



OPTICS

Fundamentals of physical optics

Dr. Gábor Erdei, associate professor

erdei.gabor@ttk.bme.hu



Budapest University of Technology and Economics
Faculty of Natural Sciences, Department of Atomic Physics

v. 04.09.2025

TABLE OF CONTENTS

TABLE OF CONTENTS	2
1. INTRODUCTION, LIGHT MODELS	4
1.1. History of science, light models, technical development	4
1.2. Fermat's principle (principle of the "least time")	6
1.3. Huygens' principle	7
1.4. Young's two-slit experiment	8
1.5. Huygens-Fresnel principle	8
2. ELECTRODYNAMIC MODELLING OF LIGHT	9
2.1. Electrodynamic description of isotropic, linear media	9
2.2. EM wave propagation in dielectric materials	10
2.3. Wave propagation in media of finite conductivity	11
2.4. When $k_{im} \parallel k_{re}$, e.g. light incidence normal to the surface of a conductive material	15
2.5. Energy propagation in EM waves	17
2.6. What career opportunities do physicists have with a degree in optics?	18
3. BEHAVIOR OF PLANE WAVES AT PLANAR INTERFACES	19
3.1. Polarization eigenstates at plane wave – planar surface interactions	19
3.2. Continuity conditions at the boundary of dielectric materials	20
3.3. S-polarization, when E is normal to the plane of incidence	21
3.4. P-polarization, when E is parallel with the plane of incidence	22
3.5. Discussion of Fresnel formulae	23
3.6. Fresnel formulae describing power ratios	24
3.7. Brewster's effect	26
3.8. Total internal reflection (TIR): $n' < n$	26
3.9. Plane wave refraction at planar dielectric/metal interface (supplementary)	28
4. THEORETICAL BACKGROUNDS OF GEOMETRICAL OPTICS	30
4.1. Approximations used	30
4.2. Basic considerations	30
4.3. Slowly varying amplitude approximation	31
4.4. The fundamental equation of geometrical optics, conditions of validity	32
4.5. Local plane-wave approximation	33
5. MODELLING LIGHT PROPAGATION BY RAY-OPTICAL APPROACH	35
5.1. Description of phase evolution in geometrical optics	35
5.2. Description of energy propagation in geometrical optics	36
5.3. Intensity law of geometrical optics	37
5.4. Equation of light rays according to arc-length parametrization	37
6. PARAXIAL APPROXIMATION OF GEOMETRICAL OPTICS	40
6.1. First-order (paraxial) approximation (Gaussian optics)	40
6.2. Image construction in case of a thin lens, properties of paraxial imaging	42
6.3. Ideal imaging in case of a complex optical system	43
6.4. Description of complex lens systems by using principal planes	44
7. TWO-BEAM INTERFERENCE	46
7.1. Concept of an interferogram	46
7.2. Interference of two plane waves in paraxial approximation	49
7.3. Visibility of interference fringes	51
7.4. Example: plane-parallel plate	52
7.5. What happens to light in two-beam interferometers?	53
7.6. Large-angle interference (supplementary)	55

8. MULTIPLE-BEAM INTERFERENCE	57
8.1. Interference of “N” plane waves	57
8.2. Diffraction gratings.....	59
9. SCALAR DIFFRACTION	62
9.1. Basic concepts, diffraction models, Green’s theorem.....	62
9.2. Integral theorem of Helmholtz and Kirchhoff	64
9.3. Fresnel-Kirchhoff diffraction integral	65
10. APPROXIMATIONS TO THE FRESNEL-KIRCHHOFF INTEGRAL.....	68
10.1. Approximation in the near field (Huygens-Fresnel integral)	68
10.2. Paraxial approximation (Fresnel diffraction)	68
10.3. Far-field approximation (Fraunhofer diffraction)	69
10.4. Applications of Fraunhofer diffraction	71
10.5. Applications of Fresnel diffraction	73
10.6. Application of Fraunhofer diffraction: resolution limit of a telescope	75
10.7. Application of Fresnel diffraction: resolution of a microscope (supplementary)	75
10.8. Fraunhofer diffraction pattern of the field in case of a slit (supplementary)	76
11. STATISTICAL OPTICS – TEMPORAL COHERENCE	78
11.1. Introduction – evolution of the concept of coherence	78
11.2. Two-beam interference with two frequencies.....	80
11.3. Two-beam interference for oscillations with multiple frequencies	84
12. INVESTIGATION OF TEMPORAL COHERENCE IN TIME DOMAIN	86
12.1. Quasi-monochromatic waves.....	86
12.2. Coherence function of a quasi-monochromatic oscillation.....	87
12.3. Description of temporally statistic behavior by autocorrelation function	89
12.4. Concept and types of coherence.....	91
13. POLARIZATION	93
13.1. Polarized nature of light	93
13.2. Representation by Jones vectors.....	96
13.3. Description of anisotropic optical elements by Jones matrices	97
13.4. Light including an unpolarized component (supplementary).....	98
14. PROPAGATION OF PLANE WAVES IN ANISOTROPTIC MEDIA.....	99
14.1. Mutual orientation of field vectors	99
14.2. Phase and ray velocity	101
14.3. Fresnel’s equation of wave normals.....	101
14.4. Index ellipsoid	104
14.5. Essentials in crystallography.....	106
14.6. Uniaxial birefringence	106
14.7. Phase retarder plates	108
14.8. Further anisotropic phenomena (supplementary)	108
ACKNOWLEDGEMENTS	109
REFERENCES	109

1. INTRODUCTION, LIGHT MODELS

Sources: [1], [2], [3], [4] + wikipedia

What topics are involved in optics?

- Propagation of light (electromagnetic, EM, waves λ : $10^{-4} \dots 10^{-9}$ m, i.e. 100 μm .. 1 nm)
- Light-matter interactions (e.g. nonlinear optics)
- Emission and absorption of light (e.g. quantum optics)

1.1. History of science, light models, technical development

Assyria	1000 BCE	mirrors, burning lenses (e.g. Nimrud lens? – ground crystal)
Euclid	280 BCE	“ <i>Optics</i> ”, light \equiv vision (rays originate from the eye, <u>propagate along straight lines</u> and interact with objects, only one ray goes towards a single point from the eye) $c = \infty$; law of reflection
Hero (of Alexandria)	I c. CE	“ <i>Catoptrics</i> ” rays propagate from one point to the other along the shortest path; reflection (constructed, proven by the shortest path)
Ptolemy	II c.	law of refraction (experimentally) $\theta' = a \cdot \theta - b \cdot \theta^2 \leftarrow$ at small angles OK!
Sidon (Phoenicia)	II c.	untinted glass (manganese used for bleaching)
Alhazen (Bagdad)	XI c.	Abu Ali al-Hasan <i>ibn al-Haitham</i> ; light: originates from the source, eye is only a “detector”; eye model: based on the principles of camera obscura; particle model (refraction/reflection is a consequence of surface forces); <u>finite</u> and <u>density-dependent</u> speed
Venice	XIII c.	water-clear glass, mirror made with tain (tin foil mirror)
Roger Bacon	XIII c.	magnifying lens, afterwards: eyewear in Italy (Florence, 1280)
Hans & Zacharias Janssen	1600	microscope
Leonard Digges, Hans Lipperhey	1608	telescope (England, Netherlands), Galileo Galilei scientific observations from about 1609
Johannes Kepler	1609 1611	“ <i>Dioptrice</i> ”, approx. for the law of refraction, total reflection, defined focus Keplerian telescope
Cristoph Scheiner	1619	anatomy of the eye, the image is formed on the retina
René Descartes	1637	“ <i>La Dioptrique</i> ”; <u>elastic medium</u> - aether, as if tennis balls make <u>mechanical movement</u> in it; wrong hypothesis for refraction: parallel component of velocity is conserved: $v' \sin \theta' = v \sin \theta$ (the concept of momentum is yet to be discovered); light speed is material-dependent (independently Willebrod Snell also realized it in 1621), Cartesian surfaces, working principle of the eye, explanation of rainbows, $c \approx \infty$
Francesco Maria Grimaldi	1650	observation of “diffraction” at the edge of shadows (he even coined the term), compares light to undulations of liquids (undulations of aether)
Pierre de Fermat	1657	nature takes the easiest way: concept of the least time, “optical resistance” \rightarrow <u>finite speed</u> , a new derivation of the law of refraction \rightarrow accordingly, light propagates slower in glass than in air
Robert Hooke	1665	“ <i>Micrographia</i> ”, observation of iridescence (colored patterns in colorless materials: feather, mollusk shell); independently Robert Boyle also discovered it; Hooke considered light as oscillations (1666 – Newton’s splitting and reconstruction of colors)
Isaac Newton	1672	reflective telescope
Olaf Rømer	1675	measuring light speed with 24% accuracy based on Jupiter moons (actually he used the Doppler effect: due to the speed of Earth the Jupiter-Earth distance continuously grows or reduces) James Bradley confirmed it in 1728 by telescopic measurements – “aberratio”

Christian Huygens (or Huyghens)	1679	aether (delicate, elastic material), light propagates as an elastic wave (he rather thought of it as a pulse, since he left no notes behind about periodicity by wavelength) in this (published before Newton, still it was only believed more than 100 years later!), small balls jostling each other, spherical wavelets – it is in correspondence with Fermat’s principle
Isaac Newton	1704	“ <i>Opticks</i> ”; got the corpuscular model stuck for 100 years, color: size of a particle; white light = <u>sum of colors</u> (coined the term “spectrum”); Newton fringes
Chester Moore Hall	1750	<u>achromatic doublet (compensates the wavelength dependence of focal length)</u>
Thomas Young	1801	interferometric experiment and explanation (path difference matters) → spatial coherence; <u>light is a periodic wave</u> ; color is in relationship with wavelength (explanation of the color of thin films); <u>accommodation of the eye lens (crystalline lens), basics of color vision</u>
Étienne-Louis Malus	1809	polarization (Bartholinus showed birefringence of calcite in 1669, also examined by Huygens and Newton formerly); pointed out that both reflected and scattered light are polarized; coined the term “polarization”
Joseph von Fraunhofer	1814	<u>invented the spectroscope; identification and description of lines in the solar spectrum (William Hyde Wollaston also discovered them in 1804)</u>
Thomas Young	1817	<u>transversality</u> (in the wake of Dominique François Arago’s experiments made with polarized light), linear polarization
Augustin Jean Fresnel	1818	diffraction experiments → mathematical description of the wave nature of light; due to transversality aether cannot be liquid only solid; explanation of dispersion by using the corpuscular structure of materials; crystal optics
Louis Mande Daguerre	1839	<u>photography made on metal plates</u>
József Petzvál	1840	<u>high-speed portrait objective (decreasing field curvature)</u>
Christian Doppler	1842	changes in the color of moving light sources (star), (Doppler effect)
Jean Bernard Léon Foucault	1850	measuring light speed in water, air (slower in water): corpuscular model \emptyset
Armand Hypolite Louis Fizeau	1851	light is partially dragged on by a moving material (based on Fresnel’s hypothesis), demonstrated by running water → theory of special relativity
Kohlrausch & Weber	1856	measuring the speed of EM waves = light speed
Robert Wilhelm Bunsen Gustav Kirchhoff	1861	<u>prismatic spectroscope, explanation of lines in the solar spectrum: absorption, emission → spectrum analysis (later quantum mechanics)</u>
Armand Hypolite Louis Fizeau	1862	<u>investigation of the visibility of interferograms (→ temporal coherence)</u>
James Clerk Maxwell	1862	theory of electromagnetic waves (following Michael Faraday)
Thomas Alva Edison	1879	<u>operational incandescent lamp (first demonstration: 1802, Humphry Davy)</u>
Abbe-Schott-Zeiss	1884	<u>debut of the manufacturing of optical glasses</u>
Michelson & Morley	1887	interferometric experiment → <u>no absolute aether?</u> ; $c' = c + v$?
Heinrich Hertz	1888	light is an <u>electromagnetic</u> wave – experimental proof (standing wave experiments with EM radio waves @ 55 MHz → speed, polarization)
Harold Dennis Taylor	1892	<u>discovered a ZnO thin layer on a lens contaminated by acid rain and sunlight, that partially reduced reflection</u>
Max Planck	1900	blackbody radiation, <u>EM field is quantized</u> : $\Delta E = h \cdot \nu$
Willem De Sitter	1913	examination of double stars → <u>c is independent of the speed of the light source</u> → ← Newtonian mechanics (support to special relativity)
Albert Einstein	1905	photo-electric effect; corpuscular model (photon); theory of special relativity → interaction between moving bodies and light
	1917	<u>stimulated emission</u>
Dénes Gábor	1947	principle of holograms
Theodore H. Maiman	1960	<u>laser</u>

Tab. 1 Brief history of optics in the fields of theory and technology.

1.2. Fermat's principle (principle of the “least time”)

Hero already stated the laws of reflection in the 1st century BCE. His idea was that during reflection light propagates along the path that is the shortest geometrically. This consideration however cannot be applied for refractive media due to the various propagation velocities in them. Fermat managed to make this generalization in the 17th century. In its original form Fermat's principle states: light propagates between two separate points along the path that can be taken in *the least amount of time*. This is not a law, but a useful tool to calculate and visualize the path of propagation of light. Since the concept of *path of propagation* of general light beams (i.e. not a spherical or plane wave) only bears a meaning in the framework of geometrical optics, Fermat's principle can only be used in the approximation of geometrical optics too.

Fermat assumed that denser media hinder the propagation of light more than those of lower density, and introduced the concept of “optical resistance” to characterize this property of materials. The corresponding contemporary terminology is the *refractive index* (n):

$$n \triangleq \frac{c}{v}, \quad (1)$$

where c is the light speed measured in vacuum and “ v ” in a given medium. This having a large value, light travels distances typical to optical equipment in a very short time, thus the “least time” introduced by Fermat is of little practical usability. Since the time elapsed is proportional to the path travelled (differentially), we may investigate the path length instead of time. However, light travels various distances during the same time in media of different refractive index. The treatment can be simplified by introducing the *optical path length* (OPL):

$$\begin{aligned} \Delta t &= \int_P^{P'} dt ; OPL \triangleq c \cdot \Delta t = \int_P^{P'} c dt ; v = \frac{dl}{dt} \Rightarrow \\ dt &= \frac{dl}{v} ; OPL = \int_P^{P'} \frac{c}{v(\mathbf{r})} dl = \int_P^{P'} n(\mathbf{r}) dl \end{aligned} \quad (2)$$

The optical path is proportional to the time taken by light going from point P to P' : by definition this is the distance light travels during the same time it needs to get from P to P' in vacuum. In this context Fermat's principle goes like this: light propagates between two points by the trajectory along which the optical path length is the shortest. This is a special form of the statement, valid only in so-called *regular* domains where no light rays cross each other.

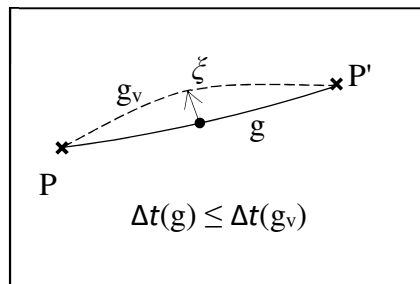


Fig. 1 Explanation to Fermat's principle by varying the trajectory.

According to the completely general phrasing that more precisely agrees with the geometrical laws of light propagation (viz. differential equations describing trajectories of light rays): light

propagates between two points along a path (g), that being slightly varied as a function of a parameter ξ that characterizes the trajectory, the OPL measured along the new trajectory (g_v) does not differ in first order from the original one (only second- or higher-order changes may occur). Mathematically speaking, the real propagation path is stationary to first order as a function of ξ :

$$\left. \frac{dOPL}{d\xi} \right|_{\xi=\xi_0} = 0 \quad (3)$$

That is, the real path can have:

- a minimum time
- a maximum time
- an inflection
- identical time for all neighboring paths.

By using Fermat's principle we can describe the laws of refraction, reflection, characteristics of the focusing effect of mirrors and lenses and so on. Its further importance lies in that it gave a first impulse to the development of *variation calculus*.

1.3. Huygens' principle

Every point of a wavefront is the starting point of new spherical wavelets. The new wavefront can be obtained by constructing the envelope surface of the elementary waves (or wavelets).

Example: propagation of a plane wave through an aperture

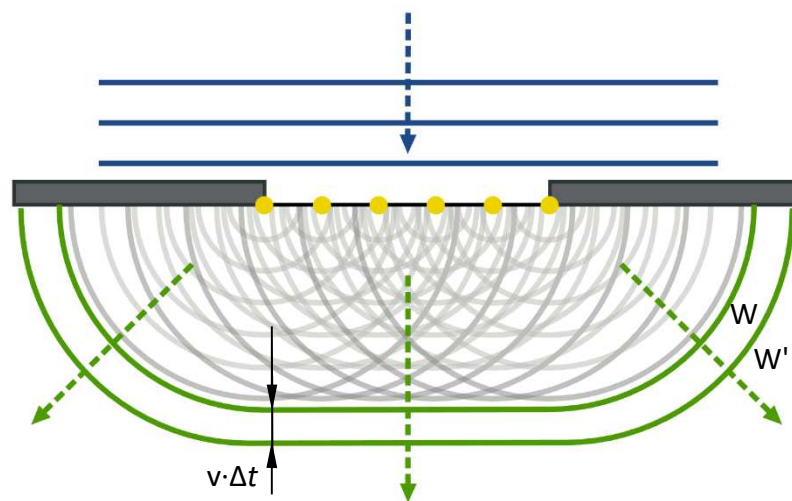


Fig. 2 Demonstration of Huygens' principle. W – wavefront at instant t , W' – new wavefront at instant $t + \Delta t$. https://en.wikipedia.org/wiki/Huygens%E2%80%93Fresnel_principle

Assumption: the envelope of wavelets going backwards should not be taken into account (the principle simply states this, not having any other way to explain the absence of backward propagation.) Repeating the previous construction on W' we can determine a newer envelope W'' . The wavefront can be found here at instant $t+2\Delta t$.

1.4. Young's two-slit experiment

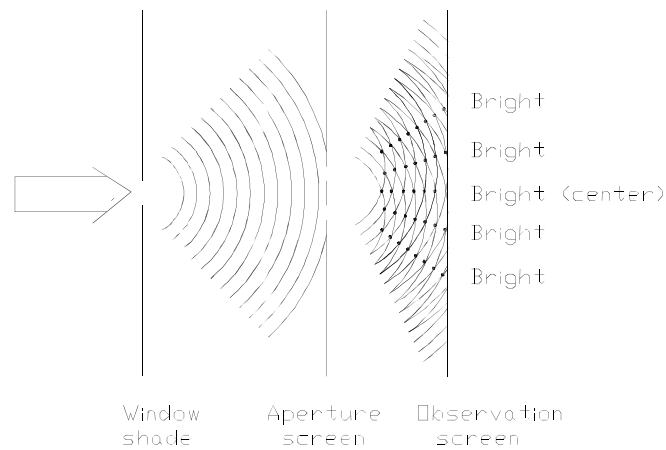


Fig. 3 Scheme of Young's two-slit experiment.

At places where the waves are in identical phase we can observe intensity maxima (if the path difference is an integer multiple of λ), therefore bright lines take shape in space. Furthermore, Young described the physical explanation of the operation of colored thin layers (e.g. soap bubble, oil film on water), and determined the value of λ for light of different colors based on Newton's interference experiments performed on thin sheets.

The envelope introduced by Huygens did not account for diffraction and interference.

Young + Fresnel: explanation of interference by wave theory.

1.5. Huygens-Fresnel principle

Every point of a wavefront is the starting point of new spherical wavelets. The value of the light signal (disturbance) can be obtained at an arbitrary point in space by superimposing the elementary waves (a more precise mathematical description will be given later in diffraction theory).

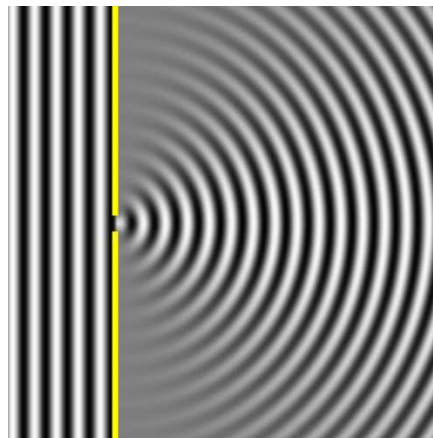


Fig. 4 Light enters the shadow region too due to diffraction. (source: wikipedia)

2. ELECTRODYNAMIC MODELLING OF LIGHT

2.1. Electrodynamic description of isotropic, linear media

Maxwell expressed his famous equations that summarize the laws of the electric field (**E**) and magnetic field (**B**) in the so-called microscopic form. To solve these, all currents and charges present in a given system have to be considered, even if these are on the atomic scale. In optics the microscopic charges-currents are of importance, since these determine the characteristics of light-matter interactions. However, the microscopic formulation yields an equation system of unsolvable complexity, thus we need another approach.

At optical frequencies ($\nu \approx 10^{14}$ Hz) the wavelength of EM radiation ($\lambda = 100\text{-}1000$ nm) is orders of magnitude larger than atomic dimensions (lattice period ≈ 0.1 nm). This means that the effects of microscopic currents and charges present in the electron cloud of single atoms can be averaged for domains comprising many atoms, since even these domains are much smaller than the wavelength of light.

The explanation for this averaging of microscopic charges and currents is as follows. Atoms in an electric field become polarized, and the sum of elementary dipoles can be taken into account by the density of electric dipole moment **P** (polarization). Similarly, the magnetic field gets the elementary magnets (orbital momentum) induced by circular currents of atomic scale aligned, the sum of which can be incorporated in the magnetic dipole moment density **M** (magnetization). The microscopic currents/charges cancel each other inside the averaging domain, thus we only observe the effect of charges/currents being on the surface of the domain of our examination. The below figure demonstrates the concept presented above:

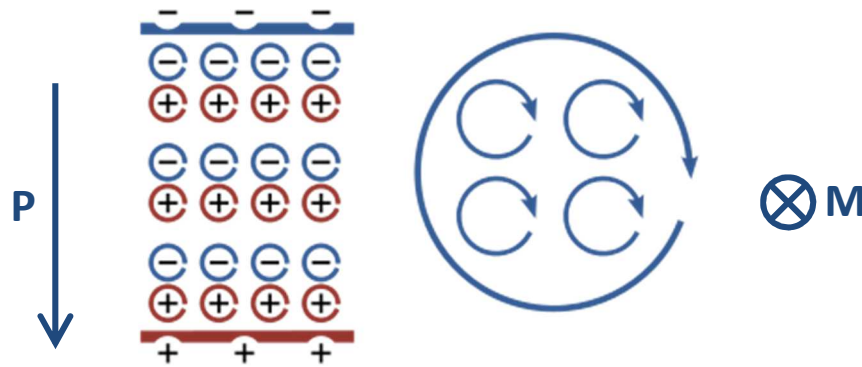


Fig. 5 Demonstration of polarization and magnetization. (source: wikipedia)

By using these new field quantities have been introduced in electrodynamics:

$$\mathbf{P} \triangleq \epsilon_0 \chi_e \mathbf{E} ; \mathbf{D} \triangleq \epsilon_0 \mathbf{E} + \mathbf{P} ; \mathbf{H} \triangleq \frac{\mathbf{B}}{\mu_0} - \mathbf{M}, \quad (4)$$

which can be used to formulate the so-called Maxwell's macroscopic equations in matter (**D** is the *dielectric displacement* and **H** is the *magnetizing field*). In these ρ and \mathbf{j} only include the density of free charges and currents, see (5).

It is worth noting that the orbital momentum of atoms has such a large inertia, that at optical frequencies there is barely magnetic interaction between the electromagnetic (EM) field and atoms. Therefore, it is almost always true that $\mu_r = 1$. This is also the reason for writing equations usually only for **E** when dealing with EM waves.

$$\left. \begin{array}{l} 1) \text{ curl } \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{j} \\ 2) \text{ curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \\ 3) \text{ div } \mathbf{B} = 0 \\ 4) \text{ div } \mathbf{D} = \rho \end{array} \right\} \quad (5)$$

$$\mathbf{B} = \mu_0 \mu_r \mathbf{H} = \mu \mathbf{H}$$

$$\mathbf{D} = \varepsilon_0 \varepsilon_r \mathbf{E} = \varepsilon \mathbf{E}$$

$$\mathbf{j} = \sigma \mathbf{E}$$

Larger context: connections between \mathbf{P} - \mathbf{E} and \mathbf{M} - \mathbf{H} are usually discussed in linear approximation, but in the general case (e.g. for strong fields and special crystals) the polarization and magnetization also depend on the electric and magnetic fields. In addition, inside inhomogeneous media ε_r , μ_r and the resulting refractive index n are position-dependent! The last equation expressing the relationship between free charge-caused currents and the electric field is called Ohm's differential law. The σ conductivity included is a macroscopic quantity for matter, similarly to \mathbf{D} and \mathbf{H} . In classical physics this expression can be interpreted by using the Drude model (not discussed here, see solid-state physics), which derives the physical properties of conductive materials from interactions between free charges (electron gas, plasma), and fixed atomic cores. In the present form this equation is only valid for frequencies much lower than the plasma frequency of the electron gas ($\omega \ll \omega_p$). In case of metals, the wavelength corresponding to plasma frequency is about $\lambda_p = 70\text{-}80 \text{ nm}$ (deeply in the UV range), thus the approximation can be used for visible wavelengths. Near plasma frequency the phase between \mathbf{j} and \mathbf{E} can shift, and material properties show a strong dependence on frequency. (At plasma frequency $\varepsilon = 0$.)

2.2. EM wave propagation in dielectric materials

Assumptions: $\sigma = 0$, $\rho = 0$ (i.e. there are no free charges and the material cannot carry a current). From these the wave equation can be derived (homogeneous differential equations):

$$\nabla^2 \mathbf{E} - \varepsilon \mu \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0 ; \nabla^2 \mathbf{B} - \varepsilon \mu \frac{\partial^2 \mathbf{B}}{\partial t^2} = 0 ; c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} ; n^2 = \varepsilon_r \mu_r \text{ (Maxwell's formula)} \quad (6)$$

where we introduced the nabla (del) operator denoted by symbol ∇ ($\hat{x}, \hat{y}, \hat{z}$ are unity vectors):

$$\nabla \triangleq \frac{\partial}{\partial x} \hat{x} + \frac{\partial}{\partial y} \hat{y} + \frac{\partial}{\partial z} \hat{z}. \quad (7)$$

The detailed development will be discussed in the subsection 2.3. The simplest solution to equations (6) is provided by a (monochromatic) plane wave of ω angular frequency, which can be written by using the complex notation as:

$$\begin{aligned} \tilde{\mathbf{E}}(\mathbf{r}, t) &= \mathbf{E}_0 \cdot e^{i(\omega t - \mathbf{k} \cdot \mathbf{r} + \varphi_1)} \\ \tilde{\mathbf{B}}(\mathbf{r}, t) &= \mathbf{B}_0 \cdot e^{i(\omega t - \mathbf{k} \cdot \mathbf{r} + \varphi_2)} \end{aligned} \quad (8)$$

Here \mathbf{E}_0 and \mathbf{B}_0 vector amplitudes have real values, i.e. we consider the wave to be linearly polarized. The general treatment of polarization will be examined in Chapter 13. The length of the wave vector \mathbf{k} can be formulated as follows (dispersion relation):

$$|\mathbf{k}| = \frac{2\pi}{\lambda} = n \frac{2\pi}{\lambda_0} = n \frac{2\pi}{c \cdot T} = n \frac{\omega}{c} = \frac{\omega}{v} \quad (9)$$

In equations (8) describing plane waves it is important to note the sign of the exponent (i.e. the phase) that characterizes spatial and temporal dependence. In the approach of physics, light carries the phase of former instants (delays) moving in the direction of propagation; this is called a wave propagation of *negative phase* (since $\mathbf{k}\mathbf{r}$ is negative). In the frequently used alternative method (we will use it too) a complex conjugate of the equations are taken in order to have a positive sign on spatial propagation. This is the so-called *positive-phase propagation*. Making aware of the currently used convention is important, having serious effects on the results if material properties are complex.

Substituting the above ansatzes into the 2nd Maxwell equation we obtain:

$$\tilde{\mathbf{B}} = \frac{\mathbf{k} \times \tilde{\mathbf{E}}}{\omega} \Rightarrow \mathbf{B}_0 \cdot e^{i(\omega t - \mathbf{k}\mathbf{r} + \varphi_2)} = \frac{\mathbf{k}}{\omega} \times \mathbf{E}_0 \cdot e^{i(\omega t - \mathbf{k}\mathbf{r} + \varphi_1)}, \quad (10)$$

which after simplification can only be fulfilled if $\varphi_1 = \varphi_2$ due to the equality of phases. Thus, the EM wave is of transverse nature indeed (since $\tilde{\mathbf{B}} \perp \tilde{\mathbf{E}} \perp \mathbf{k}$), and $\tilde{\mathbf{E}} - \tilde{\mathbf{B}}$ vibrate in phase with each other. Below we will show that in absorbing media \mathbf{k} is complex, i.e. $\varphi_1 \neq \varphi_2$.

2.3. Wave propagation in media of finite conductivity

In the general case, vectors \mathbf{D} and \mathbf{E} (\mathbf{E} and \mathbf{B} too) may be phase-shifted relative to each other, i.e. the dielectric permittivity (and the resulting wavenumber and refractive index) are complex quantities. For this case we give an example, that of materials with finite (i.e. not zero) conductivity (typically metals), by which one can easily interpret the meaning of the newly introduced complex quantities. We still suppose that $\rho = 0$, and for simplicity we present the detailed development of the wave function only for \mathbf{E} .

Let us substitute Ohm's differential law that describes the relationship between current density and \mathbf{E} into the 1st Maxwell equation:

$$\text{curl } \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{j} \rightarrow \text{curl } \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \sigma \mathbf{E} \rightarrow \text{curl } \mathbf{H} = \varepsilon \frac{\partial \mathbf{E}}{\partial t} + \sigma \mathbf{E}, \quad (11)$$

and differentiate the resulting equation by time:

$$\text{curl} \frac{\partial \mathbf{H}}{\partial t} = \varepsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} + \sigma \frac{\partial \mathbf{E}}{\partial t}. \quad (12)$$

We convert the 2nd Maxwell equation as:

$$\text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \rightarrow \text{curl } \mathbf{E} = -\frac{\partial \mu \mathbf{H}}{\partial t} \rightarrow \frac{\text{curl } \mathbf{E}}{\mu} = -\frac{\partial \mathbf{H}}{\partial t} \rightarrow \text{curl} \frac{\text{curl } \mathbf{E}}{\mu} = -\text{curl} \frac{\partial \mathbf{H}}{\partial t} \quad (13)$$

and plug into (12):

$$-\text{curl} \frac{\text{curl } \mathbf{E}}{\mu} = \varepsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} + \sigma \frac{\partial \mathbf{E}}{\partial t} \quad (14)$$

Using the $\nabla \times (f \cdot \mathbf{A}) \equiv \nabla f \times \mathbf{A} + f \cdot (\nabla \times \mathbf{A})$ identity we get:

$$-\text{grad} \frac{1}{\mu} \times \text{curl } \mathbf{E} - \frac{1}{\mu} \cdot \text{curl curl } \mathbf{E} = \varepsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} + \sigma \frac{\partial \mathbf{E}}{\partial t}. \quad (15)$$

Caution: Generally, the position dependence of μ has to be taken into account at taking curl! (Developing the wave equation pertaining to \mathbf{B} , spatial variation of ϵ and σ is to be considered.)

$$-\mu \cdot \text{grad} \frac{1}{\mu} \times \text{curl} \mathbf{E} - \text{curl} \text{curl} \mathbf{E} = \epsilon \mu \frac{\partial^2 \mathbf{E}}{\partial t^2} + \sigma \mu \frac{\partial \mathbf{E}}{\partial t}. \quad (16)$$

since $\text{grad}(\mu^{-1}) = -\text{grad}(\mu)/\mu^2$

$$\frac{\text{grad} \mu}{\mu} \times \text{curl} \mathbf{E} - \text{curl} \text{curl} \mathbf{E} = \epsilon \mu \frac{\partial^2 \mathbf{E}}{\partial t^2} + \sigma \mu \frac{\partial \mathbf{E}}{\partial t}. \quad (17)$$

For optical materials mostly used in practice it is true, that $\mu = \mu_0 = \text{const.}$, therefore the following approximation holds (upper estimate):

$$\left| \frac{\text{grad} \mu}{\mu} \times \text{curl} \mathbf{E} \right| \ll |\text{curl} \text{curl} \mathbf{E}| \Rightarrow \frac{|\text{grad} \mu|}{\mu} \ll \frac{|\text{curl} \text{curl} \mathbf{E}|}{|\text{curl} \mathbf{E}|}, \quad (18)$$

hence, after neglecting, our equations leaves us this:

$$-\text{curl} \text{curl} \mathbf{E} = \epsilon \mu \frac{\partial^2 \mathbf{E}}{\partial t^2} + \sigma \mu \frac{\partial \mathbf{E}}{\partial t}. \quad (19)$$

Similarly (though after a few more steps) we can also derive it for \mathbf{B} too:

$$-\text{curl} \text{curl} \mathbf{B} = \epsilon \mu \frac{\partial^2 \mathbf{B}}{\partial t^2} + \sigma \mu \frac{\partial \mathbf{B}}{\partial t}, \quad (20)$$

where we had to apply the following approximations: (upper estimate):

$$\frac{|\text{grad} \sigma|}{\sigma} \ll \frac{|\text{curl} \mathbf{E}|}{|\mathbf{E}|} \quad \text{and} \quad \frac{|\text{grad} \epsilon|}{\epsilon} \ll \frac{|\text{curl} \mathbf{E}|}{2|\mathbf{E}|}. \quad (21)$$

Since $\mu = \text{const.}$, we can expand the left of the second condition by $\mu + \text{Maxwell's formula (6)}$:

$$\frac{|\text{grad} n^2|}{n^2} = \frac{2|\text{grad} n|}{n} \Rightarrow \frac{|\text{grad} n|}{n} \ll \frac{|\text{curl} \mathbf{E}|}{4|\mathbf{E}|}. \quad (22)$$

In contrast to permeability, σ and ϵ_r can spatially change even in case of optical materials. Hence, in order to demonstrate the meaning of the above condition, we apply it for the simple situation when a plane wave of form (8) propagates in an isotropic medium. Since in such media $\mathbf{k} \perp \mathbf{E}$, therefore the right side of (22) can be written in the following way:

$$\frac{|\text{curl} \mathbf{E}|}{4|\mathbf{E}|} = \frac{|-i\mathbf{k} \times \mathbf{E}|}{4|\mathbf{E}|} = k \cdot \frac{|\mathbf{E}|}{4|\mathbf{E}|} = \frac{\pi}{2\lambda}, \quad (23)$$

due to which (22) simplifies to:

$$\frac{|\text{grad} n| \cdot \lambda}{n} = \frac{\Delta n}{n} \ll \frac{\pi}{2}. \quad (24)$$

This implies that the relative refractive index change should be much less than $\pi/2$ if measured along a distance of one wavelength. This is not the case for structures with spatially rapidly changing refractive index, e.g. subwavelength diffraction gratings and scattering media. Besides, the condition is even violated at the interface of different materials, which is taken into account in electrodynamics by the *continuity conditions* of Maxwell's equations (see later).

By using $\nabla \times (\nabla \times \mathbf{A}) \equiv \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}$ vector identity, equations (19)-(20) transform into:

$$\begin{aligned}\nabla^2 \mathbf{E} - \text{grad div } \mathbf{E} &= \varepsilon\mu \frac{\partial^2 \mathbf{E}}{\partial t^2} + \sigma\mu \frac{\partial \mathbf{E}}{\partial t} \\ \nabla^2 \mathbf{B} - \text{grad div } \mathbf{B} &= \varepsilon\mu \frac{\partial^2 \mathbf{B}}{\partial t^2} + \sigma\mu \frac{\partial \mathbf{B}}{\partial t}.\end{aligned}\tag{25}$$

Due to Maxwell equations 3),4) and (21), the second term on the left handside of the equations is zero in both cases, thus:

$$\begin{aligned}\nabla^2 \mathbf{E} &= \varepsilon\mu \frac{\partial^2 \mathbf{E}}{\partial t^2} + \sigma\mu \frac{\partial \mathbf{E}}{\partial t} \\ \nabla^2 \mathbf{B} &= \varepsilon\mu \frac{\partial^2 \mathbf{B}}{\partial t^2} + \sigma\mu \frac{\partial \mathbf{B}}{\partial t}.\end{aligned}\tag{26}$$

These equations describe wave propagation in media having finite conductivity (e.g. metals), together they are called the *telegram equations*. The above equations are linear, thus harmonic functions serve as a solution to them too. Let us examine now those solutions that exhibit a temporal dependence of ω angular frequency only. Using complex terminology these function can be written as:

$$\tilde{\mathbf{E}}(\mathbf{r}, t) = \tilde{\mathbf{E}}(\mathbf{r}) \cdot e^{i\omega t}.\tag{27}$$

Since the first temporal derivative corresponds to a multiplication by $i\omega$, the wave equation takes a time-invariant shape, the so-called *Helmholtz equation* (similarly for \mathbf{B}):

$$\nabla^2 \tilde{\mathbf{E}}(\mathbf{r}) + \varepsilon\mu\omega^2 \tilde{\mathbf{E}}(\mathbf{r}) - i\mu\sigma\omega \tilde{\mathbf{E}}(\mathbf{r}) = 0.\tag{28}$$

This equation formally becomes the same as the one we get for dielectric media, if we introduce the complex dielectric permittivity ($\varepsilon_{\text{im}} > 0$ for absorption):

$$\tilde{\varepsilon} \triangleq \varepsilon - i \frac{\sigma}{\omega} = \varepsilon_{\text{re}} - i \cdot \varepsilon_{\text{im}} \rightarrow \nabla^2 \tilde{\mathbf{E}}(\mathbf{r}) + \tilde{\varepsilon}\mu\omega^2 \tilde{\mathbf{E}}(\mathbf{r}) = 0.\tag{29}$$

Assuming a linearly polarized plane-wave solution of the Helmholtz equation:

$$\tilde{\mathbf{E}}(\mathbf{r}) = \mathbf{E}_0 \cdot e^{-i\mathbf{k}\mathbf{r}}\tag{30}$$

we obtain that:

$$-\mathbf{k}^2 \mathbf{E}_0 + \tilde{\varepsilon}\mu\omega^2 \mathbf{E}_0 = 0 \Rightarrow \mathbf{k}^2 = \tilde{\varepsilon}\mu\omega^2,\tag{31}$$

i.e. the wave vector is complex:

$$\begin{aligned}\tilde{\mathbf{k}} &\triangleq \mathbf{k}_{\text{re}} - i \cdot \mathbf{k}_{\text{im}}, \\ \tilde{\mathbf{E}}(\mathbf{r}, t) &= \mathbf{E}_0 \cdot e^{i(\omega t - \mathbf{k}_{\text{re}}\mathbf{r})} \cdot e^{-\mathbf{k}_{\text{im}}\mathbf{r}}.\end{aligned}\tag{32}$$

Hence, the complex wave vector implies an exponentially decaying field amplitude in the direction of \mathbf{k}_{im} . In summary: the complex permittivity and wave vector express that a material attenuates the radiation incident upon it. The degree of attenuation is characterized by the penetration depth (δ), that is by definition the distance at which the amplitude decreases to its 1/e fraction. From the above equation:

$$\delta \triangleq \frac{1}{k_{\text{im}}}.\tag{33}$$

A complex dielectric permittivity can only describe absorbing media (conductive materials, or strongly absorbing dielectrics), as opposed to the wave vector, which can even be complex in other cases, such as at total internal reflection (to be discussed in Chapter 3).

In absorbing media the components of $\tilde{\mathbf{k}}$ can be calculated by solving the equation below:

$$\tilde{\mathbf{k}}^2 = (\mathbf{k}_{\text{re}} - i \cdot \mathbf{k}_{\text{im}})^2 = k_{\text{re}}^2 - k_{\text{im}}^2 - i \cdot 2\mathbf{k}_{\text{re}}\mathbf{k}_{\text{im}} = \tilde{\epsilon}\mu\omega^2, \quad (34)$$

from which we get the following after separating the real and imaginary parts and multiplying out $k_0^2 = \omega^2/c^2$:

$$k_{\text{re}}^2 - k_{\text{im}}^2 = \frac{\omega^2}{c^2} \epsilon_{\text{re}} \mu c^2 = k_0^2 \epsilon_{\text{re}} \mu c^2 \quad \text{and} \quad 2\mathbf{k}_{\text{re}}\mathbf{k}_{\text{im}} = \frac{\omega^2}{c^2} \epsilon_{\text{im}} \mu c^2 = k_0^2 \epsilon_{\text{im}} \mu c^2. \quad (35)$$

In general, \mathbf{k}_{im} and \mathbf{k}_{re} are not parallel (e.g. they are downright normal to each other in case of the evanescent wave of total internal reflection). Their relative direction is controlled by the 4th Maxwell equation, from which the following condition can be formulated for a plane wave of ω angular frequency (30):

$$\mathbf{E}_0 \cdot \tilde{\mathbf{k}} = 0 \rightarrow \mathbf{E}_0 \cdot \mathbf{k}_{\text{re}} = i \cdot \mathbf{E}_0 \cdot \mathbf{k}_{\text{im}}. \quad (36)$$

A detailed analysis of the above relationship can be found in the 2.2.2.2 subsection of textbook [6], vol. 1. In subsection 2.4 we only discuss the simple case when \mathbf{k}_{im} and \mathbf{k}_{re} are parallel to each other.

Analogously to Maxwell's formula, we can define the complex refractive index:

$$n^2 = \epsilon_r \mu_r \rightarrow \tilde{n}^2 \triangleq \frac{\tilde{\epsilon}\mu}{\epsilon_0\mu_0} = \tilde{\epsilon}\mu c^2, \quad (37)$$

that is

$$\tilde{n} = n - i\kappa, \quad (38)$$

where κ is called the extinction coefficient ($\kappa > 0$ for absorption). The same way as the determination of k_{re} and k_{im} in subsection 2.4 we can calculate the values of n_{re} and n_{im} :

$$\kappa^2 = n_{\text{im}}^2 = \mu c^2 \left(-\frac{\epsilon_{\text{re}}}{2} + \frac{1}{2} \sqrt{\epsilon_{\text{re}}^2 + \epsilon_{\text{im}}^2} \right) ; \quad n^2 = n_{\text{re}}^2 = \mu c^2 \left(\frac{\epsilon_{\text{re}}}{2} + \frac{1}{2} \sqrt{\epsilon_{\text{re}}^2 + \epsilon_{\text{im}}^2} \right). \quad (39)$$

The complex refractive index is most frequently used in the Fresnel formulae that quantify surface reflectance, as well as at the interferometric description of thin-layer structures. Both dielectric permittivity and the refractive index strongly depend on the wavelength of the radiation subject to our investigation. In the below figure we present this wavelength-dependence for the case of a typical metal: aluminum (see also Kramers-Krönig relations).

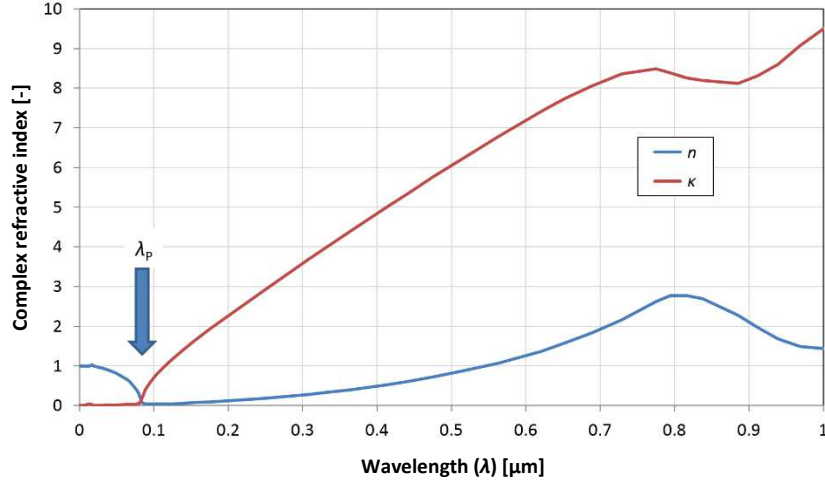


Fig. 6 Refractive index and extinction coefficient of aluminum. λ_p is the wavelength corresponding to the plasma frequency. (source: <http://refractiveindex.info>)

As we mentioned earlier, in the general case not only conductivity can be described by complex permittivity, but any kind of absorption. In addition to the plasma formed by free electrons as discussed above, single atoms, ions and molecules can also resonate with the incident radiation, resulting in absorption. Since these processes take place at different wavelengths, ϵ_{re} and ϵ_{im} exhibit diverse frequency-dependence (dispersion). The below figure presents an example for a (fictional) dielectric material.

