a is the half angle of the apparent cone of rays coming from the virtual image of the eye piece. From triangles $F_o^{obj} BC$ and $F_i^e DE$ in Figure 3.9.1 we see that

$$\mathrm{MP} = -\frac{f_i^{obj}|}{f_i^e}$$

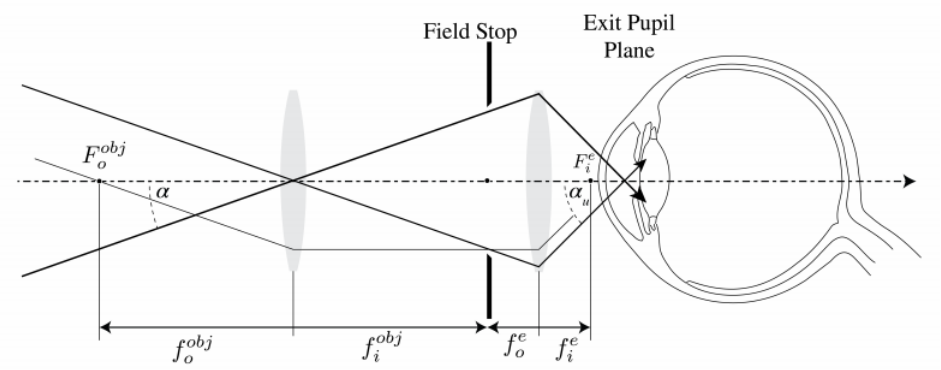


Figure 3.9.2: Ray angles for a telescope

CHAPTER OVERVIEW

4: Polarization

4.1: What you should know and be able to do after studying this chapter
4.2: Polarisation States and Jones Vectors
4.3: Creating and Manipulating Polarisation States
4.4: How to verify whether a Jones Matrix is a Linear Polariser or a Wave Plate?
4.5: Decompose Elliptical Polarisation into Linear and Circular States

---

4.1: What you should know and be able to do after studying this chapter

- Understand how different states of polarisation are caused by the ratio of the amplitudes and the difference in phase between two orthogonal components of the electric field.
- Know that elliptical polarisation is the most general state of polarisation.
- Know that linear polarisation and circular polarisation are special cases.
- Know how to compute the degree of polarisation.
- Be able to work with Jones vectors and Jones matrices.
- Know how birefringence is exploited to create wave plates and know the tyes of wave plates.
- Know how to rotate a state of linear polarisation over a given angle.
- Know how to change linear polarisation into circular polarisation and conversely.
- Be able to show that elliptical polarisation can be written as the sum of two orthogonal linear polarisations and as the sum of two circular polarisations.

4.2: Polarisation States and Jones Vectors

We have seen in Chapter 1 that light is an electromagnetic wave which satisfies Maxwell's equations and the wave equation derived therefrom. Since the electric field is a vector that oscillates in a certain direction, we say that the wave has a certain polarisation. In this chapter we look at the different types of polarisation and how the polarisation of a light beam can be manipulated.

We start with Eqs. (1.6.9) and (1.6.11) which show that the electric field $\mathcal{E}(\mathbf{r},t)$ of a plane wave is always perpendicular to the direction of propagation, which is the direction of the wave vector k. Let the wave propagate in the z-direction:

$$\mathbf{k}=\begin{pmatrix}0\\0\\k\end{pmatrix}$$

Then the electric field vector does not have a z-component and hence the real electric field at z and at time t can be written as

$$\mathcal{E}(z,t)=\begin{pmatrix}\mathcal{A}_x\cos(kz-\omega t+\varphi_x)\\ \mathcal{A}_y\cos(kz-\omega t+\varphi_y)\\ 0\end{pmatrix}.$$

where $\mathcal{A}_x$ and $\mathcal{A}_y$ are positive amplitudes and φ_x,φ_y are the phases of the electric field components. While k and ω are fixed in this case, we can vary $\mathcal{A}_x,\mathcal{A}_y,\varphi_x$ and φ_y. This degree of freedom is why different states of polarisation exist: **the state of polarisation is determined by the ratio of the amplitudes and by the phase difference $\varphi_y-\varphi_x$ between the two orthogonal components of the light wave**. Varying the quantity $\varphi_y-\varphi_x$ means that we are 'shifting' $\mathcal{E}_y(\mathbf{r},t)$ with respect to $\mathcal{E}_x(\mathbf{r},t)$. Consider the electric field in a fixed plane $z=0$:

$$\begin{aligned}\begin{pmatrix}\mathcal{E}_x(0,t)\\ \mathcal{E}_y(0,t)\end{pmatrix}&=\begin{pmatrix}\mathcal{A}_x\cos(-\omega t+\varphi_x)\\ \mathcal{A}_y\cos(-\omega t+\varphi_y)\end{pmatrix}\\&=\mathrm{Re}\left\{\begin{pmatrix}E_x(0)\\E_y(0)\end{pmatrix}e^{-i\omega t}\right\}\\&=\mathrm{Re}\left\{\begin{pmatrix}\mathcal{A}_xe^{i\varphi_x}\\ \mathcal{A}_ye^{i\varphi_y}\end{pmatrix}e^{-i\omega t}\right\}.\end{aligned}$$

The complex vector

$$\mathbf{J}=\begin{pmatrix}E_x(0)\\E_y(0)\end{pmatrix}=\begin{pmatrix}\mathcal{A}_xe^{i\varphi_x}\\ \mathcal{A}_ye^{i\varphi_y}\end{pmatrix}$$

is called the **Jones vector**. It is used to characterise the polarisation state. Let us see how, at a fixed position in space, the electric field vector behaves as a function of time for different choices of $\mathcal{A}_x,\mathcal{A}_y$ and $\varphi_y-\varphi_x$.

a) **Linear polarisation**: $\varphi_y-\varphi_x=0$ or $\varphi_y-\varphi_x=\pi$. When $\varphi_y-\varphi_x=0$ we have

$$\mathbf{J}=\begin{pmatrix}\mathcal{A}_x\\ \mathcal{A}_y\end{pmatrix}e^{i\varphi_x}$$

Equality of the phases: $\varphi_y=\varphi_x$, means that the field components $\mathcal{E}_x(z,t)$ and $\mathcal{E}_y(z,t)$ are in phase: when $\mathcal{E}_x(z,t)$ is large, $\mathcal{E}_y(z,t)$ is large, and when $\mathcal{E}_x(z,t)$ is small, $\mathcal{E}_y(z,t)$ is small. We can write

$$\begin{pmatrix}\mathcal{E}_x(0,t)\\ \mathcal{E}_y(0,t)\end{pmatrix}=\begin{pmatrix}\mathcal{A}_x\\ \mathcal{A}_y\end{pmatrix}\cos(\omega t-\varphi_x),$$

which shows that for $\varphi_y-\varphi_x=0$ the electric field simply oscillates in one direction given by real the vector $\mathcal{A}_x\hat{\mathbf{x}}+\mathcal{A}_y\hat{\mathbf{y}}$. See Figure 4.2.1a.

If $\varphi_y-\varphi_x=\pi$ we have

$$\mathbf{J}=\begin{pmatrix}\mathcal{A}_x\\ -\mathcal{A}_y\end{pmatrix}e^{i\varphi_x}$$

In this case $\mathcal{E}_x(z,t)$ and $\mathcal{E}_y(z,t)$ are out of phase and the electric field oscillates in the direction given by the real vector $\mathcal{A}_x\hat{\mathbf{x}} - \mathcal{A}_y\hat{\mathbf{y}}$.

b) **Circular polarisation**: $\varphi_y - \varphi_x = \pm\pi/2, \mathcal{A}_x = \mathcal{A}_y$. In this case the Jones vector is:

$$\mathbf{J} = \begin{pmatrix} 1 \\ \pm i \end{pmatrix} \mathcal{A}_x e^{i\varphi_x}$$

The field components $\mathcal{E}_x(z,t)$ and $\mathcal{E}_y(z,t)$ are $\pi/2$ radians (90 degrees) out of phase: when $\mathcal{E}_x(z,t)$ is large, $\mathcal{E}_y(z,t)$ is small, and when $\mathcal{E}_x(z,t)$ is small, $\mathcal{E}_y(z,t)$ is large. We can write for $z=0$ and with $\varphi_x = 0$:

$$\begin{aligned} \begin{pmatrix} \mathcal{E}_x(0,t) \\ \mathcal{E}_y(0,t) \end{pmatrix} &= \begin{pmatrix} \mathcal{A}_x \cos(-\omega t) \\ \mathcal{A}_x \cos(-\omega t \pm \pi/2) \end{pmatrix} \\ &= \mathcal{A}_x \begin{pmatrix} \cos(\omega t) \\ \pm \sin(\omega t) \end{pmatrix}. \end{aligned}$$

The electric field vector moves in a circle. When for an observer looking towards the source, the electric field is rotating anti-clockwise, the polarisation is called **left-circularly polarised** ($+$ sign in (4.2.9)), while if the electric vector moves clockwise, the polarisation is called **right-circularly polarised** (- sign in (4.2.9)).

c) **Elliptical polarisation**: $\varphi_y - \varphi_x = \pm\pi/2, \mathcal{A}_x$ **and** $\mathcal{A}_y$ **arbitrary**. The Jones vector is:

$$\mathbf{J} = \begin{pmatrix} \mathcal{A}_x \\ \pm i\mathcal{A}_y \end{pmatrix} e^{i\varphi_x}$$

In this case we get instead of (4.2.9) (again taking $\varphi_x = 0$):

$$\begin{pmatrix} \mathcal{E}_x(0,t) \\ \mathcal{E}_y(0,t) \end{pmatrix} = \begin{pmatrix} \mathcal{A}_x \cos(\omega t) \\ \pm \mathcal{A}_y \sin(\omega t) \end{pmatrix}.$$

which shows that the electric vector moves along an ellipse with major and minor axes parallel to the x - and y-axis. When the $+$sign applies, the field is called left-elliptically polarised, otherwise it is called right-elliptically polarised.

d) **Elliptical polarisation**: $\varphi_y - \varphi_x =$ **anything else**, $\mathcal{A}_x$ **and** $\mathcal{A}_y$ **arbitrary**. The Jones vector is now the most general one:

$$\mathbf{J} = \begin{pmatrix} \mathcal{A}_x e^{i\varphi_x} \\ \mathcal{A}_y e^{i\varphi_y} \end{pmatrix}$$

It can be shown that the electric field vector moves always along an ellipse. The exact shape and orientation of this ellipse of course varies with the difference in phase $\varphi_y - \varphi_x$ and the ratio of the amplitude $\mathcal{A}_x, \mathcal{A}_y$ and, except when $\varphi_y - \varphi_x = \pm\pi/2$, the major and minor axis of the ellipse are not parallel to the x - and y-axis. See Figure 4.2.1 c.

Remarks.

1. Frequently the Jones vector is normalised such that

$$|J_x|^2 + |J_y|^2 = 1.$$

The normalized vector represents of course the same polarisation state as the unnormalised one. In general, multiplying the Jones vector by a complex number does not change the polarisation state. If we multiply for example by $e^{i\theta}$, this has the same result as changing the instant that $t=0$, hence it does not change the polarisation state. In fact:

$$\mathcal{E}(0,t) = \mathrm{Re}\left[e^{i\theta}\mathbf{J}e^{-i\omega t}\right] = \mathrm{Re}\left[\mathbf{J}e^{-i\omega(t-\theta/\omega)}\right]$$

2. We will show in section 6.2 that a general time-harmonic electromagnetic field, is a superposition of plane waves with wave vectors of the same length determined by the frequency of the wave but with different directions. An example is the electromagnetic field near the focal plane of a lens. There is then no particular direction of propagation to which the electric field should be perpendicular; in other words, there is no obvious choice for a plane in which the electric field oscillates as function of time. Nevertheless, for every point in space, such a plane exists, but its orientation varies in general with position. Furthermore, the electric field in a certain point moves along an ellipse in the corresponding plane, but the shape of the ellipse and the orientation of its major axis can be arbitrary. We can conclude that in any point of an arbitrary time-harmonic electromagnetic field, the electric

(and in fact also the magnetic) field vector prescribes as function of time an ellipse in a certain plane which depends on position. In this chapter we only consider the field and polarisation state of a single plane wave.

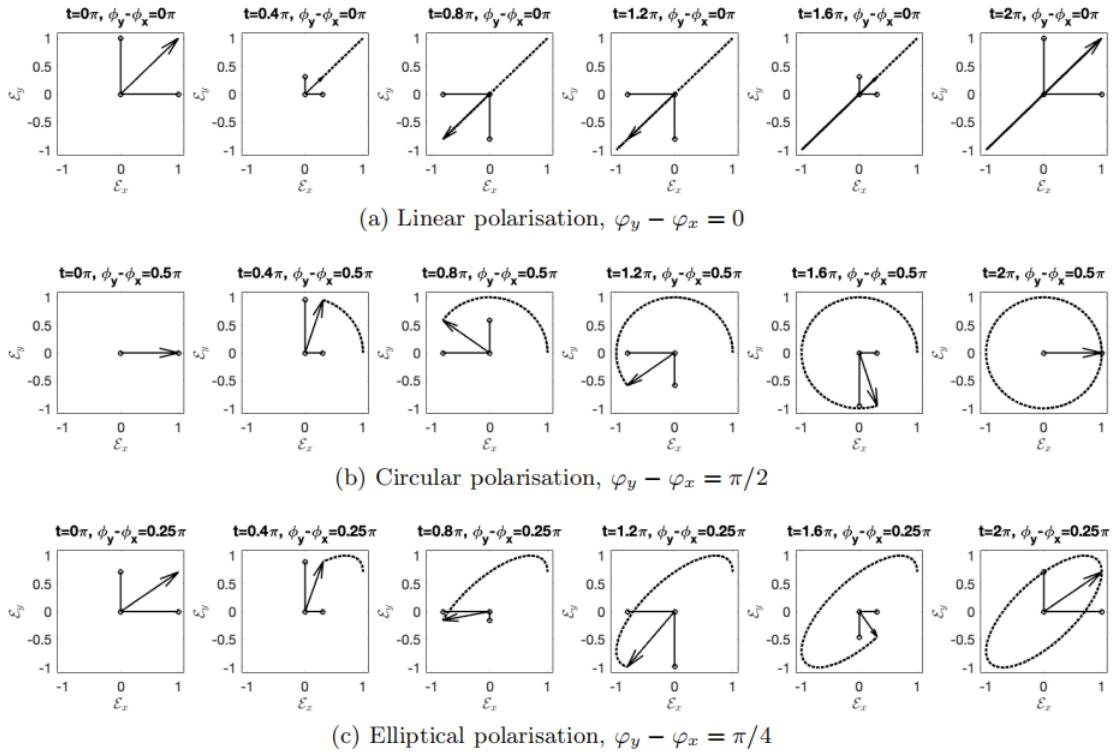


Figure 4.2.1: Illustration of different types of polarisation. Top: linear polarisation; middle: circular polarisation; bottom: elliptical polarisation. The horizonal and vertical arrows indicate the momentary field components $\mathcal{E}_x, \mathcal{E}_y$. The thick arrow indicates the vector $\mathcal{E}$. The black curve indicates the trajectory of $\mathcal{E}(t)$.

External source

KhanAcademy - Polarization of light, linear and circular: Explanation of different polarisation states and their applications.

4.3: Creating and Manipulating Polarisation States

We have seen how Maxwell's equations allow the existence of plane waves with many different states of polarisation. But how can we create these states, and how do these states manifest themselves?

Natural light often does not have a definite polarisation. Instead, the polarisation fluctuates rapidly with time. In order to turn such randomly polarised light into linearly polarised light in a certain direction, we must extinguish the light polarised in the perpendicular direction, so that the remaining light is linearly polarised along the required direction. One could do this by using light reflected under the Brewster angle (which extinguishes p-polarised light), or one could let light pass through a dichroic crystal, which is a material which absorbs light polarised perpendicular to its so-called optic axis. A third method is sending the light through a wire grid polariser, which consists of a metallic grating with sub-wavelength slits. Such a grating only transmits the electric field component that is perpendicular to the slits.

So suppose that with one of these methods we have obtained linearly polarised light. Then the question rises how the state of linear polarisation can be changed into circularly or elliptically polarised light. Or how the state of linear polarisation can be rotated over a certain angle. We have seen that the polarisation state depends on the ratio of the amplitudes and on the phase difference $\varphi_y - \varphi_x$ of the orthogonal components $\mathcal{E}_y$ and $\mathcal{E}_x$ of the electric field. Thus, to change linearly polarised light to some other state of polarisation, a certain phase shift (say $\Delta\varphi_x$) must be introduced to one component (say $\mathcal{E}_x$), and another phase shift $\Delta\varphi_y$ to the orthogonal component $\mathcal{E}_y$. We can achieve this with a **birefringent crystal**, such as calcite. What is special about such a crystal is that it has two refractive indices: light polarised in a certain direction experiences a refractive index of n_o, while light polarised perpendicular to it feels another refractive index n_e (the subscripts o and e stand for "ordinary" and "extraordinary"), but for our purpose we do not need to understand this terminology. The direction for which the refractive index is smallest (which can be either n_o or n_e) is called the **fast axis** because its phase velocity is largest, and the other direction is the **slow axis**. Because there are two different refractive indices, one can see double images through a birefringent crystal. The difference between the two refractive indices $\Delta n = n_e - n_o$ is called the **birefringence**.

Suppose $n_e > n_o$ and that the fast axis, which corresponds to n_o is aligned with $\mathcal{E}_x$, while the slow axis (which then has refractive index n_e) is aligned with $\mathcal{E}_y$. If the wave travels a distance d through the crystal, $\mathcal{E}_y$ will accumulate a phase $\Delta\varphi_y = \frac{2\pi n_e}{\lambda} d$, and $\mathcal{E}_x$ will accumulate a phase $\Delta\varphi_x = \frac{2\pi n_o}{\lambda} d$. Thus, after propagation through the crystal the phase difference $\varphi_y - \varphi_x$ has increased by

$$\Delta\varphi_y - \Delta\varphi_x = \frac{2\pi}{\lambda} d\,(n_e - n_o)\,.$$

Jones Matrices

By letting light pass through crystals of different thicknesses d, we can create different phase differences between the orthogonal field components, and this way we can create different states of polarisation. To be specific, let $\mathbf{J}$, as given by (4.1.4), be the Jones vector of the plane wave before the crystal. Then we have, for the Jones vector after the passage through the crystal:

$$\tilde{\mathbf{J}} = \mathcal{M}\mathbf{J},$$

where

$$\mathcal{M} = \begin{pmatrix} e^{\frac{2\pi i}{\lambda} d n_o} & 0 \\ 0 & e^{\frac{2\pi i}{\lambda} d n_e} \end{pmatrix} = e^{\frac{2\pi i}{\lambda} d n_o} \begin{pmatrix} 1 & 0 \\ 0 & e^{\frac{2\pi i}{\lambda} d (n_e - n_o)} \end{pmatrix}.$$

A matrix such as $\mathcal{M}$, which transfers one state of polarisation of a plane wave in another, is called a **Jones matrix**. Depending on the phase difference which a wave accumulates by traveling through the crystal, these devices are called **quarter-wave plates** (phase difference $\pi/2$), **half-wave plates** (phase difference π), or **full-wave plates** (phase difference 2π). The applications of these wave plates will be discussed in later sections.

Consider as example the Jones matrix which described the change of linear polarised light into circular polarisation. Assume that we have diagonally (linearly) polarised light, so that

$$J = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

We want to change it to circularly polarised light, for which

$$J = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix},$$

where one can check that indeed $\varphi_y - \varphi_x = \pi/2$. This can be done by passing the light through a crystal such that $\mathcal{E}_y$ accumulates a phase difference of $\pi/2$ with respect to $\mathcal{E}_x$. The transformation by which this is accomplished can be written as

$$\begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix}.$$

The matrix on the left is the Jones matrix describing the operation of a quarter-wave plate.

Another important Jones matrix is the **rotation matrix**. In the preceding discussion it was assumed that the fast and slow axes were aligned with the x - and y-direction (i.e. they were parallel to $\mathcal{E}_x$ and $\mathcal{E}_y$). Suppose now that the slow and fast axes of the wave plate no longer coincide with $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{y}}$, but rather with some other $\widehat{\mathbf{x}}'$ and $\widehat{\mathbf{y}}'$ as in Figure 4.3.1. In that case we apply a basis transformation: the electric field vector which is expressed in the $\widehat{\mathbf{x}}, \widehat{\mathbf{y}}$ basis should first be expressed in the $\widehat{\mathbf{x}}', \widehat{\mathbf{y}}'$ basis before applying the Jones matrix of the wave plate to it. After applying the Jones matrix, the electric field has to be transformed back from the $\widehat{\mathbf{x}}', \widehat{\mathbf{y}}'$ basis to the $\widehat{\mathbf{x}}, \widehat{\mathbf{y}}$ basis.

Let $\mathbf{E}$ be given in terms of its components on the $\hat{\mathbf{x}}, \hat{\mathbf{y}}$ basis:

$$\mathbf{E} = E_x \widehat{\mathbf{x}} + E_y \widehat{\mathbf{y}}.$$

To find the components $E_{x'}, E_{y'}$ on the $\widehat{\mathbf{x}}', \widehat{\mathbf{y}}'$ basis:

$$\mathbf{E} = E_{x'} \widehat{\mathbf{x}}' + E_{y'} \widehat{\mathbf{y}}',$$

we first write the unit vectors $\widehat{\mathbf{x}}'$ and $\widehat{\mathbf{y}}'$ in terms of the basis $\hat{\mathbf{x}}, \hat{\mathbf{y}}$ (see Figure 4.3.1)

$$\begin{aligned} \widehat{\mathbf{x}}' &= \cos\theta \widehat{\mathbf{x}} + \sin\theta \widehat{\mathbf{y}}, \\ \widehat{\mathbf{y}}' &= -\sin\theta \widehat{\mathbf{x}} + \cos\theta \widehat{\mathbf{y}}. \end{aligned}$$

By substituting (4.3.9) and (4.3.10) into (4.3.8) we find

$$\begin{aligned} \mathbf{E} &= E_{x'} \widehat{\mathbf{x}}' + E_{y'} \widehat{\mathbf{y}}' \\ &= E_{x'} (\cos\theta \widehat{\mathbf{x}} + \sin\theta \widehat{\mathbf{y}}) + E_{y'} (-\sin\theta \widehat{\mathbf{x}} + \cos\theta \widehat{\mathbf{y}}), \\ &= (\cos\theta E_{x'} - \sin\theta E_{y'}) \widehat{\mathbf{x}} + (\sin\theta E_x + \cos\theta E_y) \widehat{\mathbf{y}}. \end{aligned}$$

Comparing with (4.3.7) implies

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} E_{x'} \cos\theta - E_{y'} \sin\theta \\ E_{x'} \sin\theta + E_{y'} \cos\theta \end{pmatrix} = \mathcal{R}_\theta \begin{pmatrix} E_{x'} \\ E_{y'} \end{pmatrix},$$

where $\mathcal{R}_\theta$ is the rotation matrix over an angle θ in the anti-clockwise direction:

$$\mathcal{R}_\theta \equiv \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

That $\mathcal{R}(\theta)$ indeed is a rotation over angle θ in the anti-clockwise direction is easy to see by considering what happens when $\mathcal{R}_\theta$ is applied to the vector $(1, 0)^T$. Since $\mathcal{R}_\theta^{-1} = \mathcal{R}_{-\theta}$ we get:

$$\begin{pmatrix} E_{x'} \\ E_{y'} \end{pmatrix} = \mathcal{R}_{-\theta} \begin{pmatrix} E_x \\ E_y \end{pmatrix}.$$

This relationship expresses the components $E_{x'}, E_{y'}$ of the Jones vector on the $\hat{\mathbf{x}}', \widehat{\mathbf{y}}'$ basis, which is aligned with the fast and slow axes of the crystal, in terms of the components E_x and E_y on the original basis $\widehat{\mathbf{x}}, \widehat{\mathbf{y}}$. If the matrix $\mathcal{M}$ describes the Jones matrix as defined in (4.3.3), then the matrix M_θ for the same wave plate but with x' as slow and y' as fast axis, is, with respect to the $\widehat{\mathbf{x}}, \widehat{\mathbf{y}}$ basis, given by:

$$\mathcal{M}_\theta = \mathcal{R}_\theta \mathcal{M} \mathcal{R}_{-\theta}.$$

For more information on basis transformations, see Appendix F.

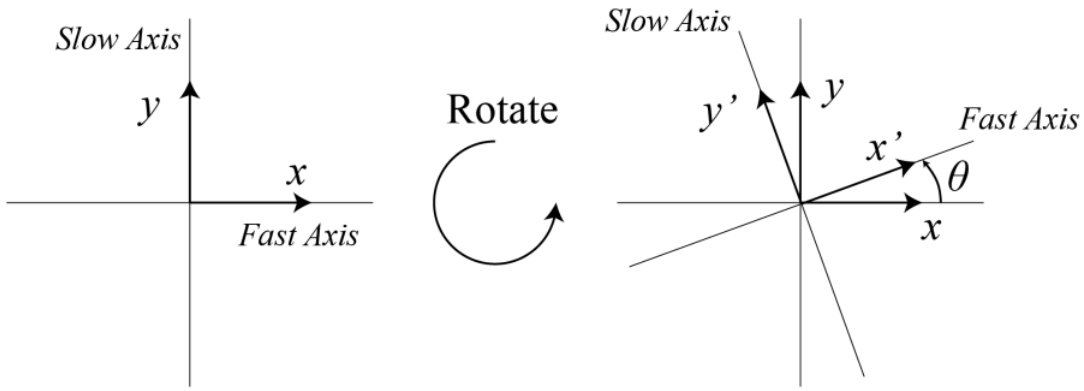


Figure 4.3.1: If the wave plate is rotated, the fast and slow axis no longer correspond to y and x. Instead, we have to introduce a new coordinate system y', x'.

4.2.2 Linear Polarisers

A polariser that only transmits horizontally polarised light is described by the Jones matrix:

$$\mathcal{M}_{LP} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Clearly, horizontally polarised light is completely transmitted, while vertically polarised light is not transmitted at all. More generally, for light that is polarised at an angle α, we get

$$\mathcal{M}_{\alpha} = \mathcal{M}_{LP} \begin{pmatrix} \cos\alpha \\ \sin\alpha \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \cos\alpha \\ \sin\alpha \end{pmatrix} = \begin{pmatrix} \cos\alpha \\ 0 \end{pmatrix}.$$

The amplitude of the transmitted field is reduced by the factor $\cos\alpha$, which implies that the intensity of the transmitted light is reduced by the factor $\cos^2\alpha$. This relation is known as **Malus' law**.

4.2.3 Degree of Polarisation

Natural light such as sun light is unpolarised. The instantaneous polarisation of unpolarised light fluctuates rapidly in a random manner. A linear polariser produces linear polarised light from unpolarised light.

Light that is a mixture of polarised and unpolarised light is called partially polarised. The **degree of polarisation** is defined as the fraction of the total intensity that is polarised:

$$\text{degree of polarisation } = \frac{I_{pol}}{I_{pol} + I_{\text{unpol}}}.$$

It follows from (4.3.17) that the intensity transmitted by a linear polariser when unpolarised light is passed incident, is the average value of $\cos^2\alpha$ namely $\frac{1}{2}$, times the incident intensity.

4.2.4 Quarter-Wave Plates

A quarter-wave plate introduces a phase shift of $\pi/2$, so its Jones matrix is

$$\mathcal{M}_{QWP} = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix},$$

because $\exp(i\pi/2) = i$. To describe the actual transmission through the quarter-wave plate, the matrix should be multiplied by some global phase factor, but because we only care about the phase difference between the field components, this global phase factor can be omitted without problem. The quarter-wave plate is typically used to convert linearly polarised light to elliptically polarised light and vice-versa. If the incident light is linearly polarised at angle α, the state of polarisation after the quater wave plate is

$$\begin{pmatrix} \cos\alpha \\ i\sin\alpha \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} \begin{pmatrix} \cos\alpha \\ \sin\alpha \end{pmatrix}.$$

In particular, if incident light is linear polarised under 45°, or equivalently, if the quarter wave plate is rotated over this angle, it will transform linearly polarised light into circularly polarised light (and vice versa).

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ i \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

A demonstration is shown.

4.2.5 Half-Wave Plates

A half-wave plate introduces a phase shift of π, so its Jones matrix is

$$\mathcal{M}_{HWP} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

because $\exp(i\pi) = -1$. An important application of the half-wave plate is to **change the orientation of linearly polarised light**. After all, what this matrix does is mirroring the polarisation state in the x-axis. Thus, if we choose our mirroring axis correctly (i.e. if we choose the orientation of the wave plate correctly), we can change the direction in which the light is linearly polarised arbitrarily. A demonstration is shown in. To give an example: the polarisation of a wave that is parallel to the x-direction, can be rotated over angle α by rotating the crystal such that the slow axis makes angle $\alpha/2$ with the x-axis. Upon propagation through the crystal, the fast axis gets an additional phase of π, due to which the electric vector makes angle α with the x-axis (see Figure 4.3.2).

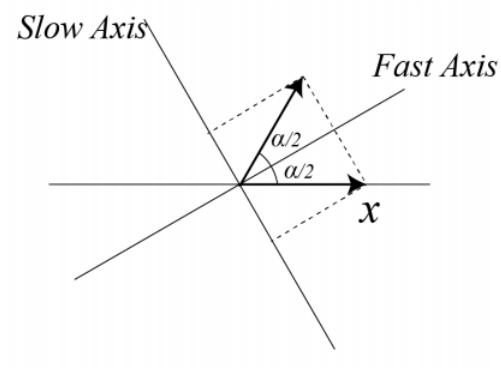


Figure 4.3.2: Rotation of horizontally polarised light over an angle α using a half-wave plate.

4.2.6 Full-Wave Plates

A full-wave plate introduces a phase difference of 2π, which is the same as introducing no phase difference between the two field components. So what can possibly be an application for a full-wave plate? We need to recall from Eq. ((4.3.1)) that the phase difference is 2π only for a particular wavelength. If we send through linearly (say vertically) polarised light of other wavelengths, these will become elliptically polarised, while the light with the correct wavelength λ_0 will stay vertically polarised. If we then let all the light pass through a horizontal polariser, the light with wavelength λ_0 will be completely extinguished, while the light of other wavelengths will be able to pass through at least partially. Therefore, **full-wave plates can be used to filter out specific wavelengths of light**.

