Modern Optics Notes

H. M. van Driel
Department of Physics
University of Toronto
Version: 2.2

Contents

List of Constants

c = speed of light; in vacuum = $2.9979 \times 10^8 ms^{-1} = 0.29979 m(ns)^{-1}$
ϵ_0 = electric permittivity of free space = $8.854 \times 10^{-12} Fm^{-1}$
μ_0 = magnetic permeability of free space = $4\pi \times 10^{-7} Hm^{-1}$
e = electric charge = $1.602 \times 10^{-19} C$
h = Planck's constant = $6.625 \times 10^{-34} Js$
$\hbar = h/2\pi = 1.054 \times 10^{-34} J\ s$
m = mass of electron = $9.109 \times 10^{-31}\ kg$
k_B = Boltzmann's constant= $1.38\ \times 10^{-23} JK^{-1}$

Simple conversions for photon energy/wavelength:
$1\ \mu$m $\equiv 1.24\ eV \equiv 1.89 \times 10^{15} rads^{-1} \equiv 3.00 \times 10^{14}\ Hz$

Some notes on typeface conventions used

A : real-valued scalar*;* $\mathcal{A}$: complex-valued scalar
$\vec{A}$: real-valued vector; $\vec{\mathcal{A}}$: complex-valued vector

Introduction

And God said:
"Let there be light. And there was light.
And God saw the light, that it was good;
And God divided the light from the darkness."
The Torah

0.1. History of optics

Light, as a phenomenon of nature, has been of fundamental interest to man since the dawn of time. To the ancients light from the sun was the source of all life, since it helped to produce crops for food. It was also regarded as a conveyor of knowledge since it allowed man to have information about his environment, while its absence created doubt, uncertainty and, consequently, fear. The sun was regarded as an object that had control over mankind, and for many peoples it was elevated to the status of a god. Because of the central role that it played in ancient lives, it is perhaps not surprising that the study of optical phenomena, or Optics, is one of the oldest fields of natural science. Nonetheless, many aspects of the nature of light, paricularly quantum properties, remain a mystery today.

Early natural philosophers such as the Greeks, Thales, Plato, Democritus and Aristotle were deeply interested in the phenomenon of vision and are known to have proposed several theories to explain it. These were centred around the observations that light travels in a straight line and that the angle of incidence is equal to the angle of reflection for a light beam striking a mirror surface. The most popular theories stated that vision took place as a two step process. In the first step a stream of "particles" or "rays" emanates from the eye in the direction it is looking. An object struck by these rays releases particles back towards the eye and forms an impression. More sophisticated versions of the theory stated that the released particles or chips from the objects were miniature versions of the object itself and therefore each carried full information about that object. However, the Greeks recognized that these theories could not be complete, since they failed to account for such simple phenomena as the refraction of light and the process of inversion in the development of mirror images.

Although the true nature of light and its propagation properties had to await the discovery of the origins of electromagnetic and quantum phenomena in the 19th and 20th centuries it was nevertheless possible for an operational or phenomenological knowledge of light to develop. *Archimedes*, perhaps the greatest scientist of antiquity, is known to have discovered the focusing properties of lenses and mirrors. Legend has it that he used this knowledge to focus sunbeams to set fire to Roman ships about to invade Syracuse in what is now Sicily. Around the year 1000 A.D., the Arab *Alhazen* made sweeping discoveries in optics and offered explanations for the phenomena of reflection from non-planar surfaces and refraction at the boundary between two different media. During the Renaissance, the Italian, Leonardo *da Vinci* showed tremendous insight into how light scatters in the atmosphere and applied much of this knowledge to his paintings. He is also known for his development of the pinhole camera working inside a room-sized *camera obscura* to make perspective and landscape studies. His countryman, *Galileo Galilei* is credited with the development of the telescope, following suggestions by the Dutch lens grinders Lipperschey and DeWard.

The seventeenth century marks the beginnings of optics as a field of pure and applied science, as opposed to a field of natural philosophy. What triggered optics as a quantitative science was the empirical discovery of the law of refraction by the Dutchman Snell in 1621. The phenomenon of diffraction was noted by Grimaldi while such scientific giants as *Christiaan Huygens* and *Isaac Newton* debated the origin of colour and whether light consisted of a stream of particles or a superposition of waves. This controversy raged for more than two centuries before the quantum mechanical foundations and wave/particle duality of light were established in the twentieth century.

During the nineteenth century, thanks to the work of *Thomas Young*, *Augustin Fresne*l, *Simeon Poisson* and *Gustav Kirchhoff* interference and diffraction of light were explained as manifestations of the wave character of light. The crowning achievement of physics in the nineteenth century, however, was the establishment of the electromagnetic theory of light by James Maxwell and verified experimentally by Heinrich Hertz. The electromagnetic theory of light, as presented through the four Maxwell's equations, is capable of explaining most macroscopic manifestations of light

FIGURE 0.1.1. Al-hazen, Huygens, Newton and Young

and its interaction with matter. This includes the propagation, reflection, refraction, diffraction and interference of light. Maxwell's theory, as even he himself knew, could never be regarded as complete since it fails to answer, for example, why objects have the colour they do, or, at a more fundamental level, why objects absorb or emit different wavelengths of light. This, of course, we now understand as being related to the microscopic manifestations of how light interacts with matter, a subject which came to be understood with the development of the atomic and quantum theory of matter.

In the early part of this century *Max Planck*, *Neils Bohr* and *Albert Einstein* laid the groundwork for a microscopic theory of light-matter interactions and which culminated in the modern quantum theory. The interaction of light with atoms was postulated to occur via the absorption or emission of quanta of energy with an electron in an atom making a transition from one energy level to another. Within the context of the quantum theory a light beam can display a particle-like interaction with matter as well as the wave-like interaction known already from Maxwell's theory. Although in any one experiment or observation only one of these facets is revealed, this wave-particle duality finally resolved the controversy launched by Newton and Huygens and showed them both to be right. In 1917 Einstein showed that the emission and absorption of light quanta by matter could take place via three elementary processes; spontaneous emission, absorption and stimulated emission.

The year 1960 marks the beginning of a new revolution in optics. Until that time, optics as a science grew because of man's desire to understand the many phenomena whereby light interacts with matter. This pursuit dealt with natural sources of light such as sunlight and candlelight as well as artificial sources such as mercury, tungsten and xenon lamps. In 1960 *Theodore Maiman* developed an optical source based on the stimulated emission of light and which has come to be known as the laser. In so doing, Theodore Maiman verified the theoretical work of Einstein as well as that performed during the previous decade by Charles Townes, Art Schawlow (a graduate of U of T !) N. Basov, and A. Prokhorov. The laser has provided a source of highly monochromatic (or coherent), intense radiation. The monochromaticity of the laser has opened up new vistas in important areas such as spectroscopy and communications, while the high intensity has led to discoveries of new optical effects such as harmonic generation and frequency mixing and several other nonlinear optical effects.

From the historical perspective, all phenomena that can be explained in terms of the electromagnetic theory of light comprise what is now referred to as *classical optics*. Those phenomena that can only be explained in terms of the microscopic theory of light-matter interaction belong to what has come to be called *quantum optics*. This, of course, includes the laser. Besides offering new fundamental areas to be studied and exploited, laser light has greatly enhanced the understanding of and implementation of classical optical phenomena. For this reason, the era since 1960 has been termed the era of *modern optics*.

0.2. Outline of the notes

In this set of notes, I attempt to outline the basis for modern optics. The historical references now having been noted, the subject will be developed in a manner that starts with the basic ideas and proceeds to the more complex ideas. Since most of our dealings with light occur in a macroscopic context, I begin with a discussion of the properties of light as an electromagnetic wave. In most cases the optical power is sufficiently high ($> 10^{-16}$ W) that the "graininess" of the beam can be ignored and one can treat light as a classical electromagnetic wave. In dealing with optical waves, as one particular class of electromagnetic waves, I somewhat arbitrarily restrict myself to waves of wavelength $0.1\mu m < \lambda < 1.0\mu m$. Keep this length scale in mind, since it is the source of many approximations.

Although at a fundamental level the interaction between light and matter is indeed a quantum phenomenon, we can proceed to discuss this interaction without recourse to many of the details of a quantum mechanical model. In the next chapter a classical model of the light-atom interaction is outlined and this so-called Lorentz model, together with Maxwell's equations, gives us a completely classical description of many optical phenomena such as refraction, reflection, absorption, and diffraction. The classical model, in essence, permits us to phenomenologically account for the resonance and off-resonance optical phenomena through a wavelength dependent electric susceptibility.

In general all classical optical phenomena can be understood in terms of Maxwell's equations and boundary conditions. In principle, then, once the optical geometry - the set of optical surfaces - is defined, one can "solve" for any part of the problem. For most cases, this would be too cumbersome and present us with an overabundance of information. In solving optical problems therefore we are led to make approximations or to confine ourselves to certain parts of the problem, depending on what information is desired. For certain subjects we may not be concerned with the the fact that a beam of light has a certain wavelength but rather concerned with the fact that it has electric and magnetic fields associated with it as in reflection, refraction and polarization phenomena. In other cases we are able to consider light as a bundle of rays which travel in a straight line. Here, where we don't concern ourselves with the electromagnetic or wave aspects, we can work in the realm of *geometrical optics or ray optics* a subfield of interest in imaging. If the wave character of light becomes of concern as it is, for example, in interference or diffraction, one speaks of *physical optics* or *wave optics*. Finally if we must consider the microscopic details of the interaction of light with matter we consider *quantum optics*.

In chapters 1 to 4 we consider light as an electromagnetic wave and present propagation, refraction, reflection, energy and polarization characteristics. In chapter 5 and 6 we introduce geometrical optics and adopt a ray optics picture in discussing elementary optical elements and their properties in transforming or imaging optical beams. Physical optical phenomena are presented in chapters 7 to 11 where interference, diffraction, Gaussian beams and waveguides are discussed. Finally in the last three chapters we concentrate on the microscopic manifestation of the interaction of light with matter in a semi-classical model where the atoms are treated quantum mechanically and light is treated classically. This allows us to consider more basic aspects of light emission and absorption properties. Particular emphasis will be given to the phenomenon of stimulated emission and the laser principle.

The notes are reasonably self-contained. They presume a familiarity with elementary electrodynamics and quantum mechanics, as would normally be obtained from 3rd year undergraduate courses.

References

M. Born and E Wolf, *Principles of Optics*, Cambridge Press, Toronto, 2002

Part 1

Light as an Electromagnetic Phenomenon

CHAPTER 1

Propagation of Light

Do not go gentle into that good night.
Rage, rage against the dying of the light.
Dylan Thomas

In this chapter the foundations of classical optics are presented with the introduction of light as an electromagnetic wave predicted by Maxwell's equations. The terminology used to describe light waves is defined. Plane, cylindrical and spherical light waves are introduced as elementary optical disturbances. Throughout the notes most of the emphasis will be on plane waves. As a result this chapter concentrates on the propagation characteristics of these waves and their dispersion relation, which couples the wavelength and the frequency. The Lorentz model of atomic oscillators provides the basis for formulating a classical model of the dielectric function of both dielectrics and metals. This function governs the velocity and absorption of light beams in matter. Finally, the superposition principle is used to construct pulses of light, which are optical disturbances localized in space and time. The propagation characteristics of these pulses in dielectrics is outlined.

1.1. Maxwell's Equations and the Constitutive Relations

We begin with a statement of *Maxwell's equations;* which are the generalization of empirical laws for electrodynamic phenomena discovered by Gauss, Ampere, and Faraday. In vacuum these four equations govern and connect the spatial and temporal evolution of two vector fields, the electric field $\vec{E}(\vec{r},t)$ and the magnetic field $\vec{H}(\vec{r},t)$ $(= \vec{B}(\vec{r},t)/\mu_0)$ that depend on space and time. For notational convenience, we sometimes suppress the explicit $\vec{r}$ and t dependence from time to time. In matter, Maxwell's equations relate five vector fields. These are the electric and magnetic fields, two fields that represent the response of the material to an applied electric field and one field that represents the magnetic response of the medium to the applied fields. For the electric field, the two response fields are the bound electron polarization field $\vec{P}$ and the free charge current density $\vec{J}$, while the magnetic response is represented by the magnetization density $\vec{M}$. In SI units Maxwell's equations for fields in matter are given by:

$$\vec{\nabla} \times \vec{E} = -\frac{\partial}{\partial t}(\vec{B}) \tag{1.1.1}$$

$$c^2\vec{\nabla} \times \vec{B} = \frac{\partial}{\partial t}\vec{E} + \frac{1}{\epsilon_0}\frac{\partial}{\partial t}(\vec{P}) + \frac{1}{\epsilon_0}\vec{J} + \frac{1}{\epsilon_0}(\vec{\nabla} \times \vec{M}) \tag{1.1.2}$$

$$\epsilon_0\vec{\nabla} \cdot \vec{E} = -\vec{\nabla} \cdot \vec{P} + \rho \tag{1.1.3}$$

$$\vec{\nabla} \cdot \vec{B} = 0 \tag{1.1.4}$$

where ρ is the free charge density. The first is also known as Faraday's law, the second, as Ampere's law, while the last two are forms of Gauss' law. The constant ϵ_0 is the dielectric constant of free space and has a value

$$\epsilon_0 \approx \frac{1}{36\pi} \times 10^{-9}\,\text{SI Units.}$$

The other constant occurring in Maxwell's equations, μ_0, the magnetic permeability of free space is defined to have the value

$$\mu_0 = 4\pi \times 10^{-7}\,\text{SI Units.}$$

A glance at Maxwell's equations reveals that the four equations alone cannot determine five vector fields. Additional relations must be provided and are done so by the *constitutive relations*, which connect the induced polarization and current fields to the electric field and the induced magnetization to the magnetic field. These relations completely reflect the way in which a particular material responds. In some cases these relations may be quite complex. For

example, the induced fields at a particular point in space and time, may depend not only on the driving fields at that point but also on the history of the field and the full spatial characteristics of the field. The functional relationship may be highly nonlinear in the driving field. For the moment we assume that the response of the medium is instantaneous and local in character and that inducing fields are sufficiently weak that a linear relation exists between the driving field and the induced fields. In the case of the electric field, the most general relation satisfying these conditions is of the form

$$\vec{P}(\vec{r},t) = \epsilon_0 \overleftrightarrow{\chi}_e \cdot \vec{E}(\vec{r},t) \tag{1.1.5}$$

where $\overleftrightarrow{\chi}_e$ is a second rank tensor or a matrix, referred to as the *electric susceptibility*. If the material has cubic or isotropic symmetry the matrix is diagonal with equal diagonal elements and is therefore a scalar. One then has

$$\vec{P} = \epsilon_0 \chi_e \vec{E}$$

where χ_e may be dependent on position $\vec{r}$. The assumptions leading to this particularly simple constitutive relation are summarized by stating that the material is *linear*, *isotropic*, and *dispersion-free*.

It may seem that the assumptions are overly restrictive and the simple relations are of little more than academic interest. This is not the case. One might expect deviations from linearity only if the applied field becomes comparable to the field binding electrons to atoms, molecules or solids. Using the hydrogen atom as an example the binding field of an electron separated from the proton by one Bohr radius is approximately 10^{10} *V/m,* a large field indeed, and one not generally encountered in the laboratory. The assumption of isotropy is applicable to many materials that concern us in optics, such as air, glass, most metals and semiconductors. Those that are not isotropic are the subject of a later chapter on polarization phenomena. Finally the assumption of a response that is localized in space and time (dispersion-free) is generally valid if the frequency of the applied field is not near a resonance; of the system (recall the phase lags associated with a driven harmonic oscillator near resonance). Since classical optics can not completely account for resonance phenomena in the form of absorption and emission, this is of little concern to us at present.

Similar arguments to those above lead to the simple constitutive relation between the applied magnetic field intensity, $\vec{H}$, and the magnetization density, which takes the form

$$\vec{M} = \mu_0 \chi_m \vec{H} \tag{1.1.6}$$

where χ_m is the magnetic susceptibility of the medium.

The total fields in the material are then given by

$$\vec{D} = \epsilon_0 \vec{E} + \vec{P} = \epsilon_0 (1 + \chi_e) \vec{E} = \epsilon \vec{E} \tag{1.1.7}$$

and

$$\vec{B} = \mu_0 \vec{H} + \vec{M} = \mu_0 (1 + \chi_m) \vec{H} = \mu \vec{H} \tag{1.1.8}$$

where ϵ and μ are the *electric permittivity* and *magnetic permeability* of the material respectively. In what follows we only consider nonmagnetic materials for which $\mu = \mu_0$. This does not neglect many materials of interest in optics. For notational convenience we then define $\chi_e = \chi$.

For the last constitutive relation we assume that the electric current density $\vec{J}$ is linearly dependent on the electric field $\vec{E}$ through the conductivity tensor, which in the case of an isotropic medium, reduces to a scalar. The current density is thus related to the field by

$$\vec{J} = \sigma \vec{E} \tag{1.1.9}$$

Materials for which $\sigma = 0$ are referred to as *insulators* or *dielectrics* and those for which σ is substantial are known as *conductors*. Metals are often good conductors but there are other good conductors such as gaseous plasmas and ionic solutions.

For non-magnetic materials, all the optical properties, *i.e.*, why copper is a good reflector and why water refracts light, are contained in the electrical parameters, the dielectric constant and the electrical conductivity. Knowledge of these parameters completely determines the propagation of light in a particular medium.

1.2. Wave Solutions to Maxwell's Equations in Homogeneous, Dielectric Media

The use of the constitutive relations reduces Maxwell's equations in material media to the determination of $\vec{E}$ and $\vec{H}$, similar to the vacuum case. One can then combine Faraday's and Ampere's law to show that wave solutions

to Maxwell's equations exist. We have that

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{E}) = -\mu\epsilon_0 \frac{\partial^2 \vec{E}}{\partial t^2} - \mu \frac{\partial^2 \vec{P}}{\partial t^2} - \mu \frac{\partial \vec{J}}{\partial t} \tag{1.2.1}$$

In the case of a dielectric medium this equation becomes

$$\vec{\nabla}(\vec{\nabla} \cdot \vec{E}) - \nabla^2 \vec{E} = -\mu\epsilon \frac{\partial^2 \vec{E}}{\partial t^2} \tag{1.2.2}$$

For a spatially uniform (homogeneous) dielectric with $\rho = 0$ Gauss' law becomes

$$\vec{\nabla} \cdot (\epsilon \vec{E}) = \epsilon \vec{\nabla} \cdot \vec{E} = 0 \tag{1.2.3}$$

and equation 1.2.1 reduces to

$$\nabla^2 \vec{E} - \frac{1}{v_\phi^2} \frac{\partial^2 \vec{E}}{\partial t^2} = 0 \tag{1.2.4}$$

with

$$v_\phi = 1/\sqrt{\mu\epsilon}$$

Equation 1.2.4 is a wave equation with a vector amplitude $\vec{E}$ and with v_ϕ being defined as the *phase speed* of the wave. In vacuum this speed is

$$v_\phi = c = 1/\sqrt{\epsilon_0 \mu_0} = 3.0 \times 10^8 ms^{-1}$$

and in a non-dispersive medium can be written as

$$v_\phi = c/n \tag{1.2.5}$$

where $n = \sqrt{\epsilon/\epsilon_0}$. An identical equation to Eq. 1.2.4, with $\vec{E}$ replaced by $\vec{H}$, exists for the magnetic field. The two field amplitudes are not associated with two different waves but with the same wave and the amplitudes are connected by Maxwell's equations.

Since these wave equations are linear in the electric and magnetic fields, the *superposition principle* holds and the sum of solutions of the wave equation is also a solution. That is, if $\vec{f}_i(\vec{r}, t)$ represents a class of solutions of the wave equations then

$$\vec{F}(\vec{r}, t) = \sum_i \vec{f}_i(\vec{r}, t) \tag{1.2.6}$$

is also a solution. Depending on the boundary conditions of the medium and the geometry, various eigenmodes or fundamental solutions of the wave equation can be found and a general solution can be represented as a superposition of these eigenmodes. If the electric or magnetic field can be considered to have only one independent vector component, one refers to the light wave as a *scalar wave*. The general vector light waves can be built up as a superposition of scalar light waves.

1.3. Plane, Spherical and Cylindrical Waves

One of the most convenient scalar waves is the *plane wave*, so named since the surfaces of constant phase are planes. In one dimension the most general plane solution of Eq. 1.2.4 is

$$\vec{E} = \hat{x}[f_+(z - v_\phi t) + f_-(z + v_\phi t)]$$

$$\vec{B} = \frac{n}{c}\hat{y}[f_+(z - v_\phi t) - f_-(z + v_\phi t)]$$

where $\hat{x}$ and $\hat{y}$ are unit vectors that describe a right-handed co-ordinate system and $f_\pm$ are arbitrary functions. The f_+ function is associated with a wave propagating along the $+z$ direction while the f_- is associated with a wave propagating along the $-z$ direction. The plane wave solutions to the wave equations have the property that the electric and magnetic field vectors lie in a plane which is perpendicular to the direction of propagation. One then refers to these as *transverse light waves*. The expression for the one-dimensional plane wave above can be generalized to a wave travelling along an arbitrary direction by an appropriate rotation of the frame of reference.

A *monochromatic* or harmonic *plane wave* is one in which the spatial and temporal variation of the electric and magnetic field is a circular function. A scalar, plane harmonic wave can be written in the form

$$E = E_0 cos(-\omega t \pm \vec{k} \cdot \vec{r} + \phi_0) \tag{1.3.1}$$

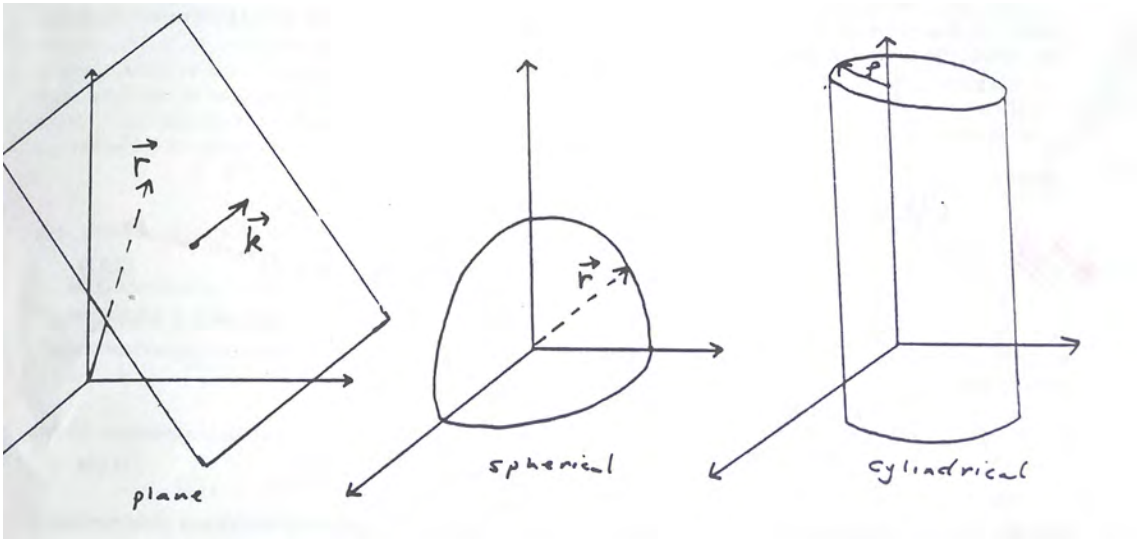


FIGURE 1.3.1. Spatial representations of plane, spherical and cylindrical waves

where ϕ_0 is an arbitrary phase, ω is the angular frequency of the wave (in radians/s) and $\vec{k}$ is the *propagation vector* with $\vec{k}/k$ defining the direction of propagation of the wave. The magnitude of the propagation vector, k, is the *propagation number* or *wave number*, and defines the periodicity or wavelength, λ, by

$$k = 2\pi/\lambda.$$

The equation

$$-\omega t \pm \vec{k} \cdot \vec{r} + \phi_0 = constant$$

defines surfaces of constant phase as shown in figure 1.3.1. From equation 1.3.1 it is easy to see that a surface of constant phase is displaced a distance dr in a time dt such that $kdr \pm \omega dt = 0$. The phase speed is then

$$v_\phi = \omega/k. \tag{1.3.2}$$

Since the frequency of a light beam in a linear medium is independent of the medium, equations 1.2.5 and 1.3.2 imply that

$$k = 2\pi n/\lambda_0.$$

Electromagnetic waves have been identified in nature over a wide spectrum of wavelengths. Our eyes however are sensitive only to those waves with wavelengths between 0.4 μm (seen as violet) and 0.7 μm (seen as red). Figure 1.3.2 shows the frequency and wavelength scale of interest to us in optics.

In some situations the plane waves may not be the most convenient basis functions to use. For example, in situations of isotropic symmetry a spherical co-ordinate system is more convenient to use. For this symmetry the Laplacian in equation 1.2.4 is given by

$$\nabla^2 = \frac{1}{r}\frac{\partial^2}{\partial r^2} r$$

where r is the radial distance from an origin. For the case of scalar waves, with U representing the magnitude of the electric or magnetic field, the wave equation then becomes

$$\frac{\partial^2}{\partial r^2}(rU) - \frac{1}{v_\phi^2}\frac{\partial^2}{\partial t^2}(rU) = 0$$

which, far from the origin, has solutions of the form

$$U = \frac{U_+(r - v_\phi t)}{r} + \frac{U_-(r + v_\phi t)}{r}.$$

The first term on the right hand side represents a *spherical wave* diverging from the origin while the second term represents a spherical wave converging on the origin. A monochromatic spherical wave in the far field ($kr \gg 1$) assumes the form

$$U = \frac{U_{+0} cos(kr - \omega t + \phi_0)}{r} + \frac{U_{-0} cos(kr + \omega t + \phi_0)}{r}$$

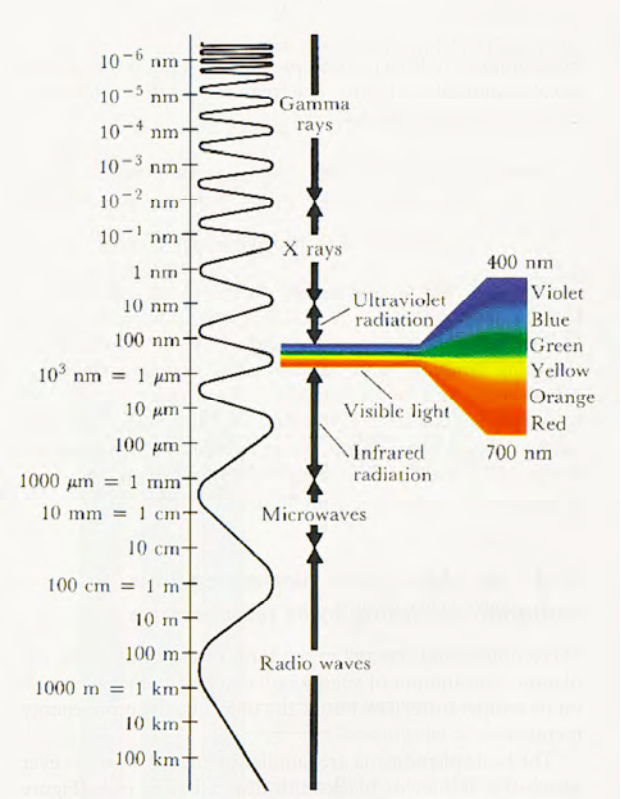


FIGURE 1.3.2. Electromagnetic spectrum

and the surfaces of constant phase are spheres. Like plane waves, spherical waves are idealizations that are never found in nature. However the cosine monochromatic spherical waves are good approximations in many situations and certainly can serve as basis functions for the expansion of more complicated optical disturbances.

For situations of cylindrical symmetry the Laplacian becomes

$$\nabla^2 = \frac{1}{\rho}\frac{\partial}{\partial\rho}(\rho\frac{\partial}{\partial\rho})$$

where ρ is the distance from the axis of symmetry. The wave equation for scalar waves now becomes

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}(\rho\frac{\partial U}{\partial\rho}) - \frac{1}{v_\phi^2}\frac{\partial^2}{\partial t^2}(U) = 0$$

which has approximate ($k\rho >> 1$) *cylindrical wave* monochromatic solutions of the form

$$U = \frac{U_{+0}cos(k\varrho - \omega t + \phi_0)}{\rho^{1/2}} + \frac{U_{-0}cos(k\varrho + \omega t + \phi_0)}{\rho^{1/2}}$$

It should be emphasized again that each of the plane wave, spherical wave and cylindrical wave solutions is only a mathematical approximation to a wave that may exist in nature, without even addressing whether their monochromatic forms can actually exist. The plane wave can never be found in nature because it has infinite

transverse extent and therefore would carry infinite energy. The spherical and cylindrical waves can't exist for $r, \rho \to 0$ since the amplitudes would become infinite. Also, no true point or line sources exist for electromagnetic waves. It might also be noted that for large r or ρ the spherical and cylindrical waves behave like plane waves. Only in this limit are these waves transverse in character.

The assumption of linear, constitutive relations has made Maxwell's equations linear and as a consequence the superposition principle is valid. A sinusoidal electromagnetic excitation of the medium therefore produces a sinusoidal electromagnetic response (a reflected or transmitted beam, for example) at the same frequency. Harmonic or monochromatic waves have a particular significance in linear optics because any transient or pulsed excitation can be written as the superposition of sinusoidal excitations. The overall response of a system can be written as a superposition of the responses to sinusoidal waves. Thus the net response of a system to a general light field can be written as the net superposition of responses the system makes to these sinusoidal excitations constituent. For this reason we offer a convenient representation and consider the properties of harmonic waves in what follows.

1.4. Phasor Representation of Waves

Monochromatic waves are most easily handled using the complex *phasor* notation. For example, the scalar harmonic function

$$U(t) = U_0 cos(\omega t - \phi_0)$$

can be represented by the equivalent phasor

$$\mathcal{U}(t) = \mathcal{U}_0 e^{-i\omega t}$$

where $\mathcal{U}_0 = U_0 e^{i\phi_0}$. The harmonic function can be retrieved from the phasor through

$$U(t) = Re[\mathcal{U}(t)].$$

Note: real quantities are denoted by *U,E,P* etc. while their phasor counterparts are denoted by $\mathcal{U}$, $\mathcal{E}$, $\mathcal{P}$, etc. For a time dependent, real, monochromatic vector field $\vec{E}(t)$ we can write

$$\begin{aligned}\vec{E}(t) &= \hat{x}E_{0x}cos(\omega t - \phi_x) + \hat{y}E_{0y}cos(\omega t - \phi_y) + \hat{z}E_{0z}cos(\omega t - \phi_z) \\ &= Re[(\hat{x}\mathcal{E}_{ox} + \hat{y}\mathcal{E}_{oy} + \hat{z}\mathcal{E}_{oz})e^{-i\omega t}] = Re(\vec{\mathcal{E}}_0 e^{-i\omega t})\end{aligned}$$

where $\vec{\mathcal{E}}_0$ is the complex vector amplitude of the field and $\mathcal{E}_{ox} = E_{0x}e^{i\phi_x}$, etc. The oscillatory spatial dependence of the wave is contained in the phase factors. In general, if ω and the amplitude of the wave are fixed, there are five free parameters (two vector components of $\vec{E}$ and three phases) that must be specified. For a transverse wave, as is shown later, it is sufficient to specify only three quantities.

1.5. Complex Form of Maxwell's Equations in Dielectric and Conducting Media

Maxwell's equations are tremendously simplified for monochromatic waves written in terms of the phasor notation. Consider then steady state or *stationary responses* of any medium to a monochromatic disturbance. The electric, magnetic and current density fields are of the form

$$\begin{aligned}&\vec{\mathcal{E}}(\vec{r})e^{-i\omega t} \\ &\vec{\mathcal{B}}(\vec{r})e^{-i\omega t} \\ &\vec{\mathcal{J}}(\vec{r})e^{-i\omega t} \\ &\vec{\mathcal{P}}(\vec{r})e^{-i\omega t}.\end{aligned}$$

Because all the field quantities have the same time variation, the time derivative is equivalent to multiplication by $-i\omega t$, *i.e.*

$$\frac{\partial}{\partial t} \equiv -i\omega$$

and integration over time is equivalent to division by $-i\omega$, *i.e.*

$$\int_t \equiv \frac{1}{-i\omega}.$$

With these simplifications Faraday's law becomes

$$\vec{\nabla} \times \vec{\mathcal{E}}(\vec{r}) = i\omega\vec{\mathcal{B}}(\vec{r}) \tag{1.5.1}$$

and Ampere's law becomes

$$\vec{\nabla} \times \vec{\mathcal{B}}(\vec{r}) = -i\omega\mu_0\epsilon\vec{\mathcal{E}}(\vec{r}) + \mu_0\vec{\mathcal{J}}(\vec{r}) \tag{1.5.2}$$

while the two Gauss' laws remain the same.

The phasor formalism also permits us to examine *dispersive media* in which the response of the medium to the excitation may not be instantaneous, as was assumed in section 1.1. For example, such a dependence may be due to inertia effects associated with the non-zero mass of the electrons or to resonances (absorption or emission) at some particular frequencies. In a linear, dispersive medium $\vec{P}$ can usually be related to the past history of a field $\vec{E}$ by a linear integral equation of the form

$$\vec{P}(\vec{r},t) = \epsilon_0 \int_{-\infty}^{\infty} R(t-t')\vec{E}(\vec{r},t')dt'$$

for which it is understood that for $t' > t$ $R(t-t') \equiv 0$. Taking the Fourier transform (see the Appendix at the end of this chapter if you are not familiar with Fourier transforms) of both sides yields

$$\vec{P}(\vec{r},\omega) = \epsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R(t-t')e^{i\omega t}\vec{E}(\vec{r},t')dt'dt$$

$$= \epsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R(t-t')e^{i\omega(t-t')}\vec{E}(\vec{r},t')e^{i\omega t'}dt'dt = \epsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R(\tau)e^{i\omega\tau}\vec{E}(\vec{r},t')e^{i\omega t'}dt'd\tau$$

$$= \epsilon_0\chi(\omega)\vec{E}(\vec{r},\omega) \tag{1.5.3}$$

where the dielectric susceptibility χ is a function of ω and is complex (reflecting a possible phase lag in the response). By a simple extension one can cast this in phasor notation through

$$\vec{\mathcal{P}}(\vec{r}) = \epsilon_0\chi(\omega)\vec{\mathcal{E}}(\vec{r})$$

For a dispersive medium the dielectric "constant" is a function of frequency and is given by

$$\hat{\epsilon}(\omega) = \epsilon_0[1+\hat{\chi}(\omega)].$$

The real and imaginary parts of the susceptibility or dielectric function are related to each other by *Kramers-Kronig relations*. The result expressed in equation 1.5.3 is very important since it shows that in a dispersive medium one has a linear, constitutive relation for the individual harmonic components of a wave. The overall response can then be obtained via the superposition principle.

The effects of conducting media may also be incorporated explicitly in a complex dielectric constant. When equation 1.1.9 is combined with equation 1.2.1 for an assumed harmonic form of the field time dependence, we have, in the case of a spatially homogeneous medium, a Helmholtz equation of the form

$$\nabla^2\vec{\mathcal{E}} + \omega^2\mu\hat{\epsilon}_{total}\vec{\mathcal{E}} = 0$$

From this equation we obtain a complex dielectric constant

$$\hat{\epsilon}_{total} = \hat{\epsilon} + i\sigma/\omega$$

where both the real and imaginary parts of the *complex dielectric function* can be frequency dependent. Hereafter we drop the "total" and refer only to a dielectric constant $\hat{\epsilon} = \hat{\epsilon_{bound}} + i\sigma/\omega$, where the first term refers to the response of the bound electrons.

1.6. Dispersion Relation for Plane Waves

The complex form of Maxwell's equations leads in a simple way to general harmonic plane-wave solutions travelling in an arbitrary direction in a dispersive medium. It is convenient to describe a propagating plane wave without reference to a particular co-ordinate system. As above, we take a propagation vector $\vec{k}$ to be normal to the phase fronts and assume a spatial dependence for the fields of the form

$$\vec{\mathcal{E}}(\vec{r}) = \vec{\mathcal{E}}_0 e^{i\vec{k}\cdot\vec{r}}$$

with

$$\vec{\mathcal{B}}(\vec{r}) = \vec{\mathcal{B}}_0 e^{i\vec{k}\cdot\vec{r}}.$$

For this form of the spatial variation of the fields the ∇ operator is equivalent to multiplication by $i\vec{k}$ and so Faraday's law becomes

$$i\vec{k}\times\vec{\mathcal{E}}_0 = i\omega\vec{\mathcal{B}}_0 \tag{1.6.1}$$

and Ampere's law becomes

$$i\vec{k}\times\vec{\mathcal{B}}_0 = -i\omega\mu\hat{\epsilon}\vec{\mathcal{E}}_0. \tag{1.6.2}$$

By cross-product multiplying equation 1.6.1 by $i\vec{k}$ we obtain

$$i\vec{k} \times (i\vec{k} \times \vec{\mathcal{E}}_0) = i\omega(i\vec{k} \times \vec{\mathcal{B}}_0)$$

and with the use of equation 1.6.2 we get

$$-\vec{k} \times (\vec{k} \times \vec{\mathcal{E}}_0) = -\vec{k}(\vec{k} \cdot \vec{\mathcal{E}}_0) + k^2\vec{\mathcal{E}}_0 = \omega^2\mu\hat{\epsilon}\vec{\mathcal{E}}_0.$$

The triple product produces a vector which is perpendicular to $\vec{k}$ and this is set equal to a vector along the direction of $\vec{\mathcal{E}}_0$. It follows that $\vec{\mathcal{E}}_0$ must be perpendicular to $\vec{k}$. We could also have obtained this result by recognizing that, at least, in a charge-free homogeneous medium, Gauss' law is given by

$$\vec{k} \cdot \vec{\mathcal{E}}_0 = 0$$

This coincidence of results is a consequence of the assumed form of solution of the wave equation. We also have that

$$k^2 = \omega^2\mu\hat{\epsilon}. \tag{1.6.3}$$

This relation between an obviously complex wave vector and the frequency of the wave is referred to as the *dispersion relation* of the plane wave. From this, we can define a complex refractive index $\hat{n}$ by

$$k^2 = \omega^2\mu\epsilon_0\hat{n}^2 = \frac{\omega^2}{c^2}\hat{n}^2. \tag{1.6.4}$$

Since we also have that

$$\vec{\mathcal{B}}_0 = \frac{1}{\omega}\vec{k} \times \vec{\mathcal{E}}_0$$

it follows that $\vec{\mathcal{B}}_0$, $\vec{\mathcal{E}}_0$ and $\vec{k}$ form a right handed coordinate system.

We finish this section by offering an insight into the relationship between the complex wave parameters and the dielectric constant. First of all, the phase velocity of the plane waves is the velocity of a surface of constant phase. Since the overall phase function of the wave can be represented as a phase function

$$\phi(\vec{r}, t) = -\omega t \pm \vec{k} \cdot \vec{r} + \phi_0$$

we have, following a previous argument, that the phase speed is

$$\frac{dr}{dt} = \frac{\omega}{k} = \frac{c}{\hat{n}} \equiv v_\phi.$$

How does one interpret a complex phase velocity and, for that matter, what does a complex refractive index mean? From equations 1.6.3 and 1.6.4 it follows that

$$\hat{\epsilon} = \epsilon_R + i\epsilon_I = \hat{n}^2\epsilon_0.$$

If we define the real and imaginary parts of the refractive index by

$$\hat{n} = n + i\kappa$$

then

$$\epsilon_R/\epsilon_0 = n^2 - \kappa^2 \qquad \epsilon_I/\epsilon_0 = 2n\kappa$$

We can obtain the real and imaginary parts of the refractive indices by inverting the last two relations to obtain

$$\epsilon_0 n^2 = \frac{1}{2}\left(\epsilon_R + \sqrt{\epsilon_R^2 + \epsilon_I^2}\right)$$

and

$$\epsilon_0 \kappa^2 = -\frac{1}{2}\left(\epsilon_R - \sqrt{\epsilon_R^2 + \epsilon_I^2}\right)$$

Because of sign conventions established, it is under stood that $n > 0$, $\kappa > 0$ if $\epsilon_I > 0$ and $n > 0$, $\kappa < 0$ if $\epsilon_I < 0$; this latter case cannot occur for classical models but only when one takes into account quantum mechanical models of the response of matter (see chapter 12). Until recently for natural materials there was never a case with $n < 0$. However, with the advent of man-made, structured, artificial material with a significant magnetic response, known as *metamaterials*, it is possible to have $n < 0$, at least over a certain spectral range.

To understand the role of the real and imaginary parts of the refractive index determine, consider a plane wave propagating in the z-direction so that

$$\vec{k} = k\hat{z}.$$

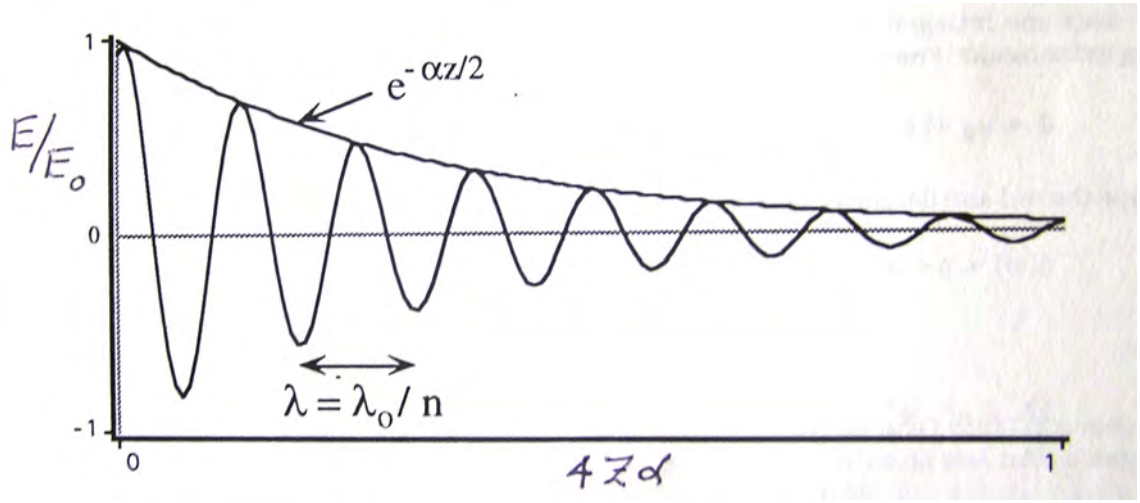


FIGURE 1.6.1. Normalized electric field as a function of distance in a medium with complex refractive index.

It follows that the spatial contribution to the phase function is given by

$$ikz = i(n + i\kappa)\frac{2\pi}{\lambda_0}z$$

and the electric field spatial variation is given by

$$e^{ikz} = \exp(in\frac{2\pi}{\lambda_0}z)exp(-\kappa\frac{2\pi}{\lambda_0}z).$$

The wave in the material propagates with a wavelength λ/n or a phase velocity c/n. At the same time the field decays at a rate which is determined by the imaginary part of the refractive index. We can introduce an attenuation or absorption constant, α, defined by the relation:

$$\frac{\alpha}{2} = \frac{2\pi\kappa}{\lambda_0}$$

The energy of a light wave is proportional to E^2 (see chapter 2); thus α^{-1} is the scale length of the spatial decay of the energy of the wave with $E^2 \propto e^{-\alpha z}$. Figure 1.6.1 illustrates this effect of the complex-valued index of refraction.

One can now interpret the phase velocity as being associated with actual motion of the surfaces of constant phase with decay of the field amplitudes in such a way that

$$\frac{k}{\omega} = \frac{1}{v_\phi} = \frac{n}{c} + i\frac{\kappa}{c} = \left[\frac{1}{v_\phi}\right]_R + \left[\frac{1}{v_\phi}\right]_I$$

1.7. Classical Model for the Dielectric Function

The complex dielectric function or the refractive index is the only parameter needed to determine how an optical disturbance propagates in a linear, isotropic medium. It is useful to introduce a simple model for the dielectric function to illustrate the most fundamental elements of the propagation of light in matter. The model we choose is an empirical model that was first offered by Lorentz towards the end of the 19th century, before the advent of quantum mechanical theories of light-atom interactions. The model is appropriate for dilute materials such as gases. We consider first the application of the *Lorentz model* to systems in which the electrons are bound such as atoms and molecules, and later show, that with a trivial extension we can consider systems in which the electrons are free to move, such as metals and plasmas.

1.7.1. Bound electron systems. Lorentz was aware of the following experimental facts concerning the interaction of an electromagnetic wave with a dielectric medium in formulating his model:

1) Media display discrete resonance frequencies at which they are capable of emitting or absorbing radiation;

2) the media respond to and emit sinusoidally oscillating waves;

3) the response of the medium is linear in the electric field amplitude.

For the sole purpose of explaining these observations, Lorentz considered the medium to be composed of a dilute set of classical electric dipoles as shown in figure 1.7.1

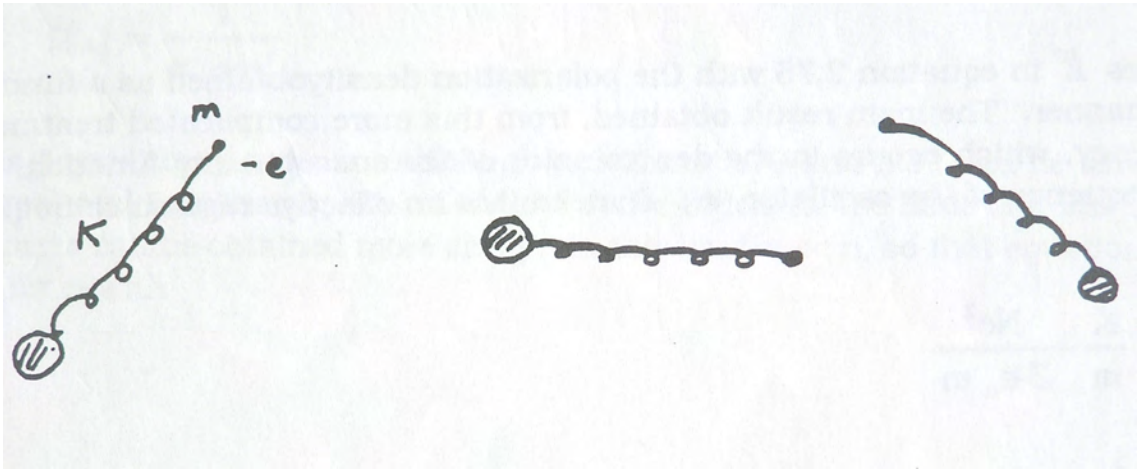


FIGURE 1.7.1. Lorentz Oscillators

The dipoles can be considered to consist of an electron of mass m bound to a massive core by a spring with spring constant K. Each dipole behaves like a simple harmonic oscillator with a *resonance frequency* ω_0 given by

$$\omega_0^2 = K/m \tag{1.7.1}$$

In the model the spring constant or the resonance frequency constitutes a phenomenological parameter, which can be adjusted to match an experimentally observed resonance frequency.

In three dimensions, the restoring force on the electron by the spring is given by

$$\vec{F} = -K\vec{r}$$

where $\vec{r}$ is the displacement of the electron from equilibrium. For such a displacement the induced microscopic dipole moment is

$$\vec{p} = -e\vec{r}$$

where e is the elementary charge. Although the Lorentz model originally considered only electronic displacements in an otherwise structureless "atom", the model can be easily extended in the same phenomenological fashion to provide a model for electric dipoles associated with nuclear degrees of freedom and thereby treat electric dipoles associated with vibrational and rotational motion.

The macroscopic response of a medium is determined by the induced polarization density. If the systems of dipoles is sufficiently dilute, their interactions can be neglected compared to their individual interaction with a macroscopic field. In this case the macroscopic polarization density is simply the number density of atomic oscillators, N, multiplied by the dipole moment for each atom. One then has

$$\vec{P} = -Ne\vec{r}$$

In the presence of a static electric field of strength $\vec{E}$ the static polarization of the gas is given by the equilibrium condition

$$-e\vec{E} - K\vec{r} = 0$$

or

$$\vec{P} = \frac{Ne^2}{K}\vec{E}.$$

If the applied field has a time dependence, the displacement of the electron from equilibrium is governed by the damped harmonic oscillator equation,

$$m\frac{d^2\vec{r}}{dt^2} + m\gamma\frac{d\vec{r}}{dt} = -K\vec{r} - e\vec{E}(t) \tag{1.7.2}$$

where the phenomenological constant, γ, can be used to account for radiation damping by an accelerating charge, energy loss due to collisions, etc. Let us consider solutions to this equation for a harmonic driving field of the form $\vec{\mathcal{E}}(t) \propto e^{-i\omega t}$. As noted earlier, after the transients associated with the solution to the homogeneous equation have died out, the time derivative is equivalent to multiplication by $-i\omega$. The differential equation is then converted into the algebraic equation

$$(-m\omega^2 - i\omega m\gamma + K)\vec{r}(t) = -e\vec{\mathcal{E}}(t)$$

Suppressing the obvious harmonic time dependence, the polarization density is given by

$$\vec{\mathcal{P}} = \left(\frac{Ne^2}{-m\omega^2 - i\omega m\gamma + K} \right) \vec{\mathcal{E}} = \epsilon_0 \chi \vec{\mathcal{E}} \tag{1.7.3}$$

In the case of a dense dielectric medium, the electric field which appears as the driving field in equation 1.7.2 is the so-called local field. This is the sum of the applied field and the field arising from the surrounding oscillating dipoles. It can be shown that the local field, $\vec{E}_l$, is given approximately by

$$\vec{\mathcal{E}}_l = \vec{\mathcal{E}} + \frac{1}{3\epsilon_0}\vec{\mathcal{P}}$$

This field then replaces $\vec{\mathcal{E}}$ in equation 1.7.3 with the polarization density obtained as a function of $\vec{\mathcal{E}}$ in a self-consistent manner. The main result obtained from this more complicated treatment is that the resonance frequency, which occurs in the denominator of the susceptibility function, is not the natural resonance frequency of the oscillator, ω_0. Instead it is an effective resonance frequency, Ω_0 , given by

$$\Omega_0^2 = \frac{K}{m} - \frac{Ne^2}{3\epsilon_0 m}$$

indicating a shift to a lower frequency with increasing density N. The second term on the right hand side is known as the local field correction to the natural resonance frequency. Since the entire treatment of the microscopic dipoles is phenomenological and the frequencies ω_0 and Ω_0 are empirical we continue with only one such frequency, ω_0.

From equation 1.7.3 the dielectric function of the Lorentz medium is given by

$$\hat{\epsilon} = \epsilon_0 \left(1 + \frac{Ne^2}{m\epsilon_0} \frac{1}{\omega_0^2 - \omega^2 - i\omega\gamma} \right)$$

with the real part given by

$$\frac{\epsilon_R}{\epsilon_0} = n^2 - \kappa^2 = 1 + \frac{Ne^2}{m\epsilon_0} \frac{\omega_0^2 - \omega^2}{\left(\omega_0^2 - \omega^2\right)^2 + \gamma^2\omega^2}. \tag{1.7.4}$$

In the vicinity of the resonance $\omega \simeq \omega_0$ this becomes

$$\frac{\epsilon_R}{\epsilon_0} = 1 + \frac{Ne^2}{2m\epsilon_0\omega_0} \frac{\omega_0 - \omega}{\left(\omega_0 - \omega\right)^2 + \gamma^2/4}. \tag{1.7.5}$$

The imaginary part of the dielectric constant is given by

$$\frac{\epsilon_I}{\epsilon_0} = 2n\kappa = \frac{Ne^2}{m\epsilon_0} \frac{\gamma\omega}{\left(\omega_0^2 - \omega^2\right)^2 + \gamma^2\omega^2} \tag{1.7.6}$$

which for $\omega \simeq \omega_0$ becomes

$$\frac{\epsilon_I}{\epsilon_0} = \frac{Ne^2}{4m\omega_0\epsilon_0} \frac{\gamma}{\left(\omega_0 - \omega\right)^2 + \gamma^2/4}$$

Because of this result, functions of the form

$$f(x) = \frac{1}{x^2 + a^2}$$

are known as *Lorentzian functions*. Although equations 1.7.4 and 1.7.6 can be solved simultaneously to yield the real and imaginary parts of the refractive index, in the limit of a low density gas (or for $\gamma << \omega_0$) these parts can be obtained more easily. In this limit $\kappa << n$, so that equation 1.7.5 is essentially an equation for n with

$$n = \left(1 + \frac{Ne^2}{2m\omega_0\epsilon_0} \frac{\omega_0 - \omega}{\left(\omega_0 - \omega\right)^2 + \gamma^2/4} \right)^{1/2} = 1 + \frac{Ne^2}{4m\omega_0\epsilon_0} \frac{\omega_0 - \omega}{\left(\omega_0 - \omega\right)^2 + \gamma^2/4}.$$

For n near unity we then obtain κ from equation 1.7.6 with

$$\kappa = \frac{Ne^2}{8m\omega_0\epsilon_0} \frac{\gamma}{\left(\omega_0 - \omega\right)^2 + \gamma^2/4}$$

Figure 1.7.2 shows the frequency dependence of the real and imaginary part of the refractive index for a dilute Lorentz gas. The real part of the refractive index, which determines the phase velocity of light, shows typical dispersion characteristics in the vicinity of the resonance where emission or absorption can occur. The region where the refractive index increases with increasing frequency is known as regions of *normal dispersion* ($dn/d\omega > 0$) while the region near the resonance frequency where the refractive index decreases with increasing frequency is known as the region of *anomalous dispersion* ($dn/d\omega < 0$). Note that in this single resonance model of the dielectric constant,

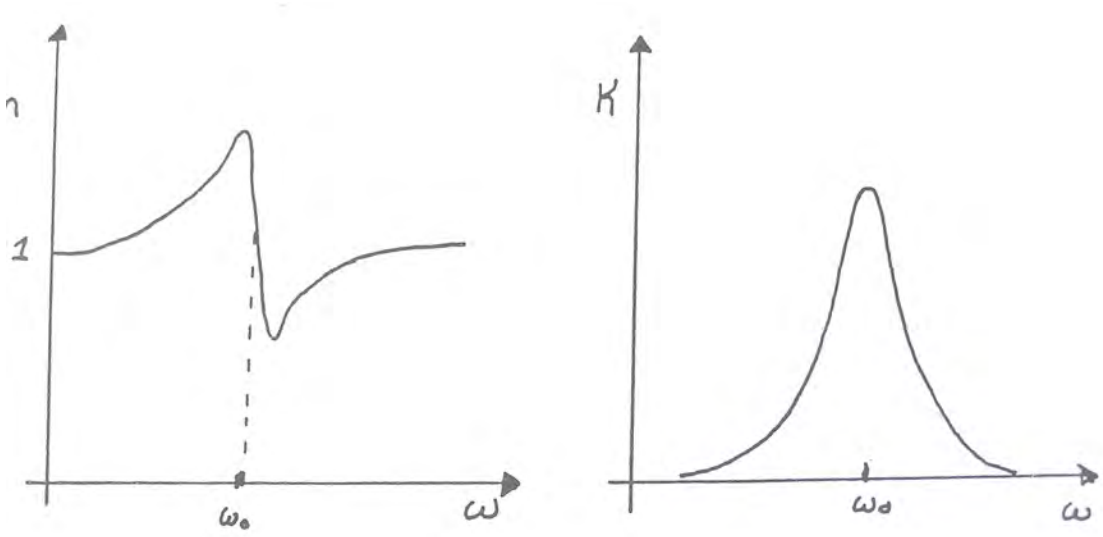


FIGURE 1.7.2. Graphs of the real and imaginary parts of the refractive index functions of a Lorentz medium

the refractive index can be less than unity, leading to a phase velocity in certain frequency regions which is greater than the velocity of light in vacuum! This is of no concern, and particularly not inconsistent with the theory of relativity since no information is actually carried at the phase velocity.

The imaginary part of the refractive index, which determines the absorption characteristics, shows the typical bell-shaped curve associated with a Lorentzian function. The peak absorption occurs at the resonance frequency, while the full width at half maximum of $\kappa(\omega)$ is exactly equal to γ. If one were to use the classical Lorentz model for actual atomic resonances then a typical value of the phenomenological damping coefficient, γ, is $10^9 s^{-1}$ leading to an atomic lifetime of $\gamma^{-1} = 1$ nanosecond. This corresponds to the time it takes for an atomic oscillator to lose $1 - e^{-2} \simeq 0.9$ of its energy. For this value, the approximation $\gamma << \omega_0$ is well justified in the visible region of the spectrum.

The Lorentz model can easily be extended to take into account the multiple resonance frequencies observed in a real dielectric. This can be done in two ways, both leading to the same end result, but for quite different pictures of the phenomenological oscillators.

In the first approach we suppose that in any given medium oscillators with different frequencies can exist. A certain fraction of them, f_i, have a resonance frequency ω_i and damping coefficient γ_i where ω_i corresponds to one of the experimentally observed resonant frequencies. We can then write the dielectric response function as a superposition of the response function of all oscillators and obtain

$$\epsilon = \epsilon_0 \left(1 + \frac{Ne^2}{m\epsilon_0} \sum_i \frac{f_i}{\omega_i^2 - \omega^2 - i\omega\gamma_i} \right) \tag{1.7.7}$$

where Nf_i is the number density of all the oscillators having a resonance frequency ω_i so that

$$\sum_i f_i = 1.$$

Equation 1.7.7 can also be obtained if we suppose that each microscopic constituent is capable of responding at all the different resonance frequencies and f_i in this case is interpreted as the "strength" of response at a particular resonance frequency. This approach is much more difficult to justify classically since a single electron has at most three degrees of freedom and a "spring" can only have one spring constant. Nevertheless, this approach treats all the "atoms" alike and, in fact, is basically what comes out of a more sophisticated quantum mechanical approach to the problem. The individual f_i (even in the quantum mechanical picture) are referred to as *oscillator strengths.*

With this more extensive formula for the dielectric constant we can approximately obtain the refractive index from dc ($\omega = 0$) all the way up to the hard x-ray region ($\omega = 10^{19} s^{-1}$) for many common materials. Experimentally it has been found that the typical variation of the *refractive index* for a dielectric medium is that shown in figure 1.7.3, a variation which is consistent with equation 1.7.7.

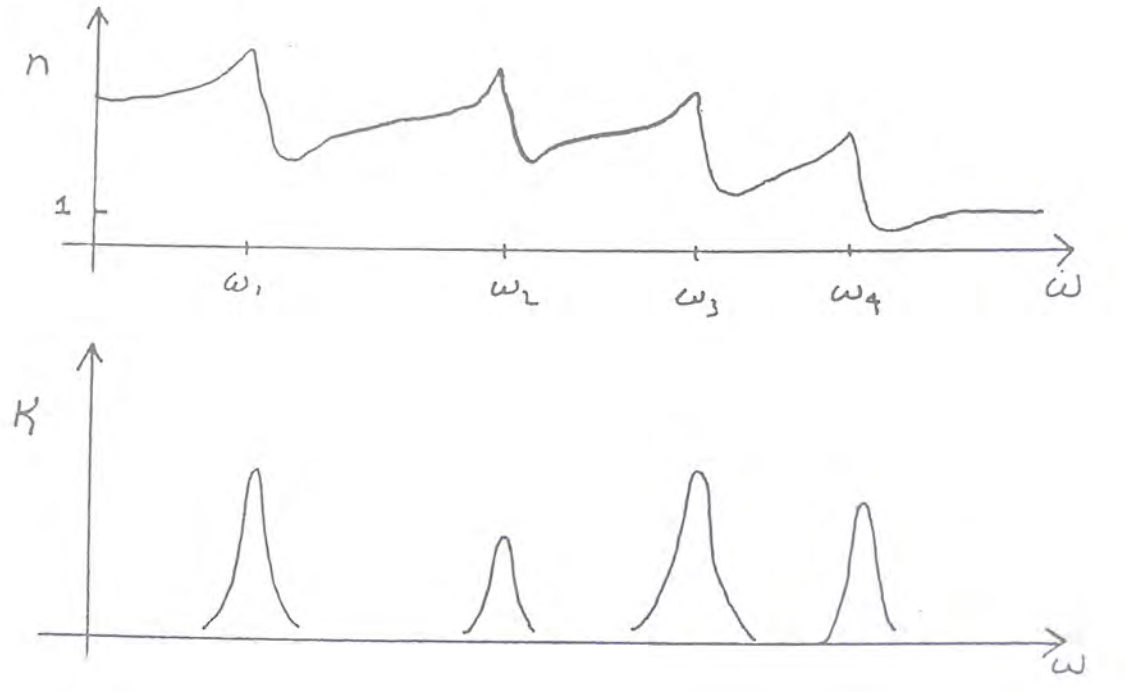


FIGURE 1.7.3. Schematic variation of the real and imaginary parts of the refractive index for a multiple resonance dilute gas

The low frequency resonances typically correspond to rotational ($10^{10} < \omega_0 < 10^{13} s^{-1}$) and vibrational ($10^{13} < \omega_0 < 10^{15} s^{-1}$) *degrees of freedom* while the high frequency resonances ($\omega_0 > 10^{16} s^{-1}$) are associated with electronic degrees of freedom. The electronic resonances have much higher frequencies, in essence, because the oscillating entity, the electron, has a mass much less than that of a nucleus.

The imaginary part of the refractive index shows absorption peaks for each of the rotational, vibrational and electronic resonances. Between resonances κ typically drops close to zero indicating that the medium is transparent. In the highly schematic graph the number of resonances displayed is small and their width is exaggerated for clarity. The actual strength and width of the resonances reflect the properties of the individual resonances. In the case of solids and liquids in particular, there may be a *band* or continuum of resonances particularly for the electronic resonances. In such a case there is considerable overlap between the individual resonances and one only observes a broad, smeared-out resonance. The width of the band then is simply a measure of the distribution of resonance frequencies and not related to the lifetime of an individual oscillator.

The real part of the refractive index is always greater than zero and the overall trend is for it to decrease with increasing frequency, displaying strong dispersive behaviour in the vicinity of a resonance. At the highest frequency, far from any resonances in the system, the refractive index is purely real and equal to unity. In this limit the system cannot respond to the very rapid field variations and light passes through the collection of atoms as if it were propagating in a vacuum. For $\omega = 0$, in the case of a dilute "gas" the real part of the refractive index is given by

$$n^2 = 1 + \frac{Ne^2}{m\epsilon_0} \sum_i \frac{f_i}{\omega_i^2}$$

If many of the low resonance frequencies have high oscillator strengths then the "dc" refractive index is high, as is the dielectric constant, which is purely real for $\omega = 0$. For example, water has many active resonances in the infrared, and its relative dielectric constant, $\epsilon(0)/\epsilon_0$ is 81. In the visible region of the spectrum, far from any resonances, the relative dielectric constant has a value 1.75. In the case of glass whose resonances have their greatest oscillator strengths in the ultraviolet region of the spectrum, the relative dielectric function is nearly constant with a value of $\simeq 2.3$ up to the visible region of the spectrum.

Table 1 lists refractive indices of several common optical materials in the mid-visible region of the spectrum ($\lambda = 0.60\ \mu m$).

Figure 1.7.4 shows the dispersion in the refractive index of common glasses in the visible region of the spectrum. Because the optical resonances associated with electrons lie in the ultraviolet region of the spectrum and vibrational

material	$\hat{n}$	material	$\hat{n}$
vacuum	1.0000	air	1.00029
water	1.333	ethyl alcohol	1.361
carbon	1.628	sodium chloride	1.50
diamond	2.419	fused quartz	1.46
light flint glass	1.57	medium flint glass	1.63
dense flint glass	1.66	extra dense flint	1.77
sapphire	1.77	magnesium fluoride	1.38
zirconium dioxide	2.1	titanium oxide	2.4
cerium fluoride	1.63	Pyrex glass	1.47
silicon(crystalline)	3.8+0.4i	silver	2.3+3.8i

TABLE 1. Refractive index of some common materials at 0.6 μm

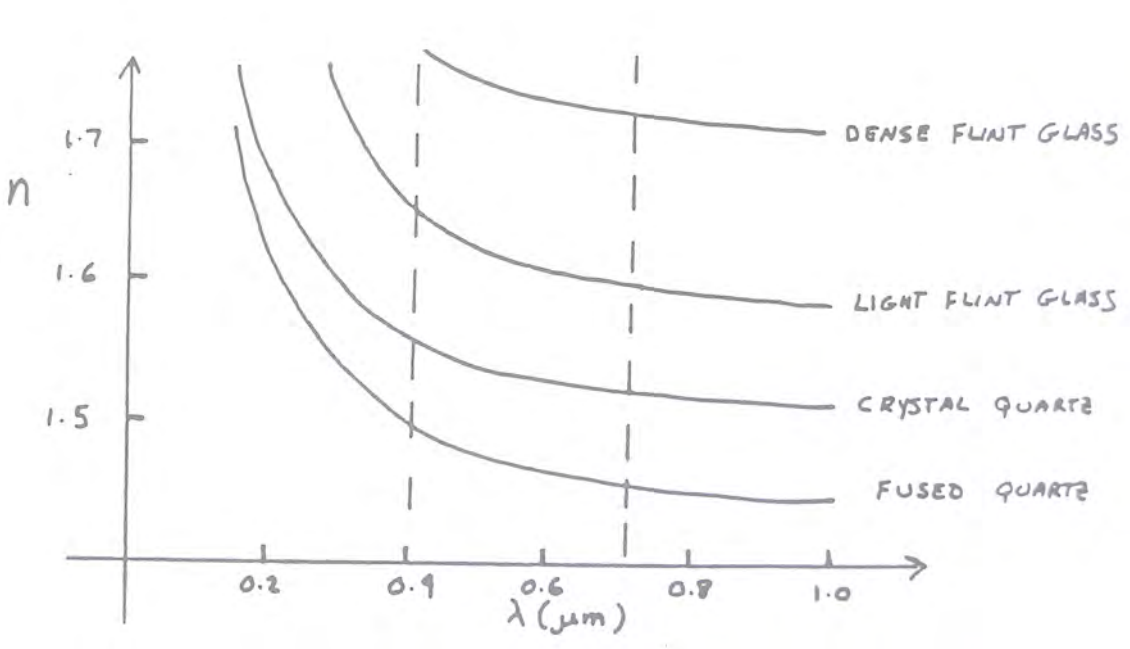


FIGURE 1.7.4. Dispersion of n for common glass in the visible

resonances lie in the infrared, all the materials are nominally transparent and show a real part of the refractive index that increases with increasing frequency.

1.7.2. Free electron systems. The Lorentz model can also be used to obtain expressions for the contribution to the dielectric constant of free electrons in *conducting materials* such as *metals*. The optical properties of metals, particularly in the visible and infrared portions of the spectrum are dominated by the *free electrons*, those which are not attached to any atoms but are free to move about the solid. These free, or conduction electrons, have $\omega_0 = 0$ if we set $K = 0$ in equation 1.7.1. Setting all the resonance frequencies equal to zero in equation 1.7.7 gives the following expressions for the dielectric constant:

$$\hat{\epsilon}/\epsilon_0 = 1 - \frac{Ne^2}{m\epsilon_0}\frac{1}{\omega^2 + i\omega\gamma} = 1 - \frac{\omega_p^2}{\omega^2 + i\omega\gamma}$$

where

$$\omega_p^2 = \frac{Ne^2}{m\epsilon_0}.$$

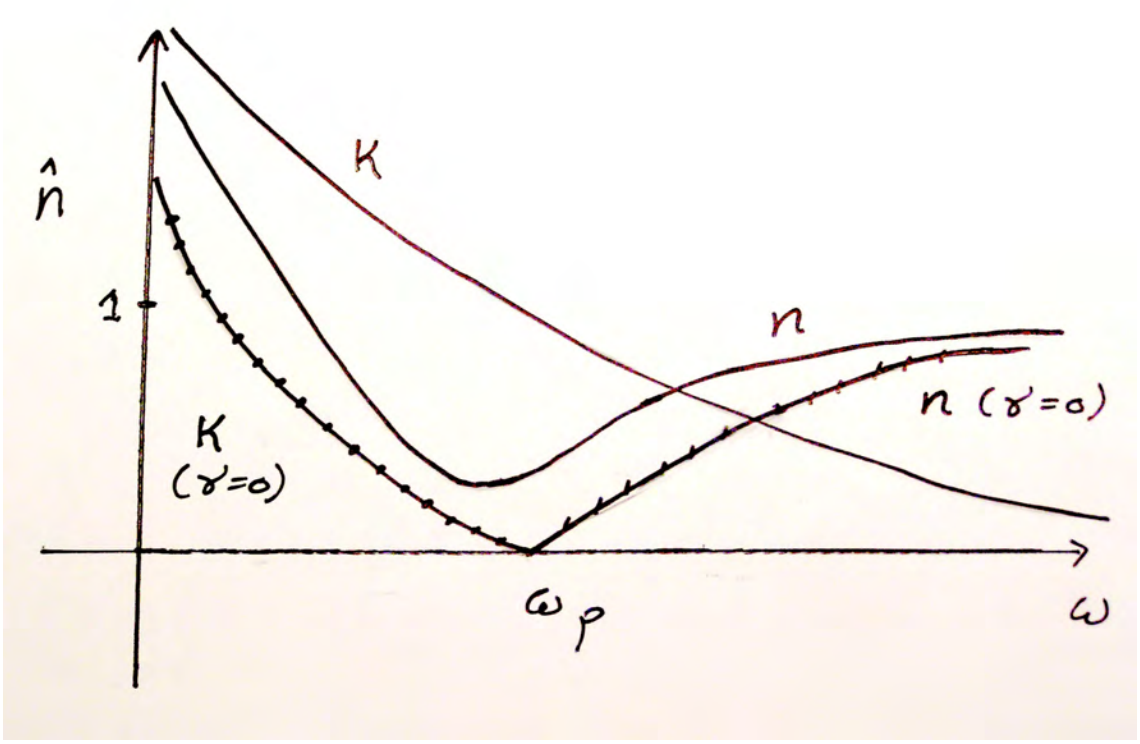


FIGURE 1.7.5. Variation of the real and imaginary parts of the refractive index for a metal in the visible region of the spectrum. If $\gamma = 0$, then $n = 0$ below the plasma frequency, ω_p, while $\kappa = 0$ above it.

The frequency ω_p is referred to as the *plasma frequency* for reasons that will become apparent below. The real and imaginary parts of the dielectric constant are then given by

$$\frac{\epsilon_R}{\epsilon_0} = n^2 - \kappa^2 = 1 - \frac{\omega_p^2}{\omega^2 + \gamma^2}$$

$$\frac{\epsilon_I}{\epsilon_0} = 2n\kappa = \frac{\gamma}{\omega}\frac{\omega_p^2}{\omega^2 + \gamma^2}$$

The dotted line illustrates $n(\omega)$ for the special case $\gamma = 0$ for which $\kappa = 0$ beyond the plasma frequency and $n = 0$ below the plasma frequency.

Typically, for a metal such as aluminium or silver the value of the characteristic plasma frequency is $\omega_p \sim 10^{16} s^{-1}$ and the value of the damping parameter which is governed by electron-lattice collisions is $\gamma \sim 10^{14} s^{-1}$. In terms of this model it is easy to show that the real part of the (frequency dependent) conductivity of the metal is given by

$$\sigma(\omega) = \epsilon_0 \gamma \frac{\omega_p^2}{\omega^2 + \gamma^2}$$

Bound electron resonances in the ultraviolet region of the spectrum can make a contribution to the dielectric constant of metals and their effect can be included by adding the bound electron contribution from equation 1.7.7 to the dielectric constant of the free electron metal above.

Figure 1.7.5 illustrates the typical variation of the real and imaginary part of a metal in the visible portion of the spectrum. The imaginary part of the refractive index is seen to decrease with increasing frequency. This is an indication of the fact that metals become more opaque in the infrared region of the spectrum where the electrons can absorb energy from the electric field and dissipate it to the lattice. In the ultraviolet region, if we neglect the contribution of a few bound electron resonances, most metals are transparent.

In the limit of negligible damping it can be seen that the real part of the refractive index goes to zero when $\omega = \omega_p$ (Figure 1.7.5). A zero refractive index corresponds to an infinite wavelength of the electric field excitation in the metal. What this means is that the entire electron gas is responding to a spatially uniform electric field oscillating at the plasma frequency. The plasma frequency therefore corresponds to a uniform collective oscillation

or "sloshing" of the electrons inside the metal, rather than the resonance frequency of an individual electron. This is the *only* natural resonance frequency of the uniform electron gas.

1.8. Pulses

In general no wave is truly monochromatic, an attribute which implies an infinite wave train. However, any optical disturbance can be written as a superposition of plane waves of different frequencies giving rise to a *pulse* in space and time. In this section we consider the characteristics of a pulse propagating in a transparent medium.

At a constant spatial position $\vec{r}$, any optical disturbance, $\vec{E}(t)$, can be written as the superposition of monochromatic disturbances, which in the case of a continuum, becomes in phasor notation,

$$\vec{\mathcal{E}}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \vec{\mathcal{E}}(\omega)e^{-i\omega t}d\omega.$$

The amplitudes of the spectral field density, $\vec{\mathcal{E}}(\omega)$, at different frequencies is obtained from the Fourier transform relation (see the Appendix at the end of this chapter),

$$\vec{\mathcal{E}}(\omega) = \int_{-\infty}^{\infty} \vec{\mathcal{E}}(t)e^{i\omega t}dt.$$

For a plane wave scalar pulse propagating in the z direction we can separate the magnitude and phase of the amplitude to write

$$\mathcal{E}(\omega, z) = \mathcal{E}_0(\omega)e^{ik(\omega)z}$$

and the optical pulse is given by

$$\mathcal{E}(t, z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathcal{E}_0(\omega)e^{i(k(\omega)z-\omega t)}d\omega.$$

If the Fourier amplitude or spectral density is appreciable only in a narrow band of frequencies about ω_c we can expand $k(\omega)$ with $\triangle\omega = \omega - \omega_c$ to yield

$$k(\omega) = k(\omega_c) + \left.\frac{\partial k}{\partial \omega}\right|_{\omega_c} \triangle\omega = k(\omega_c) + \frac{1}{v_g}\triangle\omega$$

where we have defined a *group velocity,* whose significance will be made clear in a moment, by

$$v_g = \left.\frac{\partial \omega}{\partial k}\right|_{\omega_c}.$$

It follows that

$$\mathcal{E}(t, z) = exp\left\{i\left(k(\omega_c)z - \omega_c t\right)\right\}\frac{1}{2\pi}\int_{band} \mathcal{E}_0(\omega)exp\left\{-i\triangle\omega\left(t - z/v_g\right)\right\}d\omega.$$

The factor in front of the integral is a monochromatic carrier wave with carrier frequency ω_c with a phase speed

$$v_\phi = \frac{\omega_c}{k(\omega_c)} = \frac{c}{n(\omega_c)}$$

while the integral indicates a (real) *envelope function* propagating with the group speed, which can be shown to be

$$v_g = \frac{d\omega}{dk}\Big|_{\omega_c} = \frac{c}{n(\omega_c)}\left(1 + \frac{\omega_c}{n(\omega_c)}\left.\frac{dn}{d\omega}\right|_{\omega_c}\right)^{-1}.$$

For a nondispersive medium, or far from any resonances of the system, the phase and group velocities are the same but they can differ considerably near an optical resonance. The group velocity determines the rate at which information or energy can travel. Whereas v_g is always less than the speed of light in vacuum as required by the theory of relativity, the phase velocity, v_ϕ, is not restricted in the same way as we saw earlier.

A monochromatic beam must have an infinite extent in space and time. For such an ideal wave its bandwidth is zero but its pulse width is infinite. Similarly a pulse of zero width in time or space must have an infinite band width as the Fourier transform relations show. The pulse width and bandwidth are inversely related and it is easy to show from the Fourier transform relations that the electric field bandwidth of the pulse, $\triangle\omega_p$, and the pulse width, τ_p, are related by

$$\triangle\omega_p\tau_p \geq 2\pi.$$

Hence the time-bandwidth product of a pulse is always greater than unity. If a pulse satisfies the lower limit (unity) then the minimum pulse length is determined directly from the bandwidth and the pulse is said to be *Fourier Transform limited*. A typical pulse is shown in figure 1.8.1.

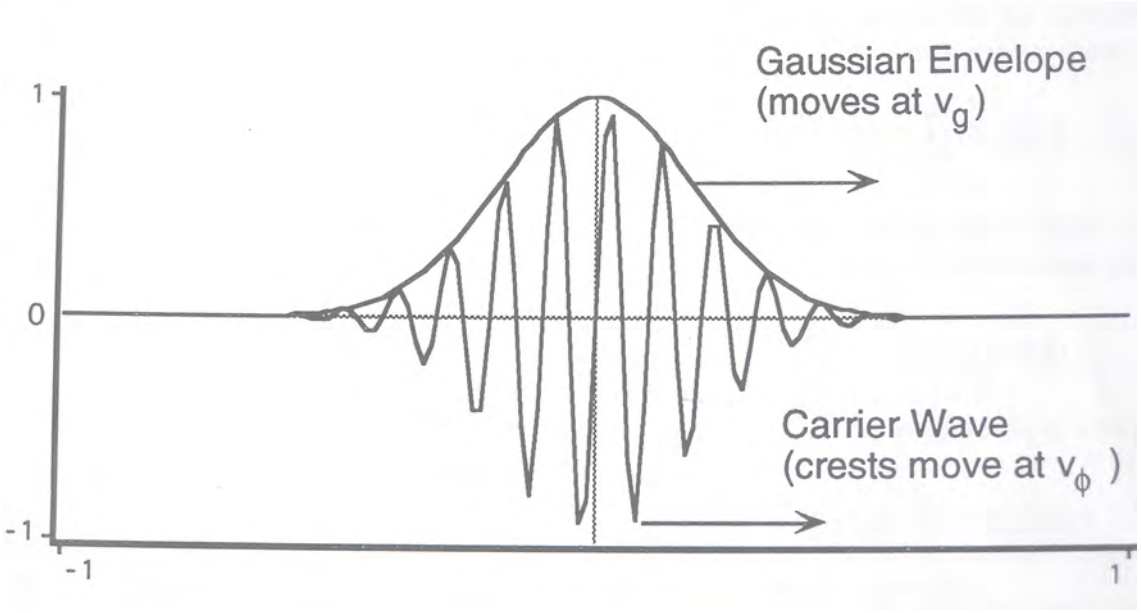


FIGURE 1.8.1. A Gaussian wave packet moves at $v_g < c$, while the carrier wave fronts move through it with phase velocity v_ϕ, which may be greater than or less than c.

The factor unity is not to be taken literally since it depends on one's definition of pulse width and bandwidth. By symmetry we can also Fourier analyze the pulse in the spatial domain from which we obtain the result that

$$\triangle k \triangle z \geq 2\pi$$

where $\triangle k$ is the spread in the propagation constant and $\triangle z$ is the spatial extent of the pulse. The Fourier transform pairs of some common pulse types are shown in Figure 1.8.2. Note: it is the real parts of the phasors that is plotted for the time dependent fields: The high frequency carrier part is suppressed for the pulses, so effectively only the envelope is plotted. For the spectrum, it is $Re\,[\mathcal{E}(\omega)] = E(\omega)$ that is shown.

Appendix 2.1: Fourier Transforms

A periodic function $f(t)$ of period T can be represented as a Fourier series

$$f(t) = \sum_{n=-\infty}^{n=\infty} F_n e^{-in\omega_p t}$$

where $\omega_f = 1/T$ is known as the fundamental frequency and the F_n are the Fourier amplitude coefficients. These coefficients can be determined from the integral relation

$$F_n = \frac{1}{2\pi T}\int_{-T/2}^{T/2} f(t) e^{in\omega_f t} dt$$

Even if a function is aperiodic (*i.e.*, not periodic) it can still be represented by an integral Fourier series. This can be done by treating the function as periodic but with an infinite period, *i.e.*, $T \to \infty$ with $\omega_f \to 0$. We then have that $\omega_f n$ becomes a continuous variable ω and

$$f(t) = \lim_{\omega_f \to 0} \sum_n F_n e^{-in\omega_p t} = \lim_{\omega_f \to 0} \sum_n \frac{F_n}{\omega_f} e^{-in\omega_p t}\omega_f = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega) e^{-i\omega t} d\omega$$

where

$$F(\omega) = \lim_{\omega_f \to 0} \frac{2\pi F_n}{\omega_f}$$

The *Fourier transform relation* gives

$$F(\omega) = \lim_{\omega_f \to 0} \frac{2\pi F_n}{\omega_f} = \lim_{T \to \infty} \frac{2\pi}{\omega_f T}\int_{-T/2}^{T/2} f(t) e^{in\omega_f t} dt = \int_{-\infty}^{\infty} f(t) e^{i\omega t} dt$$

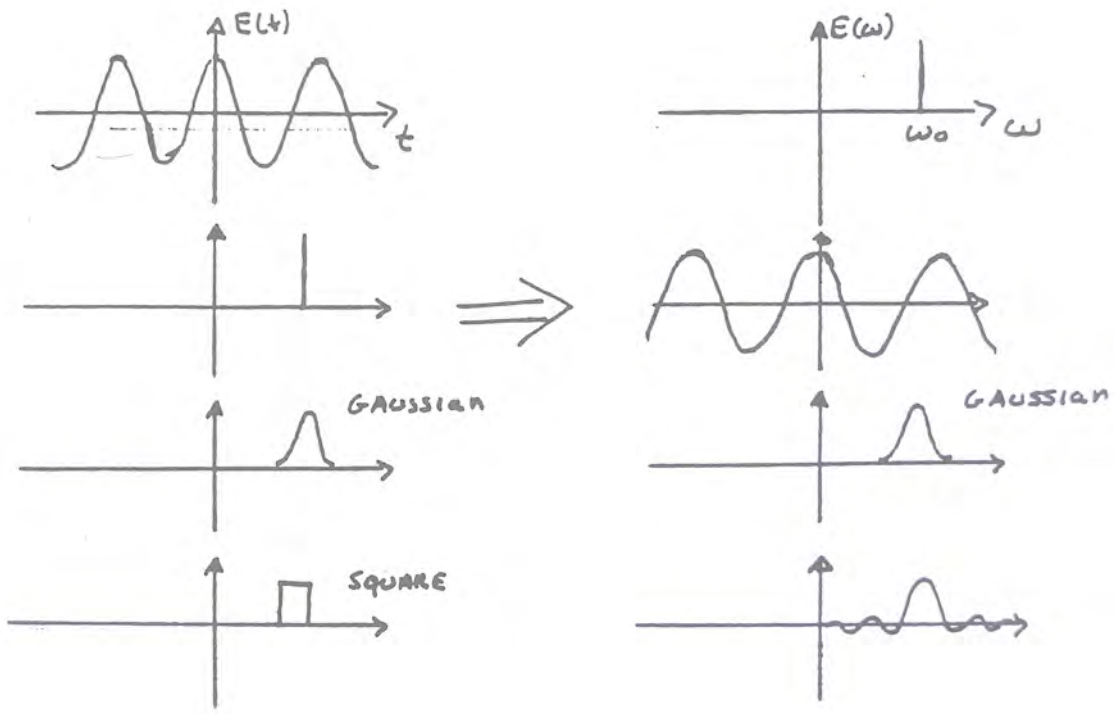


FIGURE 1.8.2. Fourier transform pairs for several common optical pulses.