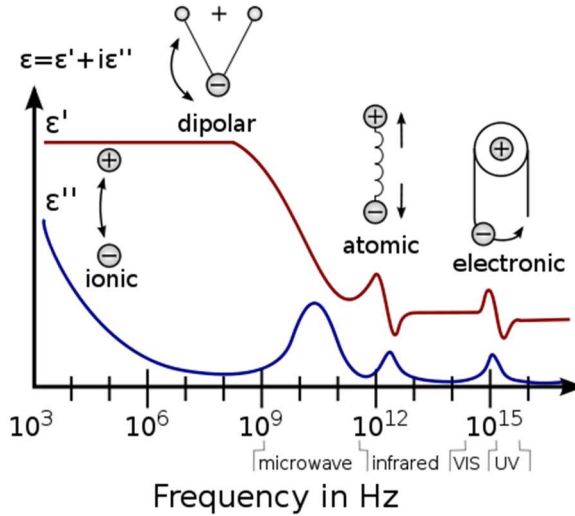


Fig. 7 Important resonances taking place in materials. (source: wikipedia)

2.4. When $\mathbf{k}_{im} \parallel \mathbf{k}_{re}$, e.g. light incidence normal to the surface of a conductive material

According to equations (37) and (31) the below relationship is always valid for plane waves:

$$\tilde{\mathbf{k}}^2 = \tilde{n}^2 \frac{\omega^2}{c^2} = \tilde{n}^2 k_0^2. \quad (40)$$

It is not necessarily true though, that $k_{re} = n_{re}k_0$ and $k_{im} = n_{im}k_0$, since the real and imaginary components of $\tilde{\mathbf{k}}$ can even point in different directions. Let us examine the case of $\mathbf{k}_{im} \parallel \mathbf{k}_{re}$ (e.g. if light impinges perpendicularly at the surface of a conductive material). From (34):

$$(\mathbf{k}_{\text{re}} - i\mathbf{k}_{\text{im}})^2 = k_{\text{re}}^2 - k_{\text{im}}^2 - i2\mathbf{k}_{\text{re}}\mathbf{k}_{\text{im}} = k_{\text{re}}^2 - k_{\text{im}}^2 - i2k_{\text{re}}k_{\text{im}} = a - ib, \quad (41)$$

where we introduced the following auxiliary notation based on (35):

$$a \triangleq k_0^2 \varepsilon_{\text{re}} \mu c^2 \text{ and } b \triangleq k_0^2 \varepsilon_{\text{im}} \mu c^2. \quad (42)$$

By using these the below two-variable equation system can be formulated:

$$\begin{cases} k_{\text{re}}^2 - k_{\text{im}}^2 = a \\ 2k_{\text{re}}k_{\text{im}} = b \end{cases} \quad (43)$$

From equation (43) we can express k_{im}^2 :

$$4(k_{\text{im}}^2)^2 + 4ak_{\text{im}}^2 - b^2 = 0. \quad (44)$$

The solution to this equation being quadratic in terms of k_{im}^2 is (the – signed alternative bears no physical meaning):

$$k_{\text{im}}^2 = \frac{-a + \sqrt{a^2 + b^2}}{2} = -\frac{k_0^2 \varepsilon_{\text{re}} \mu c^2}{2} + k_0^2 \frac{1}{2} \sqrt{(\varepsilon_{\text{re}} \mu c^2)^2 + (\varepsilon_{\text{im}} \mu c^2)^2}, \quad (45)$$

where we used (42). From this:

$$k_{\text{im}}^2 = k_0^2 \mu c^2 \left(-\frac{\varepsilon_{\text{re}}}{2} + \frac{1}{2} \sqrt{\varepsilon_{\text{re}}^2 + \varepsilon_{\text{im}}^2} \right), \quad (46)$$

as well as the real part:

$$k_{\text{re}}^2 = a + k_{\text{im}}^2 = \frac{k_0^2 \varepsilon_{\text{re}} \mu c^2}{2} + k_0^2 \frac{1}{2} \sqrt{(\varepsilon_{\text{re}} \mu c^2)^2 + (\varepsilon_{\text{im}} \mu c^2)^2}, \quad (47)$$

from this

$$k_{\text{re}}^2 = k_0^2 \mu c^2 \left(\frac{\varepsilon_{\text{re}}}{2} + \frac{1}{2} \sqrt{\varepsilon_{\text{re}}^2 + \varepsilon_{\text{im}}^2} \right). \quad (48)$$

One can see that if $\sigma = 0$, then $\varepsilon_{\text{im}} = 0$, $k_{\text{im}} = 0$ and

$$k = k_{\text{re}} = k_0 \sqrt{\varepsilon_{\text{r}} \mu_{\text{r}}}. \quad (49)$$

Comparing the above equations with (39) that expresses the refractive index, we obtain the well-known relationships:

$$\begin{aligned} \frac{k_{\text{re}}}{k_{\text{im}}} &= \frac{k_0 n}{k_0 \kappa} \quad \text{or by the other definition of } \kappa \text{ we get } \kappa = \frac{k_{\text{im}}}{k_{\text{re}}}. \end{aligned} \quad (50)$$

It is interesting to note that if we calculate the penetration depth by (46), then we obtain the same expression as in case of the “skin”-effect, which is observable at high-frequency fields propagating inside materials of finite conductivity:

$$\delta = \frac{1}{k_{\text{im}}} = \frac{1}{\omega \sqrt{\frac{\varepsilon_{\text{re}} \mu}{2} \left(\sqrt{1 + \left(\frac{\sigma}{\omega \varepsilon_{\text{re}}} \right)^2} - 1 \right)}}. \quad (51)$$

For the above-presented aluminum at 550 nm wavelength $\delta = 13$ nm.

2.5. Energy propagation in EM waves

The power density of an EM field is characterized by the so-called Poynting vector (energy traversing a unit surface placed perpedicularly to the direction of propagation in unit time):

$$\mathbf{S}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}, t) \times \mathbf{H}(\mathbf{r}, t), \quad (52)$$

where field attributes are real quantities. At high frequencies characteristic to optics it is not possible to detect instantaneous power, since the time constant of usual detectors are much larger than the period of vibration (T) for light. Therefore, we can only observe a time average:

$$\langle \mathbf{S}(\mathbf{r}, T) \rangle = \langle \mathbf{E}(\mathbf{r}, T) \times \mathbf{H}(\mathbf{r}, T) \rangle ; I = |\langle \mathbf{S}(\mathbf{r}, T) \rangle|. \quad (53)$$

The magnitude of this vector is called *intensity* (I). By using complex formalism:

$$\mathbf{E}(\mathbf{r}, t) = \text{Re}\{\tilde{\mathbf{E}}(\mathbf{r}) \cdot e^{i\omega t}\} = \frac{1}{2}(\tilde{\mathbf{E}}(\mathbf{r}) \cdot e^{i\omega t} + \tilde{\mathbf{E}}(\mathbf{r})^* \cdot e^{-i\omega t}), \quad (54)$$

similarly for \mathbf{H} . (The meaning of a complex amplitude will be interpreted later, now let us be content with knowing that with its help any radiation of linear, circular and elliptic polarization can be described.) Substituting \mathbf{E} and \mathbf{H} into (52) we can easily see that:

$$\mathbf{S}(\mathbf{r}, t) = \frac{1}{2}\text{Re}\{\tilde{\mathbf{E}}(\mathbf{r}) \times \tilde{\mathbf{H}}^*(\mathbf{r})\} + \frac{1}{2}\text{Re}\{\tilde{\mathbf{E}}(\mathbf{r}) \times \tilde{\mathbf{H}}(\mathbf{r}) \cdot e^{i2\omega t}\}. \quad (55)$$

Time-averaging cancels the second term:

$$\langle \mathbf{S}(\mathbf{r}, T) \rangle = \frac{1}{2}\text{Re}\{\tilde{\mathbf{E}}(\mathbf{r}) \times \tilde{\mathbf{H}}^*(\mathbf{r})\}. \quad (56)$$

For a plane wave propagating inside conductive media we get the following, by using (10) and the identity of $(\mathbf{a} \times (\mathbf{b} \times \mathbf{c})) \equiv (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}$:

$$\begin{aligned} \langle \mathbf{S}(\mathbf{r}, T) \rangle &= \frac{1}{2\mu\omega} \text{Re}\{\tilde{\mathbf{E}}(\mathbf{r}) \times (\tilde{\mathbf{k}}^* \times \tilde{\mathbf{E}}^*(\mathbf{r}))\} = \\ &= \frac{1}{2\mu\omega} |\tilde{\mathbf{E}}(\mathbf{r})|^2 \mathbf{k}_{\text{re}} - \frac{1}{2\mu\omega} \text{Re}\{(\tilde{\mathbf{E}}(\mathbf{r}) \cdot \tilde{\mathbf{k}}^*) \cdot \tilde{\mathbf{E}}^*(\mathbf{r})\}. \end{aligned} \quad (57)$$

Due to relationship (36) that originates from the 4th Maxwell equation, the second term is always zero for a plane wave inside isotropic media. Substituting \mathbf{E} with (32) describing plane waves, we get the following for the time-averaged Poynting vector:

$$\langle \mathbf{S}(\mathbf{r}, T) \rangle = \frac{1}{\mu} \frac{\mathbf{k}_{\text{re}}}{\omega} \frac{|\mathbf{E}_0|^2}{2} \cdot e^{-2\mathbf{k}_{\text{im}}\mathbf{r}}. \quad (58)$$

Here we suppose that μ is real (in optics this is a very good approximation). The above expression is also known as the *Lambert-Beer law*, which states that the power density and amplitude of a plane wave decreases exponentially in homogeneous absorbing media (e.g. in case of a normal angle of incidence).

2.6. What career opportunities do physicists have with a degree in optics?

Industry sectors	Application	Product, research	Hungarian connection
IT science	data transfer	networks	Furukawa
		inter chip communication	SZTAKI
	displaying	3D display	Holografika Kft
	data processing	acoustooptics	BME AFT
		neural computers	SZTAKI
Energy, heavy industry	material processing	high-power laser systems	Lasram
	nuclear fusion	simulation of fusion reactors	BME NTI
		measurement of fusion reactors	Wigner FK
Medical science	laser surgery	medical lasers	Lasram
	urine analysis, blood sugar measurement	automatic microscope, measurement of physical parameters, optical readout of chemical reagent strips	77Elektronika
	medical imaging	Positron Emission Tomography (PET)	Mediso
		confocal 3D microscope with two-photon excitation	Femtonics
	color vision correction	thin layers	BME MOGI
	water quality analysis	3D digital holography	SZTAKI
	blood analysis	diagnostic tools	Diatron
Food industry	sensors	NIR spectrometer	Siemens-BME
Semiconductor industry	semiconductor diagnostics	laser sources and measuring equipment	Semilab
Illumination technology	illumination	high-efficiency outdoor LED illumination systems	Optimal Optik
	displays	dashboards	Bosch
	illumination	illumination systems	OMI Optika
Basic research	material science, biology	femto- and attosecond laser pulses	ELI
Military industry	...		

Tab. 2 Career opportunities for physicists dealing with optics in Hungary.

3. BEHAVIOR OF PLANE WAVES AT PLANAR INTERFACES

3.1. Polarization eigenstates at plane wave – planar surface interactions

It is known from wave theory that abruptly changing the propagation speed at the boundary of two media results in the phenomena of reflection and refraction. The appearance of reflection is explicitly due to the abrupt change of refractive index, when the wave has to fulfill two conditions simultaneously at the surface: that of energy conservation and the continuous transmission of (the appropriate component of) the amplitude. In cases when the spatial index change is slow, there is no amplitude condition, and the reflected component does not occur. Below we examine the case, when optical material characteristics change abruptly (i.e. within a region much smaller than the wavelength) in a direction perpendicular to a plane surface.

When discussing the reflection/transmission of light, we cannot ignore the transverse nature of EM waves, i.e. polarization. We talk about a polarized radiation if the *direction* of \mathbf{E} and \mathbf{B} field vector oscillation shows a long-term temporal and spatial regularity (periodicity). The most well-known kind of regularity is linear polarization, when the oscillation of field vectors always remains in the same plane (the term “polarized” refers to two fixed opposing poles). A completely monochromatic plane wave is always polarized, in the general case elliptically (see Chapter 13). In order to produce *unpolarized* light we need to mix a large number of plane waves with amplitude vectors of different directions and frequency. Due to the latter condition the wavelength spectrum of the radiation broadens (it will not be monochromatic anymore). If not the frequency, but the propagation direction of the components is different, then the beam becomes divergent in the far field (it is not a plane wave anymore). A further condition is that the initial phase of the constituting plane wave components must be random (otherwise we get a pulse-like temporal behavior). If the spatial and temporal randomness take place simultaneously, we obtain the diffuse and unpolarized radiation that surrounds us. (These will be discussed at spatial and temporal coherence.)

In the following we will send an ideal (infinite) plane wave upon a plane surface, and examine the rules of reflection/refraction of light. During the derivation we assume non-absorbing dielectric media (i.e. n and k are real). Since the plane wave is an exact solution to the Maxwell equations, our results are not only valid in the realm of geometrical optics (see Chapter 4), but are general rules in wave theory. [1]

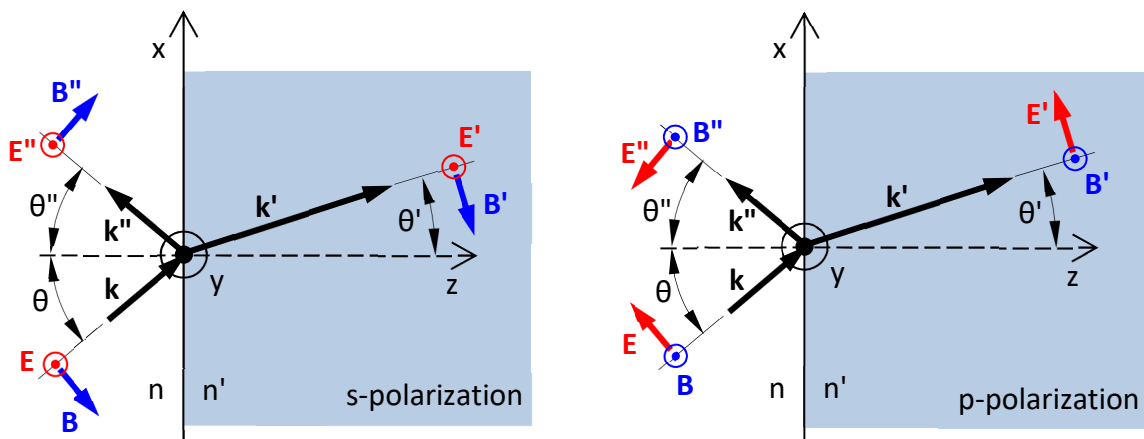


Fig. 8 Interpretation of eigenstates s- and p-, and presentation of the applied notation. The incident wave vector and the surface normal define the *plane of incidence*. The plane of reflection and refraction are similarly specified.

The derivation is partly based on the important observation that surface reflection/transmission features polarization eigenstates (see Chapter 13). For isotropic materials these are linearly polarized plane waves: the s-polarization (σ , or TE polarization) when the electric field is normal to the plane of incidence, and the p-polarization (π , or TM polarization), when the electric field is parallel with the plane of incidence. Since these are eigenstates, they *do not change* during reflection and refraction, it is only the complex amplitude of the oscillation that becomes altered (in terms of magnitude and phase).

The incident, reflected and refracted plane waves as well as their wave vectors:

$$\tilde{\mathbf{E}}(\mathbf{r}, t) = \tilde{\mathbf{E}}_0 \cdot e^{i(\omega t - \mathbf{k} \cdot \mathbf{r})} \quad (59)$$

$$\mathbf{k} = n \cdot \mathbf{k}_0 \Rightarrow \mathbf{k} = (k_x, 0, k_z) = n \cdot k_0 (\sin \theta, 0, \cos \theta) \quad (60)$$

$$\begin{aligned} \mathbf{k}'' &= (k_x'', k_y'', k_z'') ; k_x''^2 + k_y''^2 + k_z''^2 = n^2 k_0^2 \\ \mathbf{k}' &= (k_x', k_y', k_z') ; k_x'^2 + k_y'^2 + k_z'^2 = n'^2 k_0^2 \end{aligned} \quad (61)$$

The complex value of the vector amplitude ($\tilde{\mathbf{E}}_0$) expresses that the initial phase of its components can be theoretically arbitrary, by which we can generally describe (linear, circular, elliptic) polarization (this will be discussed in more detail in Chapter 13), and that the relative phase of the incident, reflected and transmitted waves can be different too (e.g. in absorbing media).

3.2. Continuity conditions at the boundary of dielectric materials

According to the continuity conditions of Maxwell's equations:

$$\begin{aligned} \mathbf{E}'_t &= \mathbf{E}_t + \mathbf{E}''_t & \rightarrow & \frac{\mathbf{B}'_t}{\mu'} = \frac{\mathbf{B}_t}{\mu} + \frac{\mathbf{B}''_t}{\mu} & \rightarrow & \mathbf{E}'_t = \mathbf{E}_t + \mathbf{E}''_t \\ \mathbf{H}'_t &= \mathbf{H}_t + \mathbf{H}''_t & & & & \mathbf{B}'_t = \mathbf{B}_t + \mathbf{B}''_t, \end{aligned} \quad (62)$$

where we assumed that the media are non-magnetic, i.e. $\mu = \mu' = \mu_0$. (Instead of the two amplitude conditions we could also use any one of them together with the law of energy conservation. In our derivation this latter will be verified based on the resulting formulae.) We need the above relationships when the material properties change over a domain that is much smaller than the wavelength (i.e. abruptly). Substituting (59) into the condition for \mathbf{E} :

$$\tilde{\mathbf{E}}'_{t0} \cdot e^{i(\omega t - \mathbf{k}' \cdot \mathbf{r})} = \tilde{\mathbf{E}}_{t0} \cdot e^{i(\omega t - \mathbf{k} \cdot \mathbf{r})} + \tilde{\mathbf{E}}''_{t0} \cdot e^{i(\omega t - \mathbf{k}'' \cdot \mathbf{r})}. \quad (63)$$

All three waves exist simultaneously along the interface, this is why we make our investigations here. Hence, $z = 0$, $k_y = 0$, and by dividing out $e^{i\omega t}$:

$$\tilde{\mathbf{E}}'_{t0} \cdot e^{-i(k'_x x + k'_y y)} = \tilde{\mathbf{E}}_{t0} \cdot e^{-i(k_x x)} + \tilde{\mathbf{E}}''_{t0} \cdot e^{-i(k''_x x + k''_y y)}. \quad (64)$$

Similarly for \mathbf{B}_t . (64) only satisfies for any arbitrary (x, y) values if the exponents are identical:

$$k'_x x + k'_y y = k_x x = k''_x x + k''_y y. \quad (65)$$

This is the so-called *phase matching*, meaning that the phase of all the incident, reflected and refracted waves change by the same rate along the interface (i.e. the wavelengths and wave vectors projected onto the x-axis are the same in all the three cases):

$$\left. \begin{array}{l} \text{I} \quad k'_y = k''_y = 0 \\ \text{II} \quad k_x = k''_x \\ \text{III} \quad k_x = k'_x \end{array} \right\} \quad (66)$$

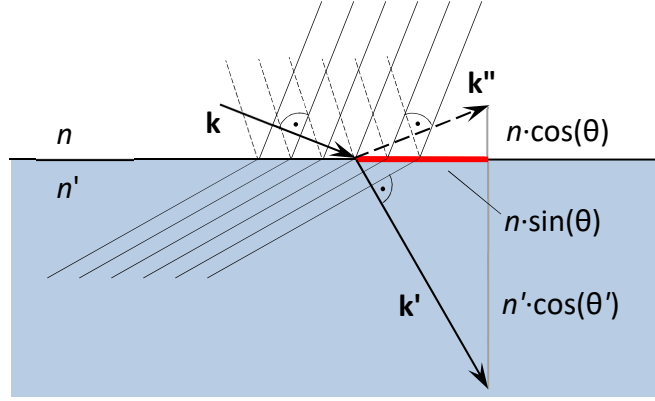


Fig. 9 Demonstration of phase matching.

(I) suggests that the plane of refraction and reflection coincide with the plane of incidence.

(II) results in the law of reflection:

$$\sin \theta = \sin \theta'' \Rightarrow \theta = \theta'' \quad (67)$$

(III) results in the law of refraction (Snell's law):

$$n \cdot k_0 \sin \theta = n' \cdot k_0 \sin \theta' \Rightarrow n \cdot \sin \theta = n' \cdot \sin \theta', \quad (68)$$

where θ and $\theta'' \in [0; \pi/2]$. If the phases are equal, we can divide them out in (64). Thus we arrive at the condition pertaining to the amplitudes:

$$\begin{aligned} \tilde{\mathbf{E}}'_{t0} &= \tilde{\mathbf{E}}_{t0} + \tilde{\mathbf{E}}''_{t0} \\ \tilde{\mathbf{B}}'_{t0} &= \tilde{\mathbf{B}}_{t0} + \tilde{\mathbf{B}}''_{t0}. \end{aligned} \quad (69)$$

3.3. S-polarization, when \mathbf{E} is normal to the plane of incidence

$\mathbf{E}_t = \mathbf{E}_x + \mathbf{E}_y$; $E_x = 0$ we have only $E_y \rightarrow E_y = E$, i.e. according to (69):

$$\tilde{E}'_0 = \tilde{E}_0 + \tilde{E}''_0. \quad (70)$$

From the 2nd Maxwell equation we determine the value of \mathbf{B} (for the incident, refracted and reflected waves alike):

$$\text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \rightarrow -i\mathbf{k} \times \tilde{\mathbf{E}}_0 = -i\omega \tilde{\mathbf{B}}_0 \rightarrow \tilde{\mathbf{B}}_0 = \frac{\mathbf{k} \times \tilde{\mathbf{E}}_0}{\omega} \quad (71)$$

$$\begin{vmatrix} \hat{x} & \hat{y} & \hat{z} \\ k_x & 0 & k_z \\ 0 & \tilde{E}_0 & 0 \end{vmatrix} = -k_z \tilde{E}_0 \hat{x} - 0 \cdot \hat{y} + k_x \tilde{E}_0 \hat{z} \rightarrow \tilde{\mathbf{B}}_0 = \frac{-k_z \tilde{E}_0 \hat{x} + k_x \tilde{E}_0 \hat{z}}{\omega}. \quad (72)$$

From (69) the condition pertaining to \mathbf{B} ($\mathbf{B}_t = \mathbf{B}_x + \mathbf{B}_y$; $B_y = 0$):

$$\tilde{B}'_{0x} = \tilde{B}_{0x} + \tilde{B}''_{0x} \quad (73)$$

By substituting the x-component of (72) into the above, and dividing out ω :

$$-k'_z \tilde{E}'_0 = -k_z \tilde{E}_0 - k''_z \tilde{E}''_0. \quad (74)$$

According to the law of reflection $k''_z = -k_z$:

$$-k'_z \tilde{E}'_0 = -k_z \tilde{E}_0 + k_z \tilde{E}''_0 \quad (75)$$

Dividing both sides by $-k_z$:

$$\frac{k'_z}{k_z} \tilde{E}'_0 = \tilde{E}_0 - \tilde{E}''_0. \quad (76)$$

(70) and (76) form an equation system with two variables, which can be solved for \tilde{E}'_0 and \tilde{E}''_0 :

$$\tilde{E}'_0 = \frac{2}{1 + \frac{k'_z}{k_z}} \tilde{E}_0 \quad ; \quad \tilde{E}''_0 = \frac{2}{1 + \frac{k'_z}{k_z}} \tilde{E}_0 - \tilde{E}_0 \quad (77)$$

We introduce the auxiliary variable a :

$$a \triangleq \frac{k'_z}{k_z}, \quad (78)$$

by which relationships (77) can be written in a more concise form:

$$\tilde{E}'_0 = \frac{2}{1 + a} \tilde{E}_0 \quad ; \quad \tilde{E}''_0 = \frac{1 - a}{1 + a} \tilde{E}_0. \quad (79)$$

By this the coefficients of transmission (τ_s) and reflection (ρ_s) defined for s-polarization are:

$$\tau_s \triangleq \frac{\tilde{E}'_0}{\tilde{E}_0} = \frac{2}{1 + a} \quad ; \quad \rho_s \triangleq \frac{\tilde{E}''_0}{\tilde{E}_0} = \frac{1 - a}{1 + a} \quad (80)$$

$$a = \frac{k'_z}{k_z} = \frac{n' \cos(\theta')}{n \cos(\theta)} \quad (81)$$

$$\tau_s = \frac{2n \cos(\theta)}{n \cos(\theta) + n' \cos(\theta')} \quad ; \quad \rho_s = \frac{n \cos(\theta) - n' \cos(\theta')}{n \cos(\theta) + n' \cos(\theta')}. \quad (82)$$

It is important to note that the values of transmission and reflection coefficients are real numbers at the interface of dielectrics, thus there is no constant phase shift between the incident, reflected and transmitted waves.

3.4. P-polarization, when E is parallel with the plane of incidence

$\mathbf{B}_t = \mathbf{B}_x + \mathbf{B}_y$; $B_x = 0$ we only have $B_y \rightarrow B_y = B$, i.e. according to (69):

$$\tilde{B}'_0 = \tilde{B}_0 + \tilde{B}''_0, \quad (83)$$

and $E_y = 0$. From these the coefficients of transmission (τ_p) and reflection (ρ_p) can be derived for p-polarization in a way similar to the above:

$$\tau_p \triangleq \frac{\tilde{E}'_0}{\tilde{E}_0} = \frac{2}{1 + b} \frac{n}{n'} \quad ; \quad \rho_p \triangleq \frac{\tilde{E}''_0}{\tilde{E}_0} = \frac{1 - b}{1 + b} \quad (84)$$

$$b \triangleq \left(\frac{n}{n'}\right)^2 \frac{k'_z}{k_z} = a \cdot \left(\frac{n}{n'}\right)^2 = \frac{n \cos(\theta')}{n' \cos(\theta)} \quad (85)$$

$$\tau_p = \frac{2n \cos(\theta)}{n' \cos(\theta) + n \cos(\theta')} \quad ; \quad \rho_p = \frac{n' \cos(\theta) - n \cos(\theta')}{n' \cos(\theta) + n \cos(\theta')}. \quad (86)$$

Relationships (82) and (86) are called the *Fresnel formulae* or *Fresnel equations*.

3.5. Discussion of Fresnel formulae

In the below figure we present the angle dependence of the reflection and transmission coefficients for an air-glass interface ($n = 1$ and $n' = 1.5$). First thing to note is that in case of external reflection ($n < n'$) the electromagnetic field is reflected in *opposite phase* relative to the incident radiation (see the sign on ρ_s). The second is that by increasing the incident angle, the magnitude of the s-reflection coefficient grows and tends to unity. The third is that at a given angle $\rho_p = 0$, i.e. the reflected wave is completely polarized, it contains *only an s-component*. This is the so-called *Brewster's effect*, to be discussed more extensively later. The fourth is that in case of internal reflection ($n > n'$) the reflection does not peak (reaches unity) at 90° , but much sooner. This is called *total internal reflection* (TIR), to be discussed later in detail too.

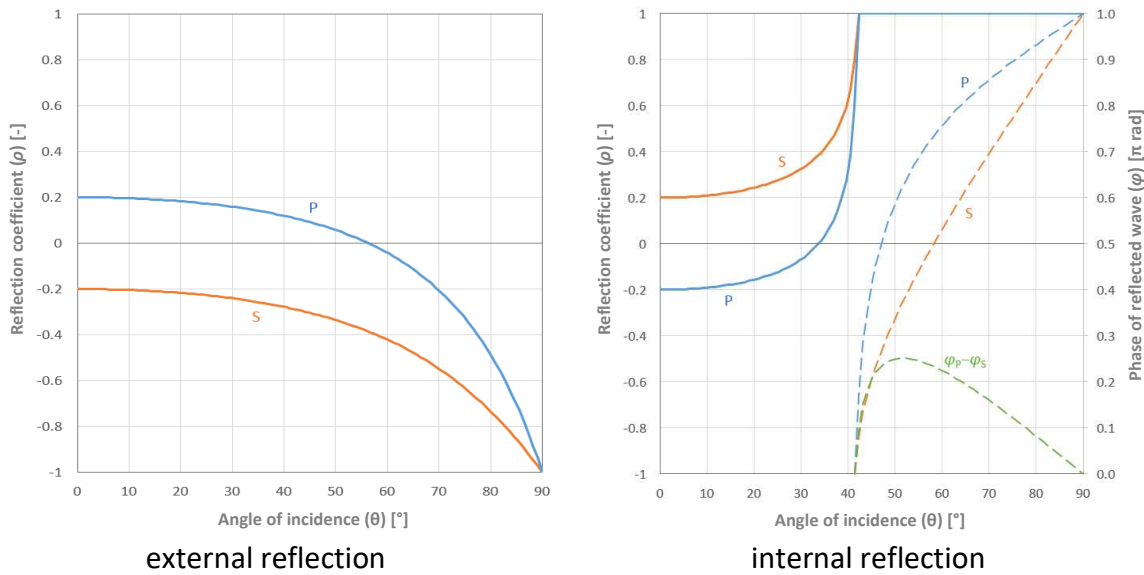


Fig. 10 Reflection coefficients of external and internal reflection as a function of angle of incidence for a glass (refractive index: 1.5) / air (refr. ind.: 1.0) interface.

In Chapter 2. we introduced the complex wave vector and refractive index that are appropriate for the description of absorbing media. For angles of incidence other than the perpendicular, the real and imaginary components of the wave vector do not point in the same direction, therefore relationship $\tilde{\mathbf{k}} = \tilde{n} \cdot \mathbf{k}_0$ is not valid either. From this it follows that (60), which we used as a starting point at the derivation of the Fresnel formulae, is not valid for any angle of incidence in case of absorbing media other than $\theta \approx 0^\circ$! Irrespective of this, the reflection and transmission coefficients become complex quantities anyway, i.e. after reflecting from a metal surface *the phase of oscillation changes*. A further consequence for linearly polarized waves is that while after reflection from a dielectric surface the non-s- or p-polarized incident radiation always remains linearly polarized, (only the plane of oscillation gets rotated), in case of metals the reflected radiation becomes elliptically polarized due to the phase difference between the s- and p-components. The general derivation of formulae applicable to an arbitrary angle can be found in references [3] pp. 611-664 and [5], as well as in the form of a brief summary at the end of this chapter.

3.6. Fresnel formulae describing power ratios

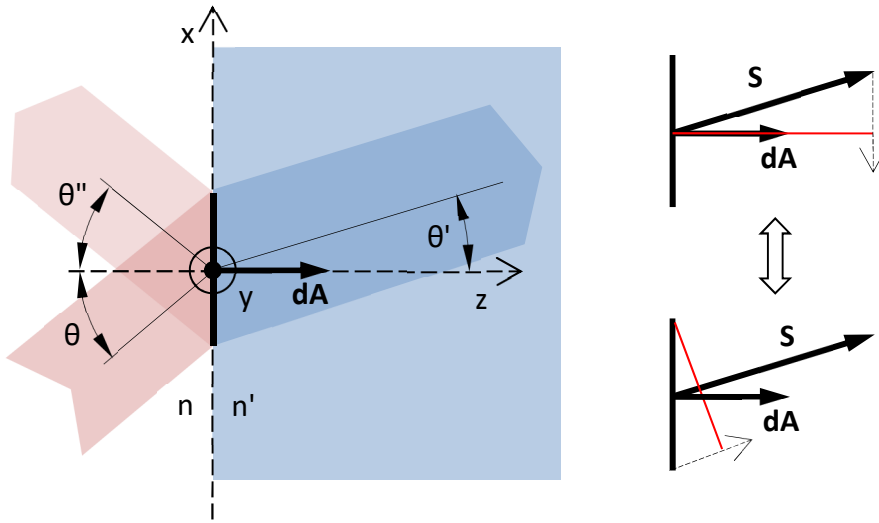


Fig. 11 Summary of notations used for describing the power conditions of a plane wave transmitting/reflecting on an elementary surface.

According to Chapter 2, the detectable time-averaged power density vector (for dielectric media with a real wave vector):

$$\langle \mathbf{S} \rangle = \frac{1}{\mu} \frac{k}{\omega} \frac{|\tilde{\mathbf{E}}_0|^2}{2}. \quad (87)$$

In the following we will make our investigation in the point $x = y = z = 0$, and still assume that the medium is non-magnetic: $\mu = \mu' = \mu_0$. Then the average power of the incident beam crossing the dA surface unit perpendicularly (which is exactly the same as if we took the projection of surface element dA perpendicularly to the incident beam):

$$dP = \langle \mathbf{S} \rangle \cdot d\mathbf{A} = \frac{1}{\mu_0} \frac{k}{\omega} \frac{|\tilde{\mathbf{E}}_0|^2}{2} \cos(\theta) dA. \quad (88)$$

Similarly for the refracted beam:

$$dP' = \langle \mathbf{S}' \rangle \cdot d\mathbf{A} = \frac{1}{\mu_0} \frac{k'}{\omega} \frac{|\tilde{\mathbf{E}}_0'|^2}{2} \cos(\theta') dA, \quad (89)$$

and the reflected beam:

$$dP'' = \langle \mathbf{S}'' \rangle \cdot d\mathbf{A} = \frac{1}{\mu_0} \frac{k}{\omega} \frac{|\tilde{\mathbf{E}}_0''|^2}{2} \cos(\theta'') dA. \quad (90)$$

The ratio of light power components arriving at and reflecting back from the surface element perpendicularly to it is called *reflectance* (R):

$$R \triangleq \frac{dP''}{dP} = \frac{|\tilde{\mathbf{E}}_0''|^2}{|\tilde{\mathbf{E}}_0|^2} = |\rho|^2. \quad (91)$$

The ratio of light power components arriving at and passing through (refracting at) the surface element perpendicularly to it is the *transmittance* (T):

$$T \triangleq \frac{dP'}{dP} = \frac{|\langle \mathbf{S}' \rangle| \cos \theta'}{|\langle \mathbf{S} \rangle| \cos \theta} = \frac{I' \cos \theta'}{I \cos \theta} = \frac{k' |\tilde{\mathbf{E}}_0'|^2 \cos \theta'}{k |\tilde{\mathbf{E}}_0|^2 \cos \theta} = |\tau|^2 \frac{n' \cos \theta'}{n \cos \theta}. \quad (92)$$

It is worth noting that the cross section of the light beam *changes* after refraction, which is expressed by the cosine factor in formula (92). For this reason, the intensity of the light beam (I' that is $\langle S' \rangle$) alters after refraction even if the transmittance is 100% (e.g. when using an anti-reflection coating): if $n < n'$ it decreases, otherwise it grows. The effect can be used among other things e.g. to expand a light beam along one direction (see the anamorphic beam shaping of laser diodes).

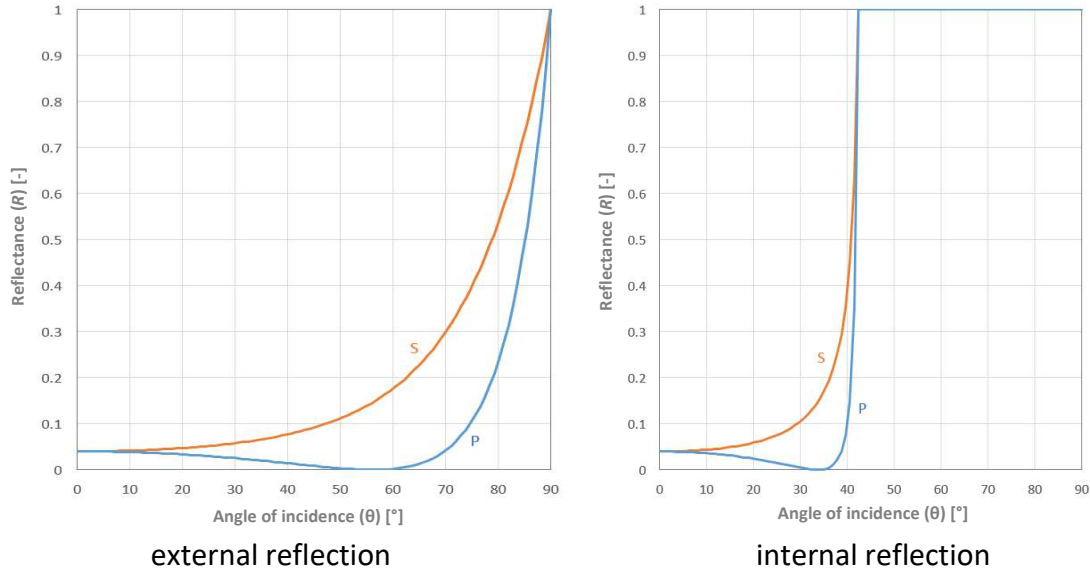


Fig. 12 Reflectances of external and internal reflection as a function of angle of incidence for a glass (refractive index: 1.5) / air (refractive index: 1.0) interface.

From the above it is easy to see, that the law of energy conservation is satisfied for T and R :

$$T + R = 1. \quad (93)$$

In case of normal incidence:

$$T = \left| \frac{2n}{n + n'} \right|^2 \frac{n'}{n} \quad \text{and} \quad R = \left| \frac{n - n'}{n + n'} \right|^2. \quad (94)$$

If the first medium is air ($n = 1$), and the second is ordinary borosilicate glass ($n' = 1.5$), then $\rho = -0.2$ and $R = 0.04$ (i.e. 4%).

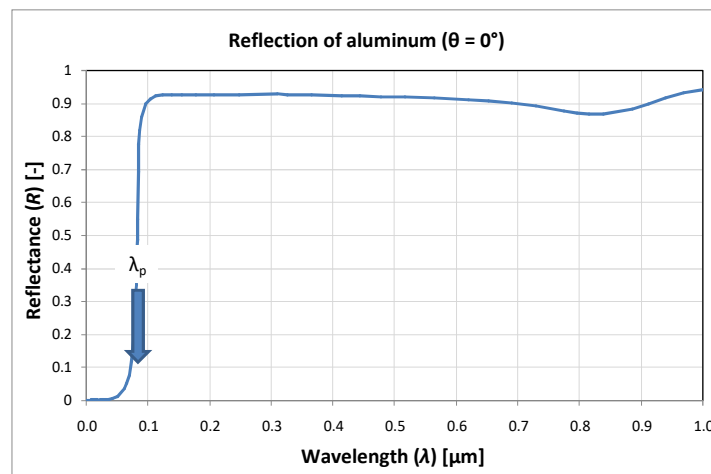


Fig. 13 Reflectance of aluminum as calculated from its complex refractive index in case of normal incidence, see (94).

3.7. Brewster's effect

The refraction and reflection of light can be interpreted on microscopic scale as if the incident light were absorbed by electric dipoles representing atoms of the second medium (which thus become oscillating), then radiated back. A radiating dipole cannot emit energy in the direction of the axis of oscillation. If the reflected radiation travels perpendicularly to the direction of the refracted beam ($\theta + \theta' = 90^\circ$), and the incident radiation is of p-polarization (i.e. it has no component normal to the plane of incidence), then dipoles in the second medium oscillate right along a direction normal to the reflected radiation. From this it follows, that a reflected p-polarized radiation transports zero energy, i.e. $R_p = 0$. This phenomenon is called the Brewster's effect after its discoverer, its characteristic angle is denoted by θ_B . From the $\rho_p = 0$ condition and using the law of refraction one can easily derive (see practice):

$$\tan \theta_B = \frac{n'}{n} \quad (95)$$

For the air-glass interface we discussed before: $\theta_B = 56.3^\circ$. If unpolarized radiation arrives at a surface at Brewster's angle, then the reflected radiation becomes linearly polarized, since it contains only s-polarized component.

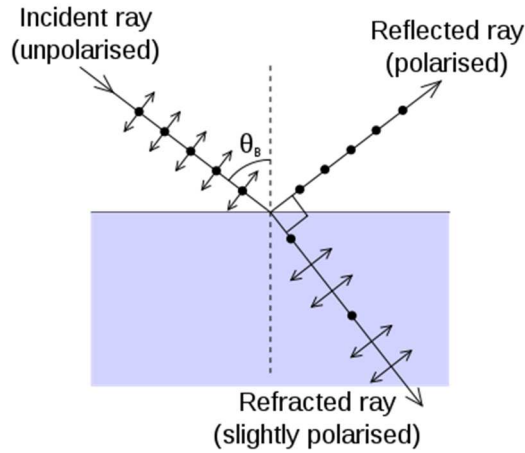


Fig. 14 Demonstration of Brewster's effect. (source: wikipedia)

3.8. Total internal reflection (TIR): $n' < n$

In case of $n' < n$, the auxiliary variable a (see derivation for s-polarization) used in Fresnel formulae (82) and (86) takes a complex value above a certain angle of incidence. Its mathematical explanation lies in the definition of a , to make which clear we rearrange the formula:

$$a = \frac{k'_z}{k_z} = \frac{k_0 \cdot n' \cos \theta'}{k_0 \cdot n \cos \theta} = \frac{n' \sqrt{1 - \sin^2 \theta'}}{n \cos \theta} = \frac{\sqrt{n'^2 - n^2 \sin^2 \theta}}{n \cos \theta}. \quad (96)$$

If the angle of incidence (θ) exceeds a critical value ($\theta = \theta_c$ when $k'_z = 0$), then we get a negative number under the square root sign in (96). The critical angle can be easily determined by checking when the expression under the root equals zero:

$$n' = n \cdot \sin \theta_c \rightarrow \theta_c = \arcsin \frac{n'}{n} \quad (97)$$

Now we examine the cases when $\theta_c < \theta$, in order to understand what it means physically for a to have a complex value. Reformulating equation (96):

$$a = \frac{\sqrt{-1 \cdot (n^2 \sin^2 \theta - n'^2)}}{n \cos \theta} = \pm i \frac{\sqrt{n^2 \sin^2 \theta - n'^2}}{n \cos \theta} \equiv -i\gamma. \quad (98)$$

Here we need to choose the negative sign, since only this has a physical meaning, as further discussion will explain. Then we get the following for the wave vector in the second medium (see the definition of $\tilde{\mathbf{k}}$ (32) in Chapter 2):

$$k'_z = -ik_z\gamma = -ik'_{\text{im},z} \Rightarrow k'_{\text{im},z} = k_z\gamma. \quad (99)$$

The above relationship can be interpreted according to our studies of the complex wave vector as a strong wave attenuation along the z axis in the medium of smaller refractive index (n'). Owing to phase matching it still holds that $k'_x = k_x$, i.e. the real component of the wave vector points in the x -direction, parallel to the surface (in other words $\mathbf{k}'_{\text{re}} \perp \mathbf{k}'_{\text{im}}$):

$$\begin{aligned} \mathbf{E}'(x, z, t) &= \mathbf{E}'_0 e^{i(\omega t - \mathbf{k}' \cdot \mathbf{r})} = \mathbf{E}'_0 e^{i(\omega t - k'_x x - k'_z z)} = \\ &= \mathbf{E}'_0 e^{i(\omega t - k'_x x + i\gamma k_z z)} = \mathbf{E}'_0 e^{i(\omega t - k'_x x)} e^{-\gamma k_z z}. \end{aligned} \quad (100)$$

Such fields are called as *evanescent waves*. According to the definition of penetration depth (δ) in Chapter 2:

$$\delta \triangleq \frac{1}{k'_{\text{im},z}} = \frac{1}{k_z\gamma} = \frac{1}{k_z} \frac{n \cdot \cos \theta}{\sqrt{n^2 \cdot \sin^2 \theta - n'^2}}. \quad (101)$$

At a glass/air interface ($n = 1.5$; $n' = 1.0$) the critical angle is $\theta_c = 41.8^\circ$. For $\theta = 45^\circ$ incidence:

$$\delta = \frac{3}{k_z} = \lambda' \cdot 0.48 = 263 \text{ nm}, \quad (102)$$

where we assumed an average (green) wavelength of $\lambda = 550 \text{ nm}$.

It is worthwhile to note that at the critical angle of total internal reflection $k'_x = k'$. Since phase matching always guarantees that $k_x = k'_x$, for angles larger than the critical $k'_x > k'$, i.e. the wavelength (x -direction periodicity) is *smaller* in the second medium than that of an ideal plane wave freely propagating in it!

Finally, let us examine what is happening with the phase. For total reflection the values of a and b introduced at Fresnel formulae become complex, so are the reflection coefficients:

$$\rho_s \triangleq \frac{\tilde{E}''_0}{\tilde{E}_0} = \frac{1-a}{1+a} = \frac{1+i\gamma}{1-i\gamma} = \frac{A \cdot e^{i\varphi_s/2}}{A \cdot e^{-i\varphi_s/2}} = e^{i\varphi_s/2} \quad (103)$$

$$\rho_p \triangleq \frac{\tilde{E}''_0}{\tilde{E}_0} = \frac{1-b}{1+b} = \frac{1+i\gamma \left(\frac{n}{n'}\right)^2}{1-i\gamma \left(\frac{n}{n'}\right)^2} = \frac{B \cdot e^{i\varphi_p/2}}{B \cdot e^{-i\varphi_p/2}} = e^{i\varphi_p/2}. \quad (104)$$

This means that the reflected light undergoes a *phase shift*. Since generally $\varphi_s \neq \varphi_p$ (but depends on the incident angle), a linearly polarized plane wave arriving at the surface not in s- or p-polarization becomes elliptically polarized after reflection. From relationships (104) we can also see that the magnitude of reflection coefficients is $R_{s,p} = 1$, i.e. all incident energy is reflected. This is why the effect is called *total internal reflection* (TIR). By drafting a phasor diagram in the complex plane, from (104) one can easily see that:

$$\varphi_s = 2 \arctan \gamma \quad (105)$$

$$\varphi_p = 2 \arctan \left(\gamma \left(\frac{n}{n'} \right)^2 \right). \quad (106)$$

The phenomenon of total reflection is used beside others e.g. in beam-steering prisms (see the eyepieces of SLR i.e. single-lens-reflex cameras, image rotators of binocular telescopes), as well as light-guiding and wave-guiding optical fibers. Utilizing the different phase shift of eigenstates a special prism can be built, the so-called Fresnel rhomb, that behaves as a wavelength-independent $\lambda/4$ phase retarder (transforms linearly polarized light into circularly polarized).