4.4: How to verify whether a Jones Matrix is a Linear Polariser or a Wave Plate?

If the direction of either the slow or fast axis is given, it is easy to write down the Jones matrix of a birefringent plate. Similarly, for a linear polariser it is simple to write down the Jones matrix if one knows the direction in which the polariser absorbs or transmits all the light. But suppose that you are given an arbitrary complex $(2,2)$-matrix:

$$\mathcal{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

How can one determine whether this matrix corresponds to a linear polariser or to a wave plate? Note that the elements of a Jones matrix are in general complex.

1. **Linear Polariser** The matrix corresponds to a linear polariser if there is a real vector which remains invariant under $\mathcal{M}$ and the real vector that orthogonal to this vector is mapped to zero. In other words, there must be an orthogonal basis of **real** eigenvectors and one of the eigenvalues must be 1 and the other 0 . Hence, to check that a given matrix corresponds to a linear polariser, one should verify that one eigenvalue is 1 and the other is 0 and furthermore that the eigenvectors are **real** vectors. It is important to check that the eigenvectors are indeed real because if they are not, they do not correspond to particular polarisation directions.
2. **Wave plate** To show that a matrix corresponds to a wave plate, there should exist two real orthogonal eigenvectors with, in general, complex eigenvalues of modulus 1. In fact, one of the eigenvectors corresponds to the fast axis with refractive index n_1, say, and the other to the slow axis with refractive index n_2, say. The eigenvalues are then

   $$e^{ikn_1 d}, \quad e^{ikn_2 d},$$

   where d is the thickness of the plate and k is the wave number. Hence to verify that a $(2,2)$-matrix corresponds to a wave plate, one has to compute the eigenvalues and check that these have modulus 1 and that the corresponding eigenvectors are real vectors and orthogonal.

4.5: Decompose Elliptical Polarisation into Linear and Circular States

Any elliptical polarisation state can be written as the sum of two perpendicular linear polarised states:

$$J = \begin{pmatrix} c\mathcal{A}_x e^{i\varphi_x} \\ \mathcal{A}_y e^{i\varphi_y} \end{pmatrix} = \mathcal{A}_x e^{i\varphi_x} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \mathcal{A}_y e^{i\varphi_y} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Alternatively, any elliptical polarisation state can be written as the sum of two circular polarisation states, one right- and the other left-circular polarised:

$$J = \begin{pmatrix} \mathcal{A}_x e^{i\varphi_x} \\ \mathcal{A}_y e^{i\varphi_y} \end{pmatrix} = \frac{1}{2}\left(\mathcal{A}_x e^{i\varphi_x} - i\mathcal{A}_y e^{i\varphi_y}\right) \begin{pmatrix} 1 \\ i \end{pmatrix} + \frac{1}{2}\left(\mathcal{A}_x e^{i\varphi_x} + i\mathcal{A}_y e^{i\varphi}\right) \begin{pmatrix} 1 \\ -i \end{pmatrix}$$

We conclude that to study what happens to elliptic polarisation, it suffices to consider two orthogonal linear polarisations, or, if that is more convenient, left- and right-circular polarised light. In a birefringent material two linear polarisations, namely the one parallel to the o-axis and the one parallel to the e-axis, each propagate with their own refractive index. To predict what happens to an arbitrary linear polarisation state which is not aligned to either of these axes, or more generally what happens to an elliptical polarisation state, we write this polarisation as a linear combination of o- and e-polarisation states, i.e. we expand the field on the o- and e-basis.

In sugar, it are the left- and right-circular polarisation states which each propagate with their own refractive index. Therefore sugars are said to be **circular birefringent**. To see what happens to an arbitrary elliptical polarisation state in such a material, the incident light is best be written as linear combination of left-and right-circular polarisations.

External sources in recommended order

1. Double Vision - Sixty Symbols: Demonstration of double refraction by a calcite crystal due to birefringence.
2. MIT OCW - Linear Polarizer: Demonstration of linear polarizers and linear polarisation.
3. MIT OCW - Polarization Rotation Using Polarizers: Demonstration of polarisation rotation using linear polarisers.
4. Demonstration of a QuarterWavePlate by Andrew Berger.
5. MIT OCW - Quarter-wave Plate: Demonstration of the quarter-wave plate to create elliptical (in particular circular) polarisation.
6. Demonstration of a HalfWavePlate by Andrew Berger.
7. MIT OCW - Half-wave Plate: Demonstration of the half-wave plate.

CHAPTER OVERVIEW

5: Interference and Coherence

5.1: What you should know and be able to do after studying this chapter
5.2: Introduction
5.3: Interference of Monochromatic Fields of the Same Frequency
5.4: Coherence
5.5: Temporal Coherence and the Michelson Interferometer
5.6: Longitudinally and Transversely Separated Points
5.7: Spatial Coherence and Young's Experiment
5.8: Increase of Spatial Coherence by Propagation
5.9: Stellar Interferometry
5.10: Fringe Visibility
5.11: Interference and polarisation

5.1: What you should know and be able to do after studying this chapter

- Understand time coherence and spatial coherence.
- Know how the degree of time coherence can be measured with a Michelson interferometer.
- Understand the connection between the time coherence and the frequency bandwidth.
- Know how the spatial coherence of two points can be measured with Young's two-slit experiment
- Understand that spatial coherence increases by propagation.
- Understand how the size of a incoherent source such as a star can be derived from measuring the spatial coherence at large distance from the source.
- Know the definition of fringe visibility.
- Know and understand the three Laws of Fresnel-Arago.

5.2: Introduction

Although the model of geometrical optics helps us to design optical systems and explains many phenomena, there are also phenomena that require a more elaborate model. For example, interference fringes observed in Young's double-slit experiment or the Arago spot indicate that light is more accurately modelled as a wave. From a previous course you may remember that the condition for interference maxima in the double-slit experiment is that the difference in path lengths to the slits is an integer multiple of the wavelength:

$$d \sin \theta = m\lambda,$$

where d is the distance between the two slits, θ is the angle of the propagation direction of the light, m is an integer, and λ is the wavelength of the light. For sufficiently narrow slits, the Huygens-Fresnel principle says that each slit acts as a point source which radiates spherical waves.

In this chapter we will find more results that follow from the wave model of light. It will be shown that the extent to which light can show interference depends on a property called coherence. In the largest part of the discussion we will assume that all light has the same polarisation, so that we can treat the fields as scalars. In the last part we will look at how polarisation affects interference, as is described by the Fresnel-Arago laws.

It is very much worth noting that the concepts of interference and coherence are not just restricted to optics. Since quantum mechanics dictates that particles have a wave-like nature, interference and coherence also play a role in e.g. solid state physics and quantum information.

External sources in recommended order

- KhanAcademy - Interference of light waves: Playlist on wave interference at secondary school level.
- Yale Courses - 18. Wave Theory of Light

5.3: Interference of Monochromatic Fields of the Same Frequency

Let us first recall the basic concepts of interference. What causes interference is the fact that light is a wave, which means that it not only has an **amplitude** but also a **phase**. Suppose for example we evaluate a time-harmonic field in two points

$$\mathcal{U}_1(t) = \cos(\omega t), \quad \mathcal{U}_2(t) = \cos(\omega t + \varphi).$$

Here φ denotes the phase difference between the fields at the two points. If $\varphi = 0$, or φ is a multiple of 2π, the fields are **in phase**, and when they are added they interfere **constructively**

$$\mathcal{U}_1(t) + \mathcal{U}_2(t) = \cos(\omega t) + \cos(\omega t + 2m\pi) = 2\cos(\omega t).$$

However, when $\varphi = \pi$, or more generally $\varphi = \pi + 2m\pi$, for some integer m, then the waves are **out of phase**, and when they are superimposed, they interfere **destructively**.

$$\begin{aligned} \mathcal{U}_1(t) + \mathcal{U}_2(t) &= \cos(\omega t) + \cos(\omega t + \pi + 2m\pi) \\ &= \cos(\omega t) - \cos(\omega t) \\ &= 0. \end{aligned}$$

We can sum the two fields for arbitrary φ more conveniently using complex notation:

$$\mathcal{U}_1(t) = \mathrm{Re}\left[e^{-i\omega t}\right], \quad \mathcal{U}_2(t) = \mathrm{Re}\left[e^{-i\omega t}e^{-i\varphi}\right].$$

Adding gives

$$\begin{aligned} \mathcal{U}_1(t) + \mathcal{U}_2(t) &= \mathrm{Re}\left[e^{-i\omega t}\left(1 + e^{-i\varphi}\right)\right] \\ &= \mathrm{Re}\left[e^{-i\omega t}e^{-i\varphi/2}\left(e^{i\varphi/2} + e^{-i\varphi/2}\right)\right] \\ &= \mathrm{Re}\left[e^{-i\omega t}e^{-i\varphi/2}2\cos(\varphi/2)\right] \\ &= 2\cos(\varphi/2)\cos(\omega t + \varphi/2) \end{aligned}$$

For $\varphi = 2m\pi$ and $\varphi = \pi + 2m\pi$ we retrieve the results obtained before. It is important to note that what we see or detect physically (say, the 'brightness' of light) does not correspond to the quantities $\mathcal{U}_1, \mathcal{U}_2$. After all, $\mathcal{U}_1$ and $\mathcal{U}_2$ can attain negative values, while there is no such thing as 'negative brightness'. What $\mathcal{U}_1$ and $\mathcal{U}_2$ describe are the **fields**, which may be positive or negative. The 'brightness' or the **irradiance** or **intensity** is given by taking an average over a long time of $\mathcal{U}(t)^2$ (see (1.7.8), we shall omit the factor $\sqrt{\epsilon/\mu_0}$). As explained in Chapter 2 , we see and measure only the long time-average of $\mathcal{U}(t)^2$, because at optical frequencies $\mathcal{U}(t)^2$

fluctuates very rapidly. We recall the definition of the time average over an interval of length T at a specific time t given in (1.8.4) in Chapter 2:

$$\langle f(t) \rangle = \frac{1}{T}\int_t^{t+T} f(t')\,\mathrm{d}t',$$

where T is a time interval that is the response time of a typical detector, which is 10^{-9} s for a very fast detector, but this is still extremely long compared to the period of visible light which is of the order of 10^{-14} s. For a time-harmonic function, the long-time average is equal to the average over one period of the field and hence **it is independent of the time t at which it is taken**. Indeed for $(5.3.5)$ we get

$$\begin{aligned} I &= \left\langle \left(\mathcal{U}_1(t) + \mathcal{U}_2(t)\right)^2 \right\rangle \\ &= 4\cos^2(\varphi/2)\left\langle \cos^2(\omega t + \varphi/2) \right\rangle \\ &= 1 + \cos(\varphi) \end{aligned}$$

where $T\omega >> 1$. It is important to note that one can use complex notation to obtain the factor $1 + \cos(\varphi)$ more easily. Let us write

$$\mathcal{U}_1(t) = \mathrm{Re}\{U_1 e^{-i\omega t}\}, \quad \mathcal{U}_2(t) = \mathrm{Re}\{U_2 e^{-i\omega t}\},$$

where

$$U_1 = 1, \quad U_2 = e^{-i\varphi}.$$

Then we find

$$\begin{aligned} |U_1 + U_2|^2 &= \left|1 + e^{-i\varphi}\right|^2 \\ &= \left(1 + e^{i\varphi}\right)\left(1 + e^{-i\varphi}\right) \\ &= 1 + 1 + e^{-i\varphi} + e^{i\varphi} \\ &= 2 + 2\cos(\varphi), \end{aligned}$$

hence

$$I = \frac{1}{2}|U_1 + U_2|^2.$$

To see why this works, recall Eq. (1.8.5) and choose $A = B = U_1 + U_2$.

Remark. To shorten the formulae, we will omit in this chapter the factor $1/2$ in front of the time-averaged intensity.

Hence we define $I_1 = |U_1|^2$ and $I_2 = |U_2|^2$, and we then find for the time-averaged intensity of the sum of U_1 and U_2 :

$$\begin{aligned} I &= |U_1 + U_2|^2 = (U_1 + U_2)^* (U_1 + U_2) \\ &= |U_1|^2 + |U_2|^2 + U_1^* U_2 + U_1 U_2^* \\ &= I_1 + I_2 + 2\,\mathrm{Re}[U_1^* U_2]. \end{aligned}$$

Here, $2\,\mathrm{Re}[U_1^* U_2]$ is known as the interference term. In the famous double-slit experiment (which we will discuss in a later section), we can interpret the terms as follows: let us say U_1 is the field that comes from slit 1 , and U_2 comes from slit 2 . If only slit 1 is open, we measure on the screen intensity I_1, and if only slit 2 is open, we measure I_2. If both slits are open, we would not measure $I_1 + I_2$, but we would observe fringes due to the interference term $2\,\mathrm{Re}[U_1^* U_2]$.

More generally, the intensity of a sum of multiple time-harmonic fields U_j all having the same frequency is given by the **coherent sum**

$$I = \left|\sum_j U_j\right|^2$$

However, we will see in the next section that sometimes the fields are unable to interfere. In that case all the interference terms of the coherent sum vanish, and the intensity is given by the **incoherent sum**

$$I = \sum_j |U_j|^2.$$

5.4: Coherence

In the discussion so far we have only considered **monochromatic** light, which means that the spectrum of the light consists of only one frequency. Although light from a laser often has a very narrow band of frequencies and therefore can be considered to be monochromatic, purely monochromatic light does not exist. All light consists of multiple frequencies and therefore is **polychromatic**. Classical light sources such as incandescent lamps and also LEDs have relatively broad frequency bands. The question then arises how differently polychromatic light behaves compared to the idealised case of monochromatic light. To answer this question, we must study the topic of coherence. One distinguishes between two extremes: fully **coherent** and fully **incoherent** light, while the degree of coherence of practical light is somewhere in between. Generally speaking, the broader the frequency band of the source, the more incoherent the light is. It is a very important observation that no light is actually completely coherent or completely incoherent. All light is **partially coherent**, but some light is more coherent than others.

An intuitive way to think about these concepts is in terms of the ability to form interference fringes. For example, with laser light, which usually is almost monochromatic and hence coherent, one can form an interference pattern with clear maxima and minima in intensities (so-called fringes) using a double slit, while with sunlight (which is incoherent) this is much more difficult. Every frequency in the spectrum of sunlight gives its own interference pattern with its own frequency dependent fringe pattern. These fringe patterns wash out due to superposition and the total intensity therefore shows little fringe contrast, i.e. the coherence is less. However, it is not impossible to create interference fringes with natural light. The trick is to let the two slits be so close together (of the order of $0.02\ \mathrm{mm}$) that the difference in distances from the slits to the sun is so small for the fields in the slits are sufficiently coherent to interfere. To understand the effect of polychromatic light, it is essential to understand that the degree to which the fields in two points are coherent, i.e. the ability to form fringes, is determined by the **difference in distances between these points and the source**. The distance itself to the source is **not** relevant. In the next sections we study these issues more quantitatively.

5.3.1 Coherence of Light Sources

In a conventional light source such as a gas discharge lamp, photons are generated by electronic transitions in the atoms of the gas. These transitions have a duration of the order of 10^{-8} to 10^{-9} s. Because the emitted wave trains are finite, the emitted light does not have a single frequency; instead, there is a band of frequencies around a centre frequency. This spread of frequencies is called the **natural linewidth**. Random thermal motions of the molecules cause further broadening due to the Doppler effect. In addition, the atoms undergo collisions that interrupt the wave trains and therefore further broaden the frequency spectrum.

We first consider a **single emitting atom**. When collisions are the dominant broadening effect and these collisions are sufficiently brief, so that any radiation emitted during the collision can be ignored, an accurate model for the emitted wave is a steady monochromatic wave train at frequency $\bar{\omega}$ at the centre of the frequency band, interrupted by random phase jumps each time that a collision occurs. The discontinuities in the phase due to the collisions cause a spread of frequencies around the centre frequency. An example is shown in Figure $5.4.1$. The average time τ_0 between the collisions is typically less than 10 ns which implies that on average between two collisions roughly 150,000 harmonic oscillations occur. The coherence time $\Delta\tau_c$ is defined as the maximum time interval over which the phase of the electric field can be predicted. In the case of collisions-dominated emission by a single atom, the coherence time is equal to the average time between subsequent collisions: $\Delta\tau_c = \tau_0$.

To understand coherence and incoherence, it is very helpful to use this model for the emission by a single atom as harmonic wave trains of thousands of periods interrupted by random phase jumps. The coherence time and the width $\Delta\omega$ of the frequency line are related as

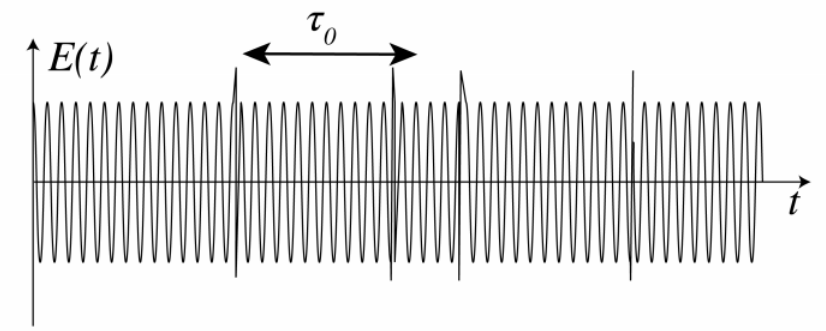


Figure 5.4.1: The electric field amplitude of the harmonic wave train radiated by a single atom at the centre frequency $\bar{\omega}$. The vertical lines are collisions separated by periods of free flight with mean duration τ_0. The quantity $\bar{\omega}\tau_0$, which is the number of periods in a typical wave train, is chosen unrealistically small (namely 60 , whereas a realistic value would be 10^5), to show the random phase changes.

$$\Delta\tau_c = \frac{2\pi}{\Delta\omega}$$

The coherence length is defined by

$$\Delta\ell_c = c\Delta\tau_c.$$

Since $\lambda\omega = 2\pi c$, we have

$$\frac{\Delta\lambda}{\bar{\lambda}} = \frac{\Delta\omega}{\bar{\omega}},$$

where λ and $\bar{\omega}$ are the wavelength and the frequency at the centre of the line. Hence,

$$\Delta\ell_c = c\frac{2\pi}{\Delta\omega} = 2\pi\frac{c}{\bar{\omega}}\frac{\bar{\omega}}{\Delta\omega} = \frac{\bar{\lambda}^2}{\Delta\lambda}$$

The coherence length and coherence time of a number of sources are listed in Table 5.4.1.

Table 5.4.1: Coherence time and length of several sources

Source	Mean wavelength	Linewidth	Coherence Length	Coherence Time
	$\bar{\lambda}$	$\Delta\lambda$	$\bar{\lambda}^2/\Delta\lambda$	$\frac{\bar{\lambda}}{\Delta\lambda}\frac{\bar{\lambda}}{c}$
Mid-IR (3-5 μm)	4.0μm	2.0μm	8.0μm	2.66×10^{-14} s
White light	550 nm	$\approx$ 300 nm	$\approx$ 900 nm	$\approx 3.0\times10^{-14}$ s.
Mercury arc	546.1 nm	$\approx$ 1.0 nm	$\approx$ 0.3 mm	$\approx 1.0\times10^{-12}$ s.
Kr ^{86} discharge lamp	605.6 nm	1.2×10^{-3} nm	0.3 m	1.0×10^{-9} s.
Stabilised He-Ne laser	632.8 nm	$\approx 10^{-6}$ nm	400 m	1.33×10^{-6} s.

5.3.2 Polychromatic Light

When dealing with coherence one has to consider fields that consist of a range of different frequencies. Let $\mathcal{U}(\mathbf{r},t)$ be the real-valued physical field component. It is always possible to write $\mathcal{U}(\mathbf{r},t)$ as an integral over time-harmonic components:

$$\mathcal{U}(\mathbf{r},t) = \mathrm{Re}\int_0^\infty A_\omega(\mathbf{r})e^{-i\omega t}\,\mathrm{d}\omega,$$

where $A_\omega(r)$ is the complex amplitude of the time-harmonic field with frequency ω. When there is only a certain frequency band that contributes, then $A_\omega = 0$ for ω outside this band. We define the **complex time-dependent field** $U(\mathbf{r},t)$ by

$$U(\mathbf{r},t) = \int_0^\infty A_\omega(\mathbf{r})e^{-i\omega t}\,\mathrm{d}\omega.$$

Then

$$\mathcal{U}(\mathbf{r},t) = \mathrm{Re}\, U(\mathbf{r},t)$$

Remark: The complex field $U(\mathbf{r},t)$ contains now the time dependence in contrast to the notation used for a time-harmonic (i.e. single frequency) field introduced in Chapter 2, where the time-dependent $e^{-i\omega t}$ was a separate factor.

Quasi-monochromatic field. If the width $\Delta\omega$ of the frequency band is very narrow, we speak of quasi-monochromatic flight. In the propagation of quasi-monochromatic fields, we use the formula for time-harmonic fields at the centre frequency $\bar{\omega}$. The quasi-monochromatic assumption simplifies the computations considerably. Nevertheless, most of the treatment that follows is more general and does not assume that the field is quasi-monochromatic.

We now compute the intensity of polychromatic light. The instantaneous energy flux is (as for monochromatic light) proportional to the square of the instantaneous real field: $\mathcal{U}(\mathbf{r},t)^2$. We average the instantaneous intensity over the integration time T of common detectors which, as stated before, is very long compared to the period at the centre frequency $2\pi/\bar{\omega}$ of the field (at least 10^5 times the period of the light). Using definition (5.2.6) and

$$\mathcal{U}(\mathbf{r},t) = \mathrm{Re}\, U(\mathbf{r},t) = \left(U(\mathbf{r},t) + U(\mathbf{r},t)^*\right)/2,$$

we get

$$\begin{aligned}
\left\langle \mathcal{U}(\mathbf{r},t)^2 \right\rangle &= \frac{1}{4}\left\langle \left(U(\mathbf{r},t) + U(\mathbf{r},t)^*\right)\left(U(\mathbf{r},t) + U(\mathbf{r},t)^*\right)\right\rangle \\
&= \frac{1}{4}\left\{ \left\langle U(\mathbf{r},t)^2 \right\rangle + \left\langle \left(U(\mathbf{r},t)^*\right)^2 \right\rangle + 2\left\langle U(\mathbf{r},t)^* U(\mathbf{r},t)\right\rangle \right\} \\
&= \frac{1}{2}\left\langle U(\mathbf{r},t)^* U(\mathbf{r},t)\right\rangle \\
&= \frac{1}{2}\left\langle \left|U(\mathbf{r},t)\right|^2 \right\rangle .
\end{aligned}$$

where the averages of $U(\mathbf{r},t)^2$ and $\left(U(\mathbf{r},t)^*\right)^2$ are zero because they are fast-oscillating functions of time. In contrast, $\left|U(\mathbf{r},t)\right|^2 = U(\mathbf{r},t)^* U(\mathbf{r},t)$ has a DC-component which does not average to zero.

Remark: In contrast to the time-harmonic case, the long time average of polychromatic light depends on the time t at which the average is taken. However, we assume in this chapter that the fields are omitted by sources that are **stationary**. The property of stationarity implies that the average over the time interval of long length T does not depend on the time that the average is taken. Many light sources, in particular conventional lasers, are stationary. (However, a laser source which emits short high-power pulses cannot be considered as a stationary source). We furthermore assume that the fields are **ergodic**, which means that taking the time-average over a long time interval amounts to the same as taking the average over the ensemble of possible fields. It can be shown that this property implies that the limit $T \to \infty$ in (5.2.6) indeed exists.

We use for the intensity again the expression without the factor $1/2$ in front, i.e.

$$I(\mathbf{r}) = \left\langle \left|U(\mathbf{r},t)\right|^2 \right\rangle .$$

The time-averaged intensity has hereby been expressed in terms of the **time-average of the squared modulus of the complex field**.

5.5: Temporal Coherence and the Michelson Interferometer

To investigate the coherence of a field, the most general approach is to make the field in two different points $\mathbf{r}_1$ and $\mathbf{r}_2$ interfere for some time delay τ and observe the fringe contrast. This means that one lets the fields $U(\mathbf{r}_1, t)$ and $U(\mathbf{r}_2, t-\tau)$ interfere. It is however customary to first look at the field in one point and let it interfere with itself but delayed in time, i.e. interfering $U(\mathbf{r}, t)$ with $U(\mathbf{r}, t-\tau)$. This special case is called **temporal coherence**. The other special case is **spatial coherence** in which the coherence of fields at two points is considered without time delay, by interfering $U(\mathbf{r}_1, t)$ and $U(\mathbf{r}_2, t)$. Spatial coherence will be treated later.

Because, when studying temporal coherence, the point $\mathbf{r}$ is always the same, we omit it from the formula. Furthermore, for easier understanding of the phenomena, we assume for the time being that the field considered is emitted by a single atom (i.e. a point source).

Temporal coherence is closely related to the spectral content of the light: if the light consists of fewer frequencies (think of monochromatic light), then it is more temporally coherent. To study the interference of $U(t)$ with $U(t-\tau)$, a Michelson interferometer, shown in Figure 5.5.1, is a suitable setup. The light that goes through one arm takes time t to reach the detector, while the light that goes through the other (longer) arm takes time $t+\tau$ which means that it was radiated earlier. Therefore, the detector observes the time-averaged intensity $\left\langle |U(t)+U(t-\tau)|^2 \right\rangle$. As remarked before, this averaged intensity does not depend on the time the average is taken, it only depends on the time difference τ between the two beams. We have

$$\begin{aligned} I(\tau) &= \left\langle |U(t)+U(t-\tau)|^2 \right\rangle \\ &= \left\langle |U(t)|^2 \right\rangle + \left\langle |U(t-\tau)|^2 \right\rangle + 2\,\mathrm{Re}\langle U(t)U(t-\tau)^* \rangle \\ &= 2\left\langle |U(t)|^2 \right\rangle + 2\,\mathrm{Re}\langle U(t)U(t-\tau)^* \rangle. \end{aligned}$$

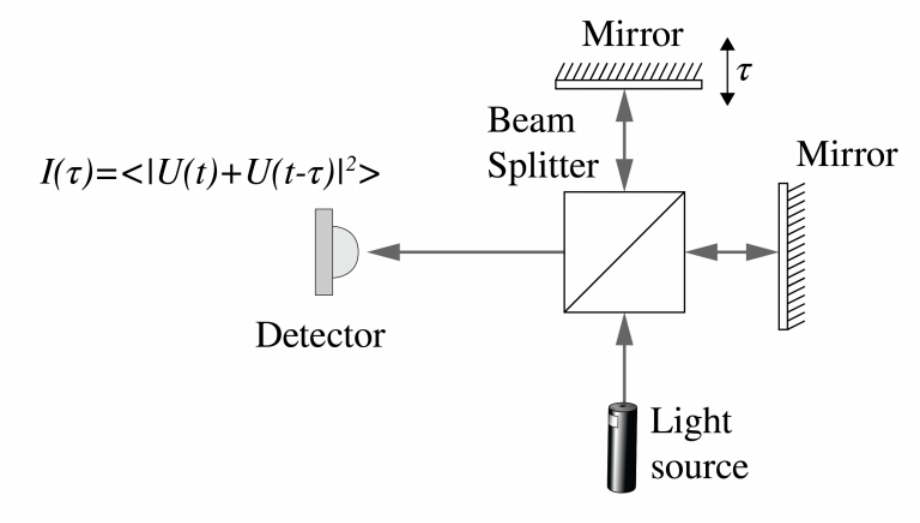


Figure 5.5.1: A Michelson interferometer to study the temporal coherence of a field. A beam is split in two by a beam splitter, and the two beams propagate over different distances which corresponds to a time difference τ and then interfere at the detector.

So far we have considered a field that originates from a single atom. The total field emitted by an extended source is the sum of fields $U_i(t)$ corresponding to all atoms i. In studying time coherence we assume that these fields are propagating more or less parallel and that the light has a fixed polarisation, so that the fields can be added algebraically. The total complex field produced by a large number N of atoms is

$$U(t) = U_1(t) + \ldots + U_N(t).$$

During the integration time of the detector the fields U_i experience thousands of random phase jumps and therefore they do not interfere: **the point sources of the extended source are mutually fully incoherent**. The detected total intensity (5.5.1) is thus the sum of the intensities of the individual atoms:

$$\begin{aligned} I(\tau) &= \left\langle |U(t)+U(t-\tau)|^2 \right\rangle \\ &= \left\langle \sum_i \Big| U_i(t)+U_i(t-\tau) \Big] \, |^2 \right\rangle \\ &= \sum_i 2 \left\{ \left\langle U_i(t)|^2 \right\rangle + \mathrm{Re}\langle U_i(t) U_i(t-\tau)^* \rangle \right\}, \end{aligned}$$

The expression $\left\langle |U(t)+U(t-\tau)|^2 \right\rangle$ does not depend on position, so it cannot describe interference fringes in space. To better observe what happens when τ is varied, we introduce interference fringes in space by tilting one beam so that the observed interference pattern is given by

$$\begin{aligned} I(x,\tau) &= \left\langle \left| U(t) e^{ik_x x} + U(t-\tau) \right|^2 \right\rangle \\ &= 2 \left\langle |U(t)|^2 \right\rangle + 2 \,\mathrm{Re} \left\langle U(t) U(t-\tau)^* e^{ik_x x} \right\rangle. \end{aligned}$$

If τ is changed, the maxima of the interference pattern translate as function of x, which is easy to observe. How interference fringes for tilted collimated beams are observed in a Michelson interferometer is demonstrated in. It is possible to obtain different fringe patterns using diverging beams instead of collimated beams, as is demonstrated in.