If $f(t)$ is a real function then

$$F(-\omega) = F^*(\omega).$$

References

M. Born and E. Wolf, *Principles of Optics*, Cambridge Press, 2002.
P. Lorrain and D. Corson, Electromagnetic Fields and Waves, W.H. Freeman, 1988
H.A. Haus, *Waves and Fields in Optoelectronics*, Prentice-Hall, Englewood Cliffs NJ, 1984.
G.R. Fowles, *Introduction to Modern Optics*, Holt, Rinehart and Winston, New York, 1968.
E. Hecht, *Optics*, Pearson, New York, 2002.

Problems

1. Show that neither the spherical wave nor the cylindrical wave solutions given in this chapter satisfy Maxwell's equations at the origin.

2. A plane harmonic wave has an electric field whose amplitude varies in space and time according to

$$E(t,z) = E_0 cos\left\{\pi 10^{15}\left(t - z/0.75c\right)\right\}$$

Determine:
a) the frequency of the light;
b) the wavelength;
c) the index of refraction of the material in which the light is propagating.

3. The DC conductivity of silver is 6.8×10^7 SI units and the free electron density is $1.5 \times 10^{28} m^{-3}$. Considering silver to be a free electron metal what is its plasma frequency? What is the electron damping time?

4. Consider a fictitious material with a single resonance at $\lambda = 0.3\mu m$ and with a damping coefficient of $10^{13} s^{-1}$. Determine the percentage difference between the group and phase velocity at $\lambda = 0.3\mu m$ and $\lambda = 0.5\mu m$. What is the maximum absorption coefficient of this material?

5. Augustin Cauchy (1789-1857) was able to determine an empirical equation for the refractive index for a variety of materials that were transparent in the visible region of the spectrum. He found that n could be represented by the power series expansion

$$n = C_1 + \frac{C_2}{\lambda^2} + \frac{C_4}{\lambda^4} + ...$$

What is the physical significance of C_1 ? What is a typical value of C_2 for glass?

6. In 1871 Sellmeir derived the following equation for the refractive index of transparent media

$$n^2 = 1 + \sum_i A_i \frac{\lambda^2}{\lambda^2 - \lambda_i^2}$$

where the A_i are constant and the λ_i are associated with the natural resonances. Show that for very long wavelengths, Cauchy's equation is an approximation of Sellmeir's equation.

7. A Gaussian pulse has a full width at half maximum of 1 picosecond and a carrier wavelength of $1\mu m$. Determine the frequency and wavelength bandwidth of the pulse if the pulse is Fourier transform limited. If the pulse is passing through a piece of glass with n = 1.5 with a dispersion of $dn/d\omega = 10^{-17}s$ at $\lambda = 1\mu m$ determine the difference between the group velocity and the phase velocity.

Special Note on Phase Conventions. The total phase ϕ associated with a monochromatic beam is written here as

$$\phi(\vec{r}, t) = \vec{k} \cdot \vec{r} - \omega t + \phi_0$$

and in section 1.4 it is used in $cos(\omega t - \phi)$, which is equivalent since cosine is an even function with $cos(-x) = cos(x)$. This is also the case in phasor notation, where it is understood that the real part is used. It is also possible to use the representation

$$\phi(\vec{r}, t) = -\vec{k} \cdot \vec{r} + \omega t + \phi_0$$

which effectively reverses the sign of the phase: as time passes (t increases) the total phase is increasing. This causes differences in the subject when we consider relative phase changes; particularly it alters the mathematical definition of right and left-handed circular polarization. *Be careful when switching between conventions.* Usage is not universal.

CHAPTER 2

Energy and Linear Momentum in an Electromagnetic Wave.

Over all, rocks, woods, and water, brooded
the spirit of repose, and the silent energy of
nature stirred the soul to its inmost depths
Thomas Cole

In this chapter we consider light as an energetic beam. Starting with Maxwell's equations we illustrate what the sources of this energy are and derive an expression for the energy or power carried by the beam in terms of basic electromagnetic quantities. Loss or absorption of energy is a fundamental result of the interaction of light with matter and we derive some general results concerning this phenomenon. The different units of light are reviewed and the circumstances in which they might be important are discussed. Finally we consider the linear momentum properties of light beams and the radiation pressure an electromagnetic wave possesses.

2.1. Transport of Energy in an Electromagnetic Wave

In the previous chapter we saw that an electromagnetic wave travelling in vacuum propagates without loss (or gain) of amplitude. Here the main results are given for the energy in electromagnetic fields *in non-dispersive media*. In section 2.3 we consider dispersive media.

If we dot multiply Faraday's law by $\vec{H}$ and add to this the dot multiplication of Ampere's law by $-\vec{E}$ we obtain

$$\vec{\nabla}\cdot(\vec{E}\times\vec{H})+\frac{\partial}{\partial t}(\frac{1}{2}\epsilon_0 E^2)+\frac{\partial}{\partial t}(\frac{1}{2}\mu_0 H^2)+\vec{E}\cdot(\frac{\partial\vec{P}}{\partial t}+\vec{J})+\vec{H}\cdot\frac{\partial(\mu_0\vec{M})}{\partial t}=0 \tag{2.1.1}$$

which can be written in the form

$$\vec{\nabla}\cdot\vec{S}+\frac{\partial}{\partial t}(w_e+w_m)+\vec{E}\cdot\vec{J}=0 \tag{2.1.2}$$

where

$$\vec{S}=\vec{E}\times\vec{H}$$
$$w_e=\frac{1}{2}\epsilon_0(1+\chi)E^2=\frac{1}{2}\epsilon E^2$$
$$w_m=\frac{1}{2}\mu_0(1+\chi_m)H^2=\frac{1}{2}\mu H^2.$$

Equation 2.1.2 is a continuity equation and amounts to a statement of the law of power conservation for electromagnetic fields. It is generally referred to as *Poynting's theorem*. The vector $\vec{S}$ is referred to as the Poynting vector and gives the magnitude and direction of the power flux or irradiance (units of $\mathrm{W/m^2}$) in the electromagnetic wave. For an electric field in a medium of susceptibility χ or dielectric constant ϵ the stored energy density associated with the electric field is given by the quantity w_e while the energy density associated with the magnetic field is w_m The units of the energy density are J/m^3.

A few comments are in order concerning the transition from equation 2.1.1 to 2.1.2. First of all the term interpreted as the energy density in the electric field incorporates the term depending on the time dependent polarization density. Indeed the term $\vec{E}\cdot(\frac{\partial\vec{P}}{\partial t})$ in fact represents the power per unit volume delivered to the polarization current density (associated with the movement of the bound electrons). In general it is not possible to determine whether the time integral of this term is stored, dissipated or both. Such a criterion can be developed for a given constitutive law and for the simple one adopted in chapter 1, the power per unit volume delivered to the polarization process is

$$\vec{E}\cdot(\frac{\partial\vec{P}}{\partial t})=\vec{E}\cdot\frac{\partial}{\partial t}(\epsilon_0\chi\vec{E})=\frac{\partial}{\partial t}(\frac{1}{2}\epsilon_0\chi E^2)$$

Therefore the w_e term incorporates the energy stored in the polarization field as well as the energy stored in the electric field. A similar argument accounts for the form of w_m.

The quantity $\vec{E}\cdot\vec{J}$ in equation 2.1.2 is a sink term for the continuity equation and is the power per unit volume imparted to the free current density, $\vec{J}$, in the medium. For the linear constitutive relation between $\vec{J}$ and $\vec{E}$ the power density dissipated to the current is σE^2 which, in general, is referred to as the Joule heating term. In vacuum or in a loss-less medium where $\vec{J}=0$ this term is zero. Hence for a nonconducting medium Poynting's theorem becomes

$$\vec{\nabla}\cdot\vec{S}+\frac{\partial w}{\partial t}=0$$

where $w=w_e+w_m$ is the total energy stored in the field.

2.2. Time Average of Sinusoidal Quantities

For light waves with $\lambda\sim 1\mu m$ the corresponding frequency ω is $\sim 10^{16}s^{-1}$. Neither the human eye nor any electronic detector can respond to such rapid variations in the field. What is usually detected, or measured, is a time-averaged quantity such as average optical power, average energy, etc. As such, most detectors are sensitive to the square of the electric field, hence the name *square law detectors*. For radiation of constant amplitude this time average can be over a duration as long as we wish. For pulsed radiation, if we want information about details of the pulse shape, successive averages must be taken on time scales which are long compared to an optical period but short compared to the pulse width.

Here we establish an expression for the time averaged Poynting vector and electromagnetic field density in terms of the amplitudes of the waves. We do so in the case of monochromatic waves. It is a matter of a simple extension to obtain expressions for more complex waveforms via the superposition principle.

Consider then the electric field associated with a monochromatic plane wave to have the time dependence

$$\vec{E}(t)=Re\left\{\vec{\mathcal{E}}_0e^{-i\omega t}\right\}=\frac{1}{2}\left(\vec{\mathcal{E}}_0e^{-i\omega t}+\vec{\mathcal{E}}_0^{*}e^{i\omega t}\right)$$

The time averaged energy density for the electric field is then

$$\langle w_e\rangle=\frac{1}{T}\int_0^T\frac{1}{2}\epsilon\left[E(t)\right]^2dt=\frac{\epsilon}{2T}\int_0^T\frac{1}{4}\left[\vec{\mathcal{E}}_0\cdot\vec{\mathcal{E}}_0e^{-2i\omega t}+\vec{\mathcal{E}}_0^{*}\cdot\vec{\mathcal{E}}_0^{*}e^{2i\omega t}+2\vec{\mathcal{E}}_0\cdot\vec{\mathcal{E}}_0^{*}\right]dt=\frac{\epsilon}{4}\left[\vec{\mathcal{E}}_0\cdot\vec{\mathcal{E}}_0^{*}\right]$$

for $T>>\omega^{-1}$. A similar argument gives

$$\langle w_m\rangle=\frac{\mu}{4}\left[\vec{\mathcal{H}}_0\cdot\vec{\mathcal{H}}_0^{*}\right]$$

Similarly, the time averaged Poynting vector or *irradiance* is given by

$$\left\langle\vec{S}\right\rangle=I\hat{k}=\frac{1}{T}\int_0^T\vec{E}(t)\times\vec{H}(t)dt=\frac{1}{2}Re\left[\vec{\mathcal{E}}_0\times\vec{\mathcal{H}}_0^{*}\right]$$

which for plane waves in a dispersionless medium leads to

$$I=Re\left[\frac{1}{\sqrt{\mu\epsilon}}\langle w\rangle\right]=\frac{c}{n}\epsilon_0\left\langle E(t)^2\right\rangle=\frac{1}{2}cn\epsilon_0E_0^2$$

with $\hat{k}$ the unit vector along the direction of plane wave propagation. Here we have used the fact that the time averaged electric and magnetic field energy densities are the same in non-dispersive media. In SI units $E_0\sim 30\sqrt{I}$. Hence a $1\,\mathrm{Wm}^{-2}$ plane wave has an associated peak field strength of 30 Volts/m.

2.3. Poynting's Theorem for Dispersive Media

A complex Poynting theorem may be derived from the complex form of Maxwell's equations for monochromatic waves in the case of dispersive (or lossy) media. Because the time averages of the sinusoidally varying functions are evaluated from the products of the complex vector amplitude of one vector with the complex conjugate of the other vector, the complex Poynting theorem aims at a relation for $\vec{\mathcal{E}}\times\vec{\mathcal{H}}^{*}$ and $\vec{E}\cdot\vec{J}^{*}$. Dot multiplication of equation 1.5.1by $\vec{\mathcal{H}}$ and of the complex conjugate of equation 1.5.2by $-\vec{\mathcal{E}}$ and adding gives

$$\vec{\nabla}\cdot(\vec{\mathcal{E}}\times\vec{\mathcal{H}}^{*})-i\omega\left(\mu\vec{\mathcal{H}}\cdot\vec{\mathcal{H}}^{*}-\epsilon^{*}\vec{\mathcal{E}}\cdot\vec{\mathcal{E}}^{*}\right)+\vec{E}\cdot\vec{J}^{*}=0$$

where it is understood that the dielectric constant and magnetic permeability are evaluated at frequency ω. The term $Re\left[i\omega\left(\mu\vec{\mathcal{H}}\cdot\vec{\mathcal{H}}^{*}-\epsilon^{*}\vec{\mathcal{E}}\cdot\vec{\mathcal{E}}^{*}\right)\right]$contributes to the real part of the divergence of the complex Poynting vector in the same way as the density of power dissipated by $\vec{\mathcal{J}}$, $Re\left[\vec{E}\cdot\vec{J}^{*}\right]$contributes to it. Therefore we interpret it as the

TABLE 1. Some Optical Units

Radiometric name	symbol	unit	definition
Energy	E	J	energy radiated into 4πsteradians
Power	P	Js^{-1}	energy per unit time
Fluence	F	Jm^{-2}	energy per unit area
Brightness	B	$J(sr)^{-1}$	energy per unit solid angle
Spectral Brightness	B_λ	$J(sr)^{-1}m^{-1}$	energy per unit solid angle per unit wavelength
intensity	I_Ω	$Js^{-1}sr^{-1}$	power per unit solid angle
irradiance	I	$Js^{-1}m^{-2}$	power per unit area

power per unit volume dissipated in the dielectric and magnetic medium. In a lossless medium, for a real frequency ω, one must therefore have that

$$Im\left[\epsilon^*\vec{\mathcal{E}}\cdot\vec{\mathcal{E}}^*\right] = Im\left[\epsilon^*\right]\vec{\mathcal{E}}\cdot\vec{\mathcal{E}}^* = 0$$

and

$$Im\left[\mu\vec{\mathcal{H}}\cdot\vec{\mathcal{H}}^*\right] = Im\left[\mu\right]\vec{\mathcal{H}}\cdot\vec{\mathcal{H}}^* = 0$$

meaning that the imaginary parts of the dielectric constant and magnetic permeability must be zero.

2.4. Optical Units

The description of energy and its flow represents one of the main sources of language confusion in optics. Words like intensity, brightness and power flux have assumed a vernacular meaning which may or may not have anything to do with the technical terms used by people who work with optics on a day-to-day basis. The confusion is further exacerbated if one considers historical units associated with light such as candle-power, foot candles; and lumens. Many of the units of optics have their origin in attempts to quantify the power per unit area at a surface, or per steradian into a solid angle. The radiometric system of units is derived from the knowledge that light is an electromagnetic wave and the SI system of units gives a precise meaning to quantities that describe it. Table 1 lists the different units attached to various measures of light parameters of interest to us as physicists.

There are other units as well but they are of minor usage. The quantity irradiance is commonly called the intensity in the laboratory. The irradiance determines the rate at which the temperature of a medium increases while the fluence determines its final temperature.

2.5. Linear Momentum and Radiation Pressure of Light

As long ago as 1619 Kepler pointed out that sunlight might account for the broad extent of a comet's tail since it exerted pressure on the tiny particles evaporating from the comet. This, of course, helped the corpuscular theory of light gain favor in the 17th and 18th century although it didn't put a stranglehold on the wave theory of the day since no one was able to measure the "*radiation pressure*" in the laboratory. Maxwell revived the idea of light pressure in the context of his wave theory and in vacuum, showed its magnitude to be equal to the total energy density. The pressure is applied in the direction of the Poynting vector and indeed is related to the Poynting vector by

$$\vec{P}r = \vec{S}/c \tag{2.5.1}$$

with the time averaged pressure given by

$$\left\langle \vec{P}r \right\rangle = \frac{I}{c}\hat{k}$$

Since pressure is force per unit area and force is related by Newton's second law to the rate of change of *linear momentum*, $\vec{p}$, we have that for an area A of an object

$$\frac{d\vec{p}}{dt} = A\left\langle \vec{P}r \right\rangle = AI\hat{k}/c = \frac{1}{c}\frac{dE}{dt}\hat{k}$$

and so the momentum delivered by the light beam is

$$\vec{p} = \frac{E}{c}\hat{k}$$

where E is the energy in the beam. Note that if the light beam is reflected by the object the momentum imparted to the object is twice that of the light beam!

In terms of the quantum theory of light,in which a monochromatic beam of light carries quanta of energy of magnitude $E = \hbar\omega$, it is interesting to see that this relation gives for the momentum carried by a quantum

$$\vec{p} = \frac{\hbar\omega}{c}\hat{k} = \hbar\vec{k}$$

where $\hbar$ is Planck's constant.

To provide an indication of the magnitude of the radiation pressure and perhaps why the scientists of a few centuries ago were unable to detect it, consider the radiation pressure exerted by sunlight at the Earth's surface. At the equator the irradiance from the sun is approximately $1400\,\mathrm{Wm}^{-2}$. From equation 2.5.1 above this gives a radiation pressure of approximately $5{\times}10^{-6}\,\mathrm{Pa}$ or approximately 10^{-11} atmospheric pressure. Across the Earth's surface however the entire force is considerable, and unlike atmospheric pressure, it acts outward from the sun. In many large stars the total radiation pressure is sufficient to support the star from collapse. Also, for many interplanetary spacecraft, which have journeys lasting years, the total effect of radiation pressure can seldom be neglected.

It is also interesting to point out that since light carries linear momentum, light can be used to cool atoms. While the details of this process requires a quantum treatment, the basic idea is that if atoms absorb light momentum from a beam propagating in a particular direction, the light beam effectively decrease the momentum of those atoms travelling in a direction opposite to that of the light beam. By directing 6 beams of light along the 6 directions associated with $\pm x, \pm y, \pm z$ one can slow down the atoms and reduce the effective temperature associated with their translational motion.

References

M. Born and E. Wolf, *Principles of Optics*, Cambridge Press, 2002.

H.A. Haus, *Waves and Fields in Optoelectronics*, Prentice-Hall, Englewood Cliffs NJ, 1984.

E. Hecht, *Optics*, Pearson, 2002.

Problems

1. Determine the energy densities w_e and w_m and the Poynting vector $\vec{E} \times \vec{H}$ for a plane wave $E_0 cos(kz - \omega t)$ propagating in free space. Check that Poynting's theorem for energy conservation is satisfied.

2. The Nova laser fusion system at the Lawrence Livermore Laboratory is capable of delivering 10^{13} W of optical power at a wavelength of 1.06 μm onto a pellet $100\mu m$ in diameter.

a) Find the rms electric field associated with uniform illumination of the pellet surface for linearly polarized and circularly polarized light.

b) Compare the answer to part a with the field which binds an electron to a proton in the hydrogen atom at the Bohr radius.

3. Calculate the radiation pressure due to sunlight on Voyager II's solar panel if the panel has an area of $10m^{-2}$ and is oriented facing the sun at the orbital position of Uranus. Assume the panel is 15% reflective and 85% absorptive. Assuming the same constant orientation since it left the earth, estimate the total amount of work done on the probe by solar radiation since it left earth, assuming that it has been travelling at a constant speed and has taken five years to reach Uranus.

4. An optical beam of wavelength 0.8 μm and irradiance $10^4\,\mathrm{Wm}^{-2}$is incident on a thick slab of silicon. If the absorption coefficient of the silicon is $10^3\,cm^{-1}$approximately how much does the surface temperature rise in a 1 microsecond interval. Neglect the reflectivity of the surface.

5. Construct the complex vector expression for the electric field for a right circularly polarized plane wave propagating in free space in the +z direction with a peak amplitude E_0 occurring at z = 0, t = 0 along the x direction. Derive an expression for the associated complex magnetic field and the complex Poynting vector.

CHAPTER 3

The Vector Nature of Light: Polarization Effects

Let us bathe in this crystalline light!
Edgar Allan Poe

3.1. Introduction

In chapter 1 we saw that light could be described classically as a propagating electric and magnetic field disturbance. For nonmagnetic media we saw that the propagation properties are entirely governed by the dielectric constant and that for isotropic media the orientation of the electric field vector is constant. In isotropic media one can therefore treat light or its associated electric field as a scalar. There are many situations in nature, however, which do not present isotropic symmetry and one must therefore be cognisant of the vector character of the light field. For example, many crystalline materials like quartz and calcite are not optically isotropic in the sense that the induced polarization density depends on the electric field orientation. An angled, planar interface (to be considered in the next chapter) also presents a situation of local anisotropy since the boundary conditions and hence the electromagnetic response are different for the electric field components oriented parallel or perpendicular to the interface.

In this chapter we provide a description and influence of the vector characteristics of light beams, also referred to as *polarization phenomena*. We begin with a discussion of the polarization state of plane, monochromatic waves for which the tip of the electric field vector is confined to move in a plane as we saw earlier. The description of more complicated light beams would have to be considered as a superposition of plane wave states.

3.2. Description of Polarization States

Without loss of generality, let's consider a vector, monochromatic light beam travelling along the z direction. For a plane wave we saw in section 1.4 that the electric field vector is confined to an *(x-y)* plane perpendicular to the direction of propagation. The two components of the complex electric field vector can be used to obtain the real components of the electric field vector through

$$Re\left[\mathcal{E}_x(z,t)\right] = E_{0x}cos(kz-\omega t+\phi_{x0}) = E_{0x}cos(-\omega t+\phi_x)$$
$$Re\left[\mathcal{E}_y(z,t)\right] = E_{0y}cos(kz-\omega t+\phi_{y0}) = E_{0x}cos(-\omega t+\phi_y)$$

where the propagation contributions (*i.e.*, kz) as well as any initial phase difference between the two beams have been lumped into the phase terms ϕ_x and ϕ_y. The vector field is then given by

$$Re\left[\vec{\mathcal{E}}(z,t)\right] = Re\left[\mathcal{E}_x(z,t)\right]\hat{x} + Re\left[\mathcal{E}_y(z,t)\right]\hat{y}.$$

The locus of the tip of the real electric field vector at a fixed $z = z_0$ is an ellipse as shown in figure 3.2.1.

Three parameters can be used to specify the *polarization ellipse*. They, in one way or another, are related to the length of the semi-major (a) and semi-minor (b) axes and the orientation (Ψ) of the ellipse relative to the chosen x and y- axes. Some possible triads which can be used to specify the ellipse are

$$(a, b, \delta = \phi_y - \phi_x)$$
$$(a, b, \Psi)$$

or

$$(E_{0x}, E_{0y}, \delta).$$

The influence of the relative phase, δ, between the two fields components can be seen in figure 3.2.2 which shows how the tip of the electric field vector moves for different values of δ to define the general polarization state of light, which is *elliptically polarized light*.

The arrows indicate the sense of movement of the electric field vector. The convention in depicting these pictures is that the observer is looking at the beam as it approaches him/her. This is also consistent with the choice of orientation of the x and y- axes, for which the z-axis, the direction of propagation, is coming out of the paper.

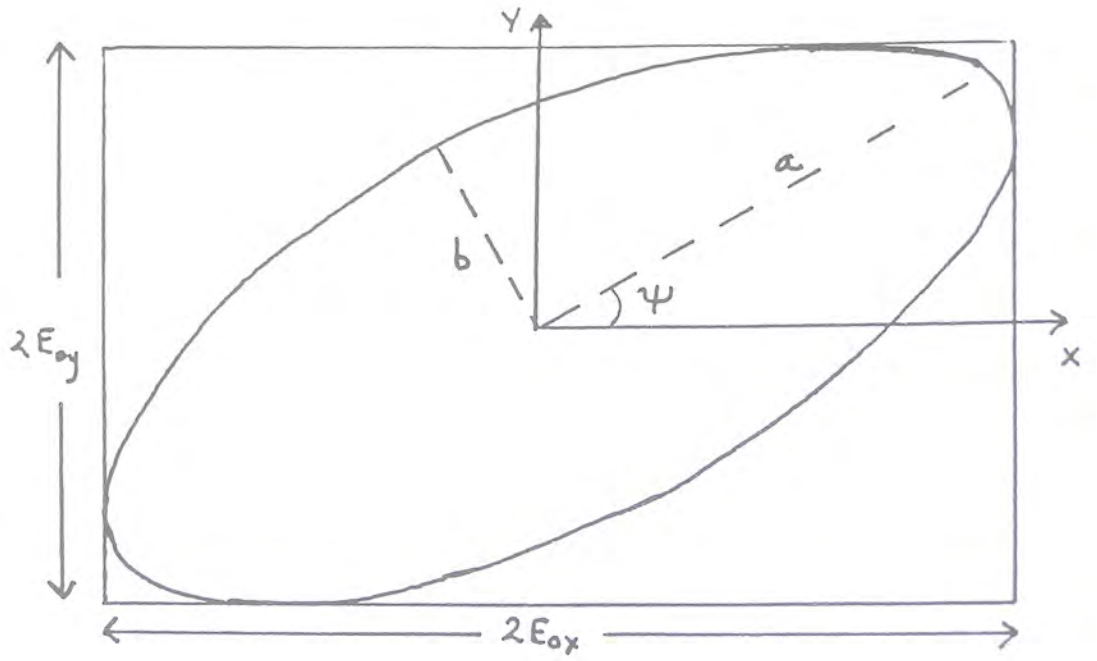


FIGURE 3.2.1. The polarization ellipse.

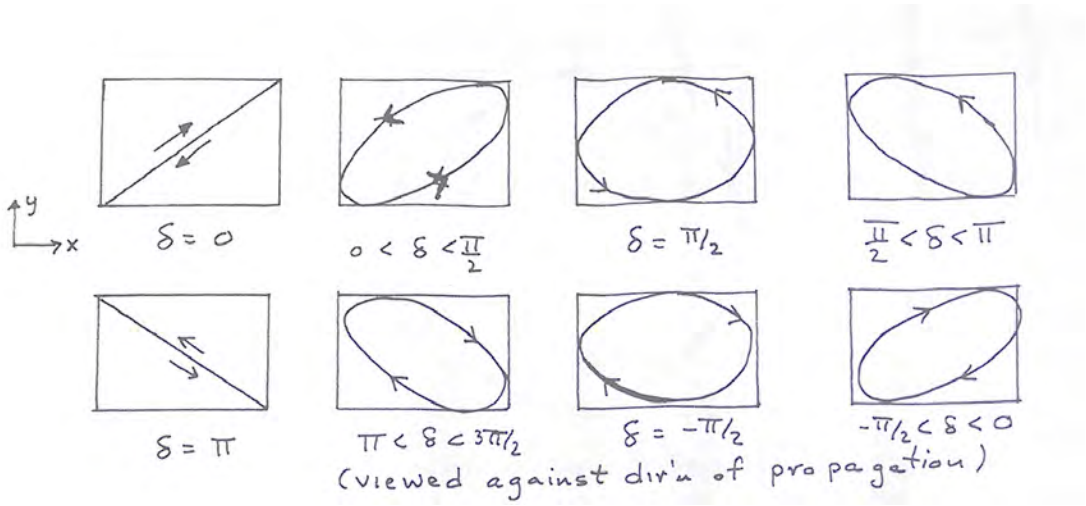


FIGURE 3.2.2. Loci of the tip of the polarization vector for fixed z and different values of δ

Two special cases occur frequently in dealing with polarized light.

1) *Linearly polarized light:* this occurs if $\delta = m\pi$ for $m = 0, \pm 1, \pm 2, ...$ In this case we have

$$\frac{\mathcal{E}_y}{\mathcal{E}_x} = \frac{E_{oy}}{E_{ox}} e^{i\delta} = \pm \frac{E_{oy}}{E_{ox}}$$

and the ellipse degenerates into a straight line as shown in figure 3.2.3.

In the case where one of the components of the field vector is identically zero, the light is linearly polarized along the other axis.

2) *Circularly polarized light:* If $E_{ox} = E_{oy}$ and the phase difference $\delta = \pm\pi/2$, the light is said to be circularly polarized for the reasons indicated in figure 3.2.4.

With our convention that $\phi = kz - \omega t + \phi_0$, the - sign is associated with right circularly polarized light since, *for fixed z* the tip of the electric field vector describes the motion of a right-handed screw with increasing z, *i.e.*, towards the observer. The $-$ sign is then associated with right-handed polarization. For circularly polarized light we have

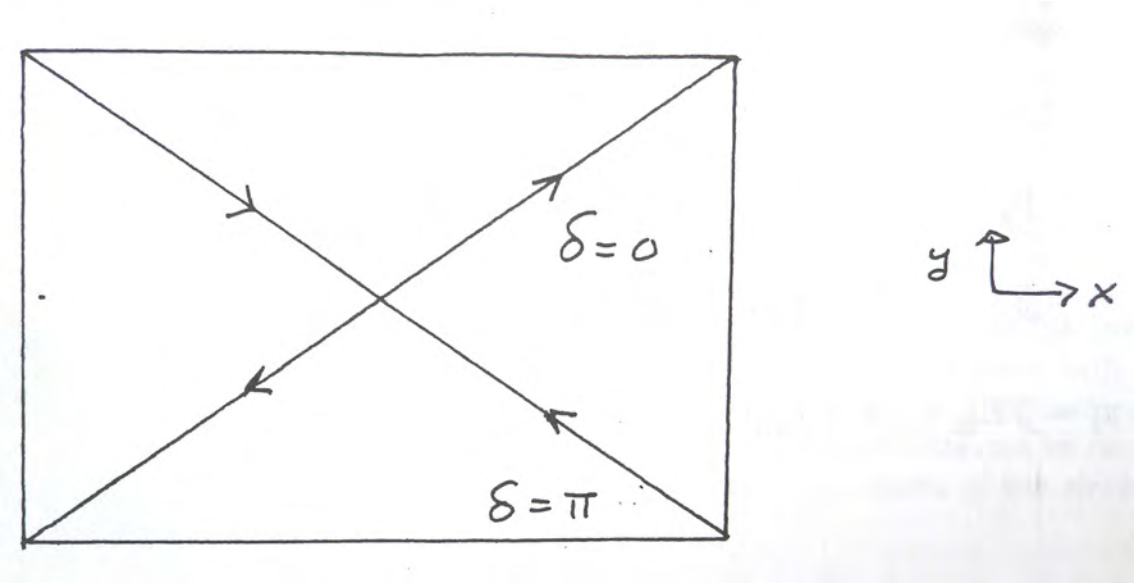


FIGURE 3.2.3. Polarization ellipse corresponding to linearly polarized light.

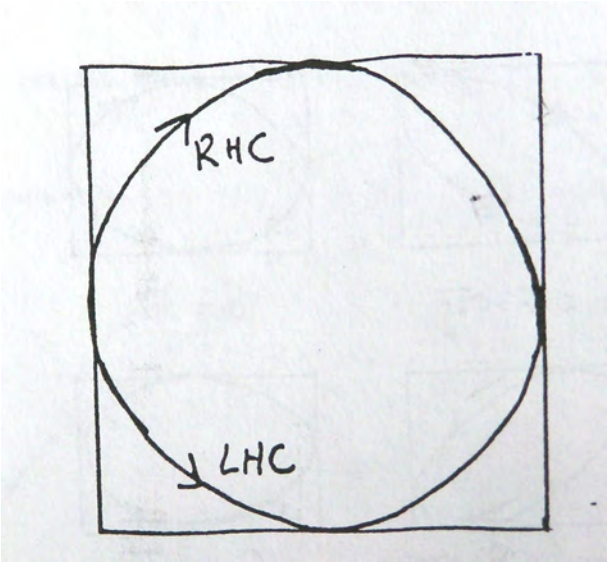


FIGURE 3.2.4. Polarization ellipse corresponding to circularly polarized light.

that

$$\frac{\mathcal{E}_y}{\mathcal{E}_x} = e^{i\delta} = \pm i.$$

Ir can be shown that light with different degrees of polarization carry different amount of *angular momentum,* with the extremes, for a single homogeneous beam, being right and left circularly polarized light.

Unpolarized light is characterized by having a random orientation of the electric field vector. Since a vector always has a specific orientation the randomness can only be defined on a suitable time scale during which the orientation of the vector is indeterminate. That is, there is no definite phase relation between the two electric field components on the time scale of interest.

In many practical situations it is often convenient to characterize the polarization state and light intensity by 4 parameters developed by G. Stokes in 1852 and which are referred to as (surprise, surprise) the *Stokes parameters.* Working from our convention, they are defined by

$$s_0 = E_{0x}^2 + E_{0y}^2 \propto I$$

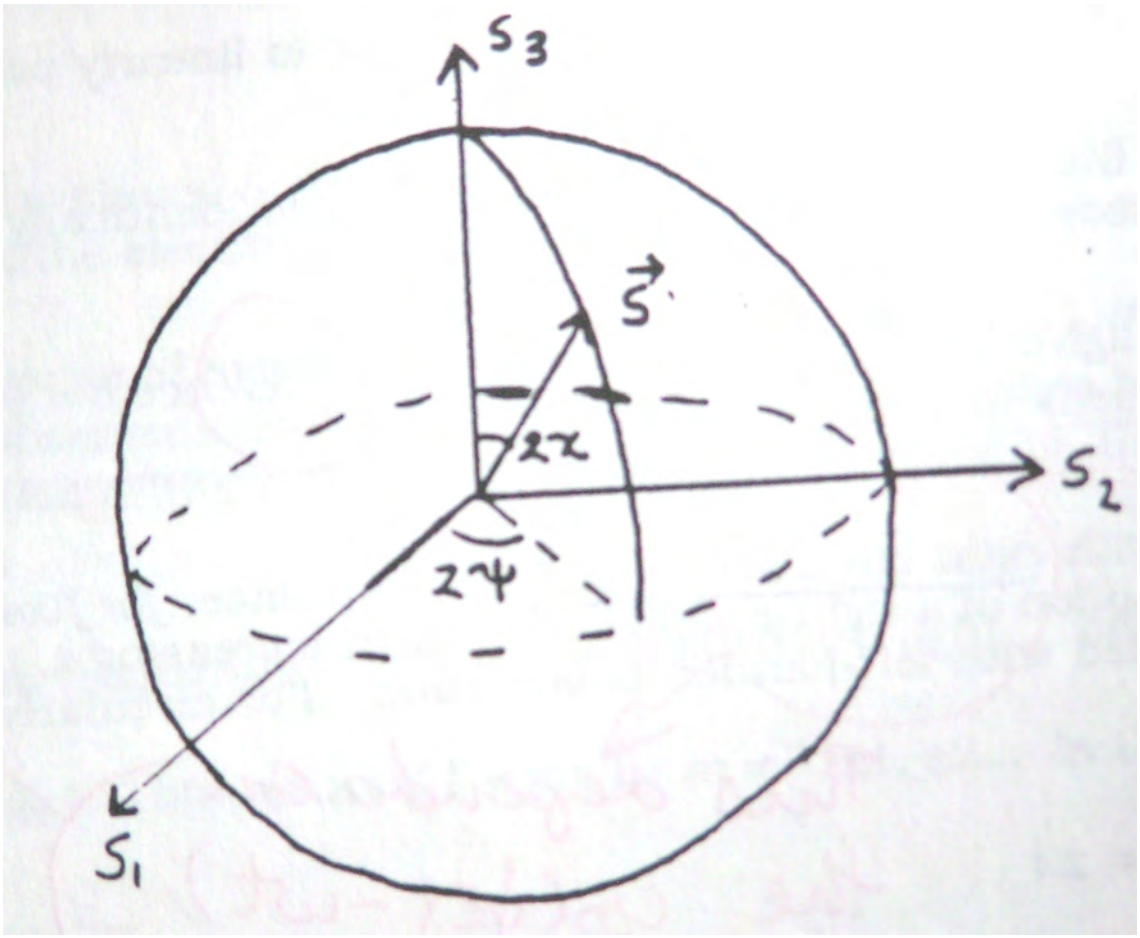


FIGURE 3.2.5. The Poincaré sphere.

$$s_1 = E_{0x}^2 - E_{0y}^2$$
$$s_2 = 2E_{0x}E_{0y}cos(\delta)$$
$$s_3 = 2E_{0x}E_{0y}sin(-\delta)$$

Since $s_0^2 = s_1^2 + s_2^2 + s_3^2$ only three of the components are independent, as was the case in the previous treatment. In particular the s_i can be related to one of the previous triads used to discuss polarization states. The s_i are related to the angle Ψ that characterizes the orientation of the polarization ellipse and an angle φ where

$$tan\varphi = \pm\frac{a}{b} \qquad (0 < \varphi < \pi)$$

and the $\pm$ sign is determined by the sense of rotation of the electric field vector. In terms of these angles we have

$$s_1 = s_0 \sin(2\varphi)\cos(2\Psi)$$
$$s_2 = s_0 \sin(2\varphi)\sin(2\Psi)$$
$$s_3 = s_0 \cos(2\varphi)$$

These equations can be interpreted simply as giving the Cartesian co-ordinates (s_1, s_2, s_3) of a point on a sphere (for s_0 constant) in terms of the spherical polar co-ordinates $(s_0, 2\varphi, 2\psi)$ where 2Ψ is the azimuthal angle and 2φ is the polar angle. This sphere, referred to as the *Poincaré sphere*, is depicted in figure 3.2.5. For every possible polarization state of a plane monochromatic wave of a given intensity there corresponds one and only one point on the Poincaré sphere. Right handed circularly polarized light corresponds to the north pole of the sphere ($\varphi = 0$) while left-handed circularly polarized light is at the south pole ($\varphi = \pi/2$). Linearly polarized light is associated with all those points which lie on the circle defined by the intersection of the plane $s_3 = 0$ with the sphere.

3.3. Anisotropic Optical Media

The importance of the vector character of light is evident when one considers the propagation of light through *anisotropic media* such as crystalline solids. Until now we have only considered isotropic optical media in which the response of the medium to an incident electric field is independent of the direction of the field. Lorentz models for anisotropic media can be constructed by allowing the spring constants to have different values for different directions of the electron motion, but we shall not pursue these here.

In general, the constitutive relations which indicate the relation between the material response, represented by the polarization density, and the applied field is of a tensor nature and the susceptibility is a second rank tensor. In this case

$$\vec{P} = \epsilon_0 \overleftrightarrow{\chi} \cdot \vec{E} \tag{3.3.1}$$

with $\overleftrightarrow{\chi}$ being the susceptibility tensor. The dielectric constant is also therefore a tensor. The tensor nature of the constitutive relation indicates that the magnitude and orientation of the induced polarization density depends on the magnitude and direction of the inducing field.

For reasons related to conservation of energy, the dielectric constant and susceptibility are symmetric tensors, and therefore can be reduced to diagonal form in a principal axis co-ordinate system. There are therefore at most three independent non-zero elements. With respect to principal axes X, Y, Z, the tensor takes the form

$$\overleftrightarrow{\epsilon} = \begin{bmatrix} \epsilon_{XX} & 0 & 0 \\ 0 & \epsilon_{YY} & 0 \\ 0 & 0 & \epsilon_{ZZ} \end{bmatrix} \tag{3.3.2}$$

The corresponding refractive indices are then $n_X = (\epsilon_{XX}/\epsilon_0)^{1/2}$, $n_Y = (\epsilon_{YY}/\epsilon_0)^{1/2}$, $n_Z = (\epsilon_{ZZ}/\epsilon_0)^{1/2}$. If the three diagonal elements of the dielectric tensor are equal, as was considered in chapter 1, the tensor and material are said to be o*ptically isotropic*. If two of the elements are equal the material is said to be *uniaxial* since there is one preferred axis in the system; the preferred axis in such a material is referred to as an *optic axis*. For light propagating along this axis, regardless of the polarization, only one refractive index is experienced. If all diagonal elements are different the material is said to be *biaxial*. With some work it can be shown that the material has two optic axes, so that if light propagates along either of these axes, a unique refractive index is experienced.

The various tensor elements determine, for example, the absorption coefficient, phase velocity of light, etc, for different orientation of the light vector in the medium. For this reason materials which are optically anisotropic are also referred to as being birefringent since, in general, the two independent field components of the light beam experience different refractive indices and different phase velocities. For example, in an anisotropic material light which has its electric field polarized along the X axis of a material propagate with a different phase velocity than light polarized along the Y axis. Light which has its electric vector oriented in an arbitrary direction must be treated as a superposition of beams with component electric field vectors oriented along the principal axes.

The subject of the propagation of light in an anisotropic material for an arbitrary launch angle and initial polarization state is algebraically complex and won't be given here. The underlying physics is still that represented by equations 3.3.1 and 3.3.2 but the mathematics can be cumbersome. Among some of the more interesting results however, is that a light beam launched into a non-principal direction in an anisotropic crystal separate into two light beams that have orthogonal polarization states and propagate in different directions as well. For either beam the surfaces of constant phase are not parallel to the surfaces of constant energy.

In what follows we only consider situations in which light is propagating along one of the principal axes which, without loss of generality, we take to be the $Z = z$ axis. In this case an arbitrarily polarized beam can be considered to be the superposition of beams polarized along the $X = x$ and $Y = y$ axes. The phase shifts suffered by each of these component beams on propagation through a thickness L of the material is given by

$$\phi_x = k_x L + \phi_{0x} = \frac{2\pi n_x}{\lambda_0} L + \phi_{0x}$$

and

$$\phi_y = k_y L + \phi_{0y} = \frac{2\pi n_y}{\lambda_0} L + \phi_{0y} \tag{3.3.3}$$

Depending on the values of the two refractive indices and L, the relative value of ϕ_x and ϕ_y can be anywhere between 0 and 2π so that the polarization state of the wave evolves as it passes through the material. The exact form of the polarization state, of course, also depend on the magnitude of the x and y components of the field. It is also worth remembering that the above treatment also allows for the refractive index to be complex so that absorption

can be treated. Finally, one must remind oneself that the refractive index in general, displays dispersion so that the polarization properties also depend on wavelength, *i.e.*, how close the wavelength is to an absorption feature, etc.

3.4. Matrix Representation of Polarization–The Jones Calculus

In dealing with polarization states and the transformation of polarization states of monochromatic radiation through the interaction of radiation with matter, it is convenient to represent the complex amplitudes of plane harmonic waves by an algebraic expression of the form

$$\vec{\mathcal{E}} = \hat{x}\mathcal{E}_x + \hat{y}\mathcal{E}_y$$

which can also be written in terms of the column vector

$$\begin{bmatrix} \mathcal{E}_x \\ \mathcal{E}_y \end{bmatrix} = \begin{bmatrix} E_{0x}e^{i\phi_x} \\ E_{oy}e^{i\phi_y} \end{bmatrix}$$

This is referred to as the *Jones vector* of the polarization state. The mathematical manipulation of Jones vectors to describe the interaction of light with matter is known as the *Jones calculus*.

In many cases one is only interested in the orientation and not the magnitude of the Jones vector so that a normalized Jones vector is adequate. For example, the vectors

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

refer to linearly polarized states with the field polarized along the x and y axes respectively. Similarly the normalized Jones vectors corresponding to right and left circularly polarized light are given by

$$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ \pm i \end{bmatrix}$$

with the + sign referring to left circularly polarized light.

The superposition of beams of different polarization states leads to new polarization states. These can be easily determined using Jones vectors by vector addition. For example, the superposition of the left and right-handed circularly polarized beams above leads to a linearly polarized beam as seen from

$$\begin{bmatrix} 1 \\ i \end{bmatrix} + \begin{bmatrix} 1 \\ -i \end{bmatrix} = 2\begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Other linear polarization states are possible depending on the relative phase of the two beams. In general the final polarization state obtained by adding a number of Jones vectors, $\vec{V}_i$, is a Jones vector given by

$$\vec{V} = \sum_i \vec{V}_i$$

The real power in using the Jones Calculus becomes evident when one attempts to determine the influence of several optical elements on the polarization state of a beam. As we have seen in the previous section the polarization state in general changes on passage through an optically anisotropic optical element or on reflection or refraction from an interface. For an optically homogeneous and linear optical medium one would expect that a well defined input polarization state yields a well defined output polarization state. Since the transformation of a vector into a vector under these circumstances is accomplished by a matrix one can anticipate that the optical elements can be represented by 2x2 matrices which completely determine the influence of an optical element on an initial polarization state. If

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

represents the initial polarization state before the optical element and

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is the .Jones matrix of the optical element then the final polarization state is given by

$$\begin{bmatrix} \alpha' \\ \beta' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

Optical element	matrix
Linear polarizer: transmission along x axis	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
Linear polarizer :transmission along y-axis	$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$
Linear polarizer: transmission at angle θ with respect to x-axis	$\begin{bmatrix} \cos^2\theta & sin\theta cos\theta \\ sin\theta cos\theta & sin^2\theta \end{bmatrix}$
Isotropic phase retarder	$\begin{bmatrix} e^{i\phi} & 0 \\ 0 & e^{i\phi} \end{bmatrix}$
Relative phase retarder	$\begin{bmatrix} 1 & 0 \\ 0 & e^{i\phi} \end{bmatrix}$
Quarter wave-plate: fast axis along x-axis	$\begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix}$
Quarter wave-plate: fast axis along y-axis	$\begin{bmatrix} 1 & 0 \\ 0 & -i \end{bmatrix}$
Half-wave plate	$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$
Left Circular polarizer: $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ input	$\begin{bmatrix} 1 & 0 \\ 0 & -i \end{bmatrix}$
Right Circular polarizer: $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ input	$\begin{bmatrix} 1 & 0 \\ 0 & -i \end{bmatrix}$
Rotation of an optical element through θ: $\overleftrightarrow{R}\overleftrightarrow{J}\overleftrightarrow{R}^{-1}$	$\overleftrightarrow{R} = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix}$

TABLE 1. Jones Matrices for Some Optical Elements

If light is sent through a series of different optical elements represented by Jones matrices $\overleftrightarrow{J}_1, \overleftrightarrow{J}_2, ...$ the final polarization state is found by the successive transformation (in the order in which light encounters the elements!) of the Jones vectors so that the final polarization state is given by

$$\begin{bmatrix} \alpha' \\ \beta' \end{bmatrix} = \prod_i \overleftrightarrow{J}_i \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

In table 1 we list some of the Jones matrices associated with different optical elements.

The matrices for the linear polarizers need little explanation since it is clear that, independent of the input polarization state, the output polarization state corresponds to the directions indicated in the table. Note that the form of the linear polarizer can be understood as, *e.g.*, the result of anisotropic absorption in the material with the absorption coefficients for light polarized along orthogonal principle axes being very different. Similarly, the form of the matrix for the isotropic phase retarder is easy to understand. This relates to the propagation of an optical beam through an isotropic material. In such a material, both the x and y-components of the electric field vector experience the same dielectric constant and refractive index. If the path length through the material is L and the direction of propagation of the beam is along the z-axis then the phase retardation of the beam for both the x and y- components of the field vector is

$$\phi = \frac{2\pi n}{\lambda_0} L$$

Since this is the same for both components and because the absolute phase of an optical beam cannot be measured, the concept of phase retardation in an isotropic materials is somewhat of an academic point. The output polarization is the same as the input polarization.

The anisotropic phase retarders are derived from optically anisotropic materials. The matrices in the table correspond to materials (usually crystals) which are cut in such a way that two of the principal axes of the dielectric tensor (X and Y, say) lie along the x and y-reference axes respectively of the beam. This can be arranged by cutting the crystal perpendicular to the Z-axis, arranging to have the beam fall on the crystal at normal incidence, and then rotating the crystal about Z. The phase retardation suffered by a beam travelling along the z-direction is as given by equation 3.3.3. If $n_x > n_y$ then because the phase velocity of the beam polarized along the X direction is less than that of the beam polarized along the Y-axis the X-axis is referred to as the *slow axis* and the Y-axis is referred to

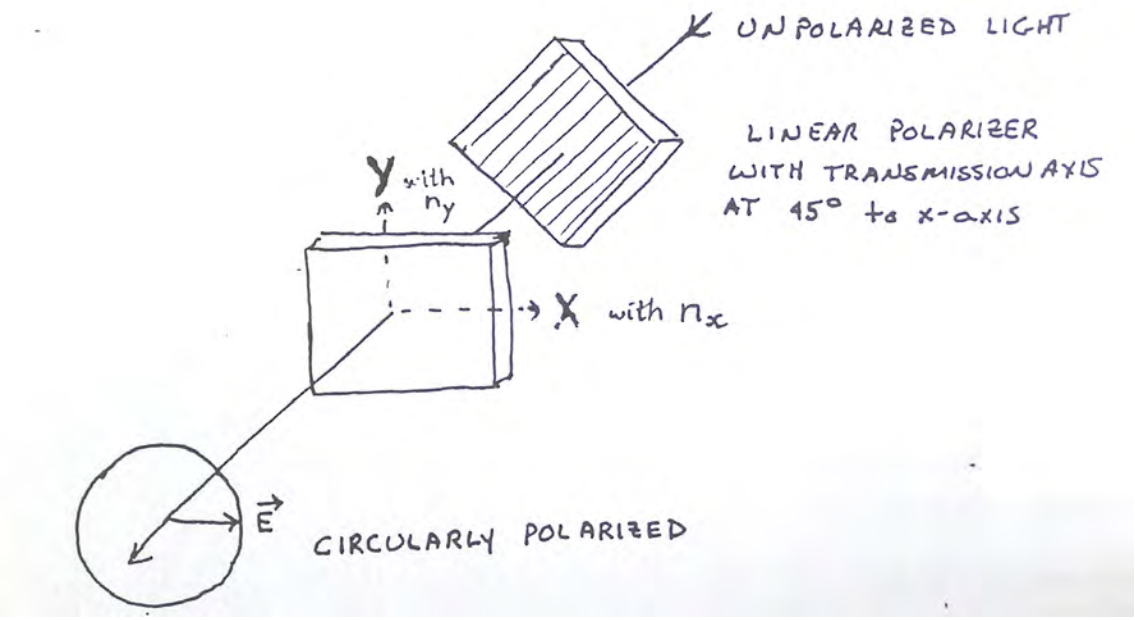


FIGURE 3.4.1. Arrangement for producing circularly polarized light.

as the *fast axis* of the crystal. In the special case where the length of the crystal is chosen such that

$$\delta = \phi_y - \phi_x = \pm\pi/2$$

the crystal is said to act as a *quarter-wave plate*. Such a plate is capable of converting linearly polarized light into circularly polarized light if the linearly polarized beam is incident with its electric field vector oriented at an angle of 45° with respect to the X-axis. The emerging light is right or left circularly polarized depending on whether the Y or the X axis is the fast axis. For a typical birefringent material used for the construction of a quarter wave plate one has $n_x \simeq n_y \simeq 1.5$ with $|n_x - n_y| = 10^{-4}$. Hence, for a wavelength of $1\mu m$ one would have to have a material thickness of approximately 1 mm. The *half-wave plate* is capable of rotating the plane of polarization of linearly polarized light by 90° if incident light again has equal components along the fast and slow axes. The plate induces a relative phase delay of

$$\delta = \phi_y - \phi_x = \pm\pi/2$$

and has a matrix representation obtained from the square of a matrix associated with a quarter wave plate.

The last two elements in the table correspond to a circular polarizer for an polarized beam. The matrices are composites or products of other Jones matrices in the table. The physical arrangement for producing circularly polarized light from an arbitrarily polarized beam is shown in figure 3.4.1.

As depicted in the figure the light is initially passed through a linear polarizer that has its transmission axis oriented at an angle of 45° relative to the fast or slow axis of a quarter wave plate. The matrices corresponding to circular polarizers therefore are products of two matrices, representing a linear polarizer with its axis oriented at 45° to the optical axis of a quarter-wave plate.

Finally, for future reference, we define two polarization states $\vec{V}_1$ and $\vec{V}_2$ to be orthogonal if

$$\vec{V}_1 \cdot \vec{V}_2^* = 0.$$

In terms of the Jones vector components this can be written as

$$\vec{V}_1 \cdot \vec{V}_2^* = \begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix} \cdot \begin{bmatrix} \alpha_2 \\ \beta_2 \end{bmatrix}^* = \alpha_1\alpha_2 + \beta_1\beta_2 = 0.$$

Thus, for example, the Jones vectors corresponding to the linearly polarized states

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

are orthogonal as are the right and left-handed circularly polarized states

$$\begin{bmatrix} 1 \\ -i \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 \\ i \end{bmatrix}.$$

3.5. Optical Activity

It has been known for many years that certain materials, regardless of thickness, possess the unusual property of being able to rotate the plane of polarization of linearly polarized beams passing through them. This phenomenon is known as *optical activity* and is unrelated to, but may occur together with, the phenomenon of birefringence discussed above. Indeed a complete understanding of the phenomenon involves taking account of magnetic field interactions with the electrons of the material as is seen below. The amount of rotation per unit length of travel is called the *specific rotary power*. If the sense of rotation is in the right-handed screw sense, the material is said to be *dextrorotary* and if in the opposite sense it is said to be *levorotary*. Certain types of sugar, turpentine and crystal quartz, among other materials, are known to be optically active and quartz has been found in both forms, depending on its crystal structure. Fused quartz on the other hand is optically isotropic. In general the specific rotary power depends on wavelength but its typical value is rarely more than $10^\circ/mm$.

Optical activity can be explained phenomenologically by assuming that the refractive index experienced by right and left circularly polarized light is different. To show this, let us denote the two refractive indices by n_r k and n_l and construct the Jones vectors corresponding to the right and left circularly polarized beams passing through a length L of optically active material. The Jones vectors (to within a constant phase factor) are given by

$$\begin{bmatrix} 1 \\ -i \end{bmatrix} e^{ik_r L} \quad \text{and} \quad \begin{bmatrix} 1 \\ i \end{bmatrix} e^{ik_l L}$$

where

$$k_r = \frac{2\pi n_r}{\lambda_0} \quad \text{and} \quad k_l = \frac{2\pi n_l}{\lambda_0}$$

Now suppose we have a linearly polarized beam passing through the material with the electric field directed along the x-axis. We can represent this polarization state by

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 \\ -i \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 1 \\ i \end{bmatrix}$$

The complex amplitude of the light wave after passage through a length L is

$$\frac{1}{2}\begin{bmatrix} 1 \\ -i \end{bmatrix} e^{ik_r L} + \frac{1}{2}\begin{bmatrix} 1 \\ i \end{bmatrix} e^{ik_l L} = e^{i\zeta}\left[\frac{1}{2}\begin{bmatrix} 1 \\ -i \end{bmatrix} e^{i\varphi} + \frac{1}{2}\begin{bmatrix} 1 \\ i \end{bmatrix} e^{-i\varphi}\right]$$

where

$$\zeta = \frac{L}{2}(k_r + k_l)$$

and

$$\varphi = \frac{L}{2}(k_l - k_r).$$

The complex amplitude can then be represented as

$$e^{i\zeta}\begin{bmatrix} \frac{1}{2}\left(e^{i\varphi} + e^{-i\varphi}\right) \\ \frac{1}{2i}\left(e^{i\varphi} - e^{-i\varphi}\right) \end{bmatrix} = e^{i\zeta}\begin{bmatrix} cos\varphi \\ sin\varphi \end{bmatrix}.$$

This represents a linearly polarized beam in which the plane of polarization has been rotated through an angle φ with

$$\varphi = \frac{\pi L(n_l - n_r)}{\lambda_0}. \tag{3.5.1}$$

It follows that the specific rotary power is given by

$$\xi = \varphi/L = \frac{\pi(n_l - n_r)}{\lambda_0}.$$

For right-handed quartz, in the mid-visible region of the spectrum, the values of the two refractive indices are

$$n_r = 1.5442 \quad \text{and} \quad \mathrm{n_l} = 1.54427$$

and the specific rotary power is $\approx 2^\circ/mm$.

We now show that the phenomenon of optical activity can be explained by considering the effect of a static magnetic field, $\vec{B}_0$, on the optical properties of a medium. (This does not involve the magnetic field of the light wave so the phenomenon cannot be explained on the basis of magnetic permeability effects. Rather, as will be seen, it should be thought of as an induced effect on the dielectric constant just as strain, pressure, etc, may also be expected

to change electronic configurations and hence dielectric constants as well.) In a classical model the equation of motion of the electron in the absence of damping is given by

$$m\frac{d^2\vec{r}}{dt^2} + K\vec{r} = -e\vec{E}(t) - e\left(\frac{d\vec{r}}{dt}\right) \times \vec{B}_0$$

Assuming a harmonic time dependence of the field $\propto e^{-i\omega t}$, the steady state solution of this equation is found from

$$-m\omega^2\vec{r} + K\vec{r} = -e\vec{E} + i\omega e\vec{r} \times \vec{B}_0$$

The induced polarization density, $\vec{P} = -Ne\vec{r}$, can be found from

$$(-m\omega^2 + K)\vec{P} = Ne^2\vec{E} + i\omega e\vec{P} \times \vec{B}_0$$

and this equation can be placed in the form

$$\vec{P} = \epsilon_0 \overleftrightarrow{\chi}_B \cdot \vec{E}$$

Here $\overleftrightarrow{\chi}_B$ is the effective susceptibility tensor which, for the magnetic field along the z direction, takes the form

$$\overleftrightarrow{\chi}_B = \begin{bmatrix} \chi_{xx} & i\chi_{xy} & 0 \\ -i\chi_{xy} & \chi_{xx} & 0 \\ 0 & 0 & \chi_{zz} \end{bmatrix} \tag{3.5.2}$$

where

$$\chi_{xx} = \frac{Ne^2}{m\epsilon_0}\frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 - \omega^2\omega_c^2}$$

$$\chi_{zz} = \frac{Ne^2}{m\epsilon_0}\frac{1}{(\omega_0^2 - \omega^2)}$$

$$\chi_{xy} = \frac{Ne^2}{m\epsilon_0}\frac{\omega\omega_c}{(\omega_0^2 - \omega^2)^2 - \omega^2\omega_c^2} \tag{3.5.3}$$

and where, as before, $\omega_0 = \sqrt{K/m}$ is the natural resonance frequency, and $\omega_c = eB_0/m$ is the cyclotron frequency of the electron.

To show that the susceptibility of equation 3.5.2 corresponds to an optically active medium consider a monochromatic wave propagating in the z direction in this material. The wave equation for the x and y components of the field becomes

$$-k^2\mathcal{E}_x + \frac{\omega^2}{c^2}\mathcal{E}_x = -\frac{\omega^2}{c^2}(\chi_{xx}\mathcal{E}_x + i\chi_{xy}\mathcal{E}_x)$$

$$-k^2\mathcal{E}_y + \frac{\omega^2}{c^2}\mathcal{E}_y = -\frac{\omega^2}{c^2}(-i\chi_{xy}\mathcal{E}_x + \chi_{xx}\mathcal{E}_y).$$

A nontrivial solution for the fields can be found if the determinant of the coefficients of the field components vanish. One then finds that two propagation constants are allowed with

$$k = \frac{\omega}{c}\sqrt{1 + \chi_{xx} \pm \chi_{xy}}$$

for which the nontrivial solutions are of the form

$$\mathcal{E}_x = \pm i\mathcal{E}_y.$$

We see therefore that the right and left circularly polarized states are polarization eigenstates of the material with refractive indices of

$$n_r = \sqrt{1 + \chi_{xx} + \chi_{xy}}$$

$$n_l = \sqrt{1 + \chi_{xx} - \chi_{xy}}.$$

Since $\chi_{xy} << \chi_{xx}$, the difference between the two refractive indices can be approximated as

$$n_r - n_l = \frac{\chi_{xy}}{\sqrt{1 + \chi_{xx}}} \approx \frac{\chi_{xy}}{n_r} \tag{3.5.4}$$

and the specific rotary power is

$$\delta = \frac{\pi\chi_{xy}}{\lambda_0 n_r}.$$

One sees then that the rotary power is directly proportional to the value of the static magnetic field. In solids like quartz the origin of this field is thought to be due to the spiral nature of the bonding. The electrons in moving through the bonds therefore establish current loops which give rise to the magnetic field. Note from equation 3.5.3

that an optically active medium, if not birefringent for purely dielectric reasons, is usually birefringent for reasons of its optical activity. This latter effect, with the exception of a few materials, is usually small however.

3.6. Magneto- and Electro-optic Effects

Optical activity and anisotropy can be induced in materials with the application of dc magnetic and electric fields. These are known as magneto-optic and electro-optic effects respectively. In this section we summarize the effects that can occur and point out their significance.

i) Faraday Rotation in Solids. Optical activity can be induced in isotropic dielectrics by placing them in a magnetic field. This phenomenon was discovered by Michael Faraday in 1845 and offered a clue to the origin of natural optical activity. When a beam of linearly polarized light is passed through the medium in the direction of the field the plane of polarization of the light beam is rotated by an amount which depends on the strength of the magnetic field, and the thickness of material traversed. The amount of rotation of the plane can therefore be expressed as

$$\varphi = V B_0 L \tag{3.6.1}$$

a result which is of the same form as equation 3.5.1 when equation 3.5.4 is considered. The constant V is known as the *Verdet constant* and for solids such as glass has a value of about 10 radians/Tesla/m in the visible region of the spectrum.

The *Faraday effect* is useful in the development of optical isolators. As a consideration of equation 3.6.1 shows, the plane of polarization is rotated in the same direction in fixed space, independent of the propagation direction of the light. One can inhibit light which leaves a source from reentering that source (due to back reflections) by using a linear polarizer and a Faraday cell in tandem. If the magnetic field is adjusted so as to give a rotation of the plane of incidence of $\pi/4$ in one pass, light re-entering the cell is not able to pass through the polarizer again.

ii) Voigt Effect. It was noted earlier that substances which are optically active are also birefringent as equation 3.5.3 shows. This effect is usually small except when the frequency of the light wave is near a resonance of the material. Dielectric materials, like atomic and molecular gases, can be made birefringent near a resonance frequency by applying a magnetic field. This phenomenon is termed the Voigt effect.

iii) Pockels Effect. When a dc electric field is applied to dielectric materials which do not possess a centre of inversion a linear change in their dielectric tensor elements can occur. This is known as the *Pockels effect*. The application of the field breaks the isotropy of space and renders an isotropic crystal anisotropic. For an already anisotropic crystal the application of the field simply alters the dielectric tensor elements. In either case, the changes are small for fields which can be produced in the laboratory. It is easy to show that the changes in the refractive indices along the three principal directions also vary linearly with the field. The exact amount of the change in a specific index depends on the orientation of the field as well as its magnitude. In some materials like potassium dihydrogen phosphate (KDP) the change is as high as 10^{-3} for an applied field of 10 kV/cm. See Fig. 3.6.1

iv) Kerr Effect. When an optically isotropic material (which does possess a centre of inversion) is placed in a dc electric field it can be made birefringent as well and behaves like a uniaxial crystal with the electric field defining the optic axis. The magnitude of the change in the index is proportional to the *square* of the applied field in this phenomenon which is known as the *Kerr effect*. The effect is attributed to the alignment of molecules in the presence of the field. For an applied field along the z direction the difference between refractive indices for light polarized parallel and perpendicular to the field is given by

$$n_z - n_x = n_z - n_y = \lambda_0 K E_{dc}^2$$

where K is known as the Kerr constant. The liquids CS_2 and nitrobenzene have particularly large Kerr constants with values of 3.5×10^{-10} and $4.4 \times 10^{-8} m V^{-2}$ respectively.

The Kerr and Pockels effects are very important in modern optical technology since they can be used to modulate a light beam. Figure 3.6.2 illustrates the use of a Kerr cell in this application. When no voltage is applied to the Kerr cell it behaves like an isotropic material. Light is not able to pass through the system which otherwise consists of two polarizers whose transmission axes are orthogonal to each other (crossed polarizers). When a voltage is applied to the Kerr cell, the light beam can develop a component of its electric field vector parallel to the transmission axis of the analyzer polarizer and some light passes through. Modulating the voltage applied to the Kerr cell therefore leads to a modulated light beam. If a voltage pulse is applied to the Kerr cell, a light pulse can be extracted from an otherwise continuous beam. The limit on the response time is determined by how fast the Kerr cell can be turned on or off. This is limited by molecular reorientation times but can be as small as a few picoseconds ($10^{-12} s$). Although the illustration is given for a Kerr cell modulator, it also works with a Pockels cell and indeed this is usually the

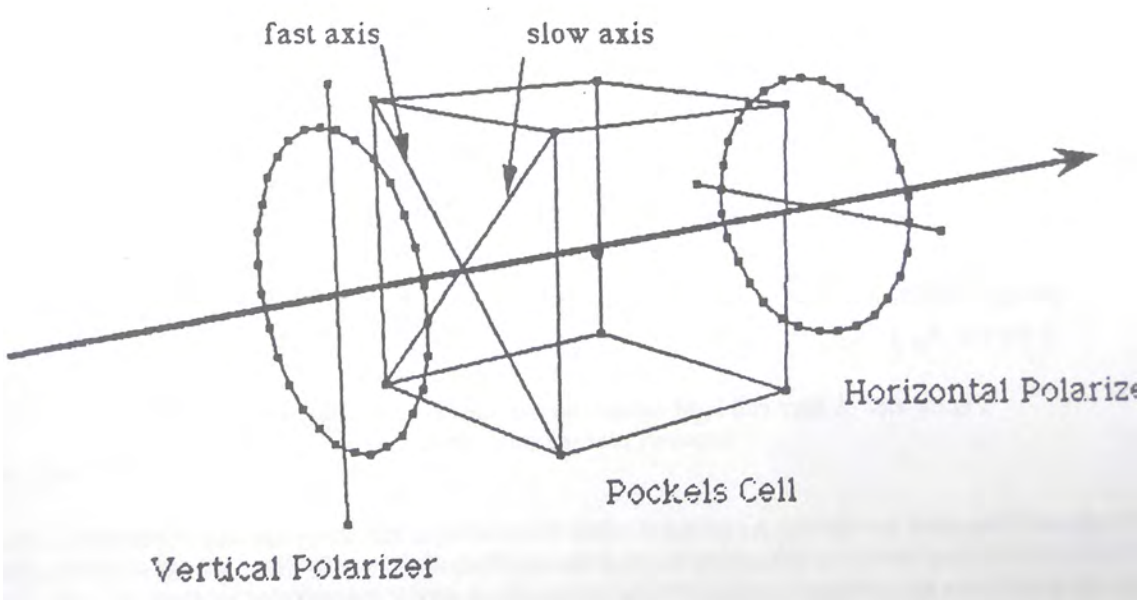


FIGURE 3.6.1. Pockels cell modulator with a Pockels crystal between crossed polarizers at a 45° orientation, and the correct applied voltage, the Pockels cell acts as a half-wave plate. Thus it converts vertically polarized light to horizontally polarized light, and passes it through the crossed polarizers.

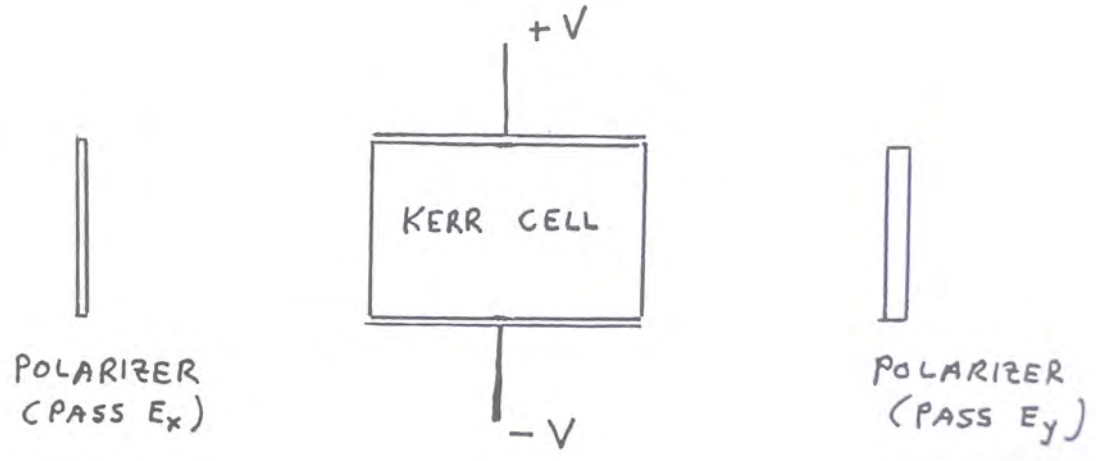


FIGURE 3.6.2. A Kerr cell light modulator; an electro-optic Kerr cell placed between crossed polarizers.

preferred arrangement since lower voltages are required in a number of suitable materials to induce the same effect as in "Kerr liquids".

Because light intensity depends on the square of the E-field, light itself can induce the Kerr effect in a material with no externally applied static field. This is called the *optical Kerr effect*; in this way an intense light pulse may modify its own polarization, or that of a second beam of light.

v) The Cotton-Mouton Effect; This effect is the magnetic analogue of the Kerr electro-optic effect and is attributed to the lining up of molecules in a magnetic field. It is of little more than academic interest since in most materials large magnetic fields are required to see even a small effect.

References

M. Born and E. Wolf, *Principles of Optics*, Cambridge Press, 2002.

P. Lorrain and D. Corson, *Electromagnetic Fields and Waves*, W.H. Freeman, San Francisco, 1970.
H. A. Haus, *Waves and Fields in Optoelectronics*, Prentice-Hall, Englewood Cliffs NJ, 1984.
G.R. Fowles, *Introduction to Modern Optics*, Holt, Rinehart and Winston, New York, 1988.
E. Hecht, *Optics*, Pearson, New York, 2002.

Problems

1. Construct the complex vector expression for the electric field for a right circularly polarized plane wave propagating in free space in the +z direction with a peak amplitude E_0 occurring at $z = 0$, $t = 0$ along the x direction. Derive an expression for the associated complex magnetic field and the complex Poynting vector.

2. a) Describe the polarization state of the waves whose Jones vectors are the following:

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 2i \end{bmatrix} \quad \begin{bmatrix} 1 \\ 1+i \end{bmatrix}$$

b) Give the Jones vectors orthogonal to each of the above and describe their polarizations.

3. Show that the wave represented by the Jones vector

$$\begin{bmatrix} A \\ Be^{i\phi} \end{bmatrix}$$

is, in general, elliptically polarized and that the major axis of the ellipse makes an angle

$$\frac{1}{2}\arctan\left[\frac{2AB\cos\phi}{A^2-B^2}\right]$$

with the x-axis.

4. a) A wire grid polarizer is an optical element which consists of a number of long, thin (compared to λ), parallel strips of metal mounted on a transparent substrate. It was first invented and used by Heinrich Hertz in 1887, working at a wavelength of 3 m, with the wires 5 cm apart. Subsequent developments have allowed these polarizers to be used into the optical domain. With unpolarized light incident on the grid, the output polarization state is, to a high degree, linearly polarized. Explain how this polarizer works and deduce the orientation of the electric field relative to the orientation of the wires.

b) How do you think Polaroid sheets polarize light?

5. How does one experimentally distinguish between right and left circularly polarized light? When might it be important to do so?

6. Consider linearly polarized light falling on a half wave plate with the direction of polarization along either the fast or slow axis. Show that if the plate is rotated through an angle θ then the plane of polarization is rotated through an angle 2θ.

7. Consider a set of imperfect quarter-wave plates each of which has the Jones matrix

$$\begin{bmatrix} 1 & \varepsilon \\ \varepsilon & i \end{bmatrix}$$

with $\varepsilon = 0.001$. For a series of ten such plates determine the final polarization state if the input beam is linearly polarized at an angle of 45° relative to the x-axis.

8. Is it possible to associate a Jones matrix with an optically active element? If so, determine the matrix in terms of the parameters associated with optical activity.

9. For CS_2 calculate the voltage that would be required to turn a 1 cm thickness of liquid into a half wave plate for $0.5\mu m$ light.

Special Note on Phase Conventions. The total phase ϕ is written here as

$$\phi(\vec{r},t) = \vec{k}\cdot\vec{r} - \omega t + \phi_0.$$

It is also possible to use the representation

$$\phi(\vec{r},t) = -\vec{k}\cdot\vec{r} + \omega t + \phi_0$$

with a phasor defined with

$$\vec{\mathcal{E}} \propto e^{-i\phi}$$

This effectively reverses the sign of the phase –as time passes (t increases) the total phase is increasing. This causes differences, particularly in the definition of the phase difference between x- and y-components, and sometimes in the mathematical definition of right and left-handed circular polarization. Be careful when switching between conventions.

CHAPTER 4

Reflection and Refraction at an Interface

But soft! What light through yonder window breaks?
William Shakespeare

In the previous chapters we outlined the propagation of optical beams in homogeneous media. In this chapter we consider the interaction of an optical beam with an optical interface, which is defined as any surface which presents a discontinuity in the dielectric constant. Initially we consider only a planar or flat interface. At such a surface an incident beam is split into a single reflected wave and a single transmitted or refracted wave. We derive the law of reflection and Snell's law of refraction which will be seen to be consequences of the translational invariance of the interface. Through the application of boundary conditions for electromagnetic waves we derive expressions for the amplitude of the reflected and refracted beams in the case of an incident plane wave. We also consider the phenomenon of total internal reflection which occurs for certain angles of incidence for a beam incident from the more dense medium. Finally, we consider reflection properties only of non-planar or optically "rough" interfaces and present some semi-quantitative results for the scattered fields which arise when a plane wave is incident on such an interface. This subject is discussed in more detail when we consider diffraction phenomena later but some of the salient features can be derived here as a motivation for diffraction theory. In all these discussions it is understood that the width of the beam is much larger than its wavelength so that plane waves can be used in the analysis.

4.1. Reflection and Refraction of a Plane Wave at a Planar Interface

We now consider the simple problem of a plane wave which is incident on a plane interface. At the interface the incident beam generally splits into a reflected beam and a refracted beam as shown in figure 4.1.1.

The goal of this section is to establish the relation between the direction of propagation of the incident wave and the two other waves. To describe the beams we choose a Cartesian co-ordinate system with the x- and y-axes in the plane of the interface so that $z = 0$ defines the interface. The medium from which the beam is incident (medium 1) is defined by $z < 0$ while $z > 0$ is the half-space containing medium 2. We assume that the dielectric constant of medium 1 is real while that of medium 2 can be complex. Let the beam be incident at an angle θ_1 with respect to a

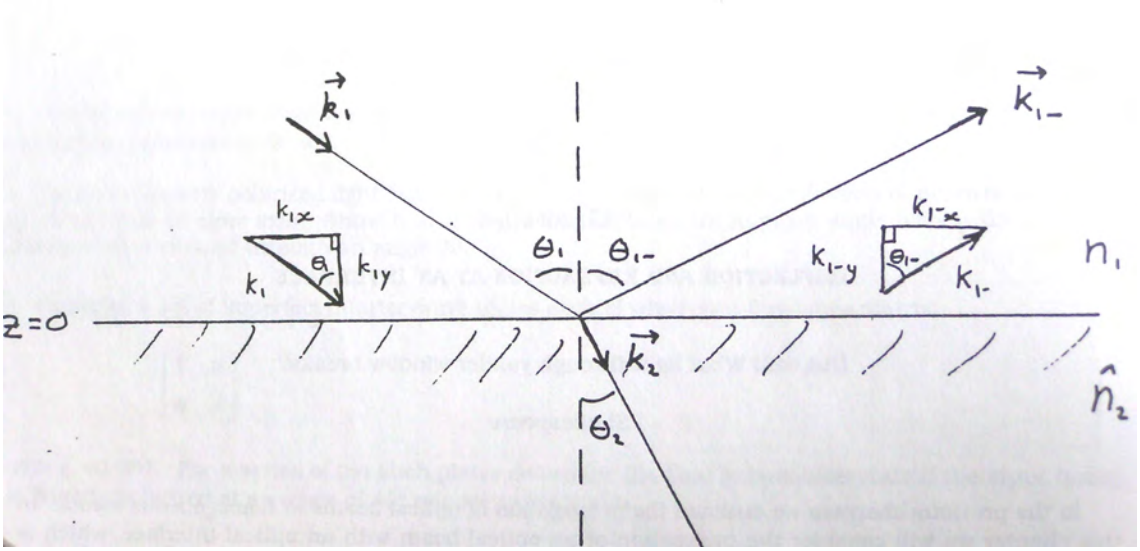


FIGURE 4.1.1. Propagation vectors associated with incident, reflected and refracted waves at a planar interface. Note: the "y" in Fig. should be "z".

normal axis so that for a propagation vector $\vec{k}_1$ we have

$$k_{1x} = k_1 \sin\theta_1 \quad \text{and} \quad \mathrm{k_{1z}} = \mathrm{k_1} \cos\theta_1$$

At the boundary between the two media the time and spatial variation of the secondary fields must be the same as for the primary wave, or

$$\vec{k}_1 \cdot \vec{r} - \omega t = \vec{k}_{1-} \cdot \vec{r} - \omega t = \vec{k}_2 \cdot \vec{r} - \omega t$$

for z = 0. Hence

$$k_{1x} x = k_{(1-)x} x = k_{2x} x \tag{4.1.1}$$

Thus $k_{1x} = k_{(1^-)x}$. Because the incident and reflected wave travel in the same medium with the same frequency we have $k_1 = k_{1^-}$. The triangles made by resolving each $\vec{k}$ into (k_x, k_z) have two sides and a right angle in common and so the triangles are congruent. Thus

$$\theta_{1-} = \theta_1$$

which is the law of .specular reflection for a planar interface: the angle of reflection is equal to the angle of incidence. This *law of specular reflection* was first discovered by *Hero of Alexandria* in antiquity. The usual convention is to measure these angles from the normal to the interface. The second equality of equation 4.1.1 gives

$$k_1 \sin\theta_1 = k_2 \sin\theta_2. \tag{4.1.2}$$

However, since

$$k_1 = \frac{2\pi}{\lambda_0} n_1 \quad \text{and} \quad \mathrm{k_2} = \frac{2\pi}{\lambda_0} \hat{\mathrm{n}}_2$$

we have

$$n_1 sin\theta_1 = \hat{n}_2 sin\theta_2. \tag{4.1.3}$$

If $\hat{n}_2$ is purely real then θ_2 is real and equation 4.1.3 is referred to as *Snell's law* for dielectrics. One then has a plane wave which is propagating in a new direction in medium 2.

If $\hat{n}_2$ is complex (as would occur for an absorbing medium) then θ_2 is also complex and it no longer has the simple interpretation of an angle of refraction. Instead we have that

$$\sin\theta_2 = \frac{n_1 \sin\theta_1}{n_2 + i\kappa_2} = \frac{n_1(n_2 - i\kappa_2)}{n_2^2 + \kappa_2^2} \sin\theta_1$$

and

$$\cos\theta_2 = \sqrt{1 - \frac{n_1^2(n_2^2 - \kappa_2^2)}{\left(n_2^2 + \kappa_2^2\right)^2} \sin^2\theta_1 + i \frac{2n_1^2 n_2 \kappa_2}{\left(n_2^2 + \kappa_2^2\right)^2} \sin^2\theta_1}$$

which is of the form

$$\cos\theta_2 = p e^{iq}$$

for p and q real. The values of p and q are easily found by comparison of the last two equations from which we have

$$p^2 \cos 2q = 1 - \frac{n_1^2(n_2^2 - \kappa_2^2)}{\left(n_2^2 + \kappa_2^2\right)^2} \sin^2\theta_1$$

$$p^2 \sin 2q = \frac{2n_1^2 n_2 \kappa_2}{(n_2^2 + \kappa_2^2)^2} \sin^2\theta_1$$

The spatial variation of the phase is therefore given by

$$\vec{k}_2 \cdot \vec{r} = \frac{\omega}{c}(n_2 + i\kappa_2)(x \sin\theta_2 + z cos\theta_2)$$

$$= \frac{\omega}{c}(n_2 + i\kappa_2)\left[x \frac{n_1(n_2 - i\kappa_2)}{n_2^2 + \kappa_2^2} \sin\theta_1 + z(p\cos q + ip \sin q\right]$$

$$= \frac{\omega}{c}\left[x n_1 \sin\theta_1 + zp\left(n_2 \cos q - \kappa_2 \sin q\right) + izp\left(\kappa_2 cosq + n_2 \sin q\right)\right]$$

From this last expression we see that the *surfaces of constant amplitude* are of the form

$$z = constant$$

and are parallel to the interface, while the *surfaces of constant phase* are given by

$$x n_1 \sin\theta_1 + zp\left(n_2 \cos q - \kappa_2 \sin q\right) = constant'$$

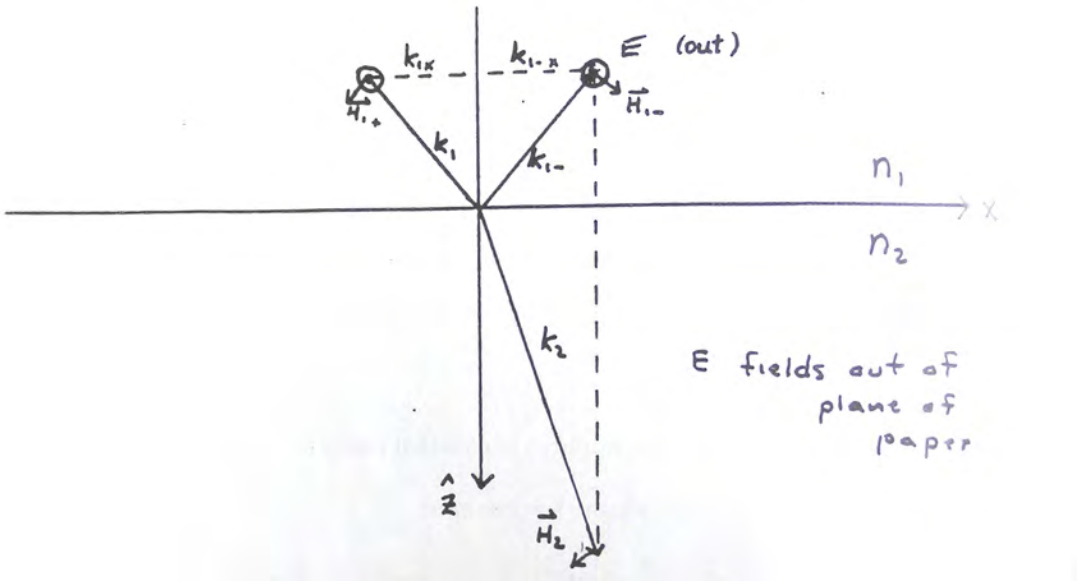


FIGURE 4.2.1. Reflection and transmission of a s-polarized beam at an interface.