3.9. Plane wave refraction at planar dielectric/metal interface (supplementary)

Let us have a dielectric of refractive index n as the incident medium, and be the refraction media a highly absorbing material (metal). As per equation (40) from Chapter 2:

$$\tilde{\mathbf{k}}'^2 = k_0 \tilde{n}'^2. \quad (107)$$

Let us introduce the N' and K' virtual refractive index and extinction coefficient of the metal:

$$\tilde{\mathbf{k}}' \triangleq k_0 (N' \hat{\mathbf{e}} - iK' \hat{\mathbf{f}}), \quad (108)$$

where the wavefront normal of the evanescent wave points in the direction of the $\hat{\mathbf{e}}$ unit vector, and the imaginary wave vector to the direction of the $\hat{\mathbf{f}}$ unit vector (the field attenuates in this direction). Separating the real and imaginary parts of (107) we can easily get:

$$\left. \begin{aligned} N'^2 - K'^2 &= n'^2 - \kappa'^2 \\ K' N' \hat{\mathbf{e}} \cdot \hat{\mathbf{f}} &= n' \kappa' \end{aligned} \right\} \quad (109)$$

where we introduce the following notation:

$$\cos \alpha \equiv \hat{\mathbf{e}} \cdot \hat{\mathbf{f}}. \quad (110)$$

From (109) the value of the virtual refractive index of (108) can be easily derived:

$$N'^2 = \frac{1}{2} \left(n'^2 - \kappa'^2 + \sqrt{(n'^2 - \kappa'^2)^2 + 4 \left(\frac{n' \kappa'}{\cos \alpha} \right)^2} \right). \quad (111)$$

Since phase matching still holds here:

$$\tilde{k}'_x = k_x \quad \Leftrightarrow \quad k_{\text{re},x} - i k_{\text{im},x} = k_x, \quad (112)$$

it follows that $k_{\text{im},x} = 0$, i.e. $\hat{\mathbf{f}}$ is always normal to the surface, and thus $\alpha = \theta'$. If the incident medium is a dielectric, the law of refraction can be written as:

$$N' \sin \theta' = n \sin \theta \quad (113)$$

$$\sqrt{\frac{1}{2} \left(n'^2 - \kappa'^2 + \sqrt{(n'^2 - \kappa'^2)^2 + 4 \left(\frac{n' \kappa'}{\cos \theta'} \right)^2} \right)} \sin \theta' = n \sin \theta. \quad (114)$$

If $\kappa' \gg n'$, which is usually true for metals, then we can apply the following approximation:

$$\sqrt{(n'^2 - \kappa'^2)^2 + 4 \left(\frac{n' \kappa'}{\cos \theta'} \right)^2} = |n'^2 - \kappa'^2| \sqrt{1 + 4 \left(\frac{n' \kappa'}{(n'^2 - \kappa'^2) \cos \theta'} \right)^2} \approx$$

$$\begin{aligned}
&\approx |n'^2 - \kappa'^2| \left(1 + 2 \left(\frac{n' \kappa'}{(n'^2 - \kappa'^2) \cos \theta'} \right)^2 \right) = \kappa'^2 - n'^2 + \frac{2n'^2 \kappa'^2}{(\kappa'^2 - n'^2) \cos^2 \theta'} \approx \\
&\approx \kappa'^2 - n'^2 + \frac{2n'^2}{\cos^2 \theta'}
\end{aligned} \tag{115}$$

Hence, the (114) refraction law becomes this really simple formula:

$$\frac{n'}{\cos \theta'} \sin \theta' \approx n \sin \theta \Rightarrow \operatorname{tg} \theta' \approx \frac{n \sin \theta}{n'}, \tag{116}$$

which approximates the actual refractive angle better than 1% in case of aluminum ($n' = 1.2$; $\kappa' = 7.0$) for the entire $0..90^\circ$ range of angle of incidence! Since the value of a tangent is always higher than that of the sine, (116) implies that the angle of refraction is always smaller in metals than dielectrics.

4. THEORETICAL BACKGROUNDS OF GEOMETRICAL OPTICS

Sources: [1], [3], [6]

Geometrical optics: an approximation, by which the propagation of light and its transmission through boundaries of different media is described with the help of geometrical entities – viz. curves. These curves are called *light rays*.

4.1. Approximations used

- direction dependence** – we consider only isotropic media, e.g. glass (in this case the 1) and 2) ray definitions yields the same, i.e. a light ray is $\parallel \mathbf{k} \parallel \mathbf{S}$, see below)
- position dependence** – we consider homogeneous and inhomogeneous media (light rays can be curved and not only straight)
- field dependence** – we consider only linear media (light rays can intersect each other without any interaction; in every point of space ω is identical with that of the excitation)
- wavelength dependence** – monochromatic case (there is only one wavelength in the spectrum, temporally coherent case) or polychromatic case (temporally incoherent light – can be decomposed into wavelength components)
- spatial coherence** – spatially coherent case, e.g. point source, plane wave (wavefronts exist – star light, laser) or spatially incoherent case (diffuse light – e.g. can be decomposed into statistically independent point sources)

4.2. Basic considerations

Maxwell's equations \rightarrow vectorial wave equation:

$$\nabla^2 \mathbf{E}(\mathbf{r}, t) - \varepsilon\mu \frac{\partial^2 \mathbf{E}(\mathbf{r}, t)}{\partial t^2} = 0 ; n^2 = \varepsilon_r \mu_r \text{ (Maxwell's formula)} \quad (117)$$

If we examine a monochromatic wave of ω angular frequency, the general form of the solution in a linear medium is:

$$\tilde{\mathbf{E}}(\mathbf{r}, t) = \tilde{\mathbf{E}}(\mathbf{r}) \cdot e^{-i\omega t}. \quad (118)$$

Since the excitation is temporally harmonic, we use the complex formalism to describe the EM field (i.e. \mathbf{E} is complex, which we denote by a tilde). Note, that for the sake of simplicity, from now on we will use the method of positive phase propagation, i.e. the phase is advancing in the direction of propagation, and is retarding in terms of time ($\tilde{\mathbf{E}} \rightarrow \tilde{\mathbf{E}}^*$). By this ansatz the equation takes the form of the time-independent, so-called Helmholtz equation:

$$\nabla^2 \tilde{\mathbf{E}}(\mathbf{r}) = -k^2 \tilde{\mathbf{E}}(\mathbf{r}). \quad (119)$$

In the above equation k denotes the length of the wave vector of a *plane wave* of ω angular frequency. Some solutions for special cases:

$$\text{Plane wave: } \tilde{\mathbf{E}}(\mathbf{r}) = \tilde{\mathbf{E}}_0 \cdot e^{i(\mathbf{k}\mathbf{r} + \varphi_0)} \quad (120)$$

$$\text{Spherical wave: } \tilde{\mathbf{E}}(\mathbf{r}) = \frac{E_0}{r} \hat{\mathbf{e}}(\mathbf{r}) \cdot e^{i(kr + \varphi_0)}, \quad (121)$$

where $k \triangleq 2\pi/\lambda$, and $\hat{\mathbf{e}}(\mathbf{r})$ refers to a unit vector locally normal to \mathbf{r} . In the general case the solution of the Helmholtz equation is (where the \mathbf{E}_0 quantity is a real vector):

$$\tilde{\mathbf{E}}(\mathbf{r}) = \mathbf{E}_0(\mathbf{r}) \cdot e^{ik_0 S(\mathbf{r})}. \quad (122)$$

Eikonal (εικών = image in Greek): a scalar quantity denoted by $S(\mathbf{r})$, describing points in space in terms of phase conditions of the electromagnetic wave. The phase difference ($\Delta\varphi$) measured between two points in space designated by \mathbf{r}_1 and \mathbf{r}_2 vectors expressed by the eikonal:

$$\Delta\varphi = k_0 S(\mathbf{r}_2) - k_0 S(\mathbf{r}_1). \quad (123)$$

In order for us to speak of phase conditions of the EM field corresponding to (122), it is necessary for the concept of “phase” to exist. Should the \mathbf{E}_0 field amplitude change in space faster than the wavelength, phase changes in space would make no sense anymore. The slow variation of amplitude occurs exclusively on one certain condition (see subsection 4.3), the role of which in optics, and wave theory in general, is significant. For this reason below we will briefly discuss this case.

4.3. Slowly varying amplitude approximation

The field can be approximated for a small ($\Delta\mathbf{r}$) change of spatial position around the \mathbf{r}_0 position vector in the next form:

$$\begin{aligned} \tilde{\mathbf{E}}(\mathbf{r}_0 + \Delta\mathbf{r}) &\approx \Delta\tilde{\mathbf{E}}(\mathbf{r}_0) + \tilde{\mathbf{E}}(\mathbf{r}_0) = \Delta(\tilde{\mathbf{E}}_0(\mathbf{r}_0) \cdot e^{ik_0 S(\mathbf{r}_0)}) + \tilde{\mathbf{E}}(\mathbf{r}_0) = \\ &= \Delta\tilde{\mathbf{E}}_0(\mathbf{r}_0) \cdot e^{ik_0 S(\mathbf{r}_0)} + \tilde{\mathbf{E}}_0(\mathbf{r}_0) \cdot \Delta e^{ik_0 S(\mathbf{r}_0)} + \tilde{\mathbf{E}}(\mathbf{r}_0). \end{aligned} \quad (124)$$

If we assume that the field amplitude changes much slower in space than the phase, then we can neglect the first term:

$$\begin{aligned} \tilde{\mathbf{E}}(\mathbf{r}_0 + \Delta\mathbf{r}) &\approx \tilde{\mathbf{E}}_0(\mathbf{r}_0) \cdot \Delta e^{ik_0 S(\mathbf{r}_0)} + \tilde{\mathbf{E}}(\mathbf{r}_0) = \\ &= \tilde{\mathbf{E}}_0(\mathbf{r}_0) \left(e^{i(k_0 \text{grad}(S(\mathbf{r}_0))\Delta\mathbf{r} + k_0 S(\mathbf{r}_0))} - e^{ik_0 S(\mathbf{r}_0)} \right) + \tilde{\mathbf{E}}(\mathbf{r}_0) \end{aligned} \quad (125)$$

The negative term just equals the last constant, thus

$$\tilde{\mathbf{E}}(\mathbf{r}_0 + \Delta\mathbf{r}) \approx \tilde{\mathbf{E}}_0(\mathbf{r}_0) \cdot e^{ik_0 S(\mathbf{r}_0)} \cdot e^{ik_0 \text{grad}(S(\mathbf{r}_0))\Delta\mathbf{r}}, \quad (126)$$

which is nothing else than the equation of a plane wave, where the local wave vector is:

$$\mathbf{k}_{\text{loc}} = k_0 \cdot \text{grad}(S(\mathbf{r}_0)). \quad (127)$$

This relationship can be interpreted as the EM field behaves locally in the slowly varying amplitude approximation as a plane wave, and field changes are primarily driven by the changes of phase and not the amplitude. Simultaneously, this also means that it makes sense to speak about *wavelength* exclusively in the slowly varying amplitude approximation. Later on, after having the geometrical optical approximation introduced we will gain more insight into (127).

Now let us examine what it means to neglect the change of field amplitude in (124). If \mathbf{J} is the Jakobian matrix, then the change can be written in the following form:

$$\begin{aligned} |\Delta\tilde{\mathbf{E}}_0(\mathbf{r}_0) \cdot e^{ik_0 S(\mathbf{r}_0)}| &\ll |\tilde{\mathbf{E}}_0(\mathbf{r}_0) \cdot \Delta e^{ik_0 S(\mathbf{r}_0)}| \Rightarrow \\ \Rightarrow |\mathbf{J}\Delta\mathbf{r} \cdot e^{ik_0 S(\mathbf{r}_0)}| &\ll |\tilde{\mathbf{E}}_0(\mathbf{r}_0) \cdot (ik_0 \cdot e^{ik_0 S(\mathbf{r}_0)} \cdot \text{grad}(S(\mathbf{r}_0)) \cdot \Delta\mathbf{r})|. \end{aligned} \quad (128)$$

Developing only the x-vector component and simplifying by the phase factor:

$$|\text{grad}(E_{0x}) \cdot \Delta \mathbf{r}| \ll E_{0x} \cdot k_0 \cdot |\text{grad}(S(\mathbf{r}_0)) \cdot \Delta \mathbf{r}|, \quad (129)$$

and overestimating the error (i.e. we assume that $\Delta \mathbf{r} \parallel \text{grad}(E_{0x})$):

$$\frac{|\text{grad}(E_{0x})|}{E_{0x}} \ll k_0 \cdot |\text{grad}(S(\mathbf{r}_0))|. \quad (130)$$

We will be able interpret this relationship in the geometrical optical approach, see below.

4.4. The fundamental equation of geometrical optics, conditions of validity

The Helmholtz equation is vectorial, but the Laplace operator is scalar, hence, the equation can be decomposed into independent vector components (E_x, E_y, E_z). Examining thus only the x-component:

$$\nabla^2 \tilde{E}_x(\mathbf{r}) = -k^2 \tilde{E}_x(\mathbf{r}). \quad (131)$$

Into this we substitute the (122) ansatz:

$$\nabla^2 (E_{0x}(\mathbf{r}) \cdot e^{ik_0 S(\mathbf{r})}) = -k^2 \cdot E_{0x}(\mathbf{r}) \cdot e^{ik_0 S(\mathbf{r})}. \quad (132)$$

It is apparent here that the real vector amplitude (E_{0x}) and the eikonal (S) are position dependent, accordingly, this will no longer be indicated. By using the below vector identity the left hand side of the equation can be transformed:

$$\nabla^2 (a \cdot b) \equiv a \cdot \nabla^2 b + 2 \nabla a \cdot \nabla b + b \cdot \nabla^2 a. \quad (133)$$

Hence the new equation (not indicating the (\mathbf{r}) position dependence):

$$E_{0x} \cdot \nabla^2 e^{ik_0 S} + 2 \nabla E_{0x} \cdot \nabla e^{ik_0 S} + e^{ik_0 S} \cdot \nabla^2 E_{0x} = -k^2 \cdot E_{0x} \cdot e^{ik_0 S}. \quad (134)$$

Based on $\nabla^2 u \equiv \nabla \cdot \nabla u$ we make an identical transformation on the first term:

$$\nabla^2 e^{ik_0 S} = \nabla (\nabla e^{ik_0 S}) = \nabla (ik_0 \cdot e^{ik_0 S} \cdot \nabla S) = -k_0^2 \cdot e^{ik_0 S} \cdot \nabla S \cdot \nabla S + \nabla^2 S \cdot ik_0 \cdot e^{ik_0 S} \quad (135)$$

where we also used $\nabla(a \cdot b) \equiv a \nabla b + b \nabla a$. The second term can be easily converted by applying the chain rule:

$$\nabla e^{ik_0 S} = ik_0 \cdot e^{ik_0 S} \cdot \nabla S. \quad (136)$$

Substituting (135), (136) into (134), and dividing out factor $e^{ik_0 S}$ from each term we get the following:

$$-k_0^2 \cdot E_{0x} \cdot \nabla S \cdot \nabla S + E_{0x} \cdot \nabla^2 S \cdot ik_0 + 2 \nabla E_{0x} \cdot ik_0 \cdot \nabla S + \nabla^2 E_{0x} = -k^2 \cdot E_{0x}. \quad (137)$$

This relationship has a real and an imaginary part, both of which must be equal. By equality of the real parts:

$$-k_0^2 \cdot E_{0x} \cdot |\nabla S|^2 + \nabla^2 E_{0x} = -k^2 \cdot E_{0x}. \quad (138)$$

If

$$|\nabla^2 E_{0x}| \ll k^2 \cdot E_{0x} \Rightarrow \frac{|\nabla^2 E_{0x}| \cdot \lambda^2}{E_{0x}} \ll 4\pi^2, \quad (139)$$

then the field is eliminated from equation (138):

$$k_0^2 \cdot |\nabla S|^2 = k^2 \Rightarrow |\nabla S|^2 = n^2. \quad (140)$$

This is the fundamental equation of geometrical optics, the so-called *eikonal equation*. Its simple form expresses that in case the approximation of geometrical optics is valid, then the phase changes of the EM field are not affected by changes of the field amplitude, but only by the refractive index distribution. It follows from the eikonal equation that

$$|\text{grad}(S(\mathbf{r}))| = n(\mathbf{r}), \quad (141)$$

where $n(\mathbf{r})$ is the local refractive index. By solving the eikonal equation using boundary conditions (i.e. we have to know the value of S along a surface) we can obtain the value of the eikonal in every point of space.

According to the above, the approximation of geometrical optics is only valid when both (130) and (139) conditions are fulfilled. Its illustrative meaning will be discussed below.

4.5. Local plane-wave approximation

If the basic condition of geometrical optics is satisfied beside condition (130), i.e. the eikonal equation holds, then $n = |\text{grad}(S)|$, hence (127) provides the length of the local wave vector:

$$k_{\text{loc}} = k_0 \cdot n(\mathbf{r}) = k(\mathbf{r}) \Rightarrow \lambda_{\text{loc}}(\mathbf{r}) = \frac{\lambda_0}{n(\mathbf{r})}. \quad (142)$$

The importance of relationships (126), (127) can be understood now: they show that in geometrical optical approximation light propagates in the direction of the local wave vector, and its phase changes are exclusively determined by the refractive index distribution. (For diffraction phenomena this is not necessarily true: the value of the phase in one specific point of space may also be affected by the field amplitude.) Since in this case the wavelength exactly equals that of a *plane wave* also having ω angular frequency and propagating in a medium of the respective refractive index, geometrical optics is often called as the *local plane-wave approximation*, though this might be somewhat misleading, since the wavefronts can be curved surfaces here too, given their radius of curvature is much larger than the wavelength.

The condition of slowly varying amplitude can be interpreted readily in geometrical optics too. According to the eikonal equation, after some rearrangements of (130) we get the following:

$$\frac{|\text{grad}(E_{0x})| \cdot \lambda}{E_{0x}} \ll 2\pi. \quad (143)$$

The above expression can be either interpreted as the relative field amplitude change measured over a displacement of a wavelength must be negligible, or as the wavelength must be much smaller than the characteristic spatial extent of field changes (approximation of short wavelengths). In order to illustrate this and assist comprehension let us assume that the field amplitude only changes in x-direction, and have (143) rearranged:

$$\lambda \ll 2\pi \left| \frac{\partial x}{\partial E_{0x}} \cdot E_{0x} \right|. \quad (144)$$

Here the quantity in the argument of the absolute value function corresponds to the distance, over which the field amplitude E_{0x} decreases to 0 (in first-order approximation).

Conditions (139) and (143) are only met simultaneously, i.e. we are within the validity of geometrical optics, if the relative change of the field amplitude (as well as the second derivative) are negligible over a displacement of a wavelength. This statement may even be clearer if reversed: we can only use geometrical optics, when the wavelength is very small relative to

extent of the characteristic spatial changes of the EM field (the amplitude does not affect the phase). For this reason, geometrical optics is often called as the “*approximation of short wavelengths*” too. This is not satisfied e.g. at the edges of shadows or in the vicinity of a focal spot, where in the former case geometrical optics would predict a discontinuity in \mathbf{E} , while in the latter it results in singularity.

The most important application field of the eikonal equation is the calculation of *inhomogeneous materials*. Examples:

- atmospheric effects (sunset, mirage)
- refraction in liquids placed in a gravity field
- lens in the eyeball (crystalline lens)
- certain planar waveguide lenses
- gradient lenses for fiber optics, laser diodes, imaging
- electron optics
- gravitational lenses.

5. MODELLING LIGHT PROPAGATION BY RAY-OPTICAL APPROACH

5.1. Description of phase evolution in geometrical optics

Wavefront: $S(\mathbf{r}) = \text{constant}$ (phasefront, surface of constant phase) ; $\mathbf{k} = k_0 \cdot \nabla S$

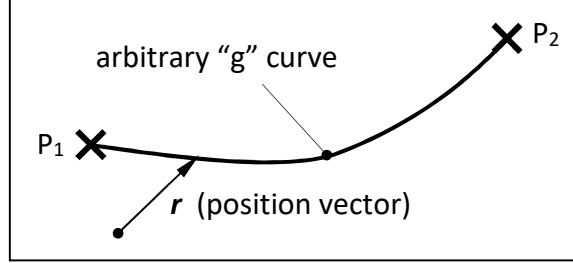


Fig. 15 Integrating the eikonal gradient along an arbitrary curve we get the phase difference.

According to (123), in geometrical optical approximation we get the phase delay $\Delta\varphi$ of the electromagnetic field between two points in space by the following integration (see Fig. 15):

$$\Delta\varphi(P_1, P_2) = k_0(S(P_2) - S(P_1)) = \int_g k_0 \cdot \text{grad}(S(\mathbf{r})) \cdot d\mathbf{r}, \quad (145)$$

where “g” is an arbitrary curve. If light is able get from P_1 to P_2 , then it is obvious that at least one g' curve must exist, for which it is true that its tangent is parallel to the $\text{grad}(S(\mathbf{r}))$ direction of local wave propagation in each of its points. For this curve the phase delay is:

$$\Delta\varphi(P_1, P_2) = k_0 \int_g |\text{grad}(S(\mathbf{r}))| dr = k_0 \int_g n(\mathbf{r}) dr = k_0 \cdot OPL(P_1, P_2), \quad (146)$$

where we made use of (141), and *OPL* is the *Optical Path Length* between P_1 and P_2 . Comparing the above equation with (145): $\Delta S = OPL$, in case of integrating along g'! A further property of the g' special curve is that in each point it is normal to the wavefront, since its tangent points in the direction of $\text{grad}(S)$, which vector is by definition normal to surfaces described by $S = \text{const.}$, i.e. the wavefronts. The g' curves defined above are called as *light rays*. The eikonal equation is insensitive for the sign of $\text{grad}(S)$ (i.e. the direction of a light ray), which corresponds to the statement that light rays are reversible.

Light ray definition 1): the light ray is a curve normal to the wavefronts in each of its points, in other words its tangent points in the direction of $\text{grad}(S)$. Since the wave vector (\mathbf{k}) is normal to the wavefronts by definition, the tangent of a light ray points in the direction of \mathbf{k} . If we measure the optical path along a light ray, then $\Delta S = OPL$.

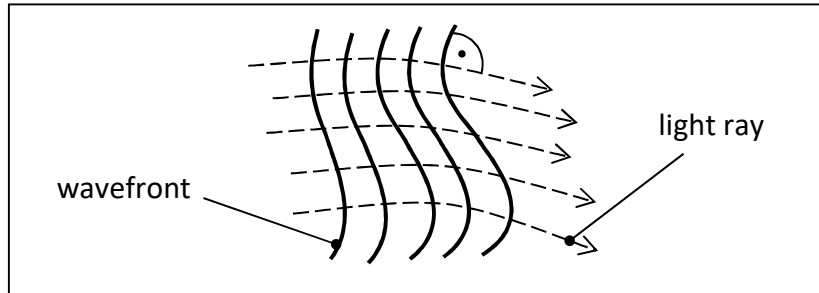


Fig. 16 Connection of wavefronts and light rays.

Light refraction at the boundary of two media: It can be proven (see e.g. [6]), that for light rays reaching a media interface Snell's law is valid in the geometrical optical approximation. Since the general validity of the refraction law can be derived for plane waves (see previously), this relationship further increases the analogy of the approximation of local plane waves.

5.2. Description of energy propagation in geometrical optics

Poynting vector: $\mathbf{S}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}, t) \times \mathbf{H}(\mathbf{r}, t)$ (power density vector, $\mathbf{E} = \text{Re}\{\tilde{\mathbf{E}}\}$!)

The time-averaged power density vector in case of a monochromatic field and isotropic dielectrics, in geometrical optical approximation:

$$\langle \mathbf{S}(\mathbf{r}, t) \rangle = \frac{\tilde{\mathbf{E}}_0(\mathbf{r}) \times \tilde{\mathbf{H}}_0^*(\mathbf{r})}{2} = \frac{\mathbf{k}}{\mu\omega} \frac{|\tilde{\mathbf{E}}_0(\mathbf{r})|^2}{2} = \frac{k}{\mu\omega} \frac{E_0(\mathbf{r})^2}{2} \cdot \frac{\nabla S}{n} = \frac{k_0}{\mu\omega} \frac{E_0(\mathbf{r})^2}{2} \cdot \nabla S. \quad (147)$$

The magnitude of this vector is called “*I*” intensity. The light power traversing an infinitesimal area dA can be written by using intensity as follows, see Fig. 17:

$$dP(\mathbf{r}) = \langle \mathbf{S}(\mathbf{r}, t) \rangle \cdot d\mathbf{A} = I(\mathbf{r}) \cdot dA \cdot \cos(\theta) \quad (148)$$

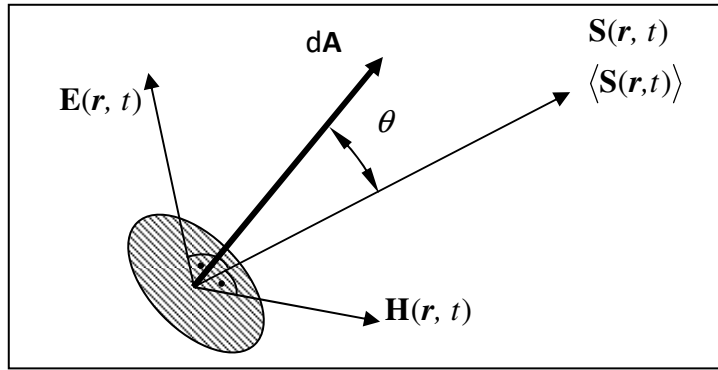


Fig. 17 Situation of the Poynting vector, field vectors and the surface normal.

Equation (147) also justifies that in case the validity conditions of the geometrical optical approximation hold, the Poynting vector and the gradient of the eikonal point in the same direction inside isotropic materials.

Light ray definition 2): a curve, the tangent of which points in the direction of the power density vector in every point, see Fig. 18. Definitions 1) and 2) are equivalent in *isotropic media*.

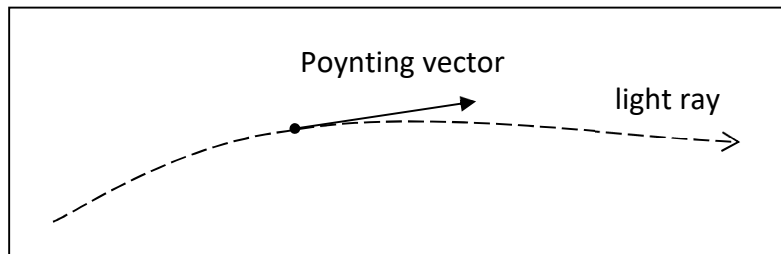


Fig. 18 The tangent of a light ray points in the direction of the Poynting vector.

Light beam: a tube-like domain delimited by rays, from which energy cannot escape except through the front and end surfaces (within the validity of geometrical optical approximation).

5.3. Intensity law of geometrical optics

From equality of the imaginary parts of equation (137) we get the following expression:

$$E_{0x} \cdot \nabla^2 S + 2\nabla E_{0x} \nabla S = 0, \quad (149)$$

which is the so-called *transport equation*, for cases when no sources are present. With its help, and having the eikonal known, the field can be determined point by point. It is worth noting that when formulating this equation we did not use the geometrical optical approximation, only that of the slowly varying amplitude.

By using simple identity transformations the (149) transport equation can be rearranged to the following form (see [7], subsection 7.3):

$$\nabla(E_{0x}^2 \cdot \nabla S) = 0, \quad (150)$$

from which we can see e.g. that for a plane wave propagating in a homogeneous medium ($\nabla S = \text{const.}$) the field amplitude does not change spatially. The term in parentheses is proportional to the time-averaged Poynting vector of the x component of the field, i.e. the intensity, see (147). Making the same equation for all three field components, and summing up the results we get the following:

$$\nabla \langle \mathbf{S} \rangle = 0 \Leftrightarrow \oint \langle \mathbf{S} \rangle \cdot d\mathbf{A} = 0. \quad (151)$$

Since light rays are parallel to the Poynting vector, equation (151) is equivalent with a relationship that is valid for infinitesimally thin light beams and also follows from the law of energy conservation, the so-called *intensity law*:

$$\langle \mathbf{S}_1 \rangle \cdot d\mathbf{A}_1 = \langle \mathbf{S}_2 \rangle \cdot d\mathbf{A}_2, \quad (152)$$

which is true in such (so-called “regular”) spatial domains, where light rays do not intersect each other.

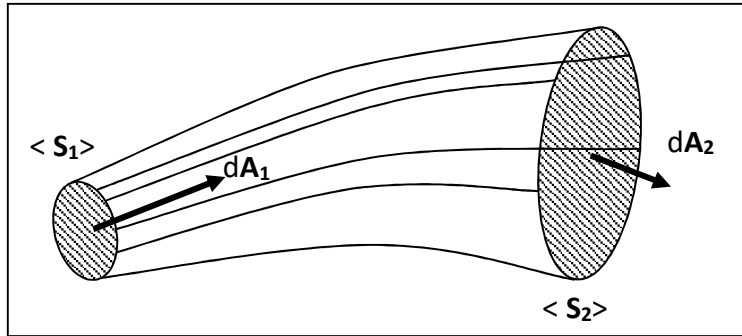


Fig. 19 Demonstration of the intensity law in case of a light beam.

5.4. Equation of light rays according to arc-length parametrization

Since we defined light rays by the eikonal, we derive their equation from the eikonal equation. Light ray = spatial curve, let us denote it by “ g ”. The specification of g as a function of the parameter p takes the following form:

$$g: \mathbf{r} = \mathbf{r}(p), \quad (153)$$

which is called the parametrization of the curve. In the above equation \mathbf{r} is a position vector scanning the trajectory of the light ray. Since light rays are parallel to $\text{grad } S$, and $\text{grad } S(\mathbf{r}) \perp S(\mathbf{r}) = \text{const.}$ (wavefront), as well as $d\mathbf{r}(p)/dp \parallel g$ curve tangent, we can write:

$$\frac{d\mathbf{r}(p)}{dp} \sim \text{grad}(S(\mathbf{r})). \quad (154)$$

Curve g has an infinite number of parametrization, because if $\mathbf{r}(p)$ is a parametrization, then $\mathbf{r}(q(p))$ it is too, given that $q(p)$ is strictly monotonic continuous together with its first derivative. For this reason the equation of light rays includes an indefinite function. The derivatives of two parametrizations of a given curve are connected by the following formula:

$$\frac{d\mathbf{r}(p)}{dp} = \frac{d\mathbf{r}(q)}{dq} \frac{dq}{dp}, \quad (155)$$

if $q(p)$ describes the relationship between the two parameters. Since dq/dp is a scalar quantity, the above expression means that all derivative vectors of a curve obtained by different parametrizations point in the same direction (the tangent of the curve), only their magnitude is different. Hence, it is certain that a specific $\mathbf{r}(q)$ parametrization must exist, for which it holds:

$$\frac{d\mathbf{r}(q)}{dq} = \text{grad}(S(\mathbf{r})). \quad (156)$$

Substituting relationship (156) into (155) we get the following:

$$\frac{d\mathbf{r}(p)}{dp} = \frac{dq}{dp} \cdot \text{grad}(S(\mathbf{r})) = f(p) \cdot \text{grad}(S(\mathbf{r})), \quad (157)$$

where $f(p)$ is an arbitrary function, not affecting the shape of the $\mathbf{r}(p)$ curve, but its parametrization. Should we choose the $f(p) = 1/n(\mathbf{r})$ function, (157) will take the following form:

$$\frac{d\mathbf{r}(p)}{dp} = \frac{\text{grad}(S(\mathbf{r}))}{n(\mathbf{r})}. \quad (158)$$

On account of the eikonal equation, this is a unit vector, and from mathematics we know that the derivative of a curve by *arc length* provides the tangent vector of unity. Thus $p = s$, which is the arc length measured along a light ray. Putting this into (158), after some rearrangement:

$$n(\mathbf{r}(s)) \frac{d\mathbf{r}(s)}{ds} = \nabla S(\mathbf{r}(s)). \quad (159)$$

In practice it is more convenient to perform calculations by using the refractive index instead of the eikonal. Based on the (140) eikonal equation we could do this if we somehow introduced $|\nabla S|^2$ into (159). For this let us differentiate it by arc length once more:

$$\frac{d}{ds} \left(n \cdot \frac{d\mathbf{r}}{ds} \right) = \frac{d}{ds} \nabla S. \quad (160)$$

Then let us convert the right side of the equation according to the below:

$$\frac{d}{ds} \nabla S = (\nabla \cdot \nabla S) \cdot \frac{d\mathbf{r}}{ds}, \quad (161)$$

and write the relationship (158) to replace $d\mathbf{r}(s)/ds$ in it:

$$(\nabla \cdot \nabla S) \cdot \frac{\nabla S}{n} = \frac{1}{n} \cdot \frac{1}{2} \cdot \nabla \cdot |\nabla S|^2. \quad \left(\text{see } \frac{df^2}{dx} = 2f \cdot \frac{df}{dx} \right) \quad (162)$$

By using the eikonal equation, the formula simplifies further:

$$\frac{1}{n} \cdot \frac{1}{2} \cdot \nabla(n^2) = \nabla n = \text{grad}(n). \quad (163)$$

Equating this with the left hand side of (160) we obtain the differential equation of light rays expressed in a parametrization by the arc length:

$$\frac{d}{ds} \left(n \cdot \frac{d\mathbf{r}}{ds} \right) = \text{grad}(n), \quad (164)$$

where $\mathbf{r}(s)$ is the parametrization of the resolution curve. Example – the case of homogeneous media ($n = \text{const.}$). Here the differential equation of light rays simplifies to the following form:

$$\frac{d^2\mathbf{r}}{ds^2} = 0, \quad (165)$$

the solution of which can be attained after two-fold integration:

$$\mathbf{r} = \mathbf{a} \cdot s + \mathbf{b}, \quad (166)$$

where \mathbf{a} and \mathbf{b} are arbitrary constant vectors (determined by the boundary conditions). This means that in homogeneous media light rays are straight lines passing through the point denoted by \mathbf{b} , and pointing in direction \mathbf{a} .

6. PARAXIAL APPROXIMATION OF GEOMETRICAL OPTICS

Application fields of geometrical optics

- imaging systems (this is what we are dealing with)
- illumination systems (see the separate subject of “Design of nonimaging optics”)

When modelling imaging and illumination systems the object/light source is decomposed into the sum of spatially coherent sources (point sources).

Geometrical optical approximations in image formation

- real ray-tracing (repeated refraction, propagation) → imaging errors, optimization
- third-order approximation (aberration theory) → analysis of imaging errors
- first-order (paraxial) approximation → magnification, object-image position, speed of lens, max. resolution

Ideal image formation – simplified definition

- images point-to-point, i.e. the image and object points are “conjugates” of each other
- the imaging system includes an axis that is imaged onto itself (e.g. it is rotationally symmetric), this is the so-called optical axis
- geometric figures in planes normal to the optical axis are imaged into similar ones (i.e. without distortion)

In case of ideal image formation we can find an infinite number of ray trajectories between the object and image points. However, on account of Fermat’s principle light can only travel along a curve, where the optical path is stationary in first order. The resolution to this apparent contradiction is that the optical path is *equal* along all rays between object and image points.

6.1. First-order (paraxial) approximation (Gaussian optics)

Conditions of the paraxial approximation:

$$\theta \approx \sin(\theta) \approx \tan(\theta) \quad [\theta] = \text{rad} ; \text{ (hence “first order” approximation)}$$
$$y \ll r \quad \text{signed quantities!}$$

Thus light rays travel in the vicinity of the optical axis at a small angle to it, and the surfaces can be *approximated by planes*. From these it follows: spherical wave ~ paraboloid surface.

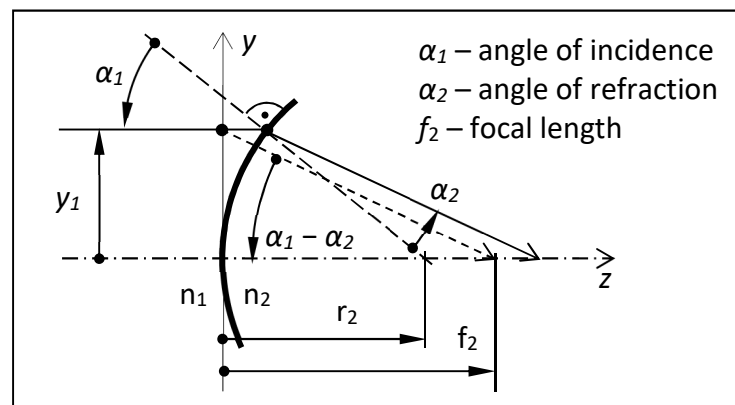


Fig. 20 Determining the focal length of a refractive spherical surface in parax. approximation.

Focal point: image of an infinitely distant object point on the optical axis (parax. approx.!)

Determination of the focal length in case of only one refractive surface:

$$\left. \begin{array}{ll} n_1 \alpha_1 = n_2 \alpha_2 & \text{refraction} \\ \alpha_1 r_2 = y_1 & \text{surface normal} \\ f_2(\alpha_1 - \alpha_2) = y_1 & \text{ideal imaging} \end{array} \right\} \Rightarrow f_2 = r_2 \frac{n_2}{n_2 - n_1} [\alpha] = \text{rad!} \quad (167)$$

Formal treatment of mirrors: $n_2 = -n_1$. *Refractive power:*

$$P_2 \triangleq \frac{n_2}{f_2} = \frac{n_2 - n_1}{r_2} [\text{diopter} = \text{m}^{-1}] \quad (168)$$

In English textbooks the refractive power is also used to be denoted by the symbol ϕ . In case of two surfaces (1 and 2) placed at zero distance one after the other:

$$P_{\text{resultant}} = P_1 + P_2 \quad (169)$$

In paraxial approximation refraction and free-space propagation can be calculated independently of each other in the x-z and y-z planes! This approximation fulfills the conditions of ideal image formation! (Derivation see in [6], vol. I., subsection 2.3.4.)

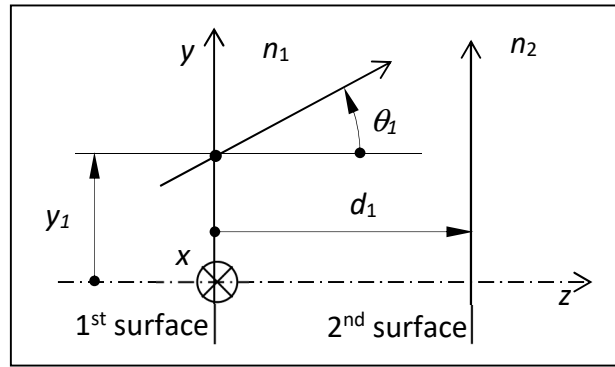


Fig. 21 Demonstration of the coordinates of a light ray.

The basic question of paraxial approximation is if the position and direction ($y_1; \theta_1$) of a light ray are given at surface 1, then how much are ($y_2; \theta_2$) at a further surface 2. Since the approximation is linear, the answer can be expressed by matrix formalism. Writing it only in the y-z plane on account of the x-z / y-z independence:

$$\begin{bmatrix} y_2 \\ n_2 \theta_2 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ n_1 \theta_1 \end{bmatrix}. \quad (170)$$

The quantity $n \cdot \theta$ describing the direction of a light ray is called the optical *direction cosine*. For axially symmetric systems the equation system written in the x-z direction has the same ABCD matrix between the 1-2 planes as above. Based on (170), the position coordinate transformation of light rays can be described by the following equations:

$$\begin{aligned} y_2 &= A \cdot y_1 + B \cdot (n_1 \theta_1) \\ n_2 \theta_2 &= C \cdot y_1 + D \cdot (n_1 \theta_1) \end{aligned} \quad (171)$$

If imaging takes place between the two members of the plane pair, then the y_2 position coordinate is independent of the starting direction of the light ray, which can only be true if $B = 0$. Then it can very easily be proven that:

$$\begin{aligned} A &= m \\ D &= m_\alpha \cdot \frac{n_2}{n_1} \quad \text{ha } y_1 = 0 \end{aligned} \quad (172)$$

Definitions of m and m_α see in 6.2. The $ABCD$ matrix of light refraction taking place at a spherical surface (if the two planes are coincident with each other and the refractive surface):

$$\mathbf{R} = \begin{bmatrix} 1 & 0 \\ -P_1 & 1 \end{bmatrix}, \quad (173)$$

where P_1 is the refractive power of the surface (interpretation see above). The $ABCD$ matrix of free-space propagation:

$$\mathbf{T} = \begin{bmatrix} 1 & d_1/n_1 \\ 0 & 1 \end{bmatrix}. \quad (174)$$

where d_1 is the distance of planes 1 and 2, n_1 is the refractive index between them. It must be noted that plane surfaces have a refractive power of zero. Evidently, the matrix of a complex system \mathbf{M} consisting of N surfaces is given by the product of the above elementary matrices:

$$\mathbf{M} = \mathbf{T}_N \cdot \mathbf{R}_N \cdot \dots \cdot \mathbf{T}_2 \cdot \mathbf{R}_2 \cdot \mathbf{T}_1. \quad (175)$$

6.2. Image construction in case of a thin lens, properties of paraxial imaging

thickness \ll radii of curvature of lens surfaces

- s, s' – object, image distance
- f' – image-space focal length ($f' = s'$, ha $s = -\infty$)
- y, y' – object, image height
- θ, θ' – object field angle, image field angle (from the law of refraction: $n \cdot \theta = n' \cdot \theta'$)
- m – lateral magnification ($m \triangleq y'/y$) (also known as transverse magnification)
- m_L – longitudinal magnification ($m_L \triangleq \partial s' / \partial s$)
- m_α – angular magnification ($m_\alpha \triangleq \alpha' / \alpha$)
- NA – object-space numerical aperture ($NA \triangleq n \cdot \sin \alpha$) – not a paraxial quantity

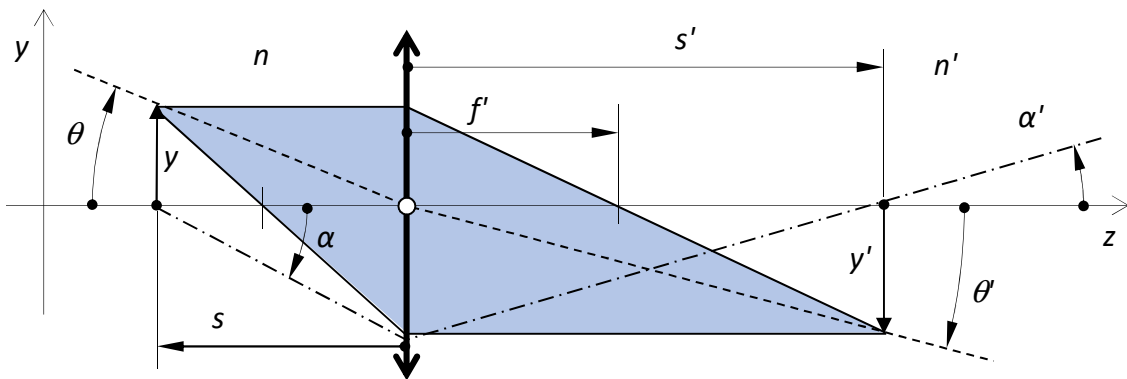


Fig. 22 Construction of imaging in case of a thin lens. Distances and angles are signed quantities! Hence $s' > 0$, but $s < 0$, and $\alpha' > 0$ but $\alpha < 0$ in the figure.

Easy-to-derive, fundamental paraxial rules (valid not only in case of thin lenses, see principal planes below):

$$\begin{aligned}
 m &= \frac{s'}{s} \cdot \frac{n}{n'} \\
 m_L &= \frac{n'}{n} \cdot m^2 \\
 m_\alpha \cdot m &= \frac{n}{n'} \quad (\text{Lagrange – Helmholtz equation}) \Rightarrow A \cdot D = 1 \\
 \frac{n}{f} &= -\frac{n'}{f'}
 \end{aligned} \tag{176}$$

6.3. Ideal imaging in case of a complex optical system

Theorem: in paraxial approximation every optical system images ideally all points of a homogeneous object space into a homogeneous image space, if at least one ray exists that intersects the optical axis both in the object and image space.

Below we discuss the derivation of this theorem. Let the z_1 - z_2 plane pair for an optical system be where this intersection occurs, and let us describe light propagation between these planes by the $ABCD$ matrix. If there is one ray running at an angle to the optical axis for which it is true that $y_1 = y_2 = 0$, then it is only possible if $B = 0$, i.e. there must be image formation between the two planes. Considering the z_1 - z_2 planes as *reference planes* now let us examine whether it is possible for another arbitrary object plane of $z_1 + \Delta z_1$ position to have an image plane. Let the position of this hypothetical image plane be $z_2 + \Delta z_2$. Since we regard the $ABCD$ matrix corresponding to the $(z_1; z_2)$ planes as given, we use it to express the $A'B'C'D'$ matrix of the translated plane pair:

$$A'B'C'D' = \mathbf{T}_2 \cdot ABCD \cdot \mathbf{T}_1. \tag{177}$$

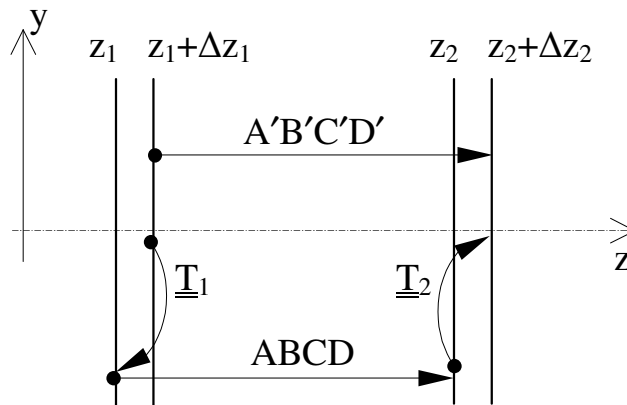


Fig. 23 Explanation to the determination of the $A'B'C'D'$ matrix.