The **self coherence function** $\Gamma(\tau)$ is defined by

$$\Gamma(\tau) = \langle U(t) U(t-\tau)^* \rangle \qquad \text{self-coherence.}$$

The intensity of $U(t)$ is

$$I_0 = \left\langle |U(t)|^2 \right\rangle = \Gamma(0).$$

The **complex degree of self-coherence** is defined by:

$$\gamma(\tau) = \frac{\Gamma(\tau)}{\Gamma(0)}. \qquad \text{complex degree of self-coherence}$$

This is a complex number with modulus between 0 and 1 :

$$0 \le |\gamma(\tau)| \le 1,$$

The observed intensity can then be written:

$$I(x,\tau) = 2 I_0 \left\{ 1 + \mathrm{Re} \left[\gamma(\tau) e^{ik_x x} \right] \right\},$$

Recall that we vary τ by varying the length of one of the arms in the Michelson interferometer.

We consider some special cases. Suppose $U(t)$ is a monochromatic wave

$$U(t) = e^{-i\omega t}.$$

In that case we get for the self-coherence

$$\Gamma(\tau) = \left\langle e^{-i\omega t} e^{i\omega(t-\tau)} \right\rangle = e^{-i\omega\tau},$$

and

$$\gamma(\tau) = e^{-i\omega\tau}.$$

Hence the interference pattern is given by

$$I(x,\tau) = 2 \left[1 + \cos(\omega\tau - k_x x) \right].$$

So for monochromatic light we expect to see a cosine interference pattern, which shifts as we change the arm length of the interferometer (i.e. change τ). No matter how large τ, a clear interference pattern should be observed.

Next we consider what happens when the light is a superposition of two frequencies:

$$U(t) = \frac{e^{-i(\bar{\omega}+\Delta\omega/2)t} + e^{-i(\bar{\omega}-\Delta\omega/2)t}}{2},$$

where $\Delta\omega \ll \bar{\omega}$. Then:

$$\begin{aligned}\Gamma(\tau) &= \frac{1}{4}\left\langle \left(e^{-i(\bar{\omega}+\Delta\omega/2)t} + e^{-i(\bar{\omega}-\Delta\omega/2)t}\right)\left(e^{i(\bar{\omega}+\Delta\omega/2)(t-\tau)} + e^{i(\bar{\omega}-\Delta\omega/2)(t-\tau)}\right)\right\rangle \\ &\approx \frac{e^{-i(\bar{\omega}+\Delta\omega/2)\tau} + e^{-i(\bar{\omega}-\Delta\omega/2)\tau}}{4} \\ &= \cos(\Delta\omega\tau/2)\frac{e^{-i-\bar{\omega}\tau}}{2}\end{aligned}$$

where in the second line the time average of terms that oscillate with time is set to zero because the averaging is done over a time interval of duration T satisfying $T\Delta\bar{\omega} \gg 1$. Hence, the complex degree of self-coherence is:

$$\gamma(\tau) = \cos(\Delta\omega\tau/2)e^{-i\bar{\omega}\tau}$$

and (5.5.8) becomes

$$I(x,\tau) = \left\{1 + \mathrm{Re}\left[\gamma(\tau)e^{ik_x x}\right]\right\} = \left[1 + \cos(\Delta\omega\tau/2)\cos(\bar{\omega}\tau - k_x x)\right].$$

The interference term is the product of the function $\cos(\bar{\omega}\tau - k_x x)$), which is a rapidly oscillating function of τ, and a slowly varying envelope $\cos(\Delta\omega\tau/2)$. It is interesting to note that the envelope, and hence $\gamma(\tau)$, vanishes for some periodically spaced τ, which means that for certain τ the degree of self-coherence vanishes and no interference fringes form. Note that if $\Delta\omega$ is increased, the intervals between the zeroes of $\gamma(\tau)$ decrease.

If more frequencies are added, the envelope function is not a cosine function but on average decreases with τ. The typical value of τ below which interferences are observed is roughly equal to half the first zero of the envelope function. This value is called the coherence time $\Delta\tau_c$. We conclude with some further interpretations of the degree of self-coherence $\gamma(\tau)$.

- In stochastic signal analysis $\Gamma(\tau) = \langle U(t)U(t-\tau)^*\rangle$ is called the autocorrelation of $U(t)$. Informally, one can interpret the autocorrelation function as the ability to predict the field U at time t given the field at time $t-\tau$.
- The Wiener-Khinchine theorem says that the **Fourier transform of the self coherence function is the spectral power density of** $U(t)$:

$$\hat{\Gamma}(\omega) = |\hat{U}(\omega)|^2.$$

This result can be proved for stationary fields using Parseval's indentity. Using the uncertainty principle, we can see that the larger the spread of the frequencies of $U(t)$ (i.e. the larger the bandwidth), the more sharply peaked $\Gamma(\tau)$ is. Thus, the light gets temporally less coherent when it consists of a broader range of frequencies.

5.6: Longitudinally and Transversely Separated Points

Consider two points that are separated along the mean direction of propagation of the light (imagine that the two points are situated along the same ray). If the phase is known in one of these points at a certain instant, then we can predict the phase in the other point at the same instant of time, provided that the distance between the points is less than the coherence length $\Delta\ell_c$. We call points separated parallel to the mean direction of propagation of the light **longitudinally separated points**. In Figure 5.6.1 at the left, the spherical wave fronts of a purely monochromatic point source are shown. The field is fully coherent. What happens in point P_1 will some time later happen in P_2 and still a bit later in P_3 etc. However, if the point source is not monochromatic, and the phase of the field changes randomly from time to time as in Figure 5.3.1, there is hardly any correlation of the wave at points that are longitudinally far apart such as P_1 and P_3. Nevertheless, the waves are not totally unpredictable: the behaviour at points closer together such as P_1 and P_2 are still somewhat correlated. When the distance between longitudinally separated points is larger than the coherence length $\Delta\ell_c$, the mutual coherence between these points is negligible. Michelson's interferometer is ideally suited to determine the coherence time and coherence length and therefore also the mutual coherence of longitudinally separated points. Because the distance between longitudinally separated points is equal to the **difference** in the distances between the source and each of points, we can alternatively say that two longitudinally separated points are coherent if the **difference in distances** from the source to these points is smaller than the coherence length.

Next consider points which are separated in a direction perpendicular to the mean direction of propagation of the light. Such points are called **transversely separated**. The mutual coherence of the fields at these points for zero (or small) time delay is called **spatial coherence**. Note that, even when the source is not monochromatic and the coherence time of the source is short, the fields in the transversely separated points P_1 and P_4 at the left of Figs. 5.6.1 are always completely correlated because they are on the same wavefront. In fact, the distances between the source and these points identical.

The situation is different when the source is not a single point source but consists of several mutually completely incoherent point sources as at the right of Figure Figs. 5.6.1. First of all, the fields from different point sources can not interfere with each other because, during the integration time of a typical detector, fields of different point sources undergo thousands of random uncorrelated phase jumps. Furthermore, the **difference** between the distances of two points such as P_1 and P_2 at the right of Figure Figs. 5.6.1, cannot be zero for all point sources. Hence, the contributions of the individual point sources to the total fields in the two points can not all be in phase or out of phase together. Stated differently, transversely separated points that are on the same wavefront emitted by a particular point source are coherent, but these points cannot be on the same wave fronts for all point sources. However, when the differences of distances from the transversely separated points P_1 and P_2 at the right of Figure Figs. 5.6.1 to each of the point sources are smaller than the coherence length $\Delta\ell_c$, the fields at the same instant (zero time delay) at the transversely separated points are nevertheless coherent.

The spatial coherence can be experimentally investigated by putting an opaque mask orthogonal to the mean direction of propagation of the light, with two small holes at the points of interest. The fields transmitted by the holes are made to interfere at a screen at a large distance, as shown in Figure 5.6.1. This is the famous Young experiment. The field due to one point source in the two holes can interfere on the screen, but contributions from different point sources cannot. Hence the fringe pattern is the sum of the intensity patterns due to the individual point sources and the contrast in the fringe pattern of this total intensity pattern is a measure of the degree of spatial coherence of the disturbances at the two transversely separated points (holes).

We conclude that for both longitudinally and transversely separated points, and in fact more generally for any pair of points,

The fields in two points considered at the same instant (no time delay) are coherent when the **differences** in distances between these points and any of the point sources in the extended source are smaller than the coherence length $\Delta\ell_c$.

This result will be derived more rigorously in Section 5.7.

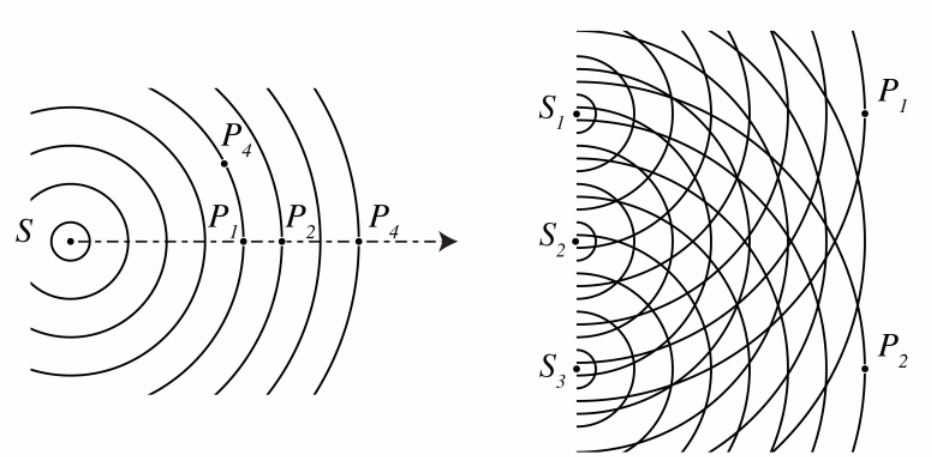


Figure 5.6.1: Left: the waves emitted by a single monochromatic point sources are perfectly (temporally and spatially) coherent. If the point source is not monochromatic, the fields at points that are longitudinally separated such as P_1, P_2 and P_3 are more coherent the smaller the difference in distance is of these points to the source. Transversely separated points such as P_1, P_4 in the left figure, are fully coherent because they are on the same wave front. When there are more point sources such as in the right figure, the degree of spatial coherence of transversally separated points such as P_1 and P_2, depends on the differences in distance of each point source to the points P_1, P_2.

5.7: Spatial Coherence and Young's Experiment

While for temporal coherence we used a **Michelson interferometer**, the natural choice to characterize spatial coherence is Young's experiment, because it allows the fields in two points separated in space to interfere with each other. In **Young's experiment**, a mask is used with two pinholes at the positions of the points P_1 and P_2 of interest. Let $\mathbf{r}_1$ and $\mathbf{r}_2$ be the position vectors of the two points P_1 and P_2, respectively. We write the field in P_1 as a superposition of time-harmonic fields as in (5.3.6):

$$U(\mathbf{r}_1, t) = \int A_\omega(\mathbf{r}_1)\, e^{-i\omega t}\, \mathrm{d}\omega.$$

According to the Huygens-Fresnel principle, a time-harmonic disturbance with frequency ω in the pinhole at $\mathbf{r}_1$ causes a radiating spherical wave behind the mask, with time-harmonic field in some point $\mathbf{r}$ given by

$$A_\omega(\mathbf{r}_1)\, \frac{e^{-i\omega(t-|\mathbf{r}-\mathbf{r}_1|/c)}}{|\mathbf{r}-\mathbf{r}_1|}.$$

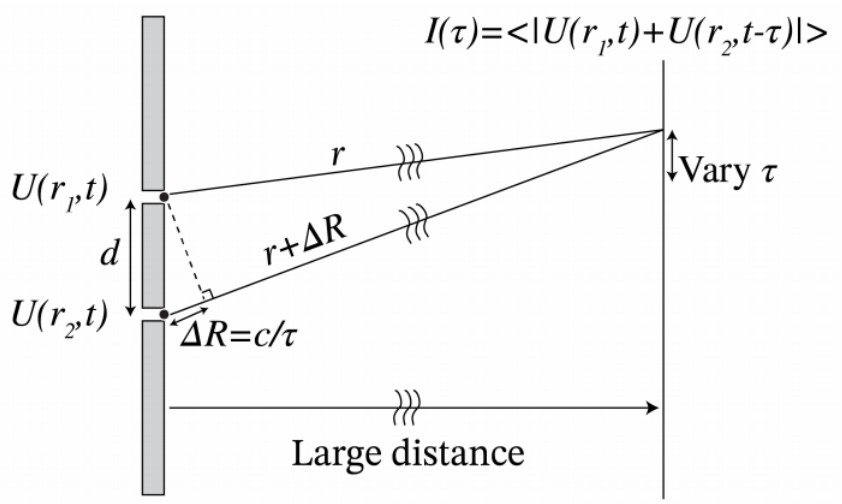


Figure 5.7.1: Young's experiment to study the spatial coherence of two points. A mask with two holes at the two points of interest, $\mathbf{r}_1$ and $\mathbf{r}_2$, is used to let the fields in these points interfere with each other on a screen at a large distance. Because the light propagates over different distances from the two holes to the point of observation, $U(\mathbf{r}_1, t)$ interferes with $U(\mathbf{r}_2, t+\tau)$, where τ is the difference in propagation time.

The total field $U_1(\mathbf{r}, t)$ in any point $\mathbf{r}$ due to the pinhole at P_1 is obtained by integrating over all frequencies:

$$U_1(\mathbf{r}, t) = \int A_\omega(\mathbf{r}_1)\, \frac{e^{-i\omega(t-|\mathbf{r}-\mathbf{r}_1|/c)}}{|\mathbf{r}-\mathbf{r}_1|} \mathrm{d}\omega = \frac{U(\mathbf{r}_1, t-|\mathbf{r}-\mathbf{r}_1|/c)}{|\mathbf{r}-\mathbf{r}_1|}.$$

In words, the field in $\mathbf{r}$ at time t due to the pinhole at $\mathbf{r}_1$ is proportional to the field at $\mathbf{r}_1$ at the earlier time $= |\mathbf{r}-\mathbf{r}_1|/c$ that it takes for the light to propagate form $\mathbf{r}_1$ to $\mathbf{r}$. The proportionality factor scales with the reciprocal distance between $\mathbf{r}$ and $\mathbf{r}_1$.

Consider the set-up shown in Figure 5.7.1. The fields U_1 and U_2 from the two pinholes at $\mathbf{r}_1$ and $\mathbf{r}_2$ interfere with each other in a point $\mathbf{r}$ at a great distance. Because of the difference in propagation distance $\Delta R = |\mathbf{r}-\mathbf{r}_2| - |\mathbf{r}-\mathbf{r}_1|$, there is a time difference τ between the two fields when they arrive at point $\mathbf{r}$ on the screen, given by

$$\tau = \frac{\Delta R}{c}.$$

Furthermore, because of the propagation, the amplitudes are reduced by a factor proportional to the reciprocal distance which is different for the two fields. But if the distance between the two screens is large enough, we can take these factors to be the same and then omit them. The interference pattern on the screen is then given by

$$
\begin{aligned}
I(\tau) &= \left\langle \left| U_1(\mathbf{r},t) + U_2(\mathbf{r},t) \right|^2 \right\rangle \\
&= \left\langle \left| U\left(\mathbf{r}_1, t - |\mathbf{r}-\mathbf{r}_1|\,/c\right) + U\left(\mathbf{r}_2, t - |\mathbf{r}-\mathbf{r}_2|\,/c\right) \right|^2 \right\rangle \\
&= \left\langle \left| U\left(\mathbf{r}_1, t\right) + U\left(\mathbf{r}_2, t-\tau\right) \right|^2 \right\rangle \\
&= \left\langle \left| U\left(\mathbf{r}_1, t\right) \right|^2 \right\rangle + \left\langle \left| U\left(\mathbf{r}_2, t-\tau\right) \right|^2 \right\rangle + 2\,\mathrm{Re}\langle U\left(\mathbf{r}_1, t\right) U\left(\mathbf{r}_2, t-\tau\right)^* \rangle .
\end{aligned}
$$

We define the **mutual coherence function** and the intensities:

$$
\begin{aligned}
\Gamma_{12}(\tau) &= \langle U\left(\mathbf{r}_1, t\right) U\left(\mathbf{r}_2, t-\tau\right)^* \rangle , \\
I_1 &= \left\langle \left| U\left(\mathbf{r}_1, t\right) \right|^2 \right\rangle = \Gamma_{11}(0), \\
I_2 &= \left\langle \left| U\left(\mathbf{r}_2, t-\tau\right) \right|^2 \right\rangle = \Gamma_{22}(0).
\end{aligned}
$$

The **complex degree of mutual coherence** is defined by using these intensities to normalise $\Gamma_{12}(\tau)$:

$$
\gamma_{12}(\tau) = \frac{\Gamma_{12}(\tau)}{\sqrt{\Gamma_{11}(0)}\sqrt{\Gamma_{22}(0)}}, \qquad \text{complex degree of mutual coherence.}
$$

The modulus of γ_{12} is smaller or equal than 1 (which can be proved by using Bessel's inequality). We can now write ($5.7.5$) as

$$
I(\tau) = I_1 + I_2 + 2\sqrt{I_1}\sqrt{I_2}\,\mathrm{Re}\{\gamma_{12}(\tau)\}.
$$

By varying the point of observation $\mathbf{r}$ over the screen, we can vary τ. By measuring the intensities, we can deduce the real part of $\gamma_{12}(\tau)$. Note that $\gamma_{12}(\tau)$ indicates the ability to form fringes.

Let us see what happens when $U(\mathbf{r},t)$ is a monochromatic field

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

In that case

$$
\begin{aligned}
\Gamma_{12}(\tau) &= \left\langle A\left(\mathbf{r}_1\right) A(\mathbf{r}_2)^* e^{-i\omega t} e^{i\omega(t-\tau)} \right\rangle \\
&= A\left(\mathbf{r}_1\right) A(\mathbf{r}_2)^* e^{-i\omega\tau}.
\end{aligned}
$$

So we get

$$
\gamma_{12}(\tau) = \frac{\Gamma_{12}(\tau)}{\left|A\left(\mathbf{r}_1\right)\right| \left|A\left(\mathbf{r}_2\right)\right|} = e^{-i\omega\tau + i\varphi},
$$

where φ is the phase difference of $A\left(\mathbf{r}_2\right)$ and $A\left(\mathbf{r}_1\right)$. In this case γ_{12} has modulus 1 , as expected for a monochromatic field. The intensity on the screen becomes

$$
I(\tau) = \left|A\left(\mathbf{r}_1\right)\right|^2 + \left|A\left(\mathbf{r}_2\right)\right|^2 + 2\left|A\left(\mathbf{r}_1\right)\right| \left|A\left(\mathbf{r}_2\right)\right| \cos(\omega\tau - \varphi).
$$

So indeed we see interference fringes, as one would expect for a monochromatic wave. If $\varphi = 0$, then interference maxima occur for

$$
\omega\tau = 0, \pm 2\pi, \pm 4\pi, \pm 6\pi, \ldots
$$

Because $\omega = c\frac{2\pi}{\lambda}$, and $\Delta R = c\tau$, we find that maxima occur when

$$
\Delta R = 0, \pm\lambda, \pm 2\lambda, \pm 3\lambda, \ldots
$$

For large distance between the mask and the screen (in the Fraunhofer limit), these path length differences correspond to directions of the maxima given by the angles θ_m (see Figure $5.7.1$):

$$
\theta_m = \frac{\Delta R}{d} = m\frac{\lambda}{d},
$$

where d is the distance between the slits and m is an integer.

Remark. We recognise $\Gamma_{12}(\tau) = \langle U(\mathbf{r}_1, t)\, U(\mathbf{r}_2, t-\tau)^* \rangle$ to be the **cross-correlation** of the two signals $U(\mathbf{r}_1, t)$ and $U(\mathbf{r}_2, t)$.

5.8: Increase of Spatial Coherence by Propagation

When light propagates, the mutual coherence at transversely separated points in general increases with propagation distance. For example, the sun consists of an incredibly large number of emitters which emit at random times. Therefore, the fields at different points at the surface of the sun are completely uncorrelated, i.e. the field at the surface of the sun is spatially fully incoherent. But after propagation to the earth, the sunlight has become partially coherent, with coherence length of roughly 50μm around the wavelength of 510 nm. The further away the sun or a star is, the more spatially coherent the field in two transversely separated points becomes. How can we understand this?

One way of looking at it is that at the star's surface, the field is determined by the closest point sources. Since all point sources emit wavetrains making thousands of uncorrelated phase jumps during an integration time of any detector, the fields in two points at the surface of the star are completely incoherent. But the field at two transversely separated points at a great distance from the star both consist of contributions from all point sources in the star, i.e. the fields of the different point sources mix spatially. Provided the **difference** in distances between each of the transverse points and the point sources of the star is smaller than the coherence length, these contributions can interfere. Interesting, when the distance of the transversely separated points to the star increases, the difference between the distances of the points to the star decreases. Therefore, the field becomes more spatially coherent when the distance to the start increases.

To quantify this phenomenon, we consider two mutually incoherent point sources S_1 and S_2 in the $z=0$ plane. We assume that their mutual coherence function is given by:

$$\begin{aligned}\Gamma_{S_1S_2}(\tau)&=0,\ \text{for all } \tau,\\ \Gamma_{S_1S_1}(\tau)&=\Gamma_{S_2S_2}(\tau)=\Gamma_0(\tau),\end{aligned}$$

where Γ_0 is the self-coherence which we assume here to be the same for both point sources, $\Gamma_0(\tau)$ is a decreasing function of the delay time τ (although it is not necessarily monotonically decreasing). Using that the long-time average does not depend on the origin of time (which was based on the assumption that the source is stationary), we find:

$$\Gamma_0(-\tau)=\langle U(S_1,t)\,U(S_1,t+\tau)^*\rangle=\langle U(S_1,t-\tau)\,U(S_1,t)^*\rangle=\Gamma_0(\tau)^*.$$

Furthermore, for $\tau=0:\Gamma_0(0)=I_0$, which is the intensity of either source.

Consider two points P_1,P_2 at large distance z from the two point sources as seen in Figure 5.8.1. We will compute the mutual coherence $\Gamma_{P_1P_2}(0)$ of these points for zero time delay $\tau=0$ (we can also compute the mutual coherence for more general time delay τ, i.e. $\Gamma_{P_1P_2}(\tau)$, but it will suffice for our purpose to take $\tau=0$). The field in P_1 is the sum of the fields emitted by S_1 and S_2 :

$$U(P_1,t)\propto\frac{U(S_1,t-|S_1P_1|/c)}{|S_1P_1|}+\frac{U(S_2,t-|S_2P_1|/c)}{|S_2P_1|},$$

where we used (5.6.3). Similarly,

$$U(P_2,t)\propto\frac{U(S_1,t-|S_1P_2|/c)}{|S_1P_2|}+\frac{U(S_2,t-|S_2P_2|/c)}{|S_2P_2|}.$$

Let us assume that z is so large that all distances $|S_iP_j|$ in the denominators can be replaced by z. Then these equal distances can be omitted. It can then be shown by substitution of (5.8.4) and (5.8.5) that the mutual coherence in P_1 and P_2 for zero time delay $\tau=0$ becomes:

$$\begin{aligned}\Gamma_{P_1P_2}(0)=&\langle(U(P_1,t)\,U(P_2,t)^*\rangle\\ =&\Gamma_{S_1S_1}\left(\frac{|S_1P_2|-|S_1P_1|}{c}\right)+\Gamma_{S_1S_2}\left(\frac{|S_2P_2|-|S_1P_1|}{c}\right)\\ &+\Gamma_{S_2S_1}\left(\frac{|S_1P_2|-|S_2P_1|}{c}\right)+\Gamma_{S_2S_2}\left(\frac{|S_2P_2|-|S_2P_1|}{c}\right).\end{aligned}$$

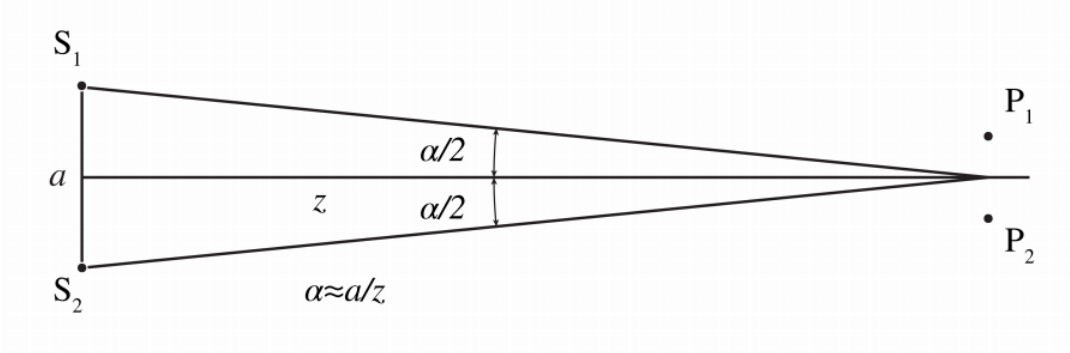


Figure 5.8.1: Two incoherent point sources S_1, S_2 and two points P_1, P_2 in a plane at large distance z from the point sources. The degree of mutual coherence at P_1 and P_2 increases to 1 in the limit $\alpha \to 0$.

Now we use (5.8.1) and (5.8.2) and get

$$\Gamma_{P_1P_2}(0) = \Gamma_0\left(\frac{|S_1P_2| - |S_1P_1|}{c}\right) + \Gamma_0\left(\frac{|S_2P_2| - |S_2P_1|}{c}\right).$$

Similarly,

$$\Gamma_{P_1P_1}(0) = \Gamma_{P_2P_2}(0) = 2\Gamma_0(0) = 2I_0.$$

The result (5.8.7) confirms what was already remarked in Section 5.5, that the mutual coherence $\Gamma_{P_1P_2}(\tau = 0)$ depends on the **difference** in distances of source point S_1 to points P_1 and P_2 and on the **difference** in distance from S_2 to P_1 and P_2.

Remark. The derivation given here is valid for any two points P_1, P_2 sufficiently far away from the points sources S_1 and S_2 and hence applies not only to transversally but also to longitudinally separated points.

Suppose that $S_1 = (a/2, 0, 0), S_2 = (-a/2, 0, 0)$ and let P_1, P_2 be given by: $P_j = (x_j, 0, z)$ for $j = 1, 2$. If z is very large, so that S_1P_1 and S_1P_2 are almost parallel, then with Figure 5.8.2 we find

$$|S_1P_2| - |S_1P_1| \approx |QP_2| = \frac{\alpha}{2}(x_1 - x_2).$$

Similarly,

$$|S_2P_2| - |S_2P_1| \approx -\frac{\alpha}{2}(x_1 - x_2).$$

Hence, with (5.8.3):

$$\Gamma_{P_1P_2}(0) = 2\,\mathrm{Re}\,\Gamma_0\left(\frac{\alpha}{2}\frac{(x_1 - x_2)}{c}\right).$$

It is thus seen that the degree of mutual coherence depends on the angle $\alpha \approx a/z$ subtended by the two point sources at the midpoint $(0, 0, z)$ on the mask. The smaller this angle, the higher the degree of spatial coherence. The reason is that for smaller α, i.e. for a smaller size of the source and/or a greater distance to the point sources, the fringes due to the point sources become almost overlapping and enforce each other, whereas for larger α the fringes are more displaced with respect to each other and hence the sum of the fringes of the two point sources has less pronounced fringe visibility.

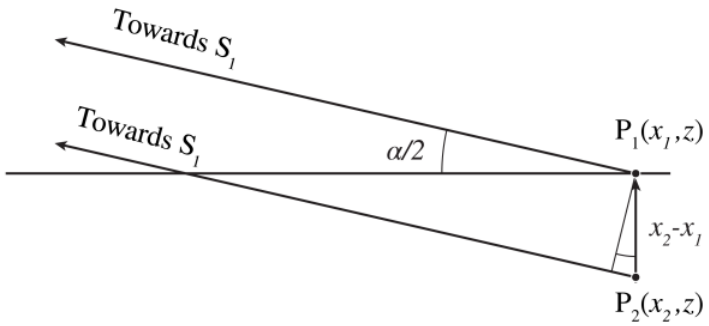


Figure 5.8.2: For z very large, S_1P_1 and S_1P_2 are almost parallel and $|S_1P_2| - |S_1P_1| \approx |QP_2| = (x_1 - x_2)\,\alpha/2$.

As example consider quasi-monochromatic light for which (see (5.4.10):

$$\Gamma_0(\tau) = I_0 e^{-i\bar{\omega}\tau}, \text{ for all } \tau$$

where $\bar{\omega}$ is the centre frequency. Then

$$\Gamma_{P_1P_2}(0) = 2I_0 \cos\left[\frac{\alpha\bar{\omega}\,(x_1 - x_2)}{c}\right],$$

and hence the degree of mutual coherence is:

$$\begin{aligned}\gamma_{P_1P_2}(0) &= \frac{\Gamma_{P_1P_2}(0)}{\sqrt{\Gamma_{P_1P_1}(0)}\sqrt{\Gamma_{P_2P_2}(0)}} \\ &= \cos\left[\frac{\alpha\bar{\omega}\,(x_1 - x_2)}{2}\right].\end{aligned}$$

We see that when

$$x_1 - x_2 = \bar{\lambda}/(2\alpha),$$

there is no interference. Hence we can say that coherence between points in the mask only occurs when there distance is smaller than $\bar{\lambda}/(2\alpha)$.

We see from (5.8.14) that by keeping P_1 fixed, we can retrieve the angle α by measuring $\mathrm{Re}\,\gamma_{P_1P_2}(0)$ for a number of different positions of P_2.

5.9: Stellar Interferometry

The property that the spatial coherence of two transversely separated points decreases for increasing angle which the source subtends at the two points, is used in **stellar interferometry**. It works as follows: we want to know the size of a certain star. The size of the star, being an extended spatially incoherent source, determines the spatial coherence of the light we receive on earth. Thus, by measuring the interference of the light collected by two transversely separated telescopes, one can effectively create a double-slit experiment, with which the degree of spatial coherence of the star light on earth can be measured, and thereby the angle which the star subtends on earth. The resolution in retrieving the angle from the spatial coherence is larger when the distance between the telescopes is larger (see 5.7.14). Then, if we know the distance of the star by independent means, e.g. from its spectral brightness, we can deduce its size from its angular size.

The method can also be used to derive the intensity distribution at the surface of the star. It can be shown that the degree of spatial coherence as function of the relative position of the telescopes is the Fourier transform of the intensity distribution. Hence, by moving the telescopes around and measuring the spatial coherence for many positions, the intensity distribution can be derived from a back Fourier transform.