Because the surfaces of constant phase and the surfaces of constant amplitude do not coincide, the wave is called an *inhomogeneous wave*.

Equations 4.1.2 and 4.1.3 have a simple interpretation if one recalls from Chapter 2 that in a quantum mechanical picture the wave vector $\vec{k}$ is associated with the momentum of the light wave. Linear momentum of course is conserved in systems which have translational symmetry, *i.e.* those which remain invariant with respect to a linear translation. In our case, the system possesses translational symmetry in the x and y directions but not in the z- direction. As a result one expects that momentum is conserved in the x and y directions but not in the z-direction for which there is an abrupt change in optical properties at $z = 0$. Equations 4.1.2 and 4.1.3 are simply statements of the fact that the x-component of the momentum is conserved at the planar interface. The laws of optical specular reflection and refraction are therefore simple consequences of this symmetry.

4.2. Amplitudes of Reflected, Transmitted and Refracted Waves

The determination of the amplitude of the reflected and transmitted electromagnetic waves at an optical interface is a straightforward application of the boundary conditions associated with the electric and magnetic fields. Because the boundary conditions are different for the electric field components polarized in the plane, and perpendicular to the plane, of the interface the reflectivity and transmissivity of the beam is, in general, different for the two components of an arbitrarily polarized beam. If an optical beam has its electric field vector polarized parallel to the interface between two media the beam is said to be a *transverse electric* (TE) wave. Optical physicists refer to such a wave as being *s-polarized*. The s is derived from the German word "senkrecht" which means perpendicular. The electric field in this case is perpendicular to the plane of incidence, which is the plane defined by the propagation vectors of the incident and reflected waves. If the optical beam is polarized with the magnetic field in the plane of the interface the beam is said to be a *transverse magnetic* (TM) wave or a *p-polarized* beam. An arbitrarily polarized beam can be written as a superposition of an s- and p-polarized wave as we saw in the previous chapter. We now treat each case in turn.

The magnitude of the fields associated with the reflected and refracted waves can be obtained by applying the condition that the tangential component of the electric and magnetic field vectors is continuous across the interface. Consider first an s-polarized wave incident on the interface as depicted in figure 4.2.1.

For an incident, monochromatic TE wave the electric field vector can be written as

$$\vec{\mathcal{E}}^{1+} = \hat{y}\vec{\mathcal{E}}_0^{1+} e^{i\vec{k}\cdot\vec{r}} e^{-i\omega t}$$

while the reflected wave is written

$$\vec{\mathcal{E}}^{1-} = \hat{y}\vec{\mathcal{E}}_0^{1-} e^{i\vec{k}_{1-}\cdot\vec{r}} e^{-i\omega t}.$$

The refracted wave is designated as

$$\vec{\mathcal{E}}^{2+} = \hat{y}\vec{\mathcal{E}}_0^{2+} e^{i\vec{k}_2\cdot\vec{r}} e^{-i\omega t}.$$

In medium 1 the total field is the superposition of the fields for the incident and reflected waves. The total electric field has a y-component only, given by

$$\mathcal{E}_y^1 = \left[\mathcal{E}_0^{1+} e^{ik_{1z}z} + \mathcal{E}_0^{1-} e^{-ik_{1z}z}\right] e^{ik_{1x}x}$$

The total magnetic field in medium 1 can be obtained from

$$\vec{k} \times \vec{\mathcal{E}}^1 = \omega\mu_0 \vec{\mathcal{H}}^1$$

with an x-component given by

$$\mathcal{H}_x^1 = -\frac{k_{1z}}{\omega\mu_0}\left[\mathcal{E}_0^{1+} e^{ik_{1z}z} - \mathcal{E}_0^{1-} e^{-ik_{1z}z}\right] e^{ik_{1x}x}.$$

We define the coefficient of the square bracket in this expression to be the *characteristic admittance*, Y_0^1, given by

$$Y_0^{(1)} = \frac{k_{1z}}{\omega\mu_0} = \left(\frac{\epsilon_0}{\mu_0}\right)^{1/2} n_1 \cos\theta_1.$$

In particular this is the characteristic admittance presented to a TE wave leaving medium 1 at an angle of inclination θ_1 with respect to the interface. The *characteristic impedance* of medium 1, Z_0^1, is given by the reciprocal of the characteristic admittance, or

$$Z_0^{(1)} = \left(Y_0^{(1)}\right)^{-1}$$

In a similar fashion one can obtain an expression for the characteristic admittance and impedance of medium 2.

Continuity of the tangential component of the $\vec{E}$ and $\vec{H}$ fields requires that the ratio

$$Z = -\frac{\mathcal{E}_y}{\mathcal{H}_x}$$

known as the wave impedance, be continuous. We therefore have that at the interface:

$$Z^{(2)} = -\frac{\mathcal{E}_y^2}{\mathcal{H}_x^2} = Z^{(1)} = -\frac{\mathcal{E}_y^1}{\mathcal{H}_x^1}$$

at $z = 0$. This gives us

$$Z_0^{(1)}\left[\frac{\mathcal{E}_0^{1+} + \mathcal{E}_0^{1-}}{\mathcal{E}_0^{1+} - \mathcal{E}_0^{1-}}\right] = Z_0^{(2)}.$$

Defining $\Gamma = re^{i\psi} = \mathcal{E}_0^{1-}/\mathcal{E}_0^{1+}$ to be the *amplitude reflection coefficient* of the wave, we obtain

$$\Gamma = \mathcal{E}_0^{1-}/\mathcal{E}_0^{1+} = \frac{Z_0^2 - Z_0^1}{Z_0^2 + Z_0^1}$$

In the case where the refractive index is real all the terms in this equation are real and apart from a possible sign change there is no phase component in Γ. If the refractive index is complex then Γ is complex and there is a phase change, ψ, on reflection. In terms of the complex refractive indices one obtains for the amplitude reflection coefficient,

$$\Gamma = \frac{n_1\cos\theta_1 - \hat{n}_2\left(1 - \frac{\epsilon_1 \sin^2\theta_1}{\hat{\epsilon}_2}\right)^{1/2}}{n_1\cos\theta_1 + \hat{n}_2\left(1 - \frac{\epsilon_1 \sin^2\theta_1}{\hat{\epsilon}_2}\right)^{1/2}} \tag{4.2.1}$$

For $\theta_1 = 0$ (normal incidence) the reflectivity is simply given by

$$\Gamma = \frac{n_1 - \hat{n}_2}{n_1 + \hat{n}_2}$$

The *energy reflectivity*, R, which is the ratio of the energies or intensities of the reflected beam to the incident beam, is defined by

$$R = |\Gamma|^2$$

Although the reflectivity expression for an interface bordering a nondielectric, or material with a complex refractive index, is more complicated than in the case of having both media dielectrics, it is worth pointing out, that at normal incidence for medium 1 air ($n_1 \simeq 1$) and medium 2 a non-dielectric, one obtains the simple expression

$$R = \frac{(n_2 - 1)^2 + \kappa_2^2}{(n_2 + 1)^2 + \kappa_2^2}$$

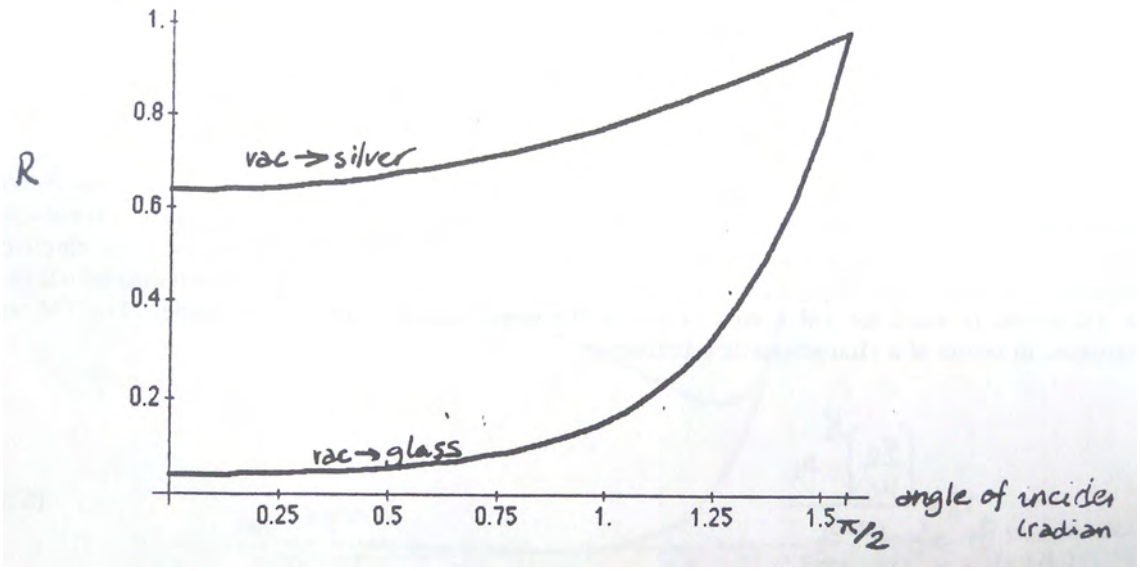


FIGURE 4.2.2. Energy reflectivity of a TE wave in the case of a glass:vacuum and silver:vacuum interface.

Figure 4.2.2 shows a typical variation with angle of incidence of the energy reflectivity for a light wave of wavelength $0.5\mu m$ incident from vacuum onto a dielectric (glass) and a metal (silver). Note that in both cases the reflectivity for the TE wave rises uniformly with angle of incidence. In the case of glass, which is transparent at wavelengths in the mid-visible, it is worth reminding oneself that the reflectivity is that of one interface only (a semi-infinite piece of glass). For a glass slide, where two interfaces are present, the total reflectivity is different than the values in the graph (depending on thickness and wavelength). At small angles of incidence where the reflectivity is small, and neglecting interference effect, the total reflectivity is approximately twice the single interface value. For materials with purely real dielectric constant, the phase change on reflection is π radians, independent of the angle of incidence.

One can also define an amplitude transmission coefficient $\Phi = te^{i\varsigma}$ through

$$\Phi = \mathcal{E}_0^{2+}/\mathcal{E}_0^{1+}$$

where t and ς represent the amplitude and phase of the complex transmissivity. Just beyond the interface the transmissivity is given by

$$\Phi = \frac{\mathcal{E}_0^{1+} + \mathcal{E}_0^{1-}}{\mathcal{E}_0^{1+}}$$

which after a bit of algebra can be reduced to

$$\Phi = \frac{2n_1 \cos\theta_1}{n_1 \cos\theta_1 + \hat{n}_2 \left(1 - \frac{\sin^2\theta_1 \epsilon_1}{\hat{\epsilon}_2}\right)^{1/2}}.$$

The *energy transmissivity* in the case where $\hat{n}_2$ is real is given by

$$T = \frac{n_2 cos\theta_2}{n_1 cos\theta_1} |\Phi|^2$$

where the factor in front of $|\Phi|^2$ accounts for energy conservation across the interface (see problem 4.4).

If both media are dielectrics it is easy to show that

$$R + T = 1$$

which is merely a statement of the conservation of energy. If $\hat{n}_2$ is complex, absorption takes place in the second medium, and since the second medium is semi-infinite, whatever energy is not reflected is totally absorbed with the deposited energy falling off exponentially as $e^{-\alpha z}$ where α is the absorption coefficient.

In the case of a TM (p-polarized) wave incident on the interface, one can construct a similar analysis leading to expressions for the reflectivity and transmissivity. However, it is easier to take advantage of a little-known theorem which states that Maxwell's equations remain the same when the electric and magnetic fields are interchanged

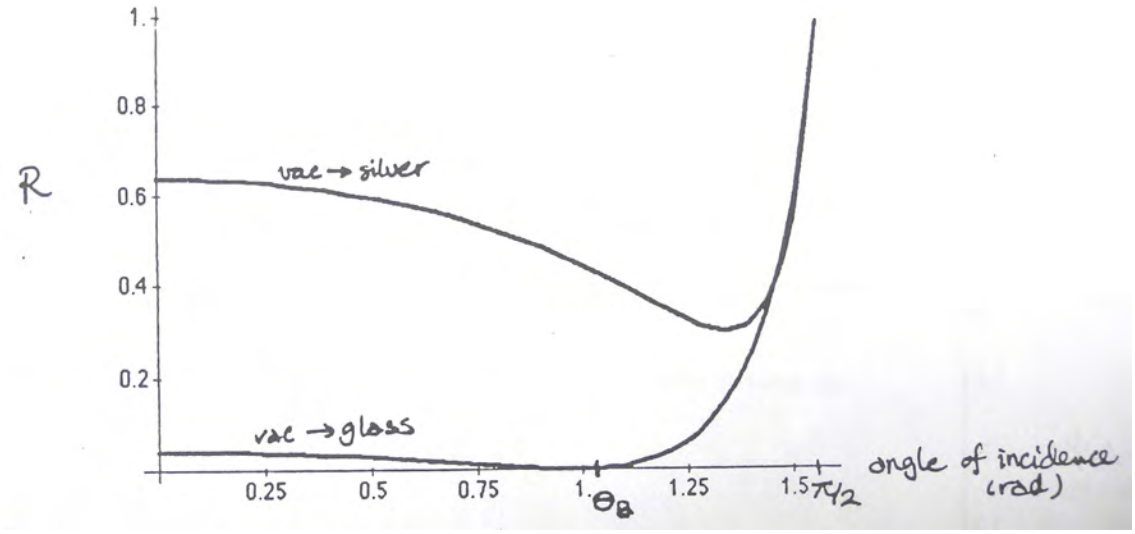


FIGURE 4.2.3. Reflectivity of a TM wave at a glass/air and silver air interface.

simultaneously with ϵ and $-\mu$. Hence any relationship which is valid for TE waves is valid for TM waves provided the appropriate changes are made. For TM waves, therefore, in terms of a characteristic admittance

$$Y_0^{(1)} = \left(\frac{\epsilon_0}{\mu_0}\right)^{1/2} n_1/\cos\theta_1$$

one has that the continuity of the tangential components of the $\vec{E}$ and $\vec{H}$ fields requires

$$Z_0^{(1)}\left[\frac{\mathcal{H}_0^{1+}+\mathcal{H}_0^{1-}}{\mathcal{H}_0^{1+}-\mathcal{H}_0^{1-}}\right] = Z_0^{(2)}$$

and the (TM) amplitude reflectivity is given by

$$\Gamma = \frac{\hat{n}_2\cos\theta_1 - n_1\left(1-\frac{\epsilon_1\sin^2\theta_1}{\hat{\epsilon}_2}\right)^{1/2}}{\hat{n}_2\cos\theta_1 + n_1\left(1-\frac{\epsilon_1\sin^2\theta_1}{\hat{\epsilon}_2}\right)^{1/2}}. \tag{4.2.2}$$

The corresponding amplitude transmissivity for TM waves is found to be

$$\Phi = \frac{2n_1\cos\theta_1}{\hat{n}_2\cos\theta_1 + n_1\left(1-\frac{\sin^2\theta_1\epsilon_1}{\hat{\epsilon}_2}\right)^{1/2}}.$$

As $\theta_1 \to 0$ and TE and TM waves become identical, note that equation 4.2.2for the TM wave does not become identical with equation 4.4.1 for the TE wave. There is an apparent sign difference, which is the subject of question 11 in this chapter. In general the equations which dictate the relations between the amplitudes of incident, reflected and refracted beams are known as the *Fresnel relations*.

In figure 4.2.3 we show the energy reflectivity of TM waves at air/glass and air/silver interfaces. Once again the wave is considered to be incident from the air side. At normal incidence it can be seen that the reflectivity coincides with the values with those of the TE wave. This is as expected, since at normal incidence one in fact can't distinguish between the TE and TM waves. With increasing angle, however, the reflectivity in the case of the TM wave incident on a dielectric drops to a value of zero at the *Brewster angle* , θ_B, where

$$\theta_B = \arctan(\frac{n_2}{n_1}).$$

For angles greater than the Brewster angle the reflectivity rises reaching the value of unity at 90° as in the case of the TE wave.

The significance of the Brewster angle can be understood as follows. At this angle of incidence the propagation vectors for the reflected and refracted waves make an angle of incidence of 90° with respect to each other. The reflected wave is just the radiation field from the electric dipoles near the interface in medium 2 and if these dipoles

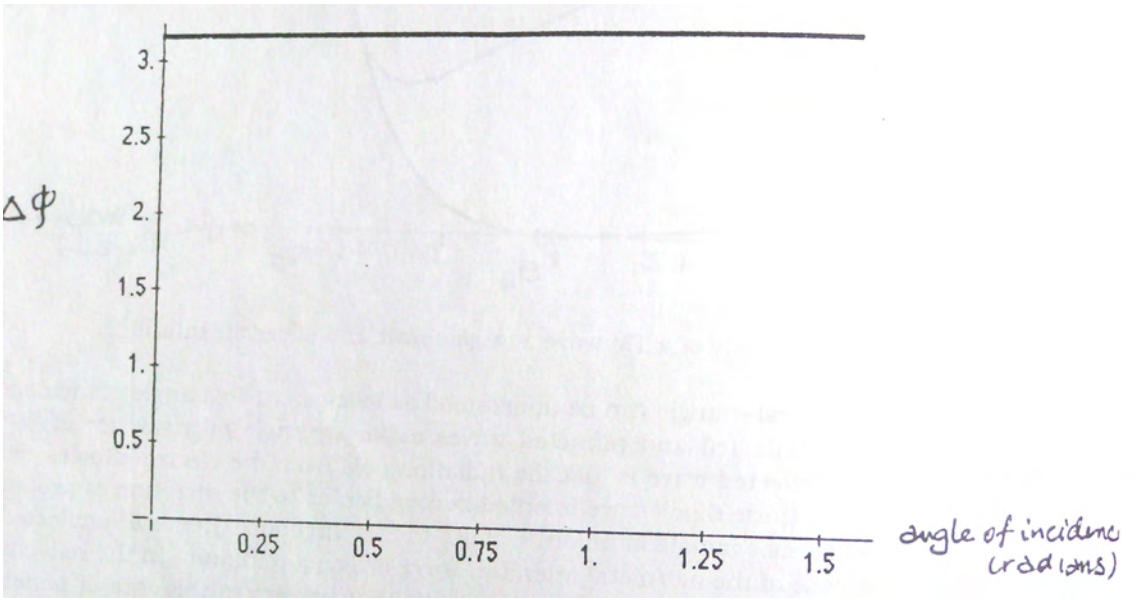


FIGURE 4.2.4. Phase change of TE waves in the case of light incident on dielectric from vacuum.

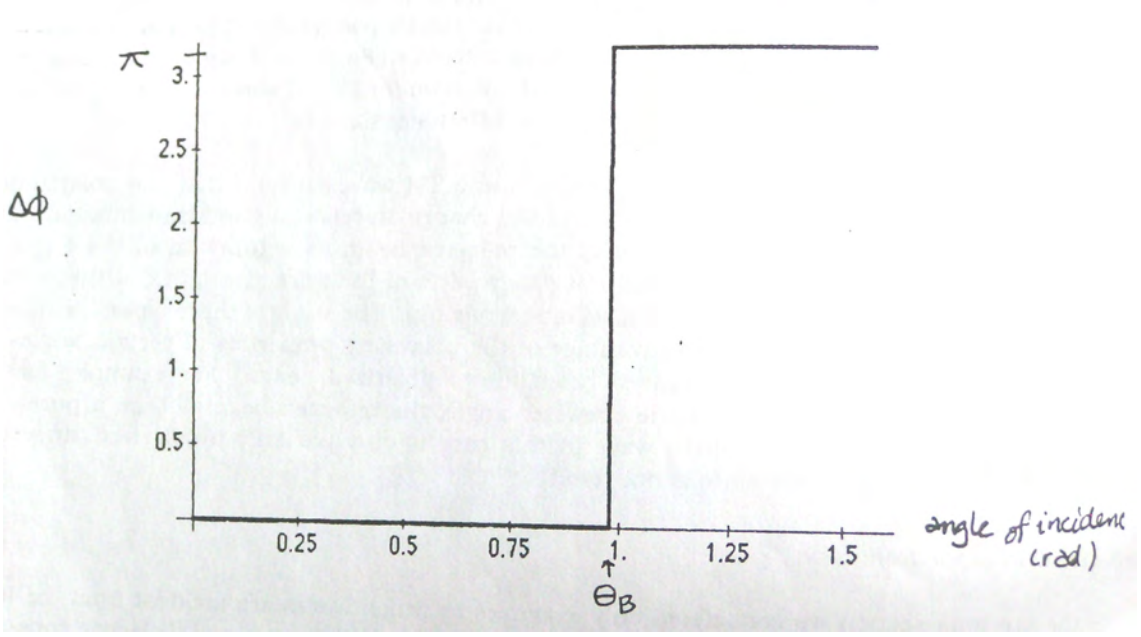


FIGURE 4.2.5. Phase change of a TM wave in the case of light incident on dielectric from vacuum.

are oriented perpendicular to the direction of propagation of the refracted wave, they cannot radiate at an angle of 90°to this direction. Hence the reflected wave amplitude is zero. In the case of the air/metal interface, there is also a minimum in the reflectivity as a function of angle of incidence. However, this minimum cannot be zero in general for a medium with complex refractive index.

Figure 4.2.4 shows the phase change that occurs for reflected s-polarized (TE) beams as a function of the angle of incidence in the case of a purely dielectric interface. Figure 4.2.5 shows the comparable results for p-polarized (TM) light; note that unlike the situation for the TE wave the phase change is not constant but rather changes from zero to π radians at Brewster's angle.

The difference in reflectivity associated with TE and TM waves means that the polarization state of an arbitrarily polarized incident plane wave changes on reflection and transmission. The analysis of the change in polarization state of the reflected beam as a function of the angle of incidence allows one to deduce the real and imaginary parts of the

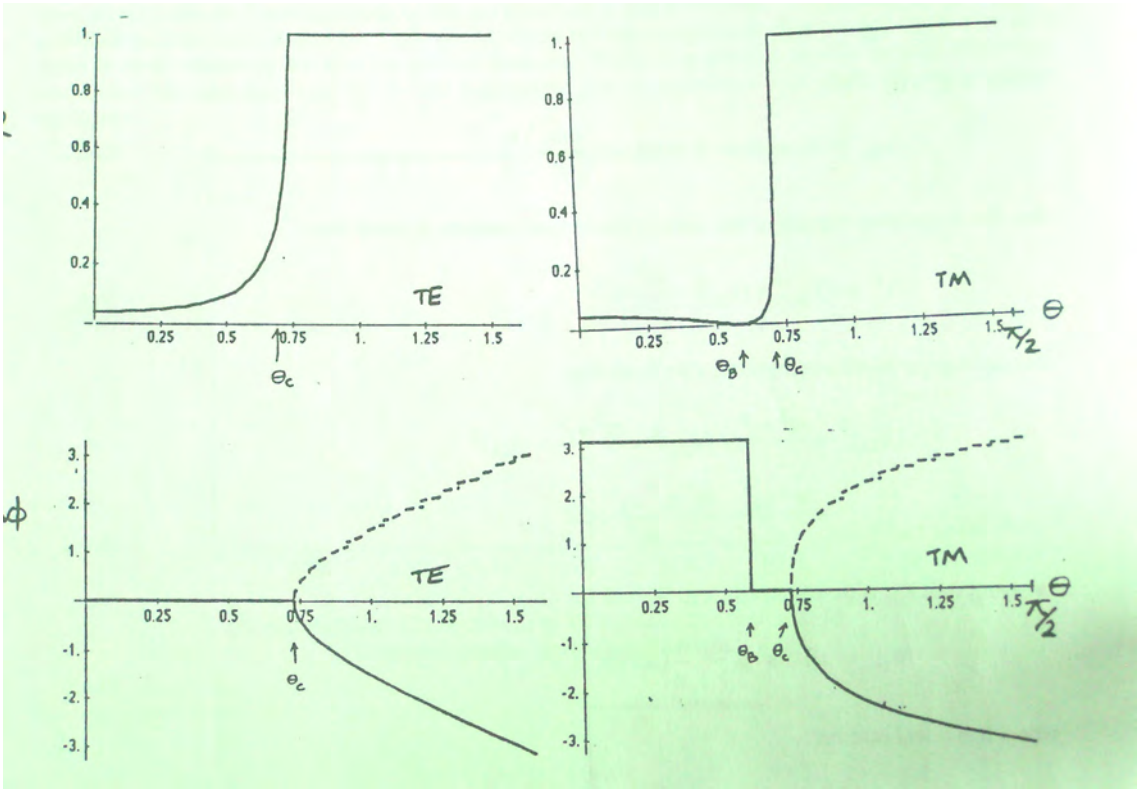


FIGURE 4.3.1. Reflectivity and phase change of TE and TM waves in the case of a beam entering a rare medium from a dense medium.

refractive index of the second medium if the first medium is air, or (for more accuracy) vacuum. The determination of the complex refractive index in this way is known as *Ellipsometry*. One can also take advantage of the polarizing properties of certain angles of incidence to render an initially unpolarized beam into a polarized beam. For example, for an incident unpolarized beam incident at the Brewster angle, the reflected beam is a purely s-polarized or TE wave. The transmitted wave in this case is only partially p-polarized since the transmission coefficient for the s-state is not zero.

4.3. Total Internal Reflection

In the previous section, in the specific examples we offered, we considered only waves incident from the less dense medium. Let us now consider the situation where both media have a real dielectric constant and the wave is incident from the denser optical medium. The Fresnel relations are still generally valid since no assumption was made concerning which medium was more dense. There is an interesting development, however, in that, if the angle of incidence is greater than a *critical angle* θ_c where

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

then a real angle, θ_2, corresponding to the refracted wave, does not exist. Application of the Fresnel relations in this case shows that the wave is totally reflected at the interface for $\theta_1 > \theta_c$. In this case there is no propagating wave in the second medium. Figure4.3.1 summarizes the energy reflectivity and phase changes occurring for s- and p-polarized beams entering a rare medium from a dense medium.

As θ_1 approaches θ_c from 0°, the angle of refraction approaches 90°, and at the critical angle the refracted wave is travelling parallel to the interface. For larger angles we still must have a field in the second medium to satisfy the boundary conditions for the electromagnetic fields. To determine the nature of the field structure in the second medium, we can begin by recalling that we still have translational symmetry along the x direction so that the x-components of the propagation vector of the incident and penetrating waves must still be equal. Hence

$$k_{1x} = k_{2x}$$

where, if $\theta_1 > \theta_c$

$$k_{1x} = k_1 \sin\theta_1 > k_1 \sin\theta_c = \frac{\omega n_1}{c}\frac{n_2}{n_1}.$$

But in the second medium

$$k^2 = (k_{2x})^2 + (k_{2z})^2.$$

Combining the last three equations we have that

$$\begin{aligned}(k_{2z})^2 &= \left(\frac{\omega n_2}{c}\right)^2 - (k_{2x})^2 \\ &= \left(\frac{\omega n_2}{c}\right)^2 - (k_{1x})^2 \\ &< \left(\frac{\omega n_2}{c}\right)^2 - \left(\frac{\omega n_1}{c}\frac{n_2}{n_1}\right)^2\end{aligned}$$

implying that k_{2z} is purely imaginary. We set

$$k_{2z} = i\beta \quad \text{for} \quad \beta = \frac{\omega}{\mathrm{c}}(\mathrm{n}_1^2 \sin^2\theta_1 - \mathrm{n}_2^2)^{1/2}$$

where β is a real number.

The field in the second medium therefore varies as

$$\mathcal{E}_2 \propto \exp(i\vec{k}_2 \cdot \vec{r}) = \exp(ik_{2x}x)\exp(-\beta z)$$

indicating that the amplitude of the field falls off exponentially with distance in the less dense medium. The wave fronts or surfaces of constant phase (in the *y-z* plane) are perpendicular to the surfaces of constant amplitude which are in the *x-y* plane. The wave is therefore an *inhomogeneous wave*. A similar situation occurs, as we saw earlier, in the case of refraction of a wave into a medium of complex dielectric constant.

The field which decays exponentially in the less dense medium is referred to as the *evanescent field*. The penetration depth depends on the difference in the two refractive indices and the wavelength. The greater the mismatch in the refractive indices the less is the penetration depth. In most cases the penetration is on the order of the wavelength of light. If one increases the refractive index of the second medium then the condition for total internal reflection can be made to disappear. Alternatively, one can make the condition for total internal reflection disappear by bringing a dense medium to within, or less than, the wavelength of light to the second medium. For example if the second medium is air and the first medium is glass, and one brings another glass piece close to the first medium, as indicated in figure 4.3.2, then of course when the two glass pieces are in perfect contact, reflection does not occur. One then speaks of a situation of frustrated total internal reflection. *Frustrated total internal reflection* occurs for separation distances less than β^{-1}, with the reflectivity being unity if the separation is $\gg \beta^{-1}$ and zero if the separation is zero. There is a formal analogy between the concept of total internal reflection and the quantum mechanical reflection of a particle moving in the vicinity of a potential barrier. Consider the one dimensional version of this problem where the particle moves along the z-axis in a potential, *V(z)*. If

$$V(z) = \begin{cases} 0 & z < 0 \\ V_0 & z \geq 0 \end{cases}$$

and the particle has total energy $E < V_o$, then the wave function is oscillatory on the negative z-axis but decays exponentially in the classically forbidden region. The range of the "evanescent" wave which is allowed by the uncertainty principle, is governed by the difference between the potential barrier and the total energy and is given by

$$\triangle z \propto (V_0 - E)^{-1/2}$$

The reflection coefficient of this barrier is unity. If the potential barrier becomes infinitely high the penetration depth also goes to zero. If we add a potential $V'(z)$ given by $V'(z) = \begin{cases} -V_0 & \text{for } z > z_0; \\ 0 & \text{for } z \leq z_0 \end{cases}$ with $z_0 = (V_0 - E)^{-1/2}$, then the particle moves in the vicinity of a potential barrier of finite width and height, and there is a finite tunneling current into the other side of the barrier. This in essence is the quantum mechanical analogue of frustrated total internal reflection, with the potential barrier being the analogue of the air gap.

Although the electromagnetic field in the less dense medium is not zero, for reasons related to the boundary conditions of electromagnetic waves, there is no energy flow in the direction normal to the interface, the z-direction. There is therefore no problem with violation of conservation of energy if the reflection is unity. Indeed, as we found above, the surfaces of constant phase are perpendicular to the interface in medium 2 and it is straightforward to show that the z-component of the time-averaged Poynting vector is zero. The time averaged x-component is not zero,

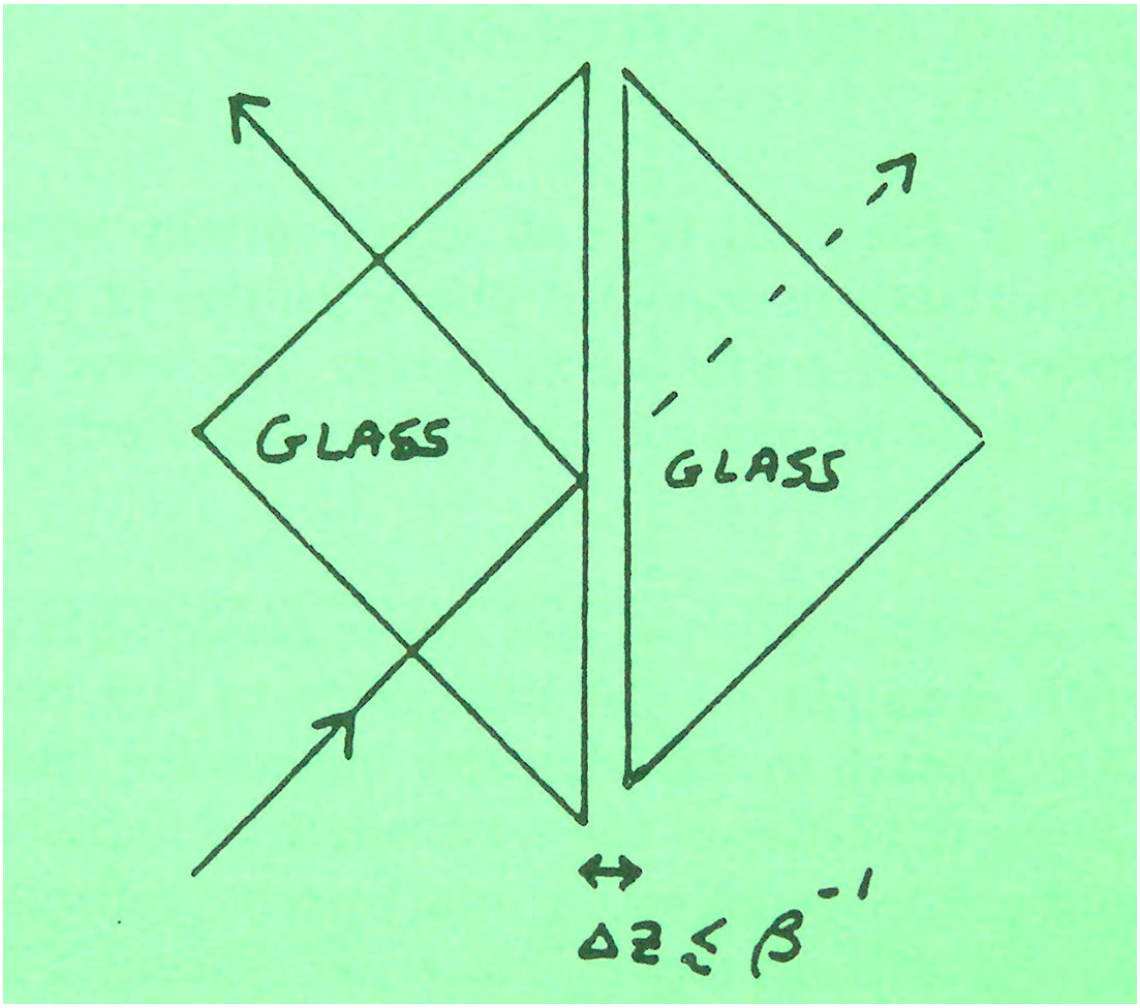


FIGURE 4.3.2. Geometry for frustrated total internal reflection.

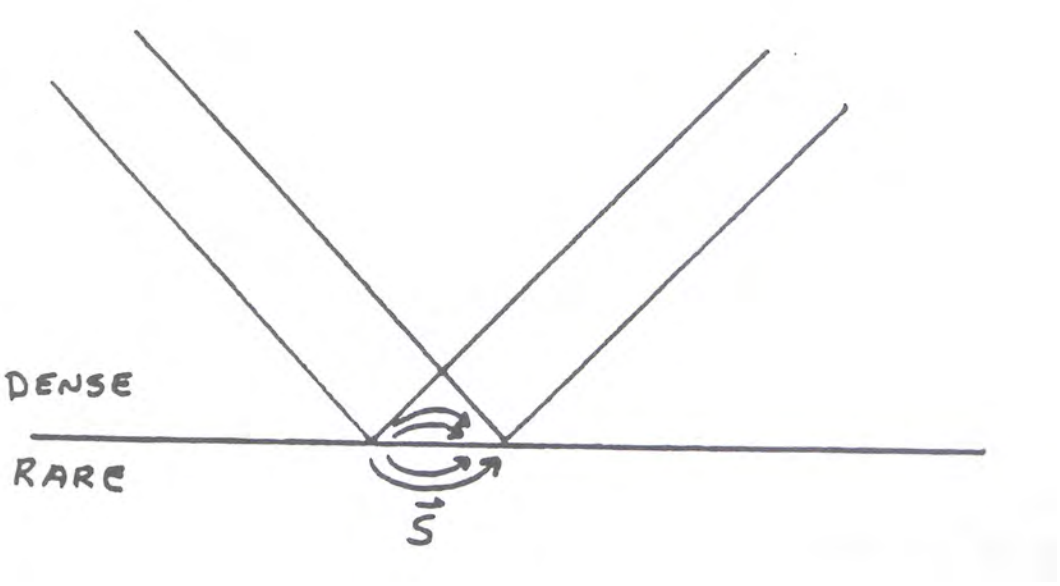


FIGURE 4.3.3. Diagram showing the direction of the Poynting vector in the case of total internal reflection.

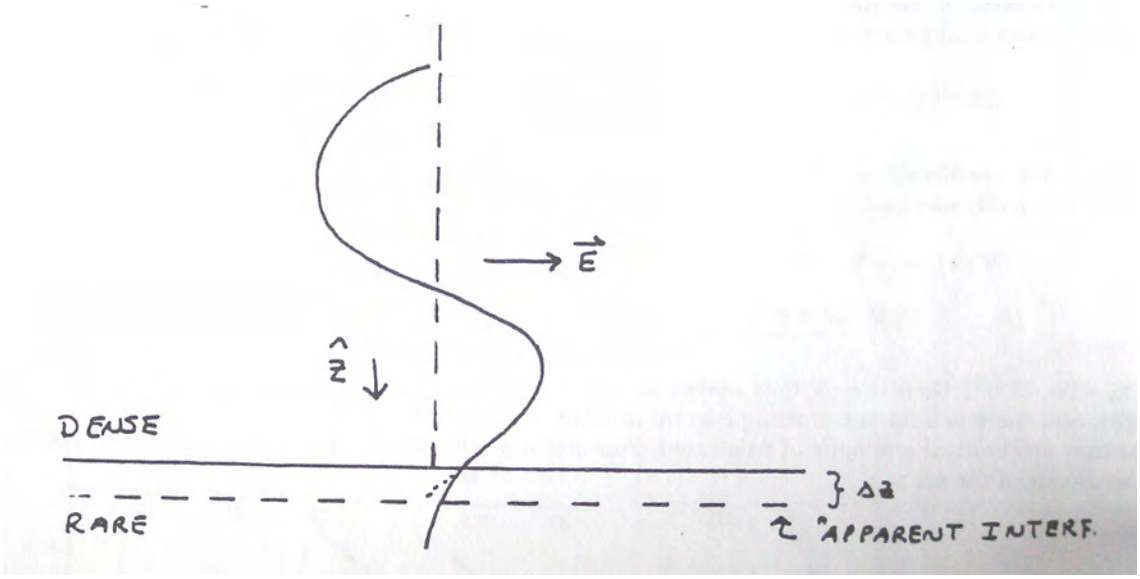


FIGURE 4.3.4. Illustration of the spatial variation of the electric field in the case of a TE wave undergoing total internal reflection.

however, and there is a component of the energy flow parallel to the interface. Figure 4.3.3 illustrates the direction of the Poynting vector in the case of total internal reflection.

The fact that the x-component of time averaged Poynting vector does not equal zero leads to a shift in the location of the origin of the reflected beam. Although for a plane wave this is of academic interest only, for beams of a finite transverse extent this shift, which is of the order of the wavelength of light can be observed. It was first discussed extensively in 1948 and is called the *Goos-Hänchen shift* after the two scientists who first pointed it out. A different way of understanding this shift is shown in figure 4.3.4. Because the electric field cannot strictly become zero at the interface, the node of the field appears to be at a plane a small distance into the second medium. It therefore appears that the reflected wave arises from this deeper plane. Hence, there is a lateral shift in the reflected beam relative to the incident wave. This can be incorporated into the description of the reflected wave by introducing a phase shift in the reflected beam phase factor.

4.4. Reflection from Rough Surfaces

The interaction of light with an optically rough surface is one of the most challenging problems in optics. Optically rough surfaces include surfaces which are periodic in amplitude or composition as well as surfaces which are statistically rough. In principle, the problem, like most problems in optics, should be easy to solve. All one has to do is to solve Maxwell's equations together with the boundary conditions for the electromagnetic fields at the interface of interest. In practice, the implementation of the boundary conditions is difficult, either because of difficulties in describing a "rough" surface or because the boundary conditions cannot be formulated in a way to make the solution of the wave equation easily tractable, apart from having the whole problem solved on a computer. Even today thousands of optical scientists are engaged in research in this area. The main interest in the work relates to applied problems in optics, such as the manufacture and characterization of diffraction gratings for spectroscopy or the development of highly specular surfaces for mirrors which are used in all kinds of optical instruments or consumer items such as consumer items which must be "nice and shiny".

The problem can be treated nearly rigorously using diffraction theory and we do so in chapters 8 and 9. In this section we illustrate some of the salient features of the interaction of a light beam with an optically rough surface with emphasis on the relaxation of the conditions that led us to the description of planar interfaces and possible modification of the laws of reflection and refraction. We only consider monochromatic plane waves and will restrict ourselves to interfaces which separate homogeneous media. We assume interface which have height fluctuations that are comparable to the wavelength of light as depicted in figure 4.4.1.

Consider a plane surface defined to be $z = 0$ (with z taken to be positive into the material) and which passes through the surface topography. One can describe the variations in the topography of the surface by a height function, $h(x, y)$, which is the (shortest) distance from the reference plane to the actual interface.

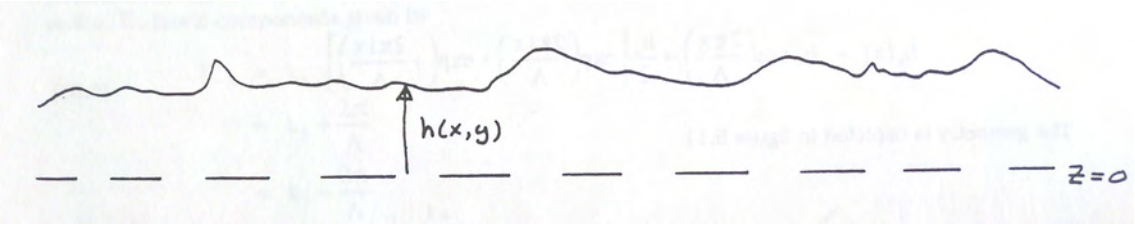


FIGURE 4.4.1. Height variations of an optically rough surface.

An arbitrarily rough surface is difficult to treat in general but the problem is much simpler to deal with if we can consider the function, h(x,y) to be a superposition of components which can be treated individually. The total solution can then be obtained by using the superposition principle, but paying attention to conservation of energy. One of the most powerful ways of treating the problem is to use Fourier analysis. The height function like any bounded, continuous, square-integrable function can be Fourier analyzed into a superposition of periodic functions. To illustrate how this applies to the problem in optics, but so as not to get bogged down in the mathematics of analyzing functions which have two independent variables, we consider a height or roughness function which is a function of x only. In this case the height function can be Fourier analyzed as

$$z = h(x) = \int_q g(q) e^{iqx} dq$$

where, since $h(x)$ is a real function,

$$g(-q) = g^*(q).$$

The function $h(x)$ can therefore be written as a superposition of periodic sinusoidal gratings with period $2\pi/q$ and (complex) amplitude $g(q)$.

If the surface has a periodic profile such as represented by a sawtooth function, then the height function can be decomposed as a discrete sum of *sinusoidal gratings* of the form

$$h(x) = \sum_{n=-\infty}^{n=\infty} G_n \exp(i2\pi nx/\Lambda)$$

where Λ is the periodicity of the surface profile.

The interaction of a plane wave with a rough surface can be reduced to its simplest element which is the interaction of a plane wave with sinusoidal varying surfaces. The overall spatial light distribution reflected from the rough surface in general is then obtained from an application of the superposition principle. In doing so one has to normalize the scattered light field amplitudes in such a way so that the total intensity in the scattered field is equal to the total incident intensity (for a totally reflective surface).

Since the essence of the problem reduces to the interaction of a plane wave with a sinusoidal grating structure, let us consider this problem in some detail. To be specific let us consider a TE wave in vacuum incident on a perfectly reflecting metallic surface which has a surface grating given by

$$h(x) = h_0 \cos(\frac{2\pi x}{\Lambda}) = \frac{h_0}{2}\left[\exp(\frac{2\pi ix}{\Lambda}) + \exp(\frac{-2\pi ix}{\Lambda})\right]$$

The geometry is depicted in figure 4.4.2.

Because the incident and reflected waves must give a total electric field at the interface in the case of a totally reflecting surface (an idealization!), we have for an incident field

$$\mathcal{E}_0^{1+} \exp\left[i\left(k_x x + k_z z\right)\right]$$

and a reflected field $\mathcal{E}^{1-}(x,z)$ that together satisfy

$$\mathcal{E}_0^{1+} \exp\left[i\left(k_x x + k_z z_s\right)\right] + \mathcal{E}^{1-}(x, z_s) = 0 \quad \text{for} \quad z_s = h(x).$$

Hence at the surface the reflected field is given by

$$\mathcal{E}^{1-}(x, z_s) = -\mathcal{E}_0^{1+} \exp\left(ik_x x\right) \exp\left(ik_z h_0 \cos\left[\frac{2\pi x}{\Lambda}\right]\right)$$

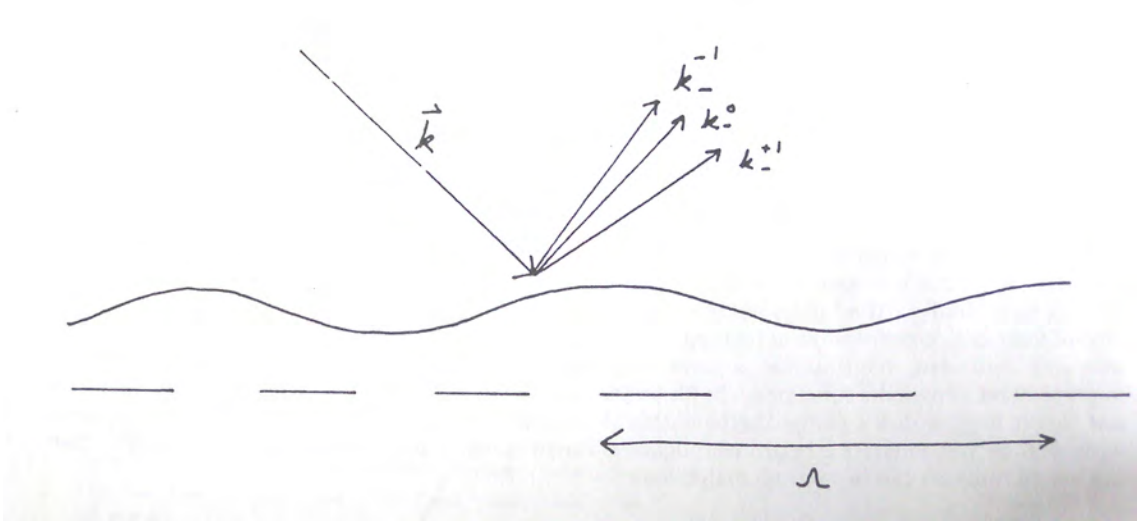


FIGURE 4.4.2. Interaction of a TE plane wave with a sinusoidal grating.

At a *planar* surface just above the sinusoidal surface, to within a constant phase factor, the reflected wave is given by

$$\mathcal{E}^{1-}(x,z) = -\mathcal{E}_0^{1+}\exp\left(ik_x x\right)\exp\left(2ik_z h_0\cos\left[\frac{2\pi x}{\Lambda}\right]\right)$$

which in the limit $h_0 k_z \ll 1$ gives

$$\mathcal{E}^{1-}(x,z) = -\mathcal{E}_0^{1+}\exp\left(ik_x x\right)\left(1+2ik_z h_0\cos\left[\frac{2\pi x}{\Lambda}\right]\right)$$

$$= -\mathcal{E}_0^{1+}\exp\left(ik_x x\right) + ik_z\mathcal{E}_0^{1+}h_0\left\{\exp\left(i\left[k_x+\frac{2\pi}{\Lambda}\right]x\right)+\exp\left(i\left[k_x-\frac{2\pi}{\Lambda}\right]x\right)\right\}.$$

Hence, in the limit of a small amplitude grating one obtains three reflected waves whose propagation vector, $\vec{k}$ has x-components given by

$$k_{-x} = k_x$$
$$= k_x + \frac{2\pi}{\Lambda}$$
$$= k_x - \frac{2\pi}{\Lambda}$$

and in general one would have (if one retained higher terms in the expansion of the exponential)

$$k_{-x} = k_x + \frac{m2\pi}{\Lambda}$$

where $m = 0, \pm1, \pm2, ...$ is referred to as the order of the reflected (or diffracted) wave. Because the reflected waves are propagating waves the z-components of the propagation constant can be found from the dispersion relation for waves in vacuum. It follows that for the m'th order wave we have

$$\left(k_x + \frac{m2\pi}{\Lambda}\right)^2 + \left(k_{-z}^m\right)^2 = k^2 = \frac{\omega^2}{c^2} \tag{4.4.1}$$

The z-component of $\vec{k}$ for each of the reflected waves can be obtained from the geometrical construction shown in figure 4.4.3. In terms of the angle of incidence, θ_1, and angle of reflection, θ_{1-}^m equation 4.4.1 can be written as

$$\sin\theta_{1-}^m = \sin\theta_1 + \frac{m\lambda_0}{\Lambda} \tag{4.4.2}$$

which is known as the *fundamental grating equation*.

In considering the derivation that led up to equation 4.4.2 we assumed that $h_0 k_z \ll 1$. If we relax this condition we obtain more scattered wave components with $|m| > 1$ which satisfy equations 4.4.2 with the amplitude of these

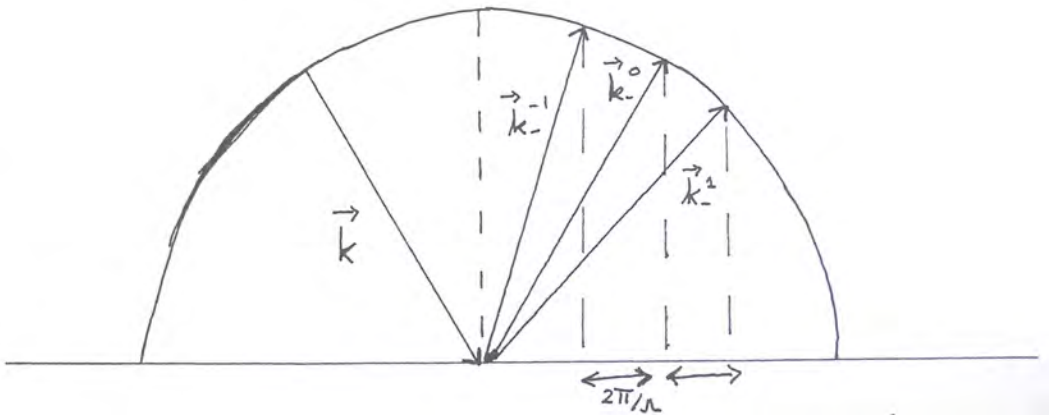


FIGURE 4.4.3. Geometrical construction used to obtain the directionality of reflected waves from a small amplitude surface grating.

"higher order" waves dropping with increasing m value. Of course, the number of such waves is finite because of the fact that

$$|k_{-x}| < |k|$$

To summarize, we see that a regular periodic surface leads to a multitude of scattered waves. As the amplitude of the grating decreases the amplitude of the higher order scattered waves drops with, of course, only the specular component (m= 0) existing in the limit of $h_0 \to 0$. In the case of an arbitrarily rough surface many scattered wave components exist. In certain cases one can use the information on the scattered light angular distribution to obtain some information about the surface roughness, but in general this is not a unique process since what is measured is the energy or intensity of the total scattered field and much of the phase information contained in the field amplitudes is lost. This is why the "rough surface" problem in optics is difficult to treat except in some statistical sense.

It is worth amplifying on the comments made on periodically rough surfaces or, in general, gratings which do not have a sinusoidal surface profile. In the case of a sinusoidal grating we can discern from the grating equation that perhaps the most important property associated with gratings is their ability to disperse light in non-zero order beams. That is, the angle of reflection for an incident plane wave depends on the wavelength. It is this property of gratings that make them useful in grating spectrometers, which are used for analyzing light into its different spectral components and enable physicists to perform spectroscopy of atomic, molecular and solid state systems. We discuss this property of gratings in more detail in the chapter on diffraction.

References

H.A. Haus, *Waves and Fields in Optoelectronics*, Prentice Hall, Englewood NJ, 1984.

M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Toronto, 1975.

Problems

1. A beam of circularly polarized light is incident on a glass surface ($n = 1.5$) at an angle of incidence of 45°. Describe the polarization state of the reflected and transmitted beams.

2. The critical angle for total internal reflection in a certain piece of glass is exactly 45°. What is the Brewster angle for a) external reflection and b) internal reflection?

3. Determine the reflectivity of a planar silver surface at an angle of incidence of 45° and for a TE beam; $\hat{n} = 2.3i + 10i$.

4. Show that the energy transmissivity of a planar surface separating a dielectric of refractive index n_2 from a dielectric of refractive index n_1 is

$$T = \frac{n_2 cos\theta_2}{n_1 cos\theta_1} |\Phi|^2$$

Hint: one has to consider conservation of energy density at the boundary.

5. The accompanying figure is a diagram of a *Mooney rhomb* for producing circularly polarized light. Show that if the index of refraction of the rhomb is 1.65, the apex angle A should be about 60°.

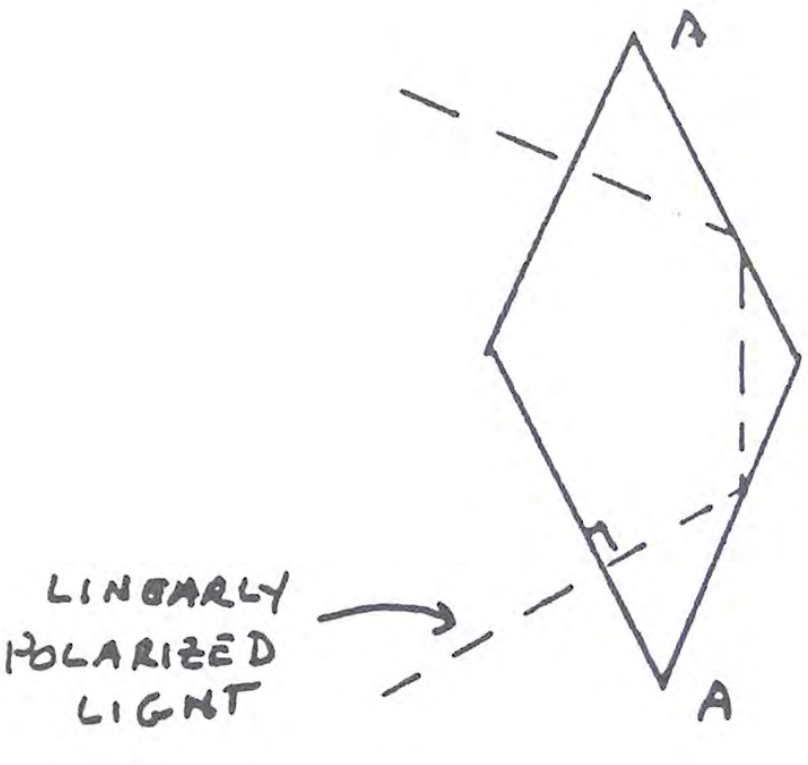


FIGURE 4.4.4. Mooney rhomb

6. To first order in a Taylor series expansion in terms of h_0/λ, find the amplitudes of the waves reflected from a perfectly conducting grating with a square wave profile of height h_0, period,Λ, for an incident wave

$$\vec{\mathcal{E}}=\mathcal{E}_0^{1+}\hat{y}\exp\left[i\left(k_x x+k_z z\right)\right]$$

with the grooves parallel to $\hat{y}$.

7. Calculate the approximate angular spread of a beam which is generated in reflection when an infinite, monochromatic, plane wave impinges on a flat surface which has a transverse extent equal to 10λ; 1000λ.

8. Consider an infinite plane wave falling on an absorbing medium with a flat surface. Show that the attenuation of the transmitted beam as a function of depth is independent of the angle of incidence. How do you reconcile this with your intuition which might indicate that the attenuation is dictated by the direction of propagation of the wave fronts?

9. In general would you expect the efficiency of metal-coated gratings to be higher for s- or p-polarized light? If the amplitude and periodicity of the teeth are the same which type of grating would have the highest efficiency for first order operation in the case of a normally incident beam; sinusoidal, rectangular, sawtooth?

10. Is it possible to define a Jones matrix for the a) transmission and b) reflectivity characteristics of a planar interface?

11. As mentioned in the text, equations for the reflection amplitudes of TM and TE waves, respectively, don't give the same sign when $\theta_1 = 0$, where TM and TE waves become the same. What is the reason for this difference? Show the details; you may wish to refer to figure 4.2.1, for $\theta_1 \to 0$.

Part 2

Ray Optics

CHAPTER 5

Geometrical Optics

Dreamer of dreams, born out of my due time
Why should I strive to set the crooked straight?
William Morris

5.1. Geometrical Optics Approximation

In discussing the properties of light and its interaction with matter up to this point we have often made explicit reference to the wave properties of light. This has been particularly true when we discussed effects related to the propagation properties and phase retardation effects associated with beam polarization. The transverse dimensions of the beam have largely been ignored, and in most cases we simply considered plane waves incident on infinitely transverse objects. Historically, it has been well recognized that many of the optical properties of light could be explained without recognition of its wave character. Such phenomena as rectilinear propagation, Snell's law, sharp shadows, etc. could be understood without a wave character of light. It was mainly the discovery of interference effects and diffraction phenomena that forced scientists to consider wave manifestations of light, contrary to the suppositions of Newton. Diffraction phenomena, to be discussed in detail in chapters 8 and 9, involve deviations from the rectilinear properties of light propagation and can be naively described as "bending of light" around an obstacle or on passing through an obstacle. As will be seen, diffraction phenomena becomes important when the wavelength of light is comparable to the characteristic distance over which the optical properties of a material are changing, e.g. the size of an aperture, the scale of roughness of a surface, etc.

In the opposite limit where the wavelength is small compared to this characteristic distance, one speaks of the regime of "*geometrical optics*" or "*ray optics*". This latter designation arises from the historical view that light was emitted in straight lines or rays from a source. A light beam of small or large transverse extent (but small compared to the scale of optical inhomogeneity of an object) can then be described in an abstract sense as a bundle of rays. The path of the rays is determined simply by the so called laws of geometrical optics, the laws reflection and refraction. Note that only information concerning the path is obtained, and no information concerning intensity, phase, etc. are described.

The influence of simple optical elements, such as prisms, mirrors, and lenses, on light beams is most easily described in the ray optics picture. The finite-size aspects of these elements in the context of diffraction is discussed later. The analysis of such elements and systems of such elements is straightforward if one considers an individual ray and the laws of reflection and refraction. In addition, we will find it useful to consider an offshoot of these laws, known as the *principle of reversibility*. This law, which is simply consistent with the other two laws, simply states that any actual ray of light in an optical system, if reversed in direction, retraces the same path backwards.

This chapter deals with rays interacting with a variety of interfaces. From a study of the interaction with elementary interfaces one can develop an understanding of the behavior of rays in more complex optical systems. We proceed then to discuss the reflected and transmitted rays generated by single interfaces, compound interfaces (prisms), and curved interfaces which form mirrors and lenses. In so doing we obtain the so-called imaging properties of simple systems which relate how a bundle of rays is transformed by the system. Finally we end with a discussion of common deviations, or aberrations, from perfect image formation.

5.2. Rays at a Plane Interface

In the last chapter we discussed the details of the reflectivity and transmissivity of plane waves at a simple plane interface. Here we restate the main results in the language of rays simply to introduce ray concepts and notation. Consider a ray incident on a plane interface as indicated in figure 5.2.1.

One can represent a ray by a unit vector having the direction of the ray. For the incident ray we have $\vec{r}_1 = (x, y, z)$. For an interface lying in the plane $z = 0$, the law of reflection gives for the reflected ray $\vec{r}_2 = (x, y, -z)$. It follows that if three planes meet mutually perpendicularly, for example, $x = 0, y = 0$ and $z = 0$, the reflected ray is given by $\vec{r}_2 = (-x, -y, -z)$ or is returned in the direction of the original ray! Such an optical element, which obviously works

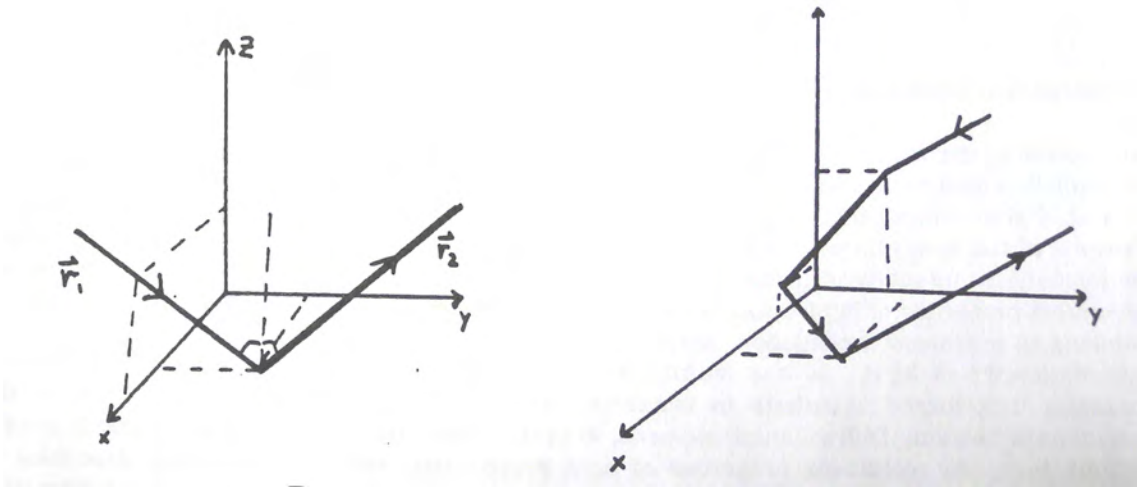


FIGURE 5.2.1. Geometry of a ray incident on a plane.

FIGURE 5.2.2. Image formation of a plane mirror.

for any angle of incidence, is referred to as a corner cube reflector. They are commonly made of glass and operate via total internal reflection at the glass/air interfaces. Additionally, one can use three mirrors which are mutually orthogonal.

Image formation of a *plane mirror* can be understood with reference to figure 5.2.2. For a *point object* in front of the mirror the reflected light appears to be coming from an *image point* behind the mirror. The object and image points are referred to, in general, as *conjugate points* of the imaging system. Simple geometry shows that the image (from which the rays seem to emanate) is located exactly the same distance behind the mirror as the object is in front of it. That is, the conjugate points are equidistant from the mirror surface. An extended object, as the illustration suggests, produces an image with the same transverse orientation but a right handed object becomes a left handed handed object. For a perfect mirror there is no distortion in the image and the apparent size is the same as that of the object located at the same distance from the observer. The magnification is therefore unity. Mirrors generate *virtual images* which means that no light actually passes through the apparent location of the image. This contrasts with *real images* (as formed by some lenses, for example) where light actually passes through the image location.

The refracted rays produced by a plane interface have directions dictated by Snell's law. Figure 5.2.3 shows the geometry of a bundle of incident and refracted waves for a variety of interfaces. In all cases for a transmitted beam we have that

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

For rays which make a small angle with respect to the normal (referred to as *paraxial rays*) one can make approximations to this law based on the fact that for θ_1 small

$$\sin\theta_1 = \tan\theta_1. \tag{5.2.1}$$

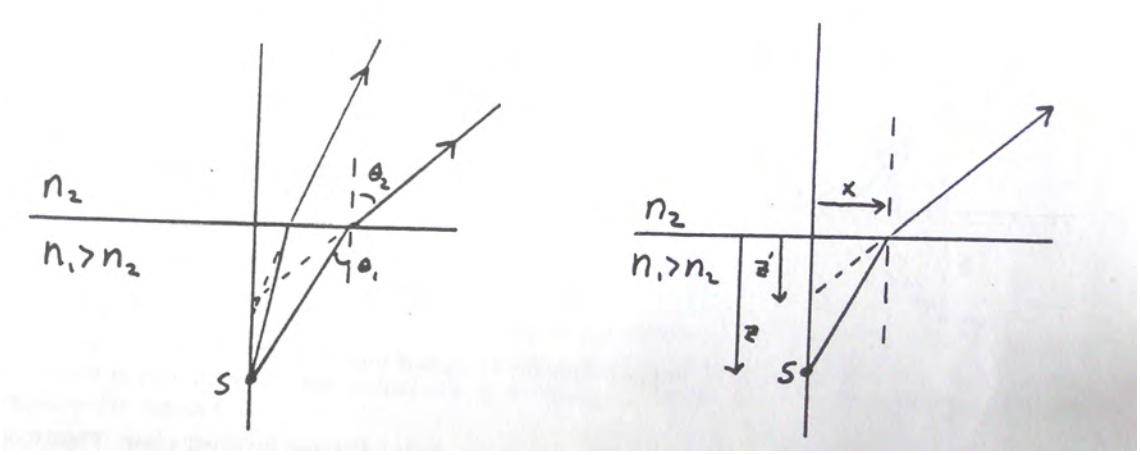


FIGURE 5.2.3. Geometry of rays refracted at a plane interface.

For paraxial rays therefore Snell's law is

$$n_1\theta_1 \simeq n_2\theta_2. \tag{5.2.2}$$

Equations 5.5.1 and 5.2.2 can be used to derive the imaging properties of paraxial beams which are transmitted through a dielectric interface. Referring to figure 5.2.3 and using

$$n_1 \tan\theta_1 = n_2 \tan\theta_2$$

we have from figure 5.2.3 that

$$n_1\left(\frac{x}{z}\right) = n_2\left(\frac{x}{z'}\right)$$

or

$$z' = \frac{n_2}{n_1 z}. \tag{5.2.3}$$

The image distance, for an object buried in a medium of higher refractive index than that of the observer, is foreshortened. This leads to a distortion of the image of an extended object. Note that there is no lateral distortion. For other than paraxial rays, the distortion is still present and, without the approximation to Snell's law, is more complicated than indicated by equation 5.2.3.

5.3. Prisms

We proceed now to discuss ray propagation through a *prism*. This is an object with a different refractive index than its surroundings and possessing two plane interfaces making an angle, α (*apex angle*) as shown in figure 5.3.1. Such elements are useful for many elementary optical functions as will become apparent after we consider how a ray is deviated from its original path on passage through the two interfaces.

To analyze a prism consider an ray incident on the first interface at an angle θ_1 as shown. Other angles relevant to the analysis are shown in the figure. Snell's law at each interface gives us

$$n_1 \sin\theta_1 = n_2 sin\theta_1^{'}.$$

Geometry also dictates that the following relations must hold between the angles:

$$\delta_1 = \theta_1 - \theta_1^{'}$$

$$\delta_2 = \theta_2 - \theta_2^{'}$$

$$\beta = \pi - \theta_1^{'} - \theta_2^{'} \quad \theta_1^{'} + \theta_2^{'} = \alpha$$

$$\delta = \delta_1 + \delta_2 = \theta_1 = \theta_2 - \theta_1^{'} - \theta_2^{'}$$

After straightforward, but tedious, algebra one finds that the *angle of deviation* of the ray is given by

$$\delta = \theta_1 + \arcsin\left[\sin\alpha\left(n^2 - \sin^2\theta_1\right)^{1/2} - \sin\theta_1\cos\alpha\right] - \alpha$$

The angle of deviation typically varies with the incidence angle in a manner which is shown in figure 5.3.2. There is an angle of minimum deviation for all prisms and its value can be used to accurately obtain the value of the refractive

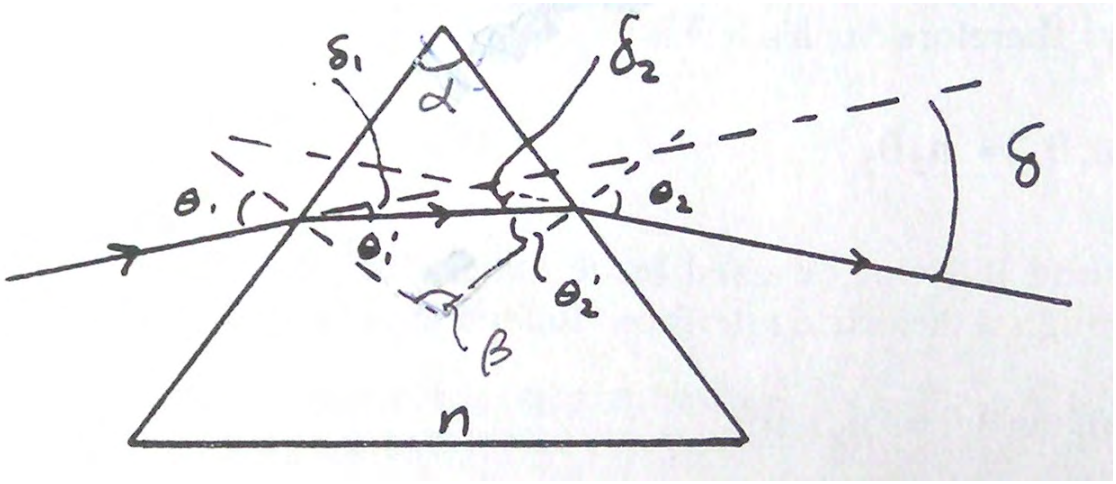


FIGURE 5.3.1. Progress of a ray through a prism.

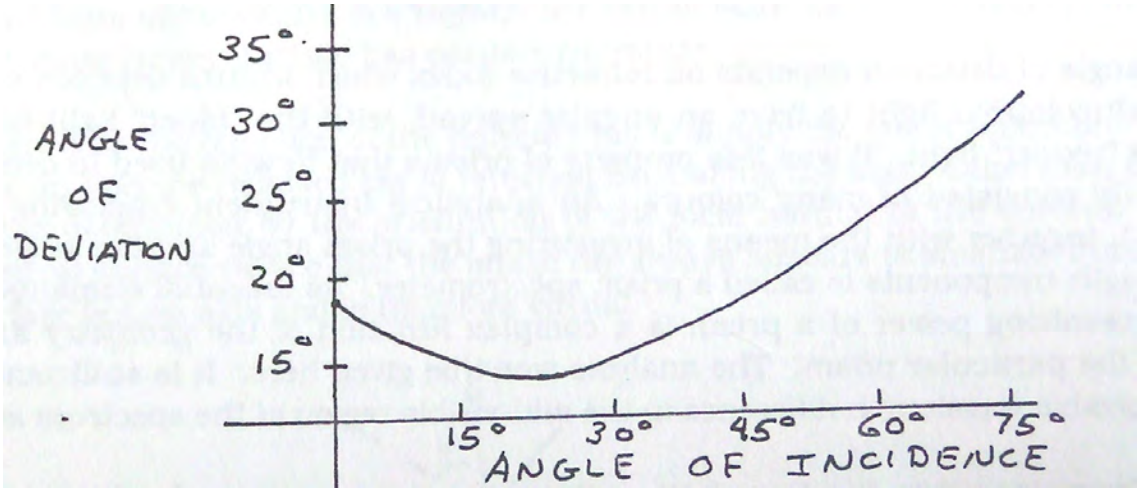


FIGURE 5.3.2. Graph of angle deviation versus angle of incidence for a light ray through a prism with $\alpha = 60°$.

index of the prism in practice. The value of this angle can be arrived at via a simple argument. It is clear that for the angle of minimum deviation the ray of light passes symmetrically through the prism, making it unnecessary to have subscripted angles. This simply follows from the principle of reversibility and the fact that the angle is a minimum. For

unscripted angles we simply have that

$$\delta = 2\theta - 2\theta'$$

and

$$\alpha = 2\theta'$$

for which

$$\theta' = \alpha/2 \quad \theta = (\alpha + \delta)/2$$

and Snell's law yields

$$\sin\left(\frac{\alpha+\delta}{2}\right) = n\sin\left(\frac{\alpha}{2}\right)$$

or

$$n = \frac{\sin\left(\frac{\alpha+\delta}{2}\right)}{\sin\left(\frac{\alpha}{2}\right)}.$$

The refractive index can be determined quite accurately by measuring the angle of deviation and the apex angle of the prism.

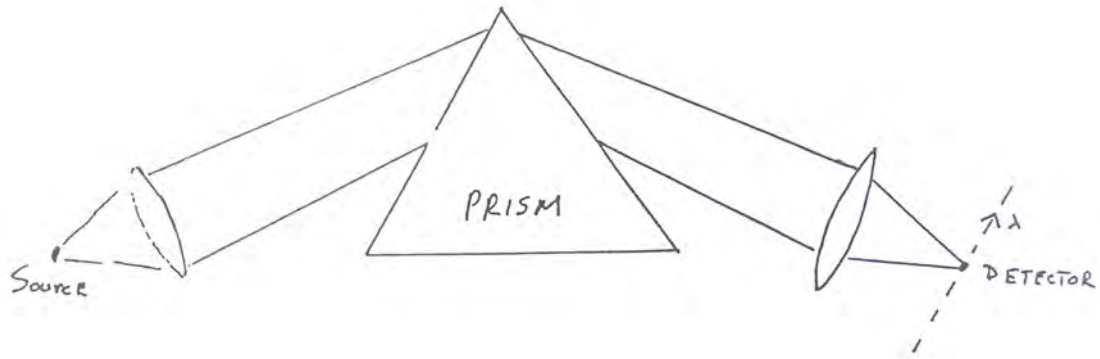


FIGURE 5.3.3. Essentials of a Prism Spectrometer

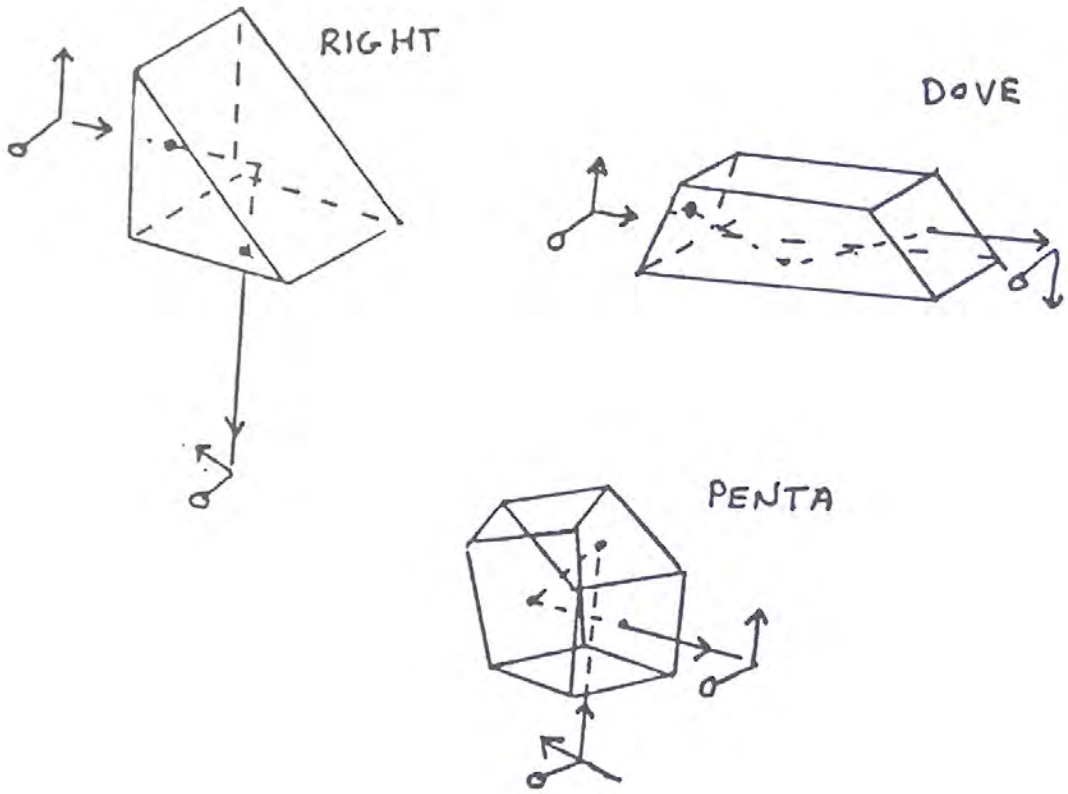


FIGURE 5.3.4. Imaging properties of common prisms.

Because the angle of deviation depends on refractive index, which in turn depends on wavelength, prisms cause multicoloured incident light to exit with an angular spread, with the "bluer" light having a larger deviation that the "redder" light. Such prisms are then referred to as *dispersing prisms*. It was this property of prisms that Newton used to demonstrate that white light actually consisted of many colours. An analytical instrument employing a prism as a dispersive element, together with the means of measuring the prism angle and the angles of deviation of various wavelength components is called a *prism spectrometer*. Its essential elements are shown in Figure 5.3.3. The resolving power of a prism is a complex function of the geometry and dispersion characteristics of the particular prism. The analysis won't be given here, but the minimum resolvable wavelength difference in the mid-visible region of the spectrum is about 0.1 nm.

There are also a wide variety of prisms which are used because of their total internal reflection properties. They are used for imaging purposes. Some of the more common types are shown in figure 5.3.4.

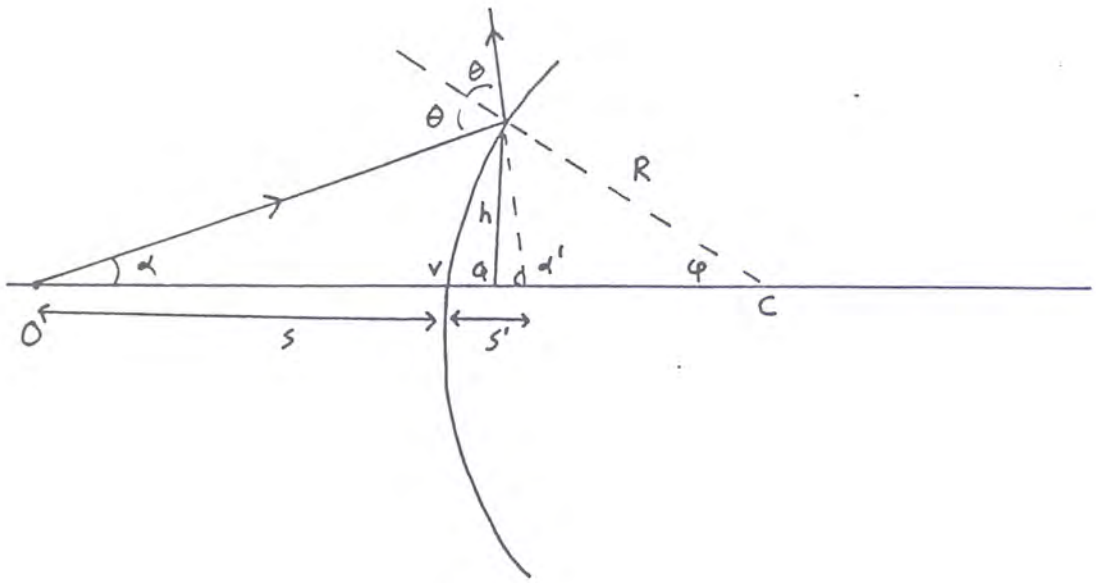


FIGURE 5.4.1. Reflection at a spherical surface.

The right angle, or *Porro prism*, causes a 90° deviation in a ray by total internal reflection and can be used in place of a mirror in many applications.

5.4. Reflection at a Curved Interface: Spherical Mirrors

We now discuss reflection and refraction at a curved surface. To begin we need to establish a sign convention for the curvature of optical surfaces and for distances. Different books on optics use different sign conventions. Any of these, including the one used here, are usually impossible to remember, and it is best to make a quick "sensibility check" of the formulae you use, by the use of a simple diagram that indicates qualitatively what occurs.

Figure 5.4.1 shows a *spherical mirror*, positioned with its centre of curvature at C. The line of cylindrical symmetry in the diagram is the *optical axis*, with the object point at O. The optical axis intercepts the optical surface at the point V , which is called the vertex.

The sign convention we use for distances and curvatures assumes that horizontal displacements associated with virtual rays are negative, and those associated with real rays are positive. The curvature of optical surfaces is signed according to the displacement of the centre of curvature from the vertex. If the centre of curvature lies beyond the vertex with respect to the incoming ray, the curvature is positive. In Fig. 5.4.1 the *convex surface* has positive curvature.

The law that governs the direction of the reflected ray is simply the law of reflection. For a ray that strikes the mirror on axis the reflected ray is returned back along the axis. Otherwise, the direction of the reflected ray is determined by the orientation of the local normal to the surface. Tracing back through the spherical surface we see that the image ray always appears to emanate from a point inside the spherical surface but on axis; it is therefore associated with a virtual image.

Consider the distance from the mirror to the object point to be s (positive) and the distance from the mirror to the image point in fig.5.4.1 to be negative. Our objective is to determine the relationship between these distances. Extension of such a simple relation would give us the general imaging properties of a spherical mirror. We obtain this relation only in the paraxial approximation in which the angles α and ϕ are assumed to be small so that

$$\sin\alpha \approx \tan\alpha \approx \alpha$$

$$\cos\alpha \approx 1.$$

For such an approximation, as we shall see, the relation between s and s' only depends on the radius of curvature, R, of the mirror.