According to the (174) general matrix of free-space propagation:

$$\mathbf{T}_1 = \begin{bmatrix} 1 & -\frac{\Delta z_1}{n_1} \\ 0 & 1 \end{bmatrix} \text{ and } \mathbf{T}_2 = \begin{bmatrix} 1 & \frac{\Delta z_2}{n_2} \\ 0 & 1 \end{bmatrix}. \tag{178}$$

After matrix multiplication the value of the B' parameter is (only this includes B):

$$B' = B + \frac{\Delta z_2}{n_2} D - \left(A + \frac{\Delta z_2}{n_2} C \right) \frac{\Delta z_1}{n_1}. \quad (179)$$

According to our definition there is imaging between the two original reference planes, thus $B = 0$, and also due to the Lagrange-Helmholtz equation (in case of imaging):

$$A \cdot D = 1 \quad (180)$$

Hence, we obtain the following for $A'B'C'D'$:

$$\begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} = \begin{bmatrix} A + \frac{\Delta z_2}{n_2} C & \frac{\Delta z_2}{A \cdot n_2} - \left(A + \frac{\Delta z_2}{n_2} C \right) \frac{\Delta z_1}{n_1} \\ C & D - \frac{\Delta z_1}{n_1} C \end{bmatrix}. \quad (181)$$

Since the above calculation results in $C' = C$, its value is independent of the object and image positions, thus C is a characteristic parameter of the optical system under investigation. Its significance will make sense in subsection 6.4. If there is imaging between the planes shifted by Δz_1 and Δz_2 as well, then $B' = 0$, from which:

$$\frac{n_2}{\Delta z_2} - \frac{1}{A^2} \frac{n_1}{\Delta z_1} = -\frac{C}{A}. \quad (182)$$

This is the general expression of the *lens formula*, written with respect to reference planes that describe an arbitrary imaging.

6.4. Description of complex lens systems by using principal planes

Principal planes: object-image plane pair, for which it holds by definition: $m = +1$. Such a pair can be determined for all optical systems. The positions of principal planes are *unambiguous*.

If we take the principal planes as references ($A = +1$), (182) takes the following simple form:

$$\frac{n_2}{\Delta z_2} - \frac{n_1}{\Delta z_1} = -C. \quad (183)$$

Let Δz_1 tend to minus infinity. Then Δz_2 gives the position of the image-space focal spot by definition:

$$f' \triangleq \Delta z_2 ; \quad C = -\frac{n'}{\Delta z_2} = -\frac{n'}{f'} ; \quad P \triangleq \frac{n'}{f'}, \quad (184)$$

where we introduced the concept of refractive power (P), and the object- and image-space refractive indices: $n \triangleq n_1$ and $n' \triangleq n_2$. The f' distance measured from the image-space principal plane is called the *effective focal length*. With these we get the well-known, Gaussian expression of the lens formula ($s \triangleq \Delta z_1$ and $s' \triangleq \Delta z_2$):

$$\frac{n'}{s'} - \frac{n}{s} = \frac{n'}{f'} \Rightarrow \frac{n'}{s'} = \frac{n'}{f'} + \frac{n}{s}. \quad (185)$$

Consequently, if we measure the object and image distances, as well as the (effective) focal length from the principal planes, then the lens formula is valid for any arbitrary lens system!

Special lens systems: **telephoto objective**, $f_{\text{effective}} > \text{construction length}$ (see photo camera)
reverse telephoto objective, back focal length $> f_{\text{effective}}$ (see projector)

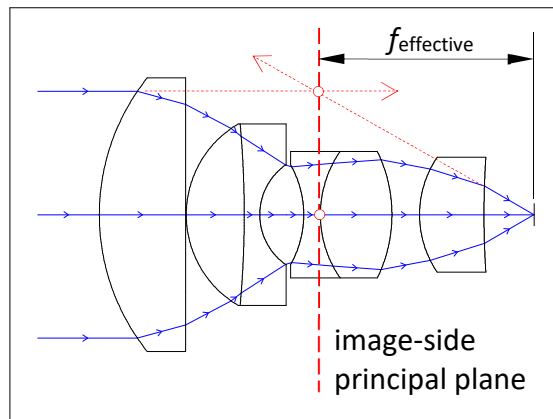


Fig. 24 Graphic determination of principal planes. See details during the lecture.

7. TWO-BEAM INTERFERENCE

Sources: [1], [3], [6], [8]

7.1. Concept of an interferogram

In the previously presented geometrical optical (i.e. short-wavelength) approximation we described the spatial phase and intensity variations of EM waves by differential equations. Since this type of treatment is based on the assumption that propagation is not affected by the field amplitude (but by the spatial distribution of the refractive index), the effect exerted by one part of a beam to another cannot be described this way. We expressed this as the differential equations of geometrical optics can only be used in regular domains, i.e. where one part of a beam coming from a given direction never overlaps with those coming from other directions (in other words: when light rays do not intersect each other). Due to this deficiency the differential equation treatment can only be used with limitations. Instead of it, the determination of light distribution is carried out by light rays representing the direction of phase and energy propagation. Every light ray can be assigned a small amount of power. By summing these up in case of incoherent (i.e. diffuse) illumination for an arbitrary surface of the optical system we can determine either the traversing power, or when referenced to a unit surface the intensity. Simply phrased: where the surface density of light rays is higher there the intensity is higher, where it is less dense the intensity is smaller. For spatially and temporally coherent radiations (e.g. laser light) the summation has to be done not for power, but complex amplitude – in such cases interference occurs.

Interference is the phenomenon when two or more discrete waves are superimposed in phase creating a spatially *standing wave*, which can be (generally) observed as dark-bright interference fringes in case of electromagnetic waves. Its modelling is usually done in geometrical optical approximation, determining the phase of the interfering beams from the optical path length measured along light rays, and matching 2π rad phase delay to the wavelength in the medium. Of course, the discrete superposition can even be used when the propagation of the interacting beams can only be described by diffraction.

In what follows we will deal with the geometrical optical description. As already seen, the spatial and temporal dependence of the phase of the field can be given with the help of the eikonal equation as:

$$\tilde{\mathbf{E}}(\mathbf{r}, t) = \mathbf{E}_0(\mathbf{r}) \cdot e^{-i\omega t} \cdot e^{ik_0 S(\mathbf{r})} = \mathbf{E}_0(\mathbf{r}) \cdot e^{-i\omega t} \cdot e^{i\varphi(\mathbf{r})}, \quad (186)$$

where $\mathbf{E}_0(\mathbf{r})$ denotes the real vector amplitude of the field. Here we applied the method of positive spatial phase propagation again. The local wave vector in geometrical optical approach is:

$$\mathbf{k}_{\text{loc}} = k_0 \cdot \text{grad}(S(\mathbf{r})). \quad (187)$$

Since $n = |\text{grad}(S)|$, its magnitude is the following:

$$k_{\text{loc}} = k_0 \cdot n(\mathbf{r}) \Rightarrow \lambda_{\text{loc}}(\mathbf{r}) = \lambda_0/n(\mathbf{r}). \quad (188)$$

The phase difference of light between two points (P_1 and P_2) is then:



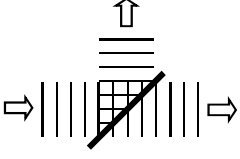
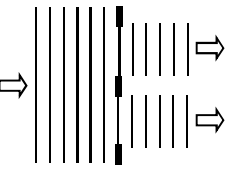
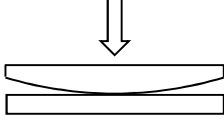
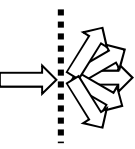
$$\Delta\varphi(P_1, P_2) = k_0 \int_g n(\mathbf{r}) dl = k_0 \cdot OPL(P_1, P_2), \quad (189)$$

where $g \mapsto g(\mathbf{r})$ is the trajectory of an arbitrary light ray. For simplicity, in what follows we will consider the field amplitude as a scalar quantity, i.e. we assume that the polarization state of the EM field remains approximately unchanged during light propagation:

$$\tilde{\mathbf{E}}(\mathbf{r}, t) \rightarrow \tilde{E}(\mathbf{r}, t) ; \mathbf{E}_0(\mathbf{r}) \rightarrow E_0(\mathbf{r}). \quad (190)$$

This is true when the light beams are not at a too steep angle relative to each other. For the sake of completeness we will briefly discuss the general case in the supplementary material at the end of this chapter.

Classification of interference phenomena

By components:	two-beam multiple-beam		
By implementation:	amplitude splitting wavefront splitting		
By standing wave form:	spatial dependence direction dependence		
By light path:	double-path single-path common-path		

Approximations, conditions, notations

Approximations: homogeneous, linear, isotropic medium, scalar approximation

Condition 1: same frequency (otherwise no standing wave)

Condition 2: same polarization (otherwise no observable superposition)

Condition 3: spatially and temporally coherent waves
(otherwise visibility drops, see later in detail)

Condition 4: intensities of the beams are identical (otherwise visibility drops)

Notations: T – period ν – frequency ($= 1/T$)
 ω – angular frequency ($= 2\pi/T$) λ – wavelength (in medium)
 k – wave number ($= 2\pi / \lambda$) v – phase velocity in medium ($= \lambda \cdot \nu = \omega / k$)

Technical examples for two-beam interference

Wavefront splitting:

- Young's double-slit experiment
- Lloyd's mirror
- Fresnel's biprism

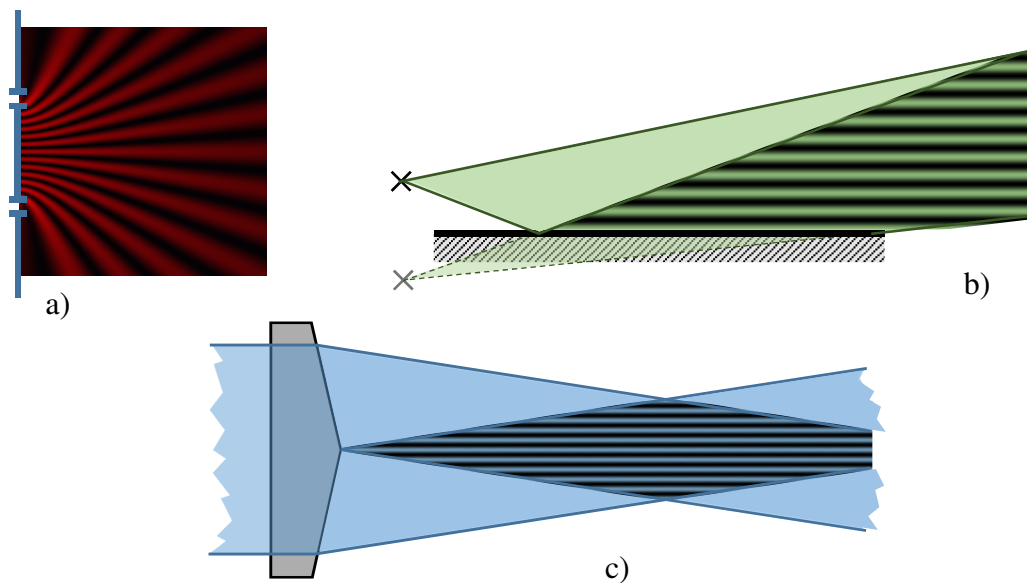


Fig. 25 Schematic structure of wavefront-splitting interferometers. a) Young's two-slit experiment, b) Lloyd's mirror, c) Fresnel's biprism.

Amplitude splitting:

- a) Michelson interferometer (see also Twyman-Green interferometer)
- b) Mach-Zehnder interferometer
- c) Fizeau interferometer (e.g. air gap between two glass sheets, Newton fringes)
- d) Plane-parallel plate (shearing interferometer)

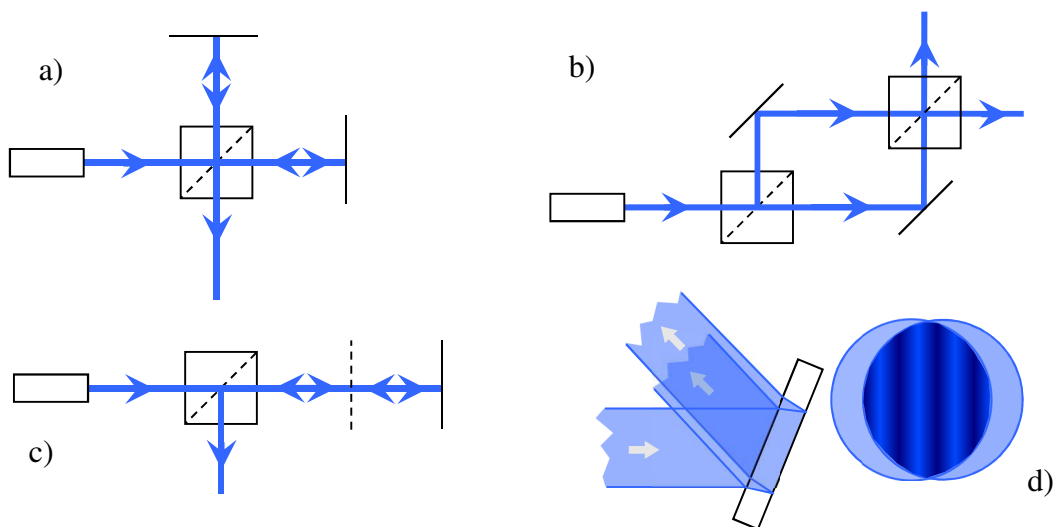


Fig. 26 Examples for amplitude-splitting interferometers. In case of d) we only see fringes with a perfect plane-parallel plate if the incoming beam has curved wavefronts.

The Michelson interferometer is used for distance measurement, Fourier transform (FTIR) spectroscopy, surface shape measurement, gravitational wave detection (LIGO), the Mach-Zehnder is applied in holographic arrangements, for refractive index measurement of gases or quantum computations, Fizeau interferometers are primarily used for surface and wavefront shape measurement, and shearing interferometers are used for the analysis/adjustment of the collimation state of lasers (i.e. in order to determine how planar their wavefront is).

7.2. Interference of two plane waves in paraxial approximation

Now let us examine the simplest case, when two plane waves of identical linear polarization but different direction of propagation interact:

$$\begin{aligned} E_1(\mathbf{r}, t) &= E_{01} \cos(-\omega t + \mathbf{k}_1 \mathbf{r} + \varphi_1) = E_{01} \cos(\Phi_1) \\ E_2(\mathbf{r}, t) &= E_{02} \cos(-\omega t + \mathbf{k}_2 \mathbf{r} + \varphi_2) = E_{02} \cos(\Phi_2) \end{aligned} \quad (191)$$

In this discussion we exceptionally use *real* terminology, since this way the effect of time averaging will be more recognizable at intensity calculations. The resultant field is thus:

$$E(\mathbf{r}, t) = E_1(\mathbf{r}, t) + E_2(\mathbf{r}, t). \quad (192)$$

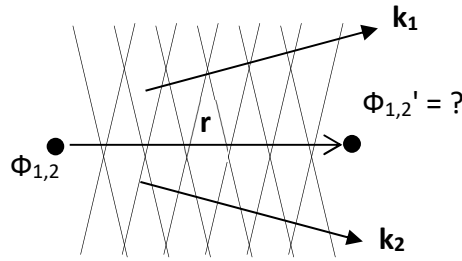


Fig. 27 Drafting the basic problem: interference of two plane waves arriving at an angle.

Power density is given by the magnitude of the Poynting vector:

$$\mathbf{S}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}, t) \times \mathbf{H}(\mathbf{r}, t), \quad (193)$$

which is of the following form for plane waves propagating in dielectric media:

$$S(\mathbf{r}, t) = \frac{k}{\mu\omega} E(\mathbf{r}, t)^2 = v\varepsilon \cdot E(\mathbf{r}, t)^2. \quad (194)$$

In the above equation the propagation (phase) velocity is denoted by “ v ”. Intensity is the time average of S taken over a duration of T :

$$I(\mathbf{r}, T) \triangleq \langle S(\mathbf{r}, t) \rangle \Rightarrow I(\mathbf{r}) = v\varepsilon \cdot \langle E(\mathbf{r}, t)^2 \rangle, \quad (195)$$

since for harmonic signals I is independent of T , if $T \rightarrow \infty$. In case \mathbf{k}_1 is approximately parallel to \mathbf{k}_2 , i.e. we are in paraxial approximation, instead of E we can write the sum of plane waves in order to determine the resultant intensity. For light propagation at larger angles \mathbf{E}_1 and \mathbf{E}_2 are not parallel, which makes the discussion more complicated (see supplementary material), although it will conclude to the same result as we get here. Hence, in paraxial, scalar approximation the intensity of the interference pattern is as follows:

$$I(\mathbf{r}) = v\varepsilon \cdot \langle (E_1(\mathbf{r}, t) + E_2(\mathbf{r}, t))^2 \rangle. \quad (196)$$

Substituting the (191) fields into the squared component:

$$(E_1(\mathbf{r}, t) + E_2(\mathbf{r}, t))^2 = E_{01}^2 \cos^2(\Phi_1) + E_{02}^2 \cos^2(\Phi_2) + 2E_{01}E_{02} \cos(\Phi_1) \cos(\Phi_2). \quad (197)$$

The mixed product can be transformed by the $2\cos(\alpha)\cos(\beta) \equiv \cos(\alpha-\beta) + \cos(\alpha+\beta)$ identity:

$$2E_{01}E_{02} \cos(\Phi_1) \cos(\Phi_2) = E_{01}E_{02}(\cos(\Phi_1 - \Phi_2) + \cos(\Phi_1 + \Phi_2)). \quad (198)$$

The term describing the phase difference is time-independent according to (191):

$$\delta(\mathbf{r}) \triangleq \Phi_1 - \Phi_2 = (\omega - \omega)t + (\mathbf{k}_1 - \mathbf{k}_2)\mathbf{r} + (\varphi_1 - \varphi_2) = \Delta\mathbf{k} \cdot \mathbf{r} + \Delta\varphi. \quad (199)$$

And the sum of the phases is:

$$\Phi_1 + \Phi_2 = 2\omega t + (\mathbf{k}_1 + \mathbf{k}_2)\mathbf{r} + (\varphi_1 + \varphi_2), \quad (200)$$

the cosine of which is zero after time averaging. Thus (196) leaves us the following:

$$I(\mathbf{r}) = \frac{v\varepsilon}{2} \cdot (E_{01}^2 + E_{02}^2 + 2E_{01}E_{02} \cos(\delta(\mathbf{r}))). \quad (201)$$

Since ω for the two beams is identical, all time-dependent terms become eliminated at time averaging, i.e. interference occurs: the intensity patterns is temporally steady, and changes periodically with δ thanks to the “cos” function.

Interference can be more easily discussed with complex notation:

$$I = \frac{v\varepsilon}{2} \cdot (\tilde{E}_1 + \tilde{E}_2) \cdot (\tilde{E}_1 + \tilde{E}_2)^* = \frac{v\varepsilon}{2} \cdot (\tilde{E}_1\tilde{E}_1^* + \tilde{E}_2\tilde{E}_2^* + \tilde{E}_1\tilde{E}_2^* + \tilde{E}_2\tilde{E}_1^*), \quad (202)$$

where * denotes the complex conjugate and

$$\tilde{E}_1 = E_{01} \cdot e^{i(-\omega t + \mathbf{k}_1\mathbf{r} + \varphi_1)} \text{ and } \tilde{E}_2 = E_{02} \cdot e^{i(-\omega t + \mathbf{k}_2\mathbf{r} + \varphi_2)}. \quad (203)$$

This method naturally concludes the same as (201), but leaves time averaging hidden from us. Depicting the fields in the complex plane, the phenomenon can be interpreted as follows:

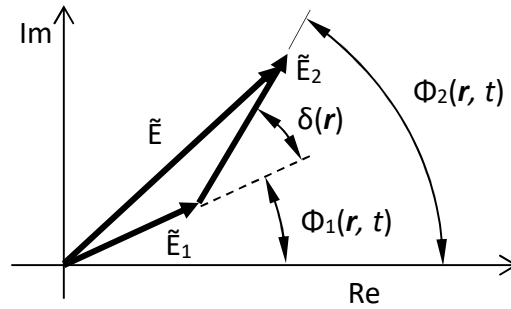


Fig. 28 Summation of phasors.

Through $\Delta\varphi$ the phase difference can be extended by either optical path difference (*OPD*), or temporal delay (Δt):

$$OPD \triangleq OPL_1 - OPL_2 ; \quad \Delta\varphi = \frac{2\pi}{k_0} OPD ; \quad \Delta t = t_1 - t_2 ; \quad \Delta\varphi = \omega\Delta t. \quad (204)$$

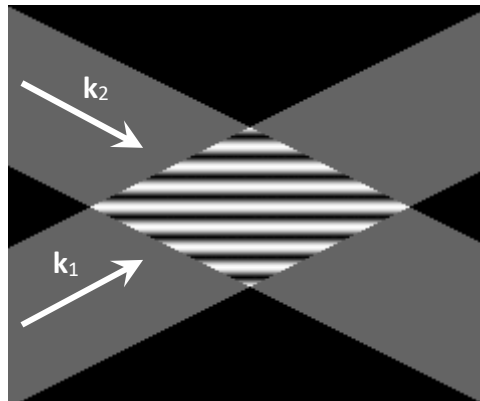


Fig. 29 Interference of plane waves in geometrical optical approximation.

Determining the period of a fringe pattern

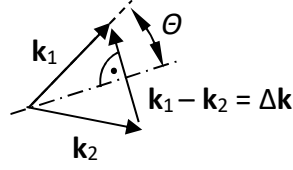


Fig. 30 Determination of the wave vector ($\Delta\mathbf{k}$) of an interferogram.

Due to (199) the formal wave vector of any fringe pattern can be considered as if it were $\Delta\mathbf{k} = \mathbf{k}_1 - \mathbf{k}_2$. From this the fringe period (Λ) can be easily obtained:

$$\Delta k = \frac{2\pi}{\Lambda} \Rightarrow \Lambda = \frac{2\pi}{\Delta k} = \frac{2\pi}{|\mathbf{k}_1 - \mathbf{k}_2|} = \frac{2\pi}{2 \cdot \frac{2\pi}{\lambda} \sin \theta} = \frac{\lambda}{2 \sin \theta} = \frac{\lambda_0}{2 \cdot n \sin \theta}. \quad (205)$$

7.3. Visibility of interference fringes

Let the intensities of two beams be I_1 and I_2 respectively. Then (201) becomes the following:

$$I_1 = \frac{v\varepsilon}{2} E_{01}^2; \quad I_2 = \frac{v\varepsilon}{2} E_{02}^2 \Rightarrow I(\mathbf{r}) = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos(\delta(\mathbf{r})). \quad (206)$$

We depicted the $I(\delta)$ function (i.e. the interferogram) in the diagram below.

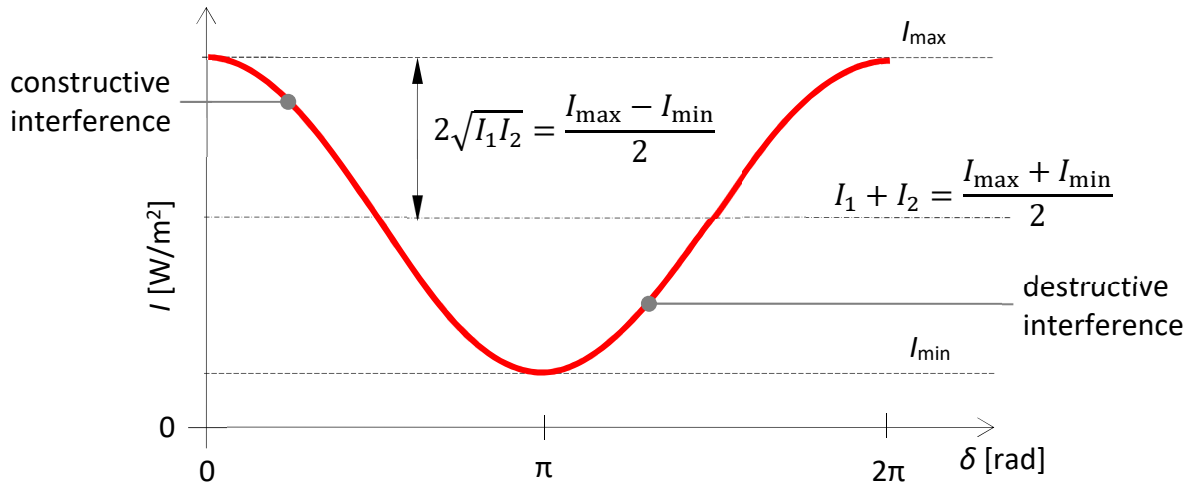


Fig. 31 Demonstration of an interferogram as a function of phase difference (δ).

The strength of interference phenomena are characterized by the so-called *visibility*. The definition of visibility is as follows, and can be calculated from (206):

$$V \triangleq \frac{(I_{\max} - I_{\min})/2}{(I_{\max} + I_{\min})/2} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2}. \quad (207)$$

$V_{\max} = 1$, if $I_1 = I_2$. In this case:

$$I(\mathbf{r}) = 2I_1 \cdot (1 + \cos(\delta(\mathbf{r}))). \quad (208)$$

This form will become of great significance later, when discussing statistical optics.

7.4. Example: plane-parallel plate

We represent the incident, transmitted and reflected plane waves by light rays respectively.

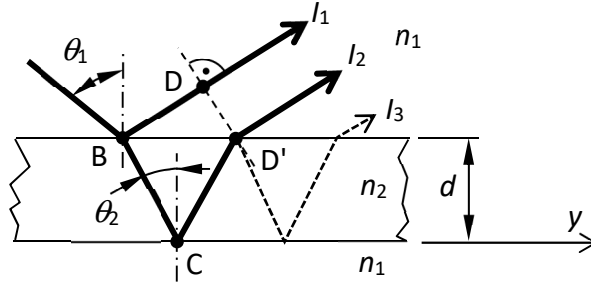


Fig. 32 Treating the interference of a plane-parallel plate in two-beam approximation.

Condition: low reflectance: $R \ll 1$ (e.g. for air-glass 4% in case of normal incidence). Then:

$$I_1 = I_I R \quad ; \quad I_2 = I_I (1 - R) \frac{\cos \theta_1}{\cos \theta_2} R (1 - R) \frac{\cos \theta_2}{\cos \theta_1} = I_I R (1 - R)^2 \approx I_I R$$

$$I_3 = I_I (1 - R) \frac{\cos \theta_1}{\cos \theta_2} R R R (1 - R) \frac{\cos \theta_2}{\cos \theta_1} = I_I R^3 (1 - R)^2 \approx I_I R^3, \quad (209)$$

where we also employed expression (92) describing the connection between intensity and transmittance for oblique incidence. Based on (209) $I_1 \approx I_2$, and $I_3 \ll I_{1,2}$, thus the interference phenomenon can be discussed in two-beam approximation. The phase difference between the two plane waves reflected by the surfaces can be calculated by the optical path length:

$$OPD_{BD'-BD} = n_2 \frac{2d}{\cos(\theta_2)} - n_1 \cdot 2d \cdot \sin(\theta_1) \tan(\theta_2) = 2d \cdot n_2 \cos(\theta_2). \quad (210)$$

Since in case of $n_2 > n_1$ the electromagnetic field undergoes a π phase jump at point B upon reflection (in the reversed case it happens at point C), the phase shift between $BD'-BD$ is:

$$\delta = OPD_{BD'-BD} \cdot \frac{2\pi}{\lambda_0} \pm \pi = \frac{4\pi \cdot n_2 d}{\lambda_0} \cos \theta_2 \pm \pi. \quad (211)$$

The above is a simplified treatment of phase delays caused by surfaces. In the general case we should write an equation such as (202) for the interference, including the occasionally complex value of the ρ reflection coefficient (see e.g. metals or total internal reflection).

Based on (206) we see constructive interference (increased reflection) if $\delta = m \cdot 2\pi$, $m = 1, 2, \dots$. The interference is destructive (reflection reduces), if $\delta = m \cdot 2\pi + \pi$, $m = 0, 1, 2, \dots$. For normal incidence, in case of $d \rightarrow 0$ we have $\delta \rightarrow \pi$, i.e. the reflection is zero, the plate “disappears”.

Haidinger fringes: if $d(y) = \text{const.}$, the interference pattern is θ_2 direction dependent.

Newton rings: if $\theta_2 = \text{const.}$, the interference pattern is y -position dependent through $d(y)$.

The temporal-coherence condition of interference

Interference occurs in case of the superposition of temporally coherent or partially coherent beams – in the latter case the width of the frequency spectrum is not of zero but finite (see Chapters 11-12). Since beams of different frequency cause beats, stationary interference can only happen between the ω angular frequency spectral components of one beam and the same frequency components of the other beam. The phase difference between the interfering wave components depends on the angular frequency as follows:

$$\delta(\mathbf{r}, \omega) = \Delta \mathbf{k} \cdot \mathbf{r} + \Delta \varphi = \frac{2\pi}{\lambda} \Delta \hat{\mathbf{s}} \cdot \mathbf{r} + \Delta \varphi = \frac{n\omega}{c} \Delta \hat{\mathbf{s}} \cdot \mathbf{r} + \omega \Delta t, \quad (212)$$

where $\Delta \hat{\mathbf{s}}$ is the difference of the unit direction vectors of the waves, Δt is the time delay between them. Let the two beams propagate in the same direction ($\Delta \hat{\mathbf{s}} = 0$), then $\delta(\omega) = \omega \cdot \Delta t$. Increasing the time difference we can see interference while the following condition is satisfied for all ω components of the spectrum (spreading from ω_1 to ω_2) (see relationship (204)):

$$2\pi \gg \delta(\omega_1) - \delta(\omega_2) = \omega_1 \Delta t - \omega_2 \Delta t = \Delta \omega \Delta t \rightarrow \Delta \nu \Delta t \ll 1. \quad (213)$$

In case of equality:

$$\tau_c \triangleq \Delta t \rightarrow \tau_c = \frac{1}{\Delta \nu} \rightarrow \Delta t \ll \tau_c \text{ v. } OPD \ll L_L = c \cdot \tau_c \quad (214)$$

where L_L denotes the *longitudinal coherence length*. This phenomenon will be discussed in detail later in Chapters 11-12 dealing with statistical optics.

7.5. What happens to light in two-beam interferometers?

The power carried by light cannot be made disappear by interference, only its spatial or temporal distribution can be changed or rearranged! (Law of energy conservation.)

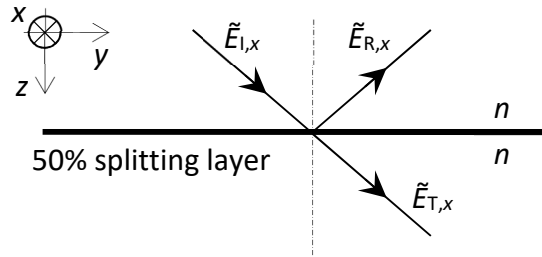


Fig. 33 Field amplitudes of a plane wave transmitting/reflecting on a 50% beam splitter layer.

Let us examine the effect of 50% intensity-splitting ratio surfaces used in most interferometers on the field vector. For simplicity we only consider the x-component of the \mathbf{E} vector being normal to the plane of incidence, which corresponds to “s” polarization. So, if the incident light (E_I) is polarized in the x-direction, after transmission (E_T) and reflection (E_R) its polarization state will not change. The analysis can be performed for the other field vector components analogously. The field at hand is monochromatic, therefore we use complex formalism for its description, discarding the temporally dependent factor $e^{-i\omega t}$. Corresponding to Fig. 33, in case of an 50% spitting layer being in a medium of refractive index n , the below equations hold for the complex field amplitudes and intensities. Since the incident, transmitted and reflected fields are complex quantities (six parameters), and from these the two parameters of the incident field are given (i.e. we have four independent parameters), we will need a total of four equations.

The sum of intensities is constant – conservation of energy:

$$|\tilde{E}_{I,x}|^2 = |\tilde{E}_{T,x}|^2 + |\tilde{E}_{R,x}|^2 \quad (215)$$

Boundary condition for the tangential component of $\tilde{\mathbf{E}}$ (two equations: real / imaginary parts):

$$\tilde{E}_{T,x} = \tilde{E}_{I,x} + \tilde{E}_{R,x} \quad (216)$$

50% intensity-splitting ratio:

$$|\tilde{E}_{T,x}|^2 = |\tilde{E}_{R,x}|^2 \quad (217)$$

From (215) and (216):

$$|\tilde{E}_{T,x}|^2 + |\tilde{E}_{R,x}|^2 = |\tilde{E}_{T,x} - \tilde{E}_{R,x}|^2 \rightarrow \tilde{E}_{T,x}\tilde{E}_{R,x}^* = -\tilde{E}_{T,x}^*\tilde{E}_{R,x} \rightarrow \tilde{E}_{T,x}^* = -\frac{\tilde{E}_{T,x}\tilde{E}_{R,x}^*}{\tilde{E}_{R,x}} \quad (218)$$

(217) corresponds to the fact that:

$$\tilde{E}_{T,x} \cdot \tilde{E}_{T,x}^* = \tilde{E}_{R,x} \cdot \tilde{E}_{R,x}^* \quad (219)$$

Substituting (218) into this we get the following:

$$\tilde{E}_{T,x}\tilde{E}_{T,x} = -\tilde{E}_{R,x}\tilde{E}_{R,x} \rightarrow \tilde{E}_{R,x} = \pm i \cdot \tilde{E}_{T,x}. \quad (220)$$

The value of the \pm sign is determined by the physical implementation of the splitting layer; we opt for the positive, this will not affect intensity values anyway. The above relationship implies that should we create a splitting layer of 50% by any means (any absorption-free dielectric mirror of appropriate thin-film structure can be used), there will always be exactly $\pi/2$ phase difference between the reflected and transmitted fields! From the above equations we can also derive:

$$\tilde{E}_{R,x} = \frac{1}{\sqrt{2}}e^{i\cdot 3/4\pi} \cdot \tilde{E}_{I,x} \text{ and } \tilde{E}_{T,x} = \frac{1}{\sqrt{2}}e^{i\cdot 1/4\pi} \cdot \tilde{E}_{I,x}. \quad (221)$$

The general formulae for an arbitrary R reflectance can be determined by transforming (217):

$$\left(|\tilde{E}_{I,x}|^2\right) = \frac{|\tilde{E}_{T,x}|^2}{1-R} = \frac{|\tilde{E}_{R,x}|^2}{R}. \quad (222)$$

Let us examine now the interference in a Mach-Zehnder interferometer, where the arm "A" of the light path includes a phase retarding plate of $\Delta\varphi$.

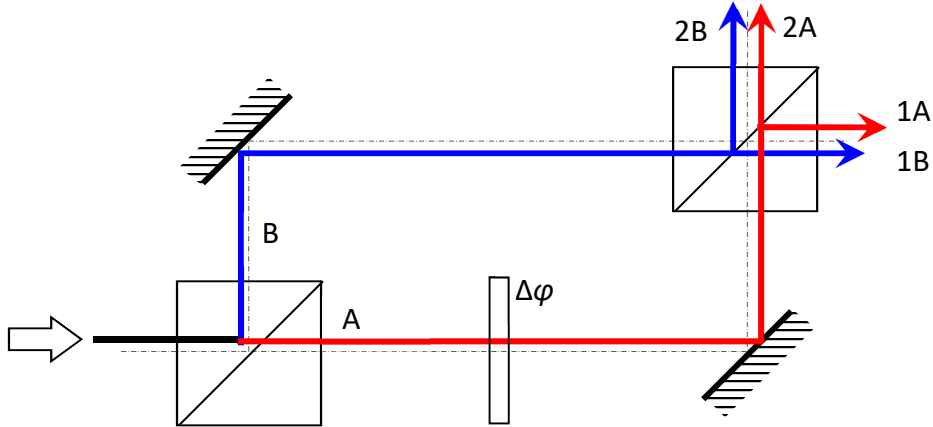


Fig. 34 Scheme of a Mach-Zehnder interferometer. A & B denotes beams after the first splitting element, and 1 & 2 denotes beams after the second splitting element.

If the input beam is \tilde{E}_I , than by (221) we can write \tilde{E}_1 and \tilde{E}_2 exit beams the following way (regarding the *OPLs* between the splitting layers of the prisms as equal, we omitted them):

$$\begin{aligned} \tilde{E}_1 &= \tilde{E}_{1A} + \tilde{E}_{1B} = \tilde{E}_I \frac{1}{\sqrt{2}}e^{i\cdot 1/4\pi} \frac{1}{\sqrt{2}}e^{i\cdot 3/4\pi}e^{i\Delta\varphi} + \tilde{E}_I \frac{1}{\sqrt{2}}e^{i\cdot 3/4\pi} \frac{1}{\sqrt{2}}e^{i\cdot 1/4\pi} \\ \tilde{E}_2 &= \tilde{E}_{2A} + \tilde{E}_{2B} = \tilde{E}_I \frac{1}{\sqrt{2}}e^{i\cdot 1/4\pi} \frac{1}{\sqrt{2}}e^{i\cdot 3/4\pi}e^{i\Delta\varphi} + \tilde{E}_I \frac{1}{\sqrt{2}}e^{i\cdot 3/4\pi} \frac{1}{\sqrt{2}}e^{i\cdot 3/4\pi} \end{aligned} \quad (223)$$

Simplifying the equations:

$$\begin{aligned}\tilde{E}_1 &= \frac{\tilde{E}_1}{2} e^{i\pi} e^{i\Delta\varphi} + \frac{\tilde{E}_1}{2} e^{i\pi} = -\frac{\tilde{E}_1}{2} (e^{i\Delta\varphi} + 1) \\ \tilde{E}_2 &= \frac{\tilde{E}_1}{2} e^{i\frac{1}{2}\pi} e^{i\Delta\varphi} + \frac{\tilde{E}_1}{2} e^{-i\frac{1}{2}\pi} = i \frac{\tilde{E}_1}{2} (e^{i\Delta\varphi} - 1)\end{aligned}\quad (224)$$

From these the intensities:

$$\begin{aligned}I_1 &= \frac{v\varepsilon}{2} |\tilde{E}_1|^2 = \frac{v\varepsilon}{2} \frac{|\tilde{E}_1|^2}{4} (1 + 1 + e^{i\Delta\varphi} + e^{-i\Delta\varphi}) = \frac{I_1}{4} (2 + e^{i\Delta\varphi} + e^{-i\Delta\varphi}) \\ I_2 &= \frac{v\varepsilon}{2} |\tilde{E}_2|^2 = \frac{v\varepsilon}{2} \frac{|\tilde{E}_1|^2}{4} (1 + 1 - e^{i\Delta\varphi} - e^{-i\Delta\varphi}) = \frac{I_1}{4} (2 - e^{i\Delta\varphi} - e^{-i\Delta\varphi})\end{aligned}\quad (225)$$

where I_1 is the intensity of the illuminating beam. From the above equations the final result is:

$$I_1 = \frac{I_1}{2} (1 + \cos \Delta\varphi) \quad \text{and} \quad I_2 = \frac{I_1}{2} (1 - \cos \Delta\varphi). \quad (226)$$

This means that the intensity variations of the two beams we get after the interferometer always have a phase difference of π , thus beams 1 and 2 are *always* in opposite phase. In other words, in case of the Mach-Zehnder interferometer (or any other type) light goes either into one of the exiting beams or into the other, driven by the path difference ($\Delta\varphi$ phase difference). If $\Delta\varphi = 0$, then all light goes into beam 1. Likewise, for a Michelson's the light either leaves the interferometer or becomes reflected back into the light source.

7.6. Large-angle interference (supplementary)

According to equation (56) of Chapter 2, the time average of the Poynting vector of a plane wave propagating in dielectric medium ($\hat{\mathbf{s}}$ denotes the unit vector pointing in the direction of the wavefront normal):

$$\langle \mathbf{S}(\mathbf{r}, T) \rangle = \frac{1}{2} \text{Re}\{\tilde{\mathbf{E}} \times \tilde{\mathbf{H}}^*\} = \frac{1}{2\mu\omega} \text{Re}\{\tilde{\mathbf{E}} \times (\mathbf{k} \times \tilde{\mathbf{E}}^*)\} = \frac{v\varepsilon}{2} \tilde{\mathbf{E}} \times (\hat{\mathbf{s}} \times \tilde{\mathbf{E}}^*). \quad (227)$$

Since \mathbf{E} and \mathbf{H} are always in phase in dielectrics, we can leave the real part operation, also we do not indicate the spatial dependence of field vectors either. We examine the interference of the following two harmonic plane waves that make an arbitrary angle:

$$\tilde{\mathbf{E}}_1(\mathbf{r}) = \mathbf{E}_1 e^{i\Phi_1(\mathbf{r})} \quad \text{and} \quad \tilde{\mathbf{E}}_2(\mathbf{r}) = \mathbf{E}_2 e^{i\Phi_2(\mathbf{r})}. \quad (228)$$

Let the electric fields lie in the x - z plane (p-polarization). The same analysis can be performed for s-polarization analogously to what is presented below.

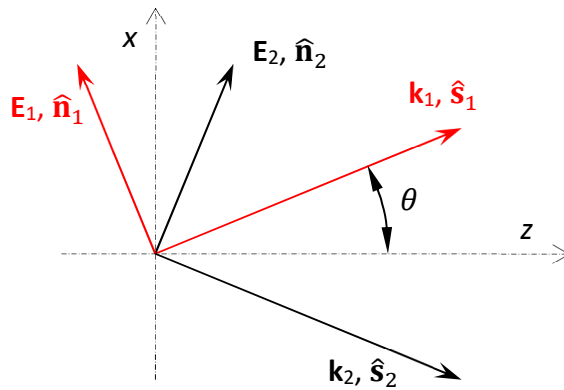


Fig. 35 Demonstration of large-angle interference.

Inserting the sum of the fields of the two plane waves (i.e. $\mathbf{E}_1 + \mathbf{E}_2$ and $\mathbf{H}_1 + \mathbf{H}_2$) into (227):

$$\langle \mathbf{S}(\mathbf{r}, T) \rangle = \frac{v\varepsilon}{2} (\tilde{\mathbf{E}}_1 + \tilde{\mathbf{E}}_2) \times (\hat{\mathbf{s}}_1 \times \tilde{\mathbf{E}}_1^* + \hat{\mathbf{s}}_2 \times \tilde{\mathbf{E}}_2^*). \quad (229)$$

Expanding the parentheses:

$$\langle \mathbf{S}(\mathbf{r}, T) \rangle = \frac{v\varepsilon}{2} (\tilde{\mathbf{E}}_1 \times (\hat{\mathbf{s}}_1 \times \tilde{\mathbf{E}}_1^*) + \tilde{\mathbf{E}}_2 \times (\hat{\mathbf{s}}_2 \times \tilde{\mathbf{E}}_2^*) + \tilde{\mathbf{E}}_1 \times (\hat{\mathbf{s}}_2 \times \tilde{\mathbf{E}}_2^*) + \tilde{\mathbf{E}}_2 \times (\hat{\mathbf{s}}_1 \times \tilde{\mathbf{E}}_1^*)) \quad (230)$$

Expanding the vectorial products according to $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) \equiv (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}$:

$$\langle \mathbf{S}(\mathbf{r}, T) \rangle = \frac{v\varepsilon}{2} ((\tilde{\mathbf{E}}_1 \tilde{\mathbf{E}}_1^*)\hat{\mathbf{s}}_1 + (\tilde{\mathbf{E}}_2 \tilde{\mathbf{E}}_2^*)\hat{\mathbf{s}}_2 + (\tilde{\mathbf{E}}_1 \tilde{\mathbf{E}}_2^*)\hat{\mathbf{s}}_2 + (\tilde{\mathbf{E}}_2 \tilde{\mathbf{E}}_1^*)\hat{\mathbf{s}}_1 - (\tilde{\mathbf{E}}_1 \tilde{\mathbf{E}}_2^*)\hat{\mathbf{s}}_2 - (\tilde{\mathbf{E}}_2 \tilde{\mathbf{E}}_1^*)\hat{\mathbf{s}}_1). \quad (231)$$

At the product of identical components the term corresponding to $(\mathbf{a} \cdot \mathbf{b})\mathbf{c}$ is zero, since \mathbf{k} is normal to \mathbf{E} . At the cross-terms we cannot say the same, thus they remain in the expression. It is these two last terms that we neglect in paraxial approximation, where making use of $\hat{\mathbf{s}}_1$ being almost parallel to $\hat{\mathbf{s}}_2$, the terms can be added up as scalars. The intensity is then

$$I(\mathbf{r}) = \langle \mathbf{S}(\mathbf{r}, T) \rangle = \frac{v\varepsilon}{2} (\tilde{\mathbf{E}}_1 \tilde{\mathbf{E}}_1^* + \tilde{\mathbf{E}}_2 \tilde{\mathbf{E}}_2^* + \tilde{\mathbf{E}}_1 \tilde{\mathbf{E}}_2^* + \tilde{\mathbf{E}}_2 \tilde{\mathbf{E}}_1^*), \quad (232)$$

which is equivalent with (196) and (201) indeed. If $|\mathbf{E}_1| = |\mathbf{E}_2| = E$, but the two waves are not of the same direction and make an arbitrary angle 2θ , (231) becomes the following:

$$\begin{aligned} \langle \mathbf{S}(\mathbf{r}, T) \rangle = \frac{v\varepsilon}{2} (2E^2 \cos(\theta) \hat{\mathbf{s}} + 2E^2 \cos(\delta) \cos(2\theta) \cos(\theta) \hat{\mathbf{s}} + \\ + \tilde{\mathbf{E}}_1 \tilde{\mathbf{E}}_2^* \sin(2\theta) \hat{\mathbf{n}}_2 - \tilde{\mathbf{E}}_2 \tilde{\mathbf{E}}_1^* \sin(2\theta) \hat{\mathbf{n}}_1). \end{aligned} \quad (233)$$

Here $\hat{\mathbf{s}}$ is a unit vector pointing in the direction of the bisector of the two wave vectors, $\hat{\mathbf{n}}_1$ and $\hat{\mathbf{n}}_2$ unit vectors point in the direction of the field vectors. The difference of these latter two points just in the direction of $\hat{\mathbf{s}}$, thus:

$$\langle \mathbf{S}(\mathbf{r}, T) \rangle = \frac{v\varepsilon}{2} (2E^2 \cos(\theta) \hat{\mathbf{s}} + 2E^2 \cos(\delta) \cos(2\theta) \cos(\theta) \hat{\mathbf{s}} + 2E^2 \cos(\delta) \sin(2\theta) \sin(\theta) \hat{\mathbf{s}}) \quad (234)$$

In the above equation we denoted phase difference as usual: $\delta \triangleq \Phi_1 - \Phi_2$. Transforming it:

$$\langle \mathbf{S}(\mathbf{r}, T) \rangle = v\varepsilon E^2 (\cos(\theta) \hat{\mathbf{s}} + \cos(\delta) \cos(2\theta) \cos(\theta) \hat{\mathbf{s}} + \cos(\delta) \sin(2\theta) \sin(\theta) \hat{\mathbf{s}}), \quad (235)$$

which gets simplified a lot by using trigonometric identities:

$$\langle \mathbf{S}(\mathbf{r}, T) \rangle = v\varepsilon E^2 (1 + \cos(\delta)) \cos(\theta) \hat{\mathbf{s}} \quad (236)$$

This relationship shows, that the power propagates in the direction of the bisector, while the magnitude of $\langle \mathbf{S} \rangle$ is the intensity:

$$I(\mathbf{r}) = v\varepsilon E^2 (1 + \cos(\delta(\mathbf{r}))) \cos(\theta). \quad (237)$$

Interpreting the above formula is quite straightforward: the larger the 2θ angle the beams make relative to each other, the less power propagates in the direction of the bisector, therefore intensity decreases by a cosine function. Since the derivation gives the same results for both s- and p-polarization, relationship (237) is valid for any two waves of arbitrary, but identical polarization. When the propagation direction of the two waves is $2\theta = 180^\circ$, i.e. they propagate opposite to each other and a perfect standing wave takes shape: the field still oscillates in a harmonic manner, but there are nodes and antinodes, and energy does not propagate in any direction.