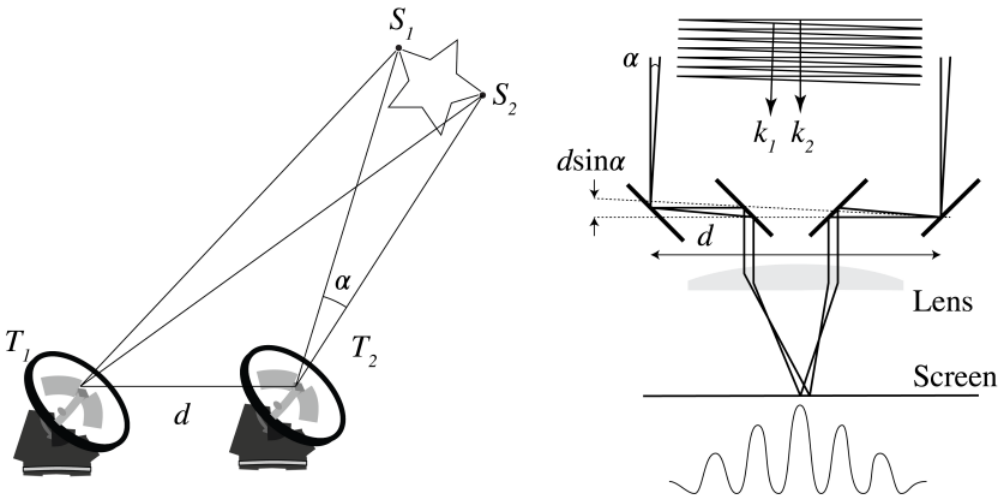


Figure 5.9.1: Left: a stellar interferometer with two telescopes that can be moved around to measure the interference at many relative positions. Right: single telescope with two outer movable mirrors. The telescope can move around it axis. The larger the distance d the higher the resolution.

5.10: Fringe Visibility

We have seen that when the interference term $\mathrm{Re}\langle U_1^* U_2\rangle$ vanishes, no fringes form, while when this term is nonzero, there are fringes. The **fringe visibility** is expressed directly in measurable quantities (i.e. in intensities instead of fields). Given some interference intensity pattern $I(x)$ as in Figure 5.10.1, the visibility is defined as

$$\mathcal{V} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}. \quad \text{fringe visibility.}$$

For example, if we have two perfectly coherent, monochromatic point sources emitting the fields U_1, U_2 with intensities $I_1 = |U_1|^2, I_2 = |U_2|^2$, then the interference pattern is with (5.6.13):

$$I(\tau) = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\omega\tau + \varphi).$$

We then get

$$I_{\max} = I_1 + I_2 + 2\sqrt{I_1 I_2}, \quad I_{\min} = I_1 + I_2 - 2\sqrt{I_1 I_2},$$

SO

$$\mathcal{V} = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2}$$

In case $I_1 = I_2$, we find $\mathcal{V} = 1$. In the opposite case, where U_1 and U_2 are completely incoherent, we find

$$I(\tau) = I_1 + I_2,$$

from which follows

$$I_{\max} = I_{\min} = I_1 + I_2,$$

which gives $\mathcal{V} = 0$.

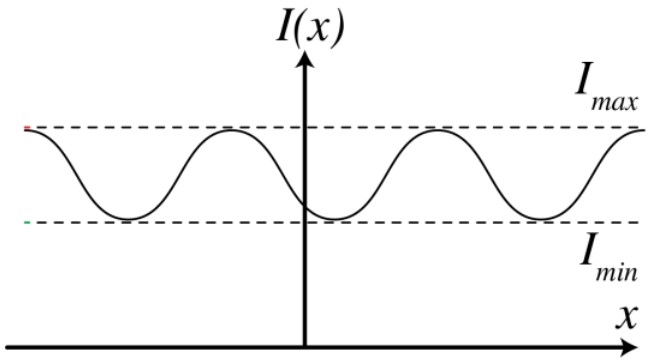


Figure 5.10.1: Illustration of $I_{\max}$ and $I_{\min}$ of an interference pattern $I(x)$ that determines the visibility $\mathcal{V}$.

5.11: Interference and polarisation

In the study of interference we have so far ignored the vectorial nature of light (i.e. its polarisation) by assuming that all the fields have the same polarisation. Suppose now that we have two real vector fields $\mathcal{E}_1, \mathcal{E}_2$. The (instantaneous) intensity of each field is (apart from a constant factor) given by

$$\mathcal{E}_1 \cdot \mathcal{E}_1, \quad \mathcal{E}_2 \cdot \mathcal{E}_2$$

If the two fields interfere, the instantaneous intensity is given by

$$(\mathcal{E}_1 + \mathcal{E}_2) \cdot (\mathcal{E}_1 + \mathcal{E}_2) = \mathcal{E}_1 \cdot \mathcal{E}_1 + \mathcal{E}_2 \cdot \mathcal{E}_2 + 2\mathcal{E}_1 \cdot \mathcal{E}_2$$

where $2\mathcal{E}_1 \cdot \mathcal{E}_2$ is the interference term. Suppose the polarisation of $\mathcal{E}_1$ is orthogonal to the polarisation of $\mathcal{E}_2$, e.g.

$$\mathcal{E}_1 = \begin{pmatrix} \mathcal{E}_{1x} \\ 0 \\ 0 \end{pmatrix}, \quad \mathcal{E}_2 = \begin{pmatrix} 0 \\ \mathcal{E}_{2y} \\ 0 \end{pmatrix}$$

Then $\mathcal{E}_1 \cdot \mathcal{E}_2 = 0$, which means the two fields can not interfere. This observation is the

First Fresnel-Arago Law: fields with orthogonal polarisation cannot interfere.

Next we write the fields in terms of orthogonal components

$$\mathcal{E}_1 = \begin{pmatrix} \mathcal{E}_{1\perp} \\ \mathcal{E}_{1\parallel} \end{pmatrix} \quad \mathcal{E}_2 = \begin{pmatrix} \mathcal{E}_{2\perp} \\ \mathcal{E}_{2\parallel} \end{pmatrix}$$

This is always possible, whether the fields are polarised or randomly polarised. Then (5.11.2) becomes

$$\mathcal{E}_1 \cdot \mathcal{E}_1 + \mathcal{E}_2 \cdot \mathcal{E}_2 + 2\mathcal{E}_1 \cdot \mathcal{E}_2 = \mathcal{E}_{1\perp}^2 + \mathcal{E}_{2\perp}^2 + 2\mathcal{E}_{1\perp}\mathcal{E}_{2\perp} + \mathcal{E}_{1\parallel}^2 + \mathcal{E}_{2\parallel}^2 + 2\mathcal{E}_{1\parallel}\mathcal{E}_{2\parallel}$$

If the fields are randomly polarised, the time average of the $\perp$-part will equal the average of the |-part, so the time-averaged intensity becomes

$$\begin{aligned} I &= 2\left\langle \mathcal{E}_{1\perp}^2 + \mathcal{E}_{2\perp}^2 + 2\mathcal{E}_{1\perp}\mathcal{E}_{2\perp} \right\rangle \\ &= 2\left\langle \mathcal{E}_{1\parallel}^2 + \mathcal{E}_{2\parallel}^2 + 2\mathcal{E}_{1\parallel}\mathcal{E}_{2\parallel} \right\rangle \end{aligned}$$

This is qualitatively the same as what we would get if the fields had parallel polarisation, e.g.

$$\mathcal{E}_1 = \begin{pmatrix} \mathcal{E}_{1\perp} \\ 0 \end{pmatrix}, \quad \mathcal{E}_2 = \begin{pmatrix} \mathcal{E}_{2\perp} \\ 0 \end{pmatrix}$$

This leads to the

Second Fresnel-Arago Law: two fields with parallel polarisation interfere the same way as two fields that are randomly polarised.

This indicates that our initial assumption in the previous sections that all our fields have parallel polarisation is not as limiting as it may have appeared at first.

Suppose now that we have some field

$$\mathcal{E} = \begin{pmatrix} \mathcal{E}_{\perp} \\ \mathcal{E}_{\parallel} \end{pmatrix}$$

which is **randomly polarised**. Suppose we separate the two polarisations, and rotate one so that the two resulting fields are aligned, e.g.

$$\mathcal{E}_1 = \begin{pmatrix} \mathcal{E}_{\perp} \\ 0 \end{pmatrix}, \quad \mathcal{E}_2 = \begin{pmatrix} \mathcal{E}_{\parallel} \\ 0 \end{pmatrix}$$

These fields can not interfere because $\mathcal{E}_{\perp}$ and $\mathcal{E}_{\parallel}$ are incoherent. This leads to

The third Fresnel-Arago Law: the two constituent orthogonal linearly polarised states of natural light cannot interfere to form a readily observable interference pattern, even if rotated into alignment.

External sources in recommended order

1. Veritasium - The original double-slit experiment, starting at 2:15 - Demonstration of an interference pattern obtained with sunlight.
2. MIT OCW - Two-beam Interference - Collimated Beams: Interference of laser light in a Michelson interferometer.
3. MIT OCW - Fringe Contrast - Path Difference: Demonstration of how fringe contrast varies with propagation distance.
4. MIT OCW - Coherence Length and Source Spectrum: Demonstration of how the coherence length depends on the spectrum of the laser light.
5. Lecture - 18 Coherence: Lecture Series on Physics - I: Oscillations and Waves by Prof. S. Bharadwaj, Department of Physics and Meteorology, IIT Kharagpur.
6. Lecture - 19 Coherence: Lecture Series on Physics - I: Oscillations and Waves by Prof. S. Bharadwaj, Department of Physics and Meteorology, IIT Kharagpur.

CHAPTER OVERVIEW

6: Scalar diffraction optics

6.1: What you should know and be able to do after studying this chapter
6.2: Introduction
6.3: Angular Spectrum Method
6.4: Rayleigh-Sommerfeld Diffraction Integral
6.5: Intuition for the Spatial Fourier Transform in Optics
6.6: Fresnel and Fraunhofer Approximations
6.7: Fraunhofer Diffraction Revisited
6.8: Fourier Optics
6.9: Super-resolution

6.1: What you should know and be able to do after studying this chapter

- Understand when the scalar wave equation can be used to propagate fields.
- Be able to derive the angular spectrum decomposition, starting from the scalar wave equation. Be able to interpret the angular spectrum method (also known as the plane wave expansion).
- Know the Rayleigh-Sommerfeld formula; in particular be able to write down the integral over spherical waves with amplitudes proportional to the field in the starting plane.
- Know how to deduce the Fresnel and Fraunhofer approximation of the RayleighSommerfeld integral.
- Understand intuitively in what sense the Fourier transform is linked to resolution.
- Understand why propagation of light leads to loss of resolution (i.e. the evanescent waves disappear).
- Know how the Fresnel and Fraunhofer propagation integrals relate to Fourier transforms.
- Understand why propagation to the far field corresponds to taking the Fourier transform.
- Understand why propagation to the focal plane of a lens corresponds to taking the Fourier transform.
- Understand why the Numerical Aperture (NA) of a lens ultimately determines the resolution of images.
- Understand how a lens can be used for Fourier filtering.

6.2: Introduction

In this chapter we will study how light propagates. The propagation of light reveals its wave-like nature: in the double-slit experiment, we concluded from the interference pattern observed on a screen that light is a wave. To demonstrate more convincingly that light is indeed a wave, we require a detailed quantitative model of the propagation of light, which gives experimentally verifiable predictions.

But a precise description of the propagation of light is not only important for fundamental science, it also has many practical applications. For example, if a sample must be analysed by illuminating it and measuring the scattered light, the fact that the detected light has not only been affected by the sample, but by both the sample and propagation has to be taken into account. Another example is lithography. If a pattern has to be printed onto a substrate using a mask that is illuminated, it has to be realised that when there is a certain distance between the mask and the photoresist, the light which reaches the resist does not have the exact shape of the mask because of propagation effects. Thus, the mask needs to be designed to compensate for this effect.

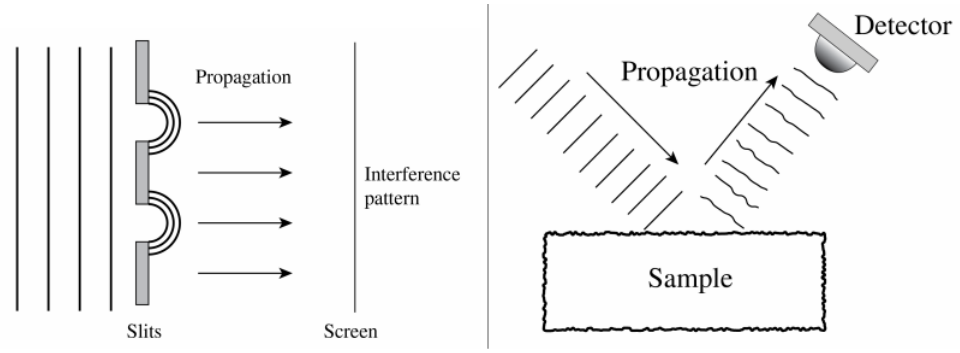


Figure 6.2.1: A quantitative model of the propagation of light may serve fundamentally scientific purposes, since it would provide predictions that can be tested, and can be applied in sample analyses or lithography.

In Section 1.4 we have derived that in homogeneous matter (i.e. the permittivity is constant), for every component of a time-harmonic electromagnetic field $\mathcal{U}(\mathbf{r},t)=\mathrm{Re}\left[U(\mathbf{r})e^{-i\omega t}\right]$, the complex field $U(\mathrm{r})$ satisfies the scalar Helmholtz equation 1.5.14

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

where $k=\omega\sqrt{\epsilon\mu_0}$ is the wave number of the light in matter with permittivity ϵ and refractive index $n=\sqrt{\epsilon/\epsilon_0}$. In Sections 6.2 and 6.3 we will describe two equivalent methods to compute the propagation of the field through homogeneous matter. Although both methods in the end describe the same, they give physical insight into different aspects of propagation, as will be seen in Sections 6.4 and 6.5. The two methods can be applied to propagate any component U of the electromagnetic field, provided the propagation is in homogeneous matter. With this assumption they give identical and rigorous results.

When the refractive index is not constant, Maxwell's equations are no longer equivalent to the wave equation for the individual electromagnetic field components and there is then coupling of the components due to the curl operators in Maxwell's equation. When the variation of the refractive index is slow on the scale of the wavelength, the scalar wave equation may still be a good approximation, but for structures that vary on the scale of the wavelength (i.e. on the scale of ten microns or less), the scalar wave equation is not sufficiently accurate.

6.3: Angular Spectrum Method

Our goal is to derive the field in some point (x, y, z) with $z = 0$, given the field in the plane $z = 0$, as is illustrated in Figure 6.3.1. The sources of the field are assumed to be in the half space $z < 0$. One way to see how light propagates from one plane to another is by using the **angular spectrum method**. We decompose the field in plane waves with a two-dimensional Fourier transform. Since we know how each plane wave propagates, we can propagate each Fourier component separately and then add them all together by taking the inverse Fourier transform. Mathematically, this is described as follows: we know the field $U(x, y, 0)$. We will write $U_0(x, y) = U(x, y, 0)$ for convenience and apply a two-dimensional Fourier transform to U_0 :

$$\mathcal{F}\left(U_0\right)(\xi, \eta) = \iint U_0(x, y) e^{-2\pi i(\xi x + \eta y)} \mathrm{d}x \, \mathrm{d}y,$$

The inverse Fourier transform implies:

$$\begin{aligned} U_0(x, y) &= \iint \mathcal{F}\left(U_0\right)(\xi, \eta) e^{2\pi i(\xi x + \eta y)} \mathrm{d}\xi \mathrm{d}\eta \\ &= \mathcal{F}^{-1}\left\{\mathcal{F}\left(U_0\right)\right\}(x, y). \end{aligned}$$

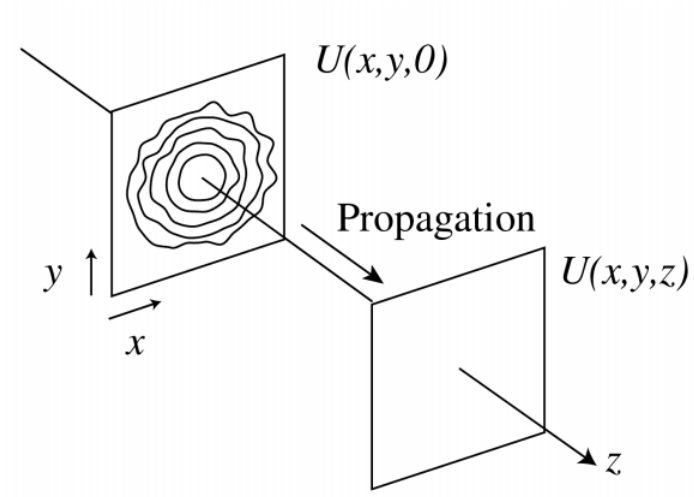


Figure 6.3.1: Given the field $U(x, y, 0)$, we want to find U in a point (x, y, z) with $z > 0$. It is assumed that the field propagates in the positive z-direction, which means that all sources are in the region $z < 0$.

The most important properties of the Fourier transform are listed in Appendix E. By defining \(k_{x}=2 \pi \xi, k_{y}=2 \pi 6.3.2 \) can be written as

$$U_0(x, y) = \frac{1}{4\pi^2} \iint \mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi}, \frac{k_y}{2\pi}\right) e^{i(k_x x + k_y y)} \mathrm{d}k_x \, \mathrm{d}k_y.$$

The variables in the Fourier plane: (ξ, η) and (k_x, k_y) are called **spatial frequencies**.

Equation (6.3.4) says that we can write $U_0(x, y) = U(x, y, z = 0)$ as an integral (a sum) of plane waves with wave vector $\mathbf{k} = (k_x, k_y, k_z)^T$, each with its own weight (i.e. complex amplitude) $\mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi}, \frac{k_y}{2\pi}\right)$. We know how each plane wave with complex amplitude $\mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi}, \frac{k_y}{2\pi}\right)$ and wave vector $\mathbf{k} = (k_x, k_y, k_z)^T$ propagates over a distance $z > 0$

$$\mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi}, \frac{k_y}{2\pi}\right) e^{i(k_x x + k_y y)} \rightarrow \mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi}, \frac{k_y}{2\pi}\right) e^{i(k_x x + k_y y + k_z z)},$$

Therefore, the field $U(x, y, z)$ in the plane z (for some $z > 0$) is given by

$$U(x, y, z) = \frac{1}{4\pi^2} \iint \mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi}, \frac{k_y}{2\pi}\right) e^{i(k_x x + k_y y + k_z z)} \mathrm{d}k_x \, \mathrm{d}k_y,$$

where

$$k_z = \sqrt{\left(\frac{2\pi}{\lambda}\right)^2 - k_x^2 - k_y^2}\,,$$

with λ the wavelength of the light as measured **in the material** (hence, $\lambda = \lambda_0/n$, with λ_0 the wavelength in vacuum). The sign in front of the square root in (\PageIndex{7}\)) could in principle be chosen negative: one would then also obtain a solution of the Helmholtz equation. The choice of the sign of k_z is determined by the direction in which the light propagates, which in turn depends on the location of the sources **and** on the convention chosen for the time dependance. We have to choose here the $+$ sign in front of the square root because the sources are in $z<0$ and the time dependence of time-harmonic fields is (as always in this book) given by $e^{-i\omega t}$ with $\omega>0$.

Eq. (6.3.6) can be written alternatively as

$$U(x,y,z) = \mathcal{F}^{-1}\left\{\mathcal{F}\left(U_0\right)(\xi,\eta)e^{ik_z z}\right\}(x,y),$$

where now k_z is to be interpreted as a function of (ξ,η) :

$$k_z = 2\pi\sqrt{\left(\frac{1}{\lambda}\right)^2 - \xi^2 - \eta^2}\,.$$

Note that one can interpret this as a diagonalisation of the propagation operator, as explained in Appendix F.

We can observe something interesting: if $k_x^2+k_y^2 > \left(\frac{2\pi}{\lambda}\right)^2$, then k_z becomes imaginary, and $\exp(ik_z z)$ decays exponentially for increasing z :

$$\exp\left\{i\left[k_x x + k_y y + z\sqrt{\left(\frac{2\pi n}{\lambda}\right)^2 - k_x^2 - k_y^2}\right]\right\} = e^{i(k_x x + k_y y)}e^{-z\sqrt{k_x^2+k_y^2-\left(\frac{2\pi n}{\lambda}\right)^2}}\,.$$

These exponentially decaying waves are **evanescent** in the positive z-direction. We have met evanescent waves already in the context of total internal reflection discussed in Section 1.9.5. The physical consequences of evanescent waves in the angular spectrum decomposition will be explained in Section 6.4.

The waves for which k_z is real have constant amplitude: only the phase changes due to propagation. These waves therefore are called **propagating waves**.

Remark. In homogeneous space, the scalar Helmholtz equation for every electric field component is equivalent to Maxwell's equations and hence we may propagate each component E_x, E_y and E_z individually using the angular spectrum method. If the data in the plane $z=0$ of these field components are physically consistent, the electric field thus obtained will automatically satisfy the condition that the electric field is free of divergence, i.e.

$$\nabla\cdot\mathbf{E} = 0$$

everywhere in $z>0$. This is equivalent to the statement that the electric vectors of the plane waves in the angular spectrum are perpendicular to their wave vectors. Alternatively, one can propagate only the E_x-and E_y-components and afterwards determine E_z from the condition that (6.3.11) must be satisfied.

6.4: Rayleigh-Sommerfeld Diffraction Integral

Another method to propagate a wave field is by using the **Rayleigh-Sommerfeld** integral. A very good approximation of this integral states that each point in the plane $z = 0$ emits spherical waves, and to find the field in a point (x, y, z), we have to add the contributions from all these point sources together. This corresponds to the Huygens-Fresnel principle postulated earlier in Section 5.6. Because a more rigorous derivation starting from the Helmholtz equation would be complicated and lengthy, we will just give the final result:

$$\begin{aligned} U(x,y,z) &= \frac{1}{i\lambda}\iint U(x',y',0)\,\frac{z e^{ik\sqrt{(x-x')^2+(y-y')^2+z^2}}}{(x-x')^2+(y-y')^2+z^2}\,\mathrm{d}x'\mathrm{d}y' \\ &= \frac{1}{i\lambda}\iint U(x',y',0)\,\frac{z e^{ikr}}{r}\,\mathrm{d}x'\mathrm{d}y' \end{aligned}$$

where we defined

$$r = \sqrt{(x-x')^2+(y-y')^2+z^2}\,.$$

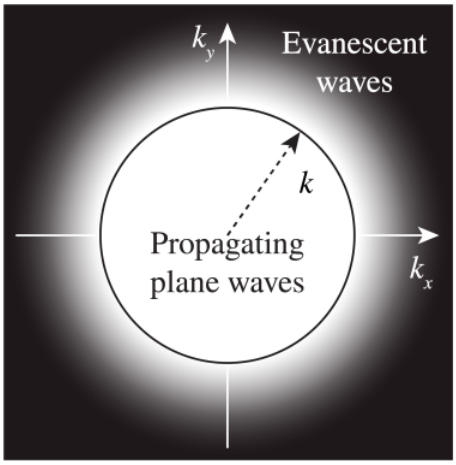


Figure 6.4.1: The spatial frequencies k_x, k_y of the plane waves in the angular spectrum of a time-harmonic field which propagates in the z-direction. There are two types of waves: the propagating waves with spatial frequencies inside the cricle: $\sqrt{k_x^2+k_y^2} < k = 2\pi/\lambda$ and which have phase depending on the propagation distance z but constant amplitude, and the evanescent waves for which $\sqrt{k_x^2+k_y^2} > k$ and of which the amplitude decreases exponentially during propagation.

Remarks.

1. The formula (6.4.1) is not completely rigorous: a term that is a factor $1/(kr)$ smaller (and in practive is therefore is very much smaller) has been omitted.
2. In (6.4.1) there is an additional factor z/r compared to the expressions for a time-harmonic spherical wave as given in (1.53) and at the right-hand side of (5.44). This factor means that the spherical waves in the Rayleigh-Sommerfeld diffraction integral have amplitudes that depend on the angle of radiation (although their wave front is spherical), the amplitude being largest in the forward direction.
3. **Equivalence of the two propagation methods**. The angular spectrum method amounts to a multiplication by $\exp(izk_z)$ in Fourier space, while the Rayleigh-Sommerfeld integral is a convolution. It is one of the properties of the Fourier transform that a multiplication in Fourier space corresponds to a convolution in real space and vice versa. Indeed a mathematical result called **Weyl's identity** implies that the rigorous version of (6.4.1) and the plane wave expansion (i.e. angular spectrum method) give identical results.

6.5: Intuition for the Spatial Fourier Transform in Optics

Since spatial Fourier transformations have played and will play a significant role in our discussion of the propagation of light, it is important to understand them not just mathematically, but also intuitively.

What happens when an object is illuminated and the reflected or transmitted light is detected at some distance from the object? Let us look at transmission for example. When the object is much larger than the wavelength, a transmission function $\tau(x,y)$ is often defined and the field transmitted by the object is then assumed to be simply the product of the incident field and the function $\tau(x,y)$. For example, for a hole in a metallic screen with diameter large compared to the wavelength, the transmission function would be 1 inside the hole and 0 outside. However, if the object has features of the size of the order of the wavelength, this simple model breaks down and the transmitted field must instead be determined by solving Maxwell's equations. This is not easy, but some software packages can do it.

Now suppose that the transmitted electric field has been obtained in a plane $z=0$ very close to the object (a distance within a fraction of a wavelength). This field is called the **transmitted near field** and it may have been obtained by simply multiplying the incident field with a transmission function $\tau(x,y)$ or by solving Maxwell's equations. This transmitted near field is a kind of footprint of the object. But it should be clear that, although it is quite common in optics to speak in terms of "imaging an object", strictly speaking we do not image an object as such, but we image the transmitted (or reflected) near field which is a kind of copy of the object. After the transmitted near field has been obtained, we apply the angular spectrum method to propagate the individual components through homogeneous matter (e.g. air) from the object to the detector plane or to an optical element like a lens.

Let $U_0(x,y)=U(x,y,0)$ be a component of the transmitted near field. The first step is to Fourier transform it, by which the field component is decomposed in plane waves. To each plane wave, characterised by the wave numbers k_x and k_y, the Fourier transform assigns a complex amplitude $\mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi},\frac{k_y}{2\pi}\right)$, the magnitude of which indicates how important the role is which this particular wave plays in the formation of the near field. So what can be said about the object field $U_0(x,y)$, by looking at the magnitude of its spatial Fourier transform $\left|\mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi},\frac{k_y}{2\pi}\right)\right|$?

Suppose $U_0(x,y)$ has sharp features, i.e. there are regions where $U_0(x,y)$ varies rapidly as a function of x and y. To describe these features as a combination of plane waves, these waves must also vary rapidly as a function of x and y, which means that the length of their wave vectors $\sqrt{k_x^2+k_y^2}$ must be large. Thus, the sharper the features that $U_0(x,y)$ has, the larger we can expect $\left|\mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi},\frac{k_y}{2\pi}\right)\right|$ to be for large $\sqrt{k_x^2+k_y^2}$, i.e. high spatial frequencies can be expected to have large amplitude. Similarly, the slowly varying, broad features of $U_0(x,y)$ are described by slowly fluctuating waves, i.e. by $\mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi},\frac{k_y}{2\pi}\right)$ for small $\sqrt{k_x^2+k_y^2}$, i.e. for low spatial frequencies. This is illustrated in Figure 6.5.1.

To investigate these concepts further we choose a certain field, take its Fourier transform, remove the higher spatial frequencies and then invert the Fourier transform. We then expect that the resulting field has lost its sharp features and only retains its broad features, i.e. the image is blurred. Conversely, if we remove the lower spatial frequencies but retain the higher, then the result will only show its sharp features, i.e. its contours. These effects are shown in Figure 6.5.1. Recall that when $k_x^2+k_y^2>\left(\frac{2\pi}{\lambda}\right)^2$, the plane wave decays exponentially as the field propagates. Losing these high spatial frequencies leads to a loss of resolution. Because by propagation through homogeneous space, the information contained in the high spatial frequencies corresponding to evanescent waves is lost (only exponentially small amplitudes of the evanescent waves remain), perfect imaging is impossible, no matter how well-designed an optical system is.

Propagation of light leads to irrecoverable loss of resolution.

It is this fact that motivates near-field microscopy, which tries to detect these evanescent waves by scanning close to the sample, thus obtaining subwavelength resolution, which otherwise is not possible.

So we have seen how we can guess properties of some object field $U_0(x,y)$ given the amplitude of its spatial Fourier transform $\left|\mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi},\frac{k_y}{2\pi}\right)\right|$. But what about the phase of $\mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi},\frac{k_y}{2\pi}\right)$? Although one cannot really guess properties of $U_0(x,y)$ by looking at the phase of $\mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi},\frac{k_y}{2\pi}\right)$ the same way as we can by looking at its amplitude, it is in fact the phase that plays a larger role in defining $U_0(x,y)$. This is illustrated in Figure 6.5.2: if the amplitude information of $\mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi},\frac{k_y}{2\pi}\right)$ is removed,

features of the original $U_0(x,y)$ may still be retrieved. However, if we only know the amplitude $|\mathcal{F}(U_0)(k_x,k_y)|$ but not the phase, then the original object is utterly lost. Thus, the phase of a field $\mathcal{F}(U_0)$ is very important, arguably sometimes even more important than its amplitude. However, we cannot measure the phase of a field directly, only its intensity $I=|\mathcal{F}(U_0)|^2$ from which we can calculate the amplitude $|\mathcal{F}(U_0)|$. It is this fact that makes **phase retrieval** an entire field of study on its own: how can we find the phase of a field, given that we can only perform intensity measurements? This question is related to a new field of optics called "lensless imaging", where amplitudes and phases are retrieved from intensity measurements and the image is reconstructed **computationally**. Interesting as this topic may be, we will not treat it in these notes and refer instead to Master courses in optics.