This is also referred to as *first-order* or *Gaussian optics*. Referring to the figure and using the fact that the exterior angle of a triangle equals the sum of its opposite interior angles, we have that

$$\theta = \alpha + \phi \qquad 2\theta = \alpha + \alpha'$$

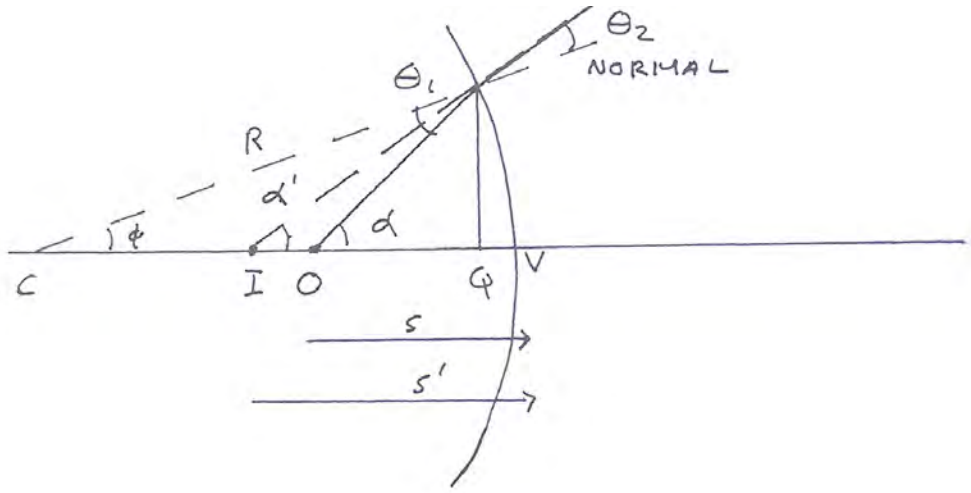


FIGURE 5.5.1. Refraction at a spherical surface.

which yields

$$\alpha - \alpha' = -2\phi.$$

We can use the small-angle approximation in which angles can be replaced by their tangents ($\alpha = h/s$, $\alpha' = -h/s'$, $\phi = h/R$) to write

$$\frac{h}{s} + \frac{h}{s'} = \frac{-2h}{R}$$

where we have also neglected the distance VQ in the figure, as being small if the angles are small. Cancelling h, and using our sign convention for s, s' and R produces the relation

$$\frac{1}{s} + \frac{1}{s'} = \frac{-2}{R}.$$

If the ray were incident on a concave spherical surface instead, the center of curvature would be to the left of the mirror and $R < 0$. In this case, for certain positions of the object point, it is possible to find a real image point also to the left of the mirror. Then s' would be positive.

The spherical mirror becomes a plane mirror if $R \to \infty$. In this case $s = -s'$ as expected. The imaging equation also shows the image and object distances entering the equation symmetrically, illustrating the interchangeability of these points as conjugate points. For an (extended) object at infinity, the incident rays are parallel and the image is a single point located at the *focal point*. The image is virtual and the corresponding distance, f, is negative if $R > 0$ with

$$f = \frac{-R}{2}.$$

With this identification, the mirror equation takes the form

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}$$

The lateral magnification associated with a spherical mirror is easily derived from simple geometry as

$$M = -\frac{s'}{s}$$

The details of the image can be found by doing *ray tracing* for different rays emerging from different parts of the object and striking the mirror at different points.

5.5. Refraction at a Curved Interface

The treatment of the transmission properties of a spherical interface that separates two media of different refractive index follows along the same lines as the previous section except that the emerging rays obey the law of refraction. Figure 5.5.1 shows such an interface of radius of curvature $R > 0$ separating a medium of refractive index n_1 from a medium of refractive index n_2. Consider an object ray emerging from a point O a distance s from the interface and passing into the medium of refractive index n_2. A ray which passes along the axis continues to do so on emerging from the interface and passes without deviation.

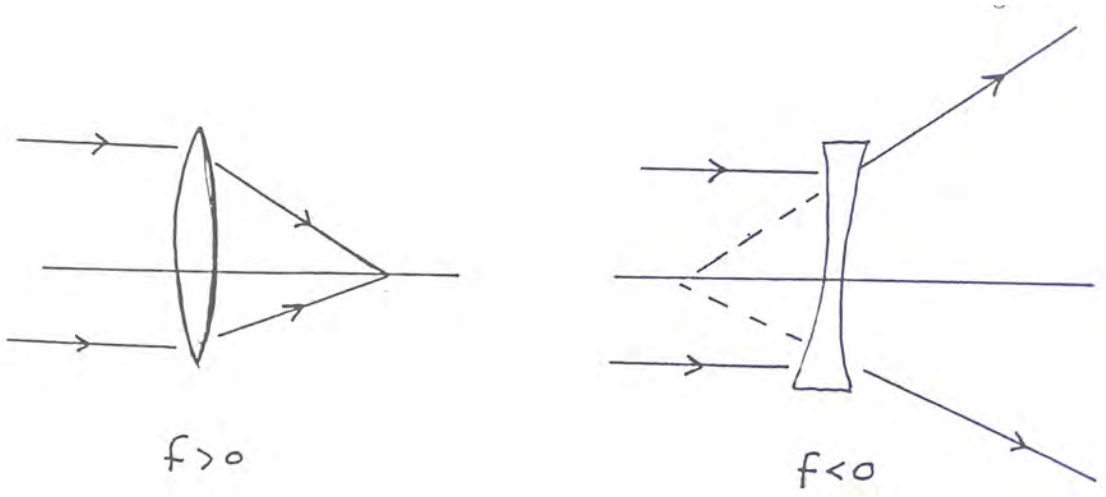


FIGURE 5.6.1. Action of thin lenses.

For a ray striking the interface at an arbitrary angle θ_1 with respect to the local normal we have that

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

The ray at angle θ_1 and the axial ray meet at their common intersection or image point I located at a distance s' from the interface. From the figure we see that

$$\theta_1 = \alpha - \phi \qquad \theta_1 - \theta_2 = \alpha' + \alpha \quad \text{so that } \theta_2 = -\phi + \alpha'$$

For paraxial rays in which the small angle approximations and sign convention of the last section are appropriate we have that

$$n_1(\alpha - \phi) = n_2(-\phi + \alpha')$$

and substituting the tangents for their angles while neglecting the distance QV

$$n_1(\frac{h}{s} + \frac{h}{R}) = n_2(\frac{h}{R} - \frac{h}{s'})$$

or

$$\frac{n_1}{s} + \frac{n_2}{s'} = \frac{n_2 - n_1}{R}$$

which holds equally well for convex or concave surfaces.

In this formula the quantity

$$\Phi = \frac{n_2 - n_1}{R} \tag{5.5.1}$$

is called the *power* of the surface. When R is measured in meters, Φ has the unit *diopter*.

When $R \to \infty$ the spherical surface becomes a plane refracting surface, and

$$s' = -\frac{n_2}{n_1} s.$$

If the object were buried in the dielectric with index n_2, then s refers to the object ray leaving the surface, and s' is the apparent depth of the virtual image of the object. The lateral magnification of the image is simply determined to be

$$M = -\frac{s'}{s} = \frac{n_2}{n_1}$$

perhaps the basis of the fisherman's honest exaggeration of "the one that got away".

5.6. Thin Lenses

Equation 5.5.1 can be applied successively to treat refraction through a series of spherical surfaces. In the situation where we have two such interfaces containing a medium of refractive index n_2 and bounded on either side by a medium of refractive index n_1, the element is called a *lens*. If the thickness of such a lens is small compared to object and image distances one refers to the lens as a *thin lens*, for which the ray transformation properties are particularly simple. Such a lens is shown in figure 5.6.1 At the first refracting surface of radius of curvature R_1 we

have

$$\frac{n_1}{s_1} + \frac{n_2}{s_1'} = \frac{n_2 - n_1}{R_1}$$

and at the second interface of radius of curvature R_2 we have that

$$\frac{n_1}{s_2} + \frac{n_2}{s_2'} = \frac{n_2 - n_1}{R_1}.$$

In the thin lens approximation whereby we neglect the thickness of the lens we have

$$s_2 = -s_1'$$

because a real intercept from the first optical surface is a virtual one in the context of the second surface, and *vice versa*. Combining the last three equations we have that

$$\frac{n_1}{s_1} + \frac{n_1}{s_2'} = n_2 - n_1 \left(\frac{1}{R_1} - \frac{1}{R_2} \right).$$

But s_1 is the original object distance and s_2' is the image distance so we can drop the subscripts to obtain

$$\frac{n_1}{s} + \frac{n_1}{s'} = \frac{n_2 - n_1}{n_1} \left(\frac{1}{R_1} - \frac{1}{R_2} \right).$$

The focal length of the lens, *f,* is defined as the image distance for an object located at infinity, giving

$$\frac{1}{f} = \frac{n_2 - n_1}{n_1} \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \tag{5.6.1}$$

Equation 5.6.1is called the *lens-makers equation* because it predicts the focal length in terms of the surface curvatures and the refractive index of the lens. In terms of the focal length the thin lens formula becomes

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}.$$

In general, wavefront analysis indicates that lenses which are thicker in the middle (regardless of the individual radius of curvature) than the edges cause convergence of incident parallel rays while in the opposite case rays diverge.

Similar to previous formulas developed, the magnification for a thin lens is given by

$$M = -\frac{s'}{s}$$

5.7. Lens Aberrations

It is worthwhile reminding ourselves that the very simple formulas developed in the previous sections were based on certain approximations and idealizations. The most significant approximation is the paraxial approximation (which for lenses is equivalent to the thin lens approximation since the image and object distances are large compared to the thickness). Beyond this approximation, if one is interested in the imaging properties of an actual system, one has to be prepared to become engaged in extensive ray tracing.

The most common aberrations of systems and an explanation of their origins will now be given in the cases of lenses, which are the most common optical elements used in imaging systems. A mathematical treatment can be developed by expanding the sine and tangent terms above. Different aberrations correspond to various terms in the expansion which represent deviation from paraxial, ideal ray theory. The different aberrations are illustrated in figure 5.7.1.

5.7.1. Spherical Aberration. This effect is related to rays which make large angles relative to the optical axis of the system, and mathematically can be shown to arise from the fact that a mirror has a spherical surface and not a parabolic surface. Rays making significantly different angles with respect to the optical axis are brought to a different focus.

5.7.2. Coma. This aberration is an off-axis effect which appears when a bundle of incident rays all make the same angle with respect to the optical axis (source at ∞) but are brought to a focus at different points on the focal plane.

5.7.3. Chromatic Aberration. Because the focal length of a lens depends on the refractive index and this in turn depends on wavelength, $n = n(\lambda)$, light of different colours emanating from an object come to a focus at different points. A white object therefore does not give rise to a white image. Rather it is distorted and has rainbow edges.

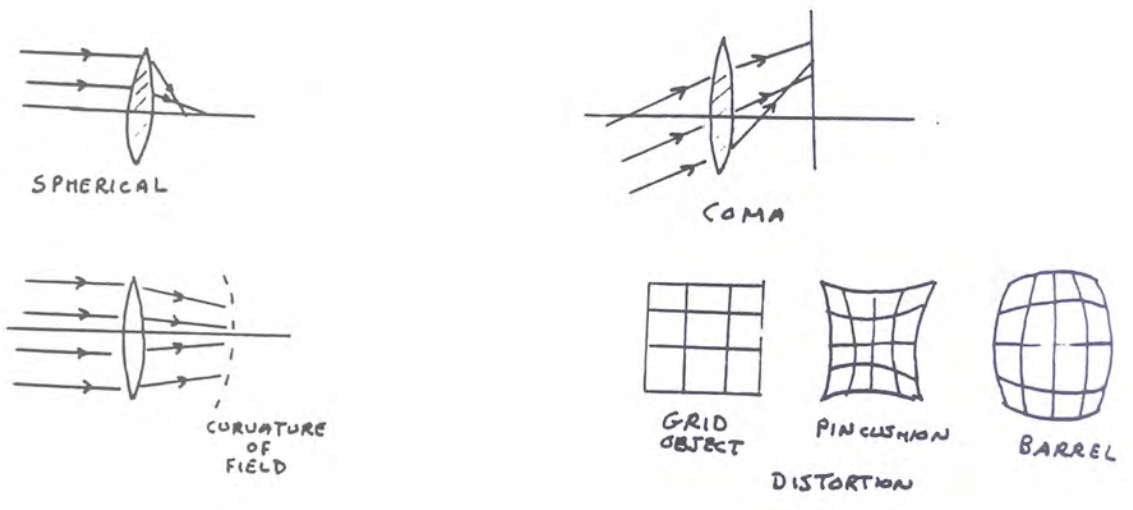


FIGURE 5.7.1. Aberrations of thin lenses.

5.7.4. Astigmatism and curvature of the field. This is an effect which yields images which have been elliptically distorted. This can arise from elliptical distortion in the manufacture of the lens, or even in the case of an ideal spherical surface, the imaging of an off-axis bundle of rays.

5.7.5. Distortion. This aberration shows up as a variation in the lateral magnification of an object. The two main types barrel and pincushion distortion are shown in the figure. It is due to variations in object to image distances across the image.

References

1) F.L. Pedrotti and L.S. Pedrotti, *Introduction to Optics*, Prentice Hall, 1987
2) J.R. Meyer-Arendt, *Introduction to Classical and Modern Optics*, Prentice Hall, 1984.

Problems

1. Under what circumstances will an object placed in front of a concave mirror produce a real image?

2. Two plane mirrors subtend a certain angle with each other. A ray of light is parallel to one of the mirrors and after four reflections it exactly retraces its path. What angle do the mirrors subtend?

3. Two thin lenses of focal length f_1 and f_2 are separated by a distance d. What is the effective length of the lens combination? Are there any restrictions on your formula? What happens if $d \ll f_1, f_2$?

4. A meter stick lies along the optical axis of a convex mirror of focal length 50 cm, with its nearer end 60 cm from the mirror surface. How long is the image of the metre stick?

5. Show that the lateral displacement s of a ray of light penetrating a rectangular plate of thickness t is given by

$$s = \frac{t \sin(\theta_1 - \theta_2)}{\cos \theta_2}$$

where θ_1 and θ_2 are the angles of incidence and refraction, respectively. Find the displacement when $t = 3$ cm, n = 1.50 and $\theta_1 = 30°$.

CHAPTER 6

Matrix Methods in Paraxial Optics

Our life is frittered away by detail...
Simplify, simplify
H.D. Thoreau

The previous chapter has indicated the usefulness of ray optics in analyzing the imaging properties of systems. Within the paraxial approximation particularly simple results were obtained for distinct optical elements which were homogeneous in a plane perpendicular to the optical axis. The problem of treating rays which pass through several optical elements could in principle be carried out with much algebraic manipulation. In this chapter we introduce a matrix method which accomplishes the same task but in a much more straightforward fashion. Transformation of rays on passage through a complex optical system then simply reduces to the multiplication of matrices associated with each optical element. At the end of the chapter we consider applying these methods to telescopes and microscopes.

6.1. Optical Rays and Transformations

Recall that a ray is defined, in the limit of geometrical optics, where $\lambda \to 0$, to be a beam of light of infinitesimal transverse extent. A ray, as a line in space, is completely defined by its distance from a given axis and the slope of that line relative to the axis. In the paraxial approximation, the slope is equivalent to the angle the ray makes with the optic axis. This is indicated in figure 6.1.1.

In the *paraxial approximation* the slope is equivalent to the angle between the ray and the axis. If we take the axis to be the z-axis, then the ray, at a certain reference plane $z = z_1$, is completely defined by $r(z_1)$which specifies the distance of the ray from the axis and $r'(z_1) = dr/dz|(z = z_1) = \theta(z_1)$. These together form what is called the *ray vector*

$$\vec{R} = \begin{bmatrix} r(z_1) \\ \left.\frac{dr}{dz}\right|_{z=z_1} = \theta(z_1) \end{bmatrix}.$$

Note that the two components of the vector do not have the same dimensionality.

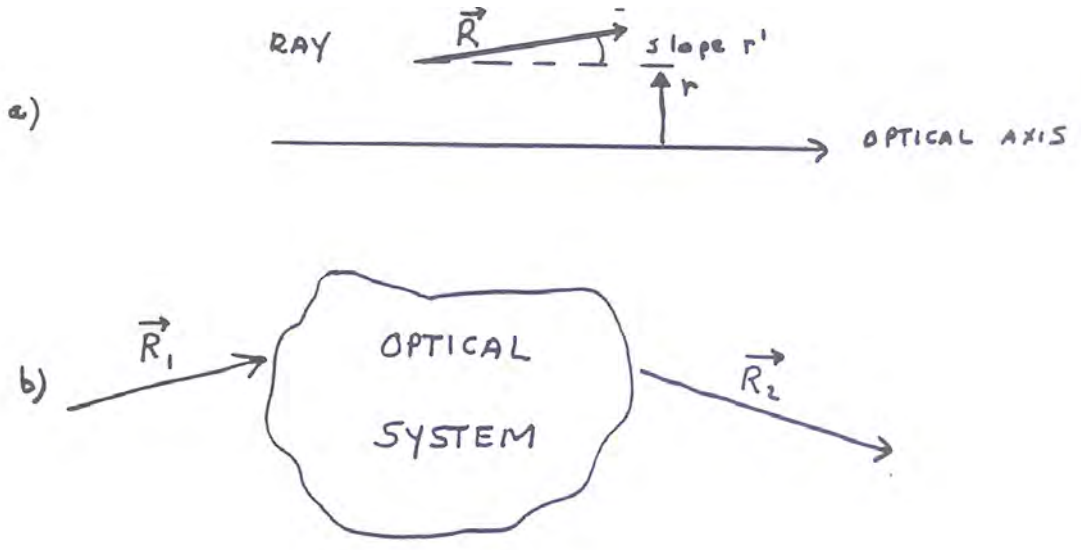


FIGURE 6.1.1. Transformation of a ray by an optical system.

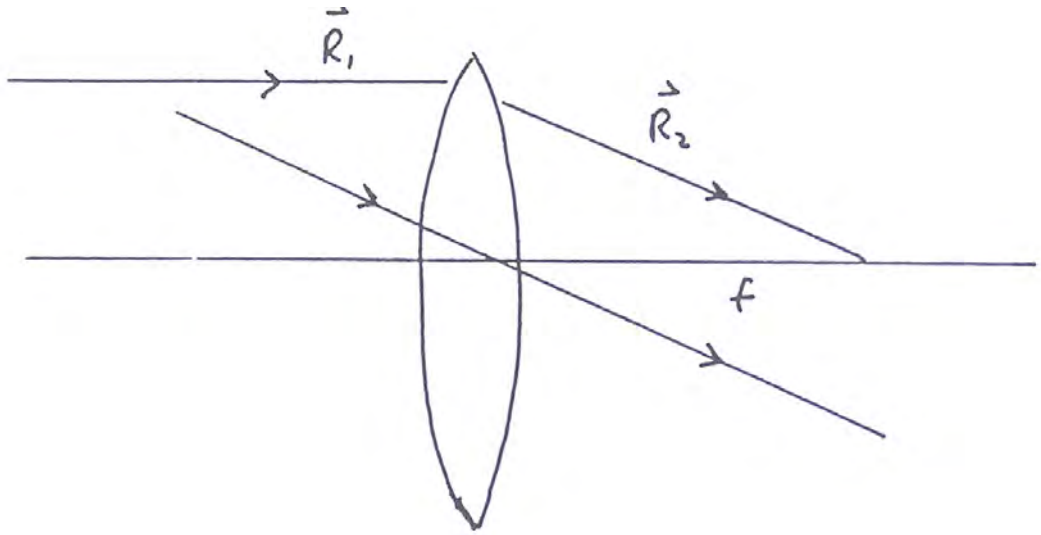


FIGURE 6.1.2. Ray transformation by a simple lens.

Upon passage through an optical system the ray vector is transformed into a new ray vector as indicated in figure 6.1.1. One expects, for a linear optical system, that there is a simple functional relationship between the incident and emerging ray vector components so that if $\vec{R}_1$ and $\vec{R}_2$ are the vectors describing the incident and emerging rays we would have

$$r_2 = f(r_1, \theta_1) \quad \text{and} \quad \theta_2 = g(r_1, \theta_1)$$

where f and g are two undetermined functions. For an imaging system (one which does not cause any transverse distortion in an incident light distribution) the transformation laws must be linear so that

$$r_2 = Ar_1 + B\theta_1$$
$$\theta_2 = Cr_1 + D\theta_1$$

or

$$\vec{R}_2 = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \vec{R}_1 \tag{6.1.1}$$

where the A , B, C and D are parameters to be determined for a particular system. They define a so-called *ABCD matri*x.

For a thin lens of focal length f we can determine the transformation matrix as follows. With respect to figure 6.1.2 it is clear that the transformation matrix must satisfy the following conditions:

1) For a thin lens we must have that $r_1 = r_2$ for all θ_1. This implies $A = 1$ and $B = 0$.

2) For a ray which passes through the center of the lens ($r_1 = 0$) we must have that the slope (angle) doesn't change so that $D = 1$.

3) An incident ray which is parallel to the axis of the lens must pass through the focal point, by definition of the focal point. Hence for $\theta_1 = 0$, as seen in the diagram, we must have

$$\theta_2' = -\frac{r_1}{f}$$

or

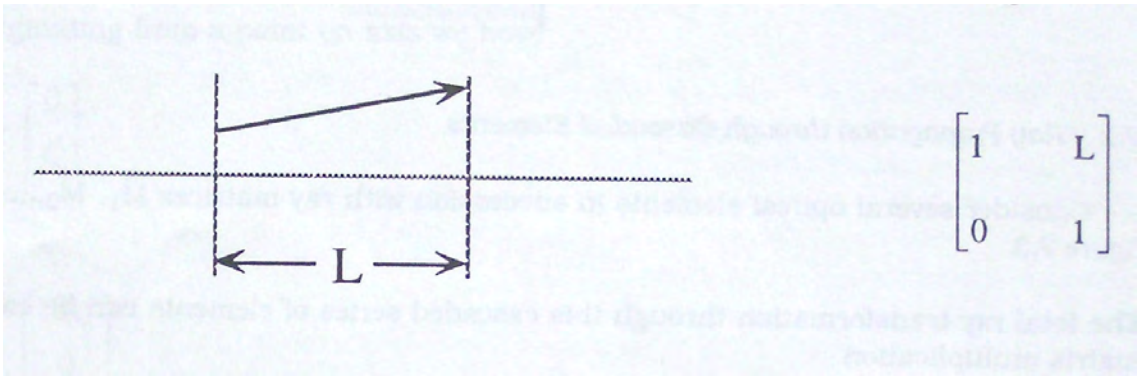
$$C = -\frac{1}{f}.$$

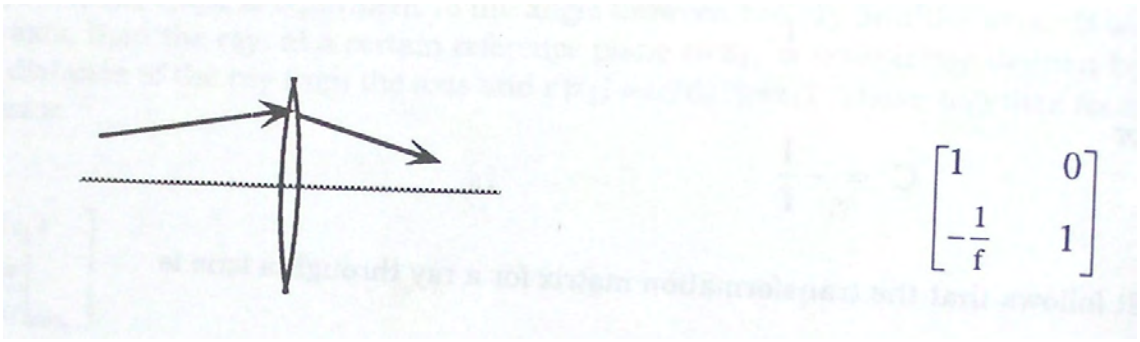
It follows that the *transformation matrix* for a ray through a lens is

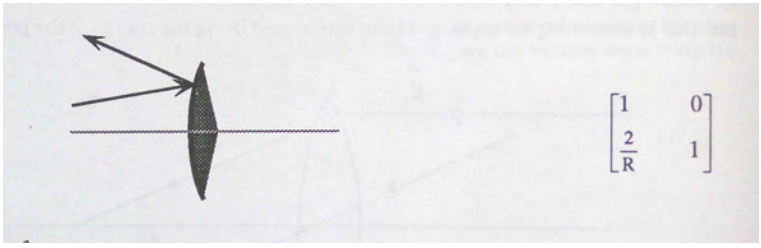
$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -f^{-1} & 1 \end{bmatrix}.$$

But this is something we could also have determined using the results of the previous chapter.

The change in displacement and slope of an optical ray upon passing through a wide variety of simple optical elements can be written in the same general form as equation 6.1.1. The matrix derived for a particular optical element is known as the *ray matrix* for the element. Most of these are derived simply by considering the results of







the previous chapter. In all cases the matrices have a determinant which is equal to n_1/n_2 where n_1 and n_2 are the refractive indices at the input and output planes.

We now proceed to list all the common ray matrices within the paraxial approximation.

a) Free space propagation (virtual rays propagate with negative distances, not to be confused with left & right):

b) Thin lens, focal length *f:*

c) Spherical mirror, radius R, (recall that virtual rays subsequently propagate with negative distance); $R > 0$ for convex incidence:

d) Curved dielectric interface: $R > 0$ for convex-surface incidence:

e) Refraction at a plane interface (from medium n_1 to medium n_2):

6.2. Ray Propagation Through Cascaded Elements

Consider several optical elements in succession with ray matrices $\overleftrightarrow{M}_1, \overleftrightarrow{M}_2, \ldots \overleftrightarrow{M}_n$.. .as shown in figure 6.2.1.

The total ray transformation through this cascaded series of elements can be calculated from simple matrix multiplication

$$\vec{R}_2 = \overleftrightarrow{M}_2 \cdot \overrightarrow{R}_1 = \overleftrightarrow{M}_2 \cdot \overleftrightarrow{M}_1 \cdot \vec{R}_0$$

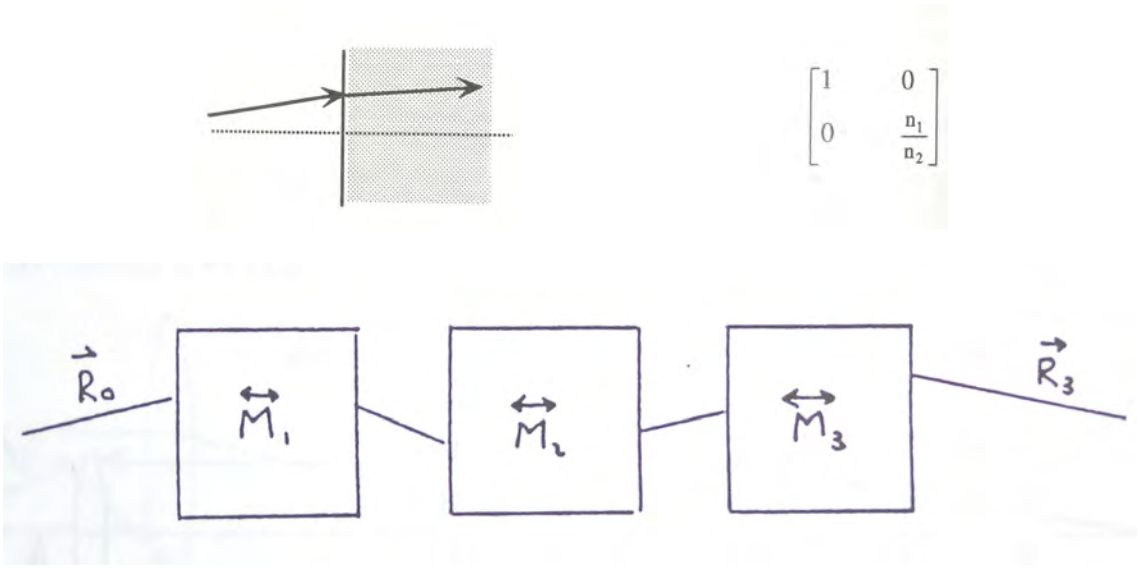


FIGURE 6.2.1. Ray matrix systems in cascade.

etc. so that the general result is

$$\vec{R}_n = \overleftrightarrow{M}_n \cdot \overleftrightarrow{M}_{n-1} ... \overleftrightarrow{M}_1 \cdot \vec{R}_1 = \overleftrightarrow{M}_{tot} \cdot \vec{R}_0$$

and the overall ray transformation matrix is given by

$$\overleftrightarrow{M}_{tot} = \overleftrightarrow{M}_n \cdot \overleftrightarrow{M}_{n-1} ... \overleftrightarrow{M}_1.$$

Note the order in which the matrices are multiplied and how this is related to the order in which the ray encounters the different optical elements.

As an illustration of how the matrix method works let us consider the following problem. A point source is located on axis 20 cm in front of a pair of lenses with focal length of 10 and -10 cm respectively and separated by 10 cm. Does the lens system lead to the formation of a real image on the far side of the second lens?

For a real image to form light must pass through the axis on the far side of the second lens for an arbitrary initial ray. This means it would have to be possible to find a location where $r = 0$ after the second lens. If it occurs let us call this distance z behind the second lens. The overall transformation matrix for the problem is seen to be

$$\overleftrightarrow{M}_{tot} = \begin{bmatrix} 1 & z \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0.1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 10 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ -0.1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 20 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -0.1z & 10 \\ -0.1 & 0 \end{bmatrix}.$$

For an initial ray originating from a point on axis we have

$$\vec{R}_0 = \begin{bmatrix} 0 \\ \theta_0 \end{bmatrix}$$

and

$$\vec{R}_f = \begin{bmatrix} 10r'_0 \\ 0 \end{bmatrix}.$$

Since this ray is independent of z and never has an $r = 0$, we conclude that it does not correspond to a real image (only the ray launched directly along the axis ($\theta_0 = 0$) remains on axis.

6.3. Telescopes and Microscopes

We close this short chapter with a discussion of two very important optical systems, the microscope and the telescope. A telescope is an instrument which images and magnifies objects far away while a microscope magnifies objects which are very small.

An astronomical *telescope*, for example, focuses for infinity and typically has a large light gathering optical element which may be a mirror or lens. The size of this element determines the brightness of the image since its area directly determines the amount of light gathered. As we shall see later in the chapter on Fraunhofer diffraction, the size of the "objective" also determines the ability of the telescope to resolve closely spaced distant objects. Figure 6.3.1illustrates a specific type of telescope which is known as the Kepler astronomical refractor.

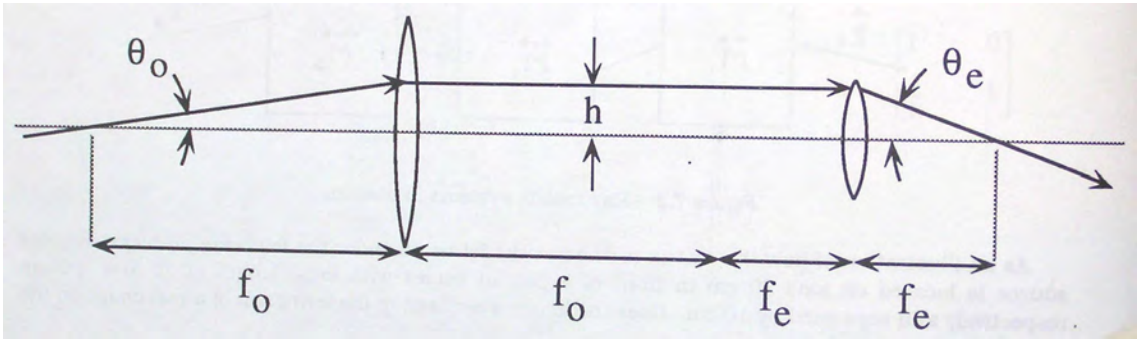


FIGURE 6.3.1. Astronomical telescope

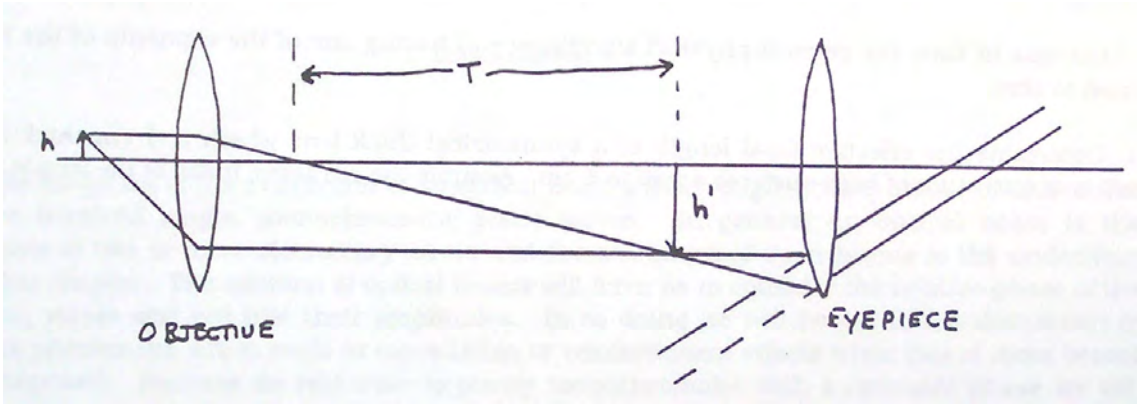


FIGURE 6.3.2. Magnification of a compound microscope.

The lenses are spaced so that the second focus of the first lens coincides with the first focus of the second lens. One of the most important characteristics of a telescope is its magnification which can be considered to be the ratio of the angular size of the image to the angular size of the object. The magnification can be determined easily using ray matrix techniques but can be determined more trivially as follows. From the figure we see that this *magnification* is given as

$$M = \frac{\theta_e}{\theta_o} = \frac{\left(\frac{h}{f_e}\right)}{\left(\frac{h}{f_o}\right)} = \frac{f_o}{f_e}.$$

We now turn our attention to microscopes. One can form a *simple microscope* from one lens of particularly short focal length but we shall not discuss this trivial case here. *Compound microscopes* like telescopes come in many different configurations and in simplest form, like a telescope, they consist of two lenses. The big difference, of course, is that telescopes try to magnify things far away while microscopes try to magnify close objects. Figure 6.3.2 shows the typical lens configuration for a compound microscope.

It is seen that the objective has a very short focal length so as to generate a large image in the focal plane of the second lens which is located a "tube length", T, away from the focal plane of the objective lens. The *magnification* of the object by the objective is easily seen to be

$$M_o = \frac{T}{f_o}.$$

The eyepiece acts as a magnifier. Since for most humans the distance of most distinct vision is 25 cm, eyepieces are designed so as to yield a virtual image 25 cm away from the eye. The magnification of the eyepiece is therefore approximately given by $25/f_e$ giving as a magnification for the whole system

$$M = \frac{25T}{f_e f_o}.$$

Microscope eyepieces are usually designated by their magnification but their focal length can be determined from the formula above.

References

J.R. Meyer-Arendt, *Introduction to Classical and Modern Optics*, Prentice-Hall, 1984.
A.E. Siegman, *Lasers*, Oxford, 1987.

Problems

1. a) Show that the transformation matrix associated with passage through a spherical dielectric interface which separates a medium of refractive index n_1 from one of refractive index n_2 is

$$\begin{bmatrix} 1 & 0 \\ \frac{n_1-n_2}{n_2 R} & \frac{n_1}{n_2} \end{bmatrix}$$

where R is the radius of curvature of the surface.

b) Show that the focal length of a thin lens made of glass of refractive index n and with surfaces possessing radius of curvature R_1 and R_2 is given by

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

if the lens is in vacuum.

2. Using the ABCD matrices determine the (in focus) transverse magnification of a telescope which has an objective of focal length f_1 separated from an eyepiece of focal length f_2.

3. An optical system when traversed by an optical ray in one direction has a certain ABCD matrix. What is the corresponding matrix if the ray enters from the opposite direction?

4. A lens guide consists of a large number of identical lenses of focal length f separated from each other by a distance d. What must the relation between d and f be such that a ray, launched into the lens system at a small angle be confined to the lens system?

5. Discuss in turn the general physical significance of having one of the elements of the ray matrix equal to zero.

6. Determine the effective focal length of a symmetrical thick lens which is 1 cm thick and has a radius of curvature of both surfaces equal to 5 cm. Assume the refractive index of the lens is 1.5.

Part 3

Wave Optics

CHAPTER 7

Superposition of Optical Waves

There was no "One, two, three and away", but they began running when they liked and left off when they liked....

Lewis Carroll

With the exception of the interaction of an optical beam with a rough surface all our efforts to this point have involved single, monochromatic plane waves. In general an optical beam is the superposition of two or more elementary waves and the treatment of such beams is the underlying theme of this chapter. The addition of optical beams forces us to consider the relative phase of the contributing waves and not just their amplitudes. In so doing we are led into a discussion of interference phenomena which leads to cancellation or reinforcement effects when two or more beams are superimposed. Because no real wave is purely monochromatic with a constant phase we introduce the concept of coherence, which describes the phase constancy or monochromaticity in a beam. The outline of this chapter is as follows. We introduce the basic ideas in interference phenomena after which the concepts of spatial and temporal coherence of optical beams are discussed. Young, Michelson and Fabry-Perot interferometers are introduced as instruments for measuring the temporal or frequency characteristics of optical beams. Finally we consider multiple beam interference in thin film optics.

7.1. Interference of Two Beams

In general, the total optical field associated with multiple beams at a given point in space is related to the amplitudes of the individual beams. In the approximation of linear optics we can use the superposition principle to determine the total field. If needed, the total instantaneous intensity is then determined by the modulus squared of the total electric field at that point. If, for example, $\vec{E}_1$, $\vec{E}_2$, $\vec{E}_3$, ...etc. are the fields associated with different waves at a point $\vec{r}$, the total field at that point, $\vec{E}_T$, is given by

$$\vec{E}_T = \vec{E}_1 + \vec{E}_2 + \vec{E}_3 + ... \tag{7.1.1}$$

and the intensity, $I(\vec{r})$, is found from

$$I = C\left|\vec{E}_T\right|^2 \tag{7.1.2}$$

where the constant C is determined by the use of Chapter 2. In what follows, since we are not primarily interested in the magnitude of I, we drop this constant for simplicity and identify the intensity with the squared modulus of the total electric field vector. Equations 7.1.1 and 7.1.2 give the basic principle which is used throughout this chapter to deal with multiple wave beams which may be discrete sums of a finite number of beams or integral sums of an infinite number of beams.

Let us begin our discussion of the treatment of multiple beams by considering the simplest possible case, the superposition of two waves. In so doing we arrive at the salient features of interference phenomena. Consider two plane, harmonic, linearly polarized waves of the same frequency represented by

$$\vec{\mathcal{E}}_1 = \vec{E}_{01}e^{i(\vec{k}_1\cdot\vec{r}-\omega t+\phi_{01})} \tag{7.1.3}$$

and

$$\vec{\mathcal{E}}_2 = \vec{E}_{02}e^{i(\vec{k}_2\cdot\vec{r}-\omega t+\phi_{02})}. \tag{7.1.4}$$

If $(\phi_{01}-\phi_{02})$ is a constant (independent of space and time), the sources and resulting waves are said to be *mutually coherent*. We shall assume this in the beginning but relax it later on. Under the present circumstances, then, we have that the total instantaneous intensity for the two waves is

$$I(\vec{r}) = \left|\vec{E}_T\right|^2 = \vec{E}_T\cdot\vec{E}_T^* = |E_{01}|^2 + |E_{02}|^2 + 2\vec{E}_{01}\cdot\vec{E}_{02}cos\delta$$

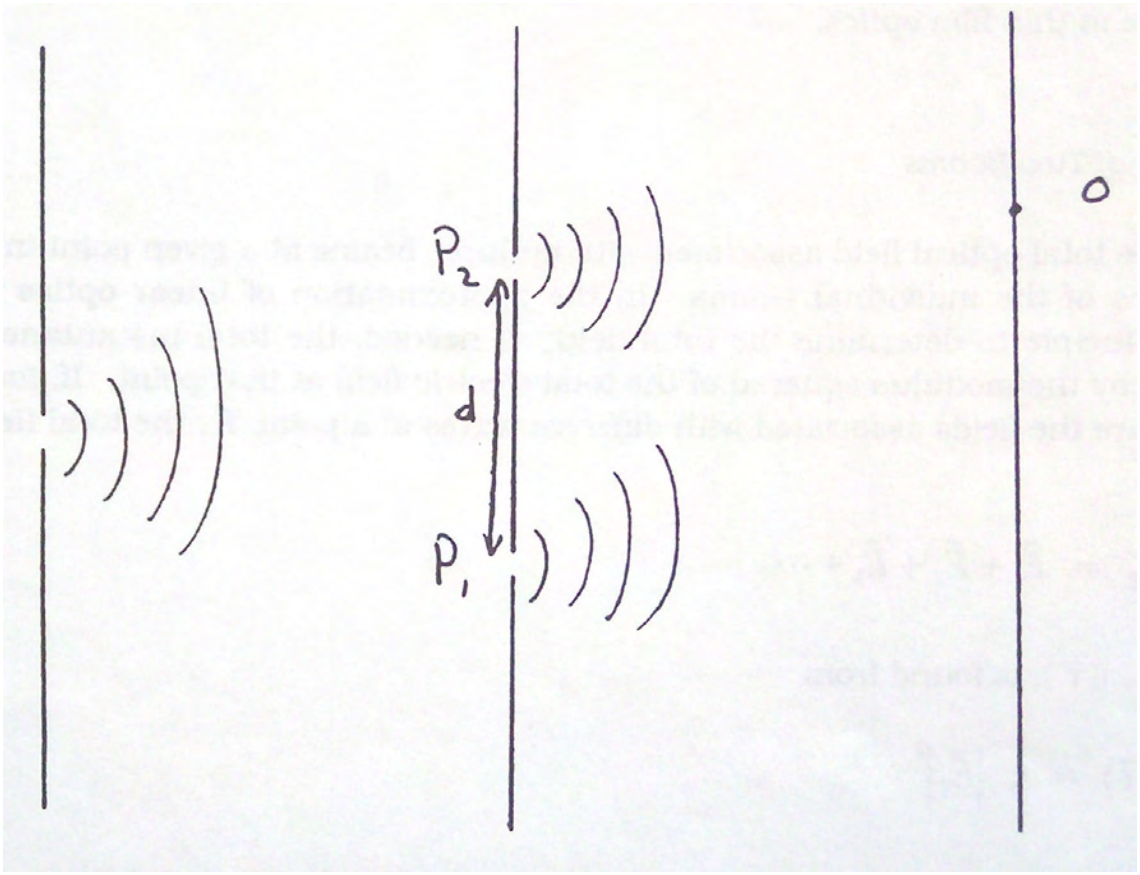


FIGURE 7.1.1. Young's experiment.

where

$$\delta = \left(\vec{k}_1 - \vec{k}_2\right) \cdot \vec{r} + \phi_{01} - \phi_{02}.$$

The term $2\vec{E}_{01} \cdot \vec{E}_{02} \cos\delta$ is referred to as the *interference term*. Its value depends on the relative phase of the two monochromatic waves; it may give a positive or negative contribution to the overall intensity and gives rise to the interference fringes. Note that the interference term depends on the polarization state of the two beams. If the two waves have orthogonal polarization states the interference term is zero.

The classic experiment which demonstrated the interference of light and which was used to argue that light was therefore a wave was performed by Thomas Young in 1802. It is illustrated schematically in figure 7.1.1. In his demonstration Young allowed colour filtered sunlight to pass through a pinhole, P, and the diverging beam from this (nearly) point source was used to illuminate two pinholes, P_1 and P_2, in a screen separated by a distance d. For simplicity we assume that the pinholes are illuminated by a linearly polarized, but not necessarily coherent source of fundamental frequency ω . The two waves which emerge from the secondary pinholes are mutually coherent since they are derived from the same source and indeed from the same wavefront striking the screen.

To view the effects of interference one can choose an observation plane which is located at a large distance from the plane containing P_1 and P_2. Although, strictly speaking, point apertures give rise to spherical waves, if the observation screen is sufficiently far from the secondary sources, the waves striking the observation screen can be approximated as plane waves. This corresponds to what is known as the far field approximation. On the observation plane the two waves have the form represented by equations 7.1.3 and 7.1.4 where, in this case, $\vec{k}_1$ and $\vec{k}_2$ are the propagation wave vectors from points P_1 and P_2 respectively to the observation point, O, and $\vec{r}$ is the position vector of a point on the observation plane. The geometry of the waves is shown in figure 7.1.2.

From the figure, and in particular with the choice of x- and y- axes, it can be seen that the difference in the propagation vectors is given approximately by

$$\vec{k}_1 - \vec{k}_2 = \hat{y}\frac{kd}{x}$$

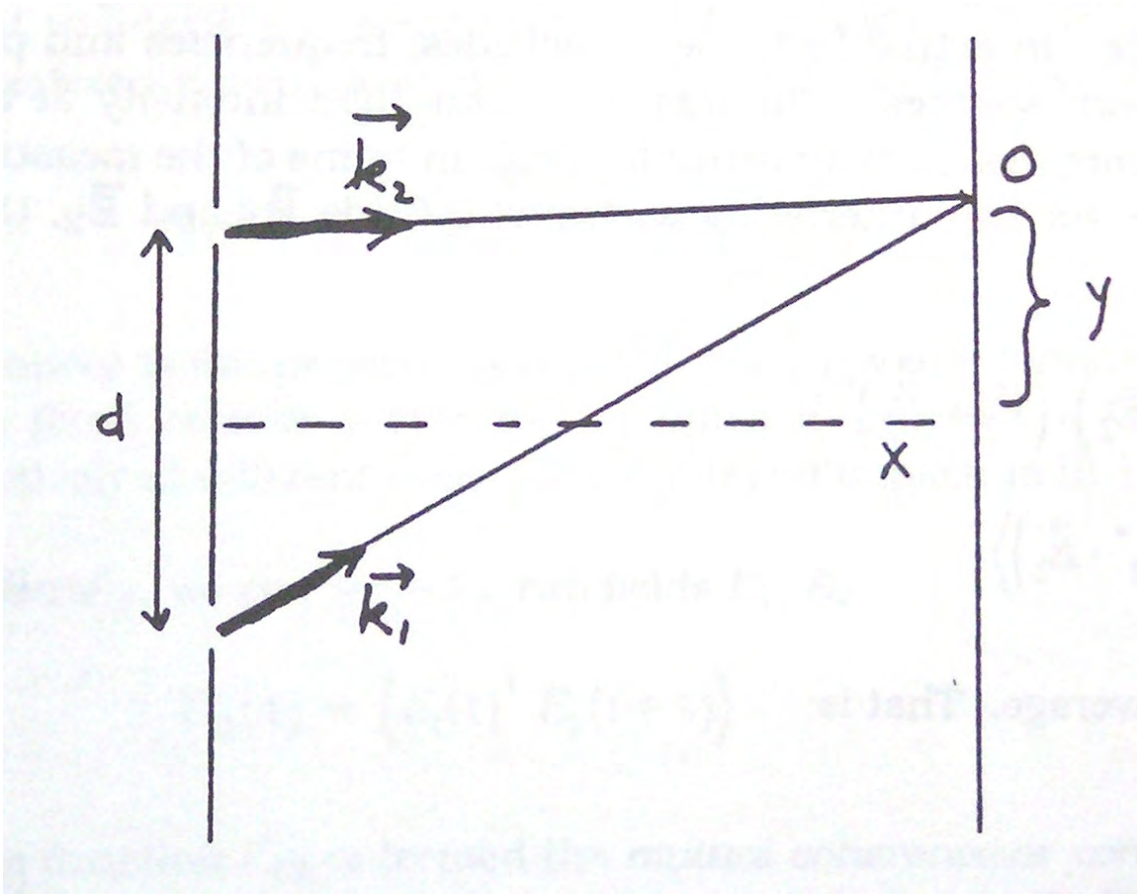


FIGURE 7.1.2. Geometry for analyzing interference in the case of dual point sources.

under the approximation that that the distance, x, from O to the plane of the apertures is large compared to y and d. In this case $\sin\theta \approx \theta \approx \frac{y}{x}$. With this approximation it follows that

$$\left(\vec{k}_1 - \vec{k}_2\right) \cdot \vec{r} = \left(\vec{k}_1 - \vec{k}_2\right) \cdot (x\hat{x} + y\hat{y}) = \frac{kyd}{x}.$$

The intensity distribution on the observation plane is then given by

$$I = |E_{01}|^2 + |E_{02}|^2 + 2\vec{E}_{01} \cdot \vec{E}_{02} \cos\left(\frac{kyd}{x} + \phi_{01} - \phi_{02}\right) = 2I_0\left(1 + \cos\left(\frac{kyd}{x}\right)\right)$$

where $I_o = |E_{o1}|^2 = |E_{o2}|^2$ for identical apertures and $\phi_{01} - \phi_{02}$ has been set to zero, since its only effect is to shift the origin of the y axis. The intensity therefore varies between 0 and $4I_0$, depending on the value of y. Interference maxima (bright fringes) occur for y = 0, $\lambda x/d$, $2\lambda x/d$, etc. If a phase retarder (which generates different polarization states) is placed over one of the apertures the range of intensity variation is not as great. If orthogonal polarization states are used, there is no intensity variation on the screen.

In Young's interference experiments it is observed that even for linearly polarized interfering waves, when the path difference between the two waves becomes large the range of intensity variations, or fringe contrast, approaches zero. This is an indication of the growing lack of coherence (phase stability) between the two waves. A more complete treatment of interference must therefore address the issue of the lack of perfect coherence, or the existence of *partial coherence*, between the two beams.

7.2. Partial Coherence

In the preceding discussion it was assumed that the optical fields were completely coherent, monochromatic, and constant in amplitude. In actual fact, the amplitudes, frequencies and phases oscillate in a random fashion for all known sources. The instantaneous light intensity at a given point therefore fluctuates rapidly. It therefore makes

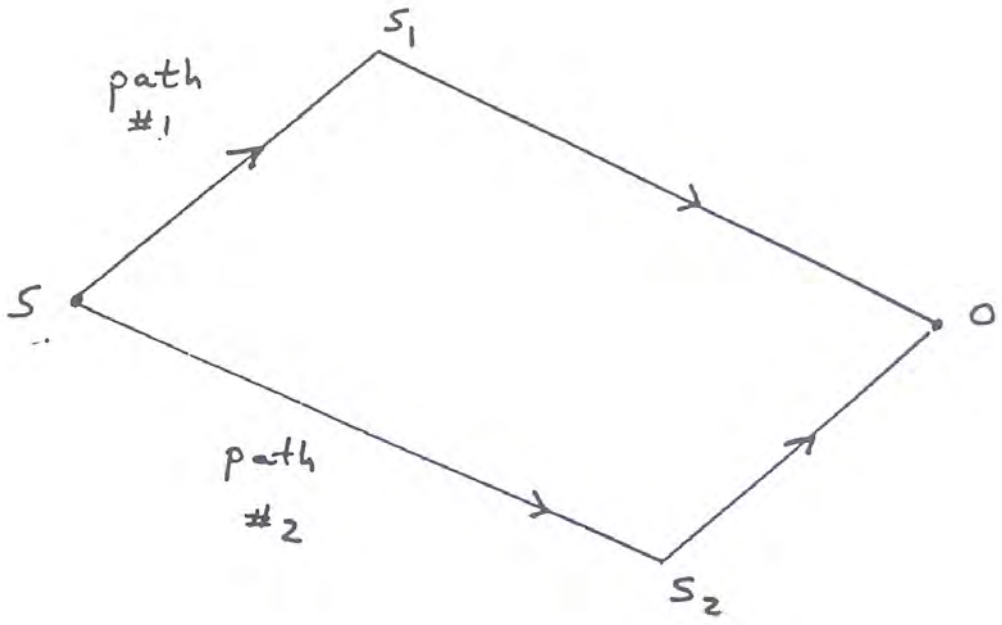


FIGURE 7.2.1. Generalized light paths in an interference experiment.

more sense to speak in terms of the measurement of a time-averaged intensity. In this case, for two interfering waves with fields $\vec{\mathcal{E}}_1$ and $\vec{\mathcal{E}}_2$, the time averaged intensity is given by

$$I = \left\langle \vec{\mathcal{E}}_T \cdot \vec{\mathcal{E}}_T^* \right\rangle = \left\langle \left(\vec{\mathcal{E}}_1 + \vec{\mathcal{E}}_1\right) \cdot \left(\vec{\mathcal{E}}_1 + \vec{\mathcal{E}}_2\right)^* \right\rangle = \left\langle \left|\vec{\mathcal{E}}_1\right|^2 + \left|\vec{\mathcal{E}}_1\right|^2 + 2Re\left(\vec{\mathcal{E}}_1^* \cdot \vec{\mathcal{E}}_2\right) \right\rangle$$

where the angle brackets indicate a time average. That is (as we used in Ch. 2)

$$\langle f \rangle = \lim_{T\to\infty} \frac{1}{T} \int_0^T f(t)dt.$$

We assume that the indicated time averages are *stationary*. By this we mean that the time averages are independent of choice of origin of time so that if we average over a sufficiently long time it doesn't matter when we begin the averaging process. We further assume that the polarization states of the two waves are identical so that we can ignore polarization effects in the discussion. We then have

$$I = I_1 + I_2 + \left\langle 2Re\left(\vec{\mathcal{E}}_1^* \cdot \vec{\mathcal{E}}_2\right) \right\rangle. \tag{7.2.1}$$

In most interference experiments the two interfering waves originate from the same source since otherwise their relative phase would be totally random in time, and interference would never be observed. The two fields in equation 7.2.1 differ, however, because, in general, they would have had to travel different paths to arrive at the point of interference. A typical schematic geometry for the two paths is shown in figure 7.2.1.

If the times to traverse the two path lengths differ by a time τ, then the two-fields as they are brought to interfere at some point had their origins at the same source but at different times τ apart. The interference term in equation 7.2.1 can be written as

$$2Re[\Gamma_{11}(\tau)] = 2Re\left[\langle \mathcal{E}_1^*(t) \cdot \mathcal{E}_1(t+\tau)\rangle\right]. \tag{7.2.2}$$

If the source is not perfectly periodic (*e.g.*, the field does not vary sinusoidal for all time), the fields $\mathcal{E}_1$ and $\mathcal{E}_2$ cannot have a fixed relative phase for all times t, except at $\tau = 0$. Therefore the fields combine constructively or destructively at different times, influence the value of the integral implicit in equation 7.2.2. As the phase fluctuates with increasing time, $\Gamma_{11}(\tau)$ decreases as τ increases.

Generally, for any two fields $\mathcal{E}_1$, $\mathcal{E}_2$ we can write

$$\Gamma_{12}(\tau) = \langle \mathcal{E}_1^*(t) \cdot \mathcal{E}_2(t+\tau)\rangle$$

and the function Γ_{12} is termed the *mutual coherence function* or (time averaged) *correlation function* of the two fields $\mathcal{E}_1$, and $\mathcal{E}_2$. From its definition we can see that $\Gamma_{11}(0) = I_1$ and $\Gamma_{22}(0) = I_2$. Since we are not usually interested in the absolute magnitude of the intensities of the interfering beams it is much more convenient to use a normalized correlation function which is defined as

$$\gamma_{12}(\tau) = \frac{\Gamma_{12}(\tau)}{\sqrt{\Gamma_{11}(0)\Gamma_{22}(0)}} = \frac{\Gamma_{12}(\tau)}{\sqrt{I_1 I_2}}.$$

It follows that the time-averaged intensity is given by

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2} Re\left(\gamma_{12}(\tau)\right).$$

The function $Re\left(\gamma_{12}(\tau)\right)$ is of the form $|\gamma(\tau)|\cos(\omega\tau)$ and therefore varies periodically with τ, with period related to $2\pi/\omega$, where ω is the frequency of the interfering beams. Therefore, one obtains an interference pattern if $|\gamma_{12}(\tau)|$ has a value other than zero. The function $|\gamma_{12}(\tau)|$ is a measure of the *degree of coherence* and conventionally one speaks of the following three cases in connection with coherence

$$|\gamma_{12}| = 1 \qquad \text{compete coherence } (e.g., \text{ infinite sinusoidal wave})$$
$$0 < |\gamma_{12}| < 1 \qquad \text{partial coherence (e.g., pulse)}$$
$$|\gamma_{12}| = 0 \qquad \text{complete incoherence (e.g., noise).}$$

The magnitude of the normalized coherence function can be directly obtained experimentally from measurements of the local contrast in the fringe visibility for a given delay time between the beams. In an interference pattern the intensity varies between two limits, $I_{max}$and I_{min} which are given by

$$I_{max} = I_1 + I_2 + 2\sqrt{I_1 I_2}\,|\gamma_{12}|$$

and

$$I_{min} = I_1 + I_2 - 2\sqrt{I_1 I_2}\,|\gamma_{12}|.$$

The fringe visibility, V, is defined as

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

It follows that

$$V = 2\sqrt{I_1 I_2}\,\frac{|\gamma_{12}|}{I_1 + I_2}$$

which, if $I_1 = I_2$ gives

$$V = |\gamma_{12}|.$$

The typical variation of the function $|\gamma_{12}|$ is indicated in figure 7.2.2.

The *characteristic decay time*, τ_c , of the $|\gamma_{12}|$ function is known as the *coherence time*. The associated *coherence length*, $\ell_c = c\tau_c$, is a measure of the corresponding path difference. For a typical incandescent source such as a mercury lamp, light associated with a typical spectral line has a coherence time of 1 nanosecond, for which the associated coherence length is 1/3 m.

7.3. Coherence Time and Coherence Length

The coherence time and coherence function are measures of the temporal characteristics of the source. In order to see how the coherence function is related to the source properties, we consider a hypothetical "quasi-monochromatic" point source with the following idealized characteristics. The oscillations of the field occur for some time τ_0 and then the phase of the wave changes abruptly to a new, random, value. The oscillations again occur for time τ_0, when a new phase change occurs to a new random value. The cycle is then repeated, *ad infinitum*, with the phase being randomly distributed in the interval $[0, 2\pi]$. Figure 7.3.1 shows the typical phase variations of the wave.

The time dependence of the field of the quasi-monochromatic wave can be expressed as

$$\mathcal{E}(t) = e^{-i\omega_0 t} e^{i\phi(t)}.$$

The random phase changes that occur in this source can be considered to be due to, for example, collisions between the atoms in the source.

The coherence function for this source is simply given by

$$\gamma_{12}(\tau) = \left\langle e^{i\omega_0 t} e^{-i\phi(t)} e^{-i\omega_0(t+\tau)} e^{i\phi(t+\tau)} \right\rangle = \left\langle e^{i\omega_0\tau} e^{-i\phi(t)} e^{i\phi(t+\tau)} \right\rangle$$

$$= e^{-i\omega_0\tau} \lim_{T\to\infty} \frac{1}{T} \int_0^T e^{-i[\phi(t)-\phi(t+\tau)]} dt. \tag{7.3.1}$$

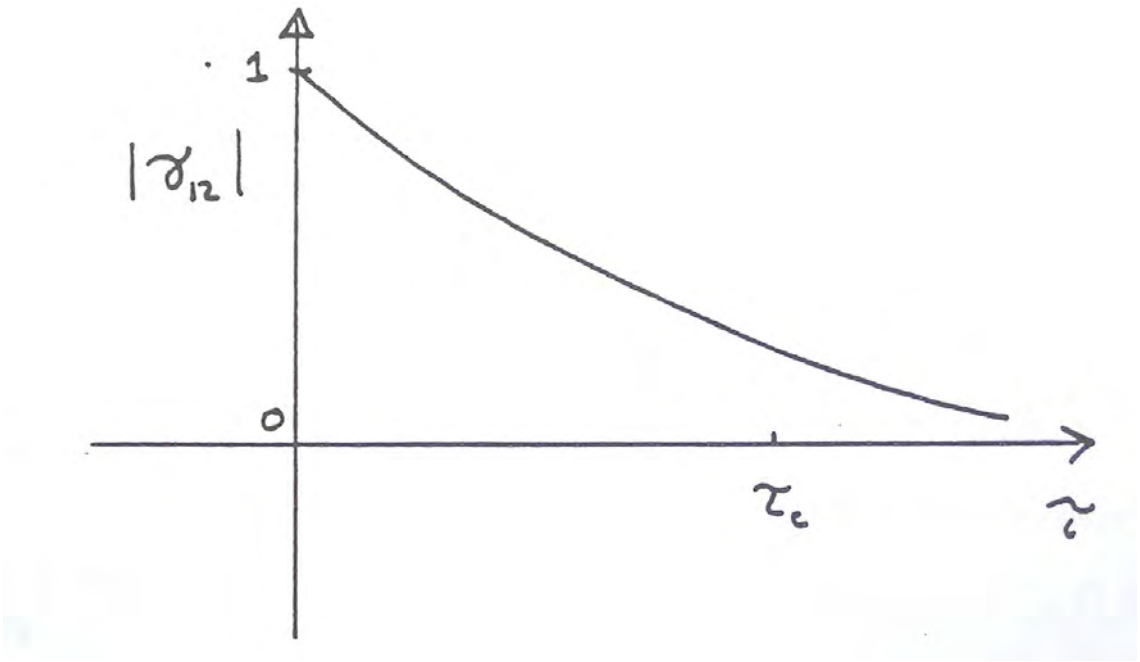


FIGURE 7.2.2. Typical variation of the magnitude of the coherence function.

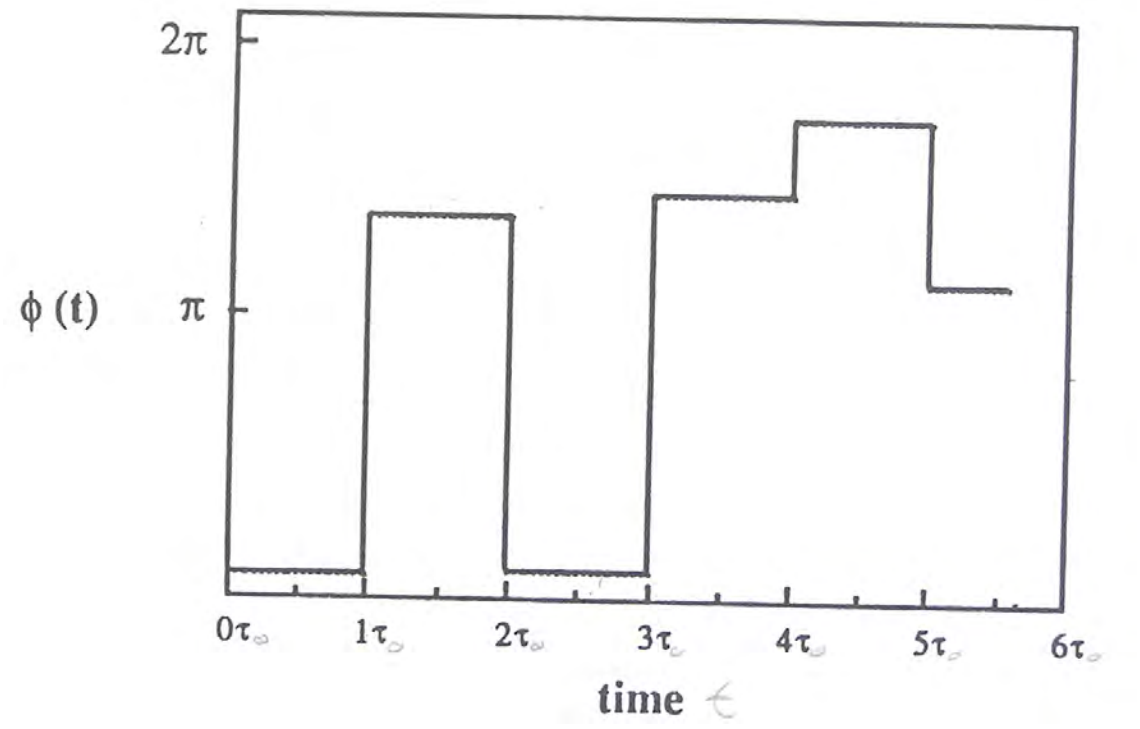


FIGURE 7.3.1. Phase function, $\phi(t)$, for a quasi-monochromatic wave.

The phase difference, $\phi(t) - \phi(t+\tau)$, determines the nature of the coherence function in general. For the particular phase fluctuations represented by figure 7.3.1, the phase difference is indicated in figure 7.3.2. If t is within the first coherence time interval where $0 \leq t \leq \tau_0$, the phase difference has the value zero if $t + \tau < \tau_0$ (*i.e.*, same "cycle"), but in the interval $2\tau_0 > (t + \tau \geq \tau_0$ (different "cycles") the phase difference assumes a random value between 0 and

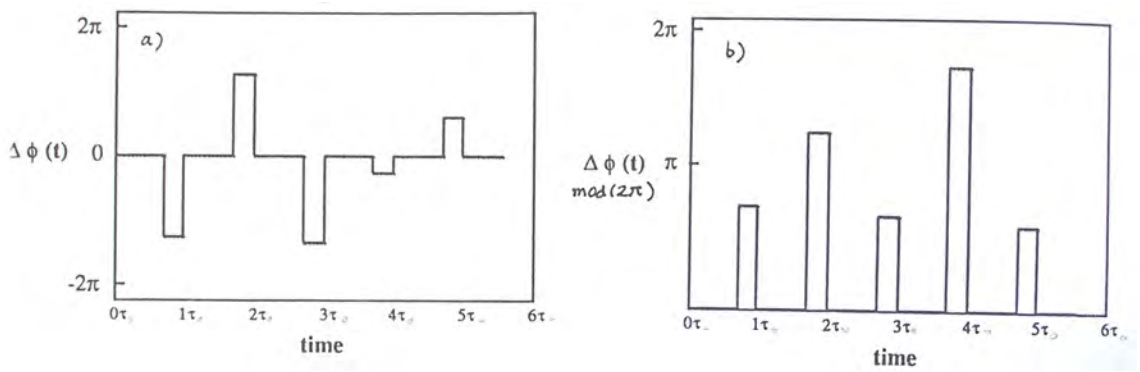


FIGURE 7.3.2. Graph of phase difference, $\triangle\phi(t,\tau) = \phi(t) - \phi(t+\tau)$.

2π. The same is true for all succeeding time intervals of width τ_0. The integral in equation 7.3.1 is then given by

$$\frac{1}{\tau_0}\int_0^{\tau_0} e^{-i[\phi(t)-\phi(t+\tau)]}dt = \frac{1}{\tau_0}\int_0^{\tau_0-\tau} dt + \frac{1}{\tau_0}\int_{\tau_0-\tau}^{\tau_0} e^{-i\triangle}dt$$

$$= \frac{\tau_0-\tau}{\tau_0} + \frac{e^{-i\triangle}}{\tau_0}\tau \tag{7.3.2}$$

where $\triangle$ is the random phase difference.

The same result is obtained for all subsequent intervals with the phase $\triangle$ varying with each interval. Since $\triangle$ varies randomly, the term involving $e^{-i\triangle}$ averages to zero. The other term in equation 7.3.2 is independent of the interval and so remains. If, however, $\tau > \tau_0$, the phase difference varies randomly so this term also averages to zero.

From this analysis we see that the coherence function is given by

$$\gamma_{12}(\tau) = \begin{cases} (1-\tau/\tau_0)\, e^{-i\omega_0\tau} & \tau \le \tau_0 \\ 0 & \tau > \tau_0 \end{cases} .$$

Hence the *degree of coherence* is given by

$$|\gamma_{12}(\tau)| = \begin{cases} (1-\tau/\tau_0) & \tau \le \tau_0 \\ 0 & \tau > \tau_0 \end{cases} .$$

A graph of the degree of coherence is plotted in figure 7.3.3. The time τ_0 in the simple case considered here can be taken as the constant collision time between the atoms. In a more realistic model where the time between collisions varies about the mean value τ_0, with a Gaussian probability function for the collision duration, the linear curve for $|\gamma_{12}(\tau)|$ becomes an exponential curve with

$$\left\langle e^{-i[\phi(t)-\phi(t+\tau)]} \right\rangle = e^{-\tau/\tau_0}.$$

In either case the coherence function decreases with increasing time separation, τ. The *coherence length* of the source again is given by

$$\ell_c = c\tau_0.$$

7.4. The Wiener-Khintchine Theorem

The coherence function gives us information about the temporal characteristics of the source. If one were to analyze the frequency content of the optical source one should be able to obtain the equivalent information in the frequency domain. This can be seen if we consider the self coherence or correlation function. We then have

$$\mathcal{F}[\Gamma(\tau)] = \int_{-\infty}^{\infty} \Gamma_{11}(\tau) e^{i\omega\tau} d\tau$$

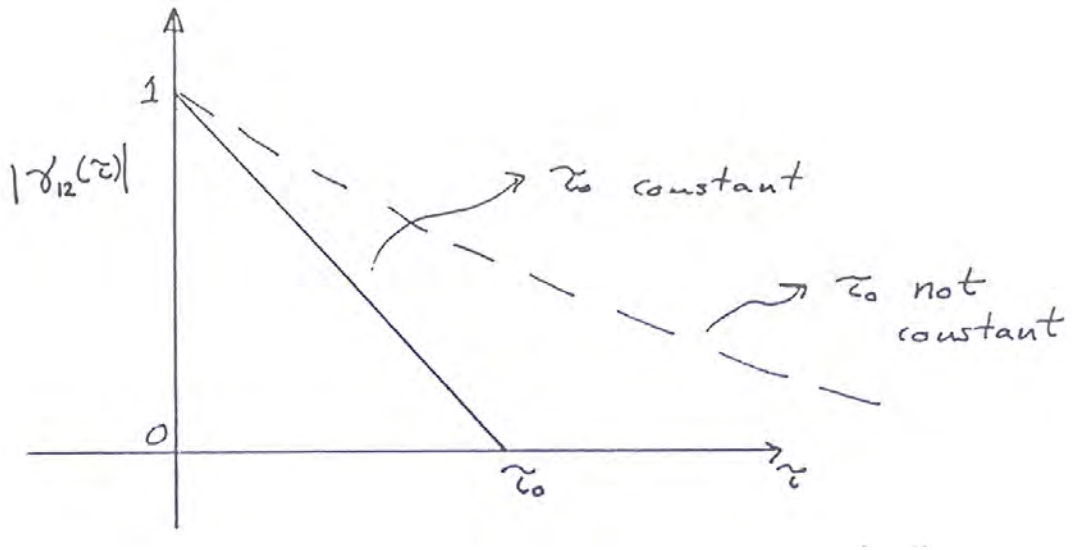


FIGURE 7.3.3. Graph of the degree of coherence for the phase function shown in figure 7.3.1 and for the case where the "jump time" has a Gaussian distribution centered on the value τ_0.

$$
\begin{aligned}
&= \lim_{T\to\infty} \tfrac{1}{T} && \int_{-\infty}^{\infty}\int_{-T/2}^{T/2} \mathcal{E}^*(t)\mathcal{E}(t+\tau)e^{i\omega\tau}dtd\tau \\
&= \lim_{T\to\infty} \tfrac{1}{T} && \int_{-\infty}^{\infty}\int_{-T/2}^{T/2} \mathcal{E}^*(t)e^{-i\omega t}\mathcal{E}(t+\tau)e^{i\omega(\tau+t)}d\tau dt \\
&= \lim_{T\to\infty} \tfrac{1}{T}\mathcal{E}^*(\omega)\mathcal{E}(\omega) && = \lim_{T\to\infty} \frac{1}{T}\left|\mathcal{E}(\omega)\right|^2
\end{aligned}
$$

where we designate $\mathcal{E}(\omega)$ as the Fourier transform of the time varying field. But since the RHS is (to within a constant) related to the power spectrum (power per unit frequency) in the field, $P(\omega)$,we have that

$$P(\omega) = \mathcal{F}\left[\Gamma(\tau)\right].$$

This result is known as the *Wiener-Khintchine theorem. The Fourier transform of the field coherence function is the power spectrum of the source.*

For example, in the case of a perfectly monochromatic source with

$$\mathcal{E}(\omega) = \mathcal{E}_0 e^{-i\omega_0 t}$$

we have that $\Gamma_{11}(\tau) = \left|\mathcal{E}_0\right|^2 e^{-i\omega_0\tau}$ and $\mathcal{F}\left[\Gamma_{11}(\tau)\right] = \left|\mathcal{E}_0\right|^2 \delta(\omega-\omega_0)$ indicating that the power function in this case is a δ-function centered on the frequency of the monochromatic wave, ω_0. A more realistic source has a self-coherence function given by

$$\Gamma_{11}(\tau) = \left|\mathcal{E}_0\right|^2 e^{-i\omega_0 t}e^{-\tau/\tau_0}. \tag{7.4.1}$$

In this case a little algebra gives a power spectrum

$$P(\omega) = \frac{\left|\mathcal{E}_0\right|^2 \left(\frac{\tau_0}{\pi}\right)}{\left[(\omega-\omega_0)^2 + \left(\frac{1}{\tau_0}\right)^2\right]}$$

which has a Lorentzian characteristic with full width at half maximum (FWHM) of $2/\tau_0$. It is no coincidence that the power spectrum in this case has the same functional form as the expression for $\kappa(\omega)$ discussed in connection with the classical model of the atom in chapter 1. In chapter 12 we show that the emission spectrum of a source has the same functional dependence on ω as the absorption spectrum.

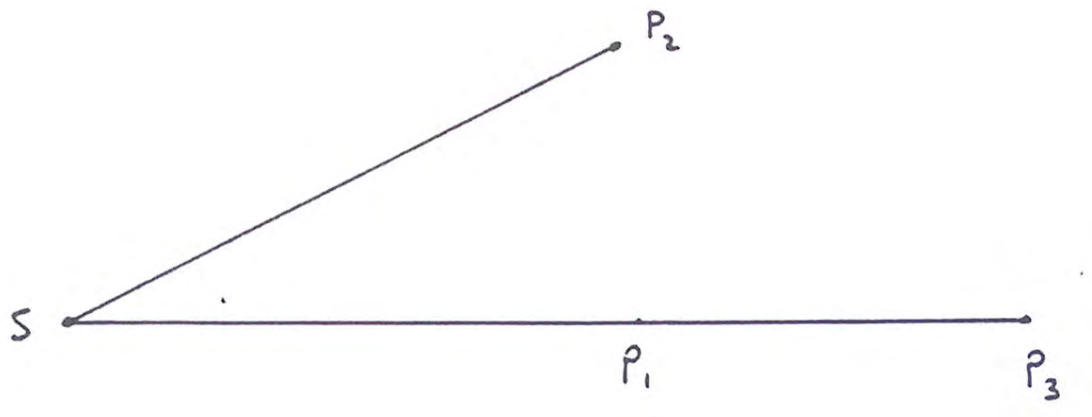


FIGURE 7.5.1. Diagram to illustrate the principal characteristics of temporal and spatial coherence.

7.5. Temporal and Spatial Coherence

It should be noted that the coherence function as defined above is a measure of the *temporal coherence* of what is usually taken to be a point source of spherical waves, and, in the far field, plane waves. In other words what the interference pattern is recording is the temporal correlation of electric fields emitted at different times from a quasi-monochromatic source which, in the case of Young's experiment, is a point source. In general, one can define two types of coherence, temporal coherence and *spatial coherence*. Whereas the former is a measure of the finite bandwidth of the source (a pure monochromatic wave as we have seen has perfect temporal coherence), the latter is a measure of the spatial extent of the source. In essence, one refers to the time-domain behavior of the source while the other refers to the spatial characteristics of the source. Figure 7.5.1 illustrates schematically how one measures these two different types of coherence.

Consider the three "observation points" P_1, P_2, and P_3. In the case of comparing the fields at point P_1 and P_3, it is evident that the coherence between the two fields only depends on the time delay between the two points. Hence a comparison of the fields from these two points is a measure of the temporal coherence. Because the points are separated only by a displacement along the direction of propagation of the field the temporal coherence is sometimes referred to as the longitudinal coherence. A comparison of the fields at P_1 and P_2 on the other hand is a measure of the spatial or lateral coherence. If the source is a point source, it is clear that if P_1 and $P_2$are equidistant from the source the coherence is independent of the separation between P_1 and P_2 and the source is said to have perfect spatial coherence. On the other hand, if the source is extended, then there are significant phase variations between P_1 and P_2 even if the source is emitting perfectly monochromatic waves. Since an extended source can be considered to be made up of a large number of independent sources the component fields have a different phase at point P_1 and in general the overall total field strength and intensity reflect some interference between these different field components. At a different point, P_2, the total intensity is different yet. A measure of the intensity as a function of lateral distance can be used to determine the spatial coherence function and yield information on the spatial extent of the source. Spatial coherence sees many applications particularly in those areas where it is not possible to gain access to the source. One of the most important applications is in the determination of the size of stars (through large-baseline interferometry) in astronomy investigations.

To illustrate how spatial coherence can be used to determine the spatial extent of sources, consider the simplest situation, namely two point sources separated by a distance s as shown in figure 7.5.2.

Consider two observation points P_1 and P_2 located such that their separations from the two point sources S_a and S_b are r_{1a}, r_{1b},r_{2a},r_{2b}. The fields at points P_1 and P_2 are then given by

$$\mathcal{E}_1 = \mathcal{E}_{1a} + \mathcal{E}_{1b}$$

$$\mathcal{E}_2 = \mathcal{E}_{2a} + \mathcal{E}_{2b}$$

where $\mathcal{E}_{1a}$ is the contribution to the field at point P_1 from source S_a, etc. The coherence function between the two observation points is thus the modulus of the function

$$\gamma_{12}(\tau) = \frac{\langle \mathcal{E}_1^*(t)\mathcal{E}_2(t+\tau)\rangle}{\sqrt{I_1 I_2}} = \frac{\langle \mathcal{E}_{1a}^*(t)\mathcal{E}_{2a}(t+\tau)\rangle}{\sqrt{I_1 I_2}} + \frac{\langle \mathcal{E}_{1b}^*(t)\mathcal{E}_{2b}(t+\tau)\rangle}{\sqrt{I_1 I_2}}$$

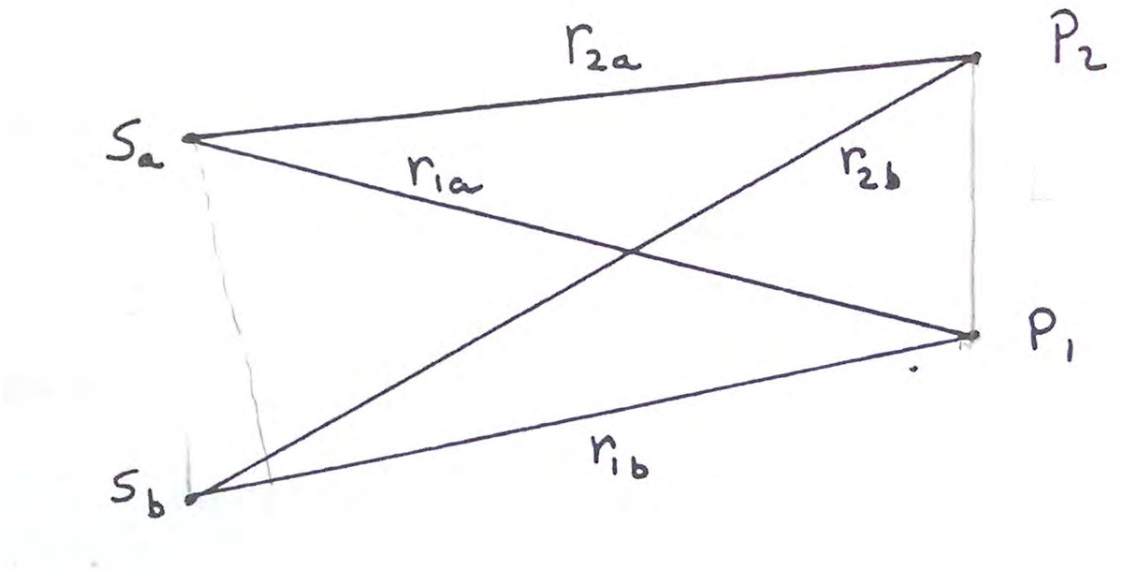


FIGURE 7.5.2. Geometry for determining the spatial coherence of two point sources.

where we have used the fact that the two sources are independent, and so incoherent. This implies

$$\langle \mathcal{E}^*_{1a}(t)\mathcal{E}_{2b}(t+\tau)\rangle = \langle \mathcal{E}^*_{1b}(t)\mathcal{E}_{2a}(t+\tau)\rangle = 0.$$

If the field is of the form

$$\mathcal{E}(t) \propto e^{-i\omega_0 t}e^{i\phi(t)}$$

then one can express the coherence function here as

$$\gamma_{12}(\tau) = \frac{1}{2}\gamma(\tau_a) + \frac{1}{2}\gamma(\tau_b)$$

where, in the case of the degree-of-coherence function represented by equation 7.4.1

$$\gamma(\tau) = e^{-i\omega_0 t}e^{-|\tau|/\tau_0}$$

and

$$\tau_a = \frac{r_{1a} - r_{2a}}{c} \quad \tau_b = \frac{r_{1b} - r_{2b}}{c}$$

The modulus of the coherence function is then given approximately by

$$|\gamma_{12}(\tau)| = \frac{1 + \cos\left[\omega_0\left(\tau_a - \tau_b\right)\right]}{2}\exp(-\frac{|\tau|}{\tau_0})$$

where we have assumed that

$$\tau_a \approx \tau_b = \tau$$

and

$$|\tau_a - \tau_b| \ll \tau_{a,b}.$$

But

$$\tau_a - \tau_b = \frac{r_{1a} - r_{1b}}{c} - \frac{r_{2b} - r_{2a}}{c} \simeq \frac{s}{c}\frac{L}{r}$$

where s is the distance between the two sources, L is the lateral distance between the two observation points and r is the mean distance between the sources and the observation points; it has been assumed that $r >> s,\ L$.