8. MULTIPLE-BEAM INTERFERENCE

8.1. Interference of “N” plane waves

Should we examine the interference of not two but many plane waves of identical ω angular frequency (still in paraxial, scalar approximation), the resultant field (192) will become the following:

$$\tilde{E} = \sum_{n=1}^N \tilde{E}_{0n} = \sum_{n=1}^N E_{0n} e^{i\Phi_n} \quad (238)$$

Since we have already seen the effect of time averaging at two-beam interference, in the following we will use the much more straightforward complex formalism.

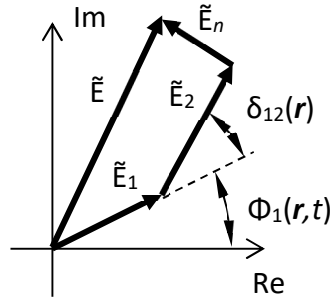


Fig. 36 Phasor summation in case of multiple-beam interference.

Assumptions: $\delta_{n+1,n} = \Phi_{n+1} - \Phi_n = \delta = \text{const.}$ and $E_{0n} = E_{01} = \text{const.}$ Then (238) can be written as:

$$\Phi_n = \Phi_1 + (n-1)\delta \rightarrow \tilde{E} = E_{01} e^{i\Phi_1} \cdot \sum_{n=1}^N e^{i\delta(n-1)} \quad (239)$$

This summation is a geometrical series, the sum formula of which is:

$$S_N = a_1 \frac{q^N - 1}{q - 1} ; a_1 = E_{01} e^{i\Phi_1} ; q = e^{i\delta} \quad (240)$$

$$\begin{aligned} \tilde{E} &= E_{01} e^{i\Phi_1} \cdot \frac{e^{i\delta N} - 1}{e^{i\delta} - 1} = E_{01} e^{i\Phi_1} \cdot \frac{e^{i\delta N/2} - 1}{e^{i\delta/2} - 1} \cdot \frac{e^{i\delta N/2} - e^{-i\delta N/2}}{e^{i\delta/2} - e^{-i\delta/2}} = \\ &= E_{01} e^{i\Phi_1} \cdot e^{\frac{i\delta(N-1)}{2}} \cdot \frac{\sin(\delta N/2)}{\sin(\delta/2)} \end{aligned} \quad (241)$$

The value of intensity is:

$$I = \langle S \rangle = \frac{v\varepsilon}{2} |\tilde{E}|^2 = v\varepsilon \frac{E_{01}^2}{2} \left(\frac{\sin(\delta N/2)}{\sin(\delta/2)} \right)^2, \quad (242)$$

that is:

$$I(\delta) = I_1 \left(\frac{\sin(\delta N/2)}{\sin(\delta/2)} \right)^2 ; m \triangleq \frac{\delta}{2\pi}. \quad (243)$$

The obtained interferogram can be seen in Fig. 37 for $N := 10$. We get constructive interference if: $\delta = m \cdot 2\pi$, where $m = 0, \pm 1, \pm 2, \dots$ – the large peaks are here with a value of $I_1 N^2$. The distance of local minima is $\Delta\delta = 2\pi/N$, there is a total of $N-1$ of them between two peaks.

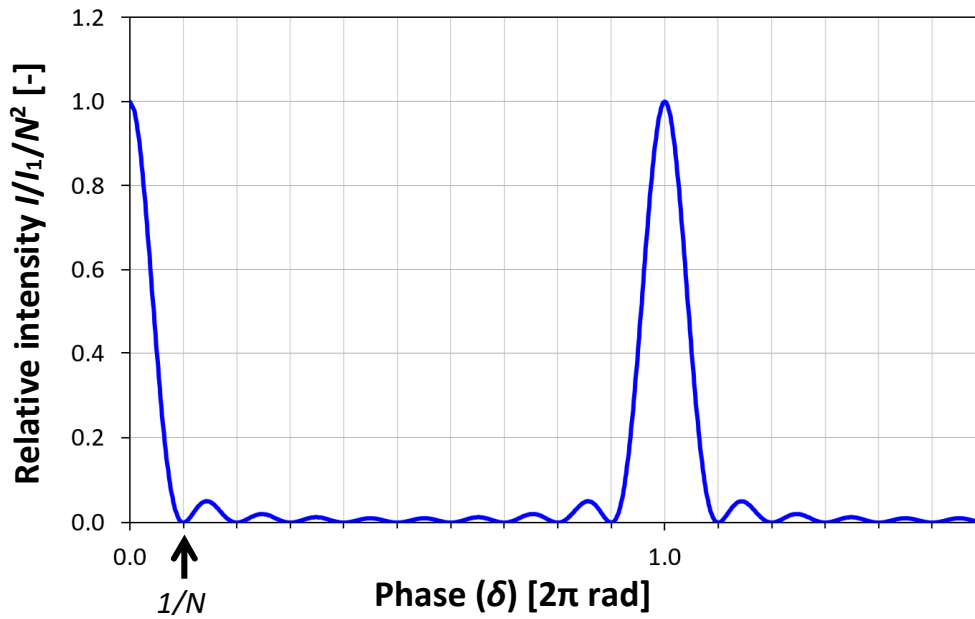


Fig. 37 Interferogram as a function of phase difference for $N = 10$ wave components.

Fig. 38 helps interpret the shape of the interferogram, showing the numerator and denominator of (241), for $N := 5$. It is clearly visible that the quotient gives zeros where $\delta/2 = \pi/N$, or an integer multiple thereof, and both the numerator and denominator are zero if $\delta/2 = \pi \cdot m$. In this latter case the quotient tends to N in a limiting case – at these points are the peaks. If N is large, the phase of the numerator changes much faster than that of the denominator, therefore in the vicinity of $\delta = 2\pi \cdot m$ the denominator can be linearly approximated: $\sin(\delta/2) \approx 2\pi \cdot m + \delta/2$, thus the quotient behaves like a $\sin(N \cdot x)/x$ function.

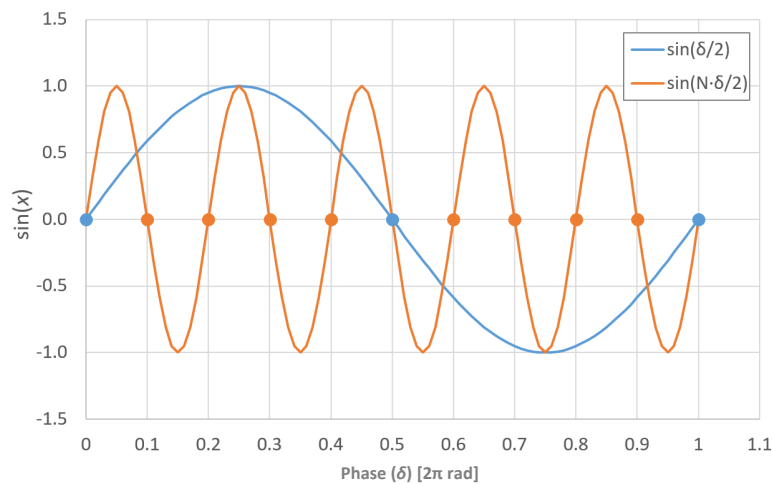


Fig. 38 Graphic interpretation of relations (241) and (242).

Examples for multiple-beam interference

- diffraction grating
- plane-parallel plate (if $R \approx 1$)
- Fabry-Perot interferometer (or etalon)
- thin-film structures (matrix description)
- temporally periodic pulses (see mode-coupled lasers)

8.2. Diffraction gratings

The construction of a simple diffraction grating can be seen in Fig. 39. The structure is one-dimensional, i.e. has no variations normal to the plane of the figure. The grating contains N slits in an opaque screen, whose

sizes are comparable to the wavelength (λ_0), thus practically cylindrical wavefronts leave them due to diffraction – we examine the interference of these *in the far field*.

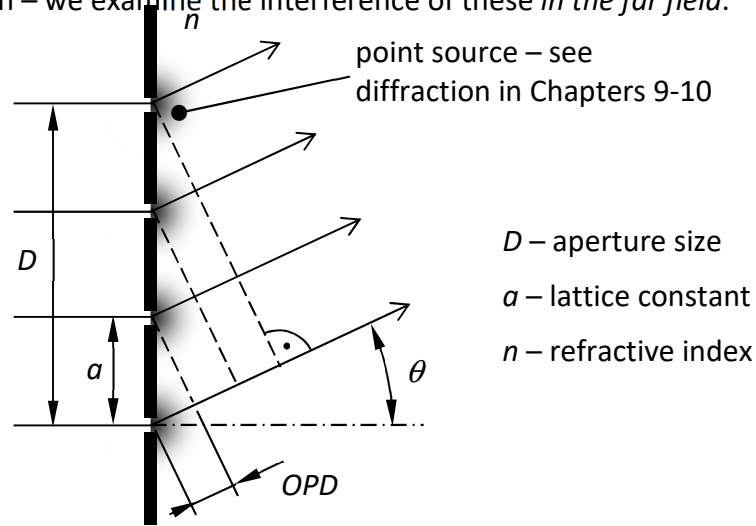


Fig. 39 Modelling a diffraction grating by using multiple-beam interference.

The task is to determine the phase difference (δ) of waves originated from neighboring point sources as they tend to infinity. If we know that, both conditions of the above-discussed multiple-beam interference hold, then the interferogram can be evaluated according to (243). In a given direction θ , at a very large distance from the grating the cylindrical wavefronts can be locally considered as plane waves. Projecting these wavefronts back to the grating (see dashed lines in the figure) the phase difference can be determined from the optical path difference between a slit and the wavefront corresponding to its neighbor:

$$\begin{aligned} OPD &= n \cdot a \cdot \sin(\theta) \\ \delta &= \frac{2\pi}{\lambda_0} OPD = \frac{2\pi}{\lambda_0} n \cdot a \cdot \sin(\theta) \end{aligned} \quad (244)$$

According to (243) constructive interference occurs in the following discrete directions:

$$\delta = m \cdot 2\pi \Rightarrow \sin(\theta_m) = m \frac{\lambda_0}{n \cdot a} = m \frac{\lambda}{a} \Leftrightarrow n \cdot \sin(\theta_m) = m \frac{\lambda_0}{a}, \quad (245)$$

where integer m denotes the serial number of the so-called *diffraction orders*. Note, that if we multiply by the refractive index, the optical direction cosine $n \cdot \sin(\theta)$ appears in (245) too.

In reality, the waves emanated by slits of the diffraction grating do not have uniform intensity in every θ direction, but it tends to zero as $\theta \rightarrow 90^\circ$ with a rate depending on the actual shape and size of the slits (in general on the spatial distribution of their transmission). Therefore, the intensity of diffraction orders also reduces gradually, as m increases. Besides, the number of diffraction orders is limited too: since the max. value of $\sin(\theta)$ is 1, the largest value of m is

$$m_{\max} = \text{floor} \left(\frac{a}{\lambda} \right), \quad (246)$$

where the $y = \text{floor}(x)$ function gives the integer closest to x while $y \leq x$; $x \in \mathbb{R}$.

For orders having larger serial number than this we get an evanescent wave, similarly to total internal reflection, i.e. the light penetrates a little into the medium behind the grating, but cannot propagate in it – every energy gets reflected from the grating. If $m_{\max} = 1$, then $a = \lambda$, i.e. a diffraction grating of lattice constant smaller than the wavelength reflects the incident radiation like a mirror (see. radar antenna, microwave oven door etc.).

If we examine the interference pattern not as a function of δ , but rather the direction $\sin(\theta)$, when $m \neq 0$ we encounter a wavelength dependence due to (245), i.e. a constructive interference can be observed in different directions for every wavelength: the larger the wavelength, the larger the difference between the direction of diffraction orders. This is exploited in spectrometers, e.g. for measuring of the discrete wavelengths of lasers, or absorption/emission spectra of materials.

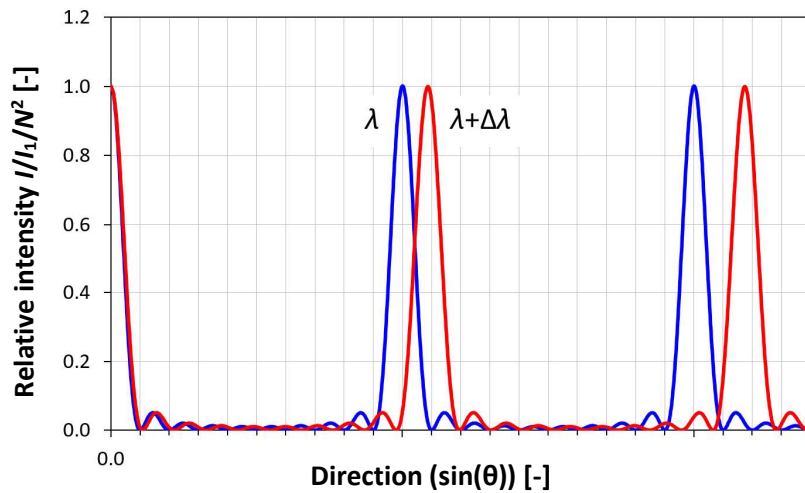


Fig. 40 Interferogram of a diffraction grating as a function of direction for $N := 10$ slits. In case of the red curve we increased the wavelength by $\Delta\lambda$ relative to the blue curve so that the Rayleigh criterion (see below) is exactly fulfilled.

Resolving power:

$$\mathcal{R} \triangleq \lambda / \Delta\lambda \quad (247)$$

The reason behind the above definition is that for a given grating this quantity is a wavelength-independent constant (see (243) and also below).

Rayleigh resolution criterion: The zero of $I(\lambda)$ coincides with the max. of $I(\lambda + \Delta\lambda)$, see Fig. 40.

Mathematically this means that:

$$\sin(\theta_m(\lambda)) + \Delta \sin(\theta_m(\lambda)) = \sin(\theta_m(\lambda + \Delta\lambda)), \quad (248)$$

where $\Delta \sin(\theta_m(\lambda))$ is the direction pointing at the first zero. From the condition \mathcal{R} can be determined:

$$\begin{aligned} \delta = m \cdot 2\pi &= \frac{2\pi}{\lambda} \cdot a \cdot \sin(\theta_m(\lambda)) \Rightarrow \sin(\theta_m(\lambda)) = \frac{m \cdot \lambda}{a} \\ \Delta\delta = \frac{2\pi}{N} &= \frac{2\pi}{\lambda} \cdot a \cdot \Delta \sin(\theta_m(\lambda)) \Rightarrow \Delta \sin(\theta_m(\lambda)) = \frac{\lambda}{a \cdot N} \\ \delta(\lambda + \Delta\lambda) &= m \cdot 2\pi \Rightarrow \sin(\theta_m(\lambda + \Delta\lambda)) = \frac{m \cdot (\lambda + \Delta\lambda)}{a} \end{aligned} \quad (249)$$

Solving the equation system:

$$\frac{\lambda}{\Delta\lambda} = \mathcal{R} = m \cdot N = m \cdot \frac{D}{a} . \quad (250)$$

That is, the larger the D aperture of the grating relative to the lattice constant (in other words the more slits are within the aperture), the higher the resolution, thus closer spectral lines can be resolved by the grating. The resolution also grows with m .

Technical application fields of interference

- spectroscopy
- measurement techniques
- holography
- interference filters
- lasers

9. SCALAR DIFFRACTION

Sources: [1], [6], [3], [9]

9.1. Basic concepts, diffraction models, Green's theorem

Diffraction: the phenomenon, when the propagation direction of light differs significantly from the one as suggested by geometrical optics (Sommerfeld). Light diffraction usually results in observable changes of light distribution if an obstacle of size comparable to the wavelength gets in the way of a radiation, or in general the properties of the medium (n , κ) vary over distances comparable to the wavelength.

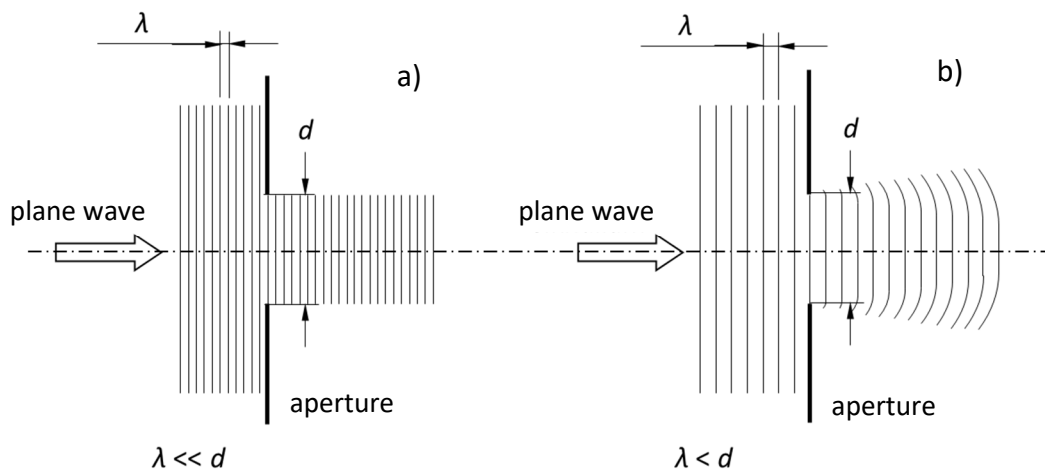


Fig. 41 Demonstration of geometrical (a) and diffraction (b) propagation of light.

Problem: If the complex field distribution is known across a surface, how can we determine the field in a point at a distance from this surface?

Conditions:

- 1) initial field distribution is given over a planar surface
- 2) phase difference between points of the surface are constant in time (spatial coherence)
- 3) illuminating beam is a monochromatic wave of ω angular frequency (temporal coherence)
- 4) scalar approximation: we consider only one field vector component, whose direction does not change significantly during light propagation (if the angle with respect to the main propagation direction is θ , then a standard rule of thumb is $\sin \theta < 0.6$ and $\cos \theta > 0.8$); we regard the components of different directions as independent of each other
- 5) changes of material properties in the propagation space can be neglected over distances of a wavelength (refractive index, absorption) \rightarrow wave equation can be used

Diffraction models

Field distributions taking shape via diffraction can be considered as the in-phase superposition of an infinite number of continuously distributed elementary wave components. For this it is practical to choose such wavelets, the propagation of which can be described in analytical form, and which satisfy the wave equation. Two basic methods exist giving equivalent results:

- decomposition into point sources (Huygens-Fresnel principle)
- decomposition into plane waves (Fourier optics)

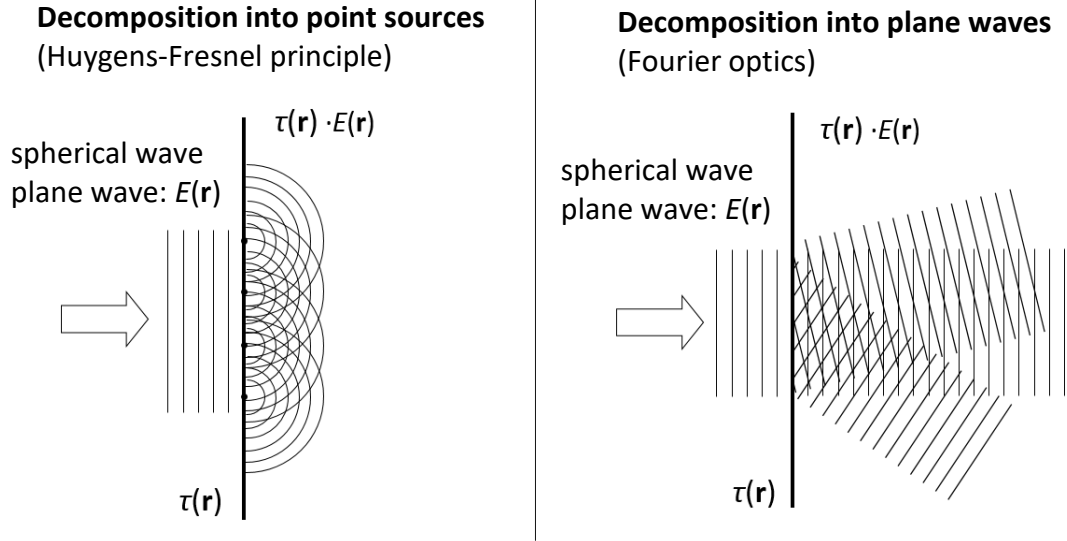


Fig. 42 Demonstration of the two best-known diffraction models.
 $\tau(\mathbf{r})$ denotes the complex transmission function interpreted at the surface.

In what follows we will discuss the decomposition to point sources.

Wave equation

The E scalar field is a complex quantity, and in our ansatz we will use the method of *positive phase propagation* (i.e. every phase term is multiplied by -1):

$$E(\mathbf{r}, t) := E(\mathbf{r})e^{-i\omega t}. \quad (251)$$

We begin our investigations from the scalar form of the time-independent wave equation ((28) Helmholtz equation), seeking its solution for specific boundary conditions:

$$\nabla^2 \mathbf{E}(\mathbf{r}, t) - \varepsilon\mu \frac{\partial^2 \mathbf{E}(\mathbf{r}, t)}{\partial t^2} = 0 \rightarrow \nabla^2 E(\mathbf{r}) + k^2 E(\mathbf{r}) = 0. \quad (252)$$

Green's theorem

The basis and starting point of the derivation is *Green's theorem* known from calculus: if we have two arbitrary complex functions given G and E , both of which are continuous as well as their first and second derivatives are also continuous over and inside a closed surface “A”, then the following expression holds:

$$\iiint_V (G \cdot \nabla^2 E - E \cdot \nabla^2 G) dV = \iint_A (G \cdot \hat{\mathbf{n}} \nabla E - E \cdot \hat{\mathbf{n}} \nabla G) dA, \quad (253)$$

where “V” is the volume enclosed by “A”. The surface integral scans through the points of “A”, which we denote by “P”. In these points the surface normal is $\hat{\mathbf{n}}$, whose directivity points always *outwards* from the examined volume “V” in accordance with the theorem.

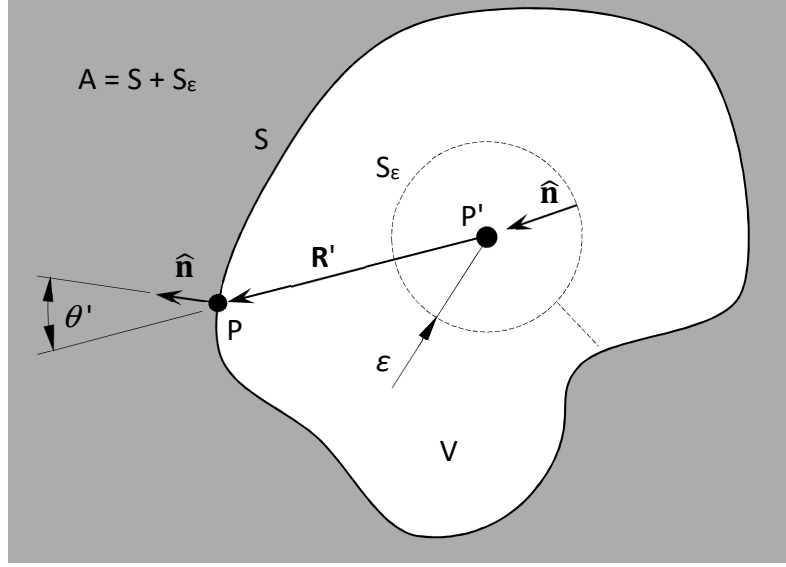


Fig. 43 Illustration of the notations used at formulating Green's theorem.

9.2. Integral theorem of Helmholtz and Kirchhoff

In case $E(P)$ is given over the surface "A", and we seek for the value of $E(P')$ at a point P' inside of it, then equation (253) needs further transformations. The below derivation serves as a basis to the diffraction integral, for the presentation of which we will follow Kirchhoff's method. (Previously Helmholtz performed the calculation for sound waves). Kirchhoff had chosen the so-called Green function G as follows:

$$G(P) := \frac{1}{R'} e^{ikR'}, \quad (254)$$

which is nothing else than a spherical wave of unity amplitude *starting* from point P' (propagation direction can be seen from the exponent sign). For this Helmholtz equation holds:

$$\nabla^2 G + k^2 G = 0 \quad (255)$$

Substituting equation (252) pertaining to E , and (255) into the left hand side of (253):

$$\iiint_V (G \cdot \nabla^2 E - E \cdot \nabla^2 G) dV = - \iiint_V (GE \cdot k^2 - EG \cdot k^2) dV = 0. \quad (256)$$

So, in case of complex E and G functions satisfying the homogeneous (source-free) wave equation Green's theorem reduces to this:

$$\iint_A (G \cdot \hat{n} \nabla E - E \cdot \hat{n} \nabla G) dA = 0. \quad (257)$$

G has a singularity at point P' , therefore it must be excluded from the investigation. For this reason it is practical to make surface "A" as a combination of two parts:

$$A = S + S_\epsilon, \quad (258)$$

where "S" is a surface enveloping the volume from the outside, and S_ϵ is a sphere of ϵ radius centered on point P' . With these (257) can be transformed in the following way:

$$- \iint_{S_\epsilon} (G \cdot \hat{n} \nabla E - E \cdot \hat{n} \nabla G) dA = \iint_S (G \cdot \hat{n} \nabla E - E \cdot \hat{n} \nabla G) dA. \quad (259)$$

Applying fraction derivation and the chain rule, the derivative of function G in direction $\hat{\mathbf{n}}$ is:

$$\begin{aligned}\hat{\mathbf{n}}\nabla G &= \hat{\mathbf{n}} \cdot \text{grad} \left(\frac{e^{ikR'}}{R'} \right) = \hat{\mathbf{n}} \cdot \frac{\partial}{\partial R'} \left(\frac{e^{ikR'}}{R'} \right) \cdot \text{grad} R' = \\ &= \cos \theta' \left(ik \frac{e^{ikR'}}{R'} - \frac{e^{ikR'}}{R'^2} \right) = \cos \theta' \cdot \left(ik - \frac{1}{R'} \right) \cdot \frac{e^{ikR'}}{R'}.\end{aligned}\quad (260)$$

The following statements are valid on surface S_ε :

$$\cos(\theta') = -1 ; \quad G(P) = \frac{e^{ik\varepsilon}}{\varepsilon} ; \quad \hat{\mathbf{n}}\nabla G = \left(\frac{1}{\varepsilon} - ik \right) \cdot \frac{e^{ik\varepsilon}}{\varepsilon}.\quad (261)$$

Taking these into account, and if $\varepsilon \rightarrow 0$, then the left hand side of (259) leaves us:

$$-\iint_{S_\varepsilon} (G \cdot \hat{\mathbf{n}}\nabla E - E \cdot \hat{\mathbf{n}}\nabla G) dA = -4\pi\varepsilon^2 \left(\frac{e^{ik\varepsilon}}{\varepsilon} \cdot \hat{\mathbf{n}}\nabla E - E \cdot \left(\frac{1}{\varepsilon} - ik \right) \cdot \frac{e^{ik\varepsilon}}{\varepsilon} \right) \rightarrow 4\pi E(P') \quad (262)$$

The area integration can be performed quite an easy way, since E and its derivative are continuous inside the examined volume, thus at P' too, and as ε tends to zero, the value of E over the surface S_ε can be regarded as constant. Substituting (262) into (259) we obtain the integral theorem of Helmholtz and Kirchhoff, which is nothing else than the unknown expression for $E(P')$, where $E(P)$ is given on surface “ S ”:

$$E(P') = \frac{1}{4\pi} \iint_S \left(\frac{e^{ikR'}}{R'} \cdot \hat{\mathbf{n}}\nabla E - E \cdot \hat{\mathbf{n}}\nabla \left(\frac{e^{ikR'}}{R'} \right) \right) dA.\quad (263)$$

9.3. Fresnel-Kirchhoff diffraction integral

In optics it is practical to determine relationship (263) for an arrangement presented in Fig. 44, where we consider the $E(P)$ field distribution caused by the illumination coming from the left as given along the plane surface of an aperture (opening) denoted by Σ .

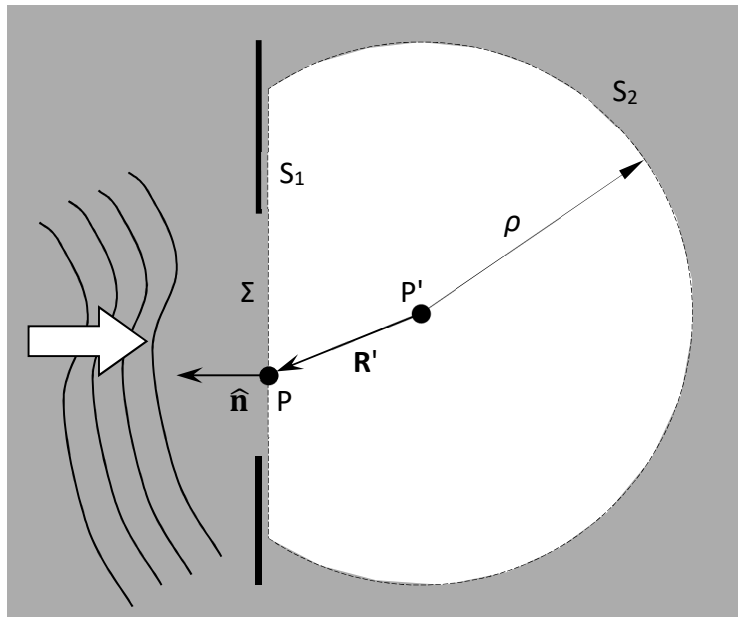


Fig. 44 Illustration of the notations used at formulating the Fresnel-Kirchhoff diffraction integral.

We perform the integration over the “S” closed surface that consists of two parts now:

$$S = S_1 + S_2. \quad (264)$$

It can be shown, if $\rho \rightarrow \infty$, and $E(P)$ vanishes on S_2 at least as rapidly as a spherical wave (*Sommerfeld radiation condition*), then the integral tends to zero on S_2 :

$$\iint_{S_2} \left(\frac{e^{ikR'}}{R'} \cdot \hat{\mathbf{n}} \nabla E - E \cdot \hat{\mathbf{n}} \nabla \left(\frac{e^{ikR'}}{R'} \right) \right) dA \rightarrow 0, \quad (265)$$

thus (263) leaves us the next:

$$E(P') = \frac{1}{4\pi} \iint_{S_1} \left(\frac{e^{ikR'}}{R'} \cdot \hat{\mathbf{n}} \nabla E - E \cdot \hat{\mathbf{n}} \nabla \left(\frac{e^{ikR'}}{R'} \right) \right) dA. \quad (266)$$

In order to avoid integrating over the infinite plane surface of S_1 , Kirchhoff set the following boundary conditions:

- 1) inside aperture Σ : $E(P) = \text{unchanged}$
- 2) outside aperture Σ : $E(P) = 0$ and $|\text{grad}(E)| = 0$

Owing to these conditions the integration has only to be performed over the aperture Σ :

$$E(P') = \frac{1}{4\pi} \iint_{\Sigma} \left(\frac{e^{ikR'}}{R'} \cdot \hat{\mathbf{n}} \nabla E - E \cdot \hat{\mathbf{n}} \nabla \left(\frac{e^{ikR'}}{R'} \right) \right) dA. \quad (267)$$

Contrary to their simplicity and advantages, both conditions pose fundamental problems. To have a field that is identical inside Σ with or without an aperture is a physical nonsense, just as it is to have $E(P)$ a zero value outside the aperture. For no matter what kind of absorbing medium we place in a radiative field, that will always alter the field distribution. However, the effect can be generally neglected if $\Sigma > \lambda$. As for the other problem, it can be mathematically proven that a continuous function having zero value and zero derivative over a finite interval must be constant zero everywhere. A resolution to the mentioned inconsistencies will be given during MSc in the Physical Optics course, where the Green function will be defined not by using (254) but in a more complex way. The result will be the Rayleigh-Sommerfeld diffraction integral, which is not significantly more precise than the one to be presented below originating from (267), only has a mathematically correct derivation.

Let us determine the Green function gradient in (267) based on (260), using the following simplifying assumption:

$$\frac{1}{R'} \ll k \Rightarrow \frac{\lambda}{2\pi} \ll R'. \quad (268)$$

Since k has quite a large value, and through scalar approximation we have already assumed that $\sin \theta' < 0.6$, i.e. size of $\Sigma < R'$, therefore condition (268) must hold anyways. By this

$$\hat{\mathbf{n}} \nabla G = \cos \theta' \cdot \left(ik - \frac{1}{R'} \right) \cdot \frac{e^{ikR'}}{R'} \approx ik \cdot \cos \theta' \cdot \frac{e^{ikR'}}{R'}, \quad (269)$$

thus the surface integral leaves us the next:

$$E(P') = \frac{1}{4\pi} \iint_{\Sigma} \frac{e^{ikR'}}{R'} (\hat{\mathbf{n}} \nabla E - E \cdot ik \cdot \cos(\theta')) dA \quad (270)$$

Keep in mind, that θ' is a signed quantity. The last step for us is to specify the radiation that illuminates aperture Σ , i.e. $E(P)$. For simplicity let it be a point source (Q), positioned to the left from the aperture.

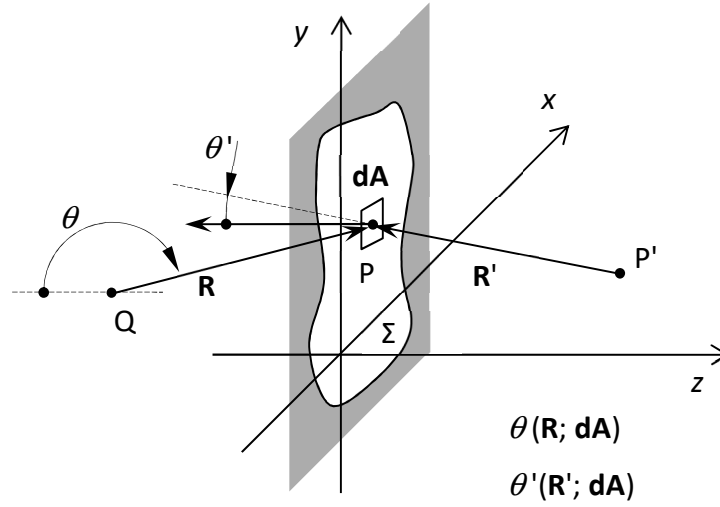


Fig. 45 Interpretation of the parameters of point source Q illuminating aperture Σ .

Point source Q and point P of the aperture span a distance R , the radiation amplitude is E_0 :

$$E(P) := \frac{E_0}{R} e^{ikR}. \quad (271)$$

We can apply condition (268) again (since size of $\Sigma < R$ due to the scalar approximation), thus:

$$\frac{1}{R} \ll k \Rightarrow \frac{\lambda}{2\pi} \ll R, \quad (272)$$

by which the gradient of E is:

$$\hat{n} \nabla E = \cos(\theta) \cdot \left(ik - \frac{1}{R} \right) \cdot \frac{e^{ikR}}{R} E_0 \approx ik \cdot \cos(\theta) \cdot \frac{e^{ikR}}{R} E_0, \quad (273)$$

Substituting this into (270) we obtain the *Fresnel-Kirchhoff diffraction integral*:

$$E(P') = \frac{i}{\lambda} \iint_{\Sigma} E_0 \frac{e^{ikR}}{R} \cdot \frac{e^{ikR'}}{R'} \cdot \frac{\cos \theta - \cos \theta'}{2} dA. \quad (274)$$

Proper interpretation of this expression first requires noting that mathematically we cannot distinguish between the phase of a point source being in P' and propagating toward P , and that of a point source being in P and propagating toward P' . Consequently, we can regard (274) as if the electric field in point P' were expressed as a sum of elementary spherical waves coming from aperture Σ , similarly to the statement of the Huygens-Fresnel principle. The key difference being that here the wavelets should be considered with cosine-direction characteristics, and a constant factor of i/λ appears. It is also worth noting how symmetric the equation is: a source placed in Q creates the same field in P' as a source placed in P' would produce in Q . This is the *Helmholtz reciprocity theorem*. Despite the mentioned mathematical inconsistencies (274) can be used surprisingly well, and usually provides rather precise results.

Integral (274) is still hard to evaluate (both analytically and numerically). In order to simplify the formula we introduce further approximations. Via these below we will arrive at the Huygens-Fresnel principle, the Fresnel, and then the Fraunhofer approximation.

10. APPROXIMATIONS TO THE FRESNEL-KIRCHHOFF INTEGRAL

10.1. Approximation in the near field (Huygens-Fresnel integral)

On behalf of scalar approximation we already required R and R' to be significantly larger than the characteristic dimension of aperture Σ . Hence, we can apply the below approximations:

$$\cos(\theta) \approx -1 \text{ and } \cos(\theta') \approx 1, \quad (275)$$

which is valid if $\sin \theta' < 0.5$; $\cos \theta' > 0.87$. The Fresnel-Kirchhoff integral thus leaves us:

$$E(P') = -\frac{i}{\lambda} \iint_{\Sigma} E_0 \frac{e^{ikR}}{R} \cdot \frac{e^{ikR'}}{R'} dA = \frac{1}{i\lambda} \iint_{\Sigma} E_0 \frac{e^{ikR}}{R} \cdot \frac{e^{ikR'}}{R'} dA \quad (276)$$

In this relationship the EM radiation that illuminates the aperture is composed of one single point source (Q). Since the $E(P)$ field distribution can be expressed as a superposition of radiations coming from several point sources, the above formula can be easily generalized:

$$E(P') = \frac{1}{i\lambda} \iint_{\Sigma} E(P) \frac{e^{ikR'}}{R'} dA. \quad (277)$$

When using this, one must take care only to get the $\Sigma < R$ condition, required by scalar diffraction, satisfied for each point source that form field $E(P)$. Save for a complex constant, approximation (277) corresponds to the Huygens-Fresnel principle.

10.2. Paraxial approximation (Fresnel diffraction)

In case we move P' away from the aperture by an adequate distance, as well as stay close to the z -axis both on aperture Σ and the examination screen, then we can make the approximation of $R' \approx z'$. As we will see, relationship (277) can be further simplified by this, at the same time this approximation also poses the first serious limitation for the applicability of the diffraction integral. For the sake of simplicity let point P' be on the z -axis, and P at the edge of the diffracting aperture. Then the above approximation can be expressed the following way:

$$z' \gg R' - z' = \sqrt{(z' \cdot \text{tg}(\theta'))^2 + z'^2} - z', \quad (278)$$

which becomes this after rearrangement and simplification:

$$\text{tg}^2(\theta') \ll 3 \Rightarrow \sin^2 \theta' \ll \frac{3}{4} \Rightarrow \sin \theta' \ll \sqrt{3/4} \rightarrow \sin \theta' \lesssim \sqrt{3/4}/10 = 0.09. \quad (279)$$

Accordingly, the condition will be $\sin(\theta') \lesssim 0.09$, so we entered the realm of *paraxial approximation*. In such cases the approximation $R' \approx z'$ can be directly used in the denominator of (277), but not in the exponent. This is because the phase changes very rapidly as a function of R' , therefore we apply its first-order Taylor series expansion instead:

$$\sqrt{1+b} \approx 1 + \frac{1}{2}b - \frac{1}{8}b^2 + \dots \quad (280)$$

Hence R' can be written as:

$$R' = \sqrt{z'^2 + (x' - x)^2 + (y' - y)^2} = z' \sqrt{1 + \left(\frac{x' - x}{z'}\right)^2 + \left(\frac{y' - y}{z'}\right)^2} \approx$$

$$\approx z' \left(1 + \frac{1}{2} \left(\frac{x' - x}{z'} \right)^2 + \frac{1}{2} \left(\frac{y' - y}{z'} \right)^2 \right) = z' + \frac{(x' - x)^2}{2z'} + \frac{(y' - y)^2}{2z'} \quad (281)$$

This approximation corresponds to considering the Huygens wavelets as parabolic surfaces of z' radius of curvature (see Fig. 46), which is only valid if the phase shift caused by quadratic or higher-order terms in (280) can be neglected:

$$\frac{2\pi z'}{\lambda} \left(\left(\frac{x' - x}{z'} \right)^2 + \left(\frac{y' - y}{z'} \right)^2 \right) \ll \pi \text{ [rad]} \Rightarrow \frac{1}{\sqrt[3]{4\lambda}} ((x' - x)^2 + (y' - y)^2)^{\frac{2}{3}} \ll z' \quad (282)$$

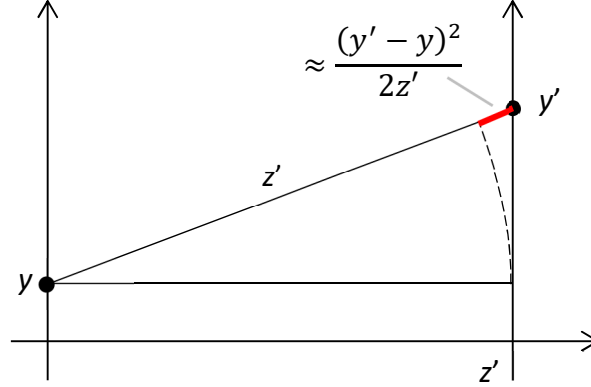


Fig. 46 Demonstration of the parabolic wavefronts occurring in the Fresnel integral.

This brings us to the *Fresnel approximation* of the diffraction integral expressed in the form of a convolution:

$$E(x', y') = \frac{e^{ikz'}}{i\lambda z'} \iint_{\Sigma} E(x, y) \cdot e^{i\frac{k}{2z'}((x' - x)^2 + (y' - y)^2)} dx dy = \frac{e^{ikz'}}{i\lambda z'} E(x, y) * e^{i\frac{k}{2z'}(x^2 + y^2)} \quad (283)$$

According to condition (282) this can be used when point P' is farther from the aperture Σ by a certain distance. E.g. for a laser beam of $\varnothing 1.0$ mm diameter and 633 nm wavelength $z' \gg 2.9$ mm (i.e. $z' > 29$ mm). Accordingly, (283) is still valid in the near field, but only starting from a larger distance than the Huygens-Fresnel integral. The complex constant we multiplied out before the integral symbol is the phase shift as calculated by *geometrical optics* for the distance z' between the aperture and point P' . It is important to note that (283) can be proven to analytically solve the paraxial wave equation (that will be presented at the practice).

10.3. Far-field approximation (Fraunhofer diffraction)

By converting (283), the Fourier transformation form of the Fresnel approximation is:

$$E(x', y') = \frac{e^{ikz'} e^{i\frac{k}{2z'}(x'^2 + y'^2)}}{i\lambda z'} \iint_{\Sigma} E(x, y) \cdot e^{i\frac{k}{2z'}(x^2 + y^2 - 2x'x - 2y'y)} dx dy \quad (284)$$

Before the integral we now have the complete phase of a spherical wave of z' radius centered on the Σ aperture. In case we are at a sufficiently large distance from the aperture, the below term in the exponent of the integrand can be neglected:

$$\frac{k}{2z'}(x^2 + y^2) \ll \pi \text{ [rad]} \Rightarrow \frac{1}{\lambda}(x^2 + y^2) \ll z'. \quad (285)$$

We interpret condition (285) again by using the example presented at the Fresnel approximation, and now we get that $z' \gg 395 \text{ mm}$ (or $z' > 4 \text{ m}$), i.e. we need to move away from the aperture by two orders of magnitude farther than at the Fresnel approximation so as to have the Fraunhofer approximation valid. Checking this condition is so important, that rearranging the formula the so-called *Fresnel number* was defined for a circular aperture of diameter D :

$$F \triangleq \frac{\left(\frac{D}{2}\right)^2}{\lambda z'} \quad (286)$$

If $F \ll 1$, then we speak of Fraunhofer diffraction, if $F \geq 1$ than of Fresnel diffraction. Should condition (285) be satisfied, we get the simplest expression of the diffraction integral:

$$E(x', y') = \frac{e^{ikz'} e^{i\frac{k}{2z'}(x'^2 + y'^2)}}{i\lambda z'} \iint_{\Sigma} E(x, y) \cdot e^{-i\frac{k}{z'}(x'x + y'y)} dx dy, \quad (287)$$

which is called far-field or *Fraunhofer* approximation. It is usually illustrated by considering the wavefronts of Huygens wavelets as planar surfaces due to the large propagation distance (see exponent). However, this is not exactly true, since the parabolic-phase coefficient outside the integral also corresponds to the elementary spherical waves, though for every one of which it is uniform. Hence, the correct interpretation is that in this approximation the wavelets can be approximated by a spherical wave of z' radius. Of course, if x' and y' are much smaller than z' , then the wavelets can be considered to be planar.

It is now that the $1/i$ constant of (287) really gains sense: the phase difference between the diffracting aperture Σ and a point being at a large z' distance from it on the z -axis is not $k \cdot z'$ as calculated by geometrical optics, but a value *advanced* by 90° relative to it! (Do not forget that in our discussion the phase increases in the direction of propagation, but in reality it decreases, i.e. delays.) This phenomenon is called *phase anomaly*, since it is like as if the wavelength became a little larger as a result of diffraction.