Remark. The importance of the phase for the field can also be seen by looking at the plane wave expansion (6.5.3). We have seen that the field in a plane $z=$ constant can be obtained by propagating the plane waves by multiplying their amplitudes by the phase factors $\exp(izk_z)$, which depends on the propagation distance z. If one leaves the evanescent waves out of consideration (since after some distance they hardly contribute to the field anyway), it follows that only the phases of the plane waves change upon propagation, while their amplitudes (the moduli of their complex amplitudes) do not change. Yet, depending on the propagation distance z, widely differing light patterns are obtained (see e.g. Figure 6.5.4).

Another aspect of the Fourier transform is the **uncertainty principle**. It states that many waves of different frequencies have to be added to get a function that is confined to a small space. Stated differently, if $U(x,y)$ is confined to a very small region, then $\mathcal{F}(U)(k_x,k_y)$ must be very spread out. This can also be illustrated by the scaling property of the Fourier transform:

$$\text{if } h(x)=f(ax) \text{ then } \mathcal{F}(h)\left(\frac{k_x}{2\pi}\right)=\frac{1}{|a|}\mathcal{F}(f)\left(\frac{k_x}{2\pi a}\right),$$

which simply states that the more $h(x)$ is squeezed by increasing a, the more its Fourier transform $\mathcal{F}(h)$ spreads out. This principle is illustrated in Figure 6.5.3. The uncertainty principle is familiar from quantum physics where it is stated that a particle cannot have both a definite momentum and a definite position. In fact, this is just one particular manifestation of the uncertainty principle just described. A quantum state $|\psi\rangle$ can be described in the position basis $\psi_x(x)$ as well as in the momentum basis $\psi_p(p)$. The basis transformation that links these two expressions is the Fourier transform

$$\psi_p(p)=\mathcal{F}\{\psi_x(x)\}(p).$$

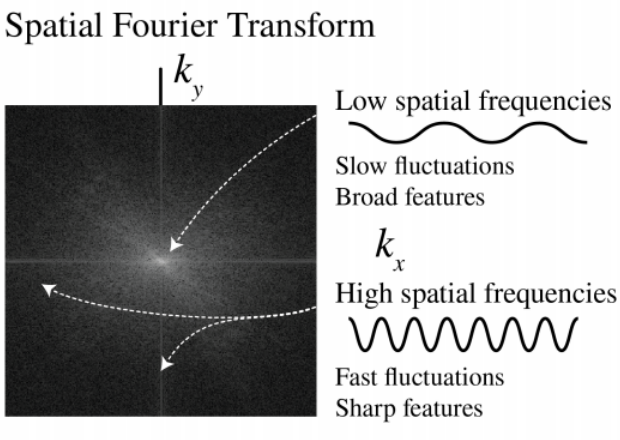


Figure 6.5.1: A qualitative interpretation of spatial Fourier transforms. The low spatial frequencies (i.e. small $\sqrt{k_x^2+k_y^2}$) represent slow fluctuations, and therefore contribute to the broad features of the real-space object. The high spatial frequencies (i.e. large $\sqrt{k_x^2+k_y^2}$) fluctuate rapidly, and can therefore form sharp features in the real-space object.

Hence, the two are obviously subject to the uncertainty principle! In fact, any two quantum observables which are related by a Fourier transform (also called conjugate variables), such as position and momentum, obey this uncertainty relation.

The **uncertainty relation** roughly says:

If a function $f(x)$ has width Δx, its Fourier transform has a width $\Delta k_x \approx 2\pi/\Delta x$.

Since after propagation over a distance z, the evanescent waves do not contribute to the Fourier transform of the field, it follows that this Fourier transform has maximum width $\Delta k_x = k$. By the uncertainty principle it follows that after propagation, the

minimum width of the field is $\Delta x, \Delta y \approx 2\pi/k = \lambda$.

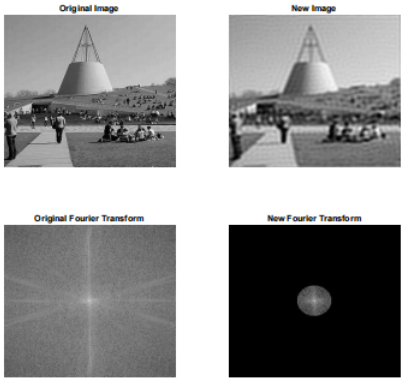
The minimum feature size of a field after propagation is of the order of the wavelength.

This poses a fundamental limit to resolution given by the wavelength of the light.

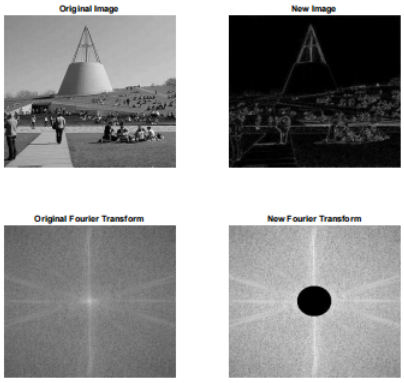
6.6: Fresnel and Fraunhofer Approximations

The Fresnel and Fraunhofer approximation are two approximations of the Rayleigh-Sommerfeld integral (6.13). The approximations are based on the assumption that the field has propagated over a sufficiently large distance z. In the Fraunhofer approximation, z has to be very large, i.e. much larger than for the Fresnel approximation to hold. Putting it differently: in order of most accurate to least accurate (i.e. only valid for large propagation distances), the diffraction integrals would rank as:

[Most accurate] Rayleigh-Sommerfeld $\rightarrow$ Fresnel $\rightarrow$ Fraunhofer **[Least accurate]**.

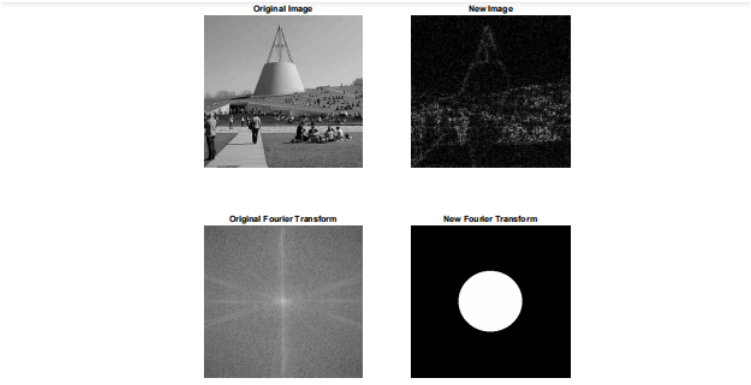


(a) Removing the high spatial frequencies

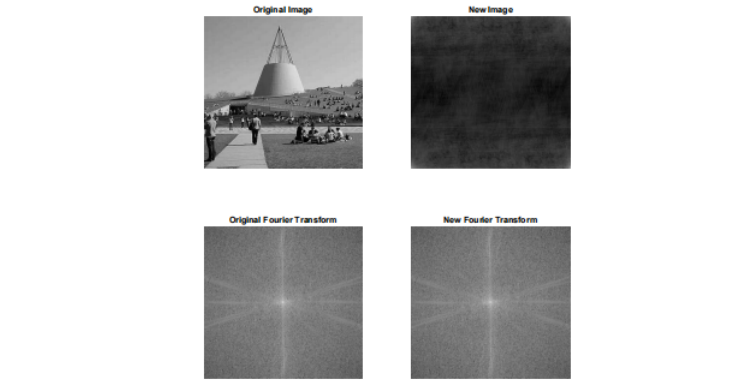


(b) Removing the low spatial frequencies

Figure 6.6.1: Demonstration of the roles of different spatial frequencies. By removing the high spatial frequencies, only the broad features of the image remain: we lose resolution. If the low spatial frequencies are removed, only the sharp features (i.e. the contours) in the image remain.



(a) Removing the amplitude information by setting the amplitude of propagating and evanescent waves to 1 and 0, respectively.



(b) Removing the phase information by setting the phase equal to 0.

Figure 6.6.2: Demonstration of the role of the phase of the spatial Fourier transform. If the amplitude information is removed, but phase information is kept, some features of the original image are still recognizable. However, if the phase information is removed but amplitude information is kept, the original image is completely lost.

6.5.1 Fresnel Approximation

For both approximations, we assume that z in Eq. (6.3.1) is so large that in the denominator we can approximate $r \approx z$

$$\begin{aligned} U(x,y,z) &= \frac{1}{i\lambda}\iint U_0\left(x',y'\right)\frac{z}{r}\frac{e^{ikr}}{r}\,\mathrm{d}x'\mathrm{d}y' \\ &\approx \frac{1}{i\lambda z}\iint U_0\left(x',y'\right)e^{ikr}\,\mathrm{d}x'\mathrm{d}y'. \end{aligned}$$

The reason why we can not apply the same approximation for r in the exponent, is that there r is multiplied by $k = 2\pi/\lambda$, which is very large, so any error introduced by approximating r would be magnified significantly by k and then can lead to a completely different value of $\exp(ikr) = \cos(kr) + i\sin(kr)$. To approximate r in the exponent $\exp(ikr)$ we must be more careful and instead apply a Taylor expansion. Recall that

$$\begin{aligned} r &= \sqrt{\left(x-x'\right)^2 + \left(y-y'\right)^2 + z^2} \\ &= z\sqrt{\frac{\left(x-x'\right)^2 + \left(y-y'\right)^2}{z^2} + 1}. \end{aligned}$$

We know that for a small number s we can expand

$$\sqrt{s+1} = 1 + \frac{s}{2} - \frac{s^2}{8} + \ldots.$$

Since we assumed that z is large, $\frac{(x-x')^2+(y-y')^2}{z^2}$ is small, so we can expand

$$\begin{aligned} r &= z\sqrt{\frac{(x-x')^2+(y-y')^2}{z^2}+1} \\ &\approx z\left(1+\frac{(x-x')^2+(y-y')^2}{2z^2}\right) \\ &= z+\frac{(x-x')^2+(y-y')^2}{2z}, \quad \text{Fresnel approximation.} \end{aligned}$$

With this approximation, we arrive at the **Fresnel diffraction integral**, which can be written in the following equivalent forms:

$$\begin{aligned} U(x,y,z) &\approx \frac{e^{ikz}}{i\lambda z}\iint U_0(x',y')\, e^{\frac{ik}{2z}\left[(x-x')^2+(y-y')^2\right]}\,\mathrm{d}x'\mathrm{d}y' \\ &= \frac{e^{ikz}e^{\frac{ik\left(x^2+y^2\right)}{2z}}}{i\lambda z}\iint U_0(x',y')\, e^{\frac{ik\left(x'^2+y'^2\right)}{2z}}\, e^{-ik\left(\frac{x}{z}x'+\frac{y}{z}y'\right)}\,\mathrm{d}x'\mathrm{d}y' \\ &= \frac{e^{ikz}e^{\frac{ik\left(x^2+y^2\right)}{2z}}}{i\lambda z}\mathcal{F}\left\{U_0(x',y')\, e^{\frac{ik\left(x'^2+y'^2\right)}{2z}}\right\}\left(\frac{x}{\lambda z},\frac{y}{\lambda z}\right). \end{aligned}$$

Especially the last expression is interesting, because it shows that

> The Fresnel integral is the Fourier transform of the field $U_0(x',y')$ multiplied by the Fresnel propagator $\exp\left(\frac{ik\left(x'^2+y'^2\right)}{2z}\right)$.

Note that this propagator depends on the distance of propagation z.

Remark. By Fourier transforming (\(6.6.6)\), one gets the plane wave amplitudes of the Fresnel integral. It turns out that these amplitudes are equal to $\mathcal{F}(U_0)$ multiplied by a phase factor. This phase factor is a paraxial approximation of the exact phase factor given by $\exp(izk_z)$, i.e. it contains as exponent the parabolic approximation of k_z. Therefore the Fresnel approximation is also called the paraxial approximation. In fact, it can be shown that the Fresnel diffraction integral is a solution of the **paraxial wave equation** and conversely, that every solution of the paraxial wave equation can be written as a Fresnel diffraction integral.

6.5.2 Fraunhofer Approximation

For the Fraunhofer approximation, we will make one further approximation to r in $\exp(ikr)$

$$\begin{aligned} r &\approx z+\frac{(x-x')^2+(y-y')^2}{2z} \quad \text{Fresnel approximation} \\ &\approx z+\frac{x^2+y^2-2xx'-2yy'}{2z} \quad \text{Fraunhofer approximation.} \end{aligned}$$

Hence we have omitted the quadratic terms $x'^2+y'^2$, and compared with respect the Fresnel diffraction integral, we simply omit the factor $\exp\left(\frac{ik\left(x'^2+y'^2\right)}{2z}\right)$ to obtain the **Fraunhofer diffraction integral**:

$$U(x,y,z) \approx \frac{e^{ikz}e^{\frac{ik\left(x^2+y^2\right)}{2z}}}{i\lambda z}\mathcal{F}(U_0)\left(\frac{x}{\lambda z},\frac{y}{\lambda z}\right)$$

This leads to the following important observation:

> The Fraunhofer far field of $U_0(x',y')$ is simply its Fourier transform with an additional quadratic phase factor.

Note that the coordinates for which we have to evaluate $\mathcal{F}(U_0)$ scale with $1/z$, and the overall field $U(x,y,z)$ is proportional to $1/z$. This means that as you choose z larger (i.e. you propagate the field further), the field simply spreads out without changing its shape, and its amplitude goes down. Stated differently, apart from the factor $1/z$ in front of the integral, the Fraunhofer field only depends on the angles x/z amd y/z. Therefore the field diverges as the propagation distance z increases.

Eventually, for sufficiently large propagation distances, i.e. in the Fraunhofer limit, light always spreads out without changing the shape of the light distribution.

Remarks.

1. The Fresnel integral is, like the Fraunhofer integral, also a Fourier transform, evaluated in spatial frequencies which depend on the point of observation:

$$\xi = \frac{x}{\lambda z}, \quad \eta = \frac{y}{\lambda z}.$$

   However, in contrast to the Fraunhofer integral, the Fresnel integral depends additionally in a different way on the propagation distance z, namely through the epxonent of the propagator in the integrand. This is the reason that the Fresnel integral does not merely depend on z through the ratios x/z and y/z but in a more complicated manner. Therefore the Fresnel integral gives quite diverse patterns depending on the value of the propagation distance z, as is shown in Figure 6.6.4.
2. Let $f_{a,b}(x,y) = f(x-a, y-b)$ be the function obtained from f by translation. From the general property of the Fourier transform:

$$\mathcal{F}\left(f_{a,b}\right)(\xi,\eta) = e^{-2\pi i(\xi a+\eta b)}\,\mathcal{F}(f)(\xi,\eta).$$

   Hence, when the field U_0 is translated, the intensity in the Fraunhofer far field is not changed. In contrast, due to the additional quadratic phase factor in the integrand of the Fresnel integral, the intensity of the Fresnel field in general changes when U_0 is translated.
3. Suppose that U_0 is the field immediately behind an aperture $\mathcal{A}$ with diameter D in an opaque screen. It can then be shown that points (x,y,z) of observation, for which the Fresnel and Fraunhofer diffraction integrals are sufficiently accurate, satisfy:

$$\frac{z}{\lambda} > \left(\frac{\max_{(x',y')\in\mathcal{A}}\sqrt{(x-x')^2+(y-y')^2}}{\lambda}\right)^{4/3}, \text{ Fresnel}$$

$$\frac{z}{\lambda} > \left(\frac{D}{\lambda}\right)^2, \text{ Fraunhofer}$$

   Suppose that $D = 1\ \text{mm}$ and the wavelength is green light: $\lambda = 550\ \text{nm}$, then Fraunhofer's approximation is accurate if $z > 2\ \text{m}$. The inequality (6.6.6) is sufficient for the Fresnel formula to be accurate, but it is not always necessary. Often the Fresnel approximation is already accurate for smaller propagation distances.
4. The points of observation where the Fraunhofer formulae can be used must in any case satisfy:

$$\frac{x}{z} < 1, \quad \frac{y}{z} < 1.$$

   When $x/z > 1$, the spatial frequency $k_x = \frac{2\pi x}{z\lambda} > k$ associated with this point corresponds to an evanescent wave. An evanescent wave obviously cannot contribute to the Fraunhofer far field because it exponentially decreases with distance z.
5. In any expression for an optical field, one may always omit factors of constant phase, i.e. an overall phase which does not depend on position. If one is only interested in the field in certain planes $z =$ constant, then a factor like $\exp(ikz)$ may also be omitted. Further, in some cases also a position dependent phase factor in front of the Fresnel and Fraunhofer diffraction integrals is omitted, namely when only the intensity is of interest. In exercises it is usually mentioned that this factor may be omitted: if this is not stated, it should be retained in the formulae.

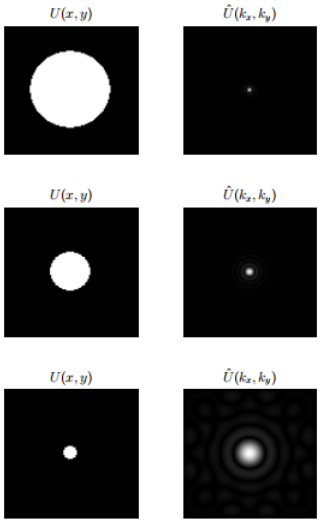


Figure 6.6.3: Demonstration of the uncertainty principle. The more confined $U(x,y)$ is, the larger the spread of $\mathcal{F}(U)$ $\left(\frac{k_x}{2\pi},\frac{k_y}{2\pi}\right)$.

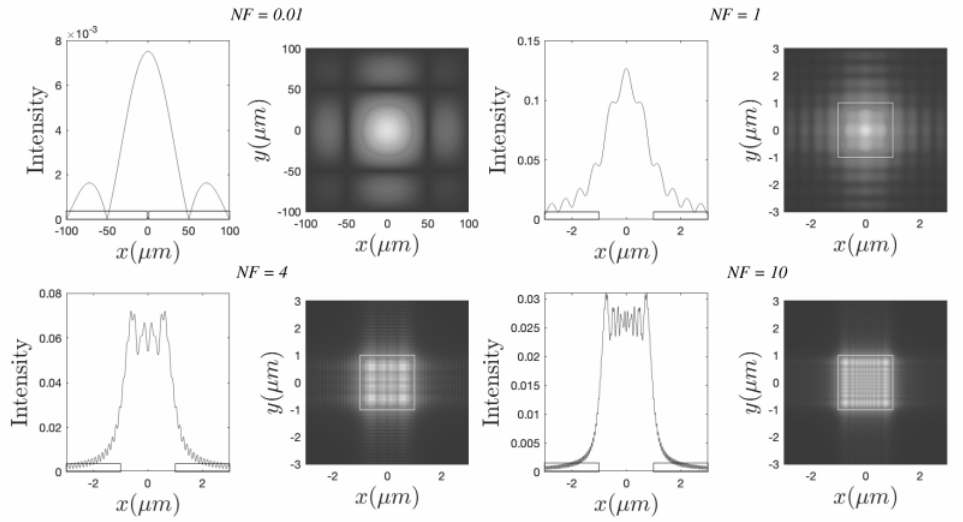


Figure 6.6.4: Example of intensities of Fresnel fields of a square aperture, shown as contours and in cross section for different distances which increase from the lower right to the upper left. The upper left pattern is equal to the Fraunhofer pattern. $N_F = D^2/\lambda z$ is the Fresnel number and $\lambda = 600$ nm.

6.5.3 Examples of Fresnel and Fraunhofer fields

Fresnel approximation of the field of two point sources.

Consider two point sources in $\mathbf{r}_s^+ = (a/2, 0, 0)$ and $\mathbf{r}_s^- = (-a/2, 0, 0)$. The fields of each of them in a point $\mathbf{r} = (x, y, z)$ are given by (5.6.2)

$$U_\pm(\mathbf{r}) = \frac{e^{ik|\mathbf{r}-\mathbf{r}_s^\pm|}}{|\mathbf{r}-\mathbf{r}_s^\pm|}$$

We apply the Fresnel approximation for large z :

$$\begin{aligned}\left|\mathbf{r}-\mathbf{r}_s^{\pm}\right| &= z\sqrt{1+\frac{(x\mp a/2)^2+y^2}{z^2}} \\ &\approx z+\frac{(x\mp a/2)^2+y^2}{2z} \\ &= z+\frac{x^2+y^2+a^2/4}{2z}\mp\frac{ax}{2z}.\end{aligned}$$

Hence,

$$U_{\pm}(\mathbf{r}) \approx \frac{e^{ikz}}{z} e^{ik\frac{x^2+y^2}{2z}} e^{ik\frac{a^2}{8z}} e^{\mp ik\frac{ax}{2z}},$$

where in the denominator we replaced $\left|\mathbf{r}-\mathbf{r}_s^{\pm}\right|$ by z. Note that the Fraunhofer approximation amounts to $e^{ika^2/(8z)} \approx 1$ while the phase factor $e^{ik\frac{x^2+y^2}{2z}}$ remains. The intensity on a screen $z=$ constant of the total field is

$$\begin{aligned}I_{\text{tot}}(\mathbf{r}) &= \left|U_+(\mathbf{r})+U_-(\mathbf{r})\right|^2 = \frac{1}{z^2}\left|e^{-ik\frac{ax}{2z}}+e^{ik\frac{ax}{2z}}\right|^2 \\ &= \frac{2}{z^2}\left[1+\cos\left(2\pi\frac{ax}{\lambda z}\right)\right]\end{aligned}$$

It is seen that the intensity results from the interference of two plane waves: $\exp[\pm ikax/(\lambda z)]$ and is given by a cosine function (see Figure $6.6.5$). Note that for two point sources, the intensity pattern is the same in the Fresnel and the Fraunhofer approximation. However, this is special for two point sources: when more than two point sources are considered, the Fresnel and Fraunhofer patterns are different. The intensity pattern is independent of y, and vanishes on lines

$$\frac{x}{z} = (2m+1)\frac{\lambda}{2a}$$

and has maxima on lines

$$\frac{x}{z} = m\frac{\lambda}{a},$$

for integer m.

Fraunhofer field of a rectangular aperture in a screen.

Let the screen be $z=0$ and the aperture be given by $-a/2<x<a/2, -b/2<y<b/2$. The transmission function $\tau(x,y)$ is:

$$\tau(x,y) = 1_{[-a/2,a/2]}(x)1_{[-b/2,b/2]}(y),$$

where

$$1_{[-a/2,a/2]}(x) = \begin{cases} 1, \text{ if } -\frac{a}{2}\le x\le\frac{a}{2}, \\ 0, \text{ otherwise }, \end{cases}$$

and similarly for $1_{[-b/2,b/2]}(y)$. Let the slit be illuminated by a perpendicular incident plane wave with unit amplitude. Then the field immediately behind the screen is:

$$U_0(x,y) = \tau(x,y) = 1_{[-a/2,a/2]}(x)1_{[-b/2,b/2]}(y),$$

We have

$$\begin{aligned}\mathcal{F}\left(1_{[-a/2,a/2]}\right)(\xi) &= \int_{-a/2}^{a/2} e^{-2\pi i\xi x}\,\mathrm{d}x \\ &= \frac{e^{\pi ia\xi}-e^{-\pi ia\xi}}{2\pi i\xi} \\ &= a\frac{\sin(\pi a\xi)}{\pi a\xi} \\ &= a\,\mathrm{sinc}(\pi a\xi),\end{aligned}$$

where $\mathrm{sinc}(u) = \sin(u)/u$. Hence,

$$\mathcal{F}\left(U_0\right)\left(\frac{x}{\lambda z}, \frac{y}{\lambda z}\right) = ab\,\mathrm{sinc}\left(\frac{\pi a x}{\lambda z}\right)\mathrm{sinc}\left(\frac{\pi b y}{\lambda z}\right).$$

The Fraunhofer far field of a rectangular aperture in a plane at large distance z is obtained by substituting (6.6.25) into (6.6.9).

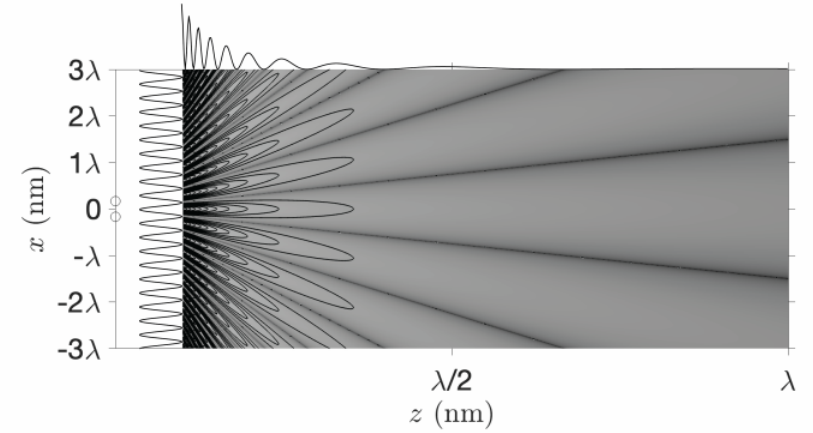


Figure 6.6.5: Fraunhofer intensity pattern of two coherent point sources 200 nm apart (circles), of equal strength for $\lambda = 600$ nm.

Remarks.

1. The first zero along the x-direction from the centre $x = 0$ occurs for

$$x = \pm\frac{\lambda z}{a}.$$

   The distance between the first two zeros along the x-axis is $2\lambda z/a$ and is thus larger when the width along the x-direction of the aperture is smaller.
2. The inequalities (6.6.14) imply that when $a < \lambda$, the far field pattern does not have any zeros as function of x. It is then difficult or even impossible to deduce the width a from the Fraunhofer intensity. This is an illustration of the fact that information about sizes less than the wavelength cannot propagate to the far field.
3. As shown in Figure 6.6.6, the Fraunhofer diffraction pattern as function of diffraction angle is narrowest in the direction in which the aperture is widest.

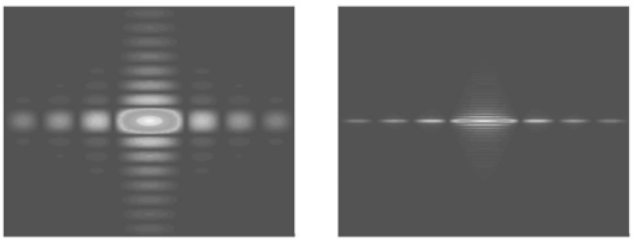


Figure 6.6.6: Fraunhofer diffraction pattern of a rectangular aperture in an opaque screen. In a) the width of the aperture in the y-direction is twice that in the x direction. In b) the width in the y-direction is 10 times that in the x-direction.

Periodic array of slits

We can now predict what the diffraction pattern of a series of slits of finite width will look like. It follows from the Fraunhofer pattern of a single rectangular aperture that, if the sides parallel to a a direction are very long, the Fraunhofer diffraction pattern as function of angle in that direction is very narrow. In Figure 6.6.6 b the Fraunhofer diffraction pattern of a rectangular aperture is shown, of which the width in the y-direction is 10 times that in the x-direction. The diffraction pattern is then strongly concentrated along the x-axis. If we only consider the Fraunhofer pattern for $y/z = 0$ while still considering it as a function of x/z, it suffices to compute the Fourier transform only with respect to x. The problem then becomes a diffraction problem for a one-dimensional slit.

We consider now an array of such slits of which the long sides are all parallel to the y-axis and neglect from now on the y-variable. Suppose $W_{\text{slit}}\left(x\right)$ is a block function describing the transmission function of a single slit. We define the Dirac comb by

$$\mathrm{II}_\Delta(x) = \sum_{m=-\infty}^{\infty} \delta(x - m\Delta).$$

Then the transmission function of an infinite series of slits with finite width is given by the convolution $\mathrm{II}_\Delta(x) * W_{\text{slit}}(x)$. To make the number of slits finite, we multiply the expression with another block function $W_{\text{array}}(x)$ and get

$$\tau(x) = (\mathrm{II}_\Delta(x) * W_{\text{slit}}(x))\, W_{\text{array}}(x).$$

The diffraction pattern in the far field is given by the Fourier transform of the transmitted near field. If the incident illumination is a perpendicular plane wave with unit amplitude, the transmitted near field is simply $\tau(x)$. Using the fact that convolutions in real space correspond to products in Fourier space and vice versa, and using the fact that

$$\mathcal{F}\{\mathrm{W}_\Delta(x)\} = (1/\Delta)\mathrm{W}_{1/\Delta}(\xi),$$

see Appendix (E.9) and (E.10), we find

$$\mathcal{F}(\tau) = \frac{1}{\Delta}\left[\mathrm{W}_{1/\Delta}\mathcal{F}\left(W_{\text{slit}}\right)\right] * \mathcal{F}\left(W_{\text{array}}\right).$$

If the slit has width a :

$$\begin{aligned}\frac{1}{\Delta}\mathrm{I}_{1/\Delta}\mathcal{F}\left(W_{\text{slit}}\right)(\xi) &= \frac{a}{\Delta}\sum_{m=-\infty}^{\infty}\delta\left(\xi - \frac{m}{\Delta}\right)\operatorname{sinc}(\pi a\xi) \\ &= \frac{a}{\Delta}\sum_{m=-\infty}^{\infty}\operatorname{sinc}\left(m\pi\frac{a}{\Delta}\right)\delta\left(\xi - \frac{m}{\Delta}\right).\end{aligned}$$

If the total width of the array is A, then

$$\mathcal{F}\left(W_{\text{array}}\right)(\xi) = A\operatorname{sinc}(\pi A\xi),$$

and we conclude that

$$\mathcal{F}(\tau)(\xi) = \frac{aA}{\Delta}\sum_{m=-\infty}^{\infty}\operatorname{sinc}\left(m\pi\frac{a}{\Delta}\right)\operatorname{sinc}\left(\pi A\left(\xi - \frac{m}{\Delta}\right)\right).$$

The Fraunhofer field of the array of slits is (omitting the quadratic phase factor):

$$\mathcal{F}(\tau)\left(\frac{x}{\lambda z}\right) = \frac{aA}{\Delta}\sum_{m=-\infty}^{\infty}\operatorname{sinc}\left(m\pi\frac{a}{\Delta}\right)\operatorname{sinc}\left(\pi A\left(\frac{x}{\lambda z} - \frac{m}{\Delta}\right)\right).$$

For the directions

$$\frac{x}{z} = \theta_m = \frac{m\lambda}{\Delta}, \quad m = 0, \pm 1, \pm 2, \ldots, \quad \text{diffraction orders}$$

the field has local maxima (peaks). These directions are called diffraction orders. Note that as explained above, there should hold: $x/z < 1$ in the Fraunhofer far field, which sets a limit to the number of the diffracted orders that occur. This limit depends on the period and the wavelength and is defined by:

$$|m| \leq \Delta/\lambda.$$

Hence, the larger the ratio of the period and the wavelength, the more diffraction orders.