Figure 7.5.3 shows a plot of the function $|\gamma_{12}(\tau)|$ as a function of the lateral separation between the two receiving sources. The pointP_1 is located symmetrically with respect to the sources. The coherence is a maximum if point P_2 coincides with P_1. On the other hand as they move apart, the degree of coherence begins to drop. When

$$\omega_0(\tau_a - \tau_b) = \frac{\omega_0 sL}{cr} = \pi$$

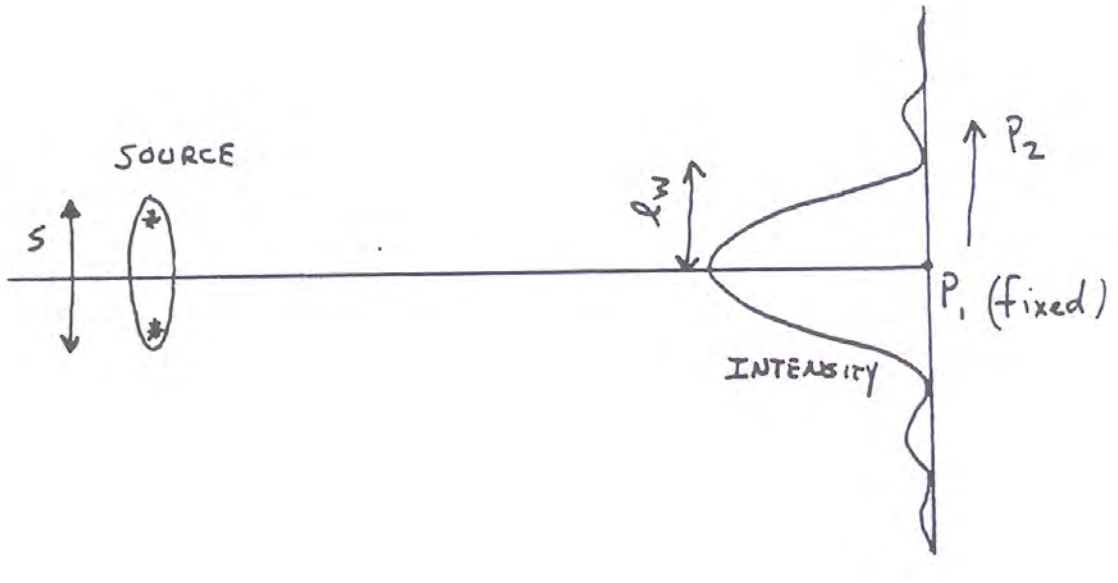


FIGURE 7.5.3. Diagram illustrating the lateral coherence of an extended source.

the degree of coherence vanishes. Because our extended source in this case consists of two isolated sources, the coherence function oscillates as a function of L. In the case of a continuous extended source, the coherence function, in most cases, decreases monotonically with

$$\ell_w = \frac{\pi c r}{\omega_0 s}$$

being a measure of the characteristic decay length of the coherence function in the observation plane. Hence, a determination of the value ℓ_w and a knowledge of the distance to the source can be used to determine the extent of the source.

7.6. Interferometer: wavefront-splitting

It is obvious from the above discussion that a considerable amount of information can be deduced about an optical source from measurements of its spatial or temporal coherence functions. In particular the temporal, spectral or spatial characteristics of the source can be determined by allowing two different fields from the source to interfere with each other and determining the intensity as a function of temporal or spatial variables. Instruments which perform this function are known as interferometers. Interferometers can be divided into two types, *amplitude-splitting* or *wavefront-splitting*. Amplitude splitting interferometers allow separate beams to be generated with something like a beam-splitter, and one beam is delayed with respect to the other before they are made to interfere. Wavefront-splitting interferometers function by taking separate beams from different parts of the wavefront emerging from the source and then bringing them to a common point where they can interfere. For example, the Young's interferometer which we considered at the beginning of this chapter is a wavefront-splitting interferometer. It can be used to measure both the spatial and temporal characteristics of a source. In the geometry we considered, where we kept the slit separation fixed and varied the observation point on a plane behind the slits, we were in fact determining the temporal characteristics of the point source. For an extended source, provided the slits are symmetrically located with respect to the source, we would still be measuring the temporal coherence of the source. On the other hand if we keep the observation point and one slit fixed while moving the other slit, we could obtain the spatial coherence function of the source.

Many different types of interferometers have been developed over the years to determine the coherence and spectral functions of the source. The type of interferometer used for a particular purpose is usually dictated by the distance from the source and the spectral width of the source. In this section we discuss two other types of interferometers, both of which are used to determine the spectral properties of the source. This property is usually the one of most concern to optical physicists. Astronomers, of course, are interested in both spatial and temporal coherence of the source. The two interferometers we discuss are the Michelson and the Fabry-Perot interferometers.

7.6.1. The Michelson Interferometer. This interferometer, developed by *Albert Michelson* in the 1880's, is perhaps most famous because it was used by Michelson to determine the velocity of the "ether", a hypothetical

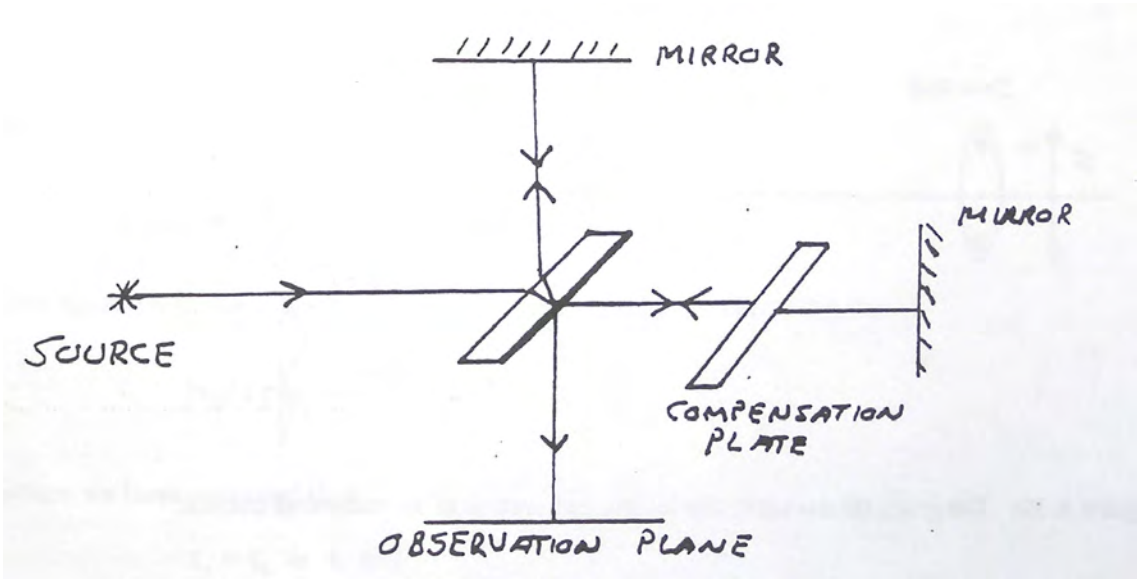


FIGURE 7.6.1. Schematic diagram of the Michelson interferometer.

material in which light was considered to propagate. Nowadays the *Michelson interferometer* is used to analyze light sources. The basic Michelson interferometer is illustrated in figure 7.6.1.

The interferometer consists of two mirrors located at right angles with respect to each other with one of the mirrors being on a translation stage. Light entering the interferometer is split into two beams by a partially silvered mirror or other form of beam-splitter. The separate beams fall on the two mirrors, are reflected back towards the beam splitter, and then combine and interfere in an exit beam with the interference pattern being observed in an observation plane. In general the Michelson interferometer is not a symmetrical instrument in terms of the paths traversed by both beams. Because the beam splitter has a finite thickness the two beams emerging from the beam splitter travel different distances, whether the beam-splitter is partially silvered on the front or the back surface. To compensate for this a glass plate is often inserted in one of the arms.

The Michelson interferometer is an amplitude-splitting interferometer and can be used to measure the temporal coherence of the source. This is done by measuring the intensity of the exit beam as a function of the path difference between the split-off beams in the interferometer arms. Let us see how this is done in the case of a source of plane waves. If ℓ_1 is the length of the fixed arm (the distance from the beam splitter to the mirror in the fixed arm) and ℓ_2 is the length of the variable arm, then the time delay between the two recombining beams is

$$\tau = \frac{2(\ell_1 - \ell_2)}{c}$$

and the intensity of light in the observation plane is given by

$$I = I_1 + I_2 + 2Re\left(\langle \mathcal{E}_1(t)^* \mathcal{E}_2(t+\tau)\rangle\right)$$

$$= I_1 + I_2 + 2Re(\Gamma_{11}(\tau)$$

where the 1 and 2 subscripts refer to the beams *emerging* from two arms of the interferometer. Depending on the characteristics of the beam splitter, the intensity of the two beams may not be the same. For $\tau = 0$, as ℓ_2 is varied relative to ℓ_1, the intensity in the observation plane varies between *0* and *4*I_1 if the intensities of the two recombining beams are the same. For values of $\tau \gg \tau_c$ the coherence function has decayed to zero and the intensity in the observation plane is given by $I = 2I_1$. Figure 7.6.2 shows a typical plot of the observation plane intensity as a function of $\ell_1 - \ell_2$.

A variation in intensity associated with passing between adjacent maxima or minima in the intensity pattern is referred to as a fringe. By counting the number of fringes in a given distance of travel of ℓ_2, one can determine the wavelength of light of the source. (Don't forget the factor of two related to the fact that one fringe corresponds to $\triangle\ell_2 = \lambda/2$). The accuracy of the wavelength determination is, of course, fundamentally limited by the coherence

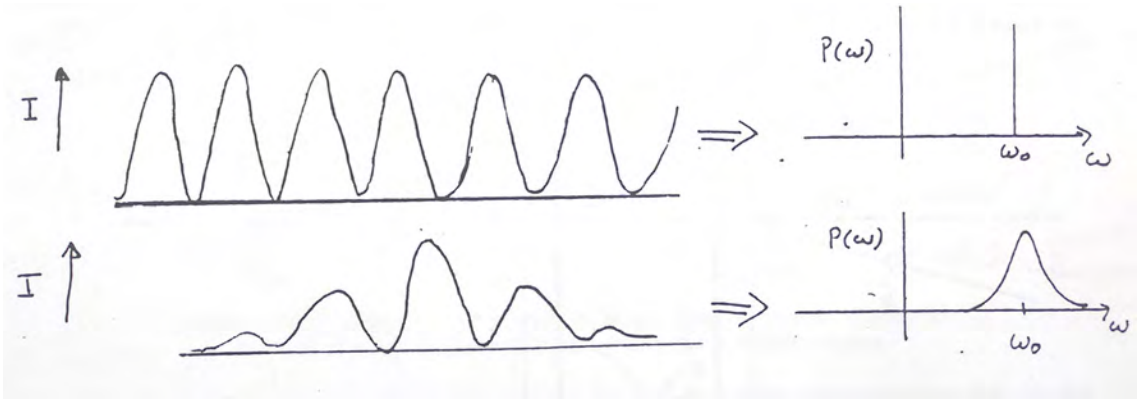


FIGURE 7.6.2. Typical intensity variations recorded by a Michelson interferometer. Note: wings in lower figure should go to $2I_1$,not zero.

length of the source, since this determines the maximum usable separation between the two recombining beams before essentially no interference occurs.

In the case of "white" light from a tungsten lamp, one can really observe only 2 or 3 fringes before the interference pattern disappears and by using a micrometer stage in the movable arm one would determine that white light corresponds to $\lambda = 0.5\,\mu m$ with $\triangle\lambda/\lambda = 0.5$. On the other hand, if one is making observations on the green line from a mercury discharge lamp, one finds that it is possible to obtain something like 107 fringes (if one has the patience or an automatic recording system). Provided one's optics and translation stage are of suitable high quality one could obtain the wavelength of the green line ($\lambda = 0.53\,\mu m$) to this accuracy. In practice, the accuracy obtained is never this high, since the travel distance in the movable arm would have to be of the order of 1 m. Micrometer translation stages with lengths of travel of 1m are not made unless one uses a corner cube reflector!

In general the source of light for a Michelson interferometer is not a plane wave source and, as a result, the intensity in the observation plane is not spatially uniform. More often than not the source of light can be approximated as a point source for which the surfaces of constant phase are spheres. When such light falls on the observation plane, the contours of constant phase are, of course, concentric circles. For light emerging from the fixed arm of the interferometer the concentric circles are fixed, while those associated with the moving arm expand as ℓ_2 increases. In this case the fringes and the regions of low and high intensity are circles which expand as ℓ_2 increases. By placing a detector at the center of the concentric circles, or for that matter over any sufficiently small area, and by varying ℓ_2 one can determine the coherence function of the source.

7.6.2. The Fabry-Perot Interferometer. The Michelson interferometer is useful in analyzing the coherence function or power spectrum of light sources with coherence lengths of the order of centimeters or less. The analysis of light sources with longer coherence lengths is not feasible with the Michelson interferometer as noted above because of the length of the arms required. On the other hand, if one could allow light to bounce back and forth many times within an arm, one would increase the effective path length while reducing the size of the instrument. This is the basis for the *Fabry-Perot interferometer* which was developed by the two French scientists Charles Fabry and Alfred Perot in 1899. The instrument is illustrated schematically in figure 7.6.3.

The device, which consists of two partially transmitting surfaces facing each other, operates on the principle of multiple beam interference. Light which enters the instrument from one side is reflected back and forth many times , with a little light escaping at each interface. The light which leaves is the superposition of all the partial beams which emerge from the partially transmitting output mirror. The added path length associated with the multiple reflections inside the "cavity" allow for the analysis of light beams with very long coherence lengths. In essence the Fabry-Perot instrument serves as a narrow bandwidth filter or a high resolution interferometer. If the separation between the plates, d, is constant one speaks of a *Fabry-Perot etalon*, otherwise it is called a Fabry-Perot interferometer. Because of the long effective path length inside the interferometer, the surfaces of the partially transmitting mirrors must be made with great precision so as to not cause phase distortion in the light beam which is being analyzed. A high quality Fabry-Perot interferometer can have a surface flatness of better than $\lambda/100$. In addition, the mirror surfaces may be curved (in particular, convex) so that the light during its many traversals is confined to the cavity.

In what follows we analyze the basic features of a Fabry-Perot interferometer with plane mirrors. In these types of instruments, so as to avoid any undesirable effects arising from reflection off the non-mirror surfaces, the mirror elements are made slightly prismatic or wedged. We therefore ignore these secondary interfaces. For ease of

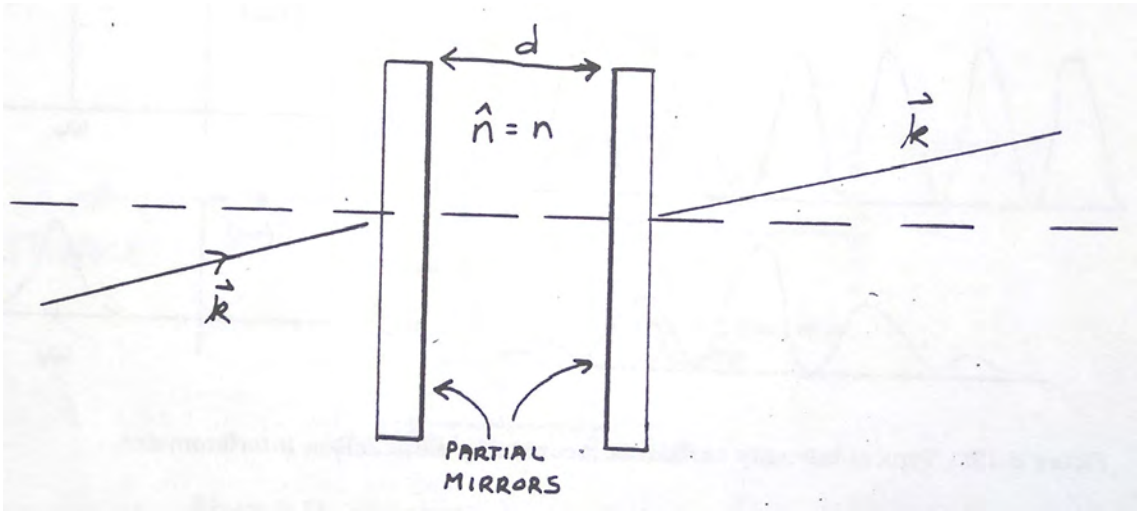


FIGURE 7.6.3. The Fabry-Perot interferometer.

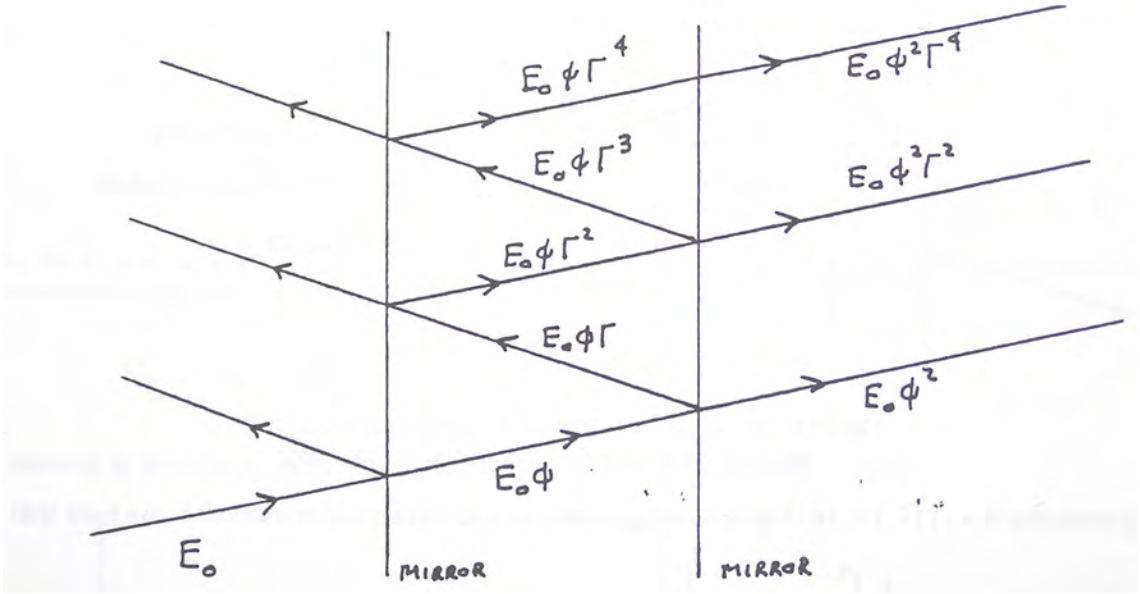


FIGURE 7.6.4. Illustration of the multiple beams involved in determining the transmission characteristics of a Fabry-Perot interferometer.

analysis we choose to analyze the symmetric Fabry-Perot instrument where both mirrors have identical reflectivity and transmissivity characteristics. We initially consider the incident waves to have planar characteristics.

Consider then a continuous light beam beam of amplitude $\mathcal{E}_0$ falling on the Fabry-Perot cavity at an angle of incidence θ as shown in figure 7.6.4.

Let the (amplitude) reflectivity of the mirror surfaces be $\Gamma = re^{i\phi}$ and the transmission be given by $\Phi = te^{i\sigma}$. The amplitude of the output beam is the superposition of the multiply reflected beams, taking into account the transmission of the end mirror and the phase delay between the different components. From figure 7.6.5 we see that the path difference between beams following two successive reflections inside the cavity is given by $2d\cos\theta$ and the corresponding phase delay is given by

$$\delta = 2kd\cos\theta = \frac{4\pi n}{\lambda_0} dcos\theta$$

where n is the refractive index of the medium between the two mirrors. Apart from phase factors, the amplitude of the beam inside the cavity following 0, 1, 2,...reflections is $\mathcal{E}_0\Phi$, $\mathcal{E}_0\Phi\Gamma$, $\mathcal{E}_0\Phi\Gamma^2$, etc. while the different components emerging from the cavity have amplitudes $\mathcal{E}_0\Phi^2$, $\mathcal{E}_0\Phi^2\Gamma^2$,$\mathcal{E}_0\Phi^2\Gamma^4$, etc. Taking into account the phase delay between

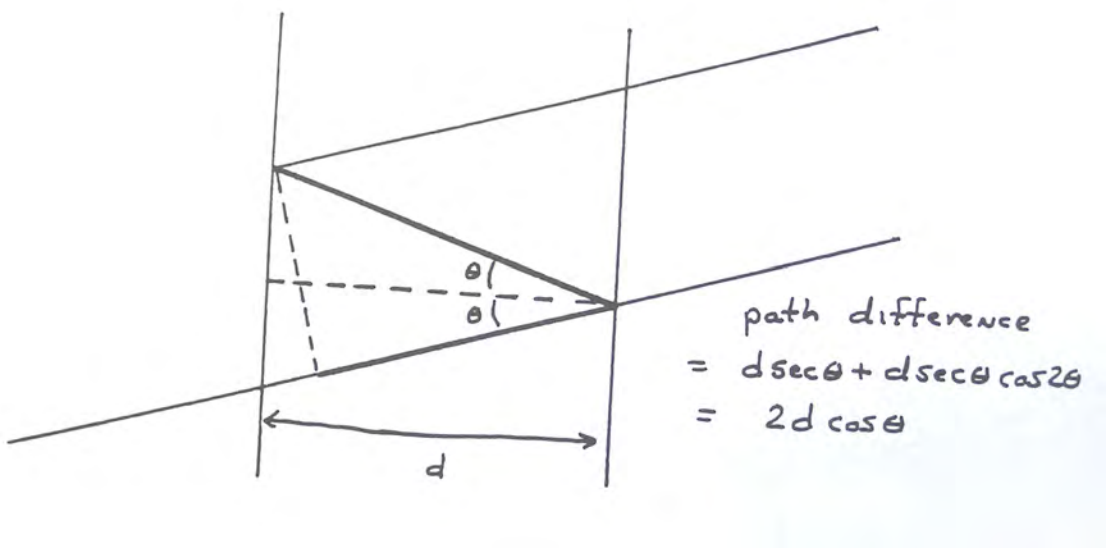


FIGURE 7.6.5. Determination of the path difference between two beams in the Fabry-Perot interferometer.

successive emerging components we have that the field associated with the output beam is given by

$$\mathcal{E} = \mathcal{E}_0\Phi^2 + \mathcal{E}_0\Phi^2\Gamma^2 e^{i\delta} + \mathcal{E}_0\Phi^2\Gamma^4 e^{2i\delta}$$

and the corresponding intensity (after summing the geometric series) is

$$I = I_0 \frac{|\Phi|^4}{|(1-\Gamma^2 e^{i\delta})|^2}.$$

If we define $R = |\Gamma|^2$, $T = |\Phi|^2$ as the energy reflection and transmission coefficients, we have that

$$I = \frac{I_0 T^2}{(1-R)^2} \frac{1}{1 + \left(\frac{4R}{(1-R)^2}\right) \sin^2\left(\frac{\triangle}{2}\right)}$$

where

$$\triangle = 2\phi + \delta.$$

If the internal medium is non-attenuating, we can write this simply as

$$I = \frac{I_0}{1 + \left(\frac{4}{\pi^2}F^2\right) \sin^2\left(\frac{\triangle}{2}\right)} \tag{7.6.1}$$

where

$$F = \frac{\pi\sqrt{R}}{1-R}$$

is known as the *finesse* of the interferometer. The functional form represented by equation 7.6.1 is known as the *Airy function*.

A plot of the transmission function as a function of $\triangle/2$ and for different values of the reflectivity is shown in figure 7.6.6. For low reflectivity, the output beam is essentially the result of the superposition of two beams of significantly different amplitude. When their path difference is increased they interfere giving rise to a small modulation of the output intensity with a nearly sinusoidal variation. On the other hand, when the reflectivity of the mirrors is high, many nearly-equal amplitude beams make up the output; constructive interference of the many components occurs only if the relative phase change is close to being a multiple of 2π, or if $\triangle/2$ is a multiple of π. High finesse or high reflectivity is associated with a sharper transmission function and a greater sensitivity to the incident wavelength. Note that at a transmission maximum the output intensity is the same as the input intensity.

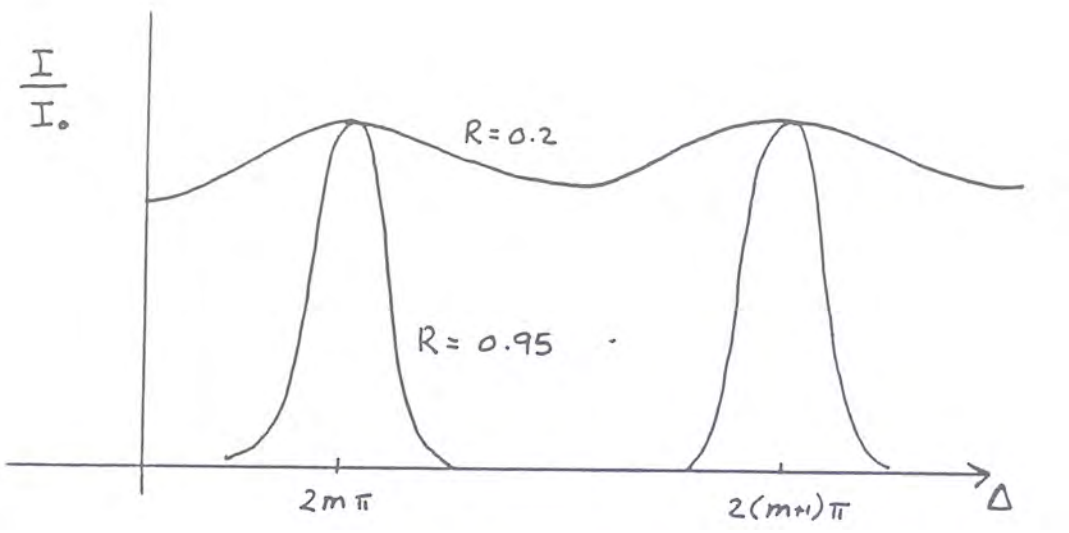


FIGURE 7.6.6. Graphs of the Airy function for different values of the finesse or reflectivity.

The intensity inside the interferometer may be much higher however since the overall output intensity is *(1-R)* of the intensity striking the rear mirror. Maxima in the transmission function are given by

$$\frac{\triangle}{2} = \pi, 2\pi, 3\pi..., m\pi$$

$$= \frac{2\pi n}{\lambda_m} dcos\theta + \phi = \frac{\omega_m n}{c} dcos\theta + \phi. \tag{7.6.2}$$

Hence, for a fixed separation between the plates and a fixed angle of incidence, adjacent maxima in the transmission function correspond to a constant frequency separation of

$$\omega_{m+1} - \omega_m = \frac{\pi c}{ndcos\theta}.$$

This quantity is referred to as the *free spectral range* of the interferometer.

For a system with large reflectivity the finesse is the spectral free range divided by the full width at half maximum of the transmission peaks. For a (typical) 10 cm separation of the mirrors with normally incident light the free spectral range corresponds to $10^{10}s^{-1}$, which is much less than the frequency of visible light. Because the output of the interferometer is periodic as a function of frequency with period equal to the frequency of the spectral free range, the analysis of light sources with bandwidths greater than the spectral free range is at best complicated and one may be better off using a Michelson interferometer. The Fabry-Perot interferometer is most useful for analyzing light sources with a bandwidth less than the free spectral range. For such sources the typical output of the interferometer is given in figure 7.6.7. In this figure the variation in $\triangle$ is achieved by scanning d. Although the output is still periodic in d, the component frequency may be obtained by comparing the output with that of a true monochromatic source of known frequency.

For plane waves incident on the interferometer, the output intensity is, of course, uniform with the intensity changing as a function of frequency of the source or separation. If the interferometer is operated with a plane waves incident, a detector located behind the interferometer can deduce the spectral content of the source when the separation between the plates is scanned as figure 7.6.7 points out. More often than not, a point source is used and the waves which interfere in the observation plane form concentric circular rings as in the case of the Michelson interferometer. In essence one is observing a scan of the interferometer with $\triangle$ changing because of the variation in θ. A scan of the intensity pattern in the observation plane gives the same information as a scan of the intensity at the center of the ring pattern as a function of d.

In order to evaluate the resolving power of the interferometer, let us consider two waves of slightly different frequencies ω and ω' falling on the interferometer. The output intensity pattern is the superposition of the two intensity patterns if the two sources are mutually incoherent. If we assume for simplicity that the two sources have

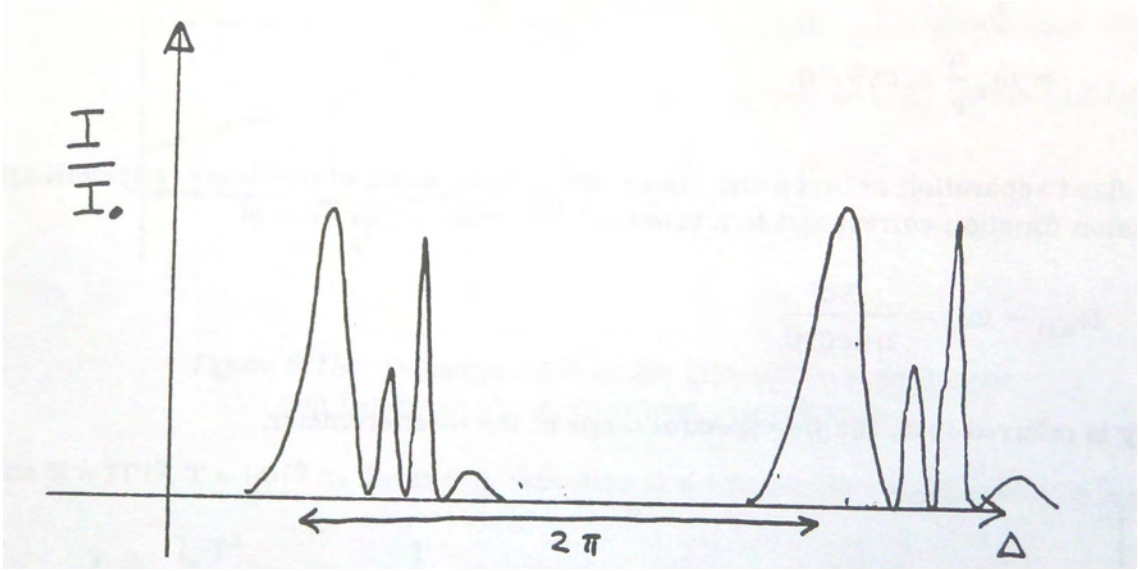


FIGURE 7.6.7. Output of the interferometer as a function of $\triangle/2$ for a multifrequency source.

the same intensity, the overall intensity is given by

$$I/I_0 = \frac{1}{1+\left(\frac{4}{\pi^2}F^2\right)\sin^2\left(\frac{\triangle}{2}\right)} + \frac{1}{1+\left(\frac{4}{\pi^2}F^2\right)\sin^2\left(\frac{\triangle'}{2}\right)}$$

where

$$\triangle = 2\phi + \frac{2\omega nd\cos\theta}{c}$$
$$\triangle' = 2\phi + \frac{2\omega' nd\cos\theta}{c}.$$

The intensity pattern as a function of $\triangle$ is shown in figure 7.6.8. In order to say that we can resolve the lines in the intensity pattern we must adopt a suitable criterion. A generally accepted criterion, which is explained more fully in a later chapter, is the *Rayleigh criterion*. According to this criterion two lines are considered to be resolved if the intensity at the saddle point is $8/\pi^2$ times the intensity at the two maxima. Therefore at the saddle point we must have

$$I/I_0 = 2\frac{1}{1+\left(\frac{4}{\pi^2}F^2\right)\sin^2\left(\frac{\triangle-\triangle'}{4}\right)} = \frac{8}{\pi^2}\left[1+\frac{1}{1+\left(\frac{4}{\pi^2}F^2\right)\sin^2\left(\frac{\triangle-\triangle'}{2}\right)}\right]$$

from which we obtain

$$\triangle - \triangle' \simeq \frac{2.1\pi}{F}.$$

Here we have assumed that the argument of the sine function is sufficiently small so that we can replace the sine function by its argument. It follows from the definition of the $\triangle$ and $\triangle'$ functions that

$$\delta\omega = \omega - \omega' = \frac{1.05\pi c}{Fnd\cos\theta}$$

and the resolving power of the interferometer is

$$R.P. = \frac{\omega}{\delta\omega} = \frac{\omega Fnd\cos\theta}{1.05\pi c}.$$

For an interferometer with a finesse of 20, $\theta = 0$, $d = 10$ cm and for an incident wavelength near $1\mu m$, the resolving power of the interferometer is $R.P. = 4\times10^6$. This is approximately 1000 times better than that of the grating spectrometer. Note that this resolving power corresponds to a frequency resolution of $10^9 s^{-1}$.

In later chapters, when we discuss lasers and feedback systems associated with lasers, we point out that the multiple reflections which occur inside a Fabry-Perot system can be used to provide feedback. In this sense a Fabry-Perot interferometer can be viewed as a standing wave resonator with the phase condition given by equation 7.6.2 being viewed as a condition for establishing standing wave or oscillation modes. The fact that the interferometer can

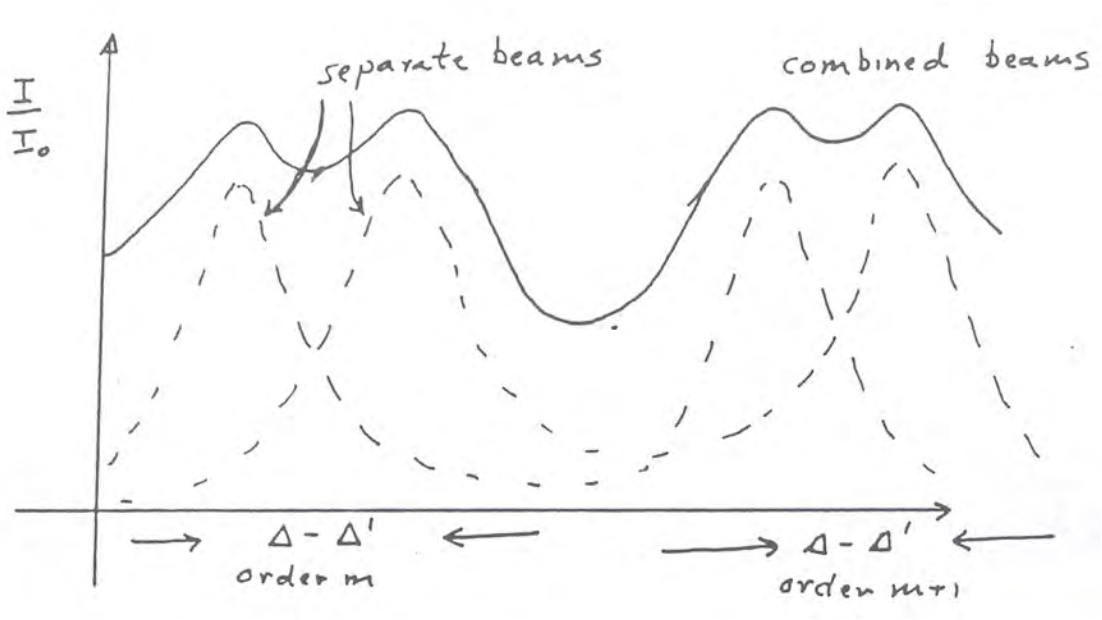


FIGURE 7.6.8. Intensity pattern for two monochromatic lines which are close in frequency.

be viewed as an oscillator which stores energy at well defined frequencies is more easily seen if we again consider the transmission or response function of the oscillator. Consider the m^{th} order transmission peak which occurs for

$$\triangle_m = 2m\pi.$$

At this phase value the energy stored in the interferometer is a maximum as pointed out above. Now consider the transmission function of a high finesse interferometer in the vicinity of a particular $\triangle_m$ with

$$\xi = \triangle - \triangle_m = \frac{2(\omega - \omega_m)nd\cos\theta}{c}.$$

It follows that

$$\sin^2(\frac{\triangle}{2}) = \sin^2(m\pi + \frac{\xi}{2}) = \sin^2(\frac{\xi}{2}) \approx \frac{\xi^2}{4}.$$

The transmission function is therefore given by

$$I/I_0 = \frac{1}{1 + \left(\frac{4}{\pi^2}F^2\right)\frac{\xi^2}{4}}$$

which represents a Lorentzian response function with a full-width at half maximum given by

$$\triangle\omega = \frac{\pi c}{Fnd\cos\theta}.$$

Hence the Fabry-Perot interferometer can be viewed as a macroscopic oscillator with a number of equally spaced resonance frequencies. These "oscillators" have energy damping rates (leakage rates) given by $\triangle\omega$ and energy lifetimes of the order of

$$\tau = \frac{1}{\triangle\omega}.$$

The higher the finesse of the system the slower the rate of energy loss from the resonator.

7.7. Multilayer Thin Films

As a last example of interference we consider the optical properties of thin films. These films, which usually consist of single or multi-layers of thin (sometimes less than a wavelength thick) dielectrics on various types of glass substrates, can be used to create low reflection or anti-reflection surfaces or high reflection surfaces. In addition one can generate all kinds of optical interference filters, including high pass (passes short wavelengths, attenuates long wavelengths), low pass and band pass. Multilayer films enjoy many practical (the windows of the Royal Bank tower in downtown Toronto are a prime example) and exotic uses these days and can be used to tailor the energy transmission and spectral properties of many surfaces.

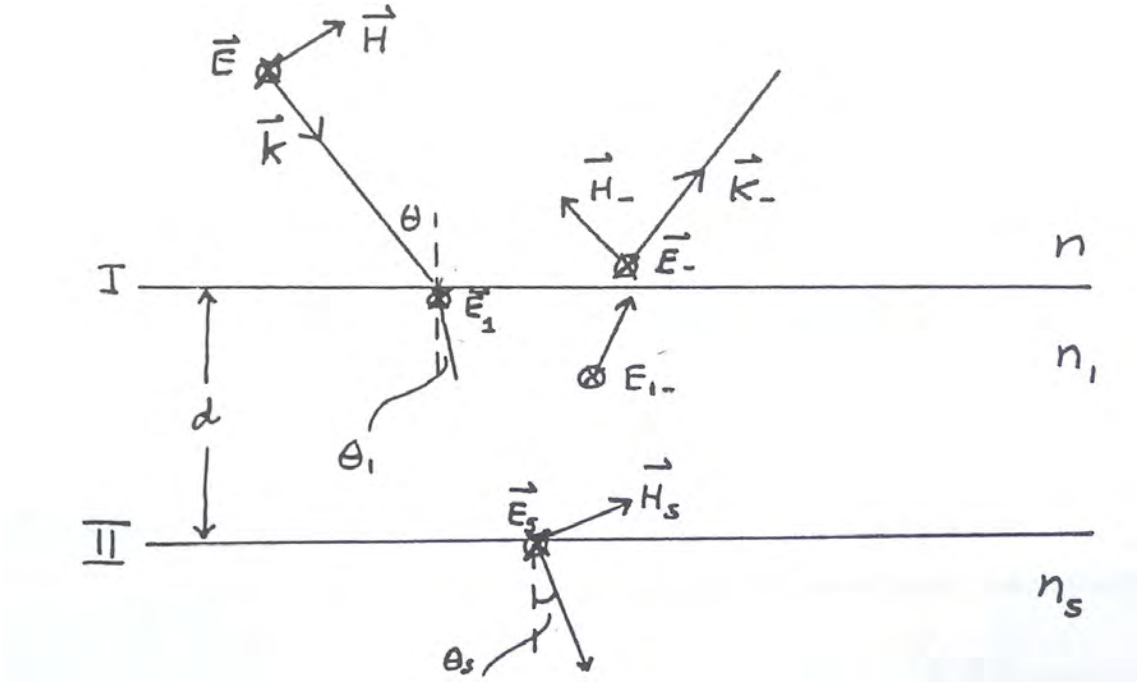


FIGURE 7.7.1. Fields at the boundary between a thin film and two adjacent media.

In our treatment of multilayer films we consider treating the total electric and magnetic fields and, through the boundary conditions associated with each, arrive at the reflection and transmission properties of different types of films. As will be seen, an individual layer can be treated using matrix techniques and multilayers can be treated using matrix multiplication.

In this section we offer an analysis of the interaction of s-polarized waves with a thin film of thickness d. The treatment for p-waves follows in a similar fashion. Consider then the interaction of the s-polarized monochromatic beam with the thin film shown in figure 7.7.1. Let us consider the refractive index of all media to be real and the refractive index of the incident medium to be n.

Consider the plane wave to be incident on the thin film with refractive index n_1 at an angle of incidence θ, and the refraction angle in the second medium to be θ_1. The refraction angle into the substrate is designated by θ_s. In general the fields in the three different media are designated with no subscript (incident medium), "1" subscript (thin film), and "s" subscript (substrate) The total reflected fields from the two interfaces (I and II) are designated by a "-" subscript.

At the first boundary, I, we have, by continuity of the electric and magnetic fields, that

$$\mathcal{E}_I = \mathcal{E} + \mathcal{E}_- = \mathcal{E}_1 + \mathcal{E}_{1-}$$

where $\mathcal{E}_1$ is the transmitted wave across the first interface and $\mathcal{E}_{1-}$ is the amplitude of the field reflected from the second interface (II) evaluated at the first interface. Continuity of the tangential component of the magnetic field requires

$$\mathcal{H}_I = \left(\frac{\epsilon_0}{\mu_0}\right)^{1/2} (\mathcal{E} - \mathcal{E}_-)\, n \cos\theta = \left(\frac{\epsilon_0}{\mu_0}\right)^{1/2} (\mathcal{E}_1 - \mathcal{E}_{1-})\, n_1 \cos\theta_1.$$

Where $\mathcal{H}_I$ is the tangential component of the magnetic field at the first interface and we have used the fact that for a plane wave

$$\vec{\mathcal{H}} = \left(\frac{\epsilon_0}{\mu_0}\right)^{1/2} n\hat{k} \times \vec{E}.$$

At the second boundary we have that

$$\mathcal{E}_{II} = \mathcal{E}_1 e^{ikh} + \mathcal{E}_{1-} e^{-ikh} = \mathcal{E}_s$$

and

$$\mathcal{H}_{II} = \left(\frac{\epsilon_0}{\mu_0}\right)^{1/2} \left(\mathcal{E}_1 e^{ikh} - \mathcal{E}_{1-} e^{-ikh}\right) n_1 \cos\theta_1 = \left(\frac{\epsilon_0}{\mu_0}\right)^{1/2} \mathcal{E}_s n_s \cos\theta_s.$$

Here

$$h = n_1 d \cos\theta_1$$

and k is the vacuum propagation constant of the light. Upon eliminating the uninteresting fields, $\mathcal{E}_1$ and $\mathcal{E}_{1_}$between these equations we arrive at two equations which relate the fields at the two boundaries. These equations are found to be

$$\mathcal{E}_I = \mathcal{E}_{II}\cos(kh) - \mathcal{H}_{II}\frac{i\sin(kh)}{Y_1}$$

$$\mathcal{H}_I = -\mathcal{E}_{II}Y_1 i\sin(kh) + \mathcal{H}_{II}\cos(kh) \tag{7.7.1}$$

where the characteristic admittance of the thin film is given (as in chapter 3) by

$$Y_1 = \left(\frac{\epsilon_0}{\mu_0}\right)^{1/2} n_1 \cos\theta_1.$$

The result of a calculation for p-polarized waves would instead have the characteristic admittance of these waves to be

$$Y_1 = \left(\frac{\epsilon_0}{\mu_0}\right)^{1/2} n_1/\cos\theta_1.$$

Equation 7.7.1 can be put in the matrix form

$$\begin{bmatrix} \mathcal{E}_I \\ \mathcal{H}_I \end{bmatrix} = \begin{bmatrix} \cos(kh) & -i\sin(kh)/Y_1 \\ -Y_1 i\sin(kh) & \cos(kh) \end{bmatrix} \cdot \begin{bmatrix} \mathcal{E}_{II} \\ \mathcal{H}_{II} \end{bmatrix}$$

or

$$\begin{bmatrix} \mathcal{E}_I \\ \mathcal{H}_I \end{bmatrix} = \overleftrightarrow{M}_1 \cdot \begin{bmatrix} \mathcal{E}_{II} \\ \mathcal{H}_{II} \end{bmatrix}.$$

The *characteristic matrix* $\overleftrightarrow{M}_1$ relates the (tangential) components of the total fields at the two interfaces. This matrix is unimodular as required by energy conservation. For a characteristic matrix of the form

$$\overleftrightarrow{M} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

the inverse matrix is easily calculated to be

$$\overleftrightarrow{M}^{-1} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

If we introduce another interface below the II interface and is the characteristic matrix which relates the field at the II and III interfaces then

$$\begin{bmatrix} \mathcal{E}_I \\ \mathcal{H}_I \end{bmatrix} = \overleftrightarrow{M}_1 \cdot \overleftrightarrow{M}_2 \cdot \begin{bmatrix} \mathcal{E}_{III} \\ \mathcal{H}_{III} \end{bmatrix}.$$

In general, for a multilayer system the input and through-put fields are simply related by

$$\begin{bmatrix} \mathcal{E}_I \\ \mathcal{H}_I \end{bmatrix} = \prod_i \overleftrightarrow{M}_i \cdot \begin{bmatrix} \mathcal{E}_{n+1} \\ \mathcal{H}_{n+1} \end{bmatrix}.$$

Let us now denote the composite matrix of a multilayer system by $\overleftrightarrow{M}$ in general with

$$\overleftrightarrow{M} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$$

Consider the calculation of the reflection and transmission coefficients of a multilayer system. Using the definitions of the characteristic admittances in the incident and substrate media to be

$$Y = \left(\frac{\epsilon_0}{\mu_0}\right)^{1/2} n\cos\theta$$

and

$$Y_s = \left(\frac{\epsilon_0}{\mu_0}\right)^{1/2} n_s \cos\theta_s$$

we have

$$\begin{bmatrix} \mathcal{E} + \mathcal{E}_- \\ (\mathcal{E} - \mathcal{E}_-)Y \end{bmatrix} = \overleftrightarrow{M}_1 \cdot \begin{bmatrix} \mathcal{E}_s \\ \mathcal{E}_s Y_s \end{bmatrix}.$$

On writing out this matrix equation using the definition of the total (*i.e.*, $\mathcal{E}_- = \Gamma\mathcal{E}$) reflection and (*i.e.*, $\mathcal{E}_s = \Phi\mathcal{E}$) transmission coefficients we have

$$1 + \Gamma = m_{11}\Phi + m_{12}Y_s\Phi$$

and

$$(1 - \Gamma)\, Y = m_{21}\Phi + m_{22}Y_s\Phi.$$

Solving these two equations simultaneously for Γ and Φ gives us

$$\Gamma = \frac{Ym_{11} + YY_sm_{12} - m_{21} - Y_sm_{22}}{Ym_{11} + YY_sm_{12} + m_{21} - Y_sm_{22}}$$

and

$$\Phi = \frac{2Y}{Ym_{11} + YY_sm_{12} + m_{21} - Y_sm_{22}}.$$

To appreciate the power of this formalism we offer several examples. As a first example, we determine the conditions that are necessary for a single thin film layer to produce zero reflectivity. Such a film is called an *anti-reflection film*. We consider only the case for a normally incident beam. In terms of the refractive indices for the different materials, the energy reflectivity is given by

$$R_1 = \Gamma_1\Gamma_1^* = \frac{n_1^2(n - n_s)^2\cos^2(kh) + (nn_s - n_1^2)^2\sin^2(kh)}{n_1^2(n + n_s)^2\cos^2(kh) + (nn_s + n_1^2)^2\sin^2(kh)}.$$

If we allow the film thickness to be an odd multiple of $\lambda/4n_1$ so that

$$kh = \frac{\pi}{2}, \frac{3\pi}{2}, ...$$

then the formula for the reflectivity is reduced to the simple form

$$R_1 = \frac{(nn_s - n_1^2)^2}{(nn_s - n_1^2)^2}.$$

The reflectivity of the single thin film system becomes zero if

$$n_1 = \sqrt{nn_s}.$$

Because of the fact that refractive indices exhibit dispersion, one can generally only fulfill this condition exactly, if at all, at a particular wavelength. For camera lenses which, of course pass a broad range of visible light wavelengths, the thickness of the film is chosen so that the reflectivity is a minimum in the yellow (mid-visible) region of the spectrum.

The derived condition for the anti-reflecting layer is of course only exactly valid for normal incidence here, and, in general, only at a particular angle; the reflectivity is higher at other angles. The condition on the refraction index of the thin film can never be fulfilled exactly. For a glass/air interface with $n_s = 1.5$, one would require $n_1 = 1.22$. Unfortunately it is not possible to find a material with the appropriate chemical stability, durability, etc. with this refractive index at $\lambda \approx 0.55\,\mu$m (mid-visible), Nonetheless, MgF_2 ($n_1 = 1.38$) and cryolite ($n_1 = 1.35$) are common low index materials which can reduce the reflectivity of glass from 0.04 to below 0.015 over the visible region of the spectrum. The behavior of the anti-reflection coating as a function of wavelength is shown in Figure 7.7.2.

By considering multi-layer films one adds more degrees of freedom to the situation and the possibility of reducing the reflectivity below that achieved by a single thin film. For a double layer anti-reflection coating one has

$$\overleftrightarrow{M} = \begin{bmatrix} 0 & -i/Y_1 \\ -iY_1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & -i/Y_2 \\ -iY_2 & 0 \end{bmatrix}$$

which at normal incidence becomes

$$\overleftrightarrow{M} = \begin{bmatrix} -n_2/n_1 & 0 \\ 0 & -n_1/n_2 \end{bmatrix}.$$

The reflectivity for this double layer film becomes

$$R_2 = \left[\frac{n_2^2n - n_sn_1^2}{n_2^2n + n_sn_1^2}\right].$$

In order that $R_2 \equiv 0$ we require

$$\left(\frac{n_2}{n_1}\right)^2 = \frac{n_s}{n}.$$

When n_2 and n_1 are as small as possible the reflectivity has a single broad minimum about the chosen wavelength. It is also clear that the refractive index of the second layer has to be larger than the reflectivity of the first layer. It

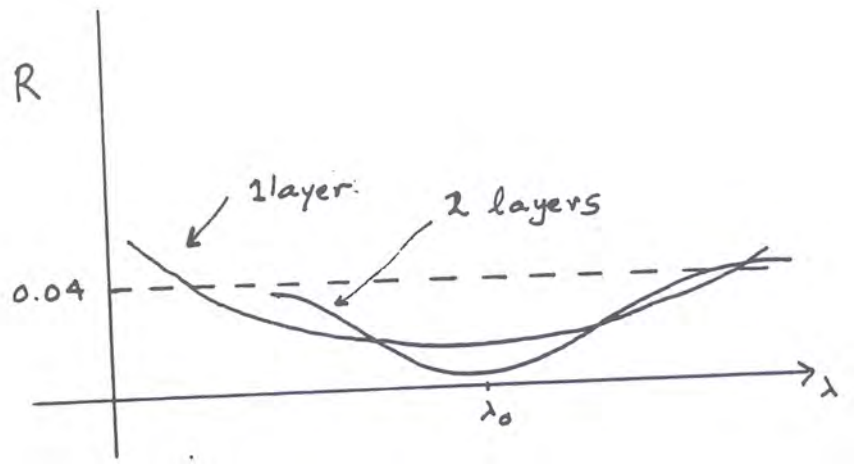


FIGURE 7.7.2. Reflectivity spectrum of glass with one and two coatings.

is common practice to designate a glass-high index-low index-air system as gHLa layer. Zirconium dioxide ($n = 2.1$), titanium dioxide ($n = 2.4$) and zinc sulfide;($n = 2.32$) are commonly used for H-layers while magnesium fluoride ($n = 1.38$) and cerium fluoride ($n = 1.63$) are often used as low index layers. Although it is still not possible with a double layer system to achieve zero reflectivity at a particular wavelength with a double layer system, the average reflectivity over the visible region of the spectrum is reduced over that of the single layer system as Figure 7.7.2 indicates.

In order to reduce the average reflectivity even further, one can consider triple, quadruple-layers, etc. with the additional films also being used to satisfy certain conditions concerning spectral response, incidence angle, etc. Although it may not seem important to increase the transmission of an interface from 0.96 to 1.00, in complex system, involving complex optical systems, such as camera lenses, the net throughput of a system can be increased by a factor of 2 by making the surfaces anti-reflective. In the case of a camera this allows for the reduction of the lens aperture by one f-stop.

Multilayer periodic systems are used to generate *high reflectance films* and *interference filters* as well as antireflection coatings. The alternate layers of high and low refractive index films is given the short-hand notation g(HL)Na where N is referred to as the period. The transfer matrix of the basic HL layer for normal incidence is found to be

$$\overleftrightarrow{M} = \begin{bmatrix} -n_H/n_L & 0 \\ 0 & -n_L/n_H \end{bmatrix}$$

and for N such elements the transfer matrix is

$$\overleftrightarrow{M}^N = \begin{bmatrix} (-n_2/n_1)^N & 0 \\ 0 & (-n_1/n_2)^N \end{bmatrix}.$$

The reflectivity for the g(HL)$^{\mathrm{N}}$a multilayer system is given by

$$R_{2N} = \left\{ \frac{\left(\frac{n_H}{n_L}\right)^{2N} - n_s}{\left(\frac{n_H}{n_L}\right)^{2N} + n_s} \right\}^2.$$

By making the ratio n_H/n_L as high as possible one can make the reflectivity of the layer approach unity. For example, for a glass substrate with $n_s = 1.5$, and for alternate layers of titanium dioxide and cryolite with $N = 7$, one finds the reflectivity to be 0.998 which is much higher, in the visible region of the spectrum, than can be achieved with bare metals. As with the anti-reflection coatings, the high reflectivity condition is valid only over a certain range of wavelengths and for a particular angle of incidence. In general, the spectral width of the high reflectance zone increases with the refractive index ratio while the maximum reflectivity is a function of the number of layers.

Interference filters, which allow for the passage of light at certain wavelengths, or in a certain wavelength region, can also be made using multilayer coatings. The advantage of using these filters over absorption filters is twofold. First of all their spectral properties are easily tailored to vary the spectral region of transmittance. Secondly, these

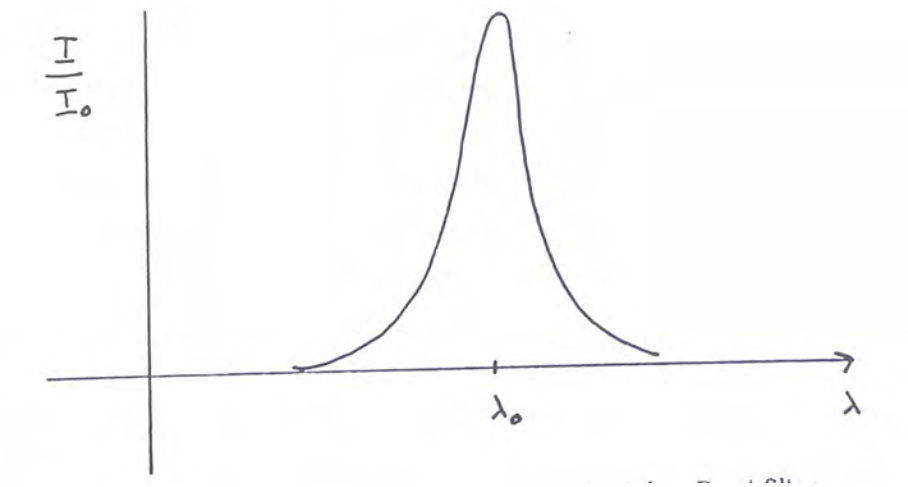


FIGURE 7.7.3. Transmission characteristics of a Fabry-Perot filter.

filters can be made from lossless dielectrics and so don't absorb any energy which falls on them. They are therefore preferred in those situations involving high intensity beams since their damage threshold is much higher.

As mentioned in the introductory comments in this section one can make low-pass, high-pass and bandpass filters. The latter type of filter can be made as a Fabry-Perot etalon. By making the separation between the plates of the etalon of the order of λ/n_e, where n_e is the refractive index of the etalon, one can achieve maximum transmission of light with wavelength λ, with additional transmission peaks well separated in wavelength. These other transmission peaks can be blocked using other types of filters such as absorption filters. The transmission peak of the etalon at the wavelength of interest may be made quite sharp by making the finesse of reflectivity of the etalon quite high.

An all-dielectric Fabry-Perot filter typically has the structure gHLHLHLLHLHLHa or gHLHLHLHHLHLHLHa where each of the layers has quarter wavelength thickness ($\lambda/4n_{1,2}$). The central double layer has a half wavelength thickness. The transmission characteristics of such filters are shown in figure 7.7.3.

For practical reasons, in general there is a trade-off in the maximum transmission and the spectral narrowness of such filters. A bandwidth as narrow as 1 nm with a peak transmission of 80% is not impossible, however, at certain wavelengths.

References

E. Hecht, *Optics*, Pearson, 2002.

G.R. Fowles, *Introduction to Modern Optics*, Holt, Reinhert and Winston, Toronto, 1968.

Problems

1. Calculate and plot the interference pattern that would be obtained if three slits instead of two were used in Young's experiment, assuming equal spacing between slits.

2. In an interference experiment of the Young's type, the distance between the slits is 0.5 mm. The wavelength of light is 0.6 μm. If it is desired to have a fringe spacing of 1 mm at the screen, what is the corresponding screen distance?

3. Fringes are observed when a parallel beam of light of wavelength of $0.5\mu m$ is incident perpendicularly onto a wedge-shaped film of refractive index 1.5. What is the angle of the wedge if the fringe separation is 0.3 cm?

4. Consider a piece of glass of uniform thickness t and refractive index n located in vacuum. If a plane wave of wavelength λ is incident on the glass at an angle of incidence θ, show that no reflected wave results if

$$2t\sqrt{n^2 - \sin^2\theta} = m\lambda$$

where m is a positive integer. What is the condition for no transmitted wave? How do the two conditions relate?

5. Light of wavelength $\lambda_0 = 0.59\,\mu m$ is incident at 45° on a thin soap film of refractive index $n = 1.35$. Parallel dark bands separated by 4 mm are observed on the film. Determine the angle between the top and bottom surfaces of the film.

6. A He-Ne laser is operating on three modes such that its power spectrum can be represented by

$$P(\omega) = A\left[\delta(\omega - \omega_0) + \delta(\omega - \omega_0 - \triangle) + \delta(\omega - \omega_0 + \triangle)\right]$$

where $\triangle \ll \omega_0$. Determine an expression for the coherence function, $\gamma(\omega)$, of the source.

7. Show that the power spectrum of a Gaussian pulse

$$G(t) = Ae^{(-at^2 - i\omega_0 t)}$$

is also a Gaussian pulse centered at the frequency ω_0.

8. The reflectivity of the plates of a Fabry-Perot interferometer is 0.9. Assuming normally incident light find the plate separation required to resolve the H_α doublet: $\lambda_0 = 0.6563\ \mu m$, $\lambda - \lambda' = 0.014$ nm . What is the resulting spectral free range?

9. What should the thickness and refractive index be for an antireflection coating for $\lambda = 0.5\mu m$, s-polarized light incident at 45°? The substrate has a refractive index of 1.5?

10. Find the peak reflectance of a high reflectance stack consisting of a) 4 and b) 16 layers of high-low index materials when $n_L = 1.4$ and $n_H = 2.8$. In the latter case, if the stack is mounted on a substrate of 1.5, by how much does the reflectivity change?

11. A piece of glass ($n = 1.5$) is antireflection coated with a single film for $\lambda_0 = 0.5\mu m$. Will such a coated film still possess a null in its reflectivity for p-polarized light for $\theta \neq 0$? If it does possess an effective Brewster angle, what is its value?

CHAPTER 8

Diffraction Phenomena

Light breaks where no sun shines
Dylan Thomas

In many of the previous chapters we have been concerned with the propagation of plane waves in infinite homogeneous media, and, with the exception of the interface problem in chapter 4 and the geometrical optics limit of chapters 5 and 6, we have not considered beams of a finite transverse extent or the interaction of plane waves with objects of a finite transverse extent. In the geometrical optics limit we considered situations in which w, the width of the beam or object, satisfies the condition $\lambda/w \ll 1$. Departures from this approximation give rise to diffraction or beam spreading effects in light propagation. Since any beam of finite extent or any segment of a beam can be treated as a superposition of plane or spherical waves, the phenomenon of diffraction can be viewed as the interference which occurs between these different components on propagation. Indeed there is no physical distinction between interference and diffraction. Which phenomena are classified as interference or diffraction is, to some extent, just a matter of taste and history.

A simple argument can be used to estimate the spreading of a beam when it is incident on an object of finite transverse extent by using the $\Delta k \Delta x \approx 2\pi$ result from Fourier analysis discussed in the second chapter. Consider a plane wave incident along the z-axis on a flat screen which has a small square hole of width w on it as shown in Figure 8.0.1. If we consider spreading along one axis only (taken to be the x-axis) then, because $\Delta x \Delta k_x \approx 2\pi$, the angular width of the spreading beam in the far field is $\Delta\theta = \Delta k_x/k_z \approx 2\pi/w(2\pi/\lambda)^{-1} = \lambda/w$. The details of the field distribution in the far field await a more rigorous treatment of diffraction.

The essential features of diffraction phenomena can be explained by a principle put forward by Christian Huygens in 1678 and named after him. The Huygens principle states that the propagation of a light wave can be predicted by assuming that each point on the wavefront acts as the source of a secondary wave which spreads in all directions. The envelope of all the secondary sources is the new wave front. This simple principle can explain such effects as propagation of light in free space and the law of reflection and refraction (provided one accounts for the different phase speeds of light in the two media). Huygens principle, however, is not able to account for diffraction of light from a grating or the fuzziness of shadows cast by opaque objects. The failure in the Huygens approach is that it does not account for interference between the different spherical wave components into which the beam can be decomposed. In 1810 Fresnel added interference to Huygens principle so that a more correct (albeit, still qualitative) formulation

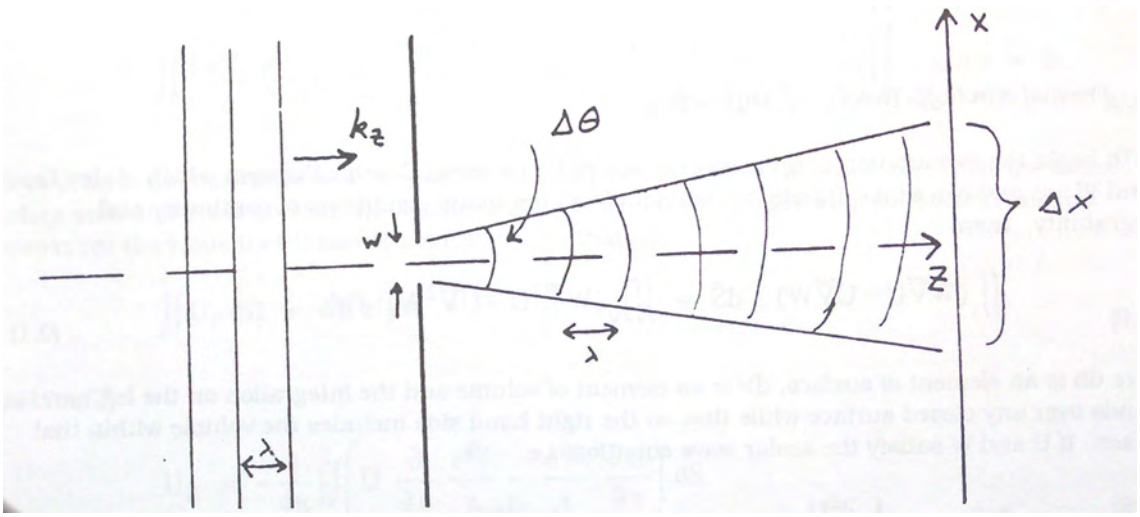


FIGURE 8.0.1. Diffraction of a plane wave from a square aperture.

could be obtained. The Huygens-Fresnel principle of diffraction states that every unobstructed point of a wavefront, at a given instant in time, serves as a source of secondary wavelets (of the same frequency as the primary wave). The amplitude of the optical field at any point beyond is the superposition of all these wavelets considering their amplitude and phase. Gustav Kirchhoff in the 1880's put this principle on a mathematical foundation by solving the appropriate differential equations which describe light propagation and indeed showed that the principle was a consequence of the wave equation. However, even Kirchhoff's theory is not rigorous. In formulating his theory Kirchhoff (as did Fresnel and Huygens) assumed scalar waves, ignoring the fact that light should be represented by a transverse vector field, and further had to constrain his theory to deal with objects and apertures whose transverse extent is much greater than the wavelength of light. In spite of these assumptions Kirchhoff's scalar theory works well for most cases of interest.

It should be stressed that the determination of an exact solution for many diffraction problems is among the most challenging in classical optics and still the subject of considerable research today (e.g., the problem of light transmission through a subwavelength aperture). The first solution utilizing the electromagnetic theory of light was published by Sommerfeld in 1896 but even he had to introduce unrealistic boundary conditions involving such things as infinitely thin, perfectly conducting, yet opaque screens. Nonetheless, the results have been extremely valuable and very close to being in agreement with experiment in most cases. Because the vector theory of Sommerfeld's only gives significantly different results from Kirchoff's in the limit of very small apertures, or in the case of an object having spatial variations on the scale of the wavelength of light (as in diffraction gratings), we adopt the simpler approach here.

8.1. Fresnel-Kirchoff Theory of Diffraction

To begin the formulation of diffraction in optics let us recall Green's theorem which states that if U and W are any two scalar functions that satisfy all the usual conditions of continuity and integrability, then

$$\int\int_S (W\vec{\nabla}U - U\vec{\nabla}W)\cdot d\vec{S} = \int\int\int_V (W\nabla^2 U - U\nabla^2 W)dV$$

where dS is an element of surface, dV is an element of volume and the integration on the left hand side extends over any closed surface while that on the right hand side includes the volume within that surface. If U and W satisfy the scalar wave equation, *i.e.*

$$\nabla^2 U = \frac{1}{v^2}\frac{\partial^2 U}{\partial t^2} \quad \text{and} \quad \nabla^2 W = \frac{1}{v^2}\frac{\partial^2 W}{\partial t^2}$$

and both functions have a harmonic time dependence of the form $e^{-i\omega t}$, it is easy to show that the right hand integral is zero so that

$$\int\int_S (W\vec{\nabla}U - U\vec{\nabla}W)\cdot d\vec{S} = 0. \tag{8.1.1}$$

Now consider that we take for W the wave function

$$W = \frac{W_0}{r}e^{i(kr-\omega t)}$$

which represents a spherical wave converging on the origin, P, as indicated in Figure 8.1.1.

Let us consider performing the integration of equation 8.1.1 over a surface which includes the origin. Because $\mathcal{E}_W$ becomes infinite at the origin we have to exclude the point P from the integration. This is accomplished by subtracting the integral over a small sphere of radius $r = \varepsilon$. The component of the gradient normal to the sphere is $\partial/\partial r$. It follows that equation 8.1.1 is the same as

$$\int\int_S \left(\frac{e^{ikr}}{r}\frac{\partial U}{\partial n} - U\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)\right)dS - \int\int_{S_\varepsilon}\left(\frac{e^{ikr}}{r}\frac{\partial U}{\partial r} - U\frac{\partial}{\partial r}\left(\frac{e^{ikr}}{r}\right)\right)\Bigg|_{r=\varepsilon}\varepsilon^2 d\Omega = 0$$

where ∂/∂n is the derivative with respect to the local normal of the outer surface, dS is the element of surface area and dΩ is an element of solid angle. As $\varepsilon \to 0$ the integrand of the second integral approaches the value that U has at point P, U_{P}, so that

$$\int\int U_P d\Omega = 4\pi U_P$$

and finally

$$U_P = \frac{1}{4\pi}\int\int_S \left(\frac{e^{ikr}}{r}\frac{\partial U}{\partial n} - U\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)\right)dS. \tag{8.1.2}$$

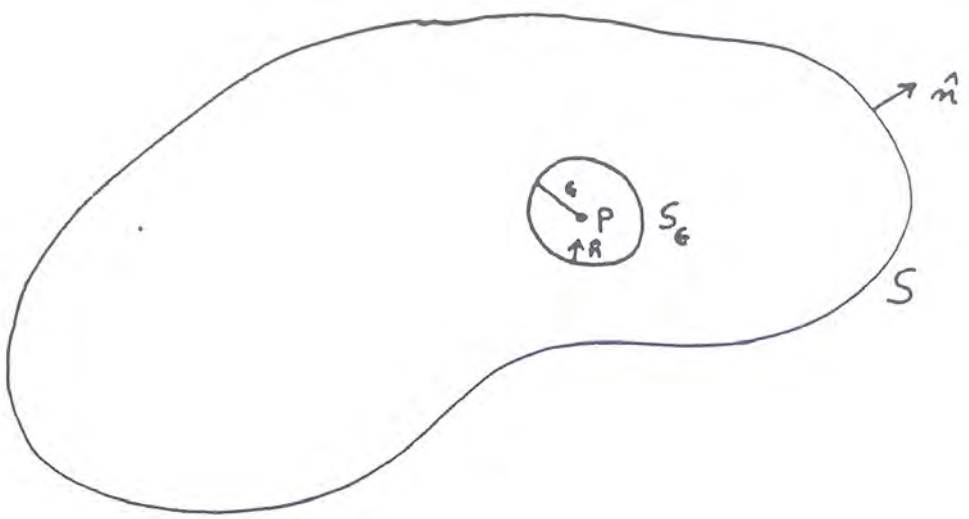


FIGURE 8.1.1. Surface of integration for proving the integral theorem.

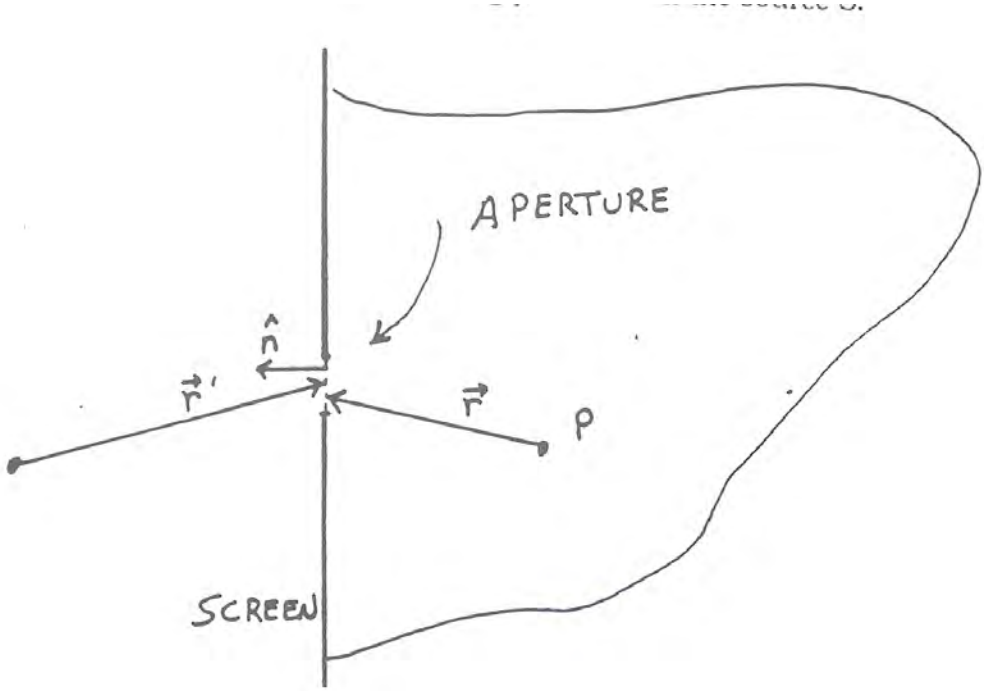


FIGURE 8.1.2. Geometry for the Fresnel-Kirchoff formula.

This is known as the Kirchoff integral theorem. It relates the value of any scalar wave function at any point P inside an arbitrary closed surface to the value of the function at the surface.

We now show how to apply Kirchoff's integral theorem to the general problem of the diffraction of light. We consider that the diffraction is produced by an aperture of arbitrary shape in an otherwise opaque, infinitely thin partition as indicated in Figure 8.1.2. The screen is taken to separate a source from a receiving point. Our task is to determine the optical disturbance or light field reaching the receiving point P from the source S.

In applying the Kirchoff integral, we choose a surface of integration such that it encloses the receiving point, and includes, as part of it the aperture as indicated. Two simplifying assumptions are introduced (following Kirchoff):

1) the function U and $\vec{\nabla}U$ contribute negligible amounts to the integral except at the aperture opening itself.

2) the values of U and $\vec{\nabla}U$ at the aperture are the same as they would be in the absence of the aperture.

The validity of these assumptions is open to debate. For example, it is not clear how an infinitely thin screen could satisfy the first condition and why edge effects in the aperture should be discounted as indicated in the second condition. The assumptions are also unrealistic in that they specify both the function and its gradient to be zero over a large area. This would imply that U is zero everywhere when the wave equation is strictly solved. Nonetheless, the result of applying the assumptions together is to give a solution in agreement with experiment, and isn't that the object of any theory?

If $\vec{r}'$ denotes the position of an arbitrary point on the aperture relative to the source S then, with the time dependence included, we have for a point source a scalar wave of the form

$$U = \mathcal{E} = \frac{U_0}{r'} e^{i(kr' - \omega t)}$$

and

$$\mathcal{E}_P = \frac{U_0}{4\pi} e^{-i\omega t} \int\int_A \left(\frac{e^{ikr}}{r} \frac{\partial}{\partial n} \left(\frac{e^{ikr'}}{r'} \right) - \frac{e^{ikr'}}{r'} \frac{\partial}{\partial n} \left(\frac{e^{ikr}}{r} \right) \right) dS$$

where the integration extends only over the aperture opening. From the modulus squared of $\mathcal{E}_P$ we can obtain the intensity of the light wave. The aperture is any surface which encloses the opening and not necessarily the surface of smallest area which would be a planar surface for a flat screen. In arriving at this equation it is easy to show that the integral over the outer surface vanishes in the limit of its area going to infinity. This is due to the cancellation of the two terms contributing to the integral of equation 8.1.2 over the outer surface. But now

$$\frac{\partial}{\partial n} \left(\frac{e^{ikr}}{r} \right) = \cos(\hat{n}, \vec{r}) \frac{\partial}{\partial r} \left(\frac{e^{ikr}}{r} \right) = \left(\frac{ike^{ikr}}{r} - \frac{e^{ikr}}{r^2} \right) \cos(\hat{n}, \vec{r})$$

$$= \left(\frac{ike^{ikr}}{r} \right) \cos(\hat{n}, \vec{r}) \quad \text{for } kr \gg 1$$

where $(\hat{n}, \vec{r})$ denotes the angle between the vectors enclosed in the brackets.

In a similar fashion we can evaluate

$$\frac{\partial}{\partial n} \left(\frac{e^{ikr'}}{r'} \right)$$

so that upon collecting all the terms we have

$$\mathcal{E}_P = \frac{-ikU_0}{4\pi} e^{-i\omega t} \int\int_A \frac{e^{ik(r+r')}}{rr'} \left(\cos(\hat{n}, \vec{r}) - \cos(\hat{n}, \vec{r}') \right) dS.$$

This equation is known as the Fresnel-Kirchoff integral formula and is basically a mathematical statement of the Huygens-Fresnel principle. This is most easily seen by applying the formula to a specific case, namely to that of a circular aperture with the source symmetrically located as shown in figure 8.1.3. The surface of integration is taken to be a spherical cap which covers the opening. In this case r' is a constant and $\cos(\hat{n}, \vec{r}') = -1$. The Fresnel-Kirchoff formula then reduces to

$$\mathcal{E}_P = \frac{-ik}{4\pi} e^{-i\omega t} \int\int_A \mathcal{E}_A \frac{e^{ikr}}{r} \left(\cos(\hat{n}, \vec{r}) + 1 \right) dS \tag{8.1.3}$$

where

$$\mathcal{E}_A = \frac{U_0}{r'} e^{ikr'}.$$

Equation 8.1.3 can be given the following simple interpretation: $\mathcal{E}_A$ is the complex amplitude of the incident primary wave at the aperture. From this primary wave, each element of area of the aperture, dA, gives rise to a secondary spherical wave of the form

$$\mathcal{E}_A Q(\vec{r}, \vec{r}') \frac{e^{i(kr - \omega t)}}{r} dA$$

where $Q(\vec{r}, \vec{r}') = [\cos(\hat{n}, \vec{r}) - \cos(\hat{n}, \vec{r}')]/2$ is known as the *obliquity factor*. The total optical disturbance at the receiving point P is obtained by summing the secondary waves from each element. In the summation it is important to take into account the obliquity factor. In the case under discussion here, the obliquity factor in the forward direction is equal to unity, but in the backward direction, where $\cos(\hat{n}, \vec{r}) = -1$, the obliquity factor is zero. This explains why there is no backward propagating wave created by the original optical wavefront. Huygens' principle, among other things, did not include this factor, and indeed had to assume the direction of advancement of the wavefront. The presence of the factor -i means that the diffracted waves are shifted in phase by $-\pi/2$ with respect to the undiffracted wave.

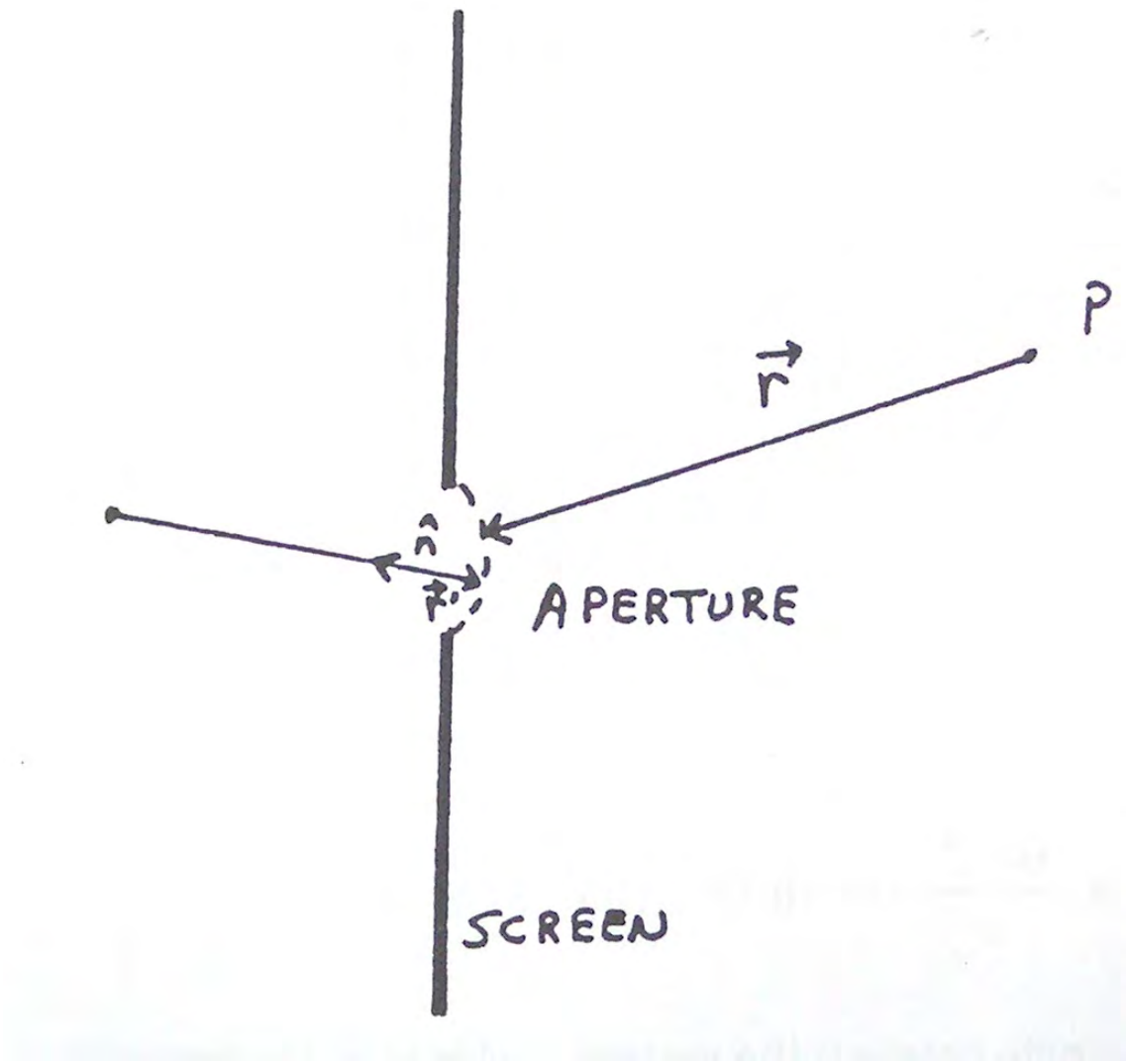


FIGURE 8.1.3. Diagram to show how Huygen's principle follows from the Kirchoff integral formula.

The integral of equation 8.1.3 is of the form

$$\mathcal{E}_P = \frac{-i}{2\lambda_0} \int\int_A \mathcal{E}_{P'} \frac{Q(\vec{r}) e^{i(kr-\omega t)}}{r} dA$$

where $\mathcal{E}_{P'}$ is related to the optical disturbance at point P' on the aperture. The integral can also be written in the form

$$\mathcal{E}_P = \int\int_A \mathcal{E}_{P'} h(P, P') dA.$$

It can be seen that the optical disturbance at the observation point is the result of a superposition of point source, or impulse response(or kernel) functions, weighted by the field distribution at the aperture.