(287) is nothing else than a 2D Fourier transform with spatial frequencies f_x and f_y . As a reminder here is the formula of the 1D Fourier transform:

$$G(f_x) = \mathcal{F}\{g(x)\} = \int_{-\infty}^{\infty} g(x) \cdot e^{-i2\pi f_x x} dx \quad ; \quad f_x = \frac{x'}{\lambda z'} \quad ; \quad f_y = \frac{y'}{\lambda z'} \quad (288)$$

We can interpret (287) more easily by using the inverse transform. For this first we rearrange (287), so that the right side of the new expression takes the exact form of Fourier transform:

$$\frac{i\lambda z'}{e^{ikz'} e^{i\frac{k}{2z'}(x'^2 + y'^2)}} E(x', y') = \iint_{\Sigma} E(x, y) \cdot e^{-i\frac{k}{z'}(x'x + y'y)} dx dy \quad (289)$$

Applying the inverse transform to both sides, and exploiting that the inverse Fourier transform of the right side is exactly $E(P)$:

$$E(x, y) = \frac{i\lambda z'}{e^{ikz'}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-i\frac{k}{2z'}(x'^2 + y'^2)} \cdot E(x', y') \cdot e^{i2\pi(f_x x + f_y y)} df_x df_y, \quad (290)$$

from which we get the expression of inverse Fourier transform by substituting the (288) formula of spatial frequencies:

$$E(x, y) = \frac{i \cdot e^{-ikz'}}{\lambda z'} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-i\frac{k}{2z'}(x'^2+y'^2)} \cdot E(x', y') \cdot e^{i\frac{k}{z'}(x'x+y'y)} dx' dy'. \quad (291)$$

The exponential expression in the integrand of (291) describes *plane waves*, which can be interpreted as if the 2D Fourier transform represented a decomposition into harmonic plane waves. The analogy is not perfect though, since the phase of a spherical wave also appears in the formula of the inverse transform, which would only vanish if points P' were not taken over a plane, but on a sphere centered on aperture Σ . The wavefront components of the constituent plane waves are as follows:

$$k'_x = k \frac{x'}{z'} ; \quad k'_y = k \frac{y'}{z'}. \quad (292)$$

10.4. Applications of Fraunhofer diffraction

The Fraunhofer diffraction integral can even be analytically evaluated in simple cases. To begin with let us examine the far-field diffraction pattern of a rectangular aperture.

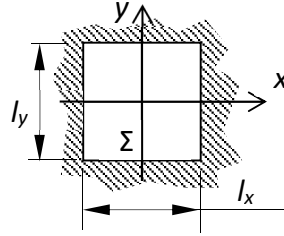


Fig. 47 Providing dimensions for the rectangular aperture.

Let the $E(P)$ field be constant 1 inside the aperture, and zero outside it:

$$E(x, y) := \text{rect}\left(\frac{x}{l_x}\right) \text{rect}\left(\frac{y}{l_y}\right) ; \quad \text{rect}(a) \triangleq \begin{cases} 1 & \text{when } |a| \leq \frac{1}{2} \\ 0 & \text{when } |a| > \frac{1}{2} \end{cases} \quad (293)$$

After passing aperture Σ , this corresponds to a homogeneous but truncated plane wave propagating along the z-axis. Putting this into (287):

$$E(x', y') = \frac{e^{ikz'} e^{i\frac{k}{2z'}(x'^2+y'^2)}}{i\lambda z'} \iint_{\Sigma} e^{-i\frac{k}{z'}(x'x+y'y)} dx dy \quad (294)$$

The integral can be analytically evaluated:

$$E(x', y') = \frac{e^{ikz'} e^{i\frac{k}{2z'}(x'^2+y'^2)}}{i\lambda z'} l_x l_y \text{sinc}\left(\frac{l_x x'}{\lambda z'}\right) \text{sinc}\left(\frac{l_y y'}{\lambda z'}\right) \quad (295)$$

$$I(x', y') = \langle S \rangle = \frac{v\varepsilon}{2} |E|^2 = \frac{v\varepsilon}{2} \frac{l_x^2 l_y^2}{\lambda^2 z'^2} \text{sinc}^2\left(\frac{l_x x'}{\lambda z'}\right) \text{sinc}^2\left(\frac{l_y y'}{\lambda z'}\right) \quad (296)$$

where the definition of function $\text{sinc}(\xi)$ is:

$$\text{sinc}(\xi) \triangleq \frac{\sin(\pi\xi)}{\pi\xi} ; \quad \xi \triangleq \frac{l_x x'}{\lambda z'} \quad \text{ill.} \quad \xi \triangleq \frac{l_y y'}{\lambda z'}. \quad (297)$$

The distinctive shape of the intensity function normalized to 1 can be seen in Fig. 48.

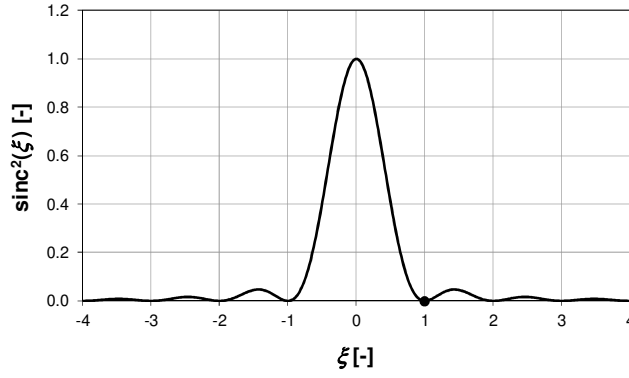


Fig. 48 Far-field diffraction intensity pattern of a rectangular aperture.

The point on the x - and y -axis, where the intensity takes its first zero is called the *Nyquist aperture*:

$$\xi = 1 \Rightarrow x_{\text{Nyquist}} = \frac{\lambda z'}{l_x}. \quad (298)$$

The far-field diffraction pattern of a circular aperture can be determined similarly to the above, only instead of a 2D Fourier transform we have to perform its cylindrical counterpart, called the *Hankel transform*, which is expressed in cylindrical coordinates.

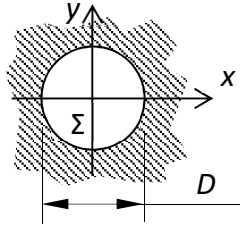


Fig. 49 Providing dimensions for a circular aperture.

Disregarding the derivation the result will be:

$$I(r') = \frac{v\varepsilon}{2} \left(\frac{kD^2}{8z'} \right)^2 \left(2 \frac{J_1(\pi\xi)}{\pi\xi} \right)^2, \quad (299)$$

where $J_1(\pi\xi)$ is a Bessel function of first kind and first order, with an argument definition of:

$$\pi\xi \triangleq \frac{kDr'}{2z'} \rightarrow \xi = \frac{Dr'}{\lambda z'}; \quad r^2 \triangleq x'^2 + y'^2 \quad (300)$$

In this case the unity-normalized intensity pattern of the diffraction spot is called the *Airy disk*. Comparing it with the diffraction pattern of a rectangular aperture we can conclude that the side maxima are much smaller here.

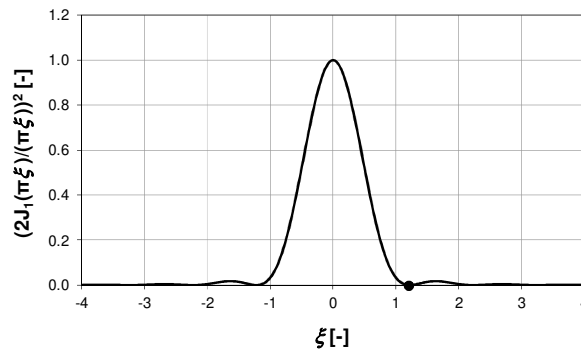


Fig. 50 Far-field diffraction intensity pattern of a circular aperture.

The name of the first zero place is *Airy radius*, based on $J_1(\pi\xi) = 0 \Rightarrow \xi = 1.22$ its value is:

$$R_{\text{Airy}} = 1.22 \frac{\lambda z'}{D}. \quad (301)$$

10.5. Applications of Fresnel diffraction

Next we will examine how the field distribution can be calculated in the x', y' focal plane of an ideal spherical wave. The radius of curvature of the wave is ρ at aperture Σ .

Condition: $\rho \gg \Sigma$ (paraxial approximation, so that the *Fresnel approximation* is valid). The illumination is thus a point source Q, *towards which* the light is converging:

$$E(x, y) := E_0 \cdot e^{-ik\sqrt{x^2+y^2+\rho^2}}, \quad (302)$$

where we considered the field amplitude as constant in aperture Σ . Due to the paraxial approximation in the exponent we can use Taylor series expansion again (up to the first order):

$$E(x, y) \approx E_0 \cdot e^{-ik\left(\rho + \frac{x^2}{2\rho} + \frac{y^2}{2\rho}\right)} \quad (303)$$

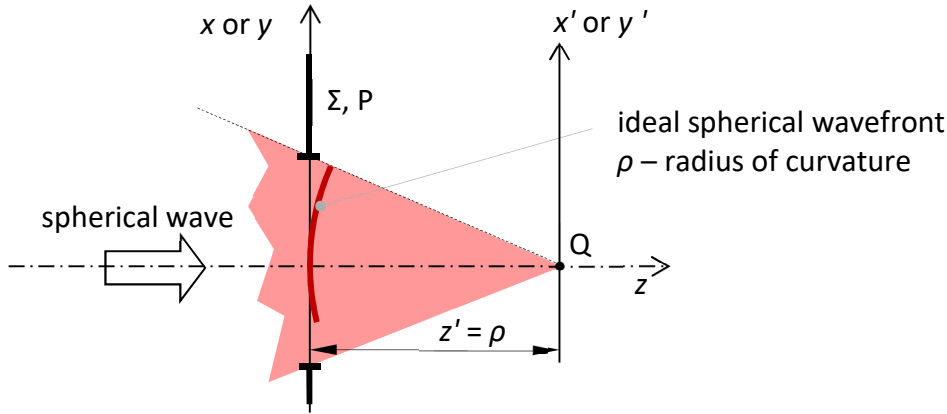


Fig. 51 Explanation for the determination of field distribution of a spherical beam in its focal plane.

This we substitute in the (284) Fourier transform formula of the *Fresnel approximation*, and let us take into account that now we are investigating right in the focal plane i.e. when $z' = \rho$:

$$E(x', y') = \frac{e^{i\frac{k(x'^2+y'^2)}{2\rho}}}{i\lambda\rho} \iint_{\Sigma} E_0 \cdot e^{-i\frac{k}{\rho}(x'x+y'y)} dx dy \quad (304)$$

This *formally* agrees with the (287) formula of Fraunhofer diffraction, except including ρ instead of z' in the equation! Since the phase of the diffracting beam (i.e. a spherical wave centered on Q) is zero in Q, the phase $k \cdot \rho$ given by geometrical optics vanishes from the formula. The phase referenced to this (zero) value is again advanced by 90° due to phase anomaly.

Generally, the output wavefront of an optical system is not ideal, but distorted by aberrations:

$$E(x, y) = E_0 \cdot e^{ik \cdot OPD(x, y)} \cdot e^{-ik\left(\rho + \frac{x^2}{2\rho} + \frac{y^2}{2\rho}\right)}, \quad (305)$$

where the optical path difference $OPD(x, y)$ describes the difference of the wavefront from an ideal sphere. From the squared magnitude of the electric field we can get the intensity distribution, which is usually the most interesting for us:

$$I = \frac{v\varepsilon}{2} \frac{E_0^2}{\lambda^2 \rho^2} \left| \iint_{\Sigma} e^{ik \cdot OPD(x,y)} \cdot e^{-i\frac{k}{\rho}(x'x+y'y)} dx dy \right|^2. \quad (306)$$

In such cases the ideal spherical wave introduced by (302) is called the *Gaussian reference sphere*. If we have a thin lens of effective focal length f inside the Σ aperture ($\rho = f$), see Fig. 52, then the Fourier transform of the radiation coming from the left will appear right in the focal plane of the lens.

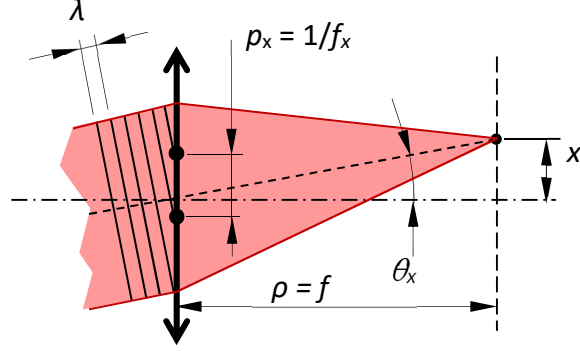


Fig. 52 Connection between points in the focal plane of a spherical wave and spatial frequencies.

According to what we have learned at Fraunhofer diffraction, the following spatial frequencies can be mapped to the different points of the focal plane:

$$f_x = \frac{x'}{\lambda \cdot f} \rightarrow k_x = 2\pi \cdot f_x = \frac{2\pi \cdot x'}{\lambda \cdot f} \rightarrow \theta_x \approx \tan \theta_x = \frac{k_x}{k} = \frac{x'}{f} \quad \text{and} \quad \theta_y \approx \frac{y'}{f} \quad (307)$$

If the field entering the lens aperture is a plane wave making θ_x and θ_y angles with the z -axis, then the field distribution in the focal plane will be a focal spot with its center offset by x' and y' relative to the optical axis. Accordingly, a plane wave of specific direction corresponds to one component of given spatial frequency in the 2D Fourier transform.

Integral (304) can be evaluated similarly to the method used at the Fraunhofer approximation, from which we obtain the already-presented Airy disk for a circular aperture and ideal spherical wavefront. Utilizing these results, the Airy disk radius measured in the image plane of a thin lens of diameter D in paraxial approximation is the following:

$$R_{\text{Airy}} = 1.22 \frac{\lambda f}{D}. \quad (308)$$

It should be noted, that by appropriate methods one can design lens systems that perform Fourier transformation not only in the paraxial approximation; these are called Fourier-transforming lenses.

10.6. Application of Fraunhofer diffraction: resolution limit of a telescope

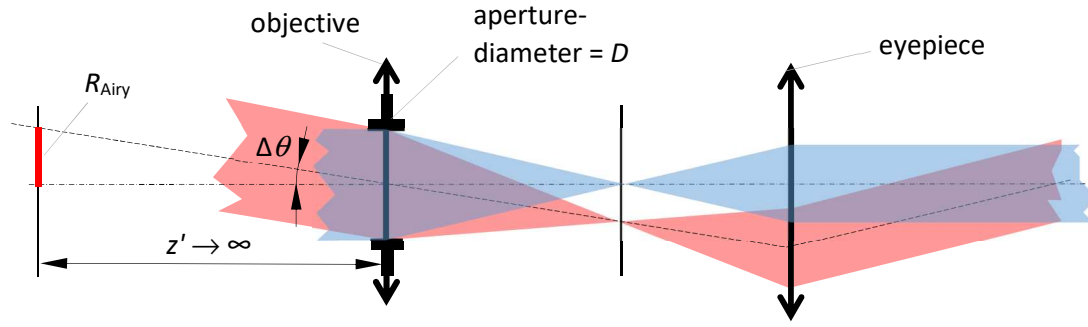


Fig. 53 Determining the resolution of a Keplerian telescope.

Rayleigh criterion of resolution: one can visually resolve two diffraction spots, if the first minimum of one of them coincides with the maximum of the other, see Fig. 54.

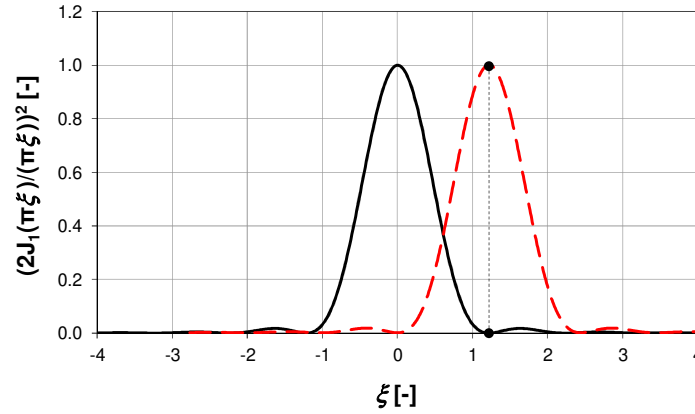


Fig. 54 Demonstration of the Rayleigh criterion of resolving power.

For telescopes, it is the aperture of the objective lens or the main mirror where light diffraction occurs. From the Fraunhofer diffraction model of circular apertures the angular resolution of a telescope is:

$$\Delta\theta \approx \text{tg}(\Delta\theta) = \frac{R_{\text{Airy}}}{z'} = 1.22 \cdot \frac{\lambda}{D} \quad (309)$$

10.7. Application of Fresnel diffraction: resolution of a microscope (supplementary)

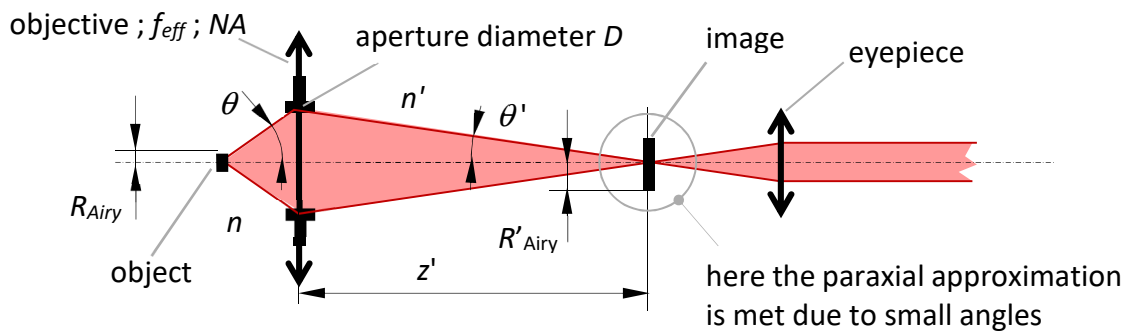


Fig. 55 Determining the resolution of a microscope objective.

The Rayleigh resolution in the image field of a small-numerical aperture lens ($\sin \theta < 0.1$) with circular aperture is, see (308):

$$R'_{\text{Airy}} = 1.22 \frac{\lambda z'}{D} \Rightarrow \text{tg } \theta' = \frac{D}{2z'} = 0.61 \frac{\lambda}{R'_{\text{Airy}}} \approx \sin \theta' \quad (310)$$

Generally we cannot use this formula directly for microscope objectives, since they are typically systems of large-numerical aperture. However, they are also aplanatic imaging systems, thus in their case the so-called Abbe sine condition is satisfied:

$$R_{\text{Airy}} \cdot n \cdot \sin \theta = R'_{\text{Airy}} \cdot n' \cdot \sin \theta' \quad (311)$$

from which the object-field resolution is:

$$R_{\text{Airy}} = 0.61 \frac{\lambda_0}{n' \sin \theta'} = 0.61 \frac{\lambda_0}{NA}. \quad (312)$$

(More explanation can be found in e.g. [3].)

10.8. Fraunhofer diffraction pattern of the field in case of a slit (supplementary)

A slit is a one-dimensional structure, whose far-field distribution can be treated as a two-dimensional diffraction problem. Unfortunately, we cannot use the limiting value of the 3D Fraunhofer diffraction field distribution of rectangular apertures, since if we increase one dimension of the aperture to infinity, then it is the Fraunhofer approximation itself that loses its validity. For this reason, the rigorous analytical solution can be derived from the 2D wave equation and 2D Green theorem, however with some clever thinking we can also find it out from (287). In order to get the solution we need to illuminate the Σ aperture of l_x and l_y width with a field, that can be written in the form of a product:

$$E(x, y) := E_x(x) \cdot E_y(y) \quad (313)$$

Then the diffraction integral can also be decomposed into the product of two factors:

$$E(x', y') = \left(\frac{e^{\frac{ikz'}{2}} \cdot e^{i\frac{k}{2z'}x'^2}}{\sqrt{i\lambda z'}} \int_{-l_x/2}^{l_x/2} E_x(x) \cdot e^{-i\frac{k}{z'}x'x} dx \right) \cdot \left(\frac{e^{\frac{ikz'}{2}} \cdot e^{i\frac{k}{2z'}y'^2}}{\sqrt{i\lambda z'}} \int_{-l_y/2}^{l_y/2} E_y(y) \cdot e^{-i\frac{k}{z'}y'y} dy \right) \quad (314)$$

In case of slits the illuminating field is of this form:

$$E(x, y) := \sqrt{E_0} \mathcal{E}(x) \cdot \sqrt{E_0} \mathcal{E}(y), \quad (315)$$

where the field magnitude E_0 in the $x = y = 0$ point, and in case of a slit of y -direction is $\mathcal{E}(y) := 1$. Then the diffraction integral will be like this:

$$E(x', y') = \left(\frac{e^{\frac{ikz'}{2}} \cdot e^{i\frac{k}{2z'}x'^2}}{\sqrt{i\lambda z'}} \int_{-l_x/2}^{l_x/2} \sqrt{E_0} \mathcal{E}(x) \cdot e^{-i\frac{k}{z'}x'x} dx \right) \cdot \left(\frac{e^{\frac{ikz'}{2}} \cdot e^{i\frac{k}{2z'}y'^2}}{\sqrt{i\lambda z'}} \sqrt{E_0} \int_{-l_y/2}^{l_y/2} e^{-i\frac{k}{z'}y'y} dy \right) \quad (316)$$

Since in y -direction the slit is infinitely long, examining the EM field in this direction it behaves as a geometrical optically propagating plane wave, i.e. for any z' coordinate it retains its $\sqrt{E_0}$ value taken in the $z' = 0$ plane, as well as its initial phase. According to these considerations, the expression in the parentheses on the right hand side must be this:

$$e^{\frac{ikz'}{2}} \sqrt{E_0} \quad (317)$$

According to (315) the field distribution in the slit is:

$$E(x) = E_0 \mathcal{E}(x) \quad (318)$$

By this we obtain the dimensionally correct final resolution:

$$E(x') = \frac{e^{ikz'} \cdot e^{i\frac{k}{2z'}x'^2}}{\sqrt{i\lambda z'}} \int_{-l_x/2}^{l_x/2} E(x) \cdot e^{-i\frac{k}{z'}x'x} dx. \quad (319)$$

This is the field of a cylindrical wave indeed if $E(x) = \text{const.}$, since the intensity decreases by $1/z'$. It is interesting to note that in case of a 2D problem the phase anomaly does not yield a 90° but 45° phase difference relative to a plane wave.

11. STATISTICAL OPTICS – TEMPORAL COHERENCE

Sources: [3], [10], [11], [12], wikipedia

11.1. Introduction – evolution of the concept of coherence

From the second half of the 17th century observations proliferated implying that concepts conceived of light up to then do not exactly correspond to reality. Grimaldi discovered a bending of light at the edges of shadows, that could not be explained by geometrical means. He coined the term “*diffraction*” (breaking up) for the effect, and explained the occurring patterns as undulations of aether (an imaginary medium carrying light particles). By his microscopic experiments Hooke examined the colorful, direction-dependent reflection (iridescence) of objects not containing color pigments (peacock feather, insect shell). At the end of the century Huygens was already capable of describing the basic characteristics of light diffraction by the help of elastic waves (more precisely disturbances akin to pulses). Though development in the 18th century was considerably hindered by Newton’s particle theory, the great scientist also contributed to the improvement of wave theory unintentionally: he was the first to take notes of the periodic patterns appearing in the thin air gap between two glass plates (Newton fringes, the color of thin sheets and pellicles), and discovered that white light is composed of the sum of colors (the term “*spectrum*” was also first used by him for the pattern of constituent colors). At the beginning of the 19th century, Young laid the wave-theory fundamentals of modern optics by explaining the unusual behavior of light traversing two slits or reflecting from thin films (e.g. a soap bubble). He associated the observed phenomena with the intensifying and weakening abilities of water and sound waves, and for the effect he introduced the concept of *interference*. By the help of his two-slit *wavefront-splitting* interferometer he established that light illuminating the slits should have originated from *the same source*, so that the interference can be perceptible. For better visibility he first transmitted sunlight through a small hole, and illuminated the next two slits producing interference with the resultant point source. He was the first to connect the concept of color to the periodicity of light as being an oscillation, and determined the approximate *wavelength* of basic colors by his interference experiments. Wave theory eventually gained its uniform structure owing to Fresnel in 1817.

Fraunhofer, making experiments with a prism spectroscopy, discovered dark lines in the solar spectrum in 1814. An explanation for these was provided in 1861 by Bunsen and Kirchhoff, stating that elements absorb and emit specific *monochromatic* (single-color) waves. For the call of chemistry spectrum analysis commenced shortly after. In 1862 Fizeau published his experimental results gained by *amplitude-splitting interferometry*. He examined monochromatic light, being thought of including one wavelength, by the amplitude-splitting two-beam interferometer named after him, where the source was the yellow flame of sodium (Na, D-line). The scientist experienced that the interferogram first almost absolutely vanished as a function of the path difference, then it reappeared again. Quite right, he concluded that the apparently single D-line is composed of two adjacent components of similar power. (According to the atom model of quantum mechanics the situation is this indeed: the spectral lines of concern are emitted by the $3p \rightarrow 3s$ transition of sodium, where the p-state has a fine structure, i.e. contains two neighboring energy levels; the corresponding wavelengths are 589.0 and 589.6 nm.) With a prism spectroscopy Fizeau got himself convinced about the correctness of his hypothesis. Thus, as soon as light considered to be monochromatic began examined by interferometers, it turned out that seemingly single-color light is not completely of one wavelength, and the visibility of the interferogram depends on the width of the spectrum.

Around 1865 Verdet reconsidered Young's two-slit experiment. He was seeking an answer to the question of *how distant* the slits can be apart until interference stays visible by using direct (spatially unfiltered) sunlight. (His calculations resulted in 20 μm for this distance.) In 1868 Fizeau made a suggestion for an interferometric arrangement by which, theoretically, the apparent diameter of remote objects (planets, stars) can be determined. In his thought experiment he masked the main mirror of a telescope so that light could reach the focal plane only through two distant apertures. The visibility of the resultant intensity pattern depends on the *size of the object* under observation. It was Michelson who later managed to implement the arrangement: by his stellar interferometer the diameter of the Galilei moons of Jupiter was first measured in 1891. From the same year started Michelson to publish his research results obtained in the field of spectral analysis: by using his interferometer invented in 1880 and later named after him he studied the spectral lines of several elements, examining the *visibility* of the interferogram as a function of path difference (this is the basis of contemporary interference or Fourier transform spectroscopy). The individual spectral lines turned out to be not indefinitely thin, instead light energy was distributed over a small wavelength range. Michelson determined the relationship between the interferogram intensity – path difference curve and the shape of the spectrum, which was nothing else than *Fourier transformation* (formulated by Sturm and Liouville around 1830). The interferogram visibility of the $\lambda_0 = 643.8\text{ nm}$ cadmium line that Michelson found to be the most monochromatic dropped to its half at about 15 cm path difference, according to which the spectral width is: $\Delta\lambda_{\text{FWHM}} = 0.0013\text{ nm}$.

In 1907 Laue published the results of his research made on the correlation of two light beams. In order to quantify it, in his paper he introduced the concept of *coherence* (the more coherent a light, the more correlated, and the more capable of producing interference it becomes). The research of Wiener, Schrödinger, van Cittert and Zernike made the concept of coherence more precise, the types of *spatial* and *temporal* coherence got separated. The former can be connected to Young's two-slit experiments and Michelson's stellar interferometer, whereas the latter to the spectral investigations of Fizeau and Michelson. In amplitude-splitting interferometers (such as a Michelson) we vary the time difference (path difference) and observe the visibility of the interferogram. From the speed of light (known after Rømer's measurements in 1675, and refined later on) it is easy to calculate how much time is necessary to cover the path difference corresponding to the specific visibility reduction. This interval is called the *coherence time*. Theoretical description of spatial coherence was accomplished by Zernike in 1939, introducing e.g. the *degree of coherence* to quantify the phenomenon. (For the sake of completeness it should be mentioned that polarization coherence also exists, for the characterization of which the *coherence matrix* has been used since Wiener's publication in 1928.) Wolf drew attention to the importance of coherence by organizing, summarizing and improving the relevant theories in the first edition of Born-Wolf: Principles of Optics in 1959.

In the wake of the development of stochastic calculus the physical interpretation and precise description of temporal coherence became possible by 1930. In terms of statistics, any signal (oscillation, wave) can be considered as either deterministic or indeterministic. For the former we can use harmonic oscillations, being very useful for the description of linear systems, as an example. In optics these are called as monochromatic waves, which can produce interference theoretically even at infinite path differences, i.e. the coherence time of such oscillations is infinite (see blue curve in Fig. 56). We can create an indeterministic signal e.g. by summing up an infinite number of harmonic oscillations of different frequency and random initial phase (see orange curve).

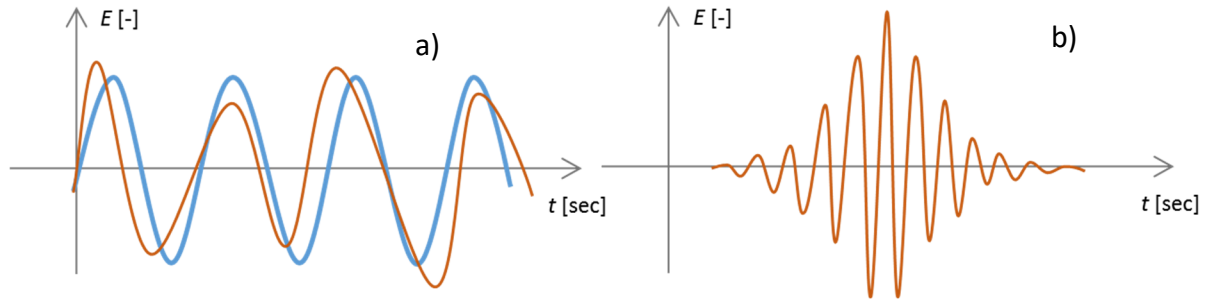


Fig. 56 Waves of identical power spectral density: in case of spectral components with random initial phase (a), and equal initial phase (b).

Should the initial phases not be different, but e.g. identical, then we will have a pulse (see figure on the right). Although pulses are deterministic in themselves, due to their finite duration larger time differences than this will not result in interference either. In what follows we will not discuss pulses any further.

One important feature of indeterministic oscillations is the finite (i.e. not zero) width of their *power spectral density*, so such a wave cannot be considered monochromatic. Another important statistical characteristic is the *autocorrelation function*, that indicates how correlated the phase states of the oscillation are between two instants at given temporal distance, how harmonic (how coherent) they are. If the width of the examined time interval is zero, then we have perfect correlation, the value of the autocorrelation function is at maximum. Deterministic signals (such as a monochromatic oscillation) have an autocorrelation function of constant value for any time interval length, on the other hand, for indeterministic signals it gradually decreases to zero. The visibility of an interferogram is in a strict relationship with the autocorrelation function of the oscillation. While examining spectral lines, Michelson realized that the wider frequency range is covered by the power spectral density of an indeterministic signal, the thinner its autocorrelation function will become (and vice versa), hence the shorter the coherence time and less coherent the oscillation will be. The completely general mathematical description of these was given by the Wiener-Khinchin theorem in the 1930s: the autocorrelation function and the power spectral density are mutual Fourier transforms of each other. The branch of modern physics that deals with the description of indeterministic light radiations is called *statistical optics*, the detailed presentation of which can be found in [13].

11.2. Two-beam interference with two frequencies

Next we will investigate the interference of two equivalently polarized plane waves (“a” and “b”) of almost identical propagation direction (i.e. we are in the paraxial scalar approximation), since this is the example by which we can most easily understand the effects that temporal coherence has on interferograms. Using real terminology the resultant intensity is:

$$I(\mathbf{r}, T) = \nu \varepsilon \langle (E_a(\mathbf{r}, t) + E_b(\mathbf{r}, t))^2 \rangle = \nu \varepsilon \frac{1}{T} \int_{-T/2}^{T/2} (E_a(\mathbf{r}, t) + E_b(\mathbf{r}, t))^2 dt \quad (320)$$

Following the footsteps of Fizeau first let us look at the interference of two waves each containing two-to-two harmonic components of different angular frequencies (ω_1 and ω_2). For the sake of better distinguishability from now on we will denote the amplitude of the interfering beams with A and B instead of E_0 :

$$\begin{aligned}
E_a(\mathbf{r}, t) &= A_1 \cos(\omega_1 t - \mathbf{k}_{a1} \mathbf{r} + \varphi_{a1}) + A_2 \cos(\omega_2 t - \mathbf{k}_{a2} \mathbf{r} + \varphi_{a2}) \\
&= A_1 \cos(\omega_1 t + \Phi_{a1}) + A_2 \cos(\omega_2 t + \Phi_{a2})
\end{aligned} \tag{321}$$

$$\begin{aligned}
E_b(\mathbf{r}, t) &= B_1 \cos(\omega_1 t - \mathbf{k}_{b1} \mathbf{r} + \varphi_{b1}) + B_2 \cos(\omega_2 t - \mathbf{k}_{b2} \mathbf{r} + \varphi_{b2}) \\
&= B_1 \cos(\omega_1 t + \Phi_{b1}) + B_2 \cos(\omega_2 t + \Phi_{b2})
\end{aligned} \tag{322}$$

We assume that $\mathbf{k}_{a1} \parallel \mathbf{k}_{a2}$ and $\mathbf{k}_{b1} \parallel \mathbf{k}_{b2}$, of course. Taking the squares:

$$\begin{aligned}
(E_a(\mathbf{r}, t) + E_b(\mathbf{r}, t))^2 &= \\
&= A_1 A_1 \cos(\omega_1 t + \Phi_{a1}) \cos(\omega_1 t + \Phi_{a1}) + A_1 A_2 \cos(\omega_1 t + \Phi_{a1}) \cos(\omega_2 t + \Phi_{a2}) + \\
&+ A_1 B_1 \cos(\omega_1 t + \Phi_{a1}) \cos(\omega_1 t + \Phi_{b1}) + A_1 B_2 \cos(\omega_1 t + \Phi_{a1}) \cos(\omega_2 t + \Phi_{a2}) + \\
&+ A_2 A_1 \cos(\omega_2 t + \Phi_{a2}) \cos(\omega_1 t + \Phi_{a1}) + A_2 A_2 \cos(\omega_2 t + \Phi_{a2}) \cos(\omega_2 t + \Phi_{a2}) + \\
&+ A_2 B_1 \cos(\omega_2 t + \Phi_{a2}) \cos(\omega_1 t + \Phi_{b1}) + A_2 B_2 \cos(\omega_2 t + \Phi_{a2}) \cos(\omega_2 t + \Phi_{a2}) + \\
&+ B_1 A_1 \cos(\omega_1 t + \Phi_{b1}) \cos(\omega_1 t + \Phi_{a1}) + B_1 A_2 \cos(\omega_1 t + \Phi_{b1}) \cos(\omega_2 t + \Phi_{a2}) + \\
&+ B_1 B_1 \cos(\omega_1 t + \Phi_{b1}) \cos(\omega_1 t + \Phi_{b1}) + B_1 B_2 \cos(\omega_1 t + \Phi_{b1}) \cos(\omega_2 t + \Phi_{a2}) + \\
&+ B_2 A_1 \cos(\omega_2 t + \Phi_{b2}) \cos(\omega_1 t + \Phi_{a1}) + B_2 A_2 \cos(\omega_2 t + \Phi_{b2}) \cos(\omega_2 t + \Phi_{a2}) + \\
&+ B_2 B_1 \cos(\omega_2 t + \Phi_{b2}) \cos(\omega_1 t + \Phi_{b1}) + B_2 B_2 \cos(\omega_2 t + \Phi_{b2}) \cos(\omega_2 t + \Phi_{a2})
\end{aligned} \tag{323}$$

Afterwards, (being a linear operation) we make time averaging for each term separately. First we choose such a short T time that only oscillations of the EM field around 10^{14} Hz get averaged (as if we had a fast detector with a cutoff below 10^{14} Hz). Those terms of the expression in which the multiplication factors have cosine functions of *identical arguments* will have a constant value. For instance:

$$\langle A_1 A_1 \cos(\omega_1 t + \Phi_{a1}) \cos(\omega_1 t + \Phi_{a1}) \rangle = \frac{A_1 A_1}{2} \tag{324}$$

Those terms having multiplication factors that contain cosine functions of *different arguments* can be separated into two parts by using the below trigonometric identity:

$$2 \cos(a) \cos(b) \equiv \cos(a + b) + \cos(a - b) \tag{325}$$

For example:

$$\begin{aligned}
&A_1 A_2 \cos(\omega_1 t + \Phi_{a1}) \cos(\omega_2 t + \Phi_{a2}) = \\
&= \frac{A_1 A_2}{2} [\cos((\omega_1 + \omega_2)t + \Phi_{a1} + \Phi_{a2}) + \cos((\omega_1 - \omega_2)t + \Phi_{a1} - \Phi_{a2})]
\end{aligned} \tag{326}$$

When time averaging such expressions, the sum-frequency term will always yield zero owing to the large frequency. Accordingly, (323) leaves us the next:

$$\begin{aligned}
\langle (E_a(\mathbf{r}, t) + E_b(\mathbf{r}, t))^2 \rangle = & \\
= & \frac{A_1 A_1}{2} + \frac{A_1 A_2}{2} \cos((\omega_1 - \omega_2)t + \Phi_{a1} - \Phi_{a2}) + \\
& + \frac{A_1 B_1}{2} \cos((\omega_1 - \omega_1)t + \Phi_{a1} - \Phi_{b1}) + \frac{A_1 B_2}{2} \cos((\omega_1 - \omega_2)t + \Phi_{a1} - \Phi_{b2}) + \\
& + \frac{A_2 A_1}{2} \cos((\omega_2 - \omega_1)t + \Phi_{a2} - \Phi_{a1}) + \frac{A_2 A_2}{2} + \\
& + \frac{A_2 B_1}{2} \cos((\omega_2 - \omega_1)t + \Phi_{a2} - \Phi_{b1}) + \frac{A_2 B_2}{2} \cos((\omega_2 - \omega_2)t + \Phi_{a2} - \Phi_{b2}) + \\
& + \frac{B_1 A_1}{2} \cos((\omega_1 - \omega_1)t + \Phi_{b1} - \Phi_{a1}) + \frac{B_1 A_2}{2} \cos((\omega_1 - \omega_2)t + \Phi_{b1} - \Phi_{a2}) + \\
& + \frac{B_1 B_1}{2} + \frac{B_1 B_2}{2} \cos((\omega_1 - \omega_2)t + \Phi_{b1} - \Phi_{b2}) + \\
& + \frac{B_2 A_1}{2} \cos((\omega_2 - \omega_1)t + \Phi_{b2} - \Phi_{a1}) + \frac{B_2 A_2}{2} \cos((\omega_2 - \omega_2)t + \Phi_{b2} - \Phi_{a2}) + \\
& + \frac{B_2 B_1}{2} \cos((\omega_2 - \omega_1)t + \Phi_{b2} - \Phi_{b1}) + \frac{B_2 B_2}{2}
\end{aligned} \tag{327}$$

Rearranging the expression:

$$\begin{aligned}
\langle (E_a(\mathbf{r}, t) + E_b(\mathbf{r}, t))^2 \rangle = & \frac{A_1 A_1}{2} + \frac{A_2 A_2}{2} + \frac{B_1 B_1}{2} + \frac{B_2 B_2}{2} + \\
& + \frac{A_1 B_1}{2} \cos(\Phi_{a1} - \Phi_{b1}) + \frac{A_2 B_2}{2} \cos(\Phi_{a2} - \Phi_{b2}) + \\
& + \frac{B_1 A_1}{2} \cos(\Phi_{b1} - \Phi_{a1}) + \frac{B_2 A_2}{2} \cos(\Phi_{b2} - \Phi_{a2}) + \\
& + \frac{A_1 A_2}{2} \cos((\omega_1 - \omega_2)t + \Phi_{a1} - \Phi_{a2}) + \frac{A_1 B_2}{2} \cos((\omega_1 - \omega_2)t + \Phi_{a1} - \Phi_{b2}) + \\
& + \frac{A_2 A_1}{2} \cos((\omega_2 - \omega_1)t + \Phi_{a2} - \Phi_{a1}) + \frac{A_2 B_1}{2} \cos((\omega_2 - \omega_1)t + \Phi_{a2} - \Phi_{b1}) + \\
& + \frac{B_1 A_2}{2} \cos((\omega_1 - \omega_2)t + \Phi_{b1} - \Phi_{a2}) + \frac{B_1 B_2}{2} \cos((\omega_1 - \omega_2)t + \Phi_{b1} - \Phi_{b2}) + \\
& + \frac{B_2 A_1}{2} \cos((\omega_2 - \omega_1)t + \Phi_{b2} - \Phi_{a1}) + \frac{B_2 B_1}{2} \cos((\omega_2 - \omega_1)t + \Phi_{b2} - \Phi_{b1})
\end{aligned} \tag{328}$$

In reality the difference of ω_1 and ω_2 is usually so large that our detectors are unable to follow even the difference-frequency oscillations. E.g. the sodium D-line Fizeau used in his 1862 experiment has wavelengths of 589.0 and 589.6 nm, where $\omega_1 - \omega_2 = 3.3 \cdot 10^{12}$ Hz. We can conclude that in our measurements every time-dependent term gets averaged out. To this single-mode lasers can be exceptions, the frequency bandwidth of which is 1-10 MHz. In such cases the above-described high-frequency beats or modulation can be revealed by using a detector of adequate bandwidth. After a subsequent rearrangement (328) leaves us this:

$$\begin{aligned}
\langle (E_a(\mathbf{r}, t) + E_b(\mathbf{r}, t))^2 \rangle = & \\
= & \left[\frac{A_1 A_1}{2} + \frac{B_1 B_1}{2} + A_1 B_1 \cos(\varphi_{a1} - \varphi_{b1}) \right] + \left[\frac{A_2 A_2}{2} + \frac{B_2 B_2}{2} + A_2 B_2 \cos(\varphi_{a2} - \varphi_{b2}) \right], \tag{329}
\end{aligned}$$

where we employed the parity of the cosine function due to which its argument can be multiplied by -1 without changing its value, and expanded Φ . For the sake of simplicity here we considered only two plane waves of identical directions, hence the spatially dependent term

cancels, but in general $\Phi(\mathbf{r})$ is spatially dependent. We can interpret this result as for the superimposed waves “a” and “b” only those components are able to interfere that have *identical angular frequency* (wavelength), and these interferograms can be summed up in *intensity*. Since we can produce radiations of identical wavelength and initial phase only by splitting the light of the same source, we can assume that magnitudes A and B are proportional to each other. Thus without a loss of generality we can write that:

$$B_1 = p \cdot A_1 ; B_2 = p \cdot A_2, \quad (330)$$

where “ p ” is a constant. By this (329) takes this form:

$$\begin{aligned} \langle (E_a(\mathbf{r}, t) + E_b(\mathbf{r}, t))^2 \rangle = \\ = A_1 A_1 \left[\frac{1}{2} + \frac{p^2}{2} + p \cos(\varphi_{a1} - \varphi_{b1}) \right] + A_2 A_2 \left[\frac{1}{2} + \frac{p^2}{2} + p \cos(\varphi_{a2} - \varphi_{b2}) \right] \end{aligned} \quad (331)$$

By somewhat limiting generality let us suppose that the two interfering beams are of identical magnitude, i.e. $p = 1$. By this:

$$\langle (E_a(\mathbf{r}, t) + E_b(\mathbf{r}, t))^2 \rangle = A_1 A_1 (1 + \cos(\varphi_{a1} - \varphi_{b1})) + A_2 A_2 (1 + \cos(\varphi_{a2} - \varphi_{b2})) \quad (332)$$

and the average intensities:

$$I_{a1} = I_{b1} = v\varepsilon \langle (A_1 \cos(\omega_1 t - \mathbf{k}_{a1} \mathbf{r} + \varphi_{a1}))^2 \rangle = v\varepsilon \frac{A_1^2}{2} \triangleq I_1 \quad (333)$$

$$I_{a2} = I_{b2} = v\varepsilon \langle (A_2 \cos(\omega_2 t - \mathbf{k}_{a2} \mathbf{r} + \varphi_{a2}))^2 \rangle = v\varepsilon \frac{A_2^2}{2} \triangleq I_2 \quad (334)$$

Finally the resultant intensity is:

$$I = 2I_1 (1 + \cos(\varphi_{a1} - \varphi_{b1})) + 2I_2 (1 + \cos(\varphi_{a2} - \varphi_{b2})). \quad (335)$$

In both wavefront- and amplitude-splitting interferometers phase difference is physically realized as a d path difference. By this the phase difference can be formulated as:

$$k = n \frac{\omega}{c} \Rightarrow \varphi_{a1} - \varphi_{b1} = \frac{2\pi}{\lambda_1} d = \frac{\omega_1}{c} n d ; \quad \varphi_{a2} - \varphi_{b2} = \frac{2\pi}{\lambda_2} d = \frac{\omega_2}{c} n d, \quad (336)$$

where n is the refractive index of the medium (now we are assuming that it is not wavelength-dependent, but in reality to a small extent it is). It is useful to change wavelength to angular frequency, since in linear media ω is constant. Substituting (336) and introducing $OPD = nd$, equation (335) leaves us:

$$I(OPD) = 2I_1 \left(1 + \cos\left(\frac{\omega_1}{c} OPD\right) \right) + 2I_2 \left(1 + \cos\left(\frac{\omega_2}{c} OPD\right) \right). \quad (337)$$

The above relationship describes the interferogram as the sum of two cosine functions (elementary interferograms). Since both components have different frequency (period, wavelength), by increasing the OPD the elementary interferograms become offset relative to each other. This is illustrated in Fig. 57 for three different wavelengths, using a Fizeau interferometer. The beats of the interferograms is clearly observable e.g. between the blue-green wavelengths: at $OPD \approx 0.7 \mu\text{m}$ the two interferograms are in opposite phase (the resultant interference vanishes here), and at $1.5 \mu\text{m}$ they become in phase again (the interference appears again). Summing up a infinite number of wavelengths, the visibility of the resultant interferogram gradually tends to zero as the path difference increases.