The width of a diffraction order is given by the width of the function (6.6.31), i.e. it is given by

$$\Delta\theta = \frac{\lambda}{A}, \quad \text{angular width of a diffraction order.}$$

Hence, the larger A, i.e. the more slits there are in the array, the narrower the peaks into which the energy is diffracted.

The property (6.6.34) that the angles of diffraction of the orders depend on wavelength is used to separate wavelengths. **Grating spectrometers** use periodic structures such as this array of slits to very accurately separate and measure wavelengths. For example, for a grating with 1000 periods one can obtain a resolution of $\Delta\lambda/\lambda = 10^{-3}$.

The amplitudes of the diffracted orders:

$$\operatorname{sinc}\left(m\pi\frac{a}{\Delta}\right),$$

are determined by the width of the slits. Hence the envelope (i.e. large features) of the Fraunhofer diffraction pattern is determined by the small-scale properties of the array, namely the width of the slits. This is illustrated in Figure 6.6.7. **Remark**. A periodic row of slits is an example of a diffraction grating. A grating is a periodic structure, i.e. the permittivity is a periodic function of position. Structures can be periodic in one, two and three directions. A crystal acts as a three-dimensional grating whose period is the period of the crystal, i.e. a few Angstrom. Electromagnetic waves of wavelength equal to one Angstrom or less are called x-rays. When a beam of x-rays illuminates a crystal, a detector in the far field measures the Fraunhofer diffraction pattern given by the intensity of the Fourier transform of the refracted near field. These diffraction orders of crystals for x-rays where discovered by Von Laue and are used to study the atomic structure of crystals.

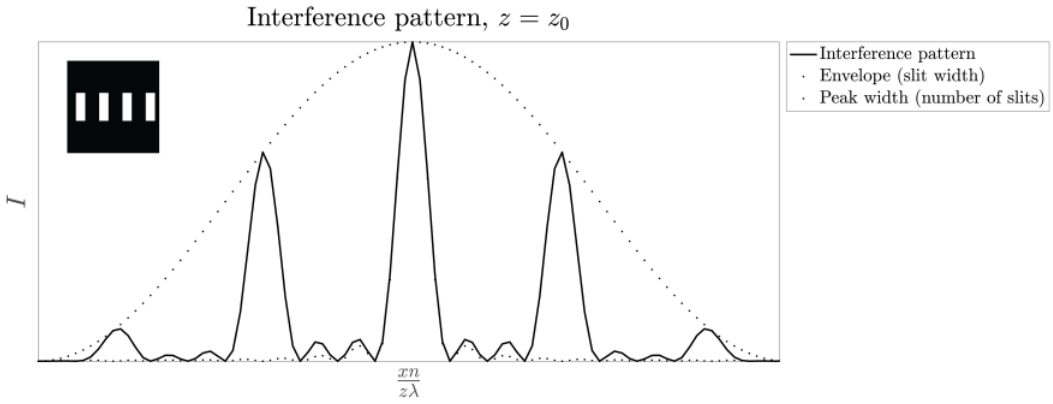


Figure 6.6.7: An illustration of a diffraction pattern of a series of four slits.

6.7: Fraunhofer Diffraction Revisited

Fraunhofer diffraction patterns can qualitatively explained by considering directions in which destructive and constructive interference occurs. Consider two mutually coherent point sources S_1, S_2 on the x-axis as shown in Figure 6.7.1. We assume that these point sources are in phase. On a screen at large distance z an interference pattern is observed. If the distance z of the screen is very large, the spherical wave fronts emitted by the point sources are almost plane at the screen. In point P on the screen at a distance x above the z-axis the optical path differences of the waves emitted by the two sources is approximately given by $S_2Q = a\theta$, where $\theta = x/z$ is assumed small. Hence constructive interference occurs for angles θ such that $S_2Q = m\lambda$ for some integer m, i.e. when

$$\theta = m\frac{\lambda}{a}, \quad \text{constructive interference.}$$

Destructive interference occurs when the path length difference satisfies $S_2Q = \lambda/2 + m\lambda$ for some integer m, hence for angles

$$\theta = (m + 1/2)\frac{\lambda}{a} \quad \text{destructive interference.}$$

If the point sources have the same strength, their fields perfectly cancel for these angles.

Now consider a slit as shown in Figure 6.7.2 which is illuminated by a perpendicular incident plane wave. By the Huygens-Fresnel principle, the field at a screen far from the slit is the sum of the fields of point sources in the aperture. Since the slit is illuminated by a plane wave at perpendicular incidence, all point sources are in phase and have equal strength. Divide the slit in two equal halves as shown in Figure 6.7.2. The point sources in the slit can be arranged into pairs, of which one point source is in the upper half of the slit and the other is at the equivalent position (at distance $a/2$ from the other point source) in the lower half of the slit. Let θ be an angle for which the two point sources of a pair cancel each other i.e.

$$\theta = (m + 1/2)\frac{\lambda}{a/2} = (1 + 2m)\frac{\lambda}{a},$$

since the distance between the point sources is $a/2$. By translating the pair of point sources through the slits, it follows that both half slits perfectly cancel each other for these angles.

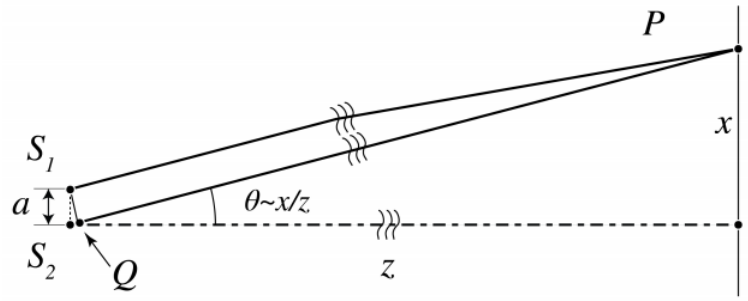


Figure 6.7.1: Interference of to mutually coherent point sources. For z very large points P where constructive and destructive interference occurs are such that for some integer m : $S_2Q = m\lambda$ and $S_2Q = (1/2 + m)\lambda$, respectively

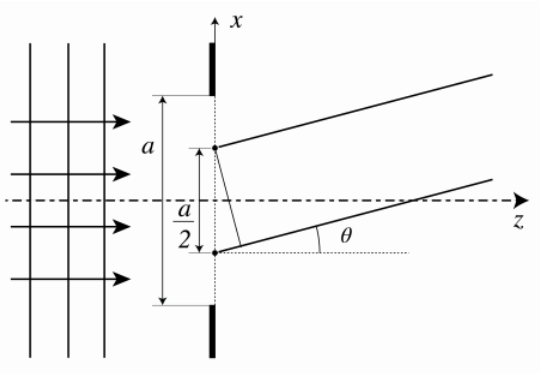


Figure 6.7.2: By dividing the slit into two slits of size $a/2$ each and considering pairs of point sources, of which one is in one half of the slit and the other is at the corresponding position in the other half, angles where destructive interference occurs between these point sources lead to minima in the diffraction pattern. Note that the point sources have corresponding positions in the two parts of the slit if their distance is $a/2$.

Next consider a diffraction grating with period Δ. It follows from Figure 6.7.3 that there will be constructive interference between adjacent periods, and hence for all periods, for angles for which the distance SQ in Fig 6.7.3 is a multiple of the wavelength, i.e. for

$$\theta = m\frac{\lambda}{\Delta},$$

which corresponds to the direction of the diffraction orders. For other angles the phases of the fields of the different periods differ widely and therefore the fields almost cancel at these angles.

This explains that for a diffraction grating consisting of many periods, the far field intensity is high only in particular directions, depending on the wavelength and the period of the grating.

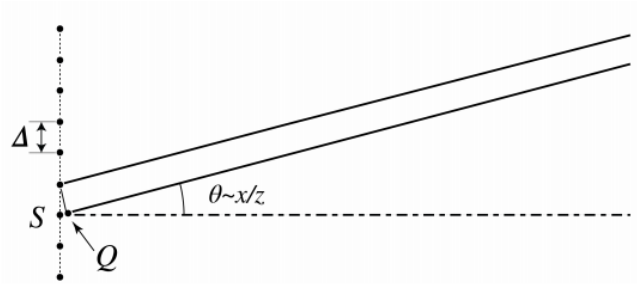


Figure 6.7.3: If the angle θ is such that SQ is a multiple of the wavelength, two adjacent periods, and hence all periods of the grating, constructively interfere. These angles correspond to the diffraction orders.

6.8: Fourier Optics

In this section we apply diffraction theory to a lens. We consider in particular the focusing of a parallel beam and the imaging of an object.

6.8.1 Focusing of a Parallel Beam

A lens induces a local phase change to an incident field which is proportional to its local thickness. Let a plane wave which propagates parallel to the optical axis be incident on the lens. According to Gaussian geometrical optics, the incident rays are parallel to the optical axis and are all focused into the focal point. According to the Principle of Fermat, all rays have travelled the **same optical distance** when they intersect in the focal point where constructive interference occurs and the intensity is maximum. In the focal region, the **wavefronts** are **spheres with centre the focal point** which are cut off by the cone with the focal point as top and opening angle $2a/f$, as shown in Figure $6.8.1$. Behind the focal point, there is a second cone where there are again spherical wavefronts, but there the light is of course propagating away from the focal point. According to Gaussian geometrical optics it is in image space completely dark outside of the two cones in Figure $6.8.1$. However, as we will show, in diffraction optics this is not true.

We assume that the lens is thin and choose as origin of the coordinate system the centre of the thin lens with the positive z-axis along the optical axis. Let f be the focal distance according to Gaussian geometrical optics. Then $(0,0,f)$ is the focal point. Let (x,y,z) be a point between the lens and the focal point. According to geometrical optics the field in (x,y,z) is

$$\begin{cases} \dfrac{e^{-ik\sqrt{x^2+y^2+(z-f)^2}-i\omega t}}{\sqrt{x^2+y^2+(z-f)^2}}, & \text{if } (x,y,z) \text{ is inside the cone,} \\ 0, & \text{if } (x,y,z) \text{ is outside the cone,} \end{cases}$$

where we have included the time-dependence. Indeed the surfaces of constant phase:

$$-\sqrt{x^2+y^2+(z-f)^2}\ -\omega t = \text{ constant},$$

are spheres with centre the focal point which for increasing time converge to the focal point, while the amplitude of the field increases in proportion to the reciprocal of the distance to the focal point so that energy is preserved.

Remark. For a point (x,y,z) to the **right** of the focal point, the spherical wave fronts propagate **away** from the focal point and therefore for $z>f$, $-ik$ should be replaced by $+ik$ in the exponent in $6.8.1$.

The exit pupil of the lens is in the plane $z=0$ where according to ($6.8.1$) the field is

$$1_{\odot_a}(x,y)\frac{e^{-ik\sqrt{x^2+y^2+f^2}}}{\sqrt{x^2+y^2+f^2}},$$

where the time dependence has been omitted and $1_{\odot_a}(x,y)=$

$$\begin{cases} 1 & \text{if } x^2+y^2<a^2, \\ 0 & \text{otherwise} \end{cases} \tag{6.8.1}$$

i.e. $1_{\odot_a}(x,y)=1$ for (x,y) in the exit pupil of the lens and $=0$ otherwise.

In diffraction optics we compute the field in the focal region using diffraction integrals instead of using ray tracing. Hence, the modification introduced by diffraction optics is due to the more accurate propagation of the field from the exit pupil to the focal region.

If a/f is sufficiently small, we may replace the distance $\sqrt{x^2+y^2+f^2}$ between a point in the exit pupil and the focal point in the denominator of ($6.8.2$) by f. This is not allowed in the exponent, however, because of the multiplication by the large wave number k. In the exponent we therefore use instead the first two terms of the Taylor series, 6.5.4:

$$\sqrt{x^2+y^2+f^2}=f\sqrt{1+\frac{x^2+y^2}{f^2}}\approx f+\frac{x^2+y^2}{2f},$$

which is valid for a/f sufficiently small. Then ($6.8.2$) becomes:

$$1 \odot_a (x,y) e^{-ik\frac{x^2+y^2}{2f}},$$

where we dropped the constant factors e^{ikf} and $1/f$. For a general incident field $U_0(x,y)$ in the entrance pupil, the lens applies a transformation such that the field in the exit plane becomes:

$$U_0(x,y) \to U_0(x,y) 1_{\odot_a}(x,y) e^{-ik\frac{x^2+y^2}{2f}},$$

transformation applied by a lens between its entrance and exit plane.

The function that multiplies $U_0(x,y)$ is the **transmission function of the lens**:

$$\tau_{lens}(x,y) = 1 \odot_a (x,y) e^{-ik\frac{x^2+y^2}{2f}}.$$

This result makes sense: in the centre $(x,y) = 0$ the lens is thickest, so the phase is shifted the most (but we can define this phase shift to be zero because only phase differences matter, not absolute phase). As is indicated by the minus-sign in the exponent, the further you move away from the centre of the lens, the less the phase is shifted. For shorter f, the lens focuses more strongly, so the phase shift changes more rapidly as a function of the radial coordinate. Note that transmission function (6.8.7) has modulus 1 so that energy is conserved. We use the Fresnel diffraction integral (6.5.6) to propagate the field (6.8.6) to the focal region:

$$U(x,y,z) = \frac{e^{ikz} e^{\frac{ik\left(x^2+y^2\right)}{2z}}}{i\lambda z} \mathcal{F}\left\{ U_0(x',y') 1 \odot_a (x',y') e^{ik\frac{x'^2+y'^2}{2}} \left(\frac{1}{z} - \frac{1}{f}\right) \right\} \left(\frac{x}{\lambda z}, \frac{y}{\lambda z}\right).$$

The intensity $I = |U|^2$ is shown at the bottom left of Figure 6.8.1. It is seen that the intensity does not monotonically increase for decreasing distance to the focal point. Instead, secondary maxima are seen along the optical axis. Also the boundary of the light cone is not sharp, as predicted by geometrical optics, but diffuse. The bottom right of Figure 6.8.1 shows the phase in the focal region. The wave fronts are close to but not exactly spherical inside the cones.

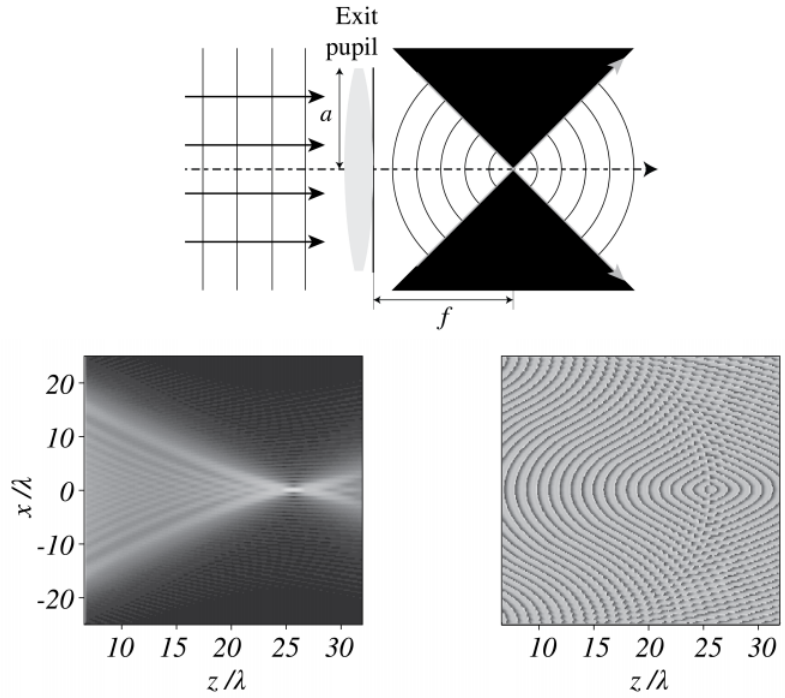


Figure 6.8.1: Top: wavefronts of the incident plane wave and the focused field according to Gaussian geometrical optics. There is no light outside of the two cones. Bottom left: amplitude as predicted by diffraction optics. The boundary of the cones is diffuse and it is not absolutely dark outside of the cones. Furthermore, the intensity does not increase monotonically with decreasing distance to the focal point, as predicted by geometrical optics. Bottom right: phase of the focused field as predicted by diffraction optics.

For points in the back focal plane of the lens, i.e. $z = f$, we have

$$U(x,y,f)=\frac{e^{ikf}e^{\frac{ik\left(x^2+y^2\right)}{2f}}}{i\lambda f}\mathcal{F}\left\{U_0\left(x',y'\right)1\odot_a\left(x',y'\right)\right\}\left(\frac{x}{\lambda f},\frac{y}{\lambda f}\right),$$

which is the same as the Fraunhofer integral! Thus, the field in the focal plane according to Gaussian geometrical optics is in diffraction optics identical to the far field of the field in the entrance pupil of the lens, or to put it differently:

The field in the entrance pupil of the lens and the field in the focal plane are related by a Fourier transform (apart from a quadratic phase factor in front of the integral).

It can be shown that the fields in the front focal plane $U(x,y,-f)$ and the back focal plane $U(x,y,f)$ are related **exactly** by a Fourier transform, i.e. without the additional quadratic phase factor.

So a lens performs a Fourier transform. Let us see if that agrees with some of the facts we know from geometrical optics.

1. We know from Gaussian geometrical optics that if we illuminate a lens with rays parallel to the optical axis, these rays all intersect in the focal point. This corresponds with the fact that for $U_0(x,y)=1$ (i.e. plane wave illumination, neglecting the finite aperture of the lens, i.e. neglecting diffraction effects due to the finite size of the pupil), its Fourier transform is a delta peak:

$$\mathcal{F}\left(U_0\right)\left(\frac{k_x}{2\pi},\frac{k_y}{2\pi}\right)=\delta\left(\frac{k_x}{2\pi}\right)\delta\left(\frac{k_y}{2\pi}\right),$$

which represents the perfect focused spot (without diffraction).

2. If in Gaussian geometrical optics we illuminate a lens with tilted parallel rays of light (a plane wave propagating in an oblique direction), then the point in the back focal plane where the rays intersect is laterally displaced. A tilted plane wave is described by $U_0(\mathbf{r})=\exp(i\mathbf{k}_0\cdot\mathbf{r})$, and its Fourier transform with respect to (x,y) is given by

$$\mathcal{F}\left\{U_0\right\}\left(\frac{k_x}{2\pi},\frac{k_y}{2\pi},z\right)=\delta\left(\frac{k_x-k_{0,x}}{2\pi}\right)\delta\left(\frac{k_y-k_{0,y}}{2\pi}\right),$$

which is indeed a shifted delta peak (i.e. a shifted focal spot).

It seems that the diffraction model of light confirms what we know from geometrical optics. But in the previous two examples we discarded the influence of the finite size of the pupil, i.e. we have left out of consideration the function $1_{\odot_a}$ in calculating the Fourier transform. If $U_0(x,y)=1$ in the entrance pupil and we take the finite size of the pupil properly into account, the δ-peaks become blurred: the focused field is then given by the Fourier transform of the circular disc with radius a. evaluated at spatial frequencies $\xi=\frac{x}{\lambda f},\eta=\frac{y}{\lambda f}$. This field is called the Airy spot and is given by (See Appendix E.17):

$$U(x,y,z)=\frac{\pi a^2}{\lambda f}\frac{2J_1\left(2\pi\frac{a}{\lambda f}\sqrt{x^2+y^2}\right)}{\frac{2\pi a}{\lambda f}\sqrt{x^2+y^2}},\qquad\text{Airy pattern for focusing,}$$

where we have omitted the phase factors in front of the Fourier transform. The pattern is shown in Figure 6.8.1. It is circular symmetric and consists of a central maximum surrounded by concentric rings of alternating zeros and secondary maxima with rapidly decreasing amplitudes. In cross-section, as function of $r=\sqrt{x^2+y^2}$, the Airy pattern is very similar (but not identical) to the sinc-function. From the uncertainty principle shown in Figure 6.5.3 it follows that the size of the focal spot decreases as a increases, and from (6.8.11) we see that the Airy function is a function of the dimensionless variable $ar/(\lambda f)$. Hence the focal spot becomes narrower as $a/(\lambda f)$ increases. The Numerical Aperture (NA) is defined by

$$\mathrm{NA}=\frac{a}{f},\qquad\text{numerical aperture.}$$

Since the first zero of the Airy pattern occurs for $2\,\mathrm{ar}/(\lambda f)=1.22$, the size of the focal spot can be estimated as

$$Size\ of\ focal\ spot\approx 0.6\frac{\lambda}{NA}$$

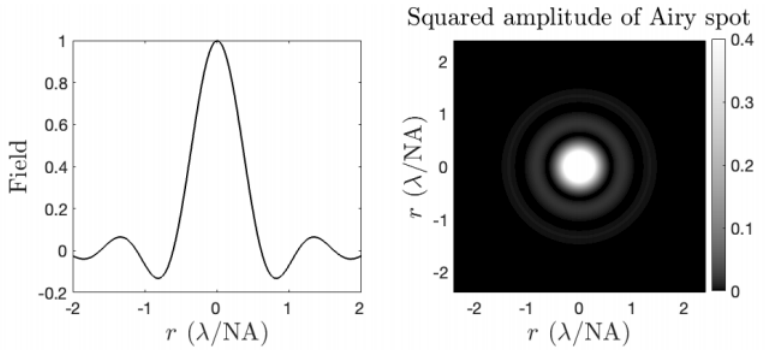


Figure 6.8.2: Left: cross section of the field of the Airy pattern. Right: intensity pattern of the Airy pattern.

6.8.2 Imaging by a lens

It follows from the derivations in the previous section that the Airy pattern is the image of a point source infinitely far in front of a lens. In this section we study the imaging of a general object. Consider first a point object at $z = s_o < 0$ in front of a lens with focal distance $f > 0$. The field in image space is derived similar to the focused field in the previous section. We postulate that the lens transforms the field radiated by the point object into a spherical wave in the exit pupil, which converges to the ideal image point of Gaussian geometrical optics. If we propagate this spherical field in the exit pupil to the image plane using the Fresnel diffraction integral, then for an object point on the optical axis we find the same Airy pattern as before and as shown in Figure 6.8.2, except that the variable $ar/(\lambda f)$ must be replaced by

$$\frac{ar}{\lambda s_i}$$

where s_i is the image position as given by the Lens Law. This field is called the Point Spread Function (PSF for short). Hence,

$$\mathrm{PSF}(x,y) = \frac{\pi a^2}{\lambda s_i} \frac{J_1\left(2\pi \frac{a}{\lambda s_i}\sqrt{x^2+y^2}\right)}{\frac{2\pi a}{\lambda s_i}\sqrt{x^2+y^2}}, \qquad \text{Airy pattern for imaging.}$$

For object points that are not on the optical axis, the PSF is translated such that it is centred on the ideal Gaussian image point.

Usually we assume that we know the field transmitted or reflected by an object in the socalled object plane, which is the plane immediately behind the object (on the side of the lens). The object plane is discretised by a set of points and the images of these points are assumed to be given by translated versions of the PSF:

$$\mathrm{PSF}(x - x_i, y - y_i)$$

where (x_i, y_i) are the transverse coordinates of the image point according to Gaussian geometrical optics.

The total image field is obtained by summing (integrating) over these PSFs, weighted by the field at the object points:

$$U_i(x, y, s_i) = \iint \mathrm{PSF}(x - Mx_o, x - My_o) U_o(x_o, y_o, s_o)\, \mathrm{d}x_o\, \mathrm{d}y_o.$$

where $x_o = x_i/M, y_o = y_i/M$ is the image point and M is the magnification. If the magnification is unity, the image field is a convolution between the PSF and the object field. If the magnification differs from unity, the integral can be made into a convolution by rescaling the coordinates in image space.

It is clear from 6.8.15 that larger radius a of the lens and smaller wavelength λ imply a narrower PSF. This in turn implies that the kernel in the convolution is more sharply peaked and hence that the resolution of the image is higher.

Remarks

1. If laser light is used to illuminate the object, the object field may in general be considered to be perfectly coherent. This implies that a detector in the image plane would measure the squared modulus of the complex field (6.8.16):

$$I_i(x, y, s_i) = \left| \iint \mathrm{PSF}(x - Mx_o, y - My_o) U_o(x_o, y_o, s_o)\, \mathrm{d}x_o\, \mathrm{d}y_o \right|^2.$$

In this case the system is called a **coherent imaging system**.

2. If the object is an spatially completely incoherent source, the images of the point sources of which the source (object) consists cannot interfere in the image plane. Therefore, in this case the intensity in the image plane is given by the incoherent sum:

$$I_i(x, y, s_i) = \iint |\mathrm{PSF}(x - Mx_o, y - Mx_o)|^2 I_o(x_o, y_o, s_o)\, \mathrm{d}x_o\, \mathrm{d}y_o,$$

where $I_o = |U_o|^2$ is the intensity of the object. Hence the image intensity is expressed in the intensity of the object by a convolution with the intensity of the PSF. This system is called a **incoherent imaging system**.

3. An object is often illuminated by a spatially incoherent light source and then imaged. The field reflected or transmitted by the object is then **partially coherent**. We have shown in Chapter 5 that the degree of mutual coherence in the object increases when the distance between the source and the object is increased. The detected intensity in the image plane can be computed by splitting up the spatially incoherent source into sufficiently many mutually incoherent point sources and considering the fields in the image plane due to the illumination by each individual point source. The total intensity in the image plane is then the sum of the individual intensities.

6.8.3 Spatial Light Modulators and Optical Fourier Filtering

- **SLM**. The field in the entrance pupil of a lens can be changed spatially by a so-called **spatial light modulator** (SLM). A SLM has thousands of pixels by which very general fields can be made. By applying three SLMs in series, one can tune the polarisation, the phase and the amplitude pixel by pixel and hence very desired pupil fields can be realised which after focusing, can give very special focused fields. An example is an electric field with only a longitudinal component (i.e. only a E_z-component) in the focal point.
- **Fourier filtering**. Suppose we have the setup as shown in Figure 6.8.3. With one lens we can create the Fourier transform of some field $U(x, y)$. If a mask is put in the focal plane and a second lens is used to refocus the light, the inverse Fourier transform of the field after the mask is obtained. This procedure is called Fourier filtering using lenses. Fourier filtering means that the amplitude and/or phase of the plane waves in the angular spectrum of the field are manipulated. An application of this idea is the phase contrast microscope.

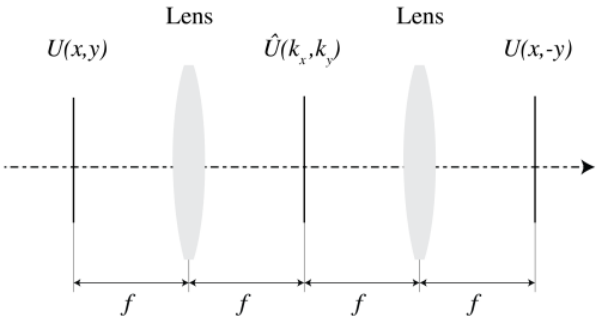


Figure 6.8.3: Set-up for Fourier filtering. The first lens creates a Fourier transform of $U(x, y)$, to which we can apply some operation (e.g. applying different phase shifts to different parts of the field). The second lens then applies another Fourier transform (which is the same as the inverse Fourier transform and a mirror transformation).

6.9: Super-resolution

We have emphasised that evanescent waves set the ultimate limit to resolution in optics. In Chapter 2 it was explained that, although within geometrical optics one can image a single point perfectly using conical surfaces, several points, let alone an extended object, cannot be imaged perfectly. It was furthermore explained that when only paraxial rays are considered, i.e. within Gaussian geometrical optics, perfect imaging of extended objects is possible. However, rays of which the angle with the optical axis is large cause aberrations. But even when perfect imaging would be possible in geometrical optics, a real image can never be perfect due to the fact that information contained in the amplitudes and phase of the evanescent waves cannot propagate. The resolution that can be obtained with an optical system consisting of lenses is less than follows from considering the loss of information due to evanescent waves because also propagating waves with spatial frequencies that are too large to be captured by the optical system cannot contribute to the image. Therefore the image of a point object has the size

$$\lambda/NA_i$$

where $NA_i = a/s_i$ is the numerical aperture in image space, i.e. it is the sinus of half the opening angle of the cone extended by the exit pupil at the Gaussian image point on the optical axis. This resolution limit is called the diffraction limit.

The size of the image of a point as given by the PSF in (6.7.15) is influenced by the magnification of the system. To characterise the resolution of a diffraction-limited system, it is therefore better to consider the numerical aperture on the object side: $NA_o = NA_i|M| = a/s_o$. The value of NA_o is the sinus of the half angle of the cone extended by the entrance pupil of the system on the object point on the optical axis. This is the cone of wave vectors emitted by this object point that can contribute to the image (they are "accepted" by the optical system). The larger the half angle of this cone, the higher the spatial frequencies that can contribute and hence information about finer details of the object is transmitted.

It should be clear by now that beating the diffraction limit is extremely difficult. Nevertheless, a lot of research in optics has been and still is directed towards realising this goal. Many attempts have been made, some successful, others not so, but, whether successful or not, most were based on very ingenious ideas. To close this chapter on diffraction theory, we will give examples of attempts to achieve what is called super-resolution.