8.2. Babinet's Principle

Consider a diffracting aperture, A, that produces a certain optical disturbance U_P at a given observation point P. Suppose, now, that the aperture is divided into two portions A_1 and A_2, such that

$$A = A_1 + A_2.$$

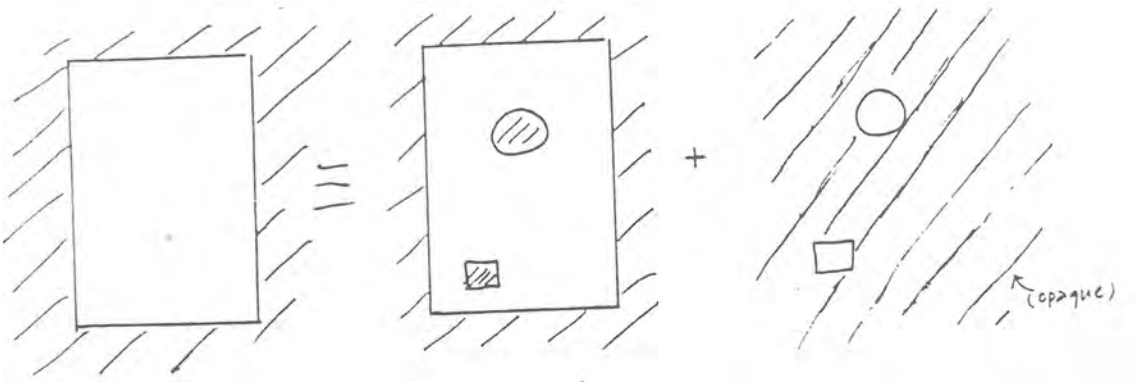


FIGURE 8.2.1. Complementary apertures.

The two apertures are said to be complementary relative to the overall aperture A. An example is shown in Figure 8.2.1. From the form of the Fresnel-Kirchoff formula, it is clear that

$$\mathcal{E}_P = \mathcal{E}_{1P} + \mathcal{E}_{2P}$$

where $\mathcal{E}_{1P}$ is the optical disturbance at P produced by A_1 alone and $\mathcal{E}_{2P}$ is the optical disturbance produced by the aperture A_2 alone. This statement is known as *Babinet's principle*. For example consider a) a fine wire, and b) a fine slit with the same width. These two apertures are complementary in the infinite open aperture. A plane wave passes undisturbed through the infinite aperture. Thus from Babinet's principle, the field produced by a plane wave diffracted by a wire and the field produce by a plane wave diffracted by its complementary slit add up to a non-diffracted plane wave so

$$1 = \mathcal{E}_{slit} + \mathcal{E}_{wire}.$$

The diffraction patterns from the slit and wire are thus quite similar.

8.3. Fresnel and Fraunhofer Diffraction

The general expression represented by equation 8.1.3 is of little practical interest and is much too cumbersome to use in most cases. Two approximations to the integral, however, have been found to be very useful and can be used to represent most situations of interest. These are known as the Fresnel and Fraunhofer approximations. In the Fresnel approximation the distances from the source and observation points to the diffracting aperture are such that one can approximate the spherical waves with surfaces of constant phase which are quadratic in the transverse variables. In the Fraunhofer approximation the distances in the problem are such that the incident and diffracted waves can be approximated by plane waves. This is valid if the distances from the source and observation points to the aperture are such that one can neglect the curvature of the wave fronts. There is no sharp distinction between these two approximations and in general one must be careful in applying either. The general idea, however, is represented by Figure 8.3.1.

An additional approximation, independent of the approximations made to the form of the wave fronts, is the paraxial approximation. It turns out that in many situations we can remove the obliquity factor if we can approximate $\cos(\hat{n}, \vec{r})$ by unity. This is valid if the vector $\vec{r}$ makes an angle of approximately less than $\pi/6$ with respect to a vector normal to the aperture plane. If we choose a Cartesian co-ordinate system as depicted in Figure 8.3.2 to label the observation point, then the approximation is equivalent to having the observation point close to the z-axis; hence the name paraxial approximation.

Provided the z co-ordinate of the observation point P is much larger than the width of the aperture this is an excellent approximation. An additional benefit of the paraxial approximation is that it allows us to replace r by z in the denominator of equation 8.1.3 so that if (x_0, y_0) labels the plane containing the aperture and z is the distance to the point of observation, we find that the impulse response function can be taken as

$$h(x, y; x_0, y_0) = \frac{e^{ikr}}{z}.$$

Note that because $k \sim 10^6\, cm^{-1}$, it would be a serious mistake to replace r by z in the exponent since even if $|r - z| = 1\,\mu m$ one would be making an error in the phase of 2π ! The main thrust of the Fresnel and Fraunhofer approximations is to make more reasonable approximations for the rapidly varying phase term in the *impulse response function*.

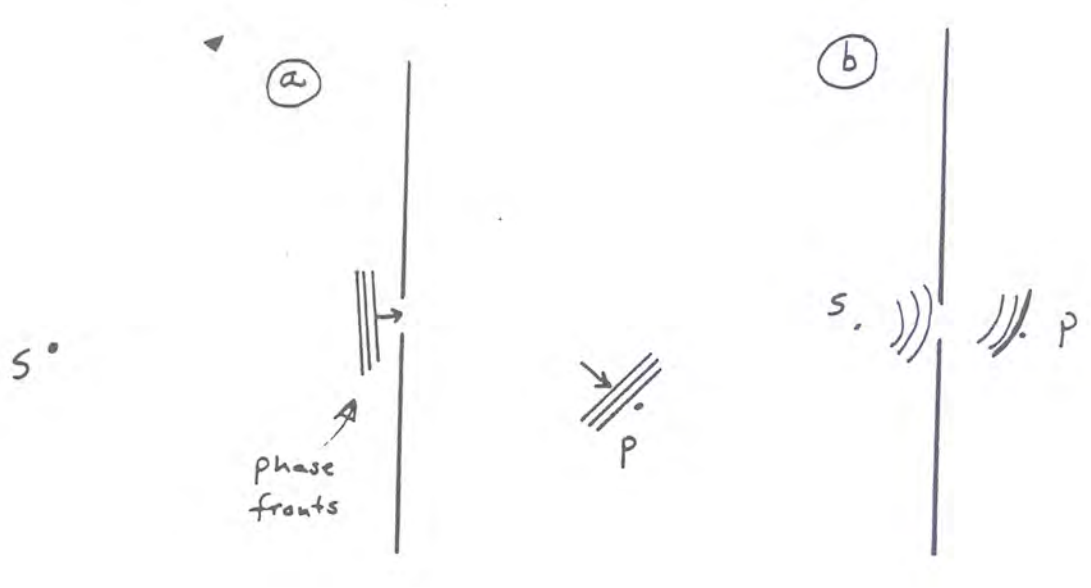


FIGURE 8.3.1. Diffraction by an aperture in a) the Fraunhofer case and b) the Fresnel case.

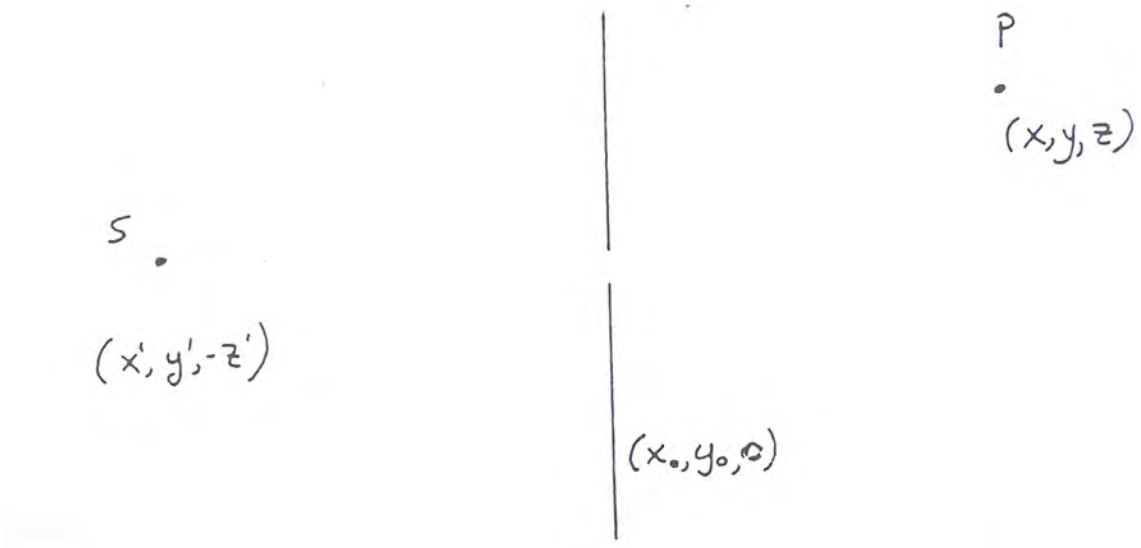


FIGURE 8.3.2. Co-ordinate system used to describe diffraction from an aperture.

8.4. Fresnel Approximation

In general

$$r = \sqrt{z^2 + (x - x_0)^2 + (y - y_0)^2}$$

and this can be rewritten as

$$r = z\left(1 + \frac{(x - x_0)^2}{z^2} + \frac{(y - y_0)^2}{z^2}\right)^{1/2}.$$

However we know that for small b, we can perform the expansion

$$\sqrt{1 + b^2} \simeq 1 + \frac{1}{2}b^2 - \frac{1}{8}b^4. \tag{8.4.1}$$

If we retain only the leading two terms of the expansion of equation 8.4.1 we have in a similar approximation for r (*the Fresnel approximation*) that

$$r \simeq z\left(1+\frac{1}{2}\frac{(x-x_0)^2}{z^2}+\frac{1}{2}\frac{(y-y_0)^2}{z^2}\right)$$

and the (Fresnel) impulse response function becomes

$$h_F(x,y;x_0,y_0)=\frac{1}{z}e^{ikz}\exp(\frac{ik}{2z}\left[(x-x_0)^2+(y-y_0)^2\right]$$

in which the Huygens spherical wavelets have been replaced by wavelets with quadratic surfaces of constant phase.

A similar approximation can be made for the source wave. If the source is a point source located as shown in Figure 8.1.3 we can, within the spirit of the present approximation, let

$$\frac{e^{ikr'}}{r'}=\frac{1}{z'}e^{ikz'}\exp(\frac{ik}{2z'}\left[(x'-x_0)^2+(y'-y_0)^2\right]$$

where *(x', y', -z')* labels the source point and *z'* is the distance of the aperture plane from the source.

If one were to use these approximations for the source and diffracted waves in equation8.1.3, one would still have a cumbersome integral to evaluate. However, a little algebraic manipulation greatly simplifies the problem. As far as the *(r + r')* term in the exponent of equation 8.1.3 is concerned, we can transform it as follows:

$$\begin{aligned} r+r' &= z+z'+1+\frac{1}{2}\frac{(x-x_0)^2+(y-y_0)^2}{z}+\frac{1}{2}\frac{(x'-x_0)^2+(y'-y_0)^2}{z'} \\ &\simeq z+z'+1+\frac{1}{2}\frac{(x-x')^2+(y-y')^2}{z+z'}+\frac{1}{2}\frac{(x_m-x_0)^2+(y_m-y_0)^2}{z_a} \end{aligned} \tag{8.4.2}$$

where

$$x_m=\frac{zx'+z'x}{z+z'} \qquad y_m=\frac{z'y+zy'}{z+z'}$$

and

$$z_a=\frac{zz'}{z+z'}.$$

The first bracketed term in equation 8.4.2 is a good approximation for the distance from the source to the observation point if one retains the leading terms in the expansion of

$$\sqrt{(z+z')^2+(x-x')^2+(y-y')^2}=|PS|\,.$$

One then has from equation 8.1.3 that the spatial part of the field at the observation point (*i.e.*, dropping the time dependence) is:

$$\mathcal{E}_P(x,y,z)=-\frac{iU_0}{\lambda_0 zz'}e^{ik|PS|}\int\int_A \exp\left(\frac{ik}{2z_a}\left[(x_m-x_0)^2+(y_m-y_0)^2\right]\right)dx_0dy_0. \tag{8.4.3}$$

This is known as the *Fresnel approximation to the Kirchoff integral formula.*

It only remains to interpret the meaning of $\vec{r}_m=(x_m,y_m,0)$. This is easily seen to be the intersection point of the aperture plane and the line that connects the source and observation points.

As a sufficient condition for accuracy in the Fresnel approximation we might require that the maximum phase change contributed by the next higher order term in the expansion of r (or r') be $<<1$ radian. For the observation point this implies

$$z^3 \gg \frac{\pi}{4\lambda_0}\left[(x-x_0)^2+(y-y_0)^2\right]^2.$$

If $|x-x_o|=|y-y_o|=1$ mm, as would occur, for example, if we are at an observation point on the axis of an aperture 1 mm in radius, and $\lambda = 1\mu$m then

$$\frac{z}{\triangle x}\gg\left(\frac{\triangle x}{\lambda_0}\right)^{1/3};\quad z\gg 1cm.$$

This condition is rather more severe than generally necessary. The key lies in the fact that the condition for accuracy is sufficient but not necessary. It turns out that there is considerable cancellation in the neglected contributing phase terms so that the actual distance from the source, for the accuracy quoted, is much less than given by the sufficient criterion. A distance an order of magnitude smaller is not atypical but, once again, each case must be addressed on its own merits.

Let us consider applying the Fresnel approximation to two simple cases, diffraction from rectangular and circular apertures. In the case of the rectangular aperture we assume that it is illuminated by a plane wave ($z' = \infty$) while for the circular aperture we assume illumination by a point source in which the curvature of the phase front cannot

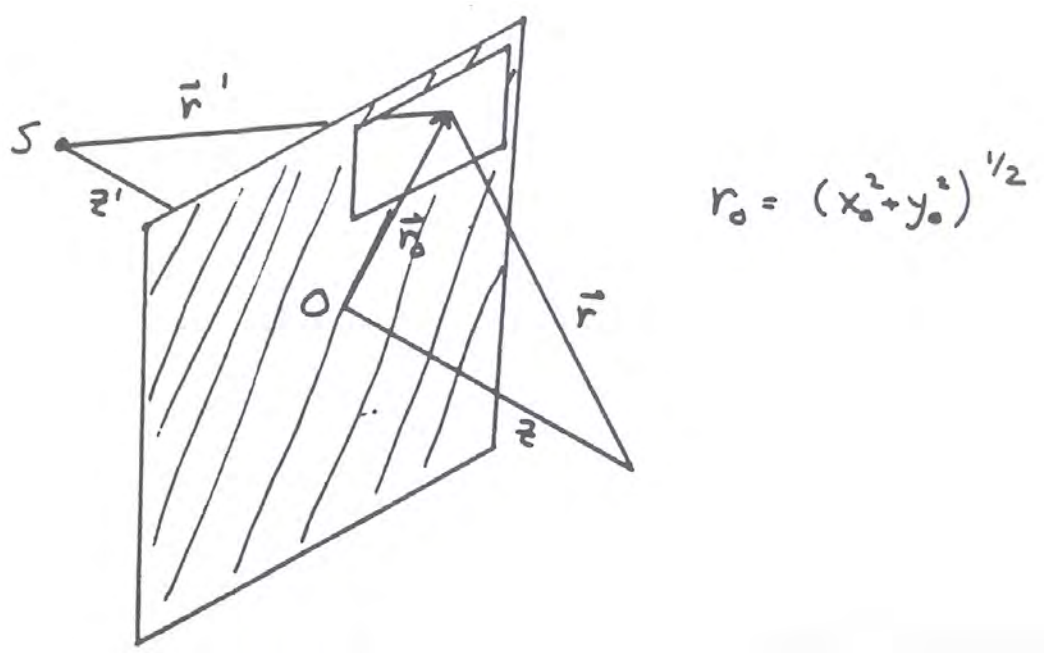


FIGURE 8.4.1. Geometry for deducing the Fresnel diffraction pattern from a square aperture.

be neglected across the aperture. Both cases are used to illustrate the basic physics and complexities(!) of working in the Fresnel limit.

In the case of diffraction from a rectangular aperture, let the rectangle be defined by the region $x_1 < x_0 < x_2$ and $y_1 < y_0 < y_2$ as indicated in Figure 8.4.1. There is no loss in generality in taking the observation point P on the z-axis, since only the relative location of the aperture and the observation point influences the diffraction pattern.

In this case we have, from equation 8.4.2 with $z' \to \infty$ and $(x, y) = (0, 0)$, that

$$r \equiv z + \frac{[x_0^2 + y_0^2]}{2z}$$

and the Fresnel-Kirchoff formula takes the form

$$\mathcal{E}_P = B \int_{x_1}^{x_2} \int_{y_1}^{y_2} \exp\left[\frac{ik(x_0^2 + y_0^2)}{2z}\right] dx_0 dy_0 = B \int_{x_1}^{x_2} \exp\left[\frac{ikx_0^2}{2z}\right] dx_0 \int_{y_1}^{y_2} \exp\left[\frac{iky_0^2}{2z}\right] dy_0$$

where B is a complex number which includes all the constant factors of little interest to us here. Upon introducing the dimensionless variables

$$u = x_0 \left(\frac{k}{\pi z}\right)^{1/2} \quad \text{and} \quad v = y_0 \left(\frac{k}{\pi z}\right)^{1/2}$$

we can write the integrals in the form

$$\mathcal{E}_P = U_1 \int_{u_1}^{u_2} e^{\frac{i\pi u^2}{2}} du \int_{v_1}^{v_2} e^{\frac{i\pi v^2}{2}} dv \tag{8.4.4}$$

where $U_1 = B\pi z/k$. The complex integrals of equation 8.4.4 can be evaluated from the integral

$$\int_0^s e^{\frac{i\pi w^2}{2}} dw = C(s) + iS(s)$$

where

$$C(s) = \int_0^s \cos(\frac{1}{2}\pi w^2) dw \quad \text{and} \quad S(s) = \int_0^s \sin(\frac{1}{2}\pi w^2) dw. \tag{8.4.5}$$

The function *C(s)* and *S(s)* are known as the *Fresnel integrals* and cannot be obtained analytically as functions of *s*. Since the C and S functions always occur together in Fresnel diffraction problems, a useful graphical representation of the Fresnel integrals can be obtained by plotting *S(s)* versus *C(s)* as a function of the parameter *s*. This is illustrated in Figure 8.4.2 which indicates a spiral in a parametric representation. The spiral is known as the *Cornu*

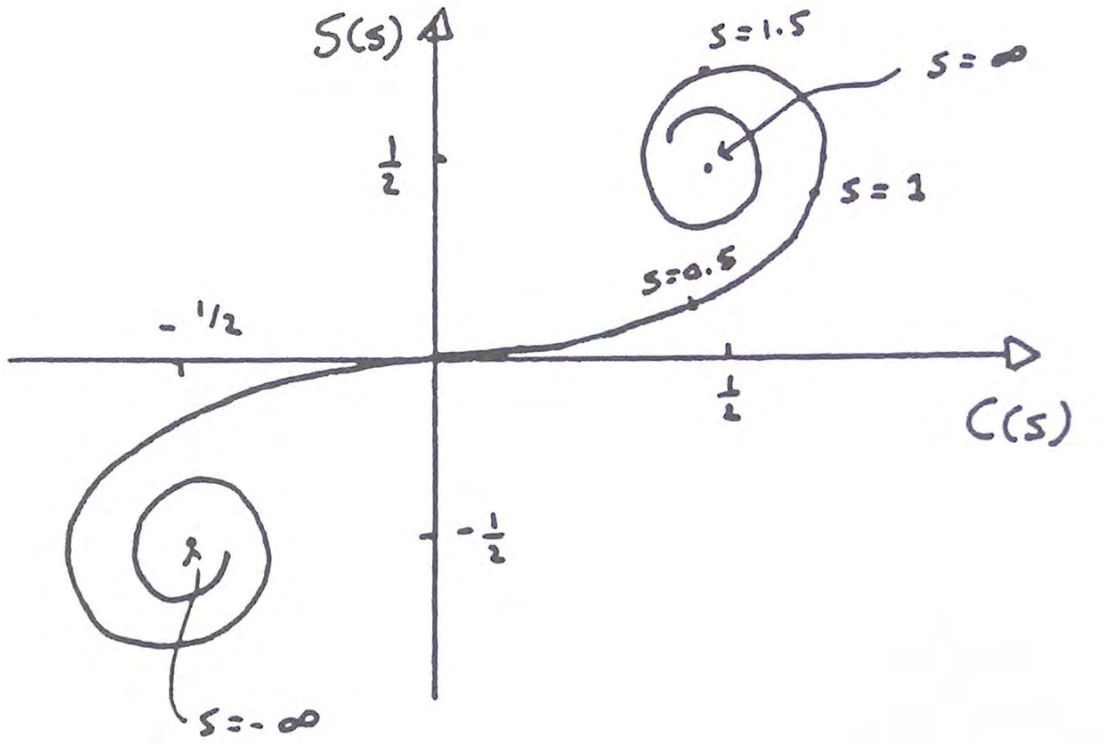


FIGURE 8.4.2. The Cornu spiral as a function of the parameter s It is the vibration curve for rectangular apertures. (cf. Fig.8.4.1, equation 8.4.5).

Spiral. This curve is an example of what is known as a *vibration curve.* These curves in general are parametric representations of the integral of equation 8.4.3 for a particular geometry. The Cornu spiral, as are all vibration curves, is useful for graphical evaluation of diffraction integrals.

Since $C(\infty) = S(\infty) = 1/2$, we obtain the value $U_1(1+i)2$ for the unobstructed optical disturbance at the observation point. Setting this equal to $\mathcal{E}_{P_o}$ we obtain for the general case

$$\mathcal{E}_P = \frac{\mathcal{E}_{P_0}}{(1+i)^2}\left[C(s)+iS(s)\right]\Big|_{u_1}^{u_2} \times \left[C(s)+iS(s)\right]\Big|_{v_1}^{v_2}.$$

The definite integrals can be obtained from the graphical representation as shown in Figure 8.4.3.

The real and imaginary parts of the complex Fresnel integral can be read off as the horizontal and vertical components of the vector which connects the appropriate s parameters on the Cornu spiral.

Diffraction from an infinite slit and a straightedge can be obtained as special cases from the general treatment above. A long horizontal slit can be treated by letting $u_1 = -\infty$ and $u_2 = \infty$. This gives the expression

$$\mathcal{E}_P = \frac{\mathcal{E}_{P_0}}{(1+i)}\left[C(s)+iS(s)\right]\Big|_{v_1}^{v_2}$$

where the v_1 and v_2 define the edges of the slit. For a straightedge (a semi-infinite obstacle) $v_1 = -\infty$ and

$$\mathcal{E}_P = \frac{\mathcal{E}_{P_0}}{(1+i)}\left[C(s)+iS(s)\right]\Big|_{-\infty}^{v_2} = \frac{\mathcal{E}_{P_0}}{(1+i)}\left[C(v_2)+iS(v_2)+\frac{1}{2}+\frac{i}{2}\right].$$

A graphical representation of the intensity associated with the diffraction pattern is shown in figure 8.4.4.

The graph is illustrated as a function of v_2, but can equally well be thought of as a function of y_2. Note that if the observation point is exactly in the geometrical shadow then $v_2 = 0$ and the intensity is one-quarter of the unobstructed value. It can be seen that the intensity falls off monotonically in the shadow zone but oscillates with diminishing amplitude in the illuminated zone. The highest intensity occurs for $v_2 \approx 1.25$ where the intensity is 1.37 times the intensity of the unobstructed beam. The region of oscillating intensity is most pronounced for long wavelengths or small distances from the obstacle.

Fresnel diffraction from circular apertures produces some surprising results. In evaluating the diffraction pattern, we consider a point source located on the axis of a circle of radius R and take advantage of the circular symmetry

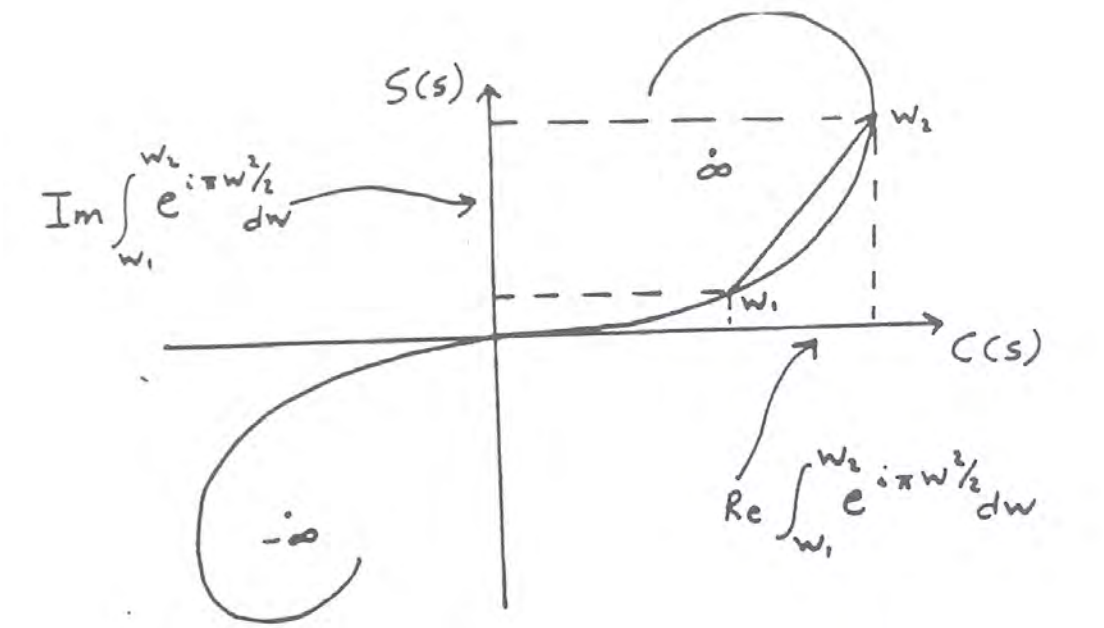


FIGURE 8.4.3. Evaluation of Fresnel integrals using Cornu spirals.

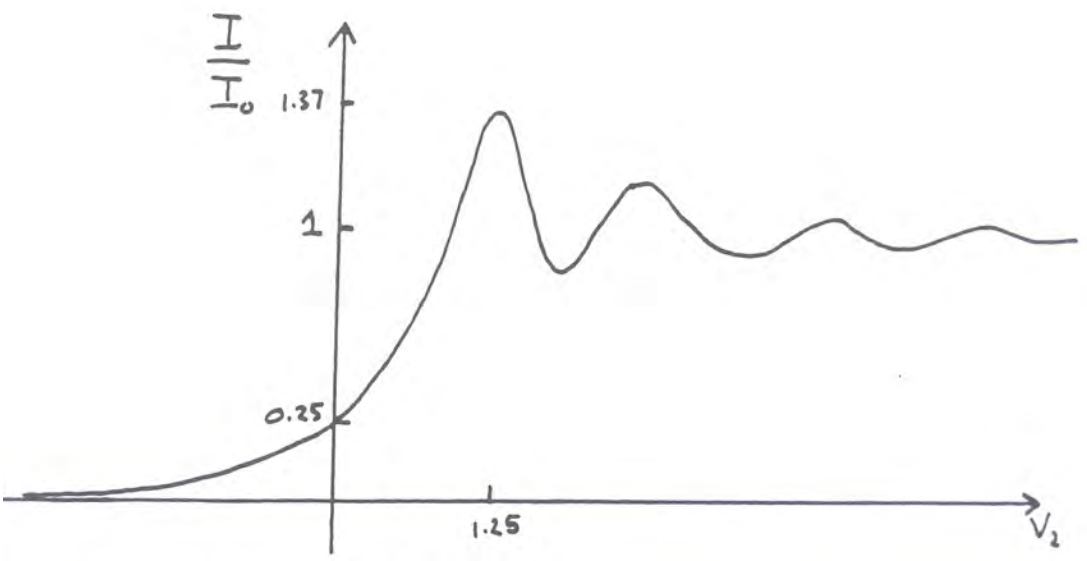


FIGURE 8.4.4. Fresnel Diffraction by a straightedge.

in the problem. Without loss of generality we take the observation point P to lie in the x-y plane. The geometry is indicated in figure 8.4.5.

In performing the calculation we rewrite equation 8.4.3 in the form

$$\mathcal{E}_P = -\frac{iU_0}{zz'}\exp(ik\,|PS|)\frac{z_0}{2}\int\int \exp(\frac{\pi i}{2}\,|\eta|^2)Qd\eta_x d\eta_y$$

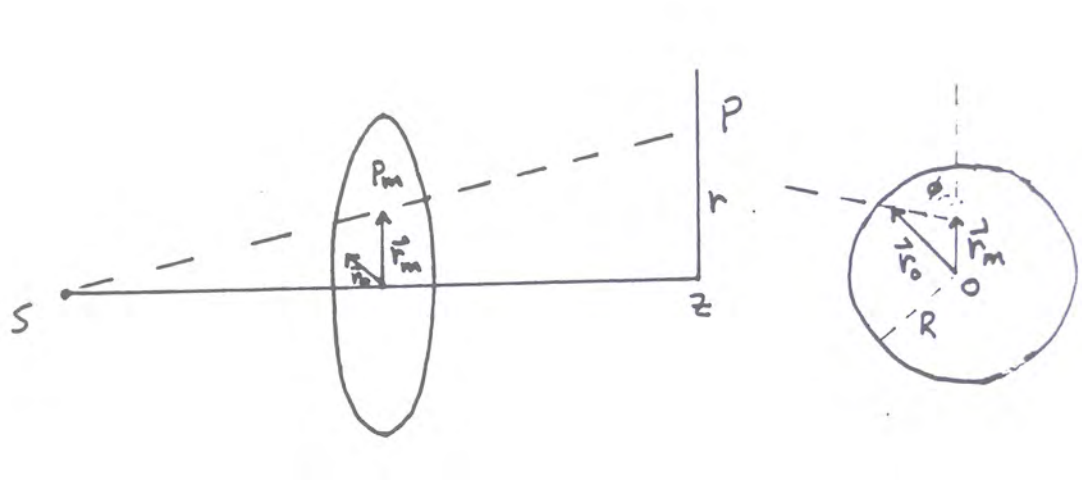


FIGURE 8.4.5. Geometry for calculating diffraction from a circular aperture.

where the obliquity factor has been reintroduced and

$$\vec{\eta} = \left(\frac{2}{\lambda_0 z_a}\right)^{1/2} (\vec{r} - \vec{r}_m).$$

We can collect all the terms in front of the integral into a constant, $-i\mathcal{E}_{P_o}/2$, where $\mathcal{E}_{P_o}$ is the optical disturbance recorded at P in the absence of the aperture. The exponential in the integral can be written as $e^{i\Psi}$ where $\Psi = \pi/2|\boldsymbol{\eta}|^2$. But we also have that

$$d\eta_x d\eta_y = \eta d\eta d\phi$$

and so

$$\mathcal{E}_P = -\frac{i\mathcal{E}_{P_0}}{2} \int\int \exp(i\Psi) Q \eta d\eta d\phi. \tag{8.4.6}$$

However, because of the azimuthal, or circular symmetry in the problem there is no dependence of Ψ on φ and we can immediately eliminate φ from the problem by

$$\eta d\eta \int_0^{2\pi} d\phi = 2\pi\eta d\eta = \pi d\eta^2 = 2d\Psi$$

and we finally obtain

$$\mathcal{E}_P = -i\mathcal{E}_{P_0} \int \exp(i\Psi) Q(\Psi) d\Psi$$

where

$$\Psi_0 = \frac{\pi R^2}{\lambda z_a}.$$

The complex integral

$$I(\Psi_0) = -i \int_0^{\Psi_0} \exp(i\Psi) Q(\Psi) d\Psi$$

is shown in figure8.4.6 .

If the obliquity factor were always unity, the curve is simply a circle. However, the obliquity factor, in general, decreases with increasing Ψ, which itself increases with aperture size. As a result the circle tends to spiral inward to the value *(0, 1)* where for $Q(\infty) = 0$ $I(\infty) = i$. The circle therefore has unit radius. Note that for an infinite aperture for which $\Psi_0 = \infty$ we have that the field intensity is that of the unobstructed wave.

In dealing with the Fresnel approximation it is useful to think of the aperture as being divided up into a number of zones, known as the *Fresnel zones* which are centered on the point P_m. These are defined in such a way that the phase difference Ψ changes by an amount π across a zone and is equal to nπ at the edge of the n'th zone. This is equivalent to a path length change (from source to point on the aperture to observation point) of $\lambda_0/2$ and $n\lambda_0/2$ respectively. Figure 8.4.7 illustrates the Fresnel zones, when the observation point is on-axis. The bright regions give phase contribution with $0 < \Psi < \pi$, while the dark regions give phase regions with $-\pi < \Psi < 0$.

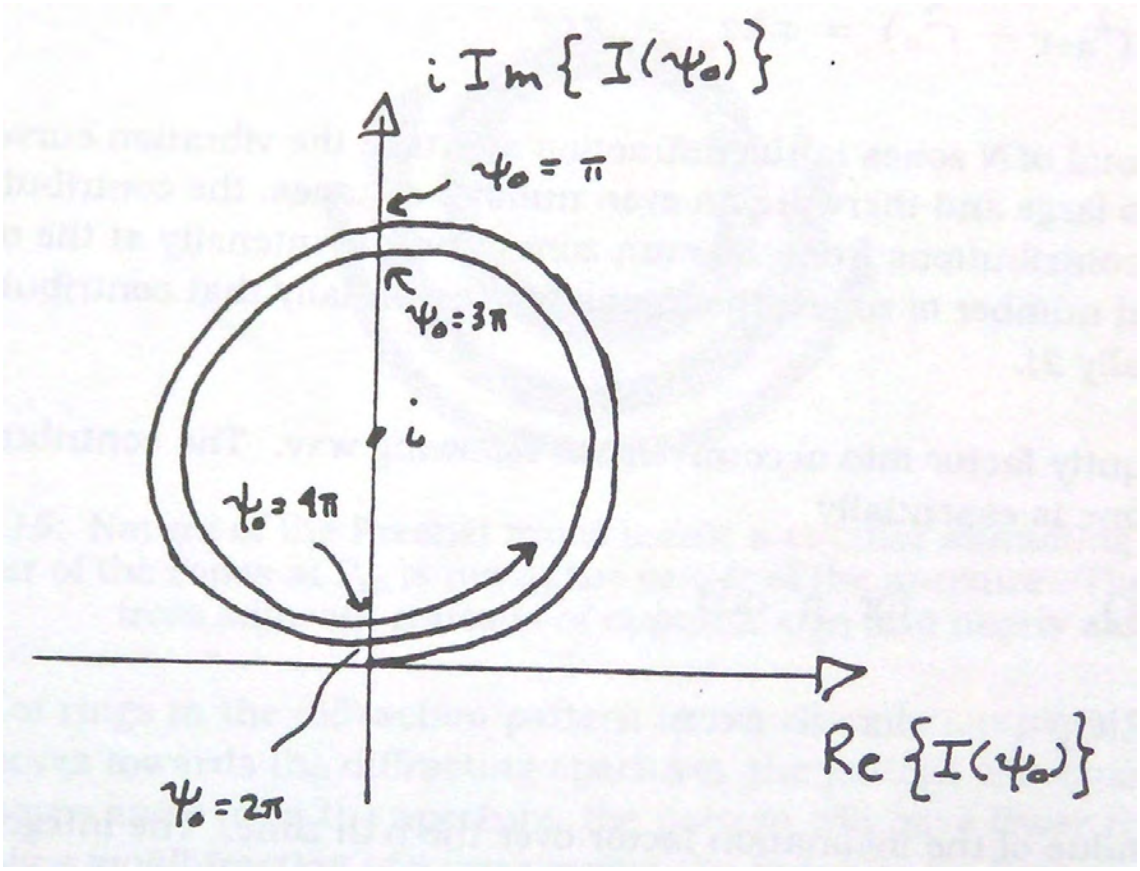


FIGURE 8.4.6. Vibration curve for the circular aperture.

FIGURE 8.4.7. Fresnel zones for a circular aperture with the observation point on axis.

From the definition of Ψ we have that the radius of the n'th zone is given by

$$r_n = \sqrt{n\lambda_0 z_a}.$$

Note that the area of each zone is independent of n and indeed the zone area is given by

$$\triangle\sigma = \pi(r_{n+1}^2 - r_n^2) = \pi\lambda_0 z_a = \pi r_1^2.$$

As we integrate across a total of N zones in the diffraction aperture the vibration curve makes $N/2$ revolutions. If N is not too large and there are an even number of zones, the contribution from the odd zones nearly cancels the contributions from the even zones and the intensity at the observation point is small. If there is an odd number of zones, the intensity is essentially that contributed by the last zone, and I(Ψo) is essentially 2i.

We may take the obliquity factor into account in the following way. The contribution to the integral I from the n'th zone is essentially

$$\begin{aligned} \triangle I_n &= 2iQ_n \quad \text{for } n \text{ odd} \\ &= -2iQ_n \quad \text{for } n \text{ even} \end{aligned}$$

where Q_n is the average value of the inclination factor over the n'th zone. The integral for $I(\psi_0)$is then given approximately by

$$\begin{aligned} I &= \sum_{n=1}^{N} \triangle I_n = 2i\left[Q_1 - Q_2 + Q_3 + ... \pm Q_N\right] \\ &= 2i\left[\frac{Q_1}{2} + \left(\frac{Q_1}{2} - Q_2 + \frac{Q_3}{2}\right) + ... \pm Q_N\right). \end{aligned}$$

Since the terms inside the round brackets are essentially zero we have that the integral is

$$\begin{aligned} I &= i(Q_1 + Q_N) \qquad N \text{ odd} \\ &= i(Q_1 - Q_N) \qquad N \text{ even.} \end{aligned}$$

For large N, Q_N approaches zero and I approaches i, just half the contribution from the first zone!

The Fresnel zones also help us understand qualitatively what happens when P is off-axis so that $r_m \neq 0$. In this case the limits on the variable φ in equation 8.4.6 depend on $|\vec{r}_0 - \vec{r}_m|$ $(= h)$ in a complicated way. But now

$$\begin{aligned} \int d\phi &= 2\pi \qquad \text{for h} < (\text{r}_0 - \text{r}_\text{m}) \\ &= 2\arccos\left[\frac{r_m^2 + h^2 - r_0^2}{2hr_m}\right] \qquad \text{for } (r_0 - r_m) < h < (r_0 + r_m). \end{aligned}$$

We get the full contributions from zones of radius less than $r_0 - r_m$ but only partial contributions from zones of radius between $(r_0 - r_m)$ and $(r_0 + r_m)$ as shown in Figure 8.4.8.

The contribution to the arc length of the vibration curve from each of the partially obstructed zones is proportional to the unobstructed area and hence is less than the value of Q_n that comes from the full n'th zone. The phase difference still changes by π across the zone. The result is that the vibration curve coils up more rapidly.

The number of rings in the diffraction pattern in the observation plane increases as the plane of observation moves towards the diffracting aperture; the pattern expands outward. As the plane of observation moves away from the aperture, the pattern has fewer rings and contract. Finally, when only a small fraction of a zone covers the aperture for P on axis, we pass over to the Fraunhofer limit which is discussed in the next chapter.

For a large aperture diameter R and/or a short distance z', the total number of zones is very large and the details of the diffraction pattern depend on the aperture being very nearly circular. If the circle has a rough edge then the contribution from the last zone is modified and the diffraction pattern depends on these details. In general a smoothing effect in the rings of the intensity distribution takes place and the intensity becomes uniform. This smoothing effect becomes increasingly important as $\lambda \to 0$. In this way the limit of geometrical optics takes over.

The Fresnel zones also allow us to calculate the diffraction pattern from a circular obstacle. In particular let us take a point P on axis. The vibration integral for this case is of the form

$$I = \int_{\Psi_1}^{\infty} \exp(i\Psi)Q(\Psi)d\Psi.$$

The tail of the vector for I is on the spiral where the phase equals Ψ_i and the head is at the point (0,i). Thus if Ψ_i is not too large so that Q is still approximately unity, we find that I should have an absolute value of 1 and the flux density is given by the value it would have if there were no obstacle! This surprising result is strictly speaking a wave phenomenon. As a matter of fact, the existence of this spot is one of the historical confirmations of the wave theory. In 1818 Poisson showed that the wave theory would lead to the "ridiculous" result that there would be a bright spot behind a circular obstacle and offered this as a deathblow to the wave-theory. The existence of this spot was almost immediately verified by Arago. As fate would have it, this spot has come to be known as *Poisson's spot* and not Arago's. More ironically it had actually been observed years earlier by Miraldi in 1723, but his work had gone unnoticed.

FIGURE 8.4.8. Nature of the Fresnel zones inside a circular diffracting aperture when the center of the zones at P_m is not at the center of the aperture. The contributions from adjacent zones is of opposite sign and nearly cancel.

Up to now we have carefully pointed out how successive zone plates tend to nullify each other. This suggests that if we block alternate zones the contribution from the remaining zones reinforce each other and we get stronger intensity at certain points in the observation plane than we did, even with no obstacle. In general a screen which alters the light, either in amplitude or phase, coming from every other half-period zone is called a zone-plate. Suppose that we construct a zone plate which passes only the first 20 odd zones and obstructs the even zones. Since the field contributions from each of the zones are equal, we obtain a field in the observation which is 20 times the field from one zone. But since the field contribution from one zone is twice that of the unobstructed wave, we would obtain an intensity which is 1600 times that of the unobstructed wave! The vibration curve for such a zone plate can be understood with the aid of figure 8.4.9 where it can be seen that the "return" portions of the vibration curves are missing.

Since energy must be conserved it is clear that the zone plate must be concentrating the energy from the source. Indeed, the zone plate operates much like a lens which focuses energy. This can be understood if we consider a general zone plate with a number of equal area zones, and the first zone having a radius of ϱ with

$$r_n = (n\lambda_0 z_a)^{1/2} = \varrho$$

or

$$\frac{n\lambda_0}{\varrho^2} = \frac{1}{z_a} = \frac{1}{z} + \frac{1}{z'}.$$

This equation has the appearance of the Gaussian lens equation with a focal length of

$$f_n = \frac{\varrho^2}{n\lambda_0}.$$

The properties of the zone plate are essentially those of a simple lens except for the multiple values of f. For example, if a plane wave is incident on the zone plate there are bright spots at f_1, $f_1/3$, $f_1/5$ etc. There are also virtual focal spots at the same distances in front of the lens so that the Fresnel zone plate acts simultaneously like a negative and positive lens of multiple focal lengths. A proof of all these properties is beyond the scope of these discussions. Suffice it to say that one can eliminate the multiple focal lengths letting $n = 1$ or by going to a phase zone plate instead of an amplitude zone plate. This latter plate is the type in which actual Fresnel lenses are constructed. These "corrugated" plastic sheets have lightweight and tremendous light gathering power over a large area (up to a square meter in area) and are useful in many non-high resolution imaging situations such as overhead projectors and rear-screen projection TV systems. These zone plates can be made from molded or extruded plastic. The amplitude zone plates are usually made by photolithographic techniques.

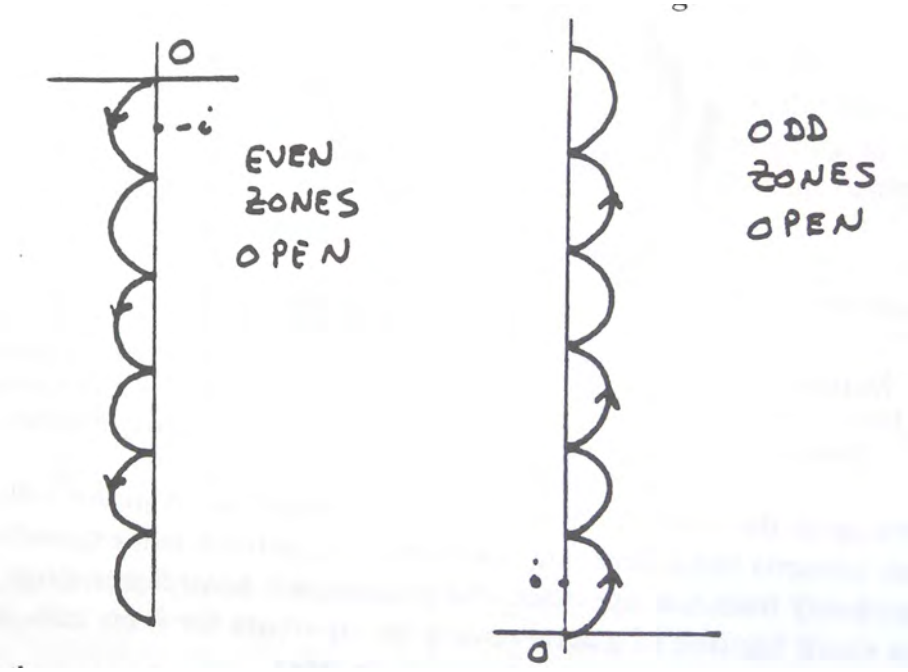


FIGURE 8.4.9. The vibration curve for a Fresnel zone plate having 12 zones with 6 of them open.

References

G.R. Fowles, *Introduction to Modern Optics*, Holt, Reinhard and Winston, Toronto, 1968.
M.V. Klein, *Optics*, John Wiley and Sons, New York, 1970.
M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, New York, 1975.

Problems

1. A point (pinhole) source is to be used in a diffraction experiment. The distance from the source to the diffracting aperture is 5 m. If the aperture is 1 mm in diameter, determine whether Fraunhofer or Fresnel diffraction applies when the screen-aperture distance is a) 10 cm b) 50 cm. Take λ_0 =0.5 µm.

2. An optical beam with a wavelength λ and a Gaussian cross section

$$\mathcal{E}_0(x_0, y_0) = \exp\left[-\frac{x_0^2 + y_0^2}{w_0^2}\right]$$

at $z = 0$ is propagating in free space. Determine the amplitude distribution of the Gaussian beam at $z > 0$ in the Fresnel approximation.

3. Determine the light distribution a) 5 mm, and b) 1 mm beyond a semi-infinite plane illuminated at normal incidence by a $\lambda_0 = 0.5$ µm plane wave.

4. A plane wave of wavelength 1.0 µm is incident on a semi-infinite thin sheet. Determine the distance behind the plane for which the Fresnel approximation is valid and determine the diffraction pattern 1 cm behind the plane.

5. Determine how a "Fresnel lens" is made and determine how it works. These lenses are used extensively in overhead projectors and large rear-projection TV screens.

CHAPTER 9

Fraunhofer Diffraction, Fourier Optics and Holography

All the effects of Nature are only mathematical results of a small number of immutable laws.
S. Laplace

Diffraction calculations (and diffraction patterns) can be simplified tremendously by adopting more stringent approximations than those offered by the Fresnel approximation. In this chapter we explore the limit of Fraunhofer diffraction introduced briefly in the last chapter. Although this approximation is valid for distances away from the aperture that are larger than those offered by Fresnel diffraction, we will see that it is nevertheless useful in a wide variety of cases. Moreover, the mathematical simplifications it offers allow us to introduce Fraunhofer diffraction as a powerful optical processing technique. Specifically, the main result we introduce is that the far field, or Fraunhofer, diffraction pattern is the Fourier transform of the field distribution at the aperture. The mathematical power of the Fourier analysis method leads to innovative ways of treating optical processes in terms of spatial frequencies. It is always exciting to uncover a new bag of analytical tricks but this toolkit becomes even more valuable if it leads to new insights into thinking about a broad range of physical problems. The intent of this chapter is to explore some of the salient features of Fourier optical methods without getting too bogged down in exotic mathematical functions like (ugh!) Bessel functions. We end the chapter with a discussion of Holography—a subject not related to Fourier optics *per se*, but one that is intimately related to diffraction effects.

9.1. The Fraunhofer Approximation

Let us reconsider the formula for the diffracted field given by equation 8.4.3,

$$\mathcal{E}_P(x,y,z) = -\frac{iU_0}{\lambda_0 z z'} e^{ik|PS|} \int\int_A \exp\left(\frac{ik}{2z_a}\left[(x_m-x_0)^2+(y_m-y_0)^2\right]\right) dx_0 dy_0.$$

In particular, consider the quadratic phase factor in the integrand which is given by

$$\frac{ik}{2z_a}\left[(x_m-x_0)^2+(y_m-y_0)^2\right] = \frac{ik}{2z_a}\left[x_0^2+y_0^2+x_m^2+y_m^2-2x_0x_m-2y_0y_m\right].$$

If

$$\frac{k}{2z_a}(x_0^2+y_0^2) \ll 1 \quad \text{or } z_a \gg \frac{\pi(x_0^2+y_0^2)}{\lambda_0} \tag{9.1.1}$$

then we can neglect this contribution to the phase factor. This corresponds to an approximation in which we assume that the source and observation points are far from the aperture. In this case the optical disturbance at the observation point is given by

$$\mathcal{E}_P(x,y,z) = -\frac{iU_0}{\lambda_0 z z'} e^{ik|PS|} \exp\left(\frac{ik}{2z_a}\left[x_m^2+y_m^2\right]\right) \int\int_A \exp\left(\frac{-ik}{z_a}\left[x_0x_m+y_0y_m\right]\right) dx_0 dy_0.$$

In this analysis we have assumed that the source is a point source with an optical field strength at the aperture which is given by $\mathcal{E}_A = U_o/z'$. If we have an extended source the complex field amplitude across the aperture may not be constant so that $\mathcal{E}_A \to \mathcal{E}(x_o, y_o)$. In this case we would have that

$$\mathcal{E}_P(x,y,z) = -\frac{i}{\lambda_0 z} e^{ik|PS|} \exp\left(\frac{ik}{2z_a}\left[x_m^2+y_m^2\right]\right) \int\int_A \mathcal{E}(x_0,y_0) \exp\left(\frac{-ik}{z_a}\left[x_0x_m+y_0y_m\right]\right) dx_0 dy_0.$$

Defining the spatial frequencies

$$u = \frac{kx_m}{z_a} \qquad v = \frac{ky_m}{z_a}$$

we have

$$\mathcal{E}_P(u,v) = -\frac{i}{\lambda_0 z} e^{ik|PS|} \exp\left(\frac{ik}{2z_a}\left[x_m^2+y_m^2\right]\right) \int\int_A \mathcal{E}(x_0,y_0) \exp\left(-i\left[ux_0+vy_0\right]\right) dx_0 dy_0. \tag{9.1.2}$$

The diffracted field distribution in this case is simply related to the field at the aperture through a Fourier transform relation! The approximation given by equation 9.1.1 is known as the *Fraunhofer approximation.* In essence, as can be seen by the lack of quadratic factors in the exponent in the integrand, it allows us to treat all the waves in the problem as plane waves. The quadratic phase factor, and indeed all phase factors outside the integral, are of no consequence since all measurements relate to the intensity of the wave. Equation 9.1.2 is known as the *Fraunhofer diffraction formula* or the far-field approximation.

At optical frequencies the Fraunhofer approximation appears to be quite demanding. For example, in the case of a source with wavelength $\lambda_0 = 6 \times 10^{-7}\, m$ (*e.g.*, a He-Ne laser) and an aperture with a diameter of 1 mm, the approximation states that one can only observe the far-field diffraction pattern if

$$z_a \gg 3m.$$

As with the Fresnel approximation the Fraunhofer approximation is sufficient but not necessary and due to the phase cancellation effects in the higher order neglected terms, the distance from the aperture to the observation plane usually doesn't have to be as great as the criterion for the approximation would dictate.

If the source is at an infinite distance from the aperture plane the Fraunhofer formula becomes even simpler. In this case

$$z_a = z \qquad \text{and} \qquad x_m = x \quad \mathrm{y_m = y}$$

so that

$$u = \frac{kx}{z} = k\sin\theta_x \qquad \text{and} \qquad \mathrm{v = \frac{ky}{z} = k\sin\theta_y}$$

where θ_x and θ_y are direction angles of the observation point relative to the z axis. The fact that u and v are dependent only on these angles and not on the individual *x,y,z* co-ordinates indicates that the far-field diffraction pattern is independent of the distance from the source and only on the direction of observation.

To illustrate the Fraunhofer approximation, we consider the far field diffraction pattern from an $\ell_x \times \ell_y$ rectangular aperture when illuminated by a plane wave ($z' \to \infty$) of uniform amplitude $\mathcal{E}_A$. Without loss of generality we consider the aperture to be centered on the origin with

$$-\frac{\ell_x}{2} < x_0 < \frac{\ell_x}{2}$$

and

$$-\frac{\ell_y}{2} < y_0 < \frac{\ell_y}{2}$$

The diffraction pattern is then given by

$$\begin{aligned}\mathcal{E}_P(u,v) &= -\frac{i}{\lambda_0 z}\exp\left(\frac{ik}{2z_a}\left[x^2+y^2\right]\right)\int_{-\frac{\ell_x}{2}}^{\frac{\ell_x}{2}}\int_{-\frac{\ell_y}{2}}^{\frac{\ell_y}{2}}\mathcal{E}_A\exp\left(-i\left[ux_0+vy_0\right]\right)dx_0dy_0 \\ &= -\frac{i\mathcal{E}_A}{\lambda_0 z}\exp\left(\frac{ik}{2z_a}\left[x^2+y^2\right]\right)\frac{\ell_x}{2}\frac{\ell_y}{2}sinc\left[\ell_x\frac{x}{\lambda_0 z}\right]sinc\left[\ell_y\frac{y}{\lambda_0 z}\right]\end{aligned}$$

where

$$sinc(\alpha) = \frac{\sin(\pi\alpha)}{\pi\alpha}$$

(pronounced "sink" function). The intensity in the diffraction pattern is therefore given by

$$\frac{I(x,y)}{I_0} = \frac{\ell_x^2\ell_y^2}{z^2\lambda_0^2}sinc^2\left[\ell_x\frac{x}{\lambda_0 z}\right]sinc^2\left[\ell_y\frac{y}{\lambda_0 z}\right]$$

where $I_o = |\mathcal{E}_A|^2$.

Figure 9.1.1 shows a graph of the normalized intensity in the diffraction pattern as a function of x. Note that the extent of the central peak (the distance between the first set of nodes) is given by

$$\triangle x = \frac{2\lambda_0 z}{\ell_x} \quad \text{or} \quad \triangle\theta_x = \frac{2\lambda_0}{\ell_x}.$$

This is in agreement with the rough calculation at the beginning of the chapter. Detailed agreement would be achieved if a rigorous definition of angular width were used for both calculations. Note that the diffracted beam intensity appears to contradict conservation of energy by varying as the square of the aperture area ($\ell_x\ell_y$). This is not really the case since the extent of the diffraction pattern in the $x-y$ plane varies as $(\ell_x\ell_y)^{-1}$. The integrated energy therefore, varies as $(\ell_x\ell_y)$, as it should.

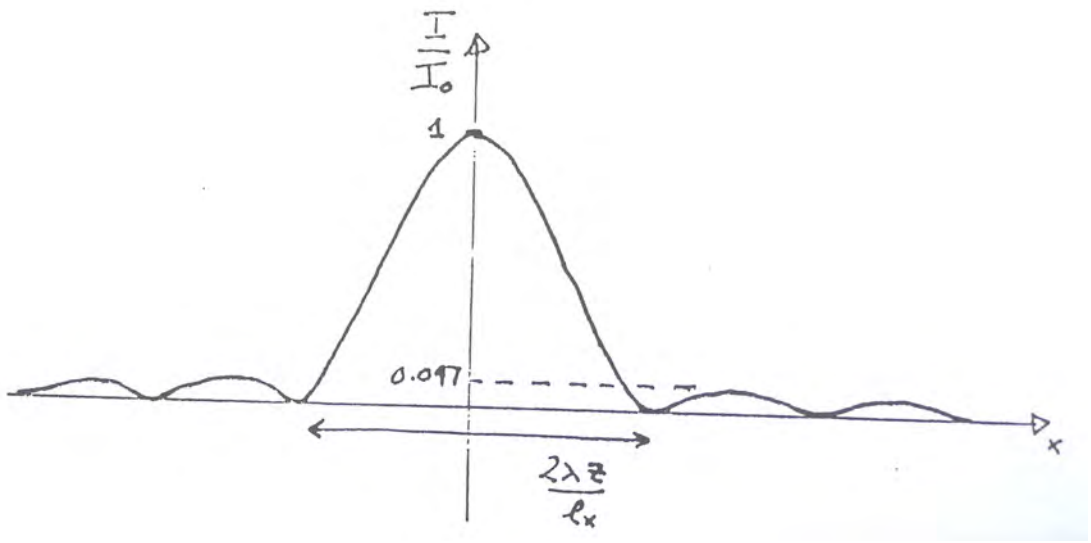


FIGURE 9.1.1. Fraunhofer Diffraction pattern from a rectangular aperture.

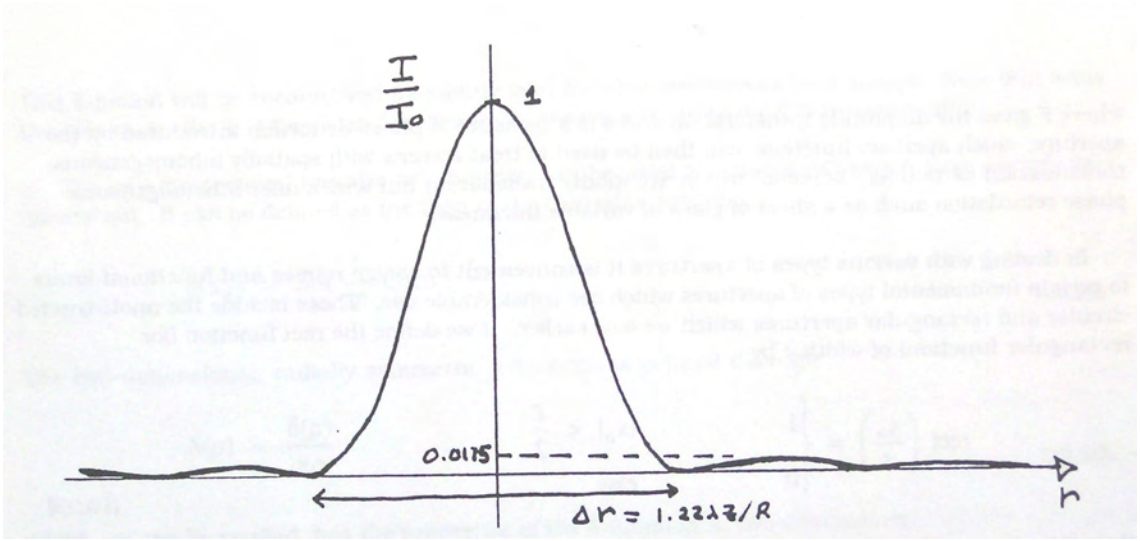


FIGURE 9.1.2. Normalized far-field diffraction from a circular aperture.

We next consider the far field diffraction pattern from a circular aperture of radius R illuminated by a wave of constant amplitude. The evaluation of the integral in this case with

$$\mathcal{E}(x_0, y_0) = \begin{cases} \mathcal{E}_A & \text{for} \quad x_0^2 + y_0^2 < R^2 \\ 0 & \text{else} \end{cases}$$

gives

$$\mathcal{E}_P(x, y) = \mathcal{E}(r, \theta) = -i\mathcal{E}_A e \frac{\exp\left(\frac{ik}{2z_a}\left[x_m^2 + y_m^2\right]\right) kR^2}{z} \frac{J_1(\frac{kRr}{z})}{\frac{kRr}{z}}$$

where $J_1(...)$ is the first order Bessel function and the last result indicates that the diffraction pattern only depends on the co-ordinate θ as it should for a situation with azimuthal symmetry. The intensity (to within the usual constant) is given by

$$I(\theta) = |\mathcal{E}(\theta)|^2 .$$

Figure 9.1.2 shows a plot of the normalized intensity as a function of the radial distance r.

The graph is similar to the previous graph for the diffraction intensity from the rectangular aperture. The diffraction pattern here, however, possesses circular symmetry. Also, the secondary maxima have significantly less intensity than the principal peak. The width of the diffraction pattern for the circular aperture (defined as the distance between the first set of nodes) is given by

$$\triangle r = \frac{1.22\lambda_0 z}{R} \quad \text{or} \quad \triangle\theta = \frac{1.22\lambda_0}{R}$$

which is slightly less than that given by a square with side R. The central bright spot in the diffraction pattern is known as the *Airy disk* and contains 0.86 of the total energy in the diffraction pattern.

9.2. Aperture Functions Made Simple

In the previous section we learned that the far-field (amplitude) diffraction pattern from an aperture is the Fourier transform of the field inside the aperture. In all our diffraction calculations so far we have only considered simple apertures involving slits and square or round hole. Since in most diffraction calculations we are not really concerned with calculating fields or intensities, but rather patterns for such. Hence it is often more convenient to speak of an aperture function, which is a dimensionless, complex number representing the field *pattern* across the aperture. Previously, when dealing with rectangular or circular apertures we assumed that the transmission function was either zero or unity. In general, the aperture, or transmission function, takes into account partial opacity or even phase distortion effects by allowing its amplitude to assume values other than zero or unity. In general then, one can define an *aperture function* by

$$A(x_0, y_0) = F(x_0, y_0)e^{i\phi(x_0,y_0)}$$

where F gives the field amplitude transmission and ϕ is a measure of phase distortion introduced by the aperture or present on the beam. Such aperture functions can then be used to treat screens with spatially inhomogeneous transmission as well as "screens" which are totally transparent, but which offer inhomogeneous phase retardation such as a sheet of glass of variable thickness. One can even break down a complex aperture (e.g. with an unusual shape or consisting of many holes) into a superposition of discrete, *non-overlapping* apertures as

$$A(x_0, y_0) = \sum_{n=1}^{N} A_n(x_0 - x_n, y_0 - y_n)$$

where the A_n defines the local complex transmission function of the *nth* component including phase effects, and which is centered on (x_n,y_n) .

Before unleashing the full power of Fourier techniques on complicated apertures it is worthwhile developing an extended "aperture basis set". In dealing with various types of apertures it is convenient to assign names and functional forms to certain fundamental types of apertures which see considerable use. These include the unobstructed circular and rectangular apertures which we saw earlier. If we define the rect function (for rectangular function) of width ℓ by

$$\text{rect}\Big(\frac{x_0}{\ell}\Big) = \begin{cases} 1 & |x_0| \le \ell/2 \\ 0 & \text{else} \end{cases}$$

then the aperture function for a ℓ_x by ℓ_y rectangular aperture, centered on the origin of the x_0,y_0 plane, can be defined as

$$A_{rect}(x_0, y_0) = rect\left(\frac{x_0}{\ell_x}\right) rect\left(\frac{y_0}{\ell_y}\right).$$

Similarly we can define the aperture function for the unobstructed circular disk of radius R to be

$$\text{cyl}\left(\frac{r}{\text{R}}\right) = \begin{cases} 1 & r \le R \\ 0 & \text{else} \end{cases} .$$

Other convenient functions to use in dealing with apertures are the following:

1) The *step function* is defined by

$$\text{step}\left(\frac{\text{x}}{\ell}\right) = \begin{cases} 0 & \frac{x}{\ell} < 0 \\ 1 & \frac{\text{x}}{\ell} \ge 0 \end{cases} .$$

This is a generalized Heaviside function and the only purpose of the ℓ is to allow the function to be reflected about the origin. The function can be used in dealing with diffraction about a straightedge.

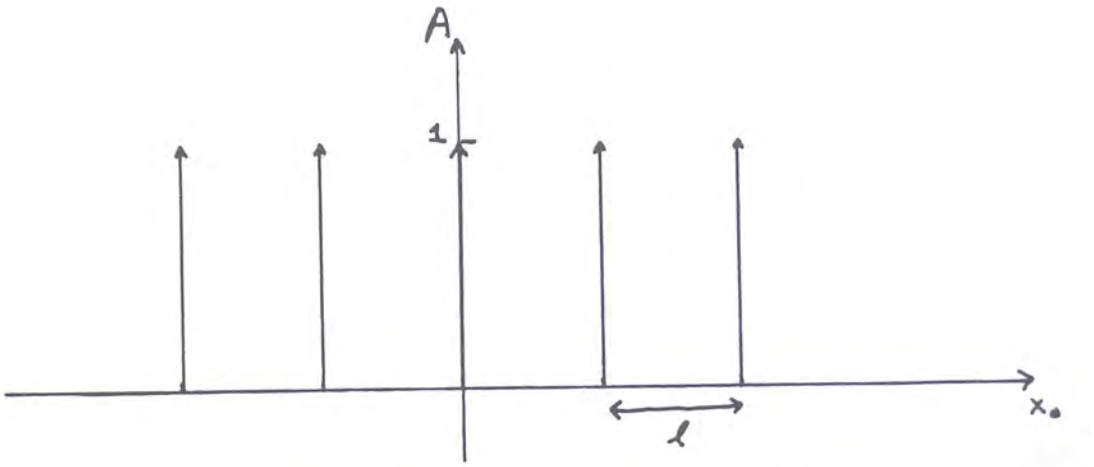


FIGURE 9.2.1. Representation of the aperture function of equation 9.2.1

2) The *Gaussian function* is defined by

$$\text{Gaus}(\frac{x}{\ell}) = \exp(-\pi(\frac{x}{\ell})^2)$$

and is normalized to unity amplitude at the origin. A similar function can be defined in polar co-ordinates as

$$\text{Gaus}(\frac{\varrho}{R}) = \exp(-\pi(\frac{\varrho}{R})^2).$$

This function is encountered frequently in the context of laser beams. Note that while the "area" under the 1-D Gaussian function is $|\ell|$ the area under the 2-D version is R^2.

3) The *one-dimensional impulse* or δ-function can be used to discuss diffraction from an infinitesimally narrow slit. It can be defined as the limit of the Gaussian function,

$$\delta(x) = \lim_{\ell \to 0} \frac{1}{|\ell|} \text{Gaus}(\frac{x}{\ell}).$$

The two-dimensional, radially symmetric δ-function is defined through

$$\delta(\varrho) = \frac{\delta(\varrho)}{\pi\varrho}$$

which, as can be verified, has the properties of the δ-function in two dimensions.

4) The one-dimensional *comb function* is defined as an infinite array of equally spaced δ-functions,

$$\text{comb}(\frac{x}{\ell}) = |\ell| \sum_{n=-\infty}^{n=\infty} \delta(x - n\ell)$$

and is useful in defining arrays of apertures of various types.

These functions can be used as building blocks to define more complicated arrays of apertures. For example the aperture function

$$A(x_0, y_0) = \text{rect}(\frac{x_0}{4\ell})\text{comb}(\frac{x_0}{\ell}) \tag{9.2.1}$$

defines an array of five slits with infinite extent along the y_o axis. These slits have infinitesimal width, are separated by ℓ and are located between $-2\ell < x_0 < 2\ell$ as illustrated in Figure 9.2.1.

In defining complicated apertures it is often convenient to make use of the convolution operation. The convolution between two one-dimensional functions *f(x)* and *g(x)* is defined by

$$f(x) \otimes g(x) = \int_{-\infty}^{\infty} f(\alpha)g(x - \alpha)d\alpha.$$

The convolution operation was used in the previous chapter to discuss the general diffraction problem. The integral is a function of the independent variable x. A simple geometrical interpretation of the convolution operation can be arrived at for real functions. The convolution can be interpreted as the area under the product of the two functions for different amounts of overlap. For example, in the case of a rectangular function convolved with itself,

it is easy to see that the convolution operation yields a triangular shaped function. The convolution operation has the following properties

1) It is *commutative*:

$$f(x) \otimes g(x) = g(x) \otimes h(x).$$

2) It is distributive:

$$[av(x) + bu(x)] \otimes h(x) = a\,[v(x) \otimes h(x)] + b\,[u(x) \otimes h(x)]$$

for scalars a and b.

3) It has *shift invariance*:

$$g(x - \beta) \otimes h(x) = g(x) \otimes h(x - \beta)$$

4) It is *associative*:

$$[v(x) \otimes h(x)] \otimes g(x) = v(x) \otimes [h(x) \otimes g(x)]$$

It is interesting to note that one of the properties which is used to define the δ-function may be interpreted in terms of a convolution. Indeed because the δ-function satisfies the following relation for any function *f(x)*,

$$\int_{-\infty}^{\infty} f(\alpha)\delta(x - \alpha)d\alpha = f(x)$$

it satisfies the following convolution relation

$$f(x) \otimes \delta(x) = f(x)$$

This is an important result and states that the convolution of any function with the δ-function merely reproduces that function; the δ-function is the identity function for convolution.

It is also possible to define the convolution for functions of two variables. If *f(x,y)* and *g(x,y)* are two arbitrary complex functions, the convolution of these two functions is defined to be

$$f \otimes g = \int\int_{-\infty}^{\infty} f(\alpha, \beta)g(x - \alpha, y - \beta)d\alpha d\beta$$

The extension of the definition of the convolution function to two dimensions can be seen to satisfy all the properties of the one-dimensional functions listed above.

With the aid of the convolution function it becomes easy to write the aperture function of an array with elements of arbitrary profile. For example the function

$$A(x_0, y_0) = e^{iksin\theta_i x_0}\mathrm{rect}(\frac{x_0}{a}) \otimes \mathrm{comb}(\frac{x_0}{\Lambda})$$

represents the aperture function of an infinite array of parallel, infinitely long slits of width a separated by distance Λ when a plane wave strikes them at an angle of incidence θ_i. This, in essence constitutes a transmission diffraction grating of infinite extent. If we wish to restrict its width (in the x_o direction) to *N+1* apertures while restricting the length of the slits to b we would have the following transmission function

$$A(x_0, y_0) = e^{iksin\theta_i x_0}\mathrm{rect}(\frac{x_0}{a})\mathrm{rect}(\frac{y_0}{b}) \otimes \left[\mathrm{rect}\left(\frac{\mathrm{x_0}}{\mathrm{N}\Lambda}\right)\mathrm{comb}(\frac{x_0}{\Lambda})\right] \tag{9.2.2}$$

This transmission function is shown in Figure 9.2.2.

In a similar manner screens of arrayed rectangular apertures as well as more exotic transmission gratings and even reflection gratings can be constructed.

For reflection gratings, or even reflection effects in general, all that has to be done is to design an appropriate aperture function which takes into account the local phase delay and amplitudes for light reflecting from a contoured surface. For example, consider a perfectly reflecting infinite surface which has a height variation $h(x_0)$. For a beam incident at an angle of incidence θ_i on the surface we have, with the use of section 4.4 , the aperture function

$$A(x_0) = \mathrm{rect}(\frac{\mathrm{x_0}}{\mathrm{w}})\exp(ik\sin\theta_i x_0)\exp[2ik\cos\theta_i h(x_0)] \tag{9.2.3}$$

where w is the width of the object and the factor of 2 accounts for the fact we are dealing with relection in determining phase delays. At a sinusoidal grating for example, $h(x_o) = h_o\sin(2\pi x_o/\Lambda)$ with Λ being the periodicity. The aperture function is, of course, different for a partially transmitting surface or one which has a non-constant amplitude reflectivity.

In this section we have used some simple definitions and the convolution operation to construct complex apertures to describe transmission and diffraction at a screen. In what follows we describe the appropriate mathematical tools to compute the far field diffraction patterns.

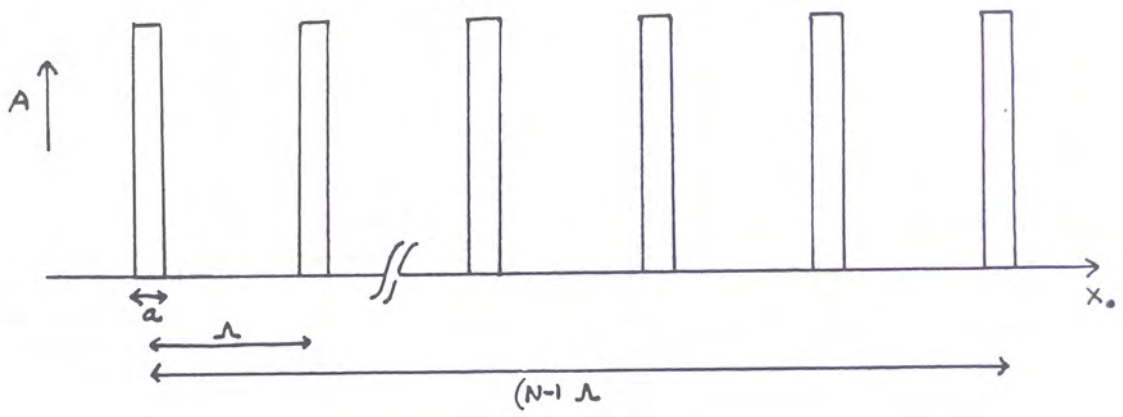


FIGURE 9.2.2. Aperture function of a finite transmission grating as a function of x_0.

9.3. Properties of Fourier Transforms

Before proceeding to diffraction calculations we review the properties of the Fourier transform which we shall find useful in analyzing diffraction problems. To begin, let us recall from the Appendix in Ch.1 that for an arbitrary, complex-valued function *f(x)* we can form the integral

$$F(u) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)e^{-iux}dx$$

where the sign difference in the exponential occurs, according to our convention, because it is associated with a spatial rather than a temporal phase effect. If this integral exists for all values of u, *F(u)* is referred to as the Fourier transform of the function *f(x)*. The definition can be extended to deal with two dimensional functions *f(x,y)* in which case the Fourier transform is defined through

$$F(u,v) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-i(ux+vy)}dxdy$$

and the function is usually designated as $F(u,v) = \mathcal{F}[f(x,y)]$. The Fourier transform function so defined is itself a complex-valued function of two spatial frequencies, u and v. The inverse Fourier transform of the function *F(u,v)* is of course $f(x,y) = \mathcal{F}[F(u,v)]$ so that

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u,v)e^{i(ux+vy)}dudv$$

Were it not for the factor of $(2\pi)^{-2}$ in the definition of the Fourier transform, the expressions for the Fourier transform and its inverse would be symmetrical. In certain conventions the defining equations for the Fourier transform and its inverse are symmetrized with each integral (for the two dimensional case) having the pre-factor of $(2\pi)^{-1}$. This is a matter of taste and convenience.

It is easy to show that the Fourier transform operation satisfies the following properties:

1) *Linearity*:

$$\mathcal{F}\{\alpha g + \beta h\} = \alpha\mathcal{F}(g) + \beta\mathcal{F}(h)$$

where α and β are complex scalars and g and h are functions of the variables *x,y*.

2) *Similarity*:

$$\text{If } \mathcal{F}\{g(x,y)\} = G(u,v) \text{ then}$$
$$\mathcal{F}\{g(ax,by)\} = \frac{1}{|ab|}G(\frac{u}{a},\frac{v}{b})$$

indicating that a stretching of the co-ordinates in the spatial domain results in a contraction of the co-ordinates in the frequency domain plus an associated change in the amplitude of the spectrum.

3) *Shift property*:

$$\text{If } \mathcal{F}\{g(x,y)\} = G(u,v) \text{ then}$$
$$\mathcal{F}\{g(x-a,y-b)\} = G(u,v)e^{-i(ua+vb)} \tag{9.3.1}$$

TABLE 1. Fourier Transform Pairs

Function	Fourier Transform$\times 2\pi$
rect(x)	$\mathrm{sinc}(\frac{u}{2\pi})$
$\delta(x)$	1
comb(x)	$\mathrm{comb}(\frac{u}{2\pi})$
Gaus(x)	$\mathrm{Gaus}(\frac{u}{2\pi})$
step(x)	$\frac{1}{2}\delta(\frac{u}{2\pi}) + \frac{1}{ui}$
cyl(r)	$J_1(u)/u$

indicating that a translation of the function in the spatial domain introduces a linear phase shift in the frequency domain. However, note that

$$\mathcal{F}\{G(u-u_0, v-v_o)\} = g(x,y)e^{i(u_0x+v_0y)}.$$

4) *Parseval's Theorem*:

$$\text{If } \mathcal{F}[g(x,y)] = G(u,v) \text{ then}$$
$$\int\int_{-\infty}^{\infty} |g(x,y)|^2\, dxdy = \int\int_{-\infty}^{\infty} |G(u,v)|^2\, dudv$$

This theorem is essentially a statement of conservation of energy in real space and Fourier space.

5) *Convolution theorem*:

$$\text{If } \mathcal{F}\{g(x,y)\} = G(u,v) \text{ and } \mathcal{F}\{h(x,y)\} = H(u,v) \text{ then}$$
$$\mathcal{F}\{g \otimes h\} = G(u,v)H(u,v)$$

That is, the convolution of two functions in real space is equivalent to a multiplication of their Fourier transforms in the frequency domain, and vice-versa.

At this point it is useful to list some of the commonly encountered functions dealt with in diffraction problems and their Fourier Transforms.

For the last case, cyl(r) is defined on the plane; both cyl(r) and its Fourier transform have radial dependence only.

9.4. Fourier Optics and Gratings

To illustrate the power of Fourier transform techniques in far field diffraction we consider the problem of diffraction from the transmission grating defined by equation 9.2.2. To simplify the problem we allow $b \to \infty$ so that we can ignore the y variable in the problem. From the above discussion we know that the diffraction pattern from the transmission grating is defined by

$$T = \mathcal{F}\left\{e^{iksin\theta_i x_0}\mathrm{rect}(\frac{x_0}{a}) \otimes \left[\mathrm{rect}\left(\frac{\mathrm{x_0}}{\mathrm{N}\Lambda}\right)\mathrm{comb}(\frac{x_0}{\Lambda})\right]\right\} \tag{9.4.1}$$

where we have, as usual, ignored the global phase factors that precede the diffraction integral. From the Fourier transform properties and the convolution theorem we have

$$T = \frac{1}{2\pi}\,|a|\, sinc(\frac{a(u-k\sin\theta_i)}{2\pi})\mathcal{F}\left\{\Lambda \sum_{n=-(N-1)/2}^{n=(N-1)/2} \delta(x-n\Lambda)\right\}$$
$$= \frac{1}{2\pi}\,|a|\, sinc(\frac{a(u-k\sin\theta_i)}{2\pi})\left\{\sum_{n=-(N-1)/2}^{n=(N-1)/2} e^{-inu\Lambda}\right\}$$

where we have implicitly assumed that N is odd. Summing the geometric series gives

$$T(\theta) = \frac{1}{2\pi}\,|a|\, sinc\left\{\frac{[\sin\theta - \sin\theta_i]\,a}{\lambda_0}\right\}\frac{\sin\left(\frac{\pi N\Lambda}{\lambda_0}\sin\theta\right)}{\sin\left(\frac{\pi\Lambda}{\lambda_0}\sin\theta\right)}$$

where θ is the direction of observation relative to the z-axis ($\sin\theta = x/z$). It can be seen that the diffraction pattern is that of a single aperture multiplied by an array factor that was arrived at by summing over the N elements of the grating. Notice that the array factor is periodic in $\sin\theta$ with a periodicity of λ_0/Λ. The intensity pattern,

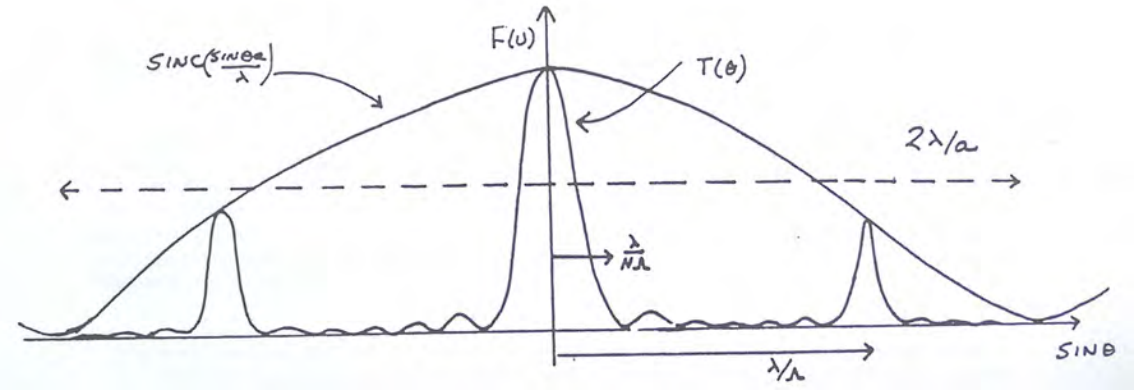


FIGURE 9.4.1. Diffraction pattern from a transmission grating.

depicted in Fig. 9.4.1, illustrates that the overall diffraction pattern can be understood as having a series of narrow peaks associated with the periodicity of the grating structure, and these peaks are modulated in amplitude by the diffraction pattern of a single elementary aperture.

In principle, the grating as a spectroscopic instrument can be used to analyze light to any desired degree of accuracy. For an infinite grating with two different plane waves at different wavelengths incident, the angular separation between the propagation of the reflected waves in the same order is slightly different. The resolution of the instrument in this case is limited only by one's ability to measure the angular separation between two beams. In practice, however, one doesn't have an infinite grating. For a finite width grating, even if a plane wave is incident, a plane wave cannot emerge as a reflected wave, since the grating has a finite transverse extent. As we have seen, the emerging wave can be Fourier analyzed as a superposition of infinite plane waves, but the reality is that the emerging beam is not well defined in terms of its direction of propagation and so, even though there is only has one wavelength incident, the beam has an angular spread. This angular spread, in the case of two plane waves with different wavelengths incident on the grating, can prevent one from resolving two closely spaced wavelengths.

The resolving power of the transmission grating can be evaluated as follows. Define $\Upsilon = \frac{\pi N\Lambda}{\lambda_0}\sin\theta$. For normal incidence light, near $x = 0$ or $\theta = 0$, the width of the peak is defined by

$$\delta\Upsilon = \frac{\pi N\Lambda}{\lambda_0}\cos\theta\delta\theta = \pi.$$

This gives

$$\delta\theta = \frac{\lambda_0}{N\Lambda\cos\theta}.$$

However the principal maxima in the diffraction pattern occur for

$$\frac{\pi\Lambda}{\lambda_0}\sin\theta = m\pi \text{ for } m = 0, \pm 1, \pm 2, ...$$

where m is referred to as the order of the diffraction pattern. Taking the derivative of this equation with respect to λ, it follows that for a given order,

$$\Lambda\cos\theta\delta\theta = \mathrm{m}\delta\lambda_0.$$

Combining the last two equations we see that the resolving power of the grating is given by

$$\frac{\lambda_0}{\delta\lambda} = \frac{\omega}{\delta\omega} = mN. \tag{9.4.2}$$

For a typical grating with total number of (illuminated) elements given by $N = 5000/cm\times 5cm$ one obtains for the first order a resolving power of 2.5×10^4 or a wavelength resolution of ≈ 0.04 nm for incident wavelengths in the vicinity of 1 μm. The resolving power quoted is considered very good. Indeed many gratings even fail to reach these values. In some cases it may be difficult to construct high quality gratings with a 5 cm width. Second, the grating may not be ruled well enough to justify the assumption of perfect periodicity. In this case the instrument function has a larger angular width and the resolving power of the grating is compromised.