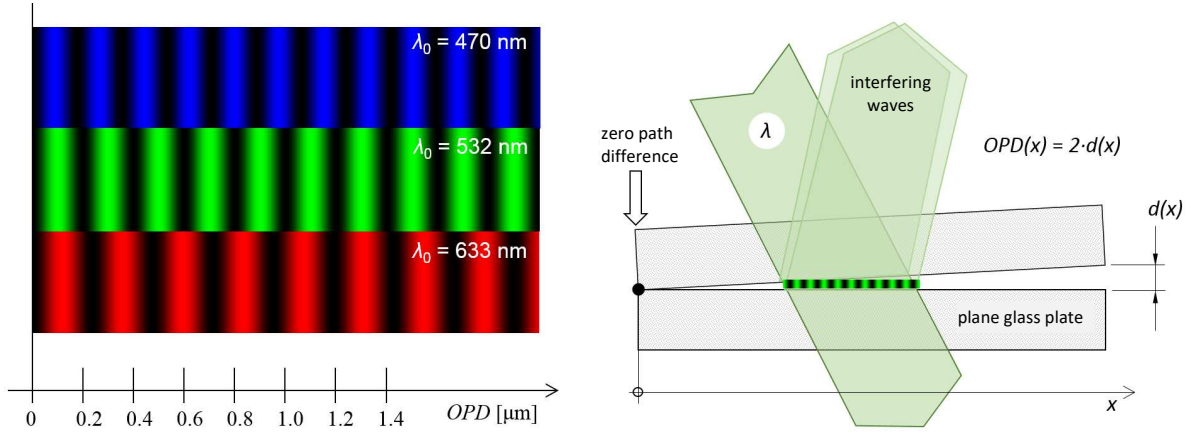


Fig. 57 Fizeau interferograms for radiations of different wavelength.

With the help of (337) we can easily move on from the two-frequency case to the interference of waves having frequency spectra containing an arbitrary number of components. Corresponding to the practice, in what follows we will examine the interferograms in a Michelson (also known as Twyman-Green) interferometer (Fig. 58), since in this we can more easily and precisely carry out the continuous and controlled variation of OPD . The I intensity analyzed is proportional to the signal (current, voltage) measured by the detector shown in Fig. 58.

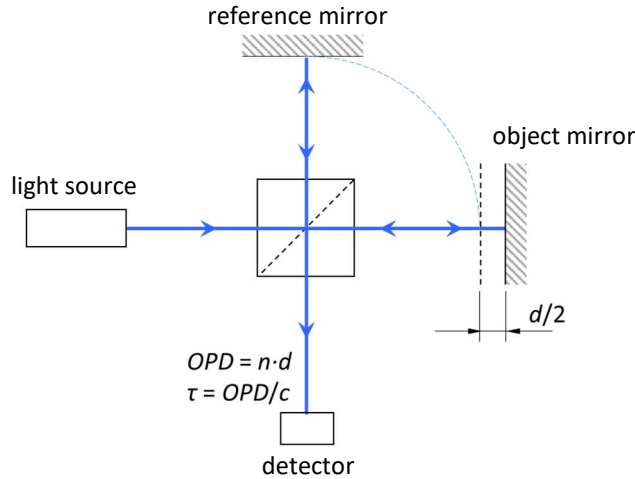


Fig. 58 Michelson (Twyman-Green) interferometer.

11.3. Two-beam interference for oscillations with multiple frequencies

If “a” and “b” waves are composed of “ N ” components of different wavelength, by generalizing the previous thoughts the intensity of an interferogram can be obtained the following way:

$$I(OPD) = 2 \sum_{i=1}^N I_i \left(1 + \cos \left(\frac{\omega_i}{c} OPD \right) \right) \quad (338)$$

In the above case N discrete angular frequencies make up the *spectrum* of the light. Should the spectrum be continuous, summation gets replaced by integration:

$$I(OPD) = 2 \int_0^{\infty} S(\nu) \cdot \left(1 + \cos \left(2\pi \frac{\nu}{c} \cdot OPD \right) \right) \cdot d\nu \quad ; \quad S(\nu) \triangleq \frac{dI(\nu)}{d\nu}, \quad (339)$$

where $S(\nu)$ is a new quantity, the spectral density. For continuous-wave light sources *power spectral density* is used, whereas a spectral density pertaining to energy is applied e.g. for pulses. In the above equation we changed angular frequency to frequency ($\nu = \omega/2\pi$). With this the total intensity of the “a” or “b” beam can be expressed as:

$$I_0 = \int_0^{\infty} S(\nu) d\nu \quad (340)$$

If we denote the time taken by light while travelling an optical path length of OPD by τ :

$$I(\tau) = 2 \int_0^{\infty} S(\nu) (1 + \cos(2\pi\nu \cdot \tau)) d\nu \quad ; \quad \tau \triangleq \frac{OPD}{c} \quad (341)$$

The intensity described by the above formula is nothing else than the *interferogram* as a function of τ . The relationship contains a constant and a τ -dependent term. Separating these:

$$I(\tau) = 2 \int_0^{\infty} S(\nu) d\nu + 2 \int_0^{\infty} S(\nu) \cos(2\pi\nu \cdot \tau) d\nu = 2I_0 \left(1 + \int_0^{\infty} \frac{S(\nu)}{I_0} \cos(2\pi\nu \cdot \tau) d\nu \right) \quad (342)$$

Introducing the concept of *normalized coherence function* (a.k.a. *degree of coherence*), denoted by $g(\tau)$ (sometimes γ), the general form of the above relationship can be written as:

$$g(\tau) \triangleq \int_0^{\infty} \frac{S(\nu)}{I_0} \cos(2\pi\nu \cdot \tau) d\nu \quad \rightarrow \quad I(\tau) = 2I_0(1 + g(\tau)). \quad (343)$$

In Chapter 12 we will discuss what the $g(\tau)$ function means in practice, how it can be determined from a specific spectrum or from temporal field changes, and what are its main characteristics.

12. INVESTIGATION OF TEMPORAL COHERENCE IN TIME DOMAIN

12.1. Quasi-monochromatic waves

Oscillations having a continuous power spectrum of finite width, containing spectral components of random initial phase behave as indeterministic signals in the time domain. By using such waves interference can only be achieved at a finite path difference. Among indeterministic signals those are of most interest from technical respect, where the interference remains visible even for path differences much longer than the wavelength. The power spectrum of such radiations can be characterized by a central frequency (ν_0), relative to which the spectral width is significantly smaller: $\Delta\nu \ll \nu_0$. Such oscillations are called *quasi monochromatic*.

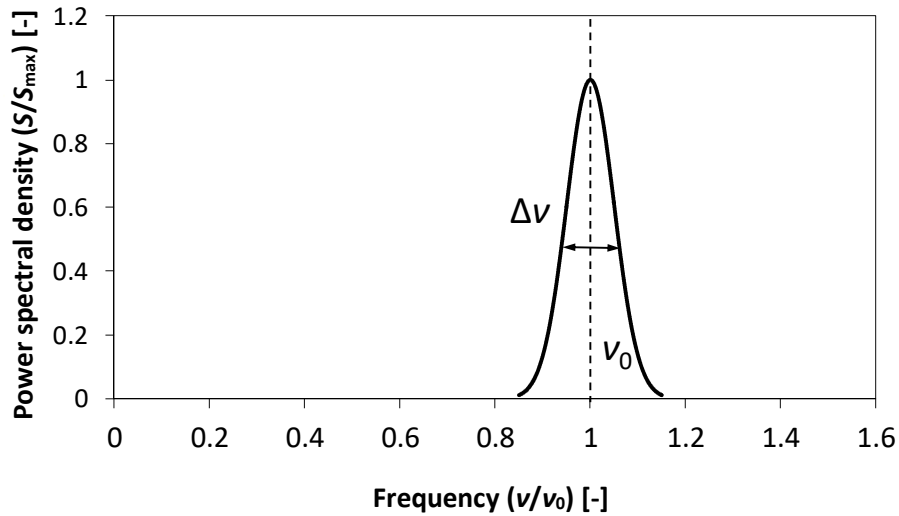


Fig. 59 Demonstration of the power spectral density of a quasi-monochromatic light source.

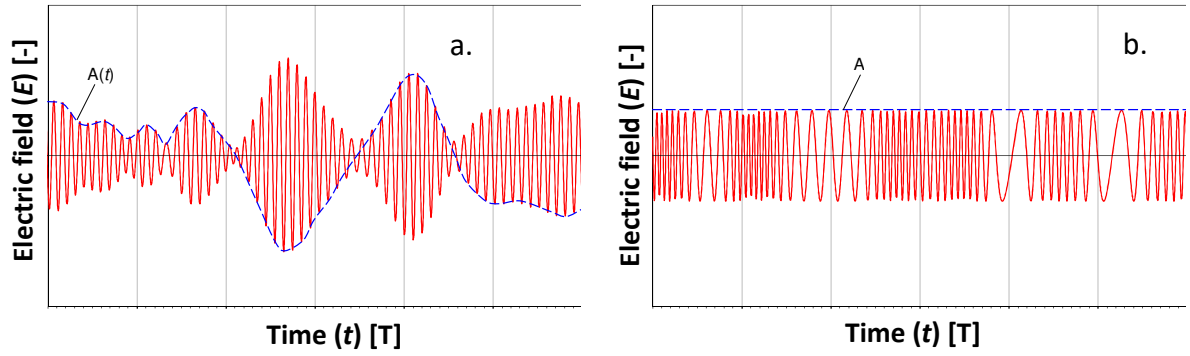
Theoretically, the temporal dependence of any signal can be expressed with the following general formula:

$$E(t) = A(t) \cos(\omega_0 t + \varphi(t)) \quad (344)$$

In quasi-monochromatic case $\Delta\nu \ll \nu_0$, therefore both magnitude (A) and phase (φ) feature much slower temporal variations than the angular frequency. Similarly to the discussion of the slowly varying amplitude approximation discussed at geometrical optics, differentiating (344) by time results in the next (T_0 denotes the period corresponding to the central frequency):

$$\left| \frac{1}{A} \frac{dA}{dt} \right| \ll \omega_0 \Leftrightarrow \left| \frac{1}{A} \frac{dA}{dt} T_0 \right| \ll 2\pi \quad \text{and} \quad \left| \frac{d\varphi}{dt} \right| \ll \omega_0 \Leftrightarrow \left| \frac{d\varphi}{dt} T_0 \right| \ll 2\pi \quad (345)$$

Two typical kinds of such oscillations can be seen in Fig. 60. The left hand side shows the so-called amplitude modulation (this corresponds to most visible radiations such as LEDs, incandescent and fluorescent sources etc.), while the right figure presents frequency modulation, which occurs at single-mode lasers, FM-modulated radio waves. In the following we will only discuss amplitude modulation.



Amplitude modulation: $E(t) = A(t) \cdot \cos(\omega_0 t + \varphi)$ Frequency modulation: $E(t) = A \cdot \cos(\omega_0 t + \varphi(t))$

Fig. 60 Presentation of the time dependence of the electric field for a chaotic, spontaneously emitted (thermal) light (a) and a light source of induced emission (b).

Depending on the light emission mechanism, a power spectrum can have multiple shapes, e.g. Lorentzian (lifetime or collision broadening or Fabry-Perot resonators), or Gaussian (Doppler broadening). The connection between the width of frequency and wavelength spectra of λ_0 central wavelength can be easily obtained:

$$\nu = \frac{c}{\lambda} ; \quad \Delta\lambda = \frac{d\lambda}{d\nu} \Delta\nu = \left[\frac{d\nu}{d\lambda} \right]^{-1} \Delta\nu \Rightarrow |\Delta\lambda| = \left. \frac{\lambda^2}{c} \right|_{\lambda_0} \cdot \Delta\nu \Leftrightarrow \left| \frac{\Delta\lambda}{\lambda_0} \right| = \left| \frac{\Delta\nu}{\nu_0} \right| \quad (346)$$

Visible light sources: Gas laser: $\Delta\lambda \sim 0.001 \text{ nm}$
Solid-state laser: $\Delta\lambda \sim 0.1 \text{ nm}$
Semiconductor laser: $\Delta\lambda \sim 1 \text{ nm}$ (+ temperature fluctuation: $2 \text{ nm}/10^\circ\text{C}$)
LED: $\Delta\lambda \sim 10 \text{ nm}$

12.2. Coherence function of a quasi-monochromatic oscillation

Based on definition (343) it is worth introducing the complex coherence function, since it will simplify the discussion of further relationships very much:

$$\tilde{g}(\tau) \triangleq \int_0^\infty \frac{S(\nu)}{I_0} \cos 2\pi\nu\tau d\nu + i \int_0^\infty \frac{S(\nu)}{I_0} \sin 2\pi\nu\tau d\nu = \int_0^\infty \frac{S(\nu)}{I_0} e^{i2\pi\nu\tau} d\nu \quad (347)$$

The integration range can be expanded to $-\infty$, since quasi monochromatic light waves always have spectra constrained to a given frequency interval, the lower boundary of which seldom reaching down to $\nu = 0 \text{ Hz}$:

$$\tilde{g}(\tau) = \int_0^\infty \frac{S(\nu)}{I_0} e^{i2\pi\nu\tau} d\nu \approx \int_{-\infty}^\infty \frac{S(\nu)}{I_0} e^{i2\pi\nu\tau} d\nu \quad (348)$$

This integral is nothing else than the formula of inverse Fourier transformation, thus:

$$\tilde{g}(\tau) = \mathcal{F}^{-1} \left\{ \frac{S(\nu)}{I_0} \right\} \quad \text{and} \quad g(\tau) = \text{Re}\{\tilde{g}(\tau)\} \quad (349)$$

Consequently, the inverse Fourier transform of power spectral density is the normalized coherence function, implying also that S/I_0 is the Fourier transform of function $\tilde{g}(\tau)$:

$$\frac{S(\nu)}{I_0} = \mathcal{F}\{\tilde{g}(\tau)\}. \quad (350)$$

Thus the power spectral density can be obtained from the normalized coherence function by Fourier transformation, and vice versa. This is what Michelson realized too, and this is the basis of present interference spectroscopy (FTIR), where the coherence function $g(\tau)$ is measured by an interferometer and the spectrum is calculated as its Fourier transform. From the properties of Fourier transformation it also follows that $\tilde{g}(\tau)$ must be hermitian since $S(\nu)$ is a real function, so

$$\tilde{g}(-\tau) = \tilde{g}^*(\tau) \Rightarrow g(-\tau) = \text{Re}\{\tilde{g}^*(\tau)\} = g(\tau), \quad (351)$$

thus the normalized coherence function is even, i.e. always symmetric to the $\tau = 0$ time difference. A further feature originates in relationship (348) being the total intensity at $\tau = 0$ normalized to I_0 , which always results in 1 according to (340). Consequently:

$$\tilde{g}(0) = g(0) = 1. \quad (352)$$

Now we examine a specific case. Let us investigate a quasi-monochromatic wave of ν_0 central frequency and $\Delta\nu$ bandwidth ($\Delta\nu \ll \nu_0$), and consider the power spectrum to be of Gaussian distribution:

$$S(\nu) := \frac{I_0\sqrt{2}}{\Delta\nu_2\sqrt{\pi}} e^{-2\left(\frac{\nu-\nu_0}{\Delta\nu_2}\right)^2}; \quad \Delta\nu \triangleq 2\Delta\nu_2 \quad (353)$$

The normalizing coefficient of the function has been chosen so that the total intensity is I_0 in accordance with (340). Based on (350) the normalized coherence function is:

$$\tilde{g}(\tau) = \mathcal{F}^{-1} \left\{ \frac{\sqrt{2}}{\Delta\nu_2\sqrt{\pi}} e^{-2\left(\frac{\nu-\nu_0}{\Delta\nu_2}\right)^2} \right\} \quad (354)$$

Inverse Fourier transforming a Gaussian function gives (see Wolfram Mathworld and [14]):

$$\mathcal{F}^{-1} \left\{ e^{-2\left(\frac{\nu-\nu_0}{\Delta\nu_2}\right)^2} \right\} = \sqrt{\frac{\pi}{2}} \cdot \Delta\nu_2 \cdot e^{-\frac{1}{2}(\pi\Delta\nu_2\tau)^2} \cdot e^{-i2\pi\nu_0\tau} \quad (355)$$

By using this, (354) can be evaluated analytically, leaving us the coherence function:

$$\tilde{g}(\tau) = e^{-\frac{1}{2}(\pi\Delta\nu_2\tau)^2} \cdot e^{-i2\pi\nu_0\tau} = g_A(\tau) \cdot e^{-i2\pi\nu_0\tau} \quad (356)$$

$$g(\tau) = \text{Re}\{\tilde{g}(\tau)\} = g_A(\tau) \cdot \cos 2\pi\nu_0\tau = e^{-\frac{1}{2}(\pi\Delta\nu_2\tau)^2} \cdot \cos 2\pi\nu_0\tau, \quad (357)$$

where $g_A(\tau)$ also has a Gaussian shape. Based on (343) the interferogram is:

$$I(\tau) = 2I_0(1 + g(\tau)) = 2I_0(1 + g_A(\tau) \cdot \cos 2\pi\nu_0\tau), \quad (358)$$

the shape of which is illustrated in the below figure.

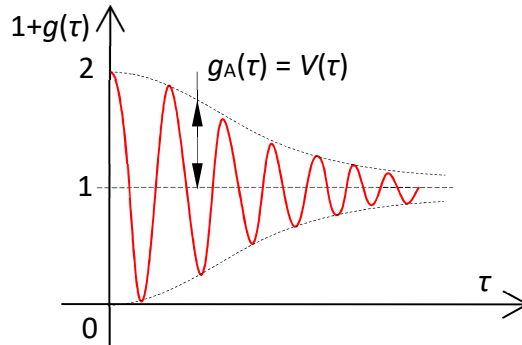


Fig. 61 Demonstration of the normalized coherence function and its appearance on the interferogram.

Comparing equation (358) with that formulated at two-beam interference

$$I(\delta) = 2I_0(1 + \cos(\delta)) \quad (359)$$

we get such an interference, where the resultant intensity changes periodically with the delay time τ , and the magnitude of the normalized coherence function can be interpreted as the localized visibility of interference fringes:

$$V(\tau) = \frac{I_{\max}(\tau) - I_{\min}(\tau)}{I_{\max}(\tau) + I_{\min}(\tau)} = \frac{2I_0(1 + g_A(\tau)) - 2I_0(1 - g_A(\tau))}{2I_0(1 + g_A(\tau)) + 2I_0(1 - g_A(\tau))} = \frac{2g_A(\tau)}{2} = g_A(\tau). \quad (360)$$

Relationship (357) nicely supports and explains the observations of Fizeau and Michelson: with increasing path (time) difference the interference of polychromatic light beams exhibits a decaying visibility. That time difference, where visibility drops below a predetermined value (by definition e.g. to its $1/e$ or $1/e^2$ fraction) is called *coherence time* (τ_c). For a known $g_A(\tau)$ we can easily determine the coherence time τ_c , we only need to solve the below equation:

$$g_A(\tau_c) = \frac{1}{2} \text{ or } \frac{1}{e} \text{ or } \frac{1}{e^2} \text{ etc.} \quad (361)$$

In our case, if $g_A(\tau_c) = e^{-2}$, then (357) results in the following coherence time:

$$\tau_c = \frac{2}{\pi\Delta\nu_2} = \frac{4}{\pi\Delta\nu} \Leftrightarrow \tau_c\Delta\nu = \frac{4}{\pi}. \quad (362)$$

12.3. Description of temporally statistic behavior by autocorrelation function

As we could see at the quasi-monochromatic case, the temporal dependence of waves with a finite spectral width shows random, indeterminate characteristics. Let us consider two waves ("a" and "b") of extensive spectra, which are obviously *not harmonic*, and let us investigate their interference in the *time domain*. For a better understanding no complex notation is used here again, though the derivation can be performed that way too. Since now we are examining temporal changes at a given point of space, the position dependence will not be indicated. Based on (320), the intensity averaged for an arbitrary T time is the following (320):

$$I(T) = v\varepsilon \langle (E_a(t) + E_b(t))^2 \rangle = v\varepsilon \frac{1}{T} \int_{-T/2}^{T/2} (E_a(t) + E_b(t))^2 dt \quad (363)$$

Let us also assume, that the time average is independent of the T averaging time, namely

$$I(T) = I(T \rightarrow \infty) = \text{const.} \triangleq I, \quad (364)$$

and that it is true for averaging both E_a and E_b . Such beams are called *statistically stationary*. Let the two beams not be independent, but let us construct E_b by delaying E_a with τ time:

$$E_b(t) := E_a(t - \tau) \quad (365)$$

Expanding the square we get the resultant intensity:

$$I(\tau) = v\varepsilon \langle (E_a(t) + E_a(t - \tau))^2 \rangle = v\varepsilon \langle E_a^2(t) + E_a^2(t - \tau) + 2E_a(t)E_a(t - \tau) \rangle \quad (366)$$

Since the beams are statistically stationary, time averaging is invariant of a τ shift:

$$I(\tau) = 2I_0 + 2v\varepsilon \cdot \langle E_a(t) \cdot E_a(t - \tau) \rangle = 2I_0 \left[1 + \frac{\langle E_a(t) \cdot E_a(t - \tau) \rangle}{\langle E_a(t)^2 \rangle} \right], \quad (367)$$

where $I_0 = v\varepsilon\langle E_a^2(t) \rangle$ is the intensity of beam E_a for $T \rightarrow \infty$. It is worth comparing the above equation with (343), which has an identical form, hence the normalized coherence function $g(\tau)$ is nothing else than the normalized autocorrelation function of $E_a(t)$:

$$g(\tau) = \frac{\langle E_a(t) \cdot E_a(t - \tau) \rangle}{\langle E_a(t)^2 \rangle}. \quad (368)$$

Now let us assume that E_a is an oscillation of ω_0 angular frequency, the magnitude $A(t)$ of which varies slowly in time relative to the period, i.e. the oscillation is quasi-monochromatic:

$$E(t) = A(t) \cos(\omega_0 t) \quad (369)$$

Since the two waves were started from the same source, the initial phase cancels from the next calculations, thus it is not indicated. Average intensity of the oscillation for $T \rightarrow \infty$ is:

$$I_0 = v\varepsilon\langle A(t)^2 \cos^2(\omega_0 t) \rangle \approx v\varepsilon\langle A(t)^2 \rangle \langle \cos^2(\omega_0 t) \rangle. \quad (370)$$

The approximation is justified by $A(t)$ slowly varying relative to $\cos(\omega_0 t)$. Thus the intensity:

$$I_0 = v\varepsilon\langle E_a^2(t) \rangle = \frac{v\varepsilon}{2} \langle A(t)^2 \rangle, \quad (371)$$

which we will need later. Substituting the quasi-monochromatic E_a field into (367):

$$I(\tau) = 2I_0 + 2v\varepsilon\langle A(t)A(t - \tau) \cos(\omega_0 t) \cos(\omega_0 t - \omega_0 \tau) \rangle \quad (372)$$

The above relationship can be transformed into this form based on (325):

$$\begin{aligned} I(\tau) &= 2I_0 + v\varepsilon\langle A(t)A(t - \tau)(\cos(2\omega_0 t - \omega_0 \tau) + \cos(\omega_0 \tau)) \rangle = \\ &= 2I_0 + v\varepsilon\langle A(t)A(t - \tau) \cos(2\omega_0 t - \omega_0 \tau) \rangle + v\varepsilon\langle A(t)A(t - \tau) \cos(\omega_0 \tau) \rangle \end{aligned} \quad (373)$$

Since $A(t)$ varies slowly relative to an oscillation of ω_0 angular frequency as per our assumption, only the $2\omega_0$ term averages out. Multiplying out $\cos(\omega_0 \tau)$ from time averaging we get:

$$I(\tau) = 2I_0 + v\varepsilon\langle A(t)A(t - \tau) \rangle \cdot \cos(\omega_0 \tau). \quad (374)$$

The above equation can be further transformed:

$$I(\tau) = 2I_0[1 + g_A(\tau) \cdot \cos(\omega_0 \tau)]. \quad (375)$$

Comparing this with (343) and (367) we arrive at the normalized coherence function:

$$g(\tau) = g_A(\tau) \cdot \cos(\omega_0 \tau) \quad \text{and} \quad g_A(\tau) \triangleq \frac{v\varepsilon\langle A(t)A(t - \tau) \rangle}{2I_0} = \frac{\langle A(t)A(t - \tau) \rangle}{\langle A(t)^2 \rangle}. \quad (376)$$

The $g_A(\tau)$ function is the magnitude of the normalized coherence function $g(\tau)$, taking its maximum value at $\tau = 0$, where the numerator is I_0 , see (371): $g(0) = 1$, as we expected. Besides, the autocorrelation function is always even, which is in correspondence again with the analysis made in frequency domain. From the above formula it also turns out that $g(\tau \rightarrow \infty) = 0$, should $A(t)$ change stochastically.

The connection between the power spectral density and the autocorrelation function is described in general by the Wiener-Khinchin theorem of signal processing:

$$\mathcal{F}^{-1} \left\{ |\mathcal{F}\{\tilde{E}_a\}|^2 \right\} = \int_{-\infty}^{\infty} \tilde{E}_a(t) \tilde{E}_a^*(t - \tau) dt \Leftrightarrow \mathcal{F}^{-1}\{S(\nu)\} = I_0 \cdot \tilde{g}(\tau), \quad (377)$$

where:

$$\tilde{E}_a(t) = A(t)e^{i\omega_0 t} ; \quad \tilde{g}(\tau) = \frac{\langle \tilde{E}_a(t)\tilde{E}_a^*(t-\tau) \rangle}{\langle \tilde{E}_a(t)\tilde{E}_a^*(t) \rangle} = g_A(\tau) \cdot e^{i\omega_0 \tau} . \quad (378)$$

Power spectral density can be interpreted based on Parseval's theorem. This states that Fourier transformation is unitarian, i.e. the *total energy* in time domain equals with that of the frequency spectrum:

$$\int_{-\infty}^{\infty} |\tilde{E}_a(t)|^2 dt = \int_{-\infty}^{\infty} |\mathcal{F}\{\tilde{E}_a\}|^2 d\nu \quad (379)$$

We get *average power* from this as follows:

$$\lim_{T \rightarrow \infty} \left[\frac{1}{T} \frac{v\varepsilon}{2} \int_{-T}^T |\tilde{E}_a(t)|^2 dt \right] = \lim_{T \rightarrow \infty} \left[\frac{1}{T} \frac{v\varepsilon}{2} \int_{-\infty}^{\infty} |\mathcal{F}\{\tilde{E}_a\}|_T^2 d\nu \right], \quad (380)$$

where the integral of the power spectrum can be seen on the right hand side. Thus the power spectral density is:

$$S(\nu) = \lim_{T \rightarrow \infty} \left[\frac{v\varepsilon}{2T} |\mathcal{F}\{\tilde{E}_a\}|_T^2 \right]. \quad (381)$$

12.4. Concept and types of coherence

In recapitulation we can state that coherence is a property characteristic of a given radiation, similarly to wavelength or intensity. Its decline is always connected to temporal uncertainties exhibited by the examined oscillation, and can be characterized by the stability of phase difference between two instants of given time difference. Three kinds of coherence can be distinguished: temporal/spatial/polarization. The latter refers to fluctuations of the polarization state of light as a function of time or position, but we do not discuss this case here. The most important is temporal coherence, which can be imagined as if we examined the phase state of a radiation in a given position of space, but in different instants, see Fig. 62. Such devices are amplitude-splitting interferometers (e.g. Michelson). As we saw earlier, the extent of temporal coherence depends on the width of the frequency spectrum of the light source.

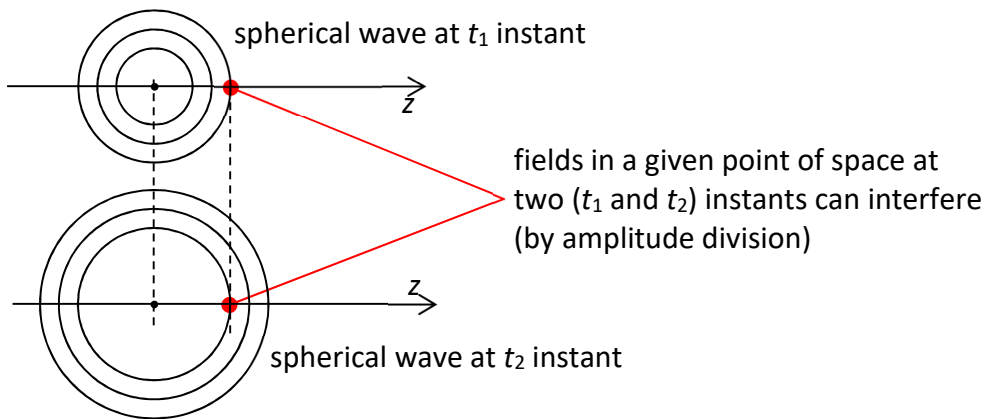


Fig. 62 Demonstration of temporal coherence.

The (longitudinal) coherence length of temporal coherence is:

$$l_L \triangleq v \cdot \tau_c \quad (382)$$

E.g.: $\lambda = 633 \text{ nm}$; $\Delta\lambda = 0.001 \text{ nm}$; $\Delta\nu = 0.75 \text{ GHz}$ (distance of longitudinal resonator modes, three modes); $l_L \approx 250 \text{ mm}$; $\tau_c \approx 0.85 \text{ nsec}$ (cca. 400000 period) according to equation (362).

This is why we cannot see standing waves when superimposing two independent laser sources: though interference takes place here too, but the pattern changes about every 1 nsec, the average of which we are only able to perceive.

Spatial coherence can be investigated by wavefront-splitting interferometers (see e.g. Young's two-slit experiment). In this case we have the light coming from two different spatial positions interfere, at the same instant. The *transverse coherence length* is a distance where the visibility of the interference drops below a specific value:

$$l_T \triangleq |\mathbf{r}_1 - \mathbf{r}_2| \quad (383)$$

By analogy of the frequency spectrum, spatial coherence is in connection with the so-called *angular spectrum*, that shows the angular extent in which plane wave components of different direction should be superimposed to obtain the given radiation. If the angle interval is Θ , i.e. the light source subtends an angle of Θ from the point where we investigate coherence, then

$$l_T \approx \frac{\lambda}{\Theta} ; [\Theta] = \text{rad}. \quad (384)$$

At formulating the above approximate equation we basically followed the straightforward method that was used to obtain equation (214) at the discussion of interference phenomena: between wavefronts being in phase at one of Young's two slits (that analyze the coherence) we require exactly λ path difference when measured at the other slit (see Fig. 63). From the formula it is clear that spatial coherence is not an absolute characteristics. Since the apparent angle of a light source decreases by the distance relative to it, the farther we go from a source, the more spatially coherent the EM wave will be. The exact equation will be derived in the Optics and photonics specialization, Examination of diffraction phase gratings measurement.

A finite frequency bandwidth is important at spatial coherence too, since completely monochromatic plane waves with random initial phases result in a temporally stationary speckle pattern. Here we will not deal with spatial coherence in any more detail.

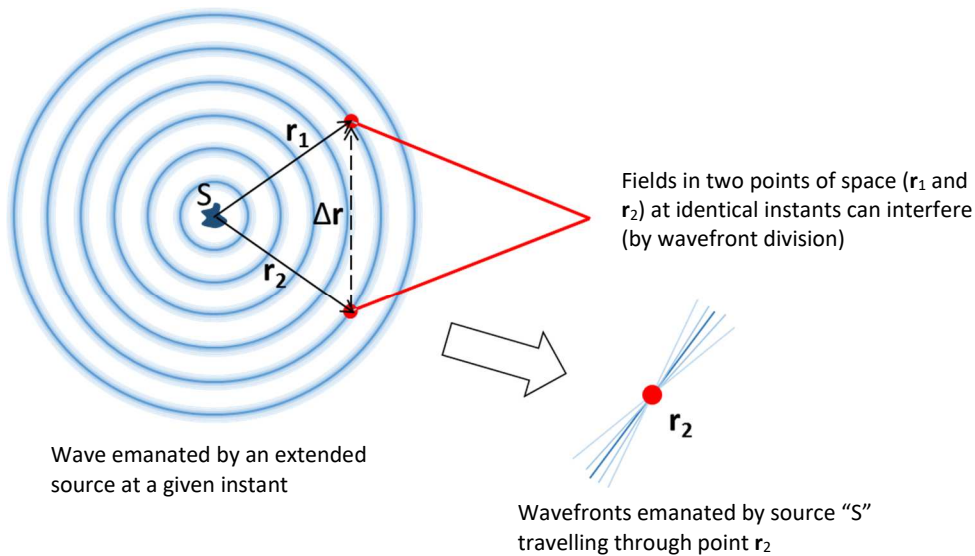


Fig. 63 Demonstration of spatial coherence.

13. POLARIZATION

Sources: [1], [6], [12]

Importance of polarization in light-matter interactions

We call an EM wave polarized, when the *direction* of field vectors makes spatially and temporally periodic oscillations. Most important examples when polarization is affected:

- reflection from the boundary of two materials is polarization-dependent (e.g. Brewster's effect)
- absorption of certain materials is polarization-dependent (dichroism)
- light scattering from materials is sensitive to polarization (see Rayleigh scattering)
- refractive index of anisotropic materials is polarization-dependent (see birefringence)
- optically active materials rotate the polarization (chiral molecules, e.g. sugar)

13.1. Polarized nature of light

When an electron returns to a lower energy state from higher one (relaxation), single atoms emit EM radiation of specific frequency with spherical wavefronts in the far field and a direction characteristics that corresponds to that of dipole radiation, see Fig. 64. Such radiations are characterized by an \mathbf{E} field vector that oscillates along a straight line at every point in space, with a direction that is constant in time.

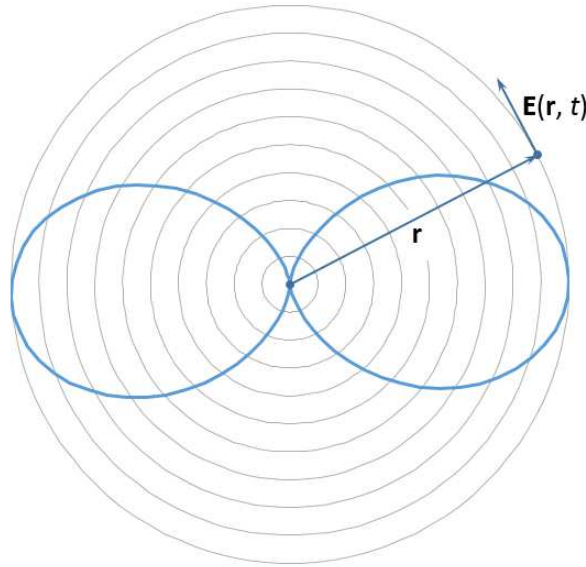


Fig. 64 Field in the far regime of a dipole source with indication of wavefronts, polarization and intensity direction characteristics.

If we examine a given propagation direction, this oscillation always *remains in a plane*, as in Fig. 65. Such a radiation is called as *linearly polarized*, thus light is naturally polarized on account of its way of origin. As an example let us consider the field vector of a harmonic plane wave propagating in the z -direction:

$$\tilde{\mathbf{E}}(t, z) = \mathbf{A} \cdot e^{i(\omega t - kz + \varphi)} \rightarrow \tilde{\mathbf{E}}(t, z) = \tilde{\mathbf{A}} \cdot e^{i(\omega t - kz)} \rightarrow \mathbf{E}(t, z) = \text{Re}\{\tilde{\mathbf{A}} \cdot e^{i(\omega t - kz)}\} \quad (385)$$

We denoted the vector amplitude by $\tilde{\mathbf{A}}$ in accordance with the traditional description of polarization phenomena.

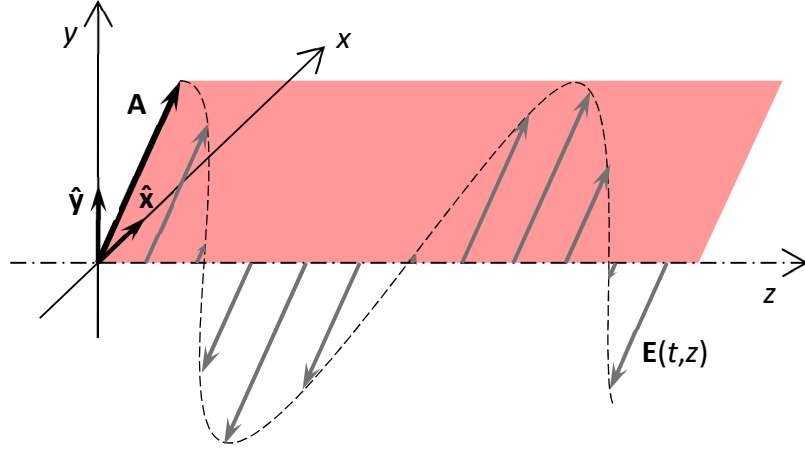


Fig. 65 Demonstration of a linearly polarized EM wave.

According to the above, each possible EM wave is composed of the sum of linearly polarized radiation. Furthermore, it is easy to prove that any linear polarization of arbitrary direction can be decomposed as the superposition of two, mutually perpendicular linearly polarized radiation, therefore any wave can be decomposed into two beams of mutually perpendicular polarization. For a plane wave let us inspect what kind of polarization states these can lead to.

$$\left. \begin{aligned} \tilde{\mathbf{E}}_x(t, z) &= \tilde{\mathbf{A}}_x \cdot e^{i(\omega t - kz)} \\ \tilde{\mathbf{E}}_y(t, z) &= \tilde{\mathbf{A}}_y \cdot e^{i(\omega t - kz)} \end{aligned} \right\} \rightarrow \tilde{\mathbf{E}}(t, z) = \tilde{\mathbf{A}}_x \cdot e^{i(\omega t - kz)} + \tilde{\mathbf{A}}_y \cdot e^{i(\omega t - kz)} \quad (386)$$

$$\tilde{\mathbf{E}}(t, z) = \tilde{A}_x \cdot \hat{\mathbf{x}} \cdot e^{i(\omega t - kz)} + \tilde{A}_y \cdot \hat{\mathbf{y}} \cdot e^{i(\omega t - kz)} \quad (387)$$

Now let us determine the shape of the curve the field vector (E_x, E_y) makes as a function of time in the x-y plane being at position z.

$$\begin{aligned} E_x &= \text{Re}\{\tilde{\mathbf{E}}_x(t, z)\} = |\mathbf{A}_x| \cos(\omega t - kz + \varphi_x) = A_x \cos(\omega t - kz + \varphi_x) \\ E_y &= \text{Re}\{\tilde{\mathbf{E}}_y(t, z)\} = |\mathbf{A}_y| \cos(\omega t - kz + \varphi_y) = A_y \cos(\omega t - kz + \varphi_y) \end{aligned} \quad (388)$$

$$\Phi \triangleq \omega t - kz + \varphi_x \quad \text{and} \quad \Delta\varphi \triangleq \varphi_y - \varphi_x \quad (389)$$

$$\begin{aligned} E_x &= A_x \cos(\Phi) \\ E_y &= A_y \cos(\Phi + \Delta\varphi) \end{aligned} \quad (390)$$

Our objective is to describe the connection of E_x, E_y with parameters A_x, A_y and $\Delta\varphi$, to which end we divide Φ out from the equations. The second equation of (390) leaves us this:

$$E_y = A_y (\cos(\Phi) \cos(\Delta\varphi) - \sin(\Phi) \sin(\Delta\varphi)) \quad (391)$$

$$\sin(\Phi) = \frac{\cos(\Phi) \cos(\Delta\varphi)}{\sin(\Delta\varphi)} - \frac{E_y}{A_y} \frac{1}{\sin(\Delta\varphi)} \quad (392)$$

$$\sin^2(\Phi) = \left(\frac{\cos(\Phi) \cos(\Delta\varphi)}{\sin(\Delta\varphi)} \right)^2 + \left(\frac{E_y}{A_y} \frac{1}{\sin(\Delta\varphi)} \right)^2 - 2 \frac{E_y \cos(\Phi) \cos(\Delta\varphi)}{A_y \sin^2(\Delta\varphi)} \quad (393)$$

And from the first equation of (390) this follows:

$$\frac{E_x}{A_x} = \cos(\Phi) \quad \text{and} \quad \left(\frac{E_x}{A_x} \right)^2 = \cos^2(\Phi) \quad (394)$$

$$1 - \left(\frac{E_x}{A_x}\right)^2 = \left(\frac{E_x \cos(\Delta\varphi)}{A_x \sin(\Delta\varphi)}\right)^2 + \left(\frac{E_y}{A_y \sin(\Delta\varphi)}\right)^2 - 2 \frac{E_y E_x \cos(\Delta\varphi)}{A_y A_x \sin^2(\Delta\varphi)} \quad (395)$$

$$1 - \left(\frac{E_x}{A_x}\right)^2 \equiv 1 - \left(\frac{E_x \sin(\Delta\varphi)}{A_x \sin(\Delta\varphi)}\right)^2 \quad (396)$$

$$1 = \left(\frac{E_x}{A_x \sin(\Delta\varphi)}\right)^2 + \left(\frac{E_y}{A_y \sin(\Delta\varphi)}\right)^2 - 2 \frac{E_y E_x \cos(\Delta\varphi)}{A_y A_x \sin^2(\Delta\varphi)} \quad (397)$$

$$\sin^2(\Delta\varphi) = \left(\frac{E_x}{A_x}\right)^2 + \left(\frac{E_y}{A_y}\right)^2 - 2 \cos(\Delta\varphi) \frac{E_x E_y}{A_x A_y}. \quad (398)$$

(398) is the general formula of ellipses, whose well-known form we get if $\Delta\varphi = \pm 90^\circ$. Thus the superposition of two beams of mutually perpendicular linear polarization propagating in the z-direction generally results in a field vector that moves along an ellipse, see Figs. 66-67.

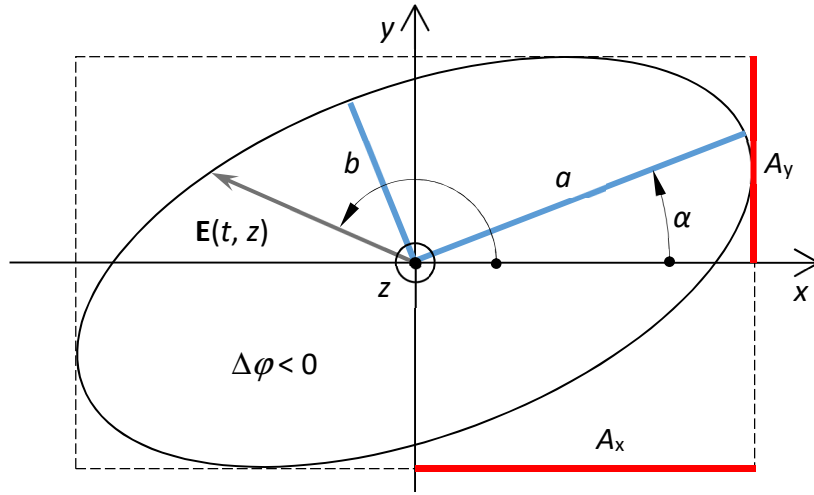


Fig. 66 Notation used for describing an elliptically polarized wave. The light propagates towards the viewer (a “physicist’s approach”).

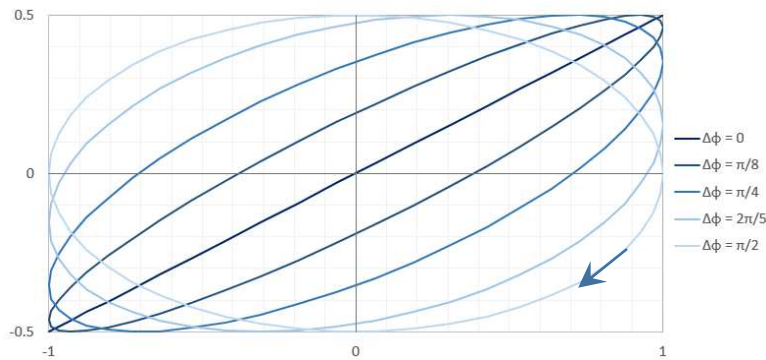


Fig. 67 Demonstration of different polarizations (note that $A_x \neq A_y$). If A_x and $A_y \neq 0$ as well as $\Delta\varphi = 0$ or π then the light is linearly polarized. When $\Delta\varphi < 0$ the field vector is rotating left ; for $\Delta\varphi > 0$ it is rotating right (if the light propagates towards us). If $A_x = A_y$ and $\Delta\varphi = \pm\pi/2$, we get circularly polarized light.

The direction of polarization is determined traditionally by the chirality of the field helix: LCP if it is left-sided (rotates counterclockwise when viewed from the front), RCP if it is right-sided (rotates clockwise when viewed from the front, see Fig. 68).

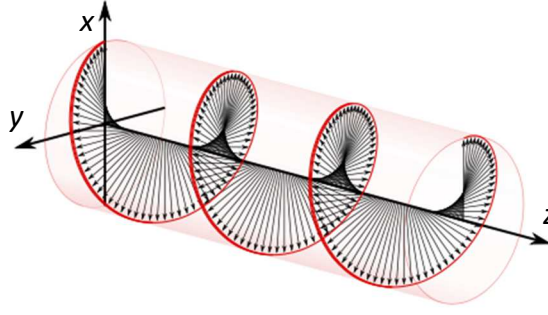


Fig. 68 Demonstrating a right-handed circularly polarized wave (RCP), wikipedia.