- **Confocal microscopy**. A focused spot is used to scan the object and the reflected field is imaged onto a small detector ("point detector"). The resolution is roughly a factor 1.5 better than for normal imaging with full field of view using the same objective. The higher resolution is achieved thanks to the illumination by oblique plane waves that are present in the spatial (Fourier) spectrum of the illuminating spot. By illumination with plane waves under large angles of incidence, higher spatial frequencies of the object which are under normal incidence not accepted by the objective, are now "folded back" into the cone of plane waves accepted by the objective. The higher resolution comes at the prize of longer imaging time because of scanning. The confocal microscope was invented by Marvin Minsky in 1957 .
- **The Perfect lens based on negative refraction**. It can be shown that when a material has **negative permittivity** and **negative permeability**, the phase velocity of a plane wave is opposite to the energy velocity. Furthermore, when a slab of such material is surrounded by material with positive permittivity and positive permeability equal to the absolute values of the permittivity and permeability of the slab, the reflection of all waves is zero for **every angle of incidence** and **every state of polarisation**. Moreover, evanescent waves **gain** amplitude inside the slab and it turns out that there are two planes, one inside the slab and one on the other side of it, where a perfect image of a point in front of the slab occurs. Note that the increase of amplitude of an evanescent wave does not violate the conservation of energy, because an evanescent wave does not propagate energy in the direction in which it is evanescent. The simple slab geometry which acts as a perfect lens was proposed by John Pendry in 2000 and presented in Figure $6.9.1$. Unfortunately, a material with negative permittivity and negative permeability has not been found in nature, although there seems to be no fundamental reason why it could not exist. Therefore, many researchers have attempted to mimic such a material using conventional materials such as metals. There are, however, more fundamental reasons why Pendry's perfect lens will not work satisfactorily, even if the material would exist. We refer to the master course Advanced Photonics for more details.

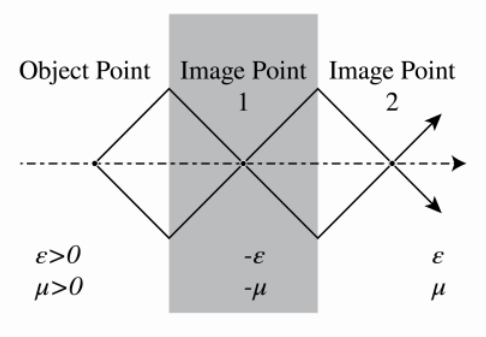


Figure 6.9.1: Pendry's perfect lens consists of a slab of a material with negative permittivity and negative permeability such that its absolute values are equal to the positive permittivity and positive permeability of the surrounding medium. Points outside the slab are imaged perfectly in two planes: one inside the slab and the other on the opposite side of the slab.

- **Hyperbolic materials**. Hyperbolic materials are anisotropic, i.e. the phase velocity of a plane wave depends on the polarisation and on the direction of the wave vector. The permittivity of an anisotropic material is a tensor (loosely speaking a (3,3)-matrix). Normally the eigenvalues of the permittivity matrix are positive; however, in a hyperbolic material two eigenvalues are of equal sign and the third has opposite sign. In such a medium all waves of the so-called extraordinary type of polarisation propagate, no matter how high the spatial frequencies are. Hence, for the extraordinary state of polarisation evanescent waves do not exist and therefore super-resolution and perfect imaging should be possible in such a medium. Natural hyperbolic media seem to exist for a few frequencies in the mid-infrared. For visible wavelengths, materials with hyperbolic behaviour are too lossy to give super-resolution. Therefore one tries to approximate hyperbolic media by so-called metamaterials which are made of very thin metallic and dielectric layers, so that the effective permittivity has the desired hyperbolic property. The success of this idea has, however, been moderate so far.
- **Nonlinear effects**. When the refractive index of a material depends on the local electric field, the material is nonlinear. At optical frequencies nonlinear effects are in general very small, but with a strong laser they can become significant. One effect is self-focusing, where the refractive index is proportional to the local light intensity. The locally higher intensity causes an increase of the refractive index, leading to a waveguiding effect due to which the beam focuses even more strongly. Hence the focused beam becomes more and more narrow while propagating, until finally the material breaks down.
- **Stimulated Emission Depletion Microscopy (STED)**. This technique was invented by V. A. Okhonin in 1986 in the USSR and was further developed by Stefan Hell and his co-workers in the nineties. Hell received the Nobel Prize in chemistry for his work in 2014. STED is a non-linear technique with which super-resolution in fluorescence microscopy can be achieved. Images made with a fluorescence microscope are blurred when the fluorescent molecules are very close together. In the STED microscope a special trick is used to ensure that fluorescing molecules are sufficiently distant from each other, so that they can be detected individually. To achieve this two focused spots are used: the first spot excites the molecules to a higher level. The second spot is slightly red-shifted and has a doughnut shape (see Figure 6.9.3). It causes decay of the excited molecules to the lower level by stimulated emission (the excited state is depleted) Because of the doughnut shape of the second spot, the molecule in the centre of the spot is not affected and will still fluorescence. Crucial is that a doughnut spot has a central dark region which is very narrow i.e. it can be much smaller than the Airy spot and this is the reason for the super-resolution.

External sources in recommended order

1. Every picture is made of waves - Sixty Symbols, 3: 33 to 7:15: Basic explanation of Fourier transforms.
2. Heisenberg's Microscope - Sixty Symbols, 0:20 to 2:38: Basic explanation of the uncertainty principle (though in the context of quantum physics).
3. E. Hecht, Optics, §7.4.4, subsection 'Fourier Analysis and Diffraction'.
4. J. Goodman, Introduction to Fourier Optics, §5.2.2: Several calculations on the Fourier transforming properties of lenses.
5. E. Hecht, Optics, §10.2.6, subsection 'Resolution of imaging systems'.

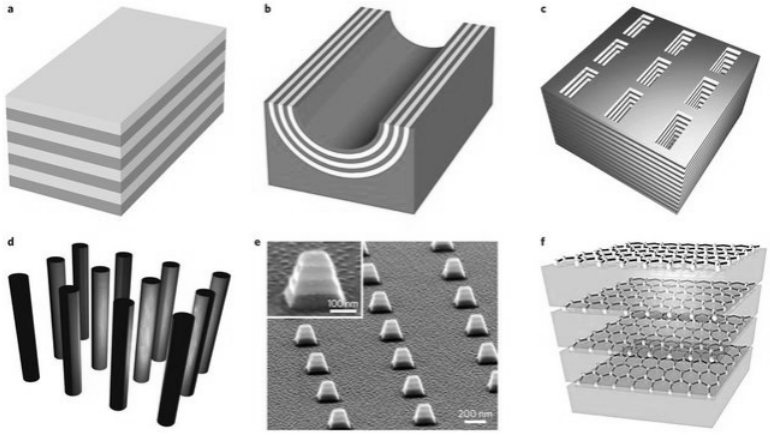


Figure 6.9.2: Examples of composite materials consisting of thin (sub-wavelength) layers of metals and dielectrics. These artificial materials are called metamaterials. (A. Poddubny, I. Iorsh, P. Belov, & Y. Kivshar, Hyperbolic metamaterials, Nat. Photon., $7(12), 948 - 957(2013)$)

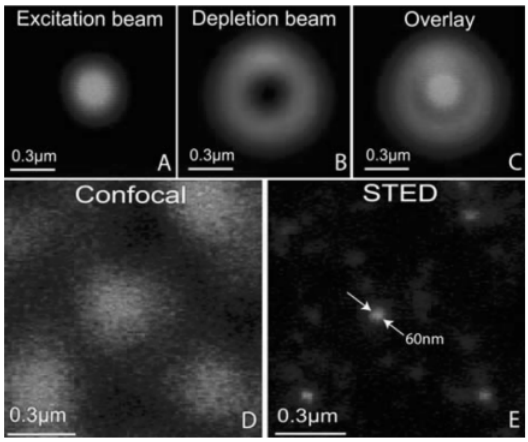


Figure 6.9.3: Spot used for excitation (top left) and for depletion (top middle). Fluorescence signal top right. In the lower figure the confocal image is compared to the STED image. (P.F. Rodriguez and al., Building a fast scanning stiumlated emission depletion microscope, Materials Science (2012))

CHAPTER OVERVIEW

7: Lasers

7.0: What you should know and be able to do after studying this chapter
7.1: Unique Properties of Lasers
7.2: Optical Resonator
7.3: Amplification
7.4: Cavities
7.5: Problems with Laser Operation
7.6: Types of Lasers

7.0: What you should know and be able to do after studying this chapter

- Know the special properties of laser sources.
- Understand the optical resonator and why it is needed.
- Understand the role of the amplifier and explain what the gain curve is.
- Explain the principle of population inversion and how it can be achieved.
- Explain how single frequency operation can be obtained.
- Understand what transverse modes are and how they can be prevented.

In the early 1950s a new source of microwave radiation, **the maser**, was invented by C.H. Townes in the USA and A.M. Prokhorov and N.G. Basov in the USSR. Maser stands for "Microwave Amplification by Stimulated Emission of Radiation". In 1958, A.L. Schawlow and Townes formulated the physical constraint to realise a similar device for visible light. This resulted in 1960 in the first optical maser by T.H. Maiman in the USA. This device was since then called **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation or laser. It has revolutionised science and engineering and has many applications, e.g.

- bar code readers,
- compact discs,
- computer printers,
- fiberoptic communication,
- sensors,
- material processing,
- non-destructive testing,
- position and motion control,
- spectroscopy,
- medical applications, such as treatment of retina detachment,
- nuclear fusion,
- holography.

7.1: Unique Properties of Lasers

The broad applications of lasers are made possible by their unique properties which distinguishes them from all other light sources. We discuss these unique properties below.

7.1.1 High Monochromaticity; Narrow Spectral Width; High Temporal Coherence

These three prperties basically are the same. Saying that the laser has high monochromaticity or that it has a very narrow spectral width means that it emits a very narrow band of frequencies, A spectral lamp, like a gas discharge lamp based on Mercury vapor, can have a spectral width $\Delta\nu = 10\text{GHz}$. Visible frequencies are around 2×10^{14} Hz , hence the spectral width of the lamp is roughly 0.02%. The line width measured in wavelengths satisfies

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta\nu}{\nu},$$

and hence for $\lambda = 550$ nm, $\Delta\lambda$ of a spectral lamp is of the order of 0.1 nm.

By contrast, a laser can easily have a frequency band that is a factor of 100 smaller, i.e. less than $10\text{MHz} = 10^7$ Hz in the visible. For a wavelength of 550 nm this means that the linewidth is only 0.001 nm. As has been explained in Chapter 7 , the coherence time τ_c of the emitted light is the reciprocal of the frequency bandwidth:

$$\tau_c = 1/\Delta\nu.$$

Light is emitted by atoms in bursts of harmonic (cosine) waves consisting of a great but finite number of periods. As will be explained in this chapter, due to the special configuration of the laser, the wave trains in laser light can be extremely long, corresponding to a very long coherence time.

7.1.2 Highly Collimated Beam

Consider a discharge lamp as shown in Figure 7.1.1. To collimate the light, the discharge lamp can be positioned in the focal plane of a lens. The spherical waves emitted by all point sources (atoms) in the lamp are collimated into plane waves whose direction depends on the position of the atoms in the source. The atoms at the edges of the source determine the overall divergence angle θ, which is given by

$$\theta = h/f,$$

where $2h$ is the size of the source and f is the focal length of the lens as shown in Figure 7.1.1. Hence the light can be collimated by either choosing a lens with large focal length or by reducing the size of the source, or both. Both methods lead, however, to weak intensities. Due to the special configuration of the laser source, which consists of a Fabry-Perot resonator in which the light bounces up and down many times before being emitted, the atomic sources are effectively all at very large distance and hence the effective size of the source is very small. The divergence of the laser beam is therefore not limited by the size of the source but by the size of its emitting surface through the inevitable effect of diffraction. As follows from Chapter 6, a parallel beam of diameter D and wavelength λ has a diffraction limited divergence given by:

$$\theta = \frac{\lambda}{D}.$$

The diffraction-limited divergence thus depends on the wavelength and decreases when the diameter of the emitting surface increases. With a laser source, the diffraction-limited convergence angle can almost be reached and therefore a collimated beam with very high intensity can be realised (Figure 7.1.2).

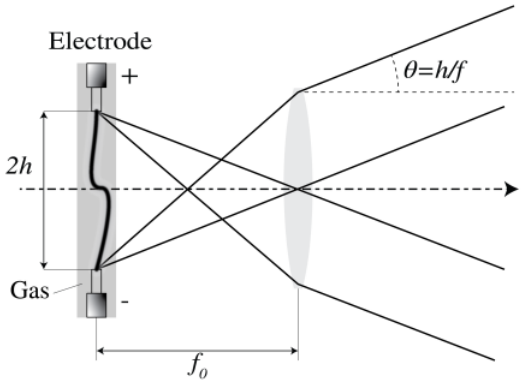


Figure 7.1.1: A discharge lamp in the focal plane of a converging lens. Every atom in the lamp emits a spherical wave during a burst of radiation, lasting on average a coherence time τ_c. The overall divergence of the beam is determined by the atoms at the extreme positions of the source.

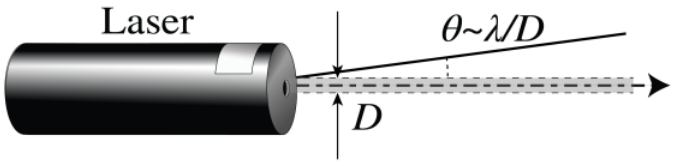


Figure 7.1.2: A laser beam can almost reach diffraction-limited collimation.

7.1.3 Diffraction-Limited Focused Spot, High Spatial Coherence

If we add a second lens after the first lens in Figure 7.1.3 a spot is obtained in the focal plane of the second lens. This spot can be very small only when the light has been almost perfectly collimated by the first lens.

What is the smallest focal spot that one can achieve? If one focuses a perfectly collimated beam with a lens with very small aberrations and with numerical aperture NA, the lateral size of the focused spot is, according to Chapter 6 , diffraction-limited and given by

$$\text{diffraction-limited spot size } = 0.6\frac{f}{D}\lambda = 0.6\frac{\lambda}{NA}.$$

With a laser one can achieve a diffraction-limited spot with a very high intensity.

As has been explained in Chapter 5, a light wave has **high spatial coherence** if at any given time, its amplitude and phase in transverse points can be predicted. The spherical waves emitted by a point source have this property. But when there are many point sources (atoms) that each emit bursts of harmonic waves that start at random times, as is the case in a classical light source, the amplitude and phase of the total emitted field at any position in space cannot be predicted. The only way to make the light spatially coherent is by making the light source very small, but then there is hardly any light. As will be explained below, by the design of the laser, the emissions by the atoms of the amplifying medium in a laser are phase-correlated, which leads to a very high temporal and spatial coherence. The property of a small spot size with high intensity is essential for many applications, such as high resolution imaging, material processing with cutting, welding and drilling spots with very high power and in retina surgery, where a very small, high-intensity spot is applied to weld the retina without damaging the surrounding healthy tissue.

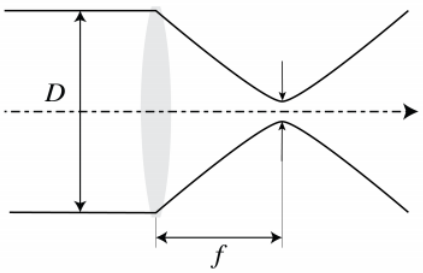


Figure 7.1.3: Diffraction-limited spot obtained by focusing a collimated beam.

7.1.4 High Power

There are two types of lasers namely continuous wave (CW) lasers, which produce a continuous output, and pulsed lasers which emit a train of pulses. These pulses can be very short: from a nanosecond to even femtoseconds $\left(10^{-15}\ \mathrm{s}\right)$. A relatively low-power CW laser is the HeNe laser which emits roughly $1\ \mathrm{mW}$ at the wavelength $632\ \mathrm{nm}$. Other lasers can emit up to a megawatt of continuous power. Pulsed lasers can emit enormous peak intensities (i.e. at the maximum of a pulse), ranging from 10^{9} to 10^{15} Watt.

There are many applications of high-power lasers such as for cutting and welding materials. To obtain EUV light with sufficient high intensity for use in photolithography for manufacturing ICs, extremely powerful $\mathrm{CO_2}$ lasers are used to excite a plasma. Extremely high-power lasers are also applied to initiate fusion and in many nonlinear optics applications. Lasers with very short pulses are also used to study very fast phenomena with short decay times, and to realise faster clocks.

7.1.5 Wide Tuning Range

For a wide range of wavelengths, from the vacuum ultra-violet (VUV), the ultra-violet (UV), the visible, the infrared (IR), the mid-infrared (MIR) up into the far infrared (FIR), lasers are available. For some type of lasers, the tuning range can be quite broad. The gaps in the electromagnetic spectrum that are not directly addressed by laser emission can be covered by techniques such as higher harmonic generation and frequency differencing.

7.2: Optical Resonator

We now explain the working of lasers. A laser consists of

1. an optical resonator;
2. an amplifying medium.

In this section we consider the resonator. Its function is to obtain a high light energy density and to gain control over the emission wavelengths.

A resonator, whether it is mechanical like a pendulum, a spring or a string, or electrical like an LRC circuit, has one or multiple resonance frequencies $\nu_{\text{res}.}$. Every resonator has losses due to which the oscillation gradually dies out when no energy is supplied. The losses cause an exponential decrease of the amplitude of the oscillation, as shown in Figure 7.2.1. The oscillation is therefore not purely monochromatic but has a finite bandwidth of order $\Delta\nu \approx 1/\tau$ as shown in Figure 7.2.1, where τ is the time at which the amplitude of the oscillation has reduced to half the initial value.

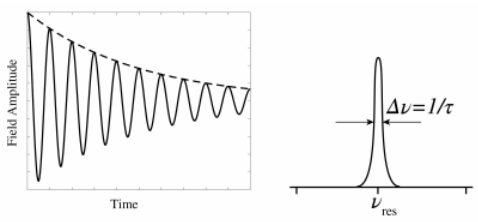


Figure 7.2.1: Damped oscillation (left) and frequency spectrum of a damped oscillation (right) with resonance wavelength and frequency width equal to the reciprocal of the decay time.

The optical resonator is a region filled with some material with refractive index n bounded by two aligned, highly reflective mirrors at a distance L. The resonator is called a **Fabry-Perot cavity**. Let the z-axis be chosen along the axis of the cavity as shown in Fig.7.2.2, and assume that the transverse directions are so large that the light can be considered a plane wave bouncing back and forth along the z-axis between the two mirrors. Let ω be the frequency and $k_0 = \omega/c$ the wave number in vacuum. The plane wave that propagates in the positive z-direction is given by:

$$E(z) = Ae^{ik_0 nz}.$$

For very good mirrors, the amplitude remains unchanged upon reflections, while the phase typically changes by π. Hence, after one round trip (i.e. two reflections) the field (7.2.1) is (the possible phase changes at the mirrors add up to 2π and hence have no effect):

$$E(z) = Ae^{2ik_0 nL}e^{ik_0 nz}.$$

A high field builds up when this wave constructively interferes with (7.2.1), i.e. when

$$k_0 = \frac{2\pi m}{2nL}, \quad \text{or} \quad \nu = \frac{kc}{2\pi} = m\frac{c}{2nL},$$

for $m = 1, 2, \ldots$. Hence, provided dispersion of the medium can be neglected (n is independent of the frequency), the resonance frequencies are separated by

$$\Delta\nu_f = c/(2nL),$$

which is the so-called **free spectral range**. For a gas laser that is 1 m long, the free spectral range is approximately 150MHz.

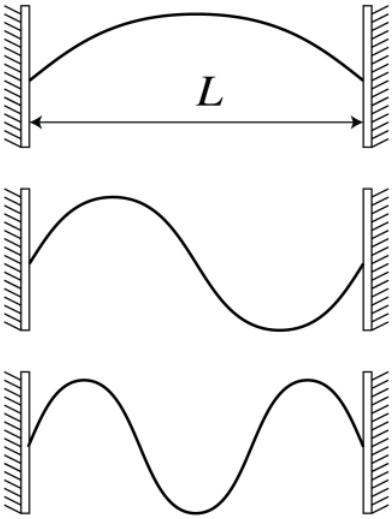


Figure 7.2.2: Fabry-Perot resonances.

Example 7.2.1

Suppose that the cavity is 100 cm long and is filled with a material with refractive index $n = 1$. Light with visible wavelength of $\lambda = 500$ nm corresponds to mode number $m = 2L/\lambda = 4 \times 10^6$ and the free spectral range is $\Delta\nu_f = c/(2L) = 150\text{MHz}$.

The multiple reflections of the laser light inside the resonator make the optical path length very large. For an observer, the atomic sources seem to be at a very large distance and the light that is exiting the cavity resembles a plane wave. The divergence of the beam is therefore not limited by the size of the source, but by diffraction due to the aperture of the exit mirror.

Because of losses caused by the mirrors (which never reflect perfectly) and by the absorption and scattering of the light, the resonances have a certain frequency width $\Delta\nu$. When a resonator is used as a laser, one of the mirrors is given a small transmission to couple the laser light out. This also contributes to the loss of the resonator. To compensate for all losses, the cavity must contain an amplifying medium. Due to the amplification, the resonance line widths inside the bandwidth of the amplifier are reduced to very sharp lines as shown in Figure 7.2.3.

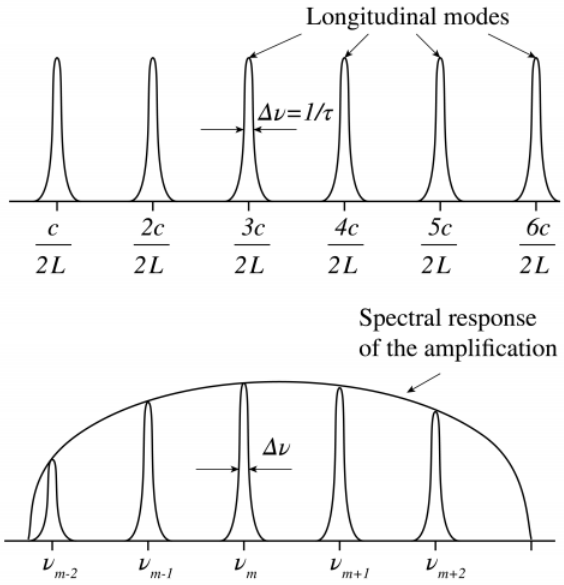


Figure 7.2.3: Resonant frequencies of a cavity of length L when the refractive index $n = 1$. With an amplifier inside the cavity, the line widths of the resonances within the bandwidth of the amplifier are reduced. The envelope is the spectral function of the amplification.

7.3: Amplification

Amplification can be achieved by a medium with atomic resonances that are at or close to one of the resonances of the resonator. We first recall the simple theory developed by Einstein in 1916 of the dynamic equilibrium of a material in the presence of electromagnetic radiation.

7.3.1 The Einstein Coefficients

We consider two atomic energy levels $E_2 > E_1$. By absorbing a photon of energy

$$\hbar\omega = E_2 - E_1,$$

an atom that is initially in the lower energy state 1 can be excited to state 2 . Here $\hbar$ is Planck's constant:

$$\hbar = \frac{6.626070040}{2\pi} \times 10^{-34} \quad \text{Js}\,.$$

Suppose $W(\omega)$ is the time-averaged electromagnetic energy density per unit of frequency interval around frequency ω. Hence W has dimension Jsm ^{3}. Let N_1 and N_2 be the number of atoms in state 1 and 2, respectively, where

$$N_1 + N_2 = N,$$

is the total number of atoms (which is constant). The rate of absorption is the rate of decrease of N_1 and is proportional to the energy density and the number of atoms in state 1:

$$\frac{dN_1}{dt} = -B_{12}N_1W(\omega), \quad \text{absorption},$$

where the constant $B_{12} > 0$ has dimension $\mathrm{m}^3\ \mathrm{J}^{-1}\ \mathrm{s}^{-2}$. Without any external influence, an atom that is in the excited state will usually transfer to state 1 within 1 ns or so, while emitting a photon of energy (7.10). This process is called spontaneous emission, since it happens also without an electromagnetic field present. The rate of spontaneous emission is given by:

$$\frac{dN_2}{dt} = -A_{21}N_2, \quad \text{spontaneous emission},$$

where A_{21} has dimension s^{-1}. The lifetime of spontaneous transmission is $\tau_{sp} = 1/A_{21}$. It is important to note that the spontaneously emitted photon is emitted in a random direction. Furthermore, since the radiation occurs at a random time, there is no phase relation between the spontaneously emitted field and the field that excites the atom.

It is less obvious that in the presence of an electromagnetic field of frequency close to the atomic resonance, an atom in the excited state can also be **stimulated** by that field to emit a photon and transfer to the lower energy state. The rate of **stimulated emission** is proportional to the number of excited atoms and to the energy density of the field:

$$\frac{dN_2}{dt} = -B_{21}N_2W(\omega), \quad \text{stimulated emission},$$

where B_{21} has the same dimension as B_{12}. It is very important to remark that stimulated emission occurs in the **same electromagnetic mode** (e.g. a plane wave) as the mode of the field that excites the transmission and that the phase of the radiated field is **identical** to that of the exciting field. This implies that stimulated emission enhances the electromagnetic field by constructive interference. This property is crucial for the operation of the laser.

7.3.2 Relation Between the Einstein Coefficients

Relations exist between the Einstein coefficients A_{21}, B_{12} and B_{21}. Consider a black body, such as a closed empty box. After a certain time, thermal equilibrium will be reached. Because there is no radiation entering the box from the outside nor leaving the box to the outside, the electromagnetic energy density is the thermal density $W_T(\omega)$, which, according to Planck's Law, is independent of the material of which the box is made and is given by:

$$W_T(\omega) = \frac{\hbar\omega^3}{\pi^2c^3}\frac{1}{\exp\left(\frac{\hbar\omega}{k_BT}\right) - 1},$$

where k_B is Boltzmann's constant:

$$k_B = 1.38064852 \times 10^{-23}\ \mathrm{m^2 kg s^{-2}\ K^{-1}}.$$

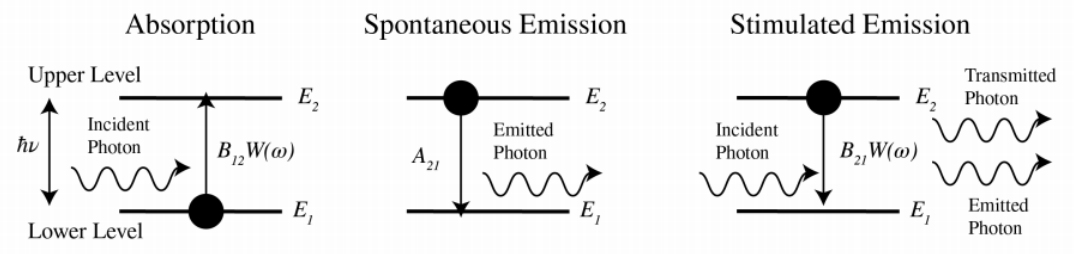


Figure 7.3.1: Absorption, spontaneous emission and stimulated emission.

The rate of upward and downwards transition of the atoms in the wall of the box must be identical:

$$B_{12}N_1W_T(\omega) = A_{21}N_2 + B_{21}N_2W_T(\omega).$$

Hence,

$$W_T(\omega) = \frac{A_{21}}{B_{12}N_1/N_2 - B_{21}}.$$

But in thermal equilibrium:

$$\frac{N_2}{N_1} = \exp\left(-\frac{E_2 - E_1}{k_BT}\right) = \exp\left(-\frac{\hbar\omega}{k_BT}\right).$$

By substituting (7.3.11) into (7.3.10), and comparing the result with (7.3.7), it follows that both expressions for $W_T(\omega)$ are identical for all temperatures only if

$$B_{12} = B_{21}, \quad A_{21} = \frac{\hbar\omega^3}{\pi^2c^3}B_{21}.$$

Example For green light of $\lambda = 550\ \mathrm{nm}$, we have $\omega/c = 2\pi/\lambda = 2.8560 \times 10^6\ \mathrm{m^{-1}}$ and thus

$$\frac{A_{21}}{B_{21}} = 1.5640 \times 10^{-15}\ \mathrm{J\ s\ m^{-3}}.$$

Hence the spontaneous and stimulated emission rates are equal if $W(\omega) = 1.5640 \times 10^{-15}\mathrm{Jsm^{-3}}$

For a (narrow) frequency band $\mathrm{d}\omega$ the time-averaged energy density is $W(\omega)\mathrm{d}\omega$ and for a plane wave the energy density is related to the intensity I (i.e. the length of the time-averaged Poynting vector) as:

$$W(\omega)\mathrm{d}\omega = I/c.$$

Table 7.3.1: Typical intensities of light sources.

	$I\ (\mathrm{W\ m^{-2}})$
Mercury lamp	10^4
Continuous laser	10^5
Pulsed laser	10^{13}

A typical value for the frequency width of a narrow emission line of an ordinary light source is: $10^{10}\ \mathrm{Hz}$, i.e. $\mathrm{d}\omega = 2\pi \times 10^{10}\ \mathrm{Hz}$. Hence, the spontaneous and stimulated emission rates are identical if the intensity is $I = 2.95 \times 10^4\ \mathrm{W/m^2}$. As seen from Table 7.1, only for laser light stimulated emission is larger than spontaneous emission. For classical light sources the spontaneous emission rate is much larger than the stimulated emission rate. If a beam with frequency width $\mathrm{d}\omega$ and energy density $W(\omega)\ \mathrm{d}\ \omega$ propagates through a material, the rate of loss of energy is proportional to:

$$(N_1 - N_2)\, B_{12}W(\omega).$$

According to (7.3.9) this is equal to the spontaneous emission rate. Indeed, the spontaneously emitted light corresponds to a loss of intensity of the beam, because it is emitted in random directions and with random phase.

When $N_2 > N_1$, the light is **amplified**. This state is called **population inversion** and it is essential for the operation of the laser. Note that the ratio of the spontaneous and stimulated emission rates is, according to (7.3.12), proportional to ω^3. Hence for shorter wavelengths such as x-rays, it is much more difficult to make lasers than for visible light.