In a spectroscopy experiment, where one attempts to obtain the relative intensities of light at different wavelengths, it is necessary to work in only one diffraction order since the efficiency is the same for all wavelengths. Unfortunately, from the point of view of optimizing the signal to noise ratio, this means that the light in other orders is wasted. The transmission grating, in general, has most of its output in the zeroth order where according to

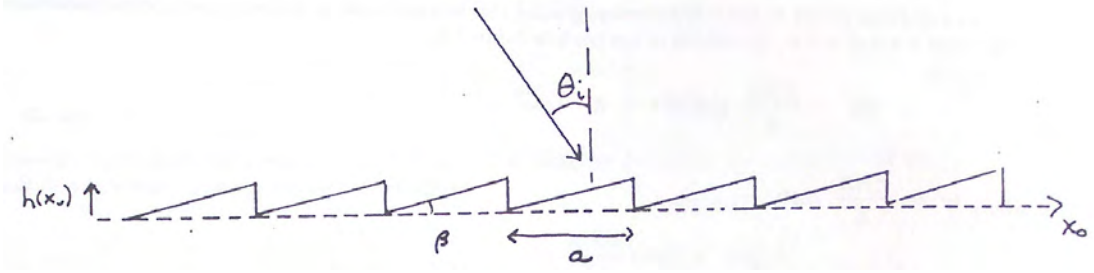


FIGURE 9.4.2. A sawtooth reflection grating ($a = \Lambda$).

equation 9.4.2 there is no resolving power. One is much better off working with reflection gratings where there are geometrical factors at one's disposal which can be used to optimize the efficiency of scattered light into a particular order.

Reflection gratings can be produced with a variety of different profiles. The sinusoidal grating of chapter 4 was useful for illustrating elementary diffraction effects, but as the coefficients of the diffracted wave amplitudes show, it probably isn't the most efficient for producing light in non-zero orders. The most efficient gratings for this purpose have sawtooth profiles. For example, consider the sawtooth grating of periodicity Λ depicted in Figure 9.4.2.

The angle of incline of the sawtooth, β, can be chosen to optimize the efficiency of light which is scattered into the -1 order. From a qualitative point of view, if the light is incident on the grating at the particular angle of incidence indicated, *i.e.* parallel to the steep part of the sawtooth, then light components that are reflected from the less-inclined parts of the sawtooth are in phase with each other if the length of the steep edge is equal to 2λ, where λ is the wavelength of the incident light. This light corresponds to light which is reflected back in the -1 order. Indeed, for this grating, since the specular components derived from the surface elements interfere with each other, nearly all the light can be directed into a particular order. Of course, the efficiency is optimum only for a certain region of wavelength, but in many applications the restricted range of wavelength is gladly traded off for the enhanced efficiency. There is a correspondence between the angle, β, and the wavelength of maximum efficiency. The angle β is referred to as the *blaze angle*, and gratings which have their efficiency optimized for a certain wavelength region are referred to as *blazed gratings*. Such gratings find considerable use in spectroscopy and also as wavelength tunable mirrors as we will see in the discussion on lasers in chapter 13.

To analyze the properties of a blazed reflection grating let us consider the sawtooth grating of Figure 9.4.2 with $a = \Lambda$, for which we can use the aperture function of equation 9.2.3 with a periodic height function.

$$h(x_0) = \begin{cases} x_0 \tan\beta & 0 \le x_0 \le \Lambda \\ 0 & \text{else} \end{cases}$$

For a grating with N periods we can rewrite the expression in equation 9.2.3 as

$$A(x_0) = \text{rect}(\frac{x_0}{\Lambda}) \exp(ik \sin\theta_i x_0) \exp[2ik \cos\theta_i h(x_0)] \otimes \left[\text{rect}\left(\frac{x_0}{N\Lambda}\right) \text{comb}(\frac{x_0}{\Lambda})\right]$$

or

$$A(x_0) = \left\{\text{rect}(\frac{x_0}{\Lambda}) \exp[ix_0 k(\sin\theta_i + 2\tan\beta cos\theta_i)]\right\} \otimes \left(\Lambda \sum_{n=-(N-1)/2}^{n=(N-1)/2} \delta(x - n\Lambda)\right)$$

With the use of the Fourier transform, and recalling the results for the transmission grating, the far field diffraction pattern is therefore of the form

$$R(u) = \frac{1}{2\pi}\Lambda sinc\left\{\frac{[\sin\theta - \sin\theta_i - 2\tan\beta\cos\theta_i]\,\Lambda}{\lambda_0}\right\} \frac{\sin\left(\frac{\pi N\Lambda}{\lambda_0}\sin\theta\right)}{\sin\left(\frac{\pi\Lambda}{\lambda_0}\sin\theta\right)}.$$

It can readily be seen that the proper choice of β allows the diffraction pattern to peak on a nonzero order of the array function. Indeed if the reflection grating is designed so that the peak of the sinc function coincides with a non-zero order of the array function one would have a situation where most of the light is diffracted into that order so that one can have peak efficiency and high dispersion at the same time. Indeed, as suggested earlier, this is why most spectroscopic gratings are blazed. For example, the condition that the peak of the sinc function coincides with the $m = -1$ order of the array function is found to be

$$(\sin\theta_i + 2\tan\beta\cos\theta_i)\Lambda = \lambda_0$$

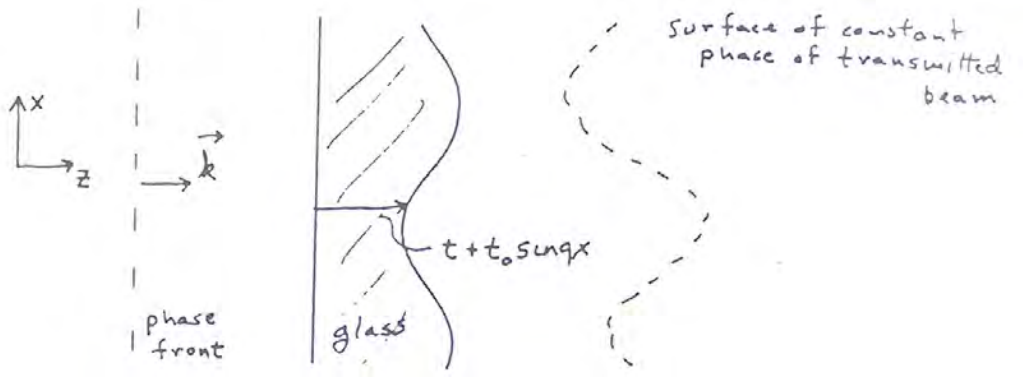


FIGURE 9.4.3. Plane wave incident on a transparent dielectric

The fact that the equation depends on wavelength indicates that, for a given blaze angle and angle of incidence, the peak efficiency only occurs near a particular wavelength, known as the *blaze wavelength* of the grating. It is also clear that the choice of β is wavelength dependent.

The grating efficiency also depends on the polarization of light. In our earlier analysis of the sinusoidal grating in chapter 4 we specifically assumed TE polarized light and used the boundary conditions associated with it. In the case of TM polarized light the analysis would have been different, and in general the reflection efficiency of TE and TM light into different orders is different.

Note that the resolving power is the same for reflection and transmission gratings. This is no coincidence since the Fourier transform of a reflecting or a transmitting screen are identical if they have the same real or complex spatial variations imposed on them. The diffraction from a reflecting object has made it clear that one doesn't require a "hole" in a screen to obtain diffraction effects. Indeed as indicated at the beginning of the previous chapter, diffraction is associated with any deviation from the propagation of a perfect plane wave (with infinite transverse extent). If there is truncation of the transverse extent of a beam, or if the wave acquires a phase factor which depends on transverse co-ordinates, diffraction effects result. It should also be clear by now that if one knows the amplitude and phase distribution of a beam (in essence the aperture function) in any plane, one can compute the phase and amplitude distribution in any other plane within the Fresnel or Fraunhofer approximation. In this sense, any plane that a beam propagates through may be viewed as an aperture plane and diffraction from this plane can be computed.

The fact that a purely phase object can cause diffraction is evident if we consider a plane wave normally incident on a transparent slab of infinite transverse extent as shown in Figure 9.4.3.

If the slab has a flat incident side but a rough exit side so that the thickness of the slab is given by $t + t_o sinqx$, (with $t_o << \lambda$) diffraction results. This can be seen if we take the incident plane wave to be of the form e^{ikz}.

To within a constant phase factor the exit beam in a plane behind the plate will have the form

$$e^{ik(t-t_0 \sin qx)} e^{it_0 knsinqx}$$

where n is the refractive index of the slab. The 1-D aperture function of the beam in a plane just beyond the exit face is therefore given by

$$A(x_0) = e^{ik(n-1)t_0 \sin qx_0} = 1 + ik(n-1)t_0 \sin qx_0 = 1 + \frac{1}{2}k(n-1)t_0 \left[e^{iqx_0} - e^{-iqx_0}\right].$$

The far-field diffraction pattern for this function (which occurs at infinity !) is given by its Fourier transform and is easily shown to be

$$F(u,v) = \frac{1}{(2\pi)^2}\delta(v)\left\{\delta(u) + \frac{k(n-1)t_0}{2}\left[\delta(u+q) - \delta(u-q)\right]\right\}.$$

That is, the diffraction pattern consists of three distinct spots, which reduce to a single spot associated with a plane wave if the slab faces are perfectly parallel.

9.5. Transmission Characteristics of a Lens

Perhaps the most important and useful nonuniform transparent medium is a lens. Although all of us have known ever since our childhood days that a lens is something which can bring a beam to a focus or cause a beam to diverge, its properties are much more general. In chapter 5 we discussed the properties of a lens from a geometrical optics

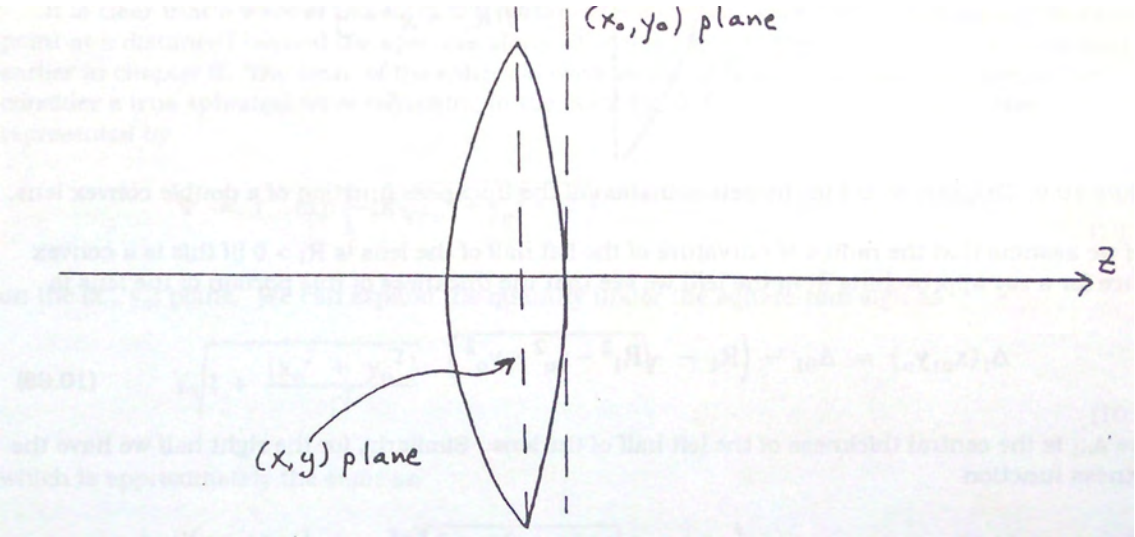


FIGURE 9.5.1. A double convex lens seen in cross section. The $x-y$ plane bisects the lens.

point of view. Here we consider a lens in the wave-optics framework where we can explicitly discuss diffraction effects. To understand the transformation properties of a lens let us consider a plane wave incident from the left on a double convex lens as depicted in Figure 9.5.1.

To determine the action of the lens on an incident plane wave we must know what the local phase delay is for a portion of the beam passing through the point *(x,y)*, on a plane bisecting the lens. This allows us to calculate the aperture function of the beam in a (x_0, y_0) plane beyond the lens. The local phase delay is determined by the thickness of the lens for co-ordinates *(x,y)*. Let us denote the local thickness by the function Δ*(x,y)* which for a double convex lens assumes its maximum value Δ_0 at the center where $(x, y) = (0, 0)$.

The net local phase delay in the beam can be determined by calculating the phase delay between parallel planes immediately before and immediately after the lens. In this case the local phase delay is given by

$$\phi(x, y) = kn\triangle(x, y) + k(\triangle_0 - \triangle(x, y))$$

where n is the refractive index of the lens. The lens may then be represented by the transmission or aperture function

$$A(x_0, y_0) = e^{i\phi(x_0,y_0)}$$

which is defined in a plane at the exit face of the lens where we define $z = 0$. To determine the thickness and hence the phase function we need a shape factor for the lens. We assume that the lens consists of two spherical caps joined together as shown in Figure 9.5.2.

If we use our convention that the radius of curvature of the left half of the lens is $R_1 > 0$ for a convex surface for a ray approaching from the left, we see that the thickness of this portion of the lens is

$$\triangle_1(x, y) = \triangle_{01} - (R_1 - \sqrt{R_1^2 - x^2 - y^2})$$

where Δ_{01} is the central thickness of the left half of the lens. Similarly, for the right half we have

$$\triangle_2(x, y) = \triangle_{02} - (R_2 - \sqrt{R_2^2 - x^2 - y^2})$$

so that the total thickness is

$$\triangle(x, y) = \triangle_1(x, y) + \triangle_2(x, y).$$

Note that for a biconvex lens, the thickness obviously goes to zero for large enough x_0, y_0. For now we ignore the region of the (x,y) plane in which part of the plane wave bypasses the lens. We also make use of the paraxial approximation in which we assume that

$$x^2 + y^2 \ll R_1^2, R_2^2$$

This is equivalent to the thin lens approximation, but it can also apply to a beam of small finite, transverse extent incident on the lens. In any event, within this approximation we have that

$$\sqrt{R_1^2 - x^2 - y^2} = R_1 \left(1 - \frac{\left[x^2 + y^2\right]}{2R_1^2}\right)$$

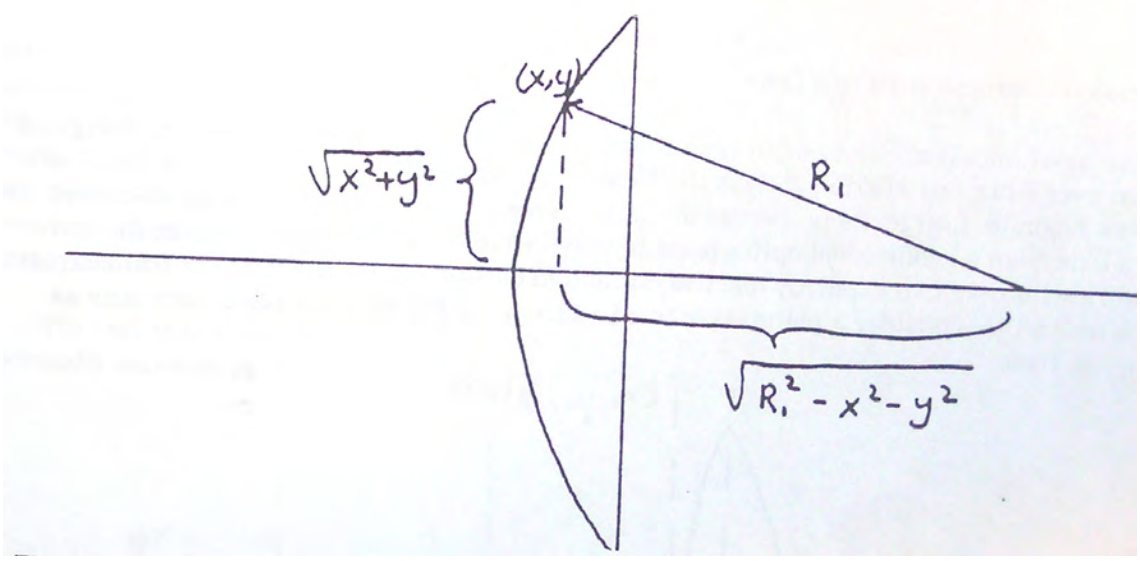


FIGURE 9.5.2. Diagram to aid in the determination of the thickness function of a double convex lens.

and a similar expression exists for $\sqrt{R_2^2 - x^2 - y^2}$. The thickness function then becomes

$$\triangle(x,y) = \triangle_0 - \frac{\left[x^2+y^2\right]}{2}\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

If we define a quantity f by the expression (it should look familiar!)

$$f^{-1} = (n-1)(R_1^{-1} - R_2^{-1})$$

then the aperture function becomes

$$A(x_0,y_0) = e^{ikn\triangle_0}\exp\left[-\frac{ik}{2f}(x_0^2+y_0^2)\right]. \tag{9.5.1}$$

It is clear that a wave of this form is a portion of a spherical wave which is collapsing to a focal point at a distance f beyond the aperture plane. The focal length here is identical to that derived earlier in chapter 5. The form of the spherical wave would be more immediately apparent if we consider a true spherical wave collapsing to the point $(0,0,f)$ from the left. Such a wave is represented by

$$\psi = r^{-1}\exp\left(-ik\sqrt{x_0^2+y_0^2+f^2}\right) \tag{9.5.2}$$

on the (x_0,y_0) plane. We can expand the quantity under the square-root sign as

$$f\sqrt{1+\frac{x_0^2+y_0^2}{f^2}} \simeq f\left(1+\frac{x_0^2+y_0^2}{2f^2}\right) \tag{9.5.3}$$

within the paraxial or thin lens approximation. Putting 9.5.3 into 9.5.2 and comparing with 9.5.1 it can be seen that, to within a constant phase factor, the two expressions for the spherical waves agree in the paraxial approximation. The action of the lens can be further understood with reference to Fig. 9.5.3.

An incident wave, whose surfaces of constant phase are planes perpendicular to the direction of propagation, is converted into a wave whose surfaces of constant phase are of spherical shape (or within the paraxial approximation, parabolic shape). Since for homogeneous waves, the direction of energy flow is normal to the local phase front the energy in the beam collapses on to a single point. Note as well that the curvature in the phase front is simply due to the fact that the portion of the plane wave which passes near the center of the lens is forced to pass through more material for which the phase velocity is low. This causes a phase retardation of the central portion of the beam relative to that in the wings.

The expressions derived above can be applied to negative, plano-convex, meniscus, etc. lenses by allowing R_1 and R_2 to vary in magnitude and sign. For example in the case of a double concave lens $f < 0$. In this case the emerging spherical wave is seen to be diverging from a point located on the axis of symmetry at a distance f in front

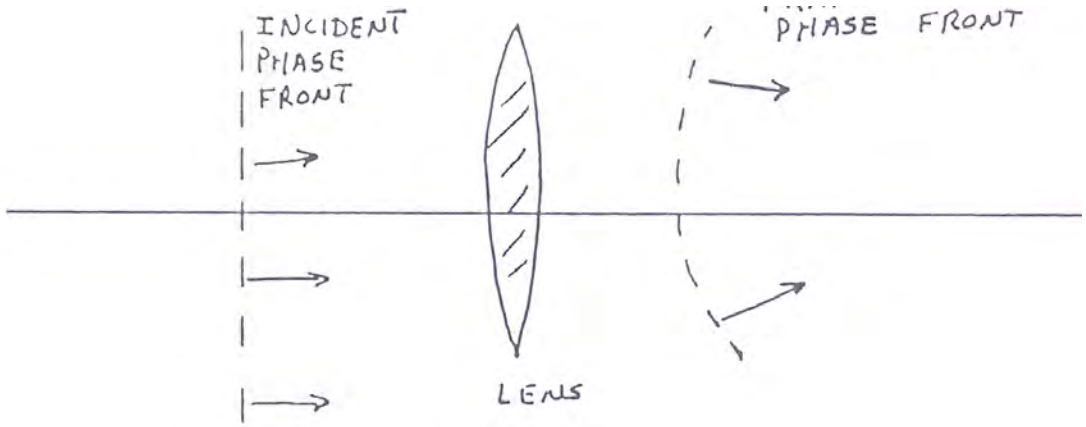


FIGURE 9.5.3. Demonstration of the action of a thin lens.

of the lens. The focal point in this case represents a virtual image of the emitted wave since no light actually emerges from this point.

9.6. The Lens as a Fourier Transforming System.

In the last section we were able to arrive at the basic features of the influence of a lens on an optical beam. However, along the way we made a number of assumptions and were even led to a somewhat unphysical result—a positive lens focuses a portion of an incident plane wave to a point. To perform a more realistic treatment of lenses we generalize the above treatment using a diffraction point of view. We concentrate our efforts on positive lenses to arrive at the salient result that a positive lens essentially performs a Fourier transform of the input beam and as such, regardless of its focal length, gives the far-field or Fraunhofer diffraction pattern of an incident beam.

Consider then a beam whose complex amplitude function immediately before a positive lens is given by $\mathcal{E}(x,y)$. If the beam is restricted in transverse extent by an aperture, by the lens, or even by itself we can account for this by introducing a pupil function $P(x,y)$ which is defined by

$$P(x,y) = \begin{cases} 1 & \text{for } (x,y) \text{ contained inside the beam} \\ 0 & \text{otherwise.} \end{cases}$$

Knowing the transmission function of the lens from the previous discussion we see that immediately after the lens the aperture function of that portion of the beam which passed through the lens is given by

$$A(x_0,y_0) = P(x_0,y_0)\mathcal{E}(x_0,y_0)\exp\left[-\frac{ik}{2f}(x_0^2+y_0^2)\right].$$

To obtain the optical field distribution in the focal plane of the lens we apply the Fresnel diffraction formula, since in some cases the focal length can be small and the Fresnel diffraction formula is certainly more general than the Fraunhofer formula. We return to the use of (x,y,z) as defining the point of observation. Doing so with $z = f$ we find that in the focal plane we have for the optical field distribution

$$\mathcal{E}_f(x,y) = -\frac{i}{\lambda f}\int\int_{-\infty}^{\infty} A(x_0,y_0)\exp\left\{\frac{ik}{2f}\left[(x-x_0)^2+(y-y_0)^2\right]\right\}dx_0dy_0$$

which upon substituting the expression for the aperture function becomes

$$U_f(x,y) = -\frac{i}{\lambda f}\exp\left[-\frac{ik}{2f}(x^2+y^2)\right]\int\int_{-\infty}^{\infty} P(x_0,y_0)t(x_0,y_0)\exp\left\{\frac{-ik}{f}\left[xx_0+yy_0\right]\right\}dx_0dy_0.$$

Apart from a global phase factor in front of the integral sign it is seen that the amplitude distribution in the focal plane is exactly the Fourier transform of the aperture function immediately in front of the lens! If the aperture function were evaluated in the front focal plane of the lens even the quadratic phase factor would disappear in the Fourier transform relation. However, this factor is of little consequence since only the intensity is ever measured.

Notice that in this more exact formulation of the lens problem an incident wave is never focused to a mathematical point. This can be understood if we consider a plane wave ($\mathcal{E}(x,y) = \mathcal{E}_0$) incident on a lens with a *pupil function*

$$P(r) = cyl(\frac{r}{R}).$$

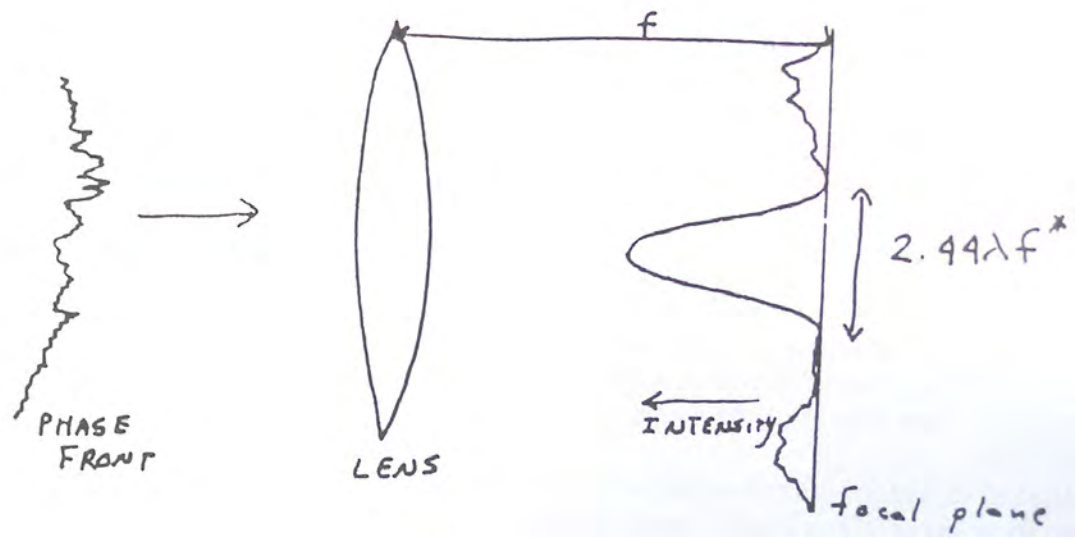


FIGURE 9.6.1. Focusing a beam with considerable amplitude variation.

Making use of the results of section 9.1 we know that the diffraction pattern in the focal plane consists of an Airy pattern and the diameter of the Airy disk is

$$D_A = 2\left(\frac{1.22\lambda f}{2R}\right) = 2.44\lambda f^*$$

where the quantity *f* (= f/2R)* is known as the *f-number* of the lens. Although, in principle, it is possible to make the f-number for the lens arbitrarily small, various constraints on the shape factor and refractive index of the lens usually restrict the f number to be greater than unity. It follows that the smallest spot size of a focused beam is of the order of λ. Note that the pupil function that determines the ultimate spot size is not necessarily associated with the diameter of the lens, but could also be the diameter of the beam incident on the lens. To obtain the smallest focal spot size from a given light source, one must maximize the transverse extent of the beam, usually to the maximum transverse extent of the lens. The focal spot size that results with a plane wave incident is referred to as the diffraction limited spot size and simply depends on the incident wavelength and the focal length and diameter of the lens. If the incident beam is not a perfect plane wave or if the lens has not been manufactured with the mathematical precision discussed here, one can show that a larger spot size occurs. A lens which is capable of producing the smallest spot size discussed above is said to yield the *diffraction limited spot size.*

What happens to the focal spot if other than a plane wave is incident on the lens? The mathematical answer to this question of course can be found by simply taking the Fourier transform of the aperture function associated with the beam immediately before the lens. Obviously if there is any spatial structure on the phase or amplitude of the beam as a function of the transverse co-ordinates this shows up in the focal plane of the lens. If there is high frequency spatial structure on the incident beam, more of the energy of the beam is located away from the focal spot than if a plane wave were incident. This is depicted schematically in Figure 9.6.1 which shows a wave with considerable "noise" in its intensity distribution. The Fourier transform of such a beam has considerable amplitude at high spatial frequencies and, as illustrated, much of the beam energy is located away from the diffraction limited spot. Note that, among other things, this implies that coherent sources always produces a smaller spot than incoherent sources. This is one of the reasons why lasers with their high degree of spatial coherence have seen tremendous use in microsurgery, micro-electronics and high density digital data recording on disks.

The ability of lenses to take the Fourier transform of the incident beam sees many applications. The potential to handle optical information either in real space or Fourier space leads to some powerful techniques in what has come to be known as optical data processing. One of the most practical examples of this is what is known as spatial filtering. Consider a well-defined beam which has small spatial frequency structures imposed on it, say those associated with an "image" of something we are interested in. During propagation from the source of the image, the light beam may be partially scattered at certain points in its aperture due to dust particles on optical surfaces, in the air, etc. This usually leads to high frequency "noise" structure on the beam which, if left there, leads to a degradation of the image. This noise is easy to remove however, by focusing the light with a lens as illustrated in Figure 9.6.2 and by placing an aperture in the focal plane to block all the high frequency components.

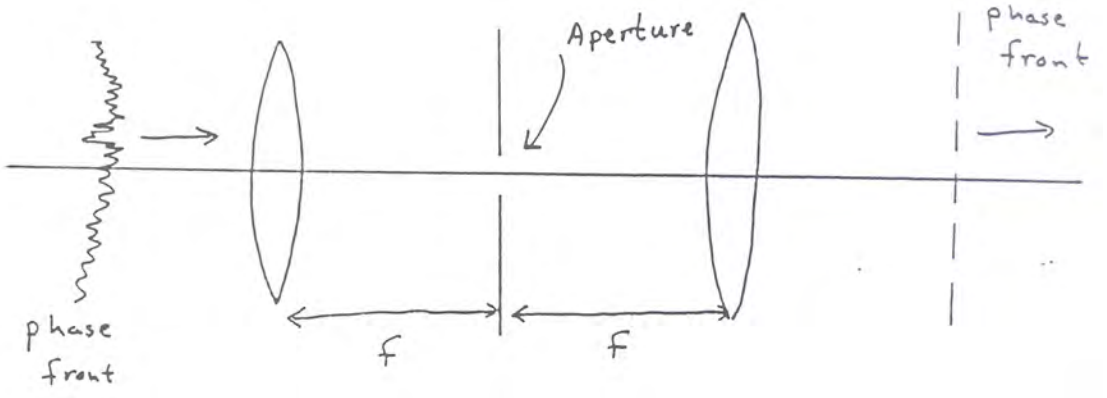


FIGURE 9.6.2. An illustration of spatial filtering.

The emerging beam is recollimated with a following lens which, in essence, takes the inverse Fourier transform without the high frequency structure, and hence "cleans up" the image. This technique is used to remove the lines from multiple, overlapping satellite photographs as well as to “clean” a beam before it is focused to a diffraction limited spot size.

9.7. Holography

Holography is popularly recognized as a technique which allows one to store three dimensional images on a two-dimensional recording medium. From a more technical point of view it allows one to record both the amplitude and phase information of a wave disturbance, in essence the full information of the wave. Note that, as such, holography can be associated with waves of any kind–acoustic, optical, etc.

The field received its start in 1948 when Dennis *Gabor* was engaged in research activities whereby he was trying to improve upon the resolving power of microscopes. He recognized (as we have just seen above) the advantages of using a coherent beam with well defined phase characteristics in improving the resolving power. In the course of his research with coherent waves he developed a technique for storing both the phase and amplitude information of an optical wave on an optical recording film although it is known that photographic recording media basically only store intensity information. In his technique the additional information represented by the phase is stored by interfering the optical disturbance to be recorded with a reference wave. In order to understand the basic principle behind holography consider Figure 10.2.1 which shows a coherent wave incident on an object.

Part of the incident beam is diffracted from the object and propagate towards a recording plate where it interferes with the undiffracted wave. If we refer to the film plane as the (x, y) plane, in this plane we can describe the disturbance emanating from the object as the "object wave", which will have the general functional form

$$O(x, y) = o(x, y)e^{i\phi(x,y)}$$

where $o(x, y)$ is the real amplitude of the object wave and $\phi(x, y)$ represents its phase. The undiffracted or "reference wave" in the same plane similarly has the functional form

$$R(x, y) = r(x, y)e^{i\psi(x,y)}.$$

It follows that the intensity on the plate is

$$I(x, y) = |O + R|^2 = |O|^2 + |R|^2 + O^*R + OR^* = |O|^2 + |R|^2 + 2orcos(\psi - \phi) \tag{9.7.1}$$

in which it is clear that not only the intensity but the relative phase of the two waves has been recorded. If the reference wave is a plane wave, we can without loss of generality choose $\psi = 0$, so that only the phase of the object wave is being recorded. The permanent recording of the intensity function is known as a hologram.

When the plate is developed it has a transmittance which depends on the intensity distribution recorded on the plate. We assume that the recording and developing processes are linear with respect to incident intensity and the emulsion has a sufficiently small grain structure so that it can faithfully reproduce all the high frequency spatial information of the beam (usually on the scale of the wavelength of light). Such assumptions are usually found not to be restrictive with present film technology. Assuming that the reference wave has uniform intensity over the film, the amplitude transmittance of the plate has the form

$$t_p = t_b + B\left(|O|^2 + O^*R + OR^*\right)$$

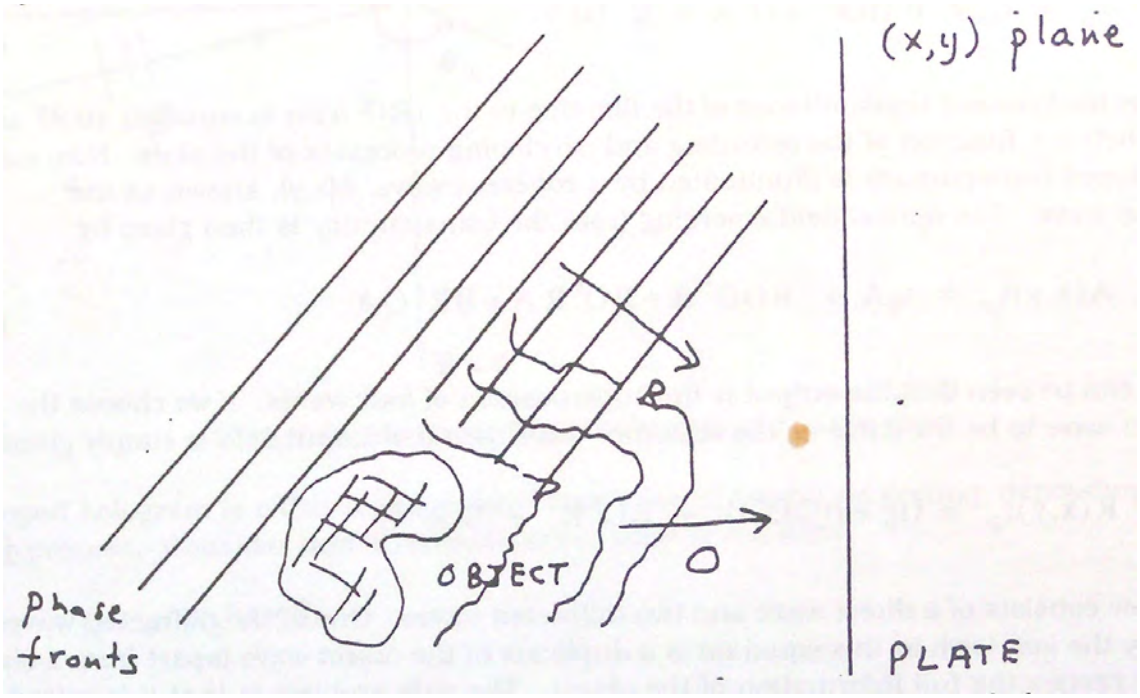


FIGURE 9.7.1. Illustration of the optical field required to record a hologram.

where t_b is the background transmittance of the film due to the $|R|^2$ term in equation 9.7.2and B is a parameter which is a function of the recording and developing processes of the plate. Now suppose that the developed transparency is illuminated by a coherent wave, $C(x, y)$, known as the reconstruction wave. The optical field emerging from the transparency is then given by

$$C(x, y)t_p = Ct_b + BOO^*C + BO^*RC + BOR^*C$$

from which it can be seen that the output is the superposition of four waves. If we choose the reconstruction wave to be the same as the reference wave then the output field is simply given by

$$R(x, y)t_p = (t_b + BOO^*)R + BO^*R^2 + B\,|R|^2\,O. \tag{9.7.2}$$

The output now consists of a direct wave and two diffracted waves. One of the diffracted waves represented by the last term in this equation is a duplicate of the object wave (apart from a change in intensity) and carries the full information of the object. The only problem is that it is mixed with two waves which have little interest for us.

To appreciate the significance of the three different types of waves emerging from a hologram, consider the simplest possible hologram produced by the interference of two plane waves on a photographic plate as shown in the figure. With reference to the co-ordinate axes used in the figure, we label the wave that has a component of its propagation constant in the -x direction the *reference wave* while the other we refer to as the *object wave*. In the $x - y$ plane of the plate we can write the object wave as

$$O(x, y) = oe^{ikx\sin\theta_1}$$

and the reference wave as

$$R(x, y) = re^{ikx\sin\theta_2}$$

where θ_1 and θ_2 are the angles between the direction of propagation and the z-axis for the object and reference wave respectively. The intensity on the plate is then given by

$$\begin{aligned} I(x, y) &= |O|^2 + |R|^2 + or\exp\left(ikx\left[\sin\theta_1 - \sin\theta_2\right]\right) + ... + or\exp\left(-ikx\left[\sin\theta_1 - \sin\theta_2\right]\right) \\ &= |O|^2 + |R|^2 + or\cos\left(kx\left[\sin\theta_1 - \sin\theta_2\right]\right). \end{aligned}$$

The developed hologram is either a simple uniform phase of amplitude grating, depending on the developing process. Consider now a reconstruction wave of the form

$$R' = r'e^{ikx\sin\theta_3}$$

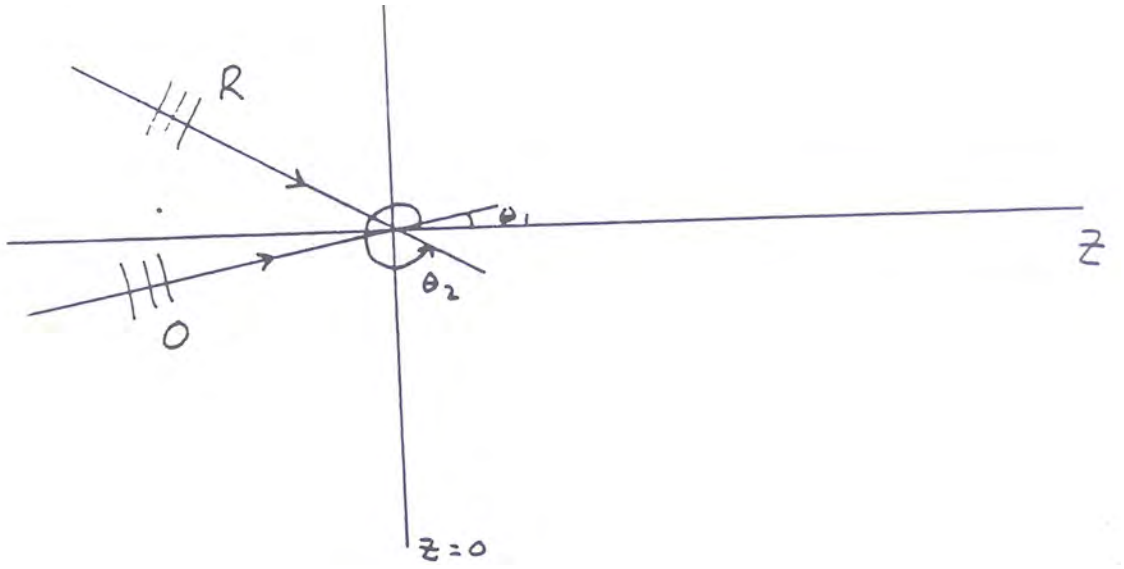


FIGURE 9.7.2. Recording a hologram of a plane wave

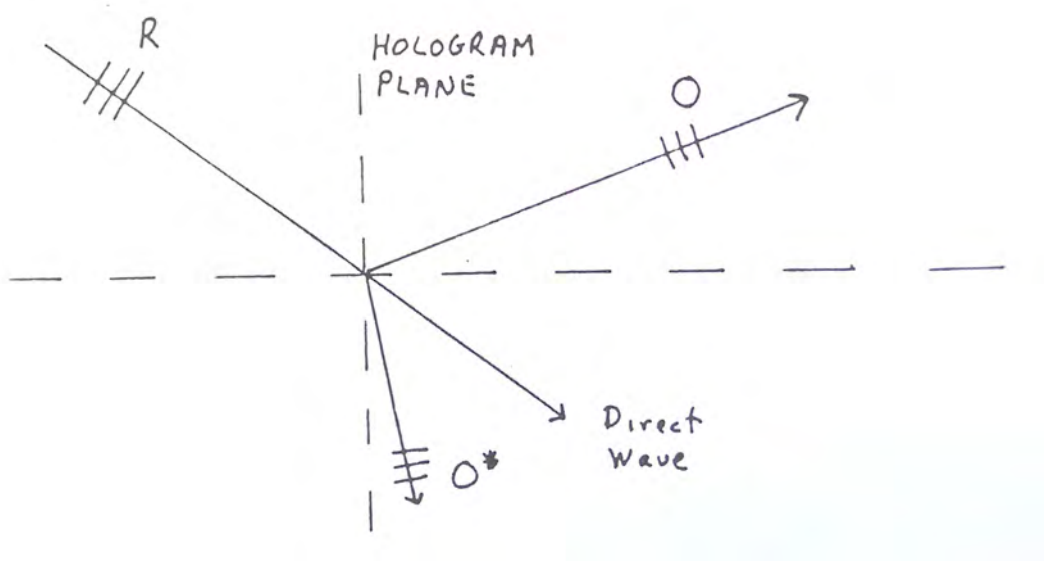


FIGURE 9.7.3. Illustrating the three waves emerging from a plane wave hologram illuminated by a reconstructing wave equivalent to the reference wave or its conjugate.

that is used to illuminate the developed hologram. For simplicity, allow t_b and B to be unity. The transmitted optical field is given by

(9.7.3)

$$R'I(x,y) = \left(|O|^2 + |R|^2 r'\right) e^{ikx\sin\theta_3} + ... + orr' \exp\left(ikx\left[\sin\theta_1 - \sin\theta_2 + \sin\theta_3\right]\right) + orr' \exp\left(ikx\left[\sin\theta_2 - \sin\theta_1 + \sin\theta_3\right]\right).$$

If the reconstruction wave has exactly the same characteristics as the reference wave (in particular $\theta_3 = \theta_2$) it is easy to see that the first wave on the right hand side of equation 9.7.3 is, apart from intensity, exactly the same as the reference wave, and is referred to as the direct wave. The second wave is the same as the original object wave and the last wave is the conjugate of the object wave travelling in a direction different from the original object wave. The three different waves are depicted in Figure 9.7.3. If the hologram is illuminated by the conjugate of the reference wave, one still obtains a direct wave, an object wave, and a conjugate wave although the direction of propagation has changed and the object and conjugate waves have switched positions as shown.

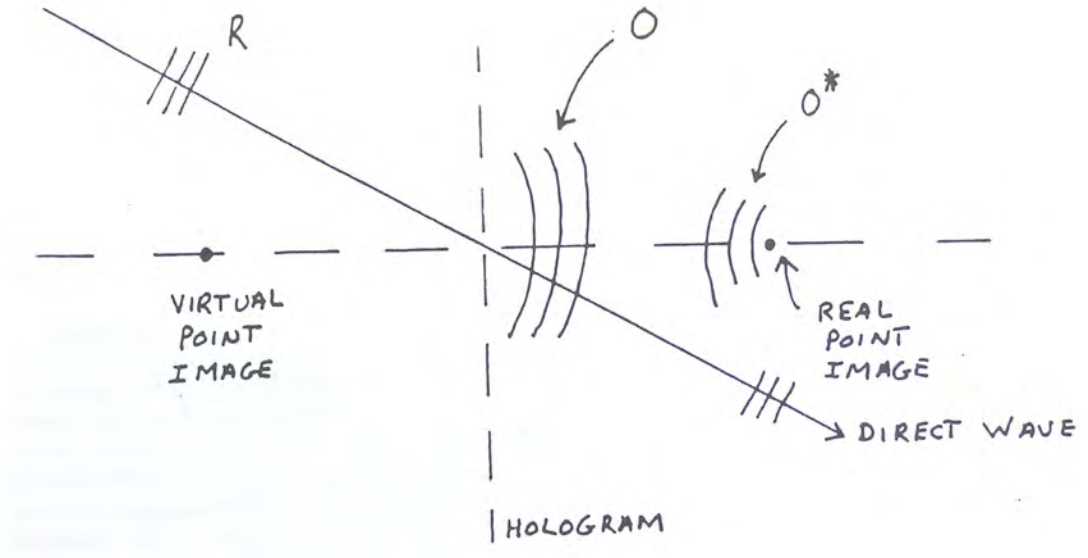


FIGURE 9.7.4. Direct, object and conjugate waves associated with a point source object.

It is obvious that for the recording geometry shown, in which the reference and object waves propagate in different directions, one can easily view the individual, separated waves. The particular geometry is referred to as off-axis holographic recording. In the on-axis configuration in which $\theta_1 = \theta_2 = \theta_3 = 0$, all the beams emerging from the hologram would be travelling in the same direction and it would be impossible to separate out the object wave from the direct and conjugate waves. This geometry was the one first proposed by Gabor to demonstrate the holographic principle, but remains as little more than a curiosity today. The off-axis technique developed in 1963 by *Leith* and *Upatneiks* of the University of Michigan is the preferred configuration. It might also be noted that Leith and Upatneiks were the first to realize the significant advantage offered by the laser as a coherent light source for recording and displaying holograms and did much to popularize the field of holography in the 1960s.

Before discussion of the practical details concerning the recording of realistic holograms it might be interesting to consider the meaning of the conjugate waves for other than plane waves. For example consider the situation in which the object wave is a spherical wave scattered from a point source in front of the recording plate while the reference wave is still a plane wave. The object wave at the plate can therefore be represented by

$$O(x, y) = oe^{ikr}$$

where $r = \sqrt{x^2 + y^2 + z^2}$ with z being the perpendicular distance of the point source from the plate. From equation 9.7.2 it is clearly seen that when the developed hologram is illuminated with the reference wave, besides the direct and object waves that emerge as plane and spherical waves respectively, the conjugate wave that emerges is of the form

$$BR^2O^*.$$

Assuming a reference wave of the form $R = re^{(ikxsin\theta_2)}$, as before, we have

$$BR^2O^* = Be^{2ikxsin\theta_2}.$$

This represents a spherical wave which is collapsing onto a point at a distance z in front of (to the right of) the hologram. The three waves, direct, object and conjugate are illustrated in Figure 9.7.4.

It is now evident that the conjugate wave is associated with the formation of a real image of the object since light actually passes through a point in front of the hologram. The object wave on the other hand represents a wave associated with a virtual image of the object, since no light actually emerges from a point in back of the hologram recording. Since any object can be viewed as being a composition of point source scatterers it is clear that we can extend the arguments made above for point sources to bulk objects.

Finally, as a matter of interest, it is worthwhile examining some geometries for recording and viewing holograms. A typical geometry is shown in Figure 9.7.5.

For simplicity, we consider only a transparent object. The geometry can be appropriately modified to deal with opaque three dimensional objects. Light from an appropriate spatially and temporally coherent source is expanded using a telescope. Part of the emerging beam is directed through the transparency while the remainder falls on a

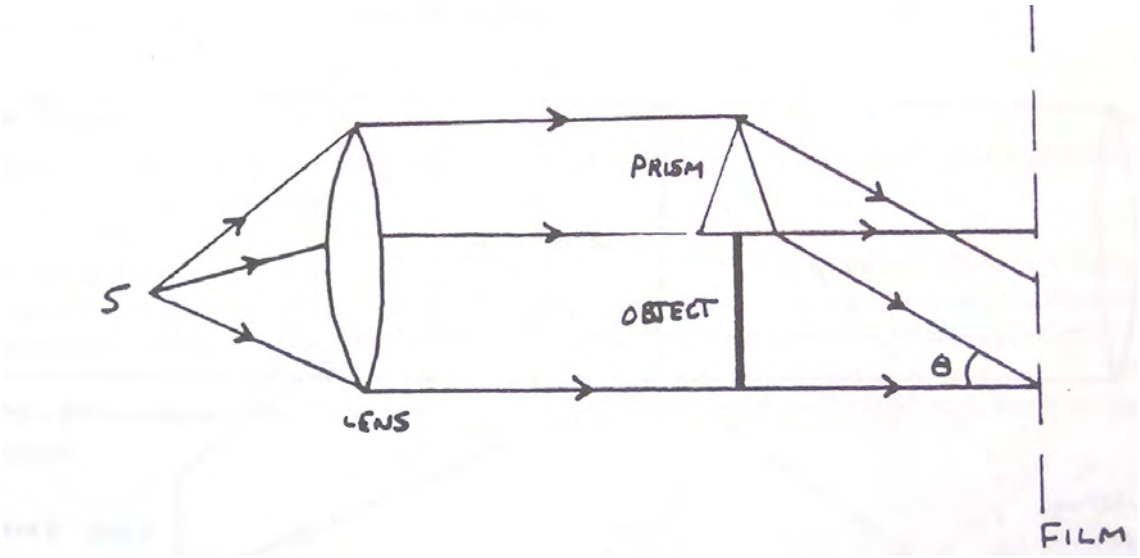


FIGURE 9.7.5. Geometry for recording an off-axis hologram of a transparent object.

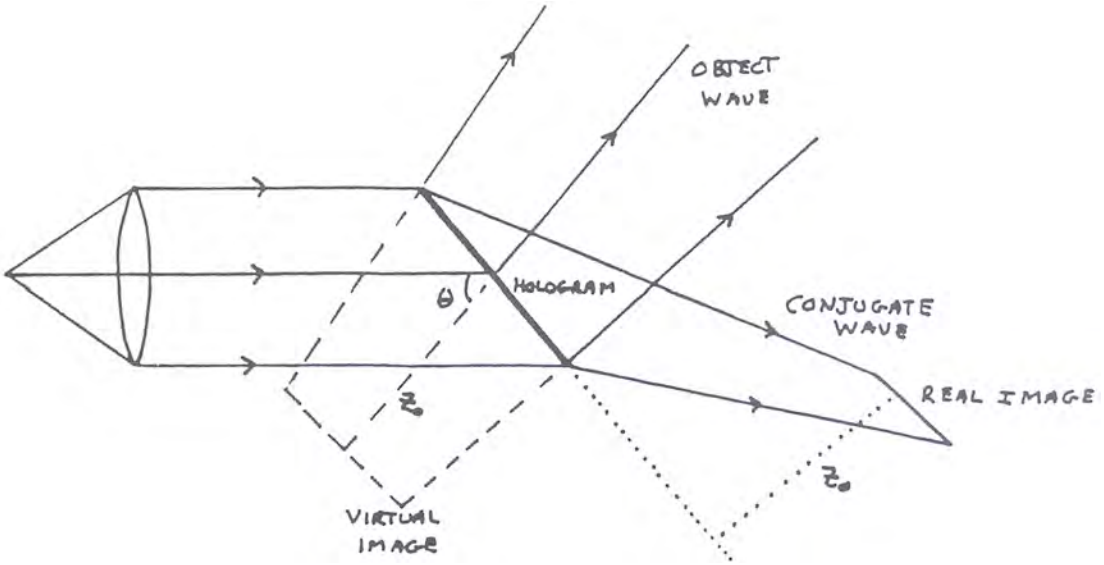


FIGURE 9.7.6. Reconstruction of the object wave using the original reference wave.

prism. The reference wave derived from the prism approaches the plane of the film at a non-zero angle of incidence and interferes with the object wave. Once the hologram is recorded and developed one can use as reconstruction geometry as illustrated in Figure 9.7.7.

As seen from the figure, if one views the holographic plate at normal incidence the object wave can be seen. At the same time a real image of the transparency is found below the direction of propagation of the reference wave. Note that if z is the distance of the transparency from the recording plate, then both the virtual and real image of the transparency are located at a distance of z from the plane of the hologram. If the object wave is constructed using a normally incident plane wave as shown in Figure 9.7.7, then once again, the distance of both the real and virtual images from the plane of the transparency is z.

Note, however, that in this case the real and virtual images present images of the object as seen from one side and so the images are distorted. The use of reference beams in reconstruction geometries different from recording geometries, in general leads to a distortion of the object and conjugate waves. The use of a reconstruction wave with a different wavelength than the reference beam gives a magnification or demagnification of the object in the ratio

$$M = \lambda/\lambda'$$

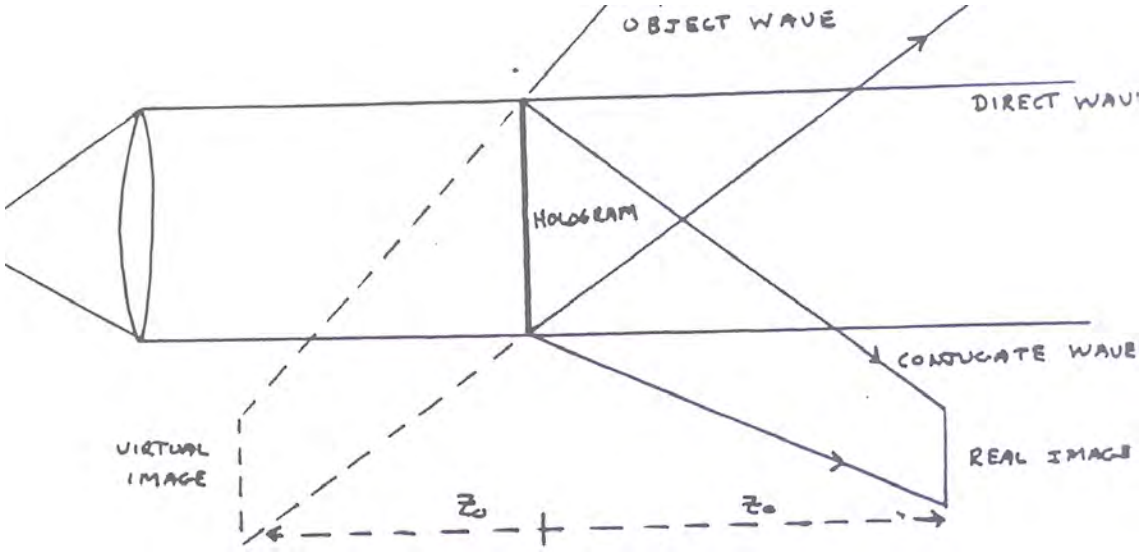


FIGURE 9.7.7. Reconstruction of the object wave using a normally incident reference wave.

where λ is the wavelength of the reconstruction wave and λ' is the wavelength of the reference wave. This relation is easily arrived at by reconsidering the recording and reconstruction process of a plane wave hologram.

Holograms these days are enjoying considerable popularity, particularly as new developments in the recording and reconstruction occur. Embossed, or reflection holograms viewed in white light are becoming cheap enough that there is serious consideration to replacing photographs with them in books and magazines. The use of white light in these applications gives a rainbow appearance to the image, since the hologram is essentially a diffraction grating which as we have seen has dispersive characteristics. There is also considerable discussion of the use of holography for data storage and retrieval and other manufacturing and industrial processes; however, to date the most prominent use of holography is simply to create realistic three dimensional images!

References

K. Iizuka, *Engineering Optics*, Springer-Verlag, Berlin, 1984.
J.D. Gaskill, *Linear Systems, Fourier Transforms and Optics*, John Wiley, Interscience, New York, 1978.
J.W. Goodman, *Introduction to Fourier Optics*, John Wiley, New York, 1968.
M.V. Klein, *Optics*, John Wiley and Sons, New York, 1970.
M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, New York, 1975.

Problems

1. What are the angular diameters of the second and third dark rings of the Fraunhofer diffraction pattern of a circular aperture.

2. A rectangular aperture is illuminated at normal incidence by parallel light of λ_0= 0.5 μm. The dimensions of the aperture are 1x3 mm. What are the dimensions of the main maximum of the diffraction pattern formed on a screen 50 m away and oriented parallel to the aperture?

3. Describe qualitatively the Fraunhofer diffraction patterns associated with each of the following apertures:

4. Describe analytically the intensity in the Fraunhofer diffraction pattern of an open rectangle with a centered rectangular obstruction. Sketch the results.

5. An optical beam with a wavelength λ and a Gaussian cross section

$$\mathcal{E}(x_0, y_0) = \mathcal{E}_0 \exp\left[-\frac{x_0^2 + y_0^2}{w_0^2}\right]$$

at z = 0 is propagating in free space. Determine the amplitude distribution of the Gaussian beam at $z > 0$ in the Fraunhofer approximation.

6. In the Fraunhofer diffraction pattern of a double slit, it is found that the sixth secondary maximum is missing. What is the ratio of the slit width to the slit separation?

FIGURE 9.7.8.

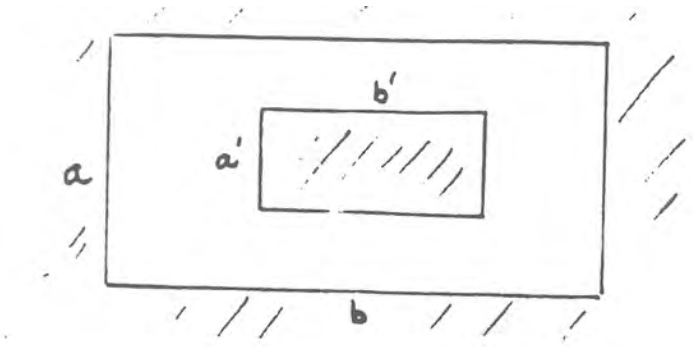


FIGURE 9.7.9. rectangle with obstruction

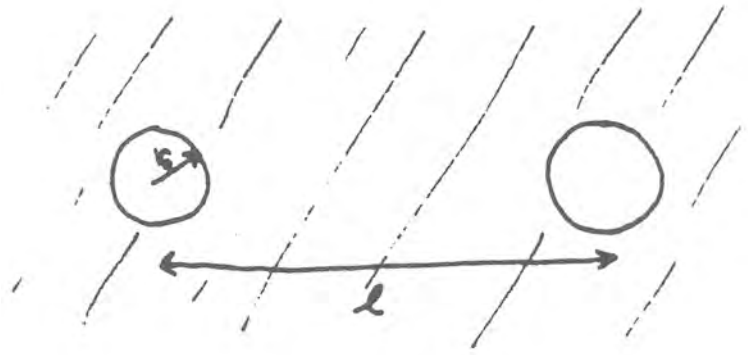


FIGURE 9.7.10.

7. An array consists of 100 point sources of coherent light of wavelength$\lambda_0 = 0.5$ µm. These are arranged in a straight line in vacuum with the separation between adjacent sources being 5 µm. The sources are anti-phased so that the relative phase between adjacent sources is π. Determine the direction and angular extent of the most intense portion of the emitted beam in the far field.

8. Calculate the diffraction pattern of an apodized slit for which the transmission function is $[0.5+0.5\cos(2\pi y/b)]$ for $-b/2 < y < b/2$ and zero otherwise. Find the relative intensity of the first secondary maximum.

9. Describe quantitatively the electric field in the Fraunhofer diffraction pattern of the double circular aperture. Sketch roughly the two-dimensional distribution of irradiance when $r_0 = 1/2$, $1/100$.

10. In a spatial filter which consists of a 3 mm diameter lens of focal length 10 mm, what size aperture must be used in the focal plane so that all light beyond the first node is blocked?

11. A simple hologram is made as follows: the object is a single narrow strip located a distance d from the base of the recording plate. The wavelength of the laser is λ. The plate is illuminated normally by the reference

beam. Show that the resulting pattern on the hologram is a one-dimensional grating with a variable spacing in the y-direction. Give the numerical values of this spacing for $\lambda_0 = 0.63$ μm, $d = 10$ cm when y = 0 and 10 cm.

12. In this chapter, the size of the hologram was considered infinite. Find the expression for the reconstructed image when the hologram dimension in the x direction is reduced to a. Assume that the object is one dimensional and is a function of only x, and assume that the angle of incidence of the reconstruction and reference beams are the same.

13. The reconstruction beam is incident along the z axis. What is the change in position of the reconstructed image when the plane of the hologram is rotated by an angle a from a position perpendicular to the z-axis?

14. A thin optical plate of thickness t has an inhomogeneous refractive index which varies with distance ρ from a central point as $n = n_0(1 - \rho^2/\rho_0^2)$. Show that for sufficiently small thicknesses this plate behaves like a lens and determine the focal length of the lens.

CHAPTER 10

Gaussian Beams

The truth, but not the whole truth
Baltasar Gracian

10.1. Paraxial Optics

In earlier chapters plane waves were found to be useful for discussions of many elementary optical effects. At the opposite extreme of plane waves, we have the optical rays, which are "pencils" of light with no width and which form the basis for geometrical optics, whereby we ignore the wavelength of light. Most realistic optical beams have a finite transverse extent, and for most practical situations this is important to consider.

Of particular interest are those waves which have finite transverse extent and relatively small phase variations along directions perpendicular to the average direction of propagation, but which can also form a complete set for describing any optical beam. We have already seen one example of such a wave in the chapters on diffraction and Fourier optics where we considered the paraxial section of a spherical wave. We saw that we could replace

$$S(r) = \frac{e^{ikr}}{r} \text{ by } h(x,y,z) = \frac{1}{z} e^{ikz} e^{ik(x^2+y^2)/2z}$$

for values of *x, y* small compared to *z*. The functional form basically represents the impulse function associated with Fresnel diffraction. This is a particular type of paraxial wave. In general, we consider solutions to the wave equation, $\Psi(x,y,z)$, to be of a paraxial nature if their phase variations in the *x,y* direction are small compared to their phase variations in the direction of propagation (z). An alternative way of saying the same thing is to say that the $\vec{k}$ vectors associated with the plane wave expansion of the optical wave make small angles with respect to the z-axis. Notice that the paraxial waves do not necessarily imply finite transverse extent of the waves. Indeed with the definition we have for paraxial waves the plane wave e^{ikz} would be considered paraxial. For the paraxial waves we could consider writing

$$\Psi(x,y,z) = u(x,y,z)e^{ikz}$$

where $u(x,y,z)$ is called the *envelope function*. When this form is substituted in the *Helmholtz equation*,

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

we obtain by direct substitution

$$\nabla_T^2 u + \frac{\partial^2 u}{\partial z^2} + 2ik\frac{\partial u}{\partial z} = 0$$

where

$$\vec{\nabla}_T = \hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y}.$$

We can consider the class of envelope functions, u, which vary only slowly along the direction of propagation. In particular, take the change in the function over a wavelength to be small, *i.e.*,

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

and also take the function to be smooth on the same scale, *i.e.*,

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

With this *slowly varying envelope approximation* (SVEA), we arrive at what is known as the paraxial wave equation:

$$\nabla_T^2 u + 2ik\frac{\partial u}{\partial z} = 0$$

which is an approximate form of the wave equation. It can be verified that the function h(x,y,z) is an exact solution of the paraxial wave equation.

One of the most important types of paraxial waves is a Gaussian beam which is a particular solution of the paraxial wave equation. In this chapter we develop the theory of Gaussian beams and consider their properties in free space and in optical resonators. We also consider a family of solutions to the paraxial equations, the Hermite-Gaussian beams, of which the Gaussian beam is a special member. These beams in general are very important in the discussion of light field distributions emerging from laser systems and Fabry-Perot resonators. There are many ways to introduce such beams, none of which is particularly insightful, and most of which are mathematically cumbersome. For example, for Fabry-Perot resonators made with curved mirrors we might solve the wave equation with appropriate boundary conditions. The natural modes of the resonator are Hermite-Gaussian beams but they can only be identified by extensive, self-consistent mathematical (computer) calculation. Our approach is much more pragmatic. We introduce a particular solution to the paraxial wave equation and later show that this particular solution satisfies the requirement of a mode of a resonator.

10.2. Gaussian Beams

To introduce the Gaussian beam we note that the paraxial wave equation is invariant with respect to a translation of the co-ordinate z to $z - z_c$ where z_c is a constant. In particular, a very interesting solution of the paraxial wave equation occurs if we consider the function $h(x, y, z)$ translated by the amount iz_0 where z_0 is a real constant. The function, $h(x, y, z - iz_0)$, which obviously satisfies the paraxial wave equation, has an envelope function with the singularity on the z-axis (at $z = 0$) removed. For reasons to be explained later, we label this function u'_{00} . It is given by

$$u'_{00}(x, y, z) = \frac{1}{z - iz_0} exp\left[ik\frac{x^2 + y^2}{2\,(z - iz_0)}\right].$$

Like h(x,y,z), the function u'_{00} is cylindrically symmetric about the z axis. It is convenient to normalize u'_{00} (to give u_{00}) through multiplication by a constant so that

$$\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, |u_{00}|^2 = 1.$$

Normalization at one cross section, say at $z = 0$, assures that the normalization is the same at other values of z, by conservation of power. When the integral is carried out it is found that

$$u_{00}(x, y, z) = \sqrt{\frac{kz_0}{\pi}}\frac{1}{z - iz_0} exp\left[ik\frac{x^2 + y^2}{2\,(z - iz_0)}\right].$$

Apart from a constant phase factor this can be put in the form

$$u_{00}(x, y, z) = \sqrt{\frac{2}{\pi}}\frac{1}{w} e^{-i\phi} \exp\left(-\frac{x^2 + y^2}{w^2}\right) \exp\left(\frac{ik(x^2 + y^2)}{2R}\right)$$

where

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

$$R(z)^{-1} = \frac{z}{z^2 + z_0^2}$$

and finally

$$\tan\phi = \frac{z}{z_0}.$$

This particular solution of the paraxial wave equation is the *fundamental Gaussian beam* solution. Note that apart from the propagation constant (or wavelength) and the location of the origin ($z = 0$) a single parameter (*e.g.* z_0) completely defines the form of the beam. Before proceeding to use this solution we should understand the various factors which contribute to this wave form this.

The properties of the Gaussian beam solution are:

1) The beam has a field and intensity profile which are a Gaussian function of the transverse variable $r = \sqrt{x^2 + y^2}$. The parameter w represents the value of r at which the field drops to e^{-1} of its value on axis. The parameter is sometimes referred to as the *fundamental spot size* since it is a measure of the transverse extent of the beam. The constant w_0 is the minimum spot size and occurs at $z = 0$. Conversely we might state that the choice of the displacement of the origin by the imaginary distance iz_0 has fixed the minimum spot size. The distance z_0, known as the *confocal parameter*, is the distance over which the spot size increases by a factor of 2 from the beam waist .

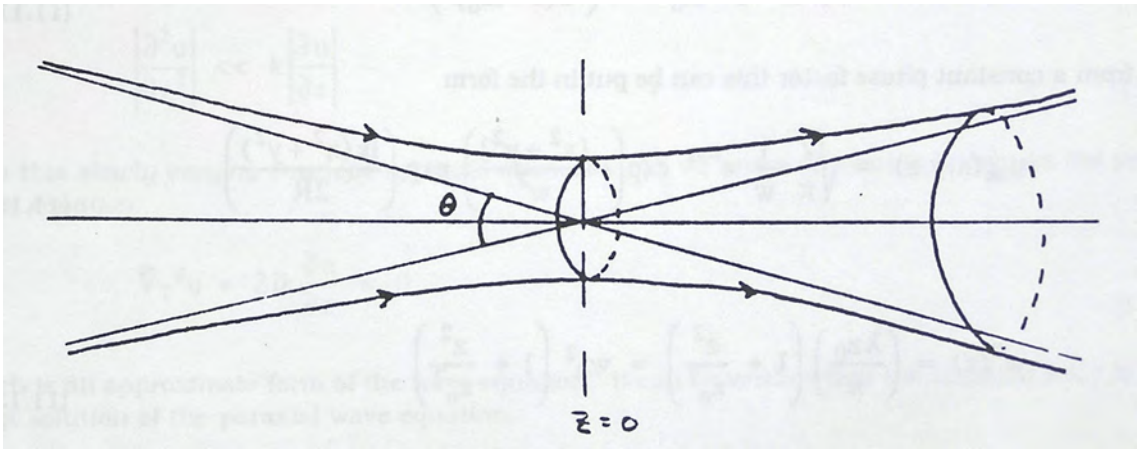


FIGURE 10.2.1. Illustration of the surfaces of equation 10.2.1.

2) Surfaces for which the intensity is a constant fraction of the on-axis intensity (at the same value of z) are defined by the equation

$$\frac{r^2}{w_0^2\left(1+\frac{z^2}{z_0^2}\right)} = C = constant$$

or

$$x^2 + y^2 - \frac{Cw_0^2}{z_0^2}z^2 = Cw_0^2. \tag{10.2.1}$$

These represent hyperboloids of revolution as illustrated in Figure 10.2.1.

Note from the figure that the confocal parameter is a measure of the distance over which the beam is quasicollimated. It is akin to the *depth of focus* or *depth of field*, terms which are used by camera savants, hence the name.. The parameter z_0 varies as w_0^2. Hence, for a more tightly focused Gaussian beam, one has a smaller depth of field over which the beam appears to be collimated. For example, if $\lambda_0 = 1$ μm and $w_0 = 1$ mm we obtain a depth of field of π m, but if $w_0 = 10$ μm we obtain a depth of field of $\pi \times 10^{-4}$ m!

3) In the far field where the hyperbolic surfaces approach asymptotes, we can calculate the uniform rate of divergence of the beam. For $z \gg z_0$ we have that $w \propto z$. It follows that if θ is the full cone angle determined by the asymptotes, then

$$tan\frac{\theta}{2} = \frac{w(z)}{z}$$

$$= \frac{w_0\frac{z}{z_0}}{z} \simeq \frac{\lambda}{\pi w_0} \simeq \frac{\theta}{2} \text{ for small } \frac{\lambda}{\pi \mathrm{w}_0}.$$

For a 1 mm fundamental spot size and $\lambda = 1$ μm, we obtain a full angle of divergence(θ) of the beam of $\approx 10^{-3}$ radians (the spot increases in size by about 1 mm for each metre of travel). For a 1 μm fundamental spot size the full angular divergence is greater than 1 radian. In this case one can question the Gaussian beam solution as being a valid solution of the paraxial wave equation.

4) The quantity *R(z)* is the radius of curvature of the surfaces of constant phase as shown in Figure 10.2.2.

At the beam waist the radius of curvature is infinite, as the defining equation for *R(z)* indicates. Alternatively, the plane for which $R = \infty$ could be used to define the location of the beam waist. For $z \gg z_0$ we find that $R(z) = z$, which means that the beam in the far field is propagating like a portion of a spherical wave. This is consistent with our starting point since if $z \gg z_0$ the Gaussian beam is essentially the same as *h(x,y,z)*. Finally, note that the surfaces of constant phase are locally perpendicular to the surfaces of constant field or intensity, as must be for solutions to the homogeneous wave equation.

5) The phase factor ϕ , referred to as the *Guoy phase shift*, influences the velocity of surfaces of constant phase. The phase speed of the Gaussian beam is not the speed of a plane wave in whatever medium the beam is propagating, which, in this case we have taken to be vacuum. We can determine the effective propagation constant from $\Psi(x,y,z)$ through the definition

$$\int_0^z k_{eff}dz = kz - \phi(z)$$

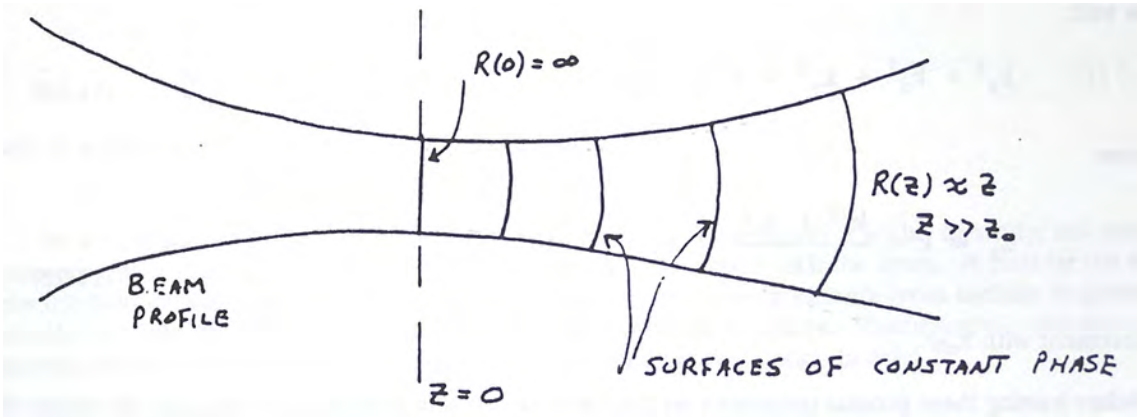


FIGURE 10.2.2. Illustration of the surfaces of constant phase for a Gaussian beam.

so that

$$k_{eff} = k - \frac{d\phi}{dz} = \frac{\omega}{c} - \frac{z_0}{z^2 + z_0^2} < \frac{\omega}{c}.$$

The phase velocity is everywhere greater than the speed of light. At $z = 0$ in particular, the effective propagation constant is

$$k_{eff} = k - \frac{2}{kw_0^2}.$$

The fact that the phase velocity is greater than c can be explained by the finite transverse extent of the beam. Such a beam can be written as a superposition of plane waves which have propagation vectors, $\vec{k}$, which are oriented at slightly different angles relative to the z- axis. If we consider the beam in the vicinity of $z = 0$ say, the typical x and y components of the propagation vector $\vec{k}$ of these waves are

$$k_x = k_y \simeq \frac{\sqrt{2}}{w_0}.$$

Thus with

$$k_z^2 + k_x^2 + k_y^2 = k^2$$

we have

$$k_z = k - \frac{k_x^2 + k_y^2}{2k} = k - \frac{2}{kw_0^2}$$

in agreement with k_{eff}.

Before leaving these general comments on Gaussian beams it is interesting to examine the range of validity of the Gaussian beam as a solution to the paraxial wave equation. The key approximations we made in arriving at the paraxial wave equation are that

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

and,

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

For the Gaussian beam solution we have that

$$\frac{\partial u_{00}}{\partial z} = -\left[\frac{1}{z - iz_0} + \frac{ik(x^2 + y^2)}{2(z - iz_0)^2}\right] u_{00}.$$

The omission of this term compared with $k|u|$ implies that

$$\frac{1}{z_0} \ll k$$

or that

$$\frac{1}{z_0 k} = \frac{\lambda^2}{2\pi^2 w_0^2} \ll 1. \tag{10.2.2}$$

We therefore require that the beam waist must be large compared to the wavelength. Further, we must have

$$\frac{x^2+y^2}{z^2+z_0^2} \ll 1.$$

Since x^2+y^2 is of the order of w^2, using the expression for w^2 in terms of z and z_0, we find that

$$\frac{w_0^2}{z_0^2} \ll 1$$

which is the same condition as derived in equation 10.2.2.

As a final point it should be noted that the function u_{00} is obviously a scalar quantity, but more importantly, it is an approximation for the electric field associated with the beam. It can't be the exact electric field because Gauss' law is not satisfied exactly. For a beam of finite cross section, in general the electric field must be a vector field that is not transverse in nature. Identifying u_{00} with the electric field is valid to the same extent that the paraxial approximation is valid.

10.3. Gaussian Beams in Resonators with Curved Mirrors: Form of the Modes

In our earlier discussion of Fabry-Perot resonators, we considered only two plane mirrors. In that discussion we were not concerned with the lateral boundaries of the mirrors; we treated the mirrors as infinite in transverse extent. In reality, the transverse amplitude distribution of a mode in an actual Fabry-Perot interferometer with planar mirrors of finite transverse extent is controlled by the diffraction of the waves at the mirror boundaries, causing diffraction losses. Curved mirror Fabry-Perot resonators practically eliminate the effects of mirror boundaries on the field distributions of resonant modes and the associated diffraction losses. This is described below.

Before we do this we might remind ourselves of how a mode is defined and further how one defines one in the context of a curved mirror resonator. For the plane mirror resonator, a transmission peak corresponded to the existence of a "standing wave" pattern in the resonator. The simple way of stating the requirement for a mode for this type of resonator is to say that the separation between the mirrors is a half-integral multiple of the wavelength, λ. Note that because we never considered resonators with mirrors of 100% reflectivity (since we wanted to get some light through the interferometer), we never had true standing waves. Only in the case where the mirrors have 100% reflectivity could we truly talk about perfect standing waves of the electromagnetic field between the mirrors with nodal planes coincident with the mirror planes. Because our imperfect modes were leaky this allowed us to couple light in and out of the resonator. Therefore, as a reference point, we adopt as definition of a mode those field structures that occur when the mirror reflectivities are 100%. To make the definition sufficiently general so that we can deal with curved mirror resonators, we define an electromagnetic mode as a field structure that satisfies the following conditions:

1) The amplitude of the field at a particular point is stationary, *i.e.* time-independent. Note that because the mirrors have perfect reflectivity, this implies no coupling with the outside world.

2) The phase associated with the field structure is stationary. This implies that the round trip phase change associated with the travelling waves that make up the standing wave is a multiple of 2π.

For the moment we neglect diffraction losses since they would give rise to loss of field amplitude. We return later to the conditions necessary to justify this assumption.

Our first question might be: how, in perfect resonators, do the radius of curvature of the mirrors determine the field structures or modes? Rather than solve the Helmholtz equation with appropriate boundary conditions for the curved mirror resonator we (with the benefit of other people's experience) guess that Gaussian beams form the appropriate basis for constructing the resonator modes. The superposition of two oppositely travelling Gaussian beams has the form

$$\left[u_{00}(x,y,z)e^{ikz} + u_{00}(x,y,-z)e^{-ikz}\right] e^{-i\omega t}$$

and forms a standing wave pattern with nodal surfaces of the electric field parallel to the phase front whose radius of curvature is given by

$$R^{-1} = \frac{z}{z^2+z_0^2}.$$

To obtain modes of the curved mirror resonator, all that is required is to find the Gaussian beam whose surfaces of constant phase match the radii of curvature of the mirrors for a separation, L, between the mirrors as shown in Figure 10.3.1.

Alternatively, if a Gaussian beam already exists, all we have to do is place our mirrors at points on the Gaussian beam where we can match the radii of curvature of the mirrors to the radii of curvature of the phase fronts. A

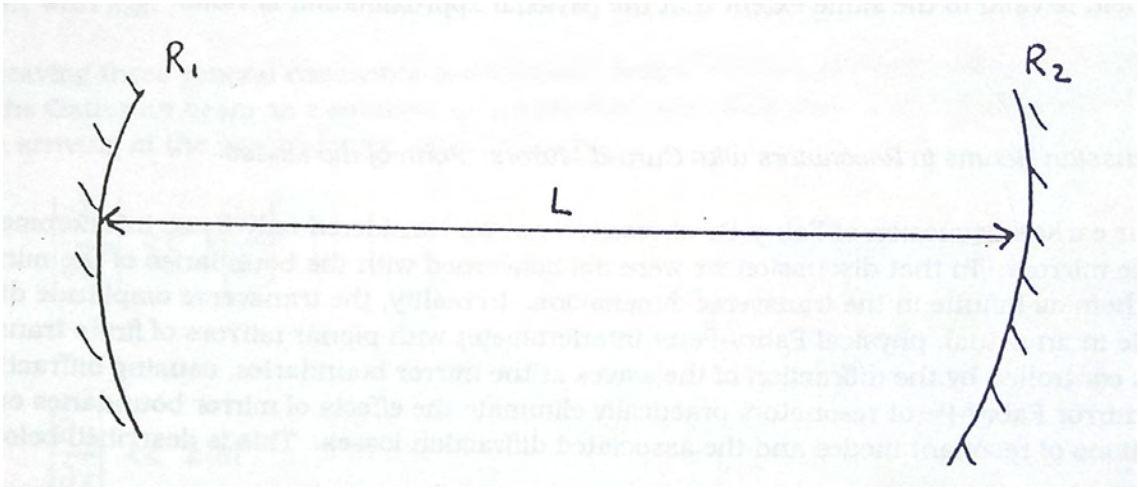


FIGURE 10.3.1. Resonator geometry.

standing wave then results. For example, with respect to the figure, where we have concave mirrors with radii of curvature R_1 and R_2, if mirror number 1 is placed at a position z_1 such that

$$-R_1^{-1} = \frac{z_1}{z_1^2 + z_0^2} \tag{10.3.1}$$

and mirror number 2 is placed at position z_2 such that

$$R_2^{-1} = \frac{z_2}{z_2^2 + z_0^2} \tag{10.3.2}$$

one may "capture" the standing wave. As we did with the radii of curvature of the lens surfaces, we take R_1 and R_2 to be positive if the mirrors are convex as they are approached, otherwise we take them to be negative. The - sign associated with R_1 in equation 10.3.1 is there because the Gaussian beam has a negative radius of curvature if it is expanding while propagating to the left along the -z axis. If the mirror diameters are chosen so that they have a transverse extent much larger than that of the Gaussian beams they are confining, they need not be infinite to make diffraction effects negligible.

Of course, the round trip phase change inside the resonator must correspond to a multiple of 2π in order to satisfy the second condition required of a mode. We return to this condition later.