The major and minor axes of the ellipse (a and b) can be calculated in the following way [15]:

$$\begin{aligned} a^2 &= \frac{1}{2} \left(A_x^2 + A_y^2 + \sqrt{(A_x^2 - A_y^2)^2 + 4 \cos^2(\Delta\phi) A_x^2 A_y^2} \right) \\ b^2 &= \frac{1}{2} \left(A_x^2 + A_y^2 - \sqrt{(A_x^2 - A_y^2)^2 + 4 \cos^2(\Delta\phi) A_x^2 A_y^2} \right). \end{aligned} \quad (399)$$

From this it is relatively easy to see that:

$$a^2 + b^2 = A_x^2 + A_y^2 \quad (400)$$

Since it is not completely straightforward, now let us consider how the intensity of an elliptically polarized radiation can be determined in the general case. Let us assume that the examined wave is a superposition of two plane waves travelling in the same direction with mutually perpendicular field vectors $\mathbf{E}_x(\mathbf{r}, t)$ and $\mathbf{E}_y(\mathbf{r}, t)$. In this case the resultant Poynting vector is:

$$\mathbf{S}(\mathbf{r}, t) = (\mathbf{E}_x(\mathbf{r}, t) + \mathbf{E}_y(\mathbf{r}, t)) \times (\mathbf{H}_x(\mathbf{r}, t) + \mathbf{H}_y(\mathbf{r}, t)), \quad (401)$$

where a normal \mathbf{H}_x is associated with \mathbf{E}_x , likewise for \mathbf{E}_y and \mathbf{H}_y . Expanding the vectorial product we get terms such as $\mathbf{E}_x \times \mathbf{H}_y$ and $\mathbf{E}_y \times \mathbf{H}_x$. Since these vectorially multiplied vectors are parallel, both of these yield zero. Hence, the result is the sum of the two Poynting vectors:

$$\mathbf{S}(\mathbf{r}, t) = \mathbf{E}_x(\mathbf{r}, t) \times \mathbf{H}_x(\mathbf{r}, t) + \mathbf{E}_y(\mathbf{r}, t) \times \mathbf{H}_y(\mathbf{r}, t) \quad (402)$$

Since these point in the same direction, the intensities can be summed up:

$$I = I_x + I_y. \quad (403)$$

Thus, using the above-introduced notation:

$$I = v\varepsilon \frac{a^2 + b^2}{2} = v\varepsilon \frac{A_x^2 + A_y^2}{2} \quad \text{or} \quad I = v\varepsilon \frac{\tilde{A}_x \tilde{A}_x^* + \tilde{A}_y \tilde{A}_y^*}{2}. \quad (404)$$

13.2. Representation by Jones vectors

The complex amplitudes of the two constituent components of a polarized wave is usually united in the so-called Jones vector. In general, the x - and y -direction linearly polarized oscillations are used as bases:

$$\mathbf{J} \triangleq \begin{bmatrix} \tilde{A}_x \\ \tilde{A}_y \end{bmatrix} \quad (405)$$

The intensity of the beam according to (404):

$$I = \frac{v\varepsilon}{2} \mathbf{J} \cdot \mathbf{J}^* \quad (406)$$

Here are some examples if we use a phase convention corresponding to (390), viz. the phase of \tilde{A}_x serves as a reference, and intensity is normalized as follows:

$$\mathbf{J} \cdot \mathbf{J}^* = 1 \quad (407)$$

Linearly polarized light:

$$\mathbf{J} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} ; \quad \mathbf{J} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} ; \quad \mathbf{J} = \begin{bmatrix} \cos(\alpha) \\ \sin(\alpha) \end{bmatrix} \quad (408)$$

Left-handed circularly polarized light:

$$\mathbf{J}_L = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix} \quad (409)$$

Right-handed circularly polarized light:

$$\mathbf{J}_R = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ +i \end{bmatrix} \quad (410)$$

We can consider two polarizations to be orthogonal if their scalar product is zero:

$$\mathbf{J}_1 \cdot \mathbf{J}_2^* = \tilde{A}_{1x}\tilde{A}_{2x}^* + \tilde{A}_{1y}\tilde{A}_{2y}^* = 0, \quad (411)$$

e.g. mutually perpendicular linearly polarized beams of identical magnitude or left- and right-handed circularly polarized beams of identical magnitude. Orthogonal Jones vectors can be properly used as new polarization bases. Any elliptically polarized beam can be described as the linear combination of two orthogonally polarized beams in the following way:

$$\mathbf{J} \rightarrow \mathbf{J}' = c_1 \mathbf{J}_1 + c_2 \mathbf{J}_2 \quad (412)$$

where:

$$c_1 = \mathbf{J} \cdot \mathbf{J}_1^* \quad \text{and} \quad c_2 = \mathbf{J} \cdot \mathbf{J}_2^* \quad (413)$$

and the magnitude of $\mathbf{J}_{1,2}$ is normalized in the following sense:

$$\mathbf{J}_1 \cdot \mathbf{J}_1^* = 1 \quad \text{and} \quad \mathbf{J}_2 \cdot \mathbf{J}_2^* = 1 \quad (414)$$

Hence, an arbitrary polarization can also be described by a left- and right-handed circularly polarized beam. E.g. for a beam that is linearly polarized in the x-direction:

$$c_1 = \mathbf{J} \cdot \mathbf{J}_L^* = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix}^* = \frac{1}{\sqrt{2}} \quad \text{and} \quad c_2 = \mathbf{J} \cdot \mathbf{J}_R^* = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ i \end{bmatrix}^* = \frac{1}{\sqrt{2}} \quad (415)$$

that is:

$$\mathbf{J} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \rightarrow \mathbf{J}' = \frac{1}{\sqrt{2}} \mathbf{J}_L + \frac{1}{\sqrt{2}} \mathbf{J}_R = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \quad (416)$$

A description in circularly polarized bases is typically applied when discussing optically active media, whereas linearly polarized bases are used e.g. at Fresnel reflection.

13.3. Description of anisotropic optical elements by Jones matrices

$$\mathbf{J}' = \mathbf{T} \cdot \mathbf{J} \quad (417)$$

E.g. a linear polarization filter of x-direction (abandoning any constant phase shifts):

$$\mathbf{T} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad (418)$$

E.g. $\lambda/4$ phase retarding plate (with principal axes lying in the x-y direction):

$$\mathbf{T} = \begin{bmatrix} 1 & 0 \\ 0 & \pm i \end{bmatrix} \quad (419)$$

E.g. $\lambda/2$ phase retarding plate (with principal axes lying in the x-y direction):

$$\mathbf{T} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (420)$$

Matrix of a rotation (coordinate transformation for rotation by angle β):

$$\mathbf{T} = \begin{bmatrix} \cos \beta & -\sin \beta \\ \sin \beta & \cos \beta \end{bmatrix} \quad (421)$$

In case the examined optical system does not contain any losses (in the form of absorption, scattering, Fresnel reflection), i.e. the transmittance is unity, then the power of the exiting wave being the result of a polarization transformation will be the same as that of the input wave. For this reason the Jones matrix is *unitarian*, which can be put mathematically as the adjoint of the Jones matrix is identical to its inverse:

$$\mathbf{T} \cdot \mathbf{T}^+ = \mathbf{1} ; \mathbf{T}^+ \triangleq \mathbf{T}^{T*} \quad (422)$$

For all eight complex parameters of the Jones matrix this gives three independent equations:

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} T_{11}^* & T_{21}^* \\ T_{12}^* & T_{22}^* \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (423)$$

$$\left. \begin{aligned} T_{11}T_{11}^* + T_{12}T_{12}^* &= 1 \\ T_{11}T_{21}^* + T_{12}T_{22}^* &= 0 \\ T_{21}T_{21}^* + T_{22}T_{22}^* &= 1 \end{aligned} \right\} \quad (424)$$

The other equation equalling zero is the same as the one already written above. Since this gives six complex equations, the Jones matrix can be described by a total of $8 - 6 = 2$, i.e. two independent parameters. These are usually the following: 1) phase shift ($\Delta\varphi$) between the two orthogonal states, 2) rotation of the coordinate system (β).

Besides, in such cases the following is also true:

$$|\det(\mathbf{T})| = 1 \quad (425)$$

It was first proven in the 1990s that any arbitrary unitarian Jones matrix, i.e. an arbitrary polarization transformation can be implemented by an optical device composed of $\lambda/4 + \lambda/2 + \lambda/4$ phase retarder plates, in which the plates can be rotated around the optical axis [16].

13.4. Light including an unpolarized component (supplementary)

Description by Stokes vectors (+ Müller matrices)

A perfect analogy with quantum mechanics:

Jones vector \leftrightarrow state vector

Jones matrix \leftrightarrow matrix of an operator

polarization \leftrightarrow spin

polarized EM wave \leftrightarrow coherent quantum state (superposition of eigenstates)

unpolarized light \leftrightarrow mixed state (incoherent sum of eigenstates)

Poincaré sphere \leftrightarrow Bloch sphere

14. PROPAGATION OF PLANE WAVES IN ANISOTROPTIC MEDIA

Sources: [3]

We presented in Chapter 1 that within the fundamental equations of electrodynamics (Maxwell's equations) the interaction between matter and the EM wave can be macroscopically described by the relative dielectric permittivity and permeability, introducing the material-dependent \mathbf{D} and \mathbf{H} quantities. In materials lacking long-range order (e.g. amorphous solids, liquids and gases) ϵ and μ are scalar quantities, implying that $\mathbf{D}-\mathbf{E}$ and $\mathbf{H}-\mathbf{B}$ point in the same direction, respectively. However, there also exist media, in which the polarization of an atom or molecule does not remain uniform as we change the direction of the electric / magnetic field vectors – these are called as *optically anisotropic materials*, for which single crystals serve as typical examples. Since the applicability of these is quite wide, in the following we briefly discuss their most important properties and the methods used for their physical description. For the sake of simplicity, we will perform our calculations in homogeneous, insulating media ($\sigma = 0$), where charge density is zero. Since most optical materials are non-magnetic, in this discussion we will also assume that $\mu = \mu_0$.

14.1. Mutual orientation of field vectors

The subject of our investigation is electric anisotropy. In the completely general case every component of the electric field affect every component of the dielectric displacement vector. In a linear approximation this can be written in the following form:

$$\left. \begin{aligned} D_x &= \epsilon_{xx}E_x + \epsilon_{xy}E_y + \epsilon_{xz}E_z \\ D_y &= \epsilon_{yx}E_x + \epsilon_{yy}E_y + \epsilon_{yz}E_z \\ D_z &= \epsilon_{zx}E_x + \epsilon_{zy}E_y + \epsilon_{zz}E_z \end{aligned} \right\} \rightarrow \mathbf{D} = \boldsymbol{\epsilon} \cdot \mathbf{E}, \quad (426)$$

where $\boldsymbol{\epsilon}$ is the dielectric permittivity tensor. It can be proven (see [3]) that in a non-magnetic medium lacking sources and absorption ($\sigma = 0$; $\rho = 0$; $\mu_r = 1$) the continuity equation that describes the conservation of energy

$$\text{div } \mathbf{S} = -\frac{\partial w}{\partial t} \quad (427)$$

can only be satisfied, if the permittivity tensor is symmetric, i.e.:

$$\epsilon_{xy} = \epsilon_{yx} ; \epsilon_{xz} = \epsilon_{zx} ; \epsilon_{yz} = \epsilon_{zy}, \quad (428)$$

where \mathbf{S} is the Poynting vector that characterizes the flow of power density, and w denotes the total energy density of the EM field. Depending on the chosen base vectors of our Cartesian coordinate system used to express (426), the parameters of $\boldsymbol{\epsilon}$ will always be different. From linear algebra we have learned that for the formulation of matrices being symmetric to their main diagonal there always exists a specific base vector system, in which the matrix is diagonal (principal axis transformation):

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_x & 0 & 0 \\ 0 & \epsilon_y & 0 \\ 0 & 0 & \epsilon_z \end{bmatrix} \quad (429)$$

Principal axes usually coincide with the main axes of symmetry of the given crystal, hence their direction is strictly connected to the carrier medium orientation. Crystals with monoclinic and triclinic unit cell (see later) are exceptions to this. Since dielectric permittivity depends on the

EM wave frequency (dispersion), in such crystals the orientation of principal axes varies with frequency. The phenomenon is known as *axial dispersion* (not discussed any further here).

Let us examine the orientation of field vectors in anisotropic media relative to each other. Our investigations will be constrained to the rigorous, wave-optical study of plane waves (i.e. the results are generally applicable within the scope of validity of geometrical optics). Since all field properties are complex, the tilde denoting this will be omitted below:

$$\begin{aligned}\mathbf{E}(\mathbf{r}, t) &= \mathbf{E}_0 \cdot e^{i(\omega t - \mathbf{k} \cdot \mathbf{r})} \\ \mathbf{B}(\mathbf{r}, t) &= \mathbf{B}_0 \cdot e^{i(\omega t - \mathbf{k} \cdot \mathbf{r})}\end{aligned}\quad (430)$$

We start from the macroscopic Maxwell's equations:

$$\left. \begin{aligned}\text{I. } \text{curl } \mathbf{H} &= \frac{\partial \mathbf{D}}{\partial t} \\ \text{II. } \text{curl } \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\ \text{III. } \text{div } \mathbf{B} &= 0 \\ \text{IV. } \text{div } \mathbf{D} &= 0\end{aligned} \right\} \quad (431)$$

In case of a harmonic plane wave the law of induction is:

$$\text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \Rightarrow \mathbf{k} \times \mathbf{E} = \omega \mathbf{B} \quad (432)$$

Also the circuital law:

$$\text{curl } \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} \Rightarrow \mathbf{k} \times \mathbf{H} = -\omega \mathbf{D} \quad (433)$$

Since $\mathbf{B} = \mu_0 \mathbf{H}$, i.e. these two vectors share the same direction, it follows from the above two equations that \mathbf{B} is normal to \mathbf{E} and \mathbf{D} too. Besides, \mathbf{B} is also normal to \mathbf{k} , thus $\mathbf{B} \perp \mathbf{D} \perp \mathbf{k}$. The Poynting vector is $\mathbf{S} = \mathbf{E} \times \mathbf{H}$, so $\mathbf{H} \perp \mathbf{E} \perp \mathbf{S}$. The above is summarized in Fig. 69 for the general case. If \mathbf{E} happens to lie in the direction of a principal axis, according to (429) it is parallel to \mathbf{D} .

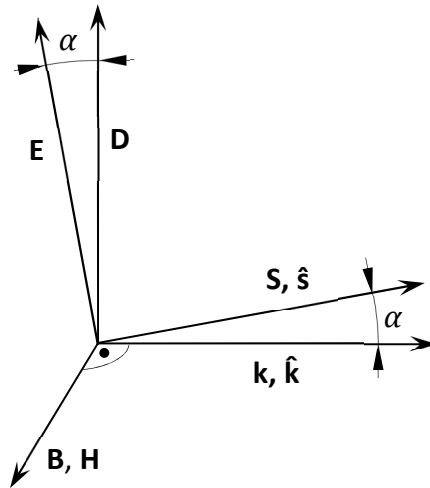


Fig. 69 Connection of the electric/magnetizing field, dielectric displacement, magnetic field, wave vector and Poynting vector in case of electric anisotropy.

14.2. Phase and ray velocity

Let us examine the propagation speed of plane waves in anisotropic media. Let us introduce the $\hat{\mathbf{k}}$ unit vector pointing in the direction of the wave vector ([3] denotes this by $\hat{\mathbf{s}}$, which might cause some confusion, this is why we use instead a symbol easier to recall):

$$\mathbf{k} = \frac{\omega}{v_p} \hat{\mathbf{k}} \quad (434)$$

Here v_p denotes *phase velocity* (the speed of wavefronts normal to the direction of $\hat{\mathbf{k}}$). It can be shown that (the derivation see in [3]):

$$w = \frac{\mathbf{S}}{v_p} \cdot \hat{\mathbf{k}} = \frac{\mathbf{S} \cdot \cos(\alpha)}{v_p} \quad (435)$$

From this the velocity of energy propagation (*ray velocity*), see Fig. 70 and 71:

$$\mathbf{v}_r = \frac{\mathbf{S}}{w} \Rightarrow v_p = v_r \cdot \cos(\alpha). \quad (436)$$

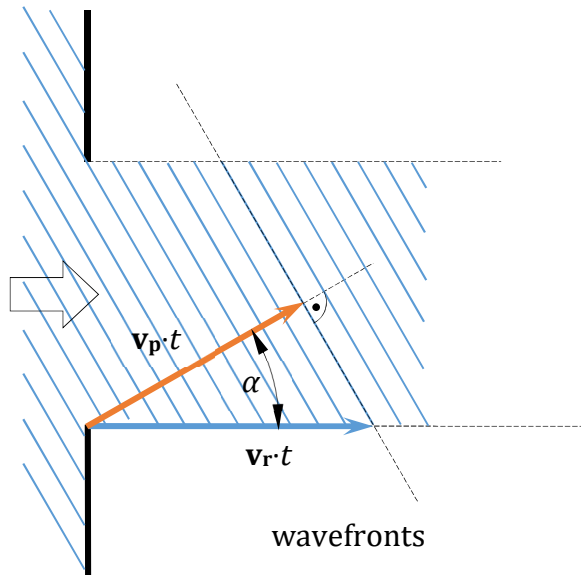


Fig. 70 Propagation of a plane wave in an anisotropic medium.

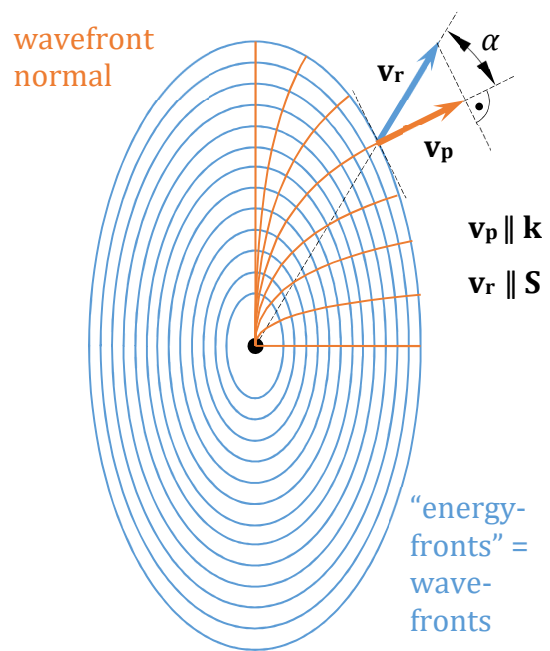


Fig. 71 Far-field of a point source in an anisotropic medium.

14.3. Fresnel's equation of wave normals

Now we are going to determine the value of v_p in different directions of \mathbf{k} . For this first we have to calculate the \mathbf{E} eigenvectors and eigenvalues corresponding to this propagation direction. Substituting (432) into (433):

$$\mathbf{k} \times \left(\mathbf{k} \times \frac{1}{\mu_0 \omega} \right) = -\omega \mathbf{D} \quad (437)$$

$$\hat{\mathbf{k}} \times (\hat{\mathbf{k}} \times \mathbf{E}) = -\mu_0 v_p^2 \mathbf{D} \quad (438)$$

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) \equiv (\mathbf{A} \cdot \mathbf{C}) \cdot \mathbf{B} - (\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C} \quad (439)$$

$$\hat{\mathbf{k}} \times (\hat{\mathbf{k}} \times \mathbf{E}) = (\hat{\mathbf{k}} \cdot \mathbf{E}) \cdot \hat{\mathbf{k}} - (\hat{\mathbf{k}} \cdot \hat{\mathbf{k}}) \cdot \mathbf{E} = (\hat{\mathbf{k}} \cdot \mathbf{E}) \cdot \hat{\mathbf{k}} - \mathbf{E} \quad (440)$$

By substituting this into (438):

$$\mathbf{D} = \frac{1}{\mu_0 v_p^2} (\mathbf{E} - (\hat{\mathbf{k}} \cdot \mathbf{E}) \cdot \hat{\mathbf{k}}) \quad (441)$$

Comparing (426) and (441) we arrive at an eigenvalue equation. This is the wave equation with a plane-wave ansatz in anisotropic case (see (31) Helmholtz equation):

$$\boldsymbol{\epsilon} \cdot \mathbf{E} = \frac{1}{\mu_0 v_p^2} (\mathbf{E} - \hat{\mathbf{k}} \cdot (\hat{\mathbf{k}} \cdot \mathbf{E})). \quad (442)$$

Let us assume that for the sake of simplicity we have performed principal axis transformation on $\boldsymbol{\epsilon}$, so that its form now corresponds to (429). With this (442) simplifies to the following:

$$\epsilon_i E_i = \frac{1}{\mu_0 v_p^2} (E_i - \hat{k}_i \cdot (\hat{\mathbf{k}} \cdot \mathbf{E})), \quad \text{where } i = x, y, z \quad (443)$$

Expressing the field components from this we get the following:

$$E_i = \frac{\hat{k}_i \cdot (\hat{\mathbf{k}} \cdot \mathbf{E})}{1 - \epsilon_i \mu_0 v_p^2} \quad (444)$$

Multiplying all three equations by \hat{k}_i and summing them up we obtain right the $\hat{\mathbf{k}} \cdot \mathbf{E}$ scalar product, what hence gives exactly 1 if we divide it by $\hat{\mathbf{k}} \cdot \mathbf{E}$:

$$\frac{\hat{k}_x^2}{1 - \epsilon_x \mu_0 v_p^2} + \frac{\hat{k}_y^2}{1 - \epsilon_y \mu_0 v_p^2} + \frac{\hat{k}_z^2}{1 - \epsilon_z \mu_0 v_p^2} = 1 \quad (445)$$

Our former statement suggests if \mathbf{E} is in the direction of a principal axis then it must be parallel with \mathbf{D} , from which it also follows that in such cases the ray and phase velocities also have a common direction and magnitude. These are called *principal velocities* with values of:

$$u_x \triangleq \frac{1}{\sqrt{\epsilon_x \mu_0}} ; \quad u_y \triangleq \frac{1}{\sqrt{\epsilon_y \mu_0}} ; \quad u_z \triangleq \frac{1}{\sqrt{\epsilon_z \mu_0}} \quad (446)$$

We chose $u_{x,y,z}$ as a notation for principal velocities to avoid any potential confusion with the x, y, z components of velocity vector \mathbf{v} . It is important to understand that velocity u_x corresponds to a field of E_x direction, u_y to E_y direction, and u_z to E_z direction. If \mathbf{E} is parallel with a principal axis, then \mathbf{D} goes the same way, and $v_p = u_{x,y,z}$. By this we get from (445):

$$\frac{\hat{k}_x^2}{1 - \frac{v_p^2}{u_x^2}} + \frac{\hat{k}_y^2}{1 - \frac{v_p^2}{u_y^2}} + \frac{\hat{k}_z^2}{1 - \frac{v_p^2}{u_z^2}} = 1 \quad (447)$$

Since $\hat{k}_x^2 + \hat{k}_y^2 + \hat{k}_z^2 \equiv 1$, the above equation can be converted into the following form:

$$\frac{\hat{k}_x^2}{u_x^2 - v_p^2} + \frac{\hat{k}_y^2}{u_y^2 - v_p^2} + \frac{\hat{k}_z^2}{u_z^2 - v_p^2} = 0. \quad (448)$$

This is Fresnel's equation of wave normals. If $u_x \neq u_y \neq u_z$ then (448) results in a quadratic equation in terms of v_p^2 (by multiplying the above expression with the product of the denominators), which means that in a general case every propagation direction has two different velocities (v'_p and v''_p). Substituting these into (444) we get real numbers for the ratios of the

two eigenvector components ($E'_x:E'_y:E'_z$ and $E''_x:E''_y:E''_z$), meaning that eigenvectors \mathbf{E}' and \mathbf{E}'' are *linearly polarized*. According to relationship (320) the same holds for \mathbf{D}' and \mathbf{D}'' too. Because of the above, the optical property described by (429) is called *birefringence*.

Principal velocities u_x ; u_y ; u_z do not indicate the velocity of wavefronts propagating in the corresponding x , y , z direction! E.g. if the phase velocity is of z direction ($\hat{k}_x = \hat{k}_y = 0$ and $\hat{k}_z = 1$), then $v_p = u_x$ and $v''_p = u_y$. Generally, if \mathbf{E} is parallel with the x , y , z principal axis, then $\mathbf{D} \parallel \mathbf{E}$, i.e. $\hat{\mathbf{k}}\mathbf{E} = 0$, thus based on (443) $v_p = u_{x,y,z}$.

The careful reader may have noticed that at deriving Fresnel's equation of wave normals we expressed (444) by dividing both sides of the previous equation by $(1 - \varepsilon_i \mu_0 v_p^2)$ without examining what happens if this expression gives zero. Based on equation-triplet (443) this is only possible when \mathbf{E} is parallel with any of the principal axes, e.g. x . Then $E_y = E_z = 0$, i.e. according to the second two equations $\hat{\mathbf{k}} \cdot \mathbf{E} = 0$. For this reason, the first equation can only be satisfied if $v_p = u_x$, i.e. when the zeros of the denominators are exact solutions of equation (448), which can be only used with these cases excluded.

Now let us determine the relative direction of the two dielectric displacement vectors corresponding to a given propagation direction. According to (444)

$$D_x = \varepsilon_x E_x = \varepsilon_x \frac{\hat{k}_x \cdot (\hat{\mathbf{k}} \cdot \mathbf{E})}{1 - \frac{v_p^2}{u_x^2}} = \varepsilon_x u_x^2 \frac{\hat{k}_x \cdot (\hat{\mathbf{k}} \cdot \mathbf{E})}{u_x^2 - v_p^2} = \frac{1}{\mu_0} \frac{\hat{k}_x \cdot (\hat{\mathbf{k}} \cdot \mathbf{E})}{u_x^2 - v_p^2} \quad (449)$$

Substituting v'_p and v''_p velocities into (449), and taking the scalar product of the resulting two dielectric displacement vectors:

$$\mathbf{D}' \cdot \mathbf{D}'' = \frac{1}{\mu_0^2} \left[\frac{\hat{k}_x(\hat{\mathbf{k}}\mathbf{E}')}{u_x^2 - v'^2_p} \cdot \frac{\hat{k}_x(\hat{\mathbf{k}}\mathbf{E}'')}{u_x^2 - v''^2_p} + \frac{\hat{k}_y(\hat{\mathbf{k}}\mathbf{E}')}{u_y^2 - v'^2_p} \cdot \frac{\hat{k}_y(\hat{\mathbf{k}}\mathbf{E}'')}{u_y^2 - v''^2_p} + \frac{\hat{k}_z(\hat{\mathbf{k}}\mathbf{E}')}{u_z^2 - v'^2_p} \cdot \frac{\hat{k}_z(\hat{\mathbf{k}}\mathbf{E}'')}{u_z^2 - v''^2_p} \right], \quad (450)$$

which after rearrangement and considering (448) leaves us:

$$\mathbf{D}' \cdot \mathbf{D}'' = \frac{1}{\mu_0^2} \frac{(\hat{\mathbf{k}}\mathbf{E}')(\hat{\mathbf{k}}\mathbf{E}'')}{v'^2_p - v''^2_p} \cdot \left[\frac{\hat{k}_x^2}{u_x^2 - v'^2_p} + \frac{\hat{k}_y^2}{u_y^2 - v'^2_p} + \frac{\hat{k}_z^2}{u_z^2 - v'^2_p} - \frac{\hat{k}_x^2}{u_x^2 - v''^2_p} - \frac{\hat{k}_y^2}{u_y^2 - v''^2_p} - \frac{\hat{k}_z^2}{u_z^2 - v''^2_p} \right] = 0 \quad (451)$$

From this it follows that the two \mathbf{D} eigenvectors pertaining to a specific propagation direction are always normal to each other. This same relationship can be proven for the \mathbf{E}' and \mathbf{E}'' vectors too, both of which are normal to the Poynting vector.

Now let us determine the Poynting vector directions that correspond to a given wavefront normal. Introducing the g auxiliary parameter:

$$g \triangleq \frac{1}{\left(\frac{\hat{k}_x}{u_x^2 - v_p^2} \right)^2 + \left(\frac{\hat{k}_y}{u_y^2 - v_p^2} \right)^2 + \left(\frac{\hat{k}_z}{u_z^2 - v_p^2} \right)^2}, \quad (452)$$

it can be proven (see [3]) that

$$v_r^2 = \frac{g}{v_p^2} + v_p^2 \quad (453)$$

Furthermore

$$\hat{s}_i = \frac{\hat{k}_i}{v_p v_r} \left(v_p^2 - \frac{g}{u_i^2 - v_p^2} \right), \quad (454)$$

where we introduced the \hat{s} unit vector that points in the direction of the Poynting vector ($\hat{s} \triangleq \mathbf{S}/S$). Since two v_p and v_r velocities correspond to a given wavefront normal, the above equation implies that there are *two different Poynting vector directions* pertaining to the \mathbf{D}' , \mathbf{D}'' eigenvectors calculated for a specific wavefront normal! The connection between ray and phase velocity vectors of plane waves is illustrated in Fig. 72.

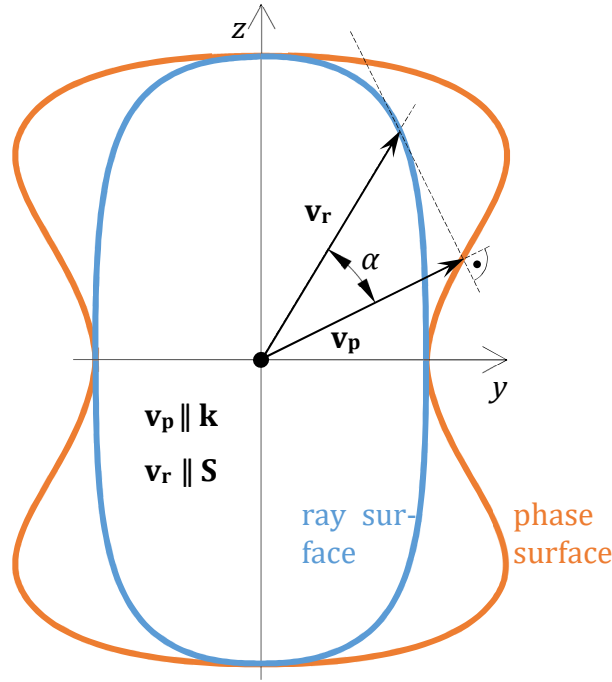


Fig. 72 Ray and phase surfaces of a plane wave in the general case, as well as the connection between phase and ray velocities.

Ray surfaces are usually not ellipsoids but fourth-order surfaces, and phase surfaces are of sixth-order, see (448). Corresponding to the two eigenvalues there are 2-2 such surfaces. Since a ray surface indicates the distance that energy can travel in t time propagating in a given direction, if we place a point source in the origin of the figure, the blue curve will indicate the discontinuity surfaces of the EM wave (kind of energy fronts). According to our previous findings, these coincide with the wavefronts.

In so-called biaxial crystals (see later) for two special wavefront normals (lying in the direction of so-called optic axes) a zero value appears in the denominator of (454). Owing to this singularity, the direction of the Poynting vector is indefinite, which leads to the interesting phenomenon of *conical refraction*.

14.4. Index ellipsoid

The energy density of an EM wave can be written as:

$$w = \mathbf{E} \cdot \mathbf{D} \quad (455)$$

In the principal axis-transformed system we are now analyzing, this can be formulated as:

$$w = \frac{D_x^2}{\varepsilon_x} + \frac{D_y^2}{\varepsilon_y} + \frac{D_z^2}{\varepsilon_z} \quad (456)$$

Now let us examine how the direction and magnitude of \mathbf{D} can be determined for plane waves travelling in different directions but having identical energy density. For the sake of simplicity let it be $w := 1$.

$$\frac{D_x^2}{\varepsilon_x} + \frac{D_y^2}{\varepsilon_y} + \frac{D_z^2}{\varepsilon_z} = 1 \quad (457)$$

This results in an ellipsoid surface called the *index ellipsoid*, see Fig. 73. Keep in mind that \mathbf{D}' , \mathbf{D}'' and $\hat{\mathbf{k}}$ form an orthogonal vector triplet. It can be proven that for a wave vector of given direction the two \mathbf{D} eigenvectors can be determined according to the below figure.

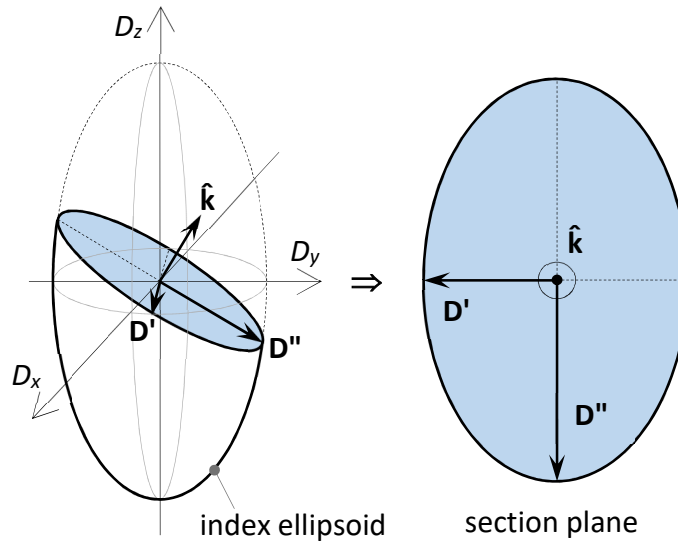


Fig. 73 Index ellipsoid (ellipsoid of wave normals, indicatrix), and the determination of polarization eigenstates.

The length of the ellipsoid half axes are respectively (according to Maxwell's formula these are proportional to the refractive index, hence the name):

$$\sqrt{\varepsilon_x} ; \sqrt{\varepsilon_y} ; \sqrt{\varepsilon_z}, \quad (458)$$

from which we get right the reciprocal of principal velocities after multiplying by $\sqrt{\mu_0}$. Based on this, one can prove in the general case that multiplying the magnitude of \mathbf{D}' , \mathbf{D}'' vectors determined above by $\sqrt{\mu_0}$ (for $w = 1$ normalization) we obtain exactly the reciprocal of phase velocities $1/v'_p$ and $1/v''_p$. If we further multiply this with c , then we get the refractive index corresponding to the \mathbf{D} vector pointing in the examined direction (since $n = c/v$).

We still have to mention the observation that a general ellipsoid has two planar sections, in which the cross section is *circular*. The normal vectors of these are called *optic axes* of the crystal. The anisotropy of crystals exhibiting such characteristics is called *biaxial birefringence*. Should the index ellipsoid be cylindrically symmetric (say around the z -axis), then these two circular sections degenerate into one, whose normal, i.e. the optic axis points in the z -direction. The anisotropy of such materials is called *uniaxial birefringence*.

14.5. Essentials in crystallography

The unit cell of a crystal is the smallest volumetric unit by periodically repeating which we can fill the whole space without gaps and overlapping. It can be proven mathematically, that crystal systems existing in three-dimensional space can have a total of 14 kinds of unit cells. These are called as Bravais lattices. The seven groups that can be formed of them by symmetry considerations are presented in the below table.

Crystal system	Principal axes x, y, z	Index ellipsoid	Optical classification	Application examples
Triclinic	CCC	general ellipsoid	biaxial	?
Monoclinic	CCF	general ellipsoid	biaxial	LYSO
Orthorombic	FFF	general ellipsoid	biaxial	BaSO ₄ , HgCl ₂ , KTP
Trigonal	FRR	circular symmetry	uniaxial	SiO ₂ , KDP, LiNbO ₃ , Al ₂ O ₃ , BBO
Tetragonal	FRR	circular symmetry	uniaxial	SiO ₂ , ADP, TeO ₂
Hexagonal	FRR	circular symmetry	uniaxial	SiO ₂ , CaCO ₃
Cubic	RRR	sphere	isotropic	NaCl, CaF ₂ , BGO, C (diamond)

Tab. 3 Crystalline groups by symmetry and their characteristics, indicating the most important materials used in optics. C – axial dispersion; F – fixed principal axis direction; R – freely-rotating or indeterminate principal axis (based on [3]); Note: quartz is able to crystallize in several different crystalline groups.

14.6. Uniaxial birefringence

Let the optic axis of the crystal be the z -axis. Then $u_x = u_y \triangleq u_o$ and $u_z \triangleq u_e$, where u_o is called as ordinary, and u_e as extraordinary principal velocity. Our aim is to determine the direction dependence of phase velocities. We may recall that at formulating equation (444) we divided both sides with $(1 - \varepsilon_i \mu_0 v_p^2)$, and at solving the Fresnel equation we excluded cases of $v_p = u_{x,y,z}$. As we will see below, at uniaxial birefringence one of the solutions will yield exactly such a result. Let us then take one step back and start from (443). Substituting the freshly introduced ordinary and extraordinary principal velocities into this we get:

$$\left. \begin{aligned} E_x &= \frac{u_o^2}{v_p^2} (E_x - \hat{k}_x \cdot (\hat{\mathbf{k}} \cdot \mathbf{E})) \\ E_y &= \frac{u_o^2}{v_p^2} (E_y - \hat{k}_y \cdot (\hat{\mathbf{k}} \cdot \mathbf{E})) \\ E_z &= \frac{u_e^2}{v_p^2} (E_z - \hat{k}_z \cdot (\hat{\mathbf{k}} \cdot \mathbf{E})) \end{aligned} \right\} \quad (459)$$

Now we look for such a solution where the electric field vector always lies in the x - y plane, thus $E_z = 0$. Then based on the third equation: $\hat{\mathbf{k}} \cdot \mathbf{E} = 0$, i.e. the wave vector is normal to \mathbf{E} . For this reason *both* first equations result in the same solution:

$$1 = \frac{u_o^2}{v_p^2} \quad (460)$$

From this we get the first solution, which would give exactly zero in the denominator of (444):

$$v_p' = u_o. \quad (461)$$

This means that if $E_z = 0$, then should the wave vector point in any direction, the phase velocity will always equal the ordinary principal velocity (which is in agreement with the solution obtained from the index ellipsoid too). In 3D this can be represented as a spherical surface.

One might ask what happens when $v_p = u_e$. In accordance with the former findings this can only occur if $E_x = E_y = 0$, i.e. the field has only z-component. In such cases the wave vector naturally remains in the x-y plane. Since this would cause a singularity in the Fresnel equation again, this case is to be excluded from our further investigations.

Correspondingly, let us determine the other solution based on (448):

$$\frac{\hat{k}_x^2}{u_o^2 - v_p^2} + \frac{\hat{k}_y^2}{u_o^2 - v_p^2} + \frac{\hat{k}_z^2}{u_e^2 - v_p^2} = 0 \quad (462)$$

Let us perform our investigations only in the x-z plane ($\hat{k}_y := 0$), since the crystal is cylindrically symmetric to z anyway, and multiply (462) with the product of the denominators (now these cannot be zero). After simplification we get this:

$$\hat{k}_x^2(u_e^2 - v_p^2) + \hat{k}_z^2(u_o^2 - v_p^2) = 0 \quad (463)$$

After some transformations:

$$v_p''^2 = \frac{\hat{k}_x^2 u_e^2 + \hat{k}_z^2 u_o^2}{\hat{k}_x^2 + \hat{k}_z^2} \quad (464)$$

since $\hat{k}_x^2 + \hat{k}_z^2 = 1$, from this we get the following:

$$v_p''^2 = \hat{k}_x^2 u_e^2 + \hat{k}_z^2 u_o^2. \quad (465)$$

This is the equation of a fourth-order curve, known as an oval (in 3D ovaloid). If $\hat{k}_z = 0$, i.e. $\hat{k}_x = 1$, this gives back even the formerly excluded case, since now $v_p'' = u_e$. If $u_o > u_e$, we speak of positive, in the opposite case negative birefringence (see Fig. 74).

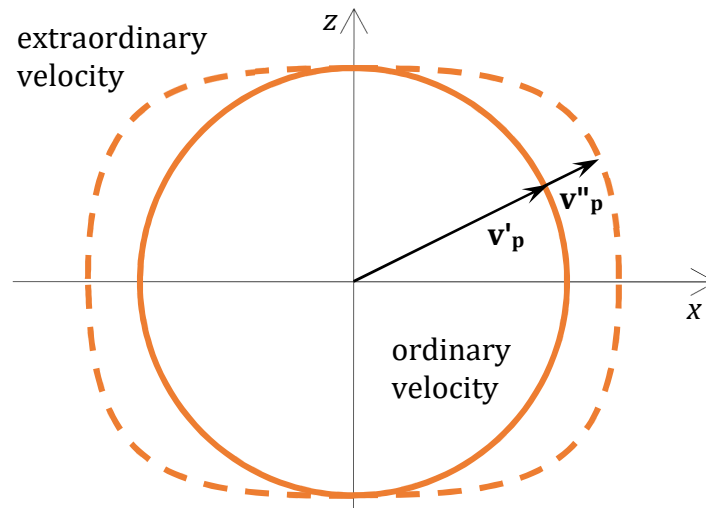


Fig. 74 Phase surfaces in case of uniaxial birefringence.

Now let us examine the direction dependence of the magnitude of wave vector k'' . Substituting the principal velocities in (465) by the appropriate refractive indices calculated by $n = c/v_p$, and dividing out c :

$$\frac{1}{n''^2} = \frac{\hat{k}_x^2}{n_e^2} + \frac{\hat{k}_z^2}{n_o^2}. \quad (466)$$

Dividing this by $(2\pi/\lambda_0)^2$ we get the magnitude of the wave vector in case of a wavefront normal of specific direction (do not forget that both n'' and k'' are direction-dependent!):

$$\frac{1}{k''^2} = \frac{\hat{k}_x^2}{k_e^2} + \frac{\hat{k}_z^2}{k_o^2} \Rightarrow 1 = \frac{k_x''^2}{k_e^2} + \frac{k_z''^2}{k_o^2}, \quad (467)$$

which is the equation of an ellipse.

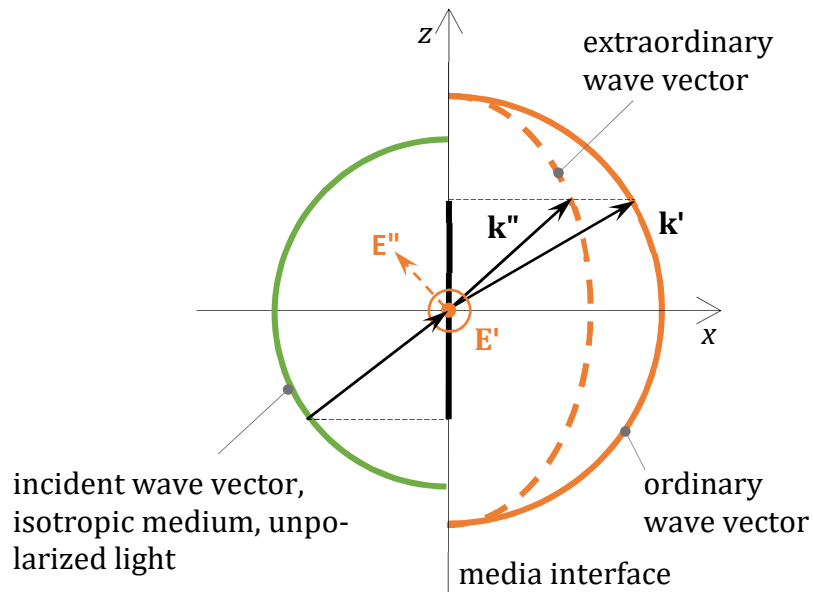


Fig. 75 Light refraction at the boundary of isotropic/anisotropic media.

Material	n_o [-]	n_e [-]
CaCO ₃ (calcite)	1.658	1.486
SiO ₂ (quartz)	1.544	1.553

Tab. 4 Examples for refractive indices in case of uniaxial birefringence (at room temperature, and $\lambda_d = 588$ nm wavelength).

14.7. Phase retarder plates

...

14.8. Further anisotropic phenomena (supplementary)

Dichroism

Optical activity

Mechanical stress-induced birefringence

ACKNOWLEDGEMENTS

I hereby wish to express my grateful thanks to my colleagues for their help offered to make these lecture notes. I thank Dr. Pál Koppa for writing the first version of the electrodynamic parts, and dr. Attila Barócsi for preparing the first version of the chapter on polarization, I owe gratitude to dr. Zoltán Füzessy for revising the chapters on statistical optics, Dr. Imre Péczeli the diffraction parts, and dr. László Kocsányi the chapter that discusses anisotropy. I am very grateful to Dr. Péter Richter for carefully revising the English version from the point-of-view of both grammar and physics.

REFERENCES

- [1] Klein-Furtak, Optics
- [2] Simonyi Károly, A fizika kultúrtörténete
- [3] M. Born and E. Wolf, Principles of Optics
- [4] John Gribbin, In Search of Schrödinger's Cat: Quantum Physics and Reality
- [5] P.C.Y Chang, J.G. Walker, K.I. Hopcraft, Ray tracing in absorbing media, Journal of Quantitative Spectroscopy & Radiative Transfer, vol. 96, pp. 327-341, 2005.
- [6] Péter Richter (ed.), Bevezetés a modern optikába
- [7] Hartmann Römer, Theoretical Optics
- [8] Füzessy Zoltán: A fotonika optikai alapjai
- [9] Goodman, Introduction to Fourier optics
- [10] Thomas Young, A Course of Lectures on Natural Philosophy and the Mechanical Arts, Vol. I., 1807
- [11] Carlos R. Stroud Jr, A jewel in the crown, Univ. of Rochester, 2004
- [12] Saleh-Teich, Fundamentals of Photonics
- [13] Goodman, Statistical Optics
- [14] Abramowitz, M. and Stegun, I. A., Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing. New York: Dover, p. 302, 1972.
- [15] <http://mathworld.wolfram.com/Ellipse.html>
- [16] R. Simon, Minimal Three-component $SU(2)$ Gadget for Polarization Optics, Phys. Let. A, Vol. 143, No. 4,5, pp. 165-169, 1990.