7.3.3 Population Inversion

For electromagnetic energy density $W(\omega)$ per unit of frequency interval, the rate equations are

$$\begin{aligned} \frac{dN_2}{dt} &= -A_{21}N_2 + (N_1 - N_2)\, B_{12}W(\omega) \\ \frac{dN_1}{dt} &= A_{21}N_2 - (N_1 - N_2)\, B_{12}W(\omega) \end{aligned}$$

Hence, for $\Delta N = N_2 - N_1$:

$$\frac{d\Delta N}{dt} = -A_{21}\Delta N - 2\Delta N B_{12}W(\omega) - A_{21}N$$

where as before: $N = N_1 + N_2$ is constant. If initially (i.e. at $t = 0$) all atoms are in the lowest state: $\Delta N(t = 0) = -N$, then it follows from 7.3.12

$$\Delta N(t) = -N\left[\frac{A_{21}}{A_{21} + 2B_{12}W(\omega)} + \left(1 - \frac{A_{21}}{A_{21} + 2B_{12}W(\omega)}\right) e^{-(A_{21}+2B_{12}W(\omega))t}\right]$$

An example where $A_{21}/B_{12}W(\omega) = 0.5$ is shown in Figure 7.3.2. We always have $\Delta N < 0$, hence $N_2(t) < N_1(r)$ for all times t. Therefore, a system with only two levels cannot have population inversion.

A way to achieve population inversion of levels 1 and 2 and hence amplification of the radiation with frequency ω with $\hbar\omega = E_2 - E_1$ is to use more atomic levels, for example three. In Figure 7.3.3 the ground state is state 1 with two upper levels 2 and 3 such that $E_1 < E_2 < E_3$. The transition of interest is still that from level 2 to level 1 . Initially almost all atoms are in the ground state 1 . Then atoms are pumped with rate R from level 1 directly to level 3 . The transition $3 \to 2$ is non-radiative and has high rate A_{32} so that level 3 is quickly emptied and therefore N_3 remains small. State 2 is called a metastable state, because the residence time in the metastable state is for every atom relatively long. Therefore its population tends to increase, leading to population inversion between the metastable state 2 and the lower ground state 1 (which is continuously being depopulated by pumping to the highest level).

Note that A_{31} has to be small, because otherwise level 1 will quickly be filled, by which population inversion will be stopped. This effect can be utilised to obtain a series of laser pulses as output, but is undesirable for a continuous output power.

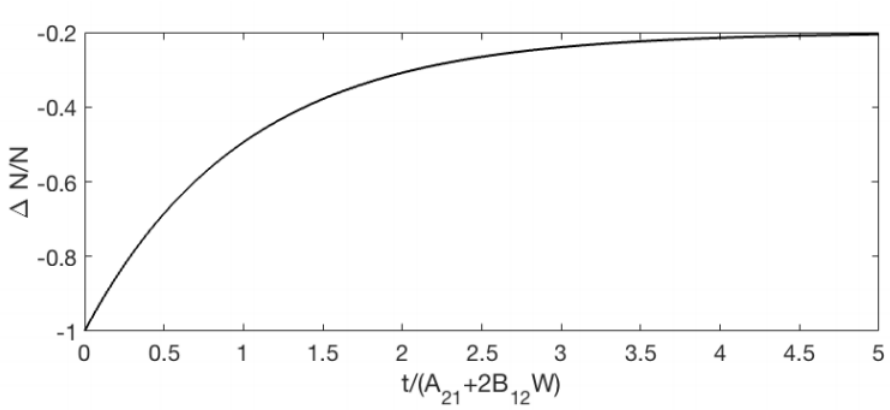


Figure 7.3.2: $\Delta N/N$ as function of $t/\left(A_{21} + 2B_{12}W\right)$ when all atoms are in the ground state at $t = 0$, i.e. $\Delta N(0) = -N$.

Pumping may be done optically as described, but the energy to transfer atoms from level 1 to level 3 can also be supplied by an electrical discharge in a gas or by an electric current. After the pumping has achieved population inversion, initially no light is emitted. So how does the laser actually start? Lasing starts by spontaneous emission. The spontaneously emitted photons stimulate

emission of the atoms in level 2 to decay to level 1 , while emitting a photon of energy $\hbar\omega$. This **stimulated emission occurs in phase with the exciting light** and hence the light continuously builds up coherently, while it is bouncing back and forth between the mirrors of the resonator. One of the mirrors is slightly transparent and in this way some of the light is leaking out of the laser.

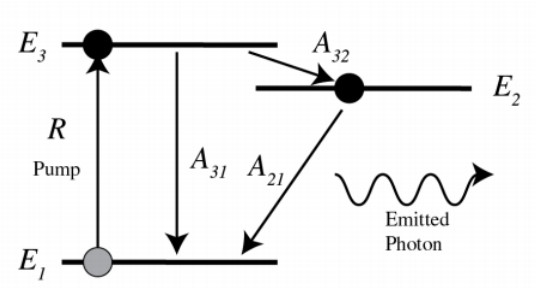


Figure 7.3.3: The three Einstein transitions and the pump.

7.4: Cavities

The amplifying medium can completely fill the space between the mirrors as the top of Figure 7.4.1, or there can be space between the amplifier and the mirrors. For example, if the amplifier is a gas, it may be enclosed by a glass cylinder. The end faces of the cylinder are positioned under the Brewster angle with respect to the axis, as shown in the middle figure of Figure 7.4.1, to minimise reflections. This type of resonator is called a **resonator with external mirrors**.

Usually one or both mirrors are convex, as shown in the bottom figure of Figure 7.4.1. We state without proof that in that case the distance L between the mirrors and the radii of curvature R_1 and R_2 of the mirrors has to satisfy

$$0 < \left(1 - \frac{L}{R_1}\right)\left(1 - \frac{L}{R_2}\right) < 1,$$

or else the laser light will ultimately leave the cavity laterally, i.e. it will escape sideways. This condition is called the **stability condition.** The curvature of a convex mirror is positive and that of a concave mirror is negative. Clearly, when both mirrors are concave, the laser is always unstable.

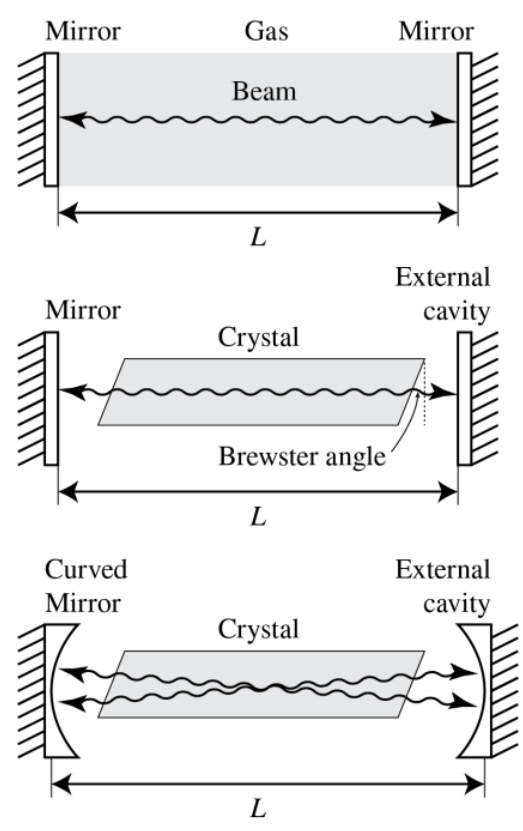


Figure 7.4.1: Three types of laser cavity. The shaded region is the amplifier. The middle case is called a laser with external cavities.

7.5: Problems with Laser Operation

In this section we consider some problems that occur with lasers and discuss what can be done to solve them.

1. **Multiple Resonance Frequencies**

In many applications such as laser communication and interferometry one needs a single wavelength. Consider a cavity of length L as shown in Figure $7.5.1$ and suppose that the amplifier has a gain curve covering many resonances of the resonator. One way to achieve single-frequency output is by taking care that there is only one frequency for which the gain is larger than the losses. One then says that the laser is above threshold for only one frequency. This can be done by choosing the length L of the cavity to be so small that there is only one mode under the gain curve for which the gain is higher than the losses. However, a small length of the amplifier means less output power and a less collimated output beam. Another method would be to reduce the pumping so that for only one mode the gain compensates the losses. But this implies again that the laser output power is relatively small. A better solution is to add a Fabry-Perot cavity inside the laser cavity as shown in Figure $7.5.2$. The cavity consists e.g. of a piece of glass of a certain thickness a.

By choosing a sufficiently small, the distance in frequency $c/(2a)$ between the resonances of the Fabry-Perot cavity becomes so large that there is only one Fabry-Perot resonance under the gain curve of the amplifier. Furthermore, by choosing the proper angle for the Fabry-Perot cavity with respect to the axis of the laser cavity, the Fabry-Perot resonance can be coupled to the desired resonance frequency. This frequency is then the frequency of the laser output. All other resonance frequencies of the resonator under the gain curve are damped, because they are not a resonance of the Fabry-Perot cavity.

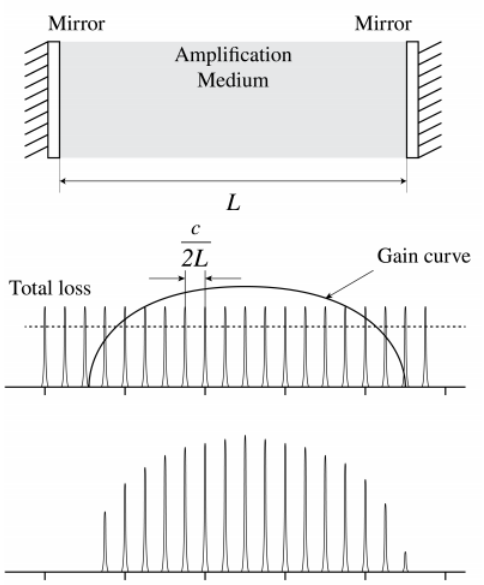


Figure $7.5.1$: Laser with cavity of length L and broad amplifier gain curve. Many resonance frequencies of the resonances are above threshold to compensate the losses.

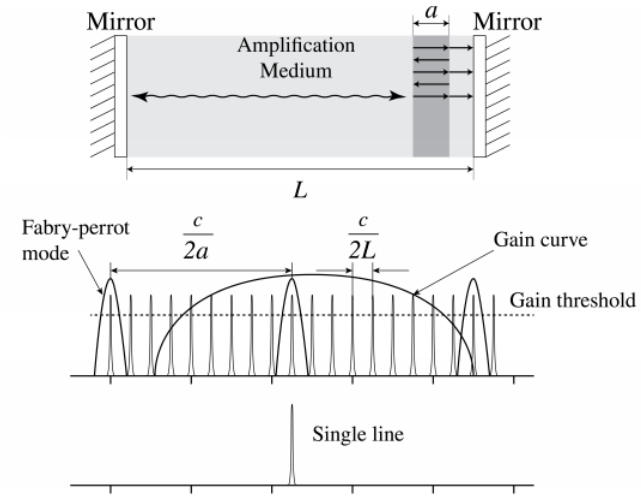


Figure 7.5.2: Laser with cavity of length L, a broad amplifier gain curve and an added Fabry-Perot cavity. The FB resonances acts as an extra filter to select only one mode of the laser.

2. **Multiple Transverse Modes**

The best-known laser mode has transverse intensity distribution, which is a Gaussian function of transverse distance to the optical axis. We call a mode with Gaussian transverse shape a **longitudinal mode** and when its frequency satisfies $\nu = mc/(2L)$, it is called the m th longitudinal mode. However, inside the laser cavity other modes with different transverse patterns can also resonate. An example is shown in Figure 7.5.3 where mode $(1,0)$ consists of two maxima. There exist many more transverse modes, as shown in Figure 7.5.4.

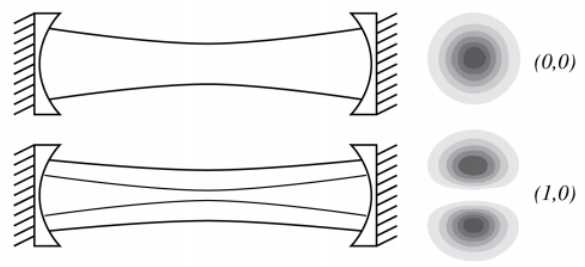


Figure 7.5.3: Laser cavity with $(0,0)$ and $(1,0)$ modes.

The transverse modes all have slightly different frequencies. So even when there is only one Gaussian mode above threshold (i.e. modes occur for only one value of m), there can be many transverse modes with frequencies very close to the frequency of the Gaussian mode, which are also above threshold. This is illustrated in Figure 7.5.5 where the frequencies of modes $(0,0)$, $(1,0)$ and $(1,1)$ all are above threshold. Usually one prefers the Gaussian mode and the transverse modes are undesired. How can one get rid of them? Because the Gaussian mode has smallest transverse width, the transverse modes can be eliminated by inserting an aperture in the laser cavity. This aperture is so small that the transverse modes suffer high-scattering losses, but is sufficiently large that the Gaussian mode is not affected.

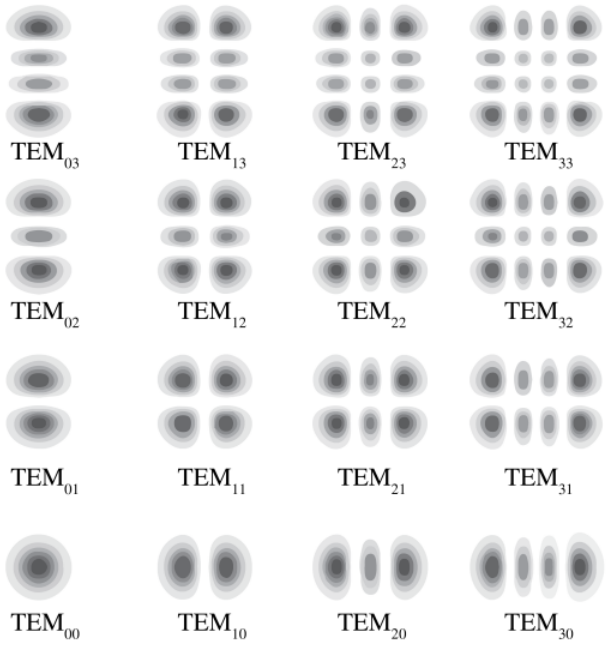


Figure 7.5.4: Intensity pattern of several transverse modes.

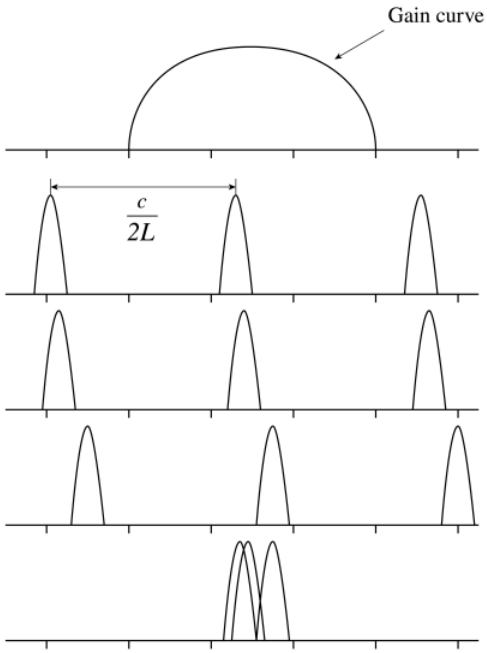


Figure 7.5.5: Resonance frequencies of transverse modes that have sufficient gain to compensate the losses.

7.6: Types of Lasers

There are many types of lasers: gas, solid, liquid, semiconductor, chemical, excimer, e-beam, free electron, fiber and even waveguide lasers. We classify them according to the pumping mechanism.

7.6.1 Optical Pumping

The energy to transfer the atom A from the ground state to the excited state is provided by light. The source could be another laser or an incoherent light source, such as a discharge lamp. If A is the atom in the ground state and A^* is the excited atom, we have

$$\hbar\omega_{02} + A \rightarrow A^*$$

where ω_{02} is the frequency for the transition $0 \rightarrow 2$ as seen in Figure 7.6.7. The Ruby laser, of which the amplifying medium consists of Al_2O_3 with 0.05 weight percent Cr_2O_3, was the first laser, invented in 1960. It emits pulses of light of wavelength $694.3\,\text{nm}$ and is optically pumped with a gas discharge lamp. Other optically pumped lasers are the YAG, glass, fiber, semiconductor and dye laser. In the dye laser the amplifier is a liquid (e.g. Rhodamine6G). It is optically pumped by an argon laser and has a huge gain width, which covers almost the complete visible wavelength range. We can select a certain wavelength by inserting a dispersive element like the Fabry-Perot cavity inside the laser cavity and rotating it at the right angle to select the desired wavelength, as explained above.

7.6.2 Electron-Collision Pump

Energetic electrons are used to collide with the atoms of the amplifier, thereby transferring some of their energy:

$$A + e\left(\mathcal{E}_1\right) \rightarrow A^* + e\left(\mathcal{E}_2\right),$$

where $e\left(\mathcal{E}_1\right)$ means an electron with energy $\mathcal{E}_1$ and where $\mathcal{E}_1 - \mathcal{E}_2$ is equal to $\hbar\omega_{02}$ so that the atom is transferred from the ground state to state 2 to obtain population inversion. Examples are the HeNe, Argon, Krypton, Xenon, Nitrogen and Copper lasers. Electrons can be created by a discharge or by an electron beam.

7.6.3 Atomic Collision

Let B^m be atom B in an excited, so-called metastable state. This means that B^m, although unstable, has a very long relaxation time, i.e. longer than $1\,\text{ms}$ or so. If B^m collides with atom A, it transfers energy to A.

$$B^m + A \rightarrow B + A^*,$$

A^* is the excited state used for the stimulated emission. If τ_{m1} is the relaxation time of metastable state B^m, then τ_{m1} is very large and hence the spontaneous emission rate is very small. This implies that the number of metastable atoms as function of time t is given by a slowly decaying exponential function $\exp(-t/\tau_{m1})$. How can one get metastable atoms? One can for example pump atom B from its ground state 1 to an excited state 3 above state m, such that the spontaneous emission rate $3 \rightarrow m$ is large. The pumping can be done electrically or by any other means. If it is done electrically, then we have

$$B + e\left(\mathcal{E}_2\right) \rightarrow B^m + e\left(\mathcal{E}_1\right),$$

Examples of these types of laser are He-Ne, which emits in the red at $632\,\text{nm}$, N_2 - CO_2 and He-Cd. All of these depend on atom or molecule collisions, where the atom or molecule that is mentioned first in the name is brought into the metastable state and lasing occurs at a wavelength corresponding to a level difference of the second mentioned atom or molecule. In the simplest case the metastable states are created by electrons generated by a discharge. The CO_2 laser emits at $10\mu\text{m}$ and can achieve huge power.

7.6.4 Chemical Pump

In some chemical reactions, a molecule is created in an excited state with population inversion. An example is:

$$A + B_2 \rightarrow (AB)^* + B$$

So in this case the lasing will take place for a transfer between states of molecule AB. The HF, DF, Ar-F, Cr-F, Xe-F and Xe-Cl lasers are all chemically pumped.

7.6.5 Semiconductor Laser

In this case pumping is done by electron current injection. It is one of the most compact lasers and yet it typically emits 20 mW of power. Transitions occur between the conduction and valence bands close to the $p-n$ junction. Electrons from the n-layer conduction band will recombine with the holes in the p-layer. A cavity is obtained by polishing the end faces that are perpendicular to the junction to make them highly reflecting. Semiconductor lasers are produced for wavelengths from 700 nm to 30μm and give continuous (CW) output.

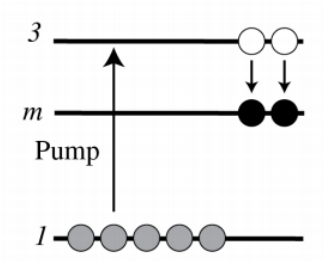


Figure 7.6.1: Optical pumping.

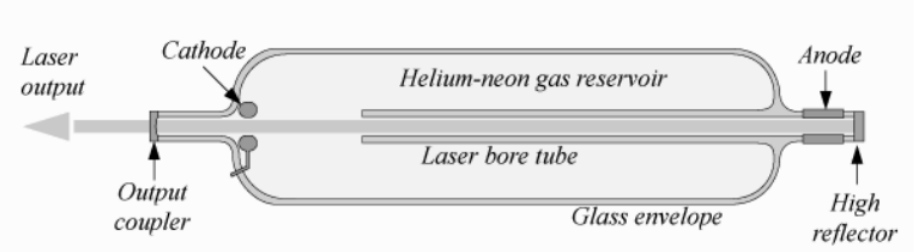


Figure 7.6.2: HeNe laser with spherical external mirrors, a discharge tube with faces at the Brewster angle to minimise reflections, and an anode and cathode for the discharge pumping(from Wikimedia Commons by DrBob/CC BY-SA 3.0).

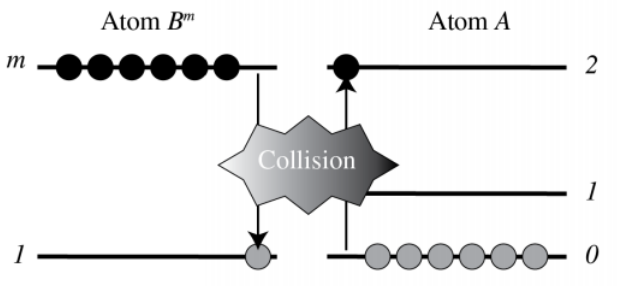


Figure 7.6.3: Pumping atoms A to state 2 by collision with metastable atoms B^m

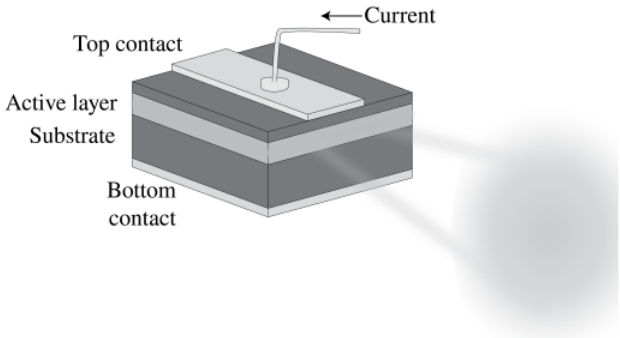


Figure 7.6.4: Semiconductor laser with active p − n junction, polished end faces and current supply for pumping.

Index

A

Airy pattern
2.8: Aberrations

B

birefringence
4.3: Creating and Manipulating Polarisation States

C

camera obscura
3.2: The Camera Obscura
coherence
5.4: Coherence
coherent sum
5.3: Interference of Monochromatic Fields of the Same Frequency
confocal microscopy
6.9: Super-resolution
Cook triplet lens
3.4: Camera in a Smart Phone

D

depth of focus
3.3: The Camera
diffraction
2.2: Introduction
diffraction limited
2.8: Aberrations
distortion
2.8: Aberrations
double Gauss lens
3.4: Camera in a Smart Phone

E

Einstein coefficients
7.3: Amplification
eye
3.5: The Human Eye

F

Fermat’s Principle of Least Action
2.4: Some Consequences of Fermat’s Principle
Fermat's principle
2.3: Principle of Fermat
focal length
3.3: The Camera
Fourier optics
6.8: Fourier Optics
Fraunhofer diffraction
6.7: Fraunhofer Diffraction Revisited
free spectral range
7.2: Optical Resonator

H

hyperbolic materials
6.9: Super-resolution

I

incoherent sum
5.3: Interference of Monochromatic Fields of the Same Frequency

J

Jones matrix
4.3: Creating and Manipulating Polarisation States
4.4: How to verify whether a Jones Matrix is a Linear Polariser or a Wave Plate?

L

lens matrix
2.6: Gaussian Geometrical Optics

M

Maxwell’s Equations
1.3: The Maxwell Equations in Vacuum

O

optical path length
2.3: Principle of Fermat

P

paraxial rays
2.6: Gaussian Geometrical Optics
perfect imaging
2.5: Perfect Imaging by Conic Sections

R

ray matrix
2.6: Gaussian Geometrical Optics
real image
2.6: Gaussian Geometrical Optics
resonator
7.2: Optical Resonator

S

Seidel aberrations
2.8: Aberrations
self coherence function
5.5: Temporal Coherence and the Michelson Interferometer
Snell’s law of refraction
1.10: Reflection and Transmission at an Interface
2.4: Some Consequences of Fermat’s Principle
spatial coherence
5.5: Temporal Coherence and the Michelson Interferometer
spatial light modulator
6.8: Fourier Optics
spherical aberration
2.8: Aberrations
stability condition
7.4: Cavities
STED
6.9: Super-resolution
stigmatic systems
2.5: Perfect Imaging by Conic Sections
stimulated emission
7.3: Amplification
Stimulated Emission Depletion Microscopy
6.9: Super-resolution

T

temporal coherence
5.5: Temporal Coherence and the Michelson Interferometer

V

virtual image
2.6: Gaussian Geometrical Optics

W

wavefront
2.2: Introduction

Detailed Licensing

Overview

Title: Optics (Konijnenberg, Adam, and Urbach)

Webpages: 77

Applicable Restrictions: Noncommercial

All licenses found:

- Undeclared: 59.7% (46 pages)
- CC BY-NC-SA 4.0: 40.3% (31 pages)

By Page

- Optics (Konijnenberg, Adam, and Urbach) - *CC BY-NC-SA 4.0*
  - Front Matter - *CC BY-NC-SA 4.0*
    - TitlePage - *CC BY-NC-SA 4.0*
    - InfoPage - *CC BY-NC-SA 4.0*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Basic Electromagnetic and Wave Optics - *CC BY-NC-SA 4.0*
    - 1.1: What You Should Know and be able to do After Studying This Chapter - *CC BY-NC-SA 4.0*
    - 1.2: Electromagnetic Theory of Optics and Quantum Optics - *CC BY-NC-SA 4.0*
    - 1.3: The Maxwell Equations in Vacuum - *CC BY-NC-SA 4.0*
    - 1.4: Maxwell Equations in Matter - *CC BY-NC-SA 4.0*
    - 1.5: The Scalar and Vector Wave Equation - *CC BY-NC-SA 4.0*
    - 1.6: Time-Harmonic Solutions of the Wave Equation - *CC BY-NC-SA 4.0*
    - 1.7: Time-Harmonic Maxwell Equations in Matter - *CC BY-NC-SA 4.0*
    - 1.8: Electromagnetic Energy - *CC BY-NC-SA 4.0*
    - 1.9: Time-Averaged Energy - *CC BY-NC-SA 4.0*
    - 1.10: Reflection and Transmission at an Interface - *CC BY-NC-SA 4.0*
    - 1.11: Fiber Optics - *CC BY-NC-SA 4.0*
  - 2: Geometrical Optics - *CC BY-NC-SA 4.0*
    - 2.1: What You Should Know and be able to do After Studying This Chapter - *CC BY-NC-SA 4.0*
    - 2.2: Introduction - *CC BY-NC-SA 4.0*
    - 2.3: Principle of Fermat - *CC BY-NC-SA 4.0*
    - 2.4: Some Consequences of Fermat's Principle - *CC BY-NC-SA 4.0*
    - 2.5: Perfect Imaging by Conic Sections - *CC BY-NC-SA 4.0*
    - 2.6: Gaussian Geometrical Optics - *CC BY-NC-SA 4.0*
    - 2.7: Beyond Gaussian Geometrical Optics - *Undeclared*
    - 2.8: Aberrations - *Undeclared*
  - 3: Optical Instruments - *CC BY-NC-SA 4.0*
    - 3.1: What you should know and be able to do after studying this chapter - *Undeclared*
    - 3.2: The Camera Obscura - *Undeclared*
    - 3.3: The Camera - *Undeclared*
    - 3.4: Camera in a Smart Phone - *Undeclared*
    - 3.5: The Human Eye - *Undeclared*
    - 3.6: Magnifying Glasses - *Undeclared*
    - 3.7: Eyepieces - *Undeclared*
    - 3.8: The Compound Microscope - *Undeclared*
    - 3.9: The Telescope - *Undeclared*
  - 4: Polarization - *CC BY-NC-SA 4.0*
    - 4.1: What you should know and be able to do after studying this chapter - *Undeclared*
    - 4.2: Polarisation States and Jones Vectors - *Undeclared*
    - 4.3: Creating and Manipulating Polarisation States - *Undeclared*
    - 4.4: How to verify whether a Jones Matrix is a Linear Polariser or a Wave Plate? - *Undeclared*
    - 4.5: Decompose Elliptical Polarisation into Linear and Circular States - *Undeclared*
  - 5: Interference and Coherence - *CC BY-NC-SA 4.0*
    - 5.1: What you should know and be able to do after studying this chapter - *Undeclared*
    - 5.2: Introduction - *Undeclared*
    - 5.3: Interference of Monochromatic Fields of the Same Frequency - *Undeclared*
    - 5.4: Coherence - *Undeclared*
    - 5.5: Temporal Coherence and the Michelson Interferometer - *Undeclared*
    - 5.6: Longitudinally and Transversely Separated Points - *Undeclared*

- 5.7: Spatial Coherence and Young's Experiment - *Undeclared*
- 5.8: Increase of Spatial Coherence by Propagation - *Undeclared*
- 5.9: Stellar Interferometry - *Undeclared*
- 5.10: Fringe Visibility - *Undeclared*
- 5.11: Interference and polarisation - *Undeclared*

6: Scalar diffraction optics - *CC BY-NC-SA 4.0*

- 6.1: What you should know and be able to do after studying this chapter - *Undeclared*
- 6.2: Introduction - *Undeclared*
- 6.3: Angular Spectrum Method - *Undeclared*
- 6.4: Rayleigh-Sommerfeld Diffraction Integral - *Undeclared*
- 6.5: Intuition for the Spatial Fourier Transform in Optics - *Undeclared*
- 6.6: Fresnel and Fraunhofer Approximations - *Undeclared*
- 6.7: Fraunhofer Diffraction Revisited - *Undeclared*
- 6.8: Fourier Optics - *Undeclared*
- 6.9: Super-resolution - *Undeclared*

7: Lasers - *CC BY-NC-SA 4.0*

- 7.0: What you should know and be able to do after studying this chapter - *Undeclared*
- 7.1: Unique Properties of Lasers - *Undeclared*
- 7.2: Optical Resonator - *Undeclared*
- 7.3: Amplification - *Undeclared*
- 7.4: Cavities - *Undeclared*
- 7.5: Problems with Laser Operation - *Undeclared*
- 7.6: Types of Lasers - *Undeclared*

Back Matter - *CC BY-NC-SA 4.0*

- Index - *CC BY-NC-SA 4.0*
- Glossary - *CC BY-NC-SA 4.0*
- Detailed Licensing - *Undeclared*