Given two mirrors with radii of curvature R_1 and R_2 separated by a distance L, we determine the Gaussian beam standing wave by attacking it as a "fitting" problem in which we exploit equations 10.3.1 and 10.3.2 together with

$$z_2 - z_1 = L$$

This gives three equations in three unknowns, z_0, z_1, z_2. Once these parameters are found one can determine all the parameters of the Gaussian beam, including the location of the beam waist (at $z = 0$). The expressions for z_0, z_1,z_2. are complicated functions of L, R_1, and R_2, and are more easily expressed in terms of the *resonator parameters*, g , defined by

$$g_1 = \left(1 - \frac{L}{R_1}\right) \qquad g_2 = \left(1 - \frac{L}{R_2}\right).$$

It follows that the confocal parameter is

$$z_0^2 = \frac{L^2 g_1 g_2 (1 - g_1 g_2)}{(g_1 + g_2 - 2 g_1 g_2)^2}.$$

Although the parameters z_1 and z_2 can also be found we won't do so here. More useful parameters to consider are the spot sizes at the two mirrors which are found to be

$$w_1^2 = \frac{L\lambda}{\pi}\left(\frac{g_2}{g_1(1 - g_1 g_2)}\right)^{1/2} \qquad w_2^2 = \frac{L\lambda}{\pi}\left(\frac{g_1}{g_2(1 - g_1 g_2)}\right)^{1/2}.$$

We can only have a confined field distribution if the spot sizes at the mirrors is finite, otherwise diffraction effects are important. It is easy to show that diffraction losses between the two mirrors are negligible if

$$F' = \frac{r_1 r_2}{L\lambda} \gg 1$$

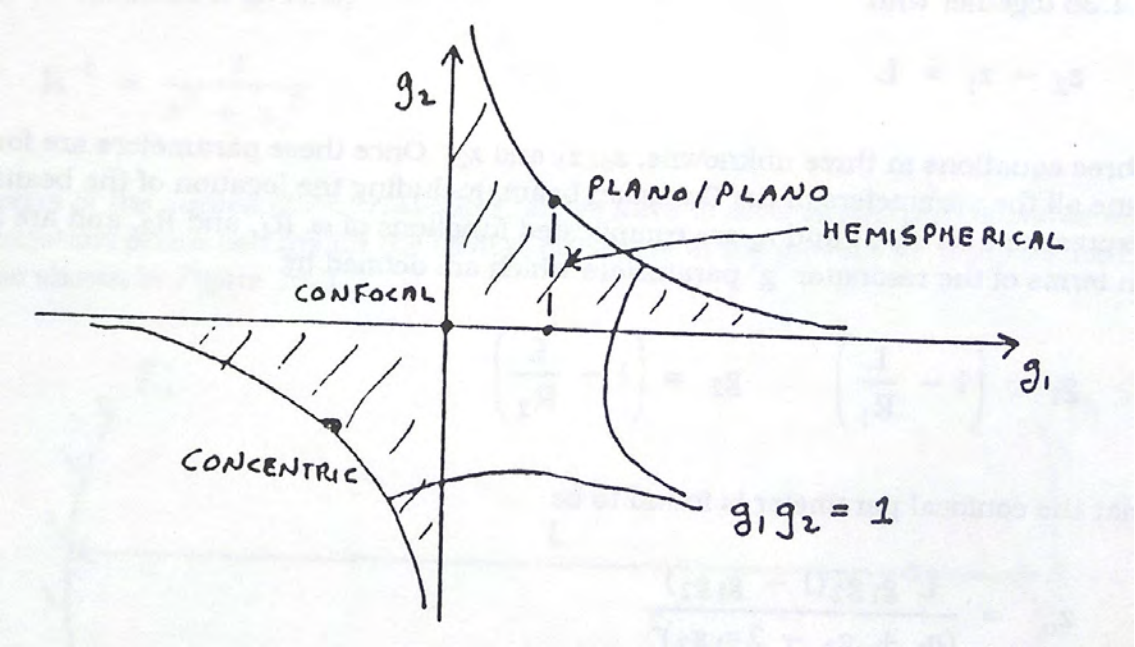


FIGURE 10.3.2. Region in g-space for which resonator solutions exist.

where r_1 and r_2 (not to be confused with the radii of curvature) are the radii of the (circular) mirrors. The parameter F' is known as the *Fresnel number*. The condition can be broken down to read

$$\frac{r_1}{L} \gg \frac{\lambda}{r_2}$$

which can be interpreted as saying that the angle subtended by mirror number 1 at mirror number 2 is much greater than the angle associated with the diffraction light cone produced when light propagates away from a "source" of width of the order of r_2. The roles of r_1 and r_2 are reversible in the expression.

However, as a sufficient condition for mode existence, we can adopt the criterion that the denominators in the expressions for w_1 and w_2 are non-zero. This guarantees the spot sizes on the mirrors are finite and, provided the mirrors are much larger than these spot sizes we always have confined beams with insignificant diffraction losses. A little thought shows that we must then have

$$0 < g_1 g_2 < 1.$$

This is the confinement condition for modes of a Fabry-Perot resonator. With respect to a two dimensional space labelled by g_1 and g_2 axes, we see that it is possible to define stable modes in the shaded regions indicated in Figure 10.3.2. This region is bounded by the hyperbola $g_1 g_2 = 1$ and the the two axes $g_1 = 0$, $g_2 = 0$.

It should be remembered that for a point inside this region we could find a Gaussian mode for the defined resonator provided the mirrors are large enough so that the spot size at the mirrors is much less than their diameter. In what follows we assume that this is the case. For all other points there is no possibility of finding a Gaussian beam mode and these regions correspond to high loss "resonators". They are usually referred to as *unstable resonators* with the other resonators being referred to as *stable resonators*.

The most common types of stable resonators encountered are indicated in Figure 10.3.2. These are:

1) Plano/plano or *planar resonator;* here $R_1 = R_2 = \infty$, *i.e.* both mirrors are flat. For this resonator $g_1 = g_2 = 1$ and the resonator is just on the verge of being a stable resonator. This resonator, which we first discussed in connection with the Fabry-Perot interferometer can only be filled by a Gaussian beam of infinite transverse extent or $z_0 = \infty$. As a result no finite sized resonator could be constructed to give negligible loss. From the ray point of view this resonator can also be seen to be unstable, since any ray that is launched at an arbitrarily small angle relative to the z-axis gradually "walks out" of a finite sized resonator.

2) *Confocal resonator* ($g_1 = g_2 = 0$). This corresponds to one that has the focal points of the two mirrors (mirrors have a focal length of $R/2$) located exactly at the center of the resonator. Note that, provided the two mirrors are truly identical such a resonator is stable against length fluctuations.

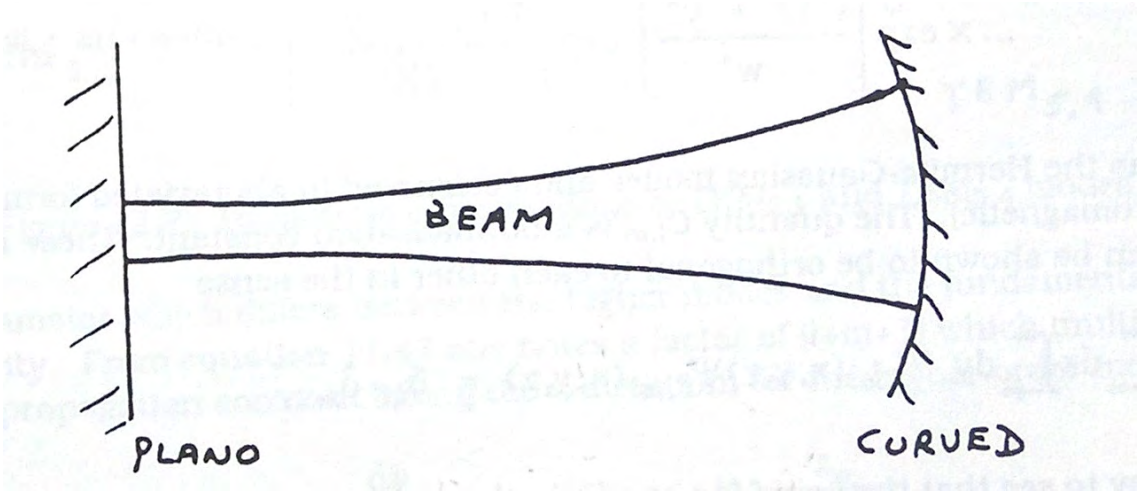


FIGURE 10.3.3. Beam profile in a hemispherical resonator.

3) *Concentric resonator* ($g_1 = g_2 = -1$). This resonator, which has identical mirrors separated by twice their radius of curvature, is unstable since it is not possible to find a Gaussian beam to satisfy this condition. This would require a value of $z_0 = 0$ which would make the spot size at the mirrors infinite.

From these comments, it is seen that some of the simplest types of symmetric resonators are in fact unstable. One can of course find symmetric resonators which are stable (like the confocal resonator which is used mainly in Fabry Perot interferometers). In lasers the type of stable resonator which is encountered most often is the

4) *Hemispherical resonator*. For this resonator there is one flat and one curved mirror with $g_1 = 1$ and $g_2 < 1$. The Gaussian beam intensity distribution inside such a resonator is shown in Figure 10.3.3.

The waist of the Gaussian beam obviously has to be at the flat mirror since this is the point where the radius of curvature is infinite. The spot size at this point is

$$w_1 = w_0 = \left(\frac{L\lambda}{\pi}\right)^{1/2} \left(\frac{g_2}{(1-g_2)}\right)^{1/4}$$

while the spot size at the curved mirror is

$$w_2 = \left(\frac{L\lambda}{\pi}\right)^{1/2} \left(\frac{1}{g_2(1-g_2)}\right)^{1/4}.$$

As $g_2 \to 1$ one obtains the planar resonator which, as we have seen, only supports plane waves and strictly speaking is unstable. For this case the spot size at both mirrors is infinite. Similarly as $g_2 \to 0$, the spot size at the first mirror goes to zero and the spot size at the second mirror approaches infinity for the rapidly diverging beam. In between, one obtains stable solutions. Because of the quartic root which occurs in the expressions for the spot size, the spot sizes and Gaussian beam parameters are quite insensitive to the value of g_2 over a broad range of g_2 away from the two singular points of g_2. For example, with $L = 1$ m, $\lambda_0 = 1$ μm and $R_2 = 5m$ one obtains $w_0 = 0.8\,mm$ and $w_2 = 0.85\,mm$. If we change R_2 to 20 m, we find that $w_0 = 1.2\,mm$ and $w_2 = 1.3\,mm$, which is only a modest increase. Unless g_1 or g_2 are very close to the range of instability the spots sizes are typically in the range of 1 mm for visible or near visible light Gaussian modes. This explains why most visible lasers have a Gaussian beam spot size close to 1 mm.

10.4. Hermite-Gaussian Modes

The Gaussian beam solution to the paraxial wave equation represents only one particular solution. There is a whole family of solutions of which the Gaussian beam is only the simplest. These modes which are functionally products of Hermite polynomials and the fundamental Gaussian solution have the form

$$\Psi_{l,m}(x,y,z) = u_{l,m}(x,y,z)e^{ikz}$$

$$= \frac{C_{l,m}}{w(z)} H_l\left[\frac{x\sqrt{2}}{w(z)}\right] H_m\left[\frac{y\sqrt{2}}{w(z)}\right] \exp\left(-\frac{x^2+y^2}{w^2}\right) \exp\left(\frac{ik(x^2+y^2)}{2R}\right) e^{-i(l+m+1)\phi} e^{ikz} \tag{10.4.1}$$

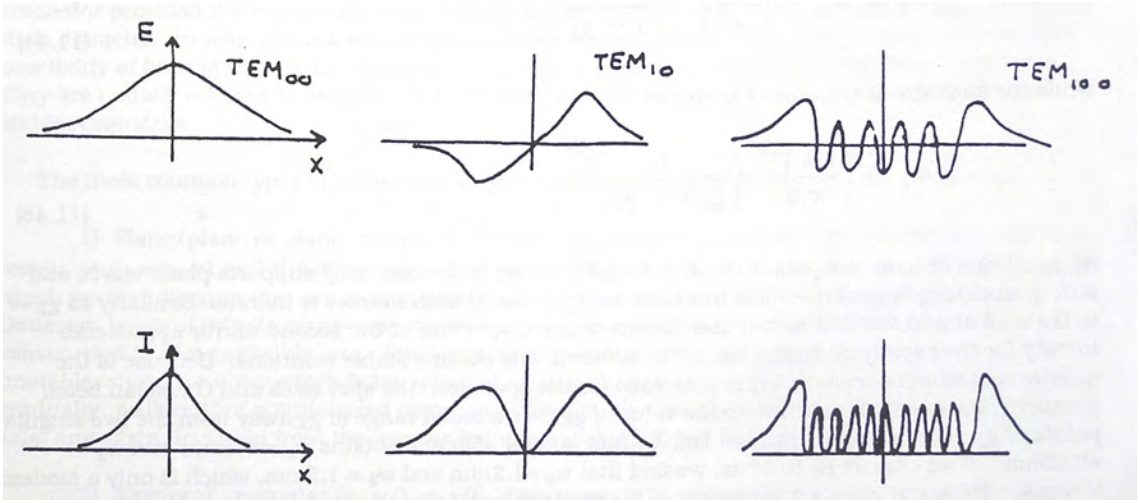


FIGURE 10.4.1. Field and intensity dependence of $TEM_{0,0}$, $TEM_{1,0}$ and $TEM_{10,0}$ modes

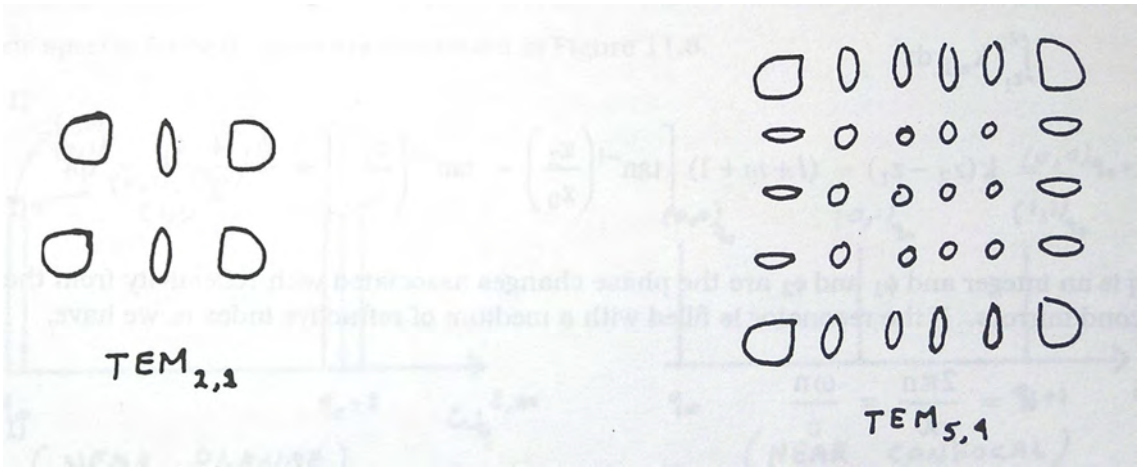


FIGURE 10.4.2. Transverse cross sections of $TEM_{2,1}$ and $TEM_{5,4}$ modes

and are known as the *Hermite-Gaussian modes* and designated in abbreviated form as $\mathrm{TEM_{l,m}}$ (for *transverse electromagnetic*) modes. The quantity $C_{l,m}$ is a normalization constant. These modes, for different values of *(l,m)* can be shown to be orthogonal to each other in the sense

$$\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \Psi_{l,m} \Psi^*_{l',m'} = \delta_{l,l'} \delta_{m,m'}.$$

It is relatively easy to see that this set of functions is also complete in all space so that any given optical disturbance can be expanded in terms of a linear combination of these functions. Note that the form of the Hermite-Gaussian beams is completely determined by the same factors that determine the propagation of Gaussian beams. The Hermite polynomials not only complicate the transverse field distribution but they also locate more of the beam energy away from the axis compared to the fundamental Gaussian beam. Typical transverse field and intensity patterns are shown in Figure 10.4.1.

It can be seen that for those modes that have one of the mode indices equal to zero the number of nodes is equal to the non-zero index. The strongest peak is the one furthest off-axis. Note, as well, that the spot-size is no longer $w(z)$, but is larger and increases with the mode number. In the case of both $l,m \neq 0$ the cross section of the beam appears as a matrix with the number of orthogonal nodal planes equivalent to l or m. Transverse cross sections of the intensity for $\mathrm{TEM_{1,2}}$ and $\mathrm{TEM_{5,4}}$ modes are illustrated in Figure 10.4.2.

The other parameter which differs from that of the fundamental Gaussian beam is the phase velocity. From equation 10.4.1 one notes a factor of $(l + m + 1)$ that multiplies ϕ. It follows that the effective propagation constant

(along the z-direction) for the higher order modes is

$$k_{eff} = k - (l+m+1)\frac{d\phi}{dz} = k - (l+m+1)\frac{z_0}{z^2+z_0^2}.$$

This is always less than the k associated with light propagation in vacuum, and indeed less than that associated with the fundamental Gaussian beam. It follows that the phase velocity of the higher order modes is higher than that of the Gaussian beam, reflecting the more rapid divergence of these modes.

Later, in chapter 13 where we discuss lasers which make use of Fabry-Perot resonators for feedback and amplification of light, we will see that it is possible to have a laser operate in a single mode or in a superposition of modes. Generally the trade-offs are as follows: One tends to use single-mode, TEM_{00}, operation if a nice "clean" (relatively uniform in intensity) beam with a uniform phase front is of concern, such as for holography. On the other hand, because the higher order modes have a greater transverse spatial extent, there is a possibility of extracting more energy from a medium because of the larger cross-section of interaction. Of course the intensity variations across the beam may be horrendous and certainly a well-defined phase front won't be possible, but if it's raw energy you want this is often the path to choose.

Finally, it might be noted that the form of the mode expressions we have derived are independent of the value of z. This is perhaps obvious since it is easy to show that the expressions for the Hermite-Gaussian modes represent solutions to the paraxial wave equation. From the point of view of diffraction theory it might be thought that the form of the fields would be different in the near-field and the far-field. However, the same forms hold for both the Fresnel and Fraunhofer diffraction limits. The latter case, in particular, is just a reflection of the fact that the Hermite-Gaussian field distributions are Fourier transforms of each other.

10.5. Hermite-Gaussian Beams in Resonators—Allowed Frequencies

In the determination of the Gaussian beam parameters for a given resonator, we neglected to consider phase effects in the modes. As was mentioned earlier, the condition we require, in general, for a mode is that the round trip phase change of the optical field has to be a multiple of 2π, or the single-pass phase change has to be a multiple of π. In terms of the Hermite-Gaussian field parameters, and a resonator whose mirrors are located at $z = z_1$ and $z = z_2$ as before, we require

$$\int_{z_1}^{z_2} k_{eff}dz = k(z_2 - z_1) - (l+m+1)\left[\arctan(\frac{z_2}{z_0}) - \arctan(\frac{z_2}{z_0})\right] + \frac{\varphi_2+\varphi_1}{2} = q\pi$$

where q is an integer and φ_1 and φ_2 are the phase changes associated with reflectivity from the first and second mirrors. If the resonator is filled with a medium of refractive index n, we have

$$k = \frac{2\pi n}{\lambda_0} = \frac{\omega n}{c}.$$

It follows that the various allowed frequencies are determined by three indices *q,l,* and m and the allowed frequencies of the standing modes are

$$\omega_q^{l,m} = \frac{q\pi c}{nL} + \frac{c}{nL}\left[(l+m+1)\left[\arctan(\frac{z_2}{z_0}) - \arctan(\frac{z_2}{z_0})\right] + \frac{\varphi_2+\varphi_1}{2}\right]. \tag{10.5.1}$$

The indices l and m within a resonator context are referred to as the *transverse mode numbers* while the index q is referred to as the longitudinal or *axial mode number*. We have neglected the dispersion of the refractive index in writing down this expression. If it were strongly dependent on frequency, the determination of the mode frequencies would be much more difficult. We return to this problem later when we discuss mode frequencies of a laser. Equation 10.5.1 is cumbersome to use, and it would be more meaningful if we could rewrite it in terms of the resonator parameters g_1 and g_1. After considerable algebraic manipulation one can show that the required expression is

$$\omega_q^{l,m} = \left[q + \frac{(l+m+1)\arccos\left(\sqrt{g_1g_2}\right)}{\pi}\right]\frac{\pi c}{nL}$$

where we have ignored phase variations due to reflection at the mirrors. For near-planar resonators where the resonator parameters are both ≈ 1, we have

$$\arccos\left(\sqrt{g_1g_2}\right) = \alpha \ll \pi$$

and

$$\omega_q^{l,m} = \left[q + \frac{(l+m+1)\alpha}{\pi}\right]\frac{\pi c}{nL}$$

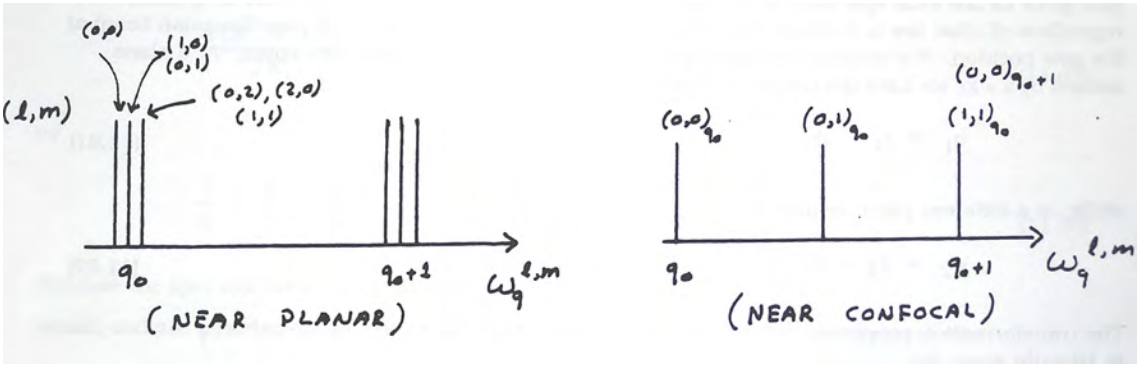


FIGURE 10.5.1. Mode spectra for near planar and near-confocal resonators.

while for the near-confocal type resonators where the resonator parameters are both ≈ 0, we have

$$\arccos\left(\sqrt{g_1 g_2}\right) \simeq \pi/2$$

and the corresponding mode frequencies are

$$\omega_q^{l,m} = \left[q + \frac{l+m+1}{2}\right]\frac{\pi c}{nL}.$$

The mode spectra for both cases are illustrated in Figure 10.5.1.

For the near-planar indices the transverse mode-numbers have little influence on the mode spectrum. As $g_{1,2} \to 1$ the modes for different *l,m* values become degenerate and the mode spectrum is determined entirely by the axial mode number with

$$\triangle\omega = \omega_{q+1} - \omega_q = \frac{c\pi}{nL} = constant.$$

Typically, for a one meter cavity and $n = 1$, the mode spacing is $10^9 s^{-1}$. The degeneracy of the transverse modes can easily be understood, since, for, near-planar mirrors all the modes resemble the same plane waves and indeed are the same plane wave for $g_{1,2} \equiv 1$.

For the near confocal resonator, the transverse mode numbers are virtually as important as the axial mode numbers in determining the spectrum. From the figure it can be seen that a change of the transverse mode numbers by a total of two units is equivalent to changing the axial mode number by 1. It is therefore seen that there is here a large amount of degeneracy in that many different combinations of *q,l* and *m* can lead to the same frequency of a mode.

10.6. Transformation of Gaussian Beams

Gaussian beams not only represent one of the most fundamental solutions of the paraxial equation but they also represent one of the most common beams encountered, particularly when dealing with lasers. We have learned in some detail the properties of Gaussian beams and how they propagate in free space or a homogeneous medium. What happens to our description of these beams when they pass into or through a different medium such as a lens? Do we have to start from scratch and re-solve the paraxial wave equation with appropriate boundary conditions? Of course we could do that, but for many common situations this is not necessary. It becomes easier to describe the transformation properties of Gaussian beams using matrix techniques.

To begin the discussion of the transformation properties recall that the parameter, $q = z - iz_0$, (known as the *Gaussian beam parameter*) completely specifies, apart from intensity, the Gaussian beam at position *z*. Indeed, we have that

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

so that the real part of *1/q* gives us the inverse radius of curvature of the beam while the imaginary part gives us the local spot size. If we can find how q transforms between two different points, regardless of what lies in between, we obviously can define the new Gaussian beam at the new position. For example, consider a Gaussian beam propagating in free space. At a plane defined by $z = z_1$ we have the Gaussian beam parameter

$$q_1 = z_1 - iz_0$$

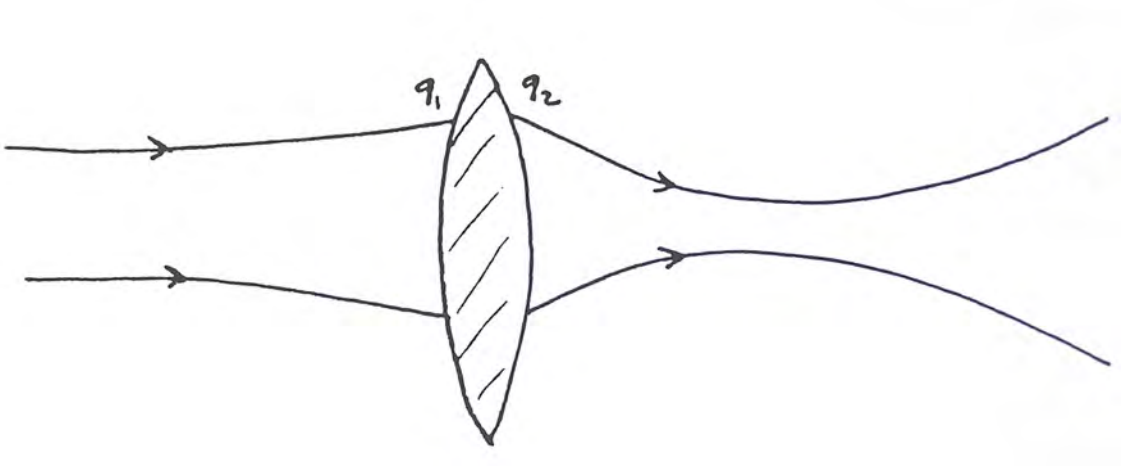


FIGURE 10.6.1. Transformation of a Gaussian beam by a lens.

while, at a different plane defined by $z = z_2$, we have

$$q_2 = z_2 - iz_0.$$

The transformation properties of a Gaussian beam propagating in free space between the two planes is trivially given by

$$q_2 = q_1 + (z_2 - z_1)$$

Let's consider now a Gaussian beam propagating through a thin lens of focal length f such that the beam has a Gaussian beam parameter q_1 immediately before the lens and a new Gaussian beam parameter q_2 immediately after the lens as shown in Figure 10.6.1. For a thin lens, the spot size of the Gaussian beam doesn't change so that

$$w_2(z) = w_1(z).$$

The lens, however, imposes a change on the phase front as we saw in chapter 9. The transmission function of a lens of focal length f is of the form

$$T(x,y) = \exp\left(-\frac{ik\left[x^2+y^2\right]}{2f}\right).$$

When we apply this to the Gaussian beam for fixed z, we have

$$T\Psi_{00}(x,y,z) = \Psi_{00}^{'}(x,y,z)$$

with the only difference between the two beams being the radius of curvature of the phase front. If the new radius of curvature is *R'*, then

$$\frac{ik(x^2+y^2)}{2R} - \frac{ik(x^2+y^2)}{2f} = \frac{ik(x^2+y^2)}{2R'}$$

or

$$\frac{1}{R} - \frac{1}{f} = \frac{1}{R'}.$$

Because the spot size does not change we have that

$$\frac{1}{q} - \frac{1}{f} = \frac{1}{q'}$$

so that the transformation of the Gaussian beam is given by

$$q' = \frac{q}{\left(\frac{-1}{f}\right)q + 1}.$$

Although it is beyond the scope of these notes, it turns out that the transformation of a Gaussian beam can be represented by an equation of the form

$$q' = \frac{Aq + B}{Cq + D}$$

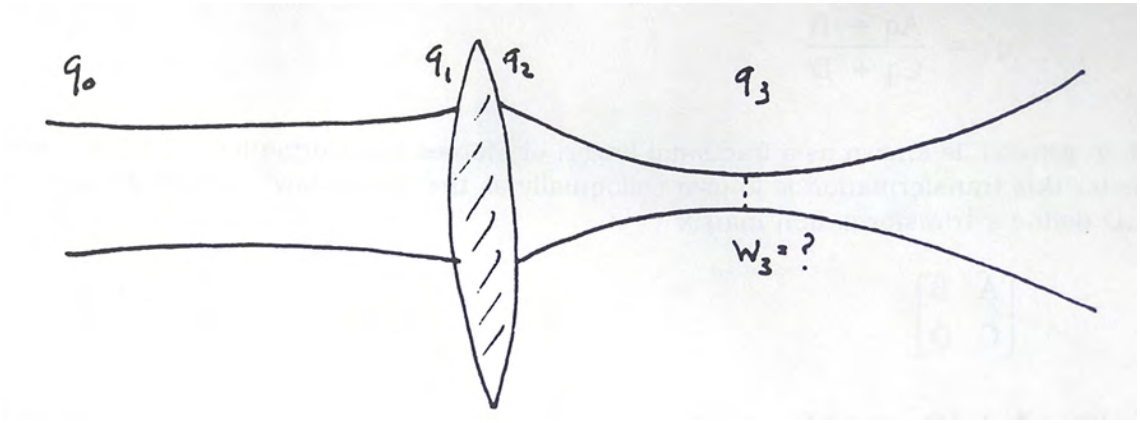


FIGURE 10.6.2. Focusing a Gaussian laser beam

which, in general, is known as a *fractional linear,* or *Möbius transformation.* This is known colloquially as the "ABCD law". The four parameters, A,B,C,D define a transformation matrix

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

which although, for Gaussian beams, it is never used like a matrix, is the same as the ABCD matrices we considered for rays in chapter 5!

For propagation in free space through a distance $z_2 - z_1$ we have seen that the transformation matrix is given by

$$\begin{bmatrix} 1 & z_2 - z_1 \\ 0 & 1 \end{bmatrix}$$

while for a lens of focal length f we have the transformation matrix

$$\begin{bmatrix} 1 & 0 \\ \frac{-1}{f} & 1 \end{bmatrix}$$

where, of course, f is positive for a converging lens and negative for a diverging lens. One of the benefits of using the Möbius transformation for the Gaussian beam parameter is that it becomes easy to treat a system by multiple, successive transformations, *e.g.*, by a lens, free-space propagation, an interface, other lenses, etc. Indeed, if a Gaussian beam is propagating through a series of N optical "elements" each of which has an associated transformation matrix $\overleftrightarrow{M}_i$then the overall transformation matrix of the system is easily shown to be

$$\overleftrightarrow{M}_S = \prod_{i=1}^{N} \overleftrightarrow{M}_i$$

where the order of the matrices, from right to left is the order in which the Gaussian beam would encounter the associated elements, *i.e.,*

$$\overleftrightarrow{M}_S = ... \overleftrightarrow{M}_{second} \overleftrightarrow{M}_{first}.$$

To illustrate the simplicity of the use of the transformation matrices let's consider the following example. Say we have a He-Ne laser producing a Gaussian beam with a divergence of 1 mR and with a beam waist at the output of the laser of 0.4 mm. What is the diffraction limited spot size we can achieve with a positive lens of focal length 2 cm, located 1m from the beam waist? The situation is depicted in Figure 10.6.2.

In considering the problem, we start with a beam parameter q_0 at the beam waist. This gets transformed into a parameter q_1 just before the lens, and a parameter q_2 just after the lens. Finally at the focal spot of the beam the parameter is q_3. If we can determine the imaginary part of q_3^{-1}, we will have the spot size at the focus. The overall transformation matrix of the system is the product of three matrices, namely those associated with propagation in free space through a distance of 1 m, propagation through the lens and propagation through a certain distance that bring us to the focal spot. This system matrix can then be used to relate q_0 to q_3 from which we could find w_3. For illustration purposes however, let's break the problem up into its elementary constituents to see what actually happens to the Gaussian beam.

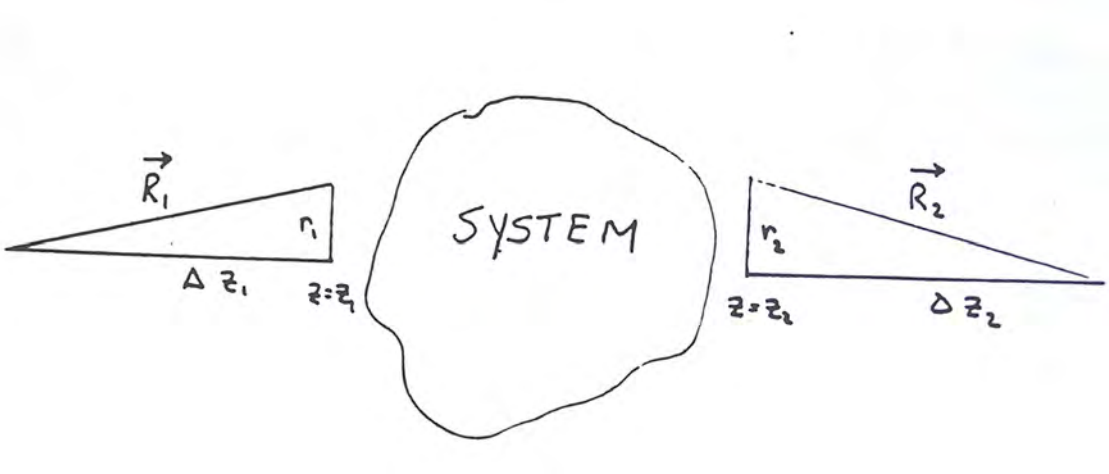


FIGURE 10.6.3. Rays entering and leaving an optical system.

To determine q_0 from the information given we recall that the divergence of a Gaussian beam is given by

$$\theta = \frac{2\lambda}{\pi w_0}$$

which, for the numbers given implies that $\lambda_0 = 0.63\,\mu m$. It follows that

$$q_0 = 0 - 0.8i \quad z_0 = 0.8m$$

and

$$q_1 = q_0 + 1$$

giving $R_1 = 1.64\,m$ and $w_1 = 0.64\,mm$. On passage through the lens we have

$$q_2^{-1} = q_1^{-1} - f^{-1}.$$

Note that the radius of curvature of the beam emerging from the lens is not 2 cm, so the beam does not focus exactly 2 cm behind the lens. Only an incident plane wave focuses at a distance f behind a lens of focal length f as we saw in the chapter on diffraction. To determine where the focal spot is in our case we note that

$$q_3 = q_2 + \ell = (-0.21 + \ell) + 2 \times 10^{-4} i$$

where ℓ is the distance to the focal point. Now the focal point is defined to be the position of the beam waist, which in turn is where the radius of curvature of the beam is infinite and the Gaussian beam parameter is purely imaginary. Hence $\ell = 2.1$ cm. We can then determine the beam waist from

$$q_3^{-1} = -i\frac{\lambda}{\pi w_3^2} = -5 \times 10^3 i$$

giving $w_3 = 6.3\,\mu m$ and also giving the depth of field, z_0¸(3) of the focused beam to be 200 µm.

It's a remarkable fact that the same set of matrices apply to rays and Gaussian beams. It also applies to paraxial portions of spherical waves ($z_0 \to 0$). This is remarkable for two reasons.

1) The transformation of Gaussian beams is governed by a fractional linear transformation while that of rays is governed by a true matrix transformation.

2) In dealing with rays one totally ignores the wave character of light while for Gaussian beams it is explicitly included.

We can remove some of the mystery of the similarity between the results if we rewrite the ray transformation law as

$$\frac{r_2}{r_2'} = \frac{A\left(\frac{r_1}{r_1'}\right) + B}{C\left(\frac{r_1}{r_1'}\right) + D}.$$

Referring to Figure 10.6.3, we can define a distance

$$\triangle z_1 = \left.\frac{r}{\left(\frac{dr}{dz}\right)}\right|_{z=z_1} = \frac{r_1}{r_1'}$$

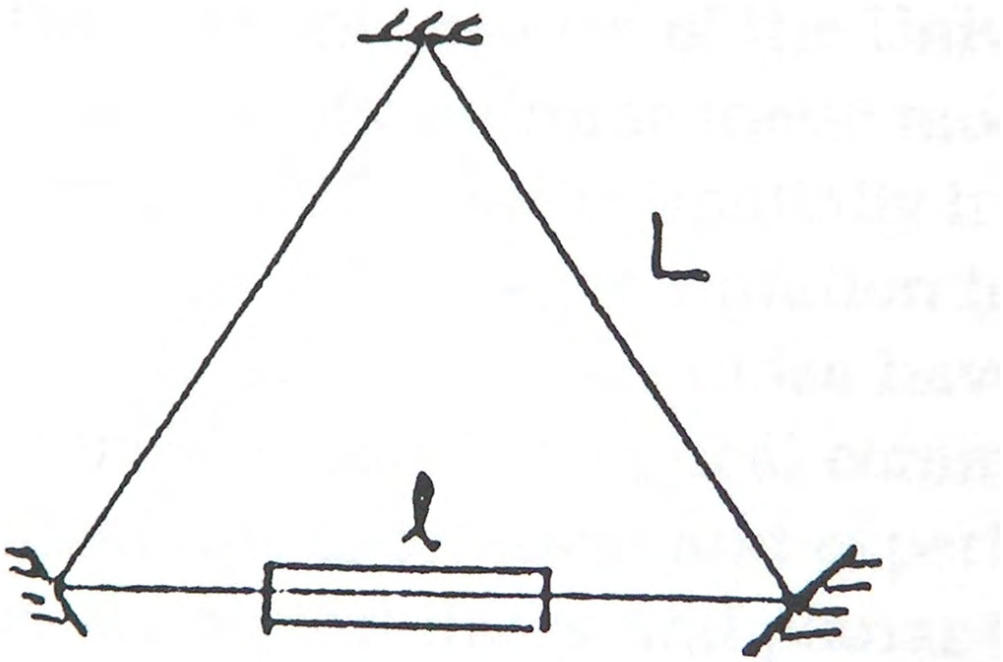


FIGURE 10.6.4. Ring laser

which is the distance between the reference plane and the intersection point of the ray with the z-axis. The intersection point represents the effective source point from which all rays with the ray parameters r_1 r_1' seem to be emanating. Such rays, of course, lie on a cone. Similarly the distance

$$\triangle z_2 = \frac{r_2}{r_2'}$$

is the effective source or, possibly convergence or focus point, associated with all rays with parameters r_2 and r_2'.

A Gaussian beam may be considered to be the paraxial limit of a solution to a wave equation for a point source with the source shifted by the imaginary amount iz_0. Without the origin shift, recall that the paraxial solution of the wave equation is a portion of a spherical wave emanating from $z = 0$. The distance $z - iz_0 = q$ measures the "complex distance" from the reference plane (location of the point source for spherical waves or beam waist for Gaussian waves) to the intersection point with the axis of the "complex ray" pertaining to the Gaussian mode. This gives some hint as to why q obeys the same transformation law as $r/r' = \Delta z$.

References

A. Yariv, *Introduction of Optical Electronics*, Holt, Reinhard, Winston, New York, 1976.

A.E. Siegman, *Introduction to Laser Physics*, Prentice Hall, New York, 1971.

H.A. Haus, *Waves and Fields in Optoelectronics*, Prentice Hall, New York, 1981.

Problems

1. Determine the approximate error made in associating u_{00} with the magnitude of the electric field for a Gaussian beam at different points on the beam. At what point on the beam is the error likely to be largest?

2. Obtain the mode confinement condition for an optical resonator formed by two identical mirrors with radius of curvature R, with separation L and with a thin lens of focal length f located at the center.

3. A ring laser consists of three mirrors at the vertices of an equilateral triangle of side L with an active medium of refractive index n and length l as shown. Such a laser can support clockwise and counterclockwise travelling wave modes.

a) Derive an expression for the wavelengths associated with the cavity modes of a stationary laser.

b) If the laser rotates in a clockwise direction at an angular velocity Ω about an arbitrary fixed axis, determine an expression for the beat frequency between pairs of travelling wave modes which are degenerate for the stationary laser. Assuming λ=0.5 µm, L=10 cm, n=1.5 andℓ = 5 cm, what is the beat frequency measured for such a laser which is located with its plane horizontal to the ground in Toronto? (If you wish not to get too bogged down in geometry, you may consider the path of the light beam to describe a circle rather than a triangle.)

4. A Gaussian beam with w_0 = 0.05 mm and λ_0 = 0.5 µm has its waist located 20 cm from a lens of focal length 2 cm. Behind the lens is a semi-infinite slab of glass with n= 1.5. Where does the beam come to a focus?

CHAPTER 11

Optical Waveguides

Optics is more than a pane in the glass
anon

11.1. Waveguides and Ray Optics

In the previous chapters we have mainly been concerned with the propagation and diffraction of plane waves. Plane waves are the fundamental solutions of the wave equation for homogeneous media of infinite extent. As such they are the "modes of the Universe". In this chapter and the next, we are concerned with the propagation of electromagnetic modes of finite transverse extent. These arise as solutions to the wave equation in the case of spatially inhomogeneous media and, in particular, are the fundamental solutions of the wave equation in media of finite transverse extent—otherwise known as optical waveguides. Optical waveguides have achieved tremendous popularity in recent years, mainly because of the implications for optical communications. There are two particular types of waveguides on which considerable theoretical and experimental research has been done over the past 40 years. These are known as *optical fibres* and *planar waveguides* with the former being, in essence, one dimensional waveguides and the latter two dimensional waveguides.

Optical fibres are simply long strands of optical glass and consists of a core which is surrounded by a cladding as shown in Figure 11.1.1

The refractive index of the core, n_2, is slightly higher than that of the claddingn_3 so that if light gets into the core at an appropriate angle of incidence into the fibre, the fibre is able to trap a light beam by total internal reflection. In terms of a simple ray picture the condition for this is that the angle θ has to exceed the critical angle, θ_c. The critical angle is related to the angle of incidence by

$$n_1 \sin\theta_1 = n_2 \sin\theta' = n_2 \sin\theta_c$$

where n_1 is the refractive index of the external medium. The critical angle, as we saw in an earlier chapter, is simply given by

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

The illustrated fibre is referred to as a *step-index fibre* since there is an abrupt discontinuity in the refractive index as a function of radial distance. It is also possible to get wave guiding action in a graded index fibre by allowing the refractive index to decrease gradually from the center outwards.

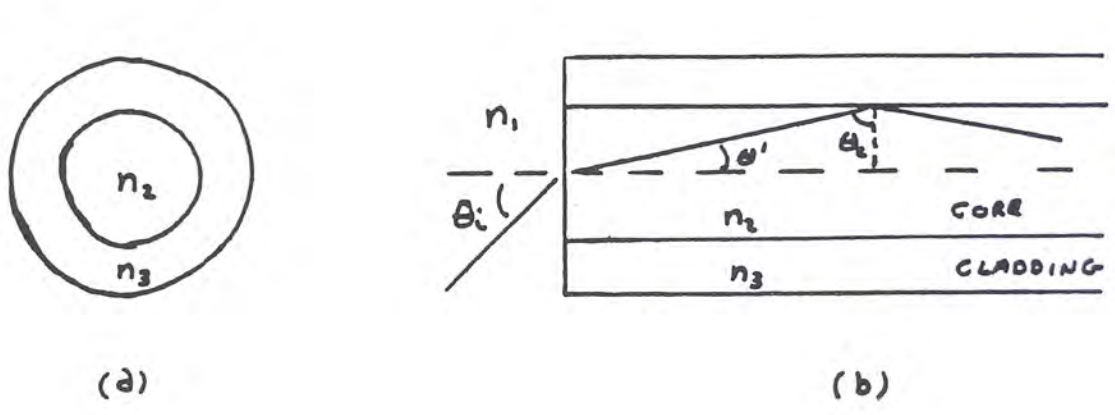


FIGURE 11.1.1. One-dimensional optical waveguide; a) cross section b) illustrating total internal reflection.

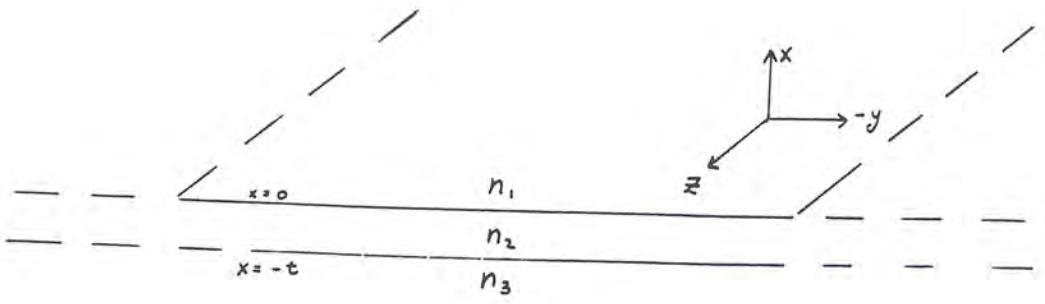


FIGURE 11.1.2. Planar optical waveguide

Light trapping and propagation through total internal reflection in a waveguide geometry has been known ever since Tyndall, in about 1850, demonstrated that a fine stream of water flowing from a hole in the bottom of a tank could "conduct" light. The field of waveguide optics, particularly in the fibre geometry has only taken off in recent years, when the technology of glass manufacturing had advanced to the state where it is possible to make glass so pure that at certain wavelengths light suffers a loss due to absorption or scattering of much less than 1% per kilometer! To appreciate the significance of this let me point out that if you were to make a slab 10 m thick out of ordinary window glass you wouldn't be able to see through it. The low loss afforded by the high purity glasses allow one to propagate information carrying optical beams over tens of kilometers before repeaters are required to boost the signals. The high frequency offered by light as a carrier also allows a tremendous number of independent channels. Indeed, if a typical voice channel has a bandwidth of $10^4 s^{-1}$, then with light's typical carrier frequency of $10^{15} s^{-1}$ one could permit as many as 10^{11} independent channels on the same optical beam. That is, one could carry all the telephone conversations of the world on a single fibre. Of course, one wouldn't want to do this for many reasons but nonetheless the high information carrying capacity of optical fibres has stimulated great interest in them and they have replaced many electronic forms of communication. The other advantages offered by fibre optics technology is that the fibres are typically small in cross section (less than 1 mm in diameter), are lightweight and flexible, are not subject to electromagnetic interference, and are not easily tapped.

In the planar, or slab waveguide geometry, illustrated in Figure 11.1.2, the waveguide consists of three layers of optical material.

The principle behind the operation of these waveguides is the same as for the fibre waveguides. However, for planar waveguides there is no restriction on the extent of the wave in two dimensions—what we have taken to be the $y - z$ plane which defines the plane of the waveguide. The interest in planar waveguides centers around the fact that they could serve as the basis for an integrated optics signal processing "chip" with two dimensional lenses, prisms, etc, manipulating beams. In practice, however, it is much more difficult to manufacture a slab wave-guide of sufficient quality to perform these functions. It is even difficult to produce a simple slab waveguide which has a loss of less than 10% per millimeter!

The physics or optics of the two types of waveguides are similar and the particular geometry only determines the details of the optical field distribution as a function of the co-ordinates involved. It turns out that the optical fibre waveguide is most easily discussed in terms of cylindrical co-ordinates. The solution of the wave equation then yields Bessel functions which provide little insight into the salient features of waveguide phenomena. Planar waveguides yield much simpler solutions for the wave equation and so we analyze these structures to understand some of the basic features associated with waveguide optics. In terms of the simple ray picture, in a planar waveguide the modes correspond to pairs of rays that have the same inclination to the z-axis; in a circular waveguide, these pairs of rays must be replaced by complicated cones of rays. The physical interpretation of the modes as standing waves is similar but much more difficult to visualize with circular symmetry. In addition there are standing waves in the azimuthal direction; consequently, two numbers are needed to describe a mode in a circular waveguide, just as two numbers would be needed in the planar guide if both transverse directions were finite. We now examine the form of the modes for the planar waveguides beyond the simple ray picture discussed so far.

11.2. Modes of Planar Dielectric Waveguides

Consider the geometry of the planar optical waveguide shown in Figure 11.1.2. The geometry indicated is the same as that of a single thin film coating. In that case we saw that in no situation, *i.e.,* for no angle of incidence

or wavelength, did we get trapping of the light in the thin film. On the other hand the arguments based on total internal reflection above indicate that under certain conditions light can be trapped in the waveguide. The problem is a matter of coupling. Certainly by what is known as "end fire coupling", where we bring the light in through the side, we can realize trapping. This is certainly not the only way this can be done but let us not be concerned with the coupling problem until we understand the types of modes which can exist for the dielectric slab.

To determine the fundamental modes of the waveguide we look for solutions of Maxwell's equation for this geometry of the form

$$\vec{\mathcal{E}}(\vec{r},t) = \vec{E}(\vec{r})e^{-i(\omega t-\phi(\vec{r}))}.$$

If, as usual, we define the vacuum propagation constant to be $k = \omega/c$, then the $\vec{E}$ field must satisfy the Helmholtz equation (which assumes $\vec{\nabla}\cdot\vec{D} = 0$)

$$\nabla^2\vec{E}(\vec{r}) + k^2n^2(\vec{r})\vec{E}(\vec{r}) = 0$$

where $n(\vec{r})$ gives the spatially dependent refractive index. A similar differential equation exists for the magnetic field $\vec{H}$. The differential equation one uses is primarily determined by which boundary conditions, those for $\vec{E}$ or $\vec{H}$ are easier to implement in the system under consideration. This is to be understood in what follows. For the moment and to avoid cumbersome notation we continue to write the differential equation in terms of $\vec{E}$ only. For simplicity we also assume that the waveguide is a lossless dielectric so that $n(\vec{r})$ can be considered to be real. Without loss of generality we can restrict ourselves to constant amplitude waves with constant phase fronts normal to the z-direction so that we can choose

$$\phi(\vec{r}) = \beta z$$

for some constant β. This describes the oscillation frequency of the field in the z-direction only i*n all three media*. It might be thought that the assumption of a single β for all three media is restrictive. This is not so. Indeed, it is a necessary requirement if we have perfectly planar waveguides, since the system has translational symmetry in the z-direction and so the z-dependence of the field cannot change in traversing the boundaries. We saw this before, albeit, in a much simpler situation when we discussed refraction and reflection at a single interface. Because of symmetry and the assumption of wave fronts normal to the z-direction the amplitude and phase of the wave can be considered to be independent of the y-co-ordinate. It is worth emphasizing that this implies field oscillations in the z, but not the y-direction; what happens in the x-direction remains to be determined. The x-dependence of the electric field can be determined from the Helmholtz equation which under the stated assumptions now takes the form

$$\frac{\partial^2\vec{E}(x)}{\partial x^2} + \left[k^2n^2(\vec{r}) - \beta^2\right]\vec{E}(x) = 0.$$

There are two fundamental types of solutions which can occur for this equation and which are consistent with Maxwell's equations; those with the $\vec{H}$ field in the plane of the waveguide (known as TM modes with $\vec{H} = H\hat{y}$) and those with the $\vec{E}$ field in the plane of the waveguide and known as TE modes. For the TE modes the magnetic field has components in the z and x-directions as does the electric field in case of TM modes. Because there is no variation of the fields in the y-direction and because we can use continuity of the $\vec{E}$ field across the boundaries for TE waves and continuity of the $\vec{H}$ field for TM waves, the differential equation to be solved is that of $\vec{E}$ for TE waves and that of $\vec{H}$ for TM waves. We return to the difference between the two types of modes later when we consider the details of the field distributions. For the moment, since the differential equation has the same form in either case let us examine only the form the solutions will take in either case and concern ourselves with the magnitudes of the fields later when we apply the boundary conditions.

To be concrete, let us choose to look for the electric fields associated with TE modes. With the direction of the field direction fixed, for each of the three sections of the waveguide we can reduce the Helmholtz equation for TE modes to a scalar differential equation given by

$$\frac{\partial^2\vec{E}(x)}{\partial x^2} + \left[k^2n_i^2 - \beta^2\right]\vec{E}(x) = 0$$

where n_i is the refractive index of one of the three regions of the waveguide. This equation can be rewritten in the form

$$-\frac{\partial^2\vec{E}(x)}{\partial x^2} - k^2n_i^2\vec{E}(x) = -\beta^2\vec{E}(x) \tag{11.2.1}$$

where we have suppressed the x-dependence of the field. Notice that this single-variable differential equation is exactly of the same form as the Schrödinger equation for a particle in a one-dimensional well. We can interpret the second derivative with respect to x as the kinetic energy operator, the quantity$k^2n_i^2$ as the potential energy

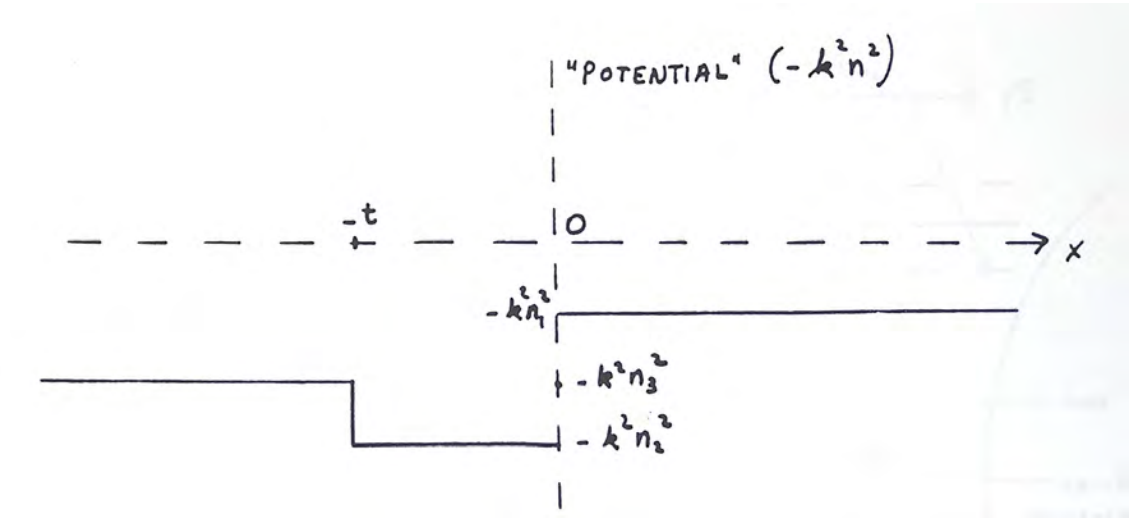


FIGURE 11.2.1. Refractive index variation for a typical planar waveguide geometry.

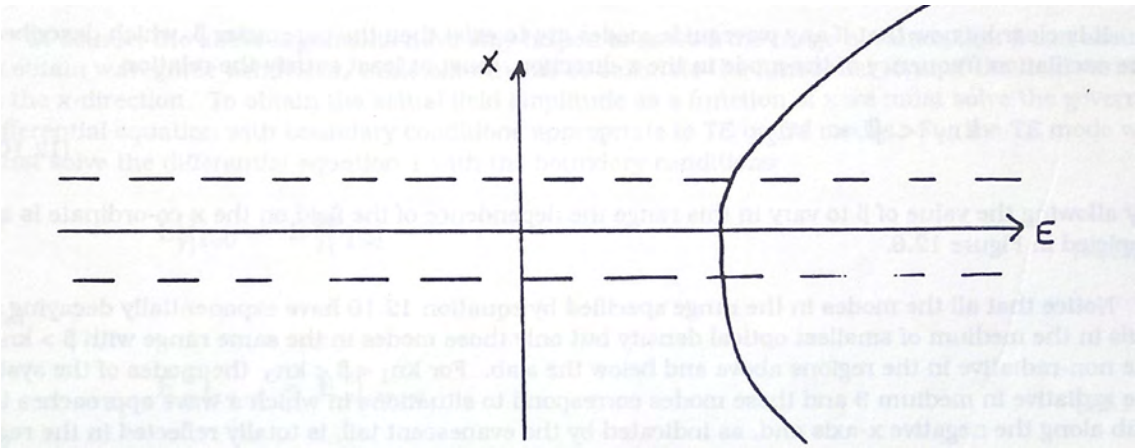


FIGURE 11.2.2. Field distribution as a function of x for $\beta > kn_2$.

operator and the quantity $-\beta^2$ as the total energy. (Once again, this is just another illustration of the fact that a wave-equation is a wave-equation is ....). Let us assume that the refractive indices are related by

$$n_2 > n_3 > n_1$$

so that the "potential energy" as a function of z is as illustrated in Figure 11.2.1.

In many cases medium 1 is simply air so that $n_1 = 1$. The form of *E(x)* can then simply be obtained by solving the second order differential equation and matching the field and its first derivative at the boundaries. Let us examine what the field distributions looks like as a function of the one free parameter we have, namely β.

If $\beta > kn_2$ (the largest allowed value of kn_i) it is easily seen that *E(x)* increases exponentially with distance $|x|$ along both the negative and positive x directions as illustrated in Figure 11.2.2.

Note that the field strength of such a mode goes to infinity as $x \to \infty$ and the mode cannot be normalized. This is obviously not a physically allowed solution. In the case of the corresponding quantum mechanical problem of a particle in a one-dimensional square well the problem is equivalent to finding eigenstates of the particle with total energy less than the lowest potential energy which implies that the kinetic energy would have to be negative everywhere — a non-realizable situation.

Similarly if $\beta < kn_1$ (the smallest allowed value of kn_i) one obtains oscillatory solution for *E(x)* everywhere as illustrated in Figure 11.2.3.

This implies that there is always a real component of the propagation vector away from the plane of the wave guide and hence energy propagates or "radiate" away from the plane of the waveguide. This type of mode could, for example correspond to a mode which is incident on the plane of the wave guide at a small angle of incidence (close

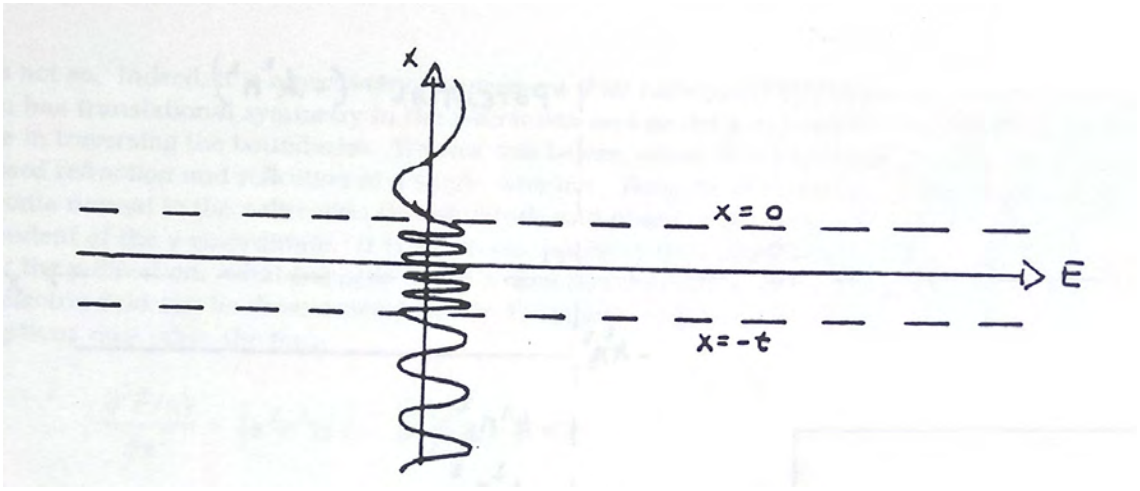


FIGURE 11.2.3. Field distribution as a function of z for $\beta <$ kn1

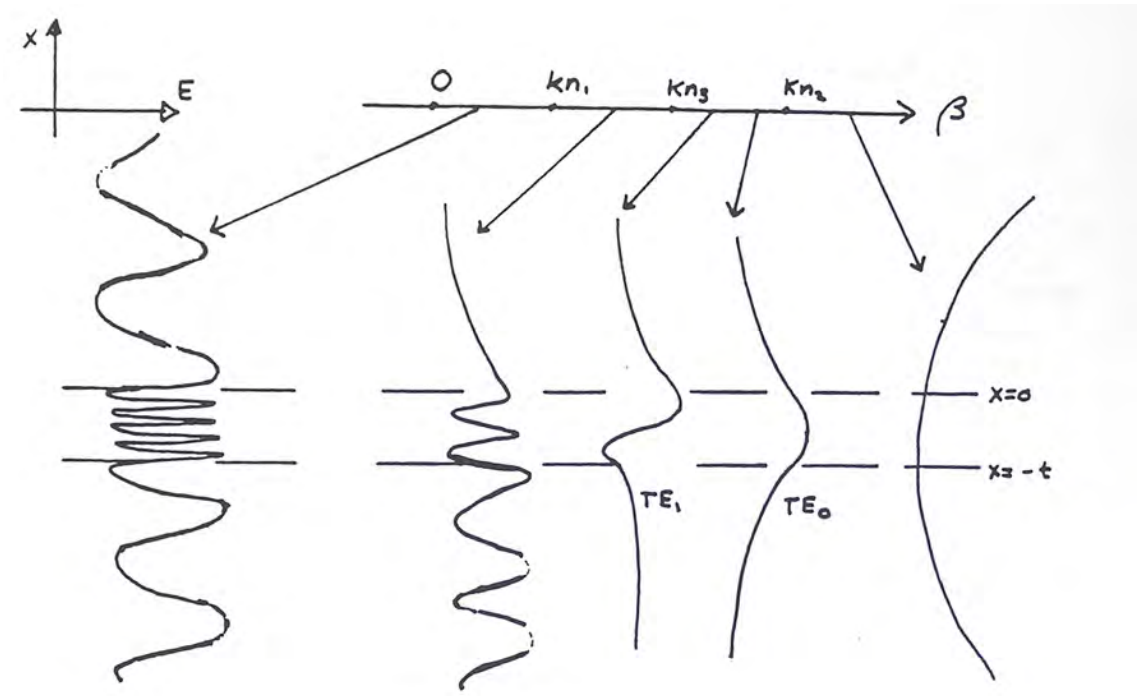


FIGURE 11.2.4. Mode spectrum and *E(x)* variations for a slab waveguide.

to normal incidence) from either medium 1 or 3. This situation is a special case of what we were concerned with in the thin-film reflection and transmission problem. The energy is not confined to the wave guide and the mode is known as a radiative mode. In the one-dimensional quantum mechanical analog this case corresponds to a situation where the total energy is never negative and so one doesn't have a particle which is bound to the potential well.

It is clear by now that if any waveguide modes are to exist then the parameter β, which describes the oscillation frequency of the mode in the x-direction, must at least satisfy the relation

$$kn_1 < \beta < kn_2. \tag{11.2.2}$$

By allowing the value of β to vary in this range the dependence of the field on the x-co-ordinate is as depicted in Figure 11.2.4.

Notice that all the modes in the range specified by equation 11.2.2 have exponentially decaying tails in the medium of smallest optical density but only those modes in the same range with $\beta > kn_3$ are non-radiative in the

regions above and below the slab. For $kn_1 < \beta < kn_3$ the modes of the system are radiative in medium 3 and these modes correspond to situations in which a wave approaches the slab along the negative x-axis and, as indicated by the evanescent tail, is totally reflected in the region of the slab. Since β is the component of the propagation constant in the z-direction, the condition is equivalent to assuming that the angle of incidence of the wave on the slab is within a certain range. For no value of β can one obtain total reflection of a wave approaching from the $+x$ direction in the sense that one obtains oscillatory behavior in region 1 and exponential decay of the field amplitude in region 3. In summary the condition required to obtain a mode which is confined to the slab is given by

$$kn_3 < \beta < kn_2.$$

In this range the field structure decays evanescently in both the negative and positive x-direction in media 3 and 1 respectively and the field oscillates in the z-direction. The energy is confined to the slab while propagating in the z-direction only.

Of course, the above arguments have only helped to narrow the range of values that β can assume to obtain waveguide behaviour, while allowing us to calculate the functional form of the field variation in the x-direction. To obtain the actual field amplitude as a function of x we must solve the governing differential equation with boundary conditions appropriate to TE or TM modes. For the TE mode we must solve the differential equation 11.2.1with the boundary conditions

$$E_y^{(1)}\Big|_{x=0} = E_y^{(2)}\Big|_{x=0}$$

and

$$E_y^{(2)}\Big|_{x=-d} = E_y^{(3)}\Big|_{x=-d}$$

where 1, 2, 3 refer to the different media. Because of Faraday's law we also have that

$$\frac{\partial E_y}{\partial x} = i\omega\mu_0 H_z.$$

But since tangential components of the $\vec{H}$ field are continuous we require that $\partial E/\partial z$ is continuous across the two boundaries. For the three media the most general solution of the differential equation given in equation 1 is therefore given by

$$E_y(x,y,z) = E_0(x)e^{-i(\omega t-\beta z)}$$

with

$$\begin{aligned} E_0(x) &= Ce^{-\alpha_1 z} \qquad 0 < z < \infty \\ &= C\left[\cos(hx) - \frac{\alpha_1}{h}\sin(hx)\right] \quad -t < x < 0 \\ &= C\left[\cos(ht) + \frac{\alpha_1}{h}\sin(ht)\right] e^{\alpha_3(x+t)} \quad -\infty < x < -t \end{aligned}$$

where

$$h = \sqrt{n_2^2k^2 - \beta^2}$$

$$\alpha_1 = \sqrt{\beta^2 - n_1^2k^2}$$

$$\alpha_3 = \sqrt{\beta^2 - n_3^2k^2}$$

.The constant C is determined by the total energy or intensity of the wave and the H field can be determined by Faraday's law which for this situation is equivalent to

$$H_x = -\frac{i}{\omega\mu_0}\frac{\partial E_y}{\partial z}$$

$$H_z = \frac{i}{\omega\mu_0}\frac{\partial E_y}{\partial x}.$$

Now since $\partial E_y/\partial x$ is continuous at $x = -t$ we have that

$$h\sin(ht) - \alpha_1\cos(ht) = \alpha_3\left[\cos(ht) + \frac{\alpha_1}{h}\sin(ht)\right]$$

which gives as a restriction on the values of β that are allowed

$$\tan(ht) = \frac{\alpha_1 + \alpha_3}{h\left(1 - \frac{\alpha_1\alpha_3}{h^2}\right)}. \tag{11.2.3}$$

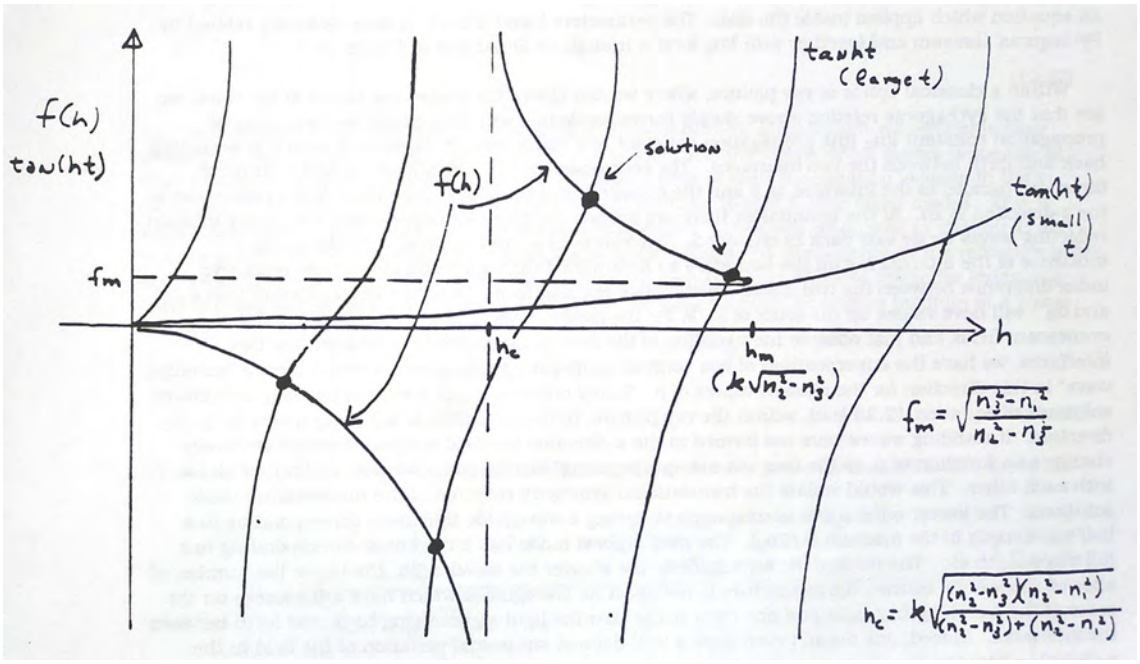


FIGURE 11.2.5. Graphical solution of equations 11.2.3 and 11.2.4.

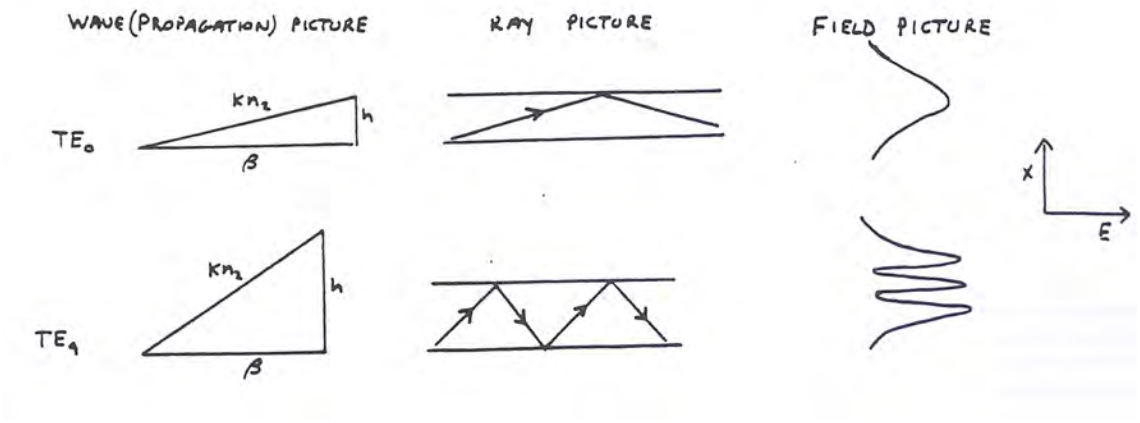


FIGURE 11.2.6. An illustration of the different components of the propagation vector.

This equation only yields a discrete number of modes or solutions (if it yields any) for β once t, n_1,n_2, n_3 and k are given. The set of modes is ordered by the finite number of values of β with the lowest order modes corresponding to the highest value of β. The determining equation is analogous to the equation one considers for the energy eigenvalues of a particle in a one-dimensional potential well of finite depth in quantum mechanics. The graphical solution of the determining equation for β is shown in Figure 11.2.5.

As can be seen, as k or the thickness of the film, t, decreases the number of modes decreases (as the depth of a quantum well increases the number of bound modes increases).

Before discussing the TE and TM modes in more detail it may be helpful to point out a simple interpretation of the parameters β, h, a_1 and α_2 defined above. First of all, note that we have

$$\beta^2 + h^2 = k^2 n_2^2$$

an equation which applied inside the slab. The parameters β and h are therefore obviously related by Pythagoras' theorem and together with kn_2 form a triangle as illustrated in Figure 11.2.6.

Within a classical optics or ray picture, where we can ignore the transverse extent of the wave, we see that the Pythagoras relation above simply corresponds to a situation where we have a ray of propagation constant kn_2 (the

propagation constant of a plane wave in medium 2) which is bouncing back and forth between the two interfaces. The component of the propagation constant in the z-direction, parallel to the interface, is β and the component in the x-direction which is perpendicular to the z-direction is $\pm h$. At the boundaries there are evanescent fields associated with the totally internal reflecting waves as we saw back in chapter 1. The values of α_1 and $\alpha\_$ depend on the angle of incidence of the internal ray on the boundary as determined through β and also on the refractive index difference between the two media which define the interface. As before we can expect that α_1^{-1} and α_3^{-1} have values on the order of λ. If, for the moment, we ignore the existence of the evanescent fields and just observe the variation of the field in the x-direction between the two interfaces, we have the superposition of two counter-propagating disturbances which form a "standing wave" in this direction for the allowed values of β. Stated differently, the values of β which are allowed solutions of equation 11.2.3 lead, within the ray picture, to the formation of standing waves in the x-direction. If standing waves were not formed in the x-direction the field amplitude would obviously change as a function of z, as the two "counter-propagating" disturbances went in and out of phase with each other. This would violate the translational symmetry required of the fundamental mode solutions. The lowest order mode corresponds to having a waveguide thickness corresponding to a half wavelength in the medium ($\lambda/2n_2$). The next highest mode has a thickness corresponding to a full wavelength etc. The thicker the waveguide or the shorter the wavelength, the larger the number of allowed modes. Of course, the ray picture is not exact for waveguides which have a thickness on the order of the wavelength of light and one can't really view the light as bouncing back and forth between the interfaces. Indeed, one doesn't even have a well-defined sinusoidal variation of the field in the x-direction between the two interfaces. Nonetheless, the above discussion allows us to see how the classical, macroscopic limit can be viewed.

Returning to the exact formulation, we can do for the TM modes what we have achieved for the TE modes. The equation for TM modes corresponding to equation 11.2.3 for TE modes is found to be

$$\tan(ht) = \frac{\alpha_1' + \alpha_3'}{h\left(1 - \frac{\alpha_1'\alpha_3'}{h^2}\right)} \tag{11.2.4}$$

where

$$\alpha_1' = \frac{n_2^2}{n_1^2}\alpha_1 \qquad \alpha_3' = \frac{n_2^2}{n_1^2}\alpha_3.$$

This equation can be solved graphically in a manner similar to that of equation . The solution for β obtained from it only differs slightly from those of the TE modes. In general only small differences between the different refractive indices are required to obtain waveguide modes. This is particularly true if one is only interested in single-mode waveguides, where within the naive ray picture the energy propagates internally at near grazing incidence with respect to the interfaces. Indeed in a typical single mode waveguide, with a thickness of the order of 1 µm, the three refractive indices would have values like: $n_1 = 1.$, $n_2 = 1.5$ and $n_3 = 1.45$.

Within the waveguide the time average Poynting vector only has non-zero components in the z-direction. It is therefore almost incidental what the wave is doing in the x-direction apart from the fact that this direction determines the number of modes and the value of β. The phase velocity of the mode in the direction of propagation is simply given by

$$v_\phi = \frac{\omega}{\beta}$$

and because β decreases for higher order modes, the modes have considerable dispersion in their velocity of propagation. Within the ray picture, it is seen that the lowest order modes propagate at near grazing incidence and hence have a larger component of their propagation velocity in the z-direction. Their phase velocity is therefore much lower than that of higher order modes. This is known as *modal dispersion* as opposed to colour dispersion, which was discussed in chapter 1, and was associated with the variation of the phase velocity through the dependence of the refractive index on wavelength. By analogy with colour dispersion we can define an effective refractive index, n_{eff} such that the particular mode can be viewed as propagating through an infinite medium at a phase velocity

$$v_\phi = \frac{c}{n_{eff}} = \frac{2\pi c}{\lambda\beta}.$$

Figure 11.2.7 illustrates the typical mode dispersion of a slab wave guide for both TE and TM modes.

The graph shows the normalized phase velocity or n_{eff}^{-1} plotted as a function of t/λ, and thickness, for $\lambda = 0.63$ µm.

Several features are to be noted. First of all, below a certain ratio of t/λ the waveguide can support no modes. Secondly, as $t/$ increases the first type of mode to be supported is a TEo, or lowest order transverse electric mode, with the next being a TM_0 or lowest order transverse magnetic mode. Each new mode which becomes active as t/λ

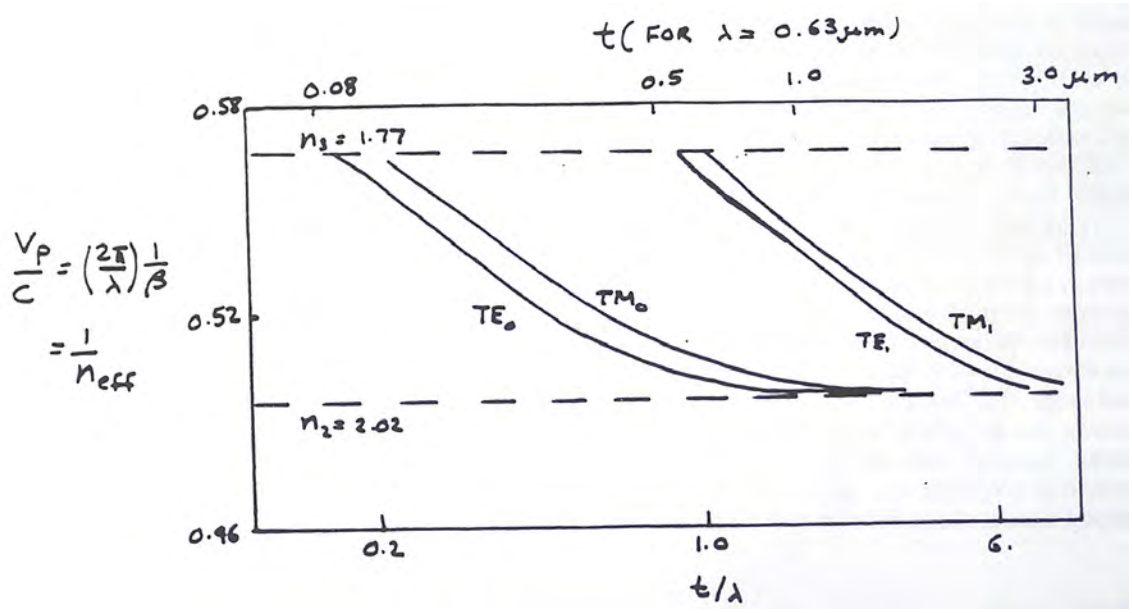


FIGURE 11.2.7. Mode dispersion of TE and TM modes in a slab waveguide.

increases does so abruptly at well defined cut-off values of this ratio. As the ratio t/λ increases, for a large value of the ratio, the number of modes supported increases. Third, for a given mode, the phase velocity decreases between a value appropriate to n_3 (that of the substrate) and the value n_2 (that of the sandwich layer). This can be understood with reference to, say the TE_0 mode. For the cutoff value of the ratio t/λ for this mode, nearly all of the energy of the mode is in the evanescent tail which, just at the cut-off value extends to x= -∞. Indeed, if the ratio were slightly less than the cutoff value, one would have a radiative mode in the substrate which would propagate with a phase velocity defined by n_3. As the ratio increases, more of the energy of the mode is propagating in the sandwich layer, which has a refractive index higher than n_2. As a result the overall phase velocity of the mode drops. As the ratio continues to increase the amount of energy in the evanescent tail decreases to the point where it becomes negligible, and the ray picture of mode behavior becomes more applicable. In the ray picture, the lowest order mode for thicker and thicker waveguides travel virtually parallel to the plane of the waveguide. Such modes therefore would travel with a phase velocity appropriate to the sandwich layer, or n_2. It follows that the phase velocity of modes asymptotically approach that of the sandwich layer as $t/\lambda \to \infty$. Finally it can be seen that for a fixed value of t/λ, different modes travel at significantly different phase velocities with the higher order modes travelling with a larger phase velocity than the lower order modes causing modal dispersion.

11.3. Modal Dispersion

The existence of modal dispersion in multimode fibers is one of their most serious drawbacks with respect to applications in optical communications. The name of the game in optical communications is information carrying capacity or information transfer rate. For most applications it has been found most practical to code information in the form of short optical pulses. The information carrying capacity is then directly determined by how densely one can pack the pulses. This in turn is determined by how short one can make the pulses and how far such pulses can propagate before they start overlapping. Two factors control the amount of overlap the pulses experience for a given waveguide length. First, with all other things being equal, if the pulses are very short they have a large frequency or wavelength bandwidth and normal colour dispersion forces pulse spreading since the different frequency components travel at different velocities. Second, if the optical pulse is launched into a superposition of waveguide modes, modal dispersion leads to pulse spreading as well. There isn't very much one can do about colour dispersion, apart from operating in wavelength regions where the colour dispersion is minimal. For typical silica based glasses two regions have been identified, one near 1.3 μm and the other near 1.55 μm. In most cases, modal dispersion has been determined to represent the most severe limitation to information carrying capacity. There have been various clever schemes proposed to tailor the refractive index profile, $n(\vec{r})$, of waveguides to minimize modal dispersion such as by choosing parabolic refractive index profiles with $n(x) = n_0(1 - \zeta x^2)$ for a limited range of x. Making use of the fact that the refractive index profile is a free parameter, modal dispersion can be reduced in "graded

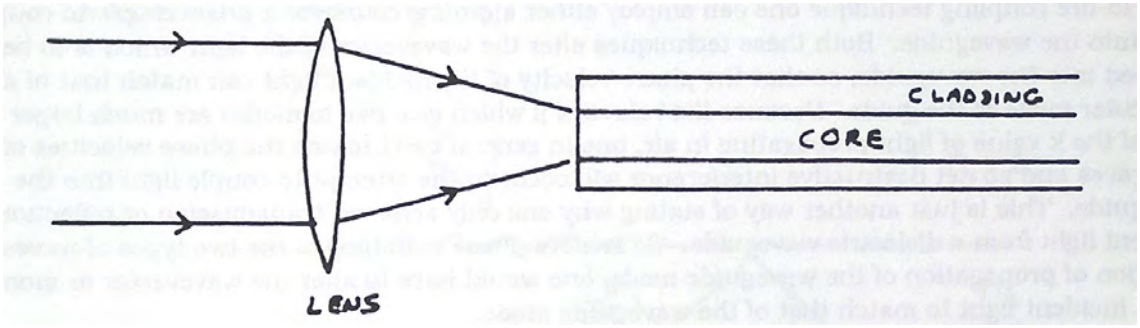


FIGURE 11.4.1. Coupling of light into fibre modes

index" waveguides over that afforded by step index waveguides. It turns out, however, with advances in optical fibre production, that the easiest way to eliminate modal dispersion is to operate in a single mode. These days virtually all optical communications over any significant distance is performed with single mode fibre waveguides.

11.4. Optical Fibres and Waveguide Coupling

Nearly all of the comments made in the previous section about planar waveguides apply to optical fibres. Only the details of the field distributions, cut-off wavelengths for single mode behavior, *etc*, depend on geometry. An additional unfortunate complication is that most of the rays that make up a mode of a fibre never pass through the axis of the fiber. Such rays are known as skew rays to distinguish them from the meridional rays which do pass through the fibre axis. Skew rays spiral around the axis of the waveguide without intersecting it. Skew rays (or more strictly speaking, modes) may be bound to the fibre, but some are only loosely bound, in the sense that they have long evanescent tails, and contribute to the loss of energy from the fibre. They are sometimes referred to as leaky modes. To couple light into a fibre there is only one technique which one can employ—end fire coupling— since the core of the fibre is uniformly surrounded by a cladding. The situation is depicted in Figure 11.4.1.

One focuses the light onto the end of a fibre in the vicinity of the core. This is often non-trivial, since, for example, in the case of single mode fibres, the diameter of the core is usually not much more than 5 μm. Within the ray picture description, and with reference to the figure, we see that we can couple light in the fibre provided the angle of incidence of the incoming rays can satisfy the relation stated at the beginning of this chapter, namely,

$$n_1 \sin\theta_1 < n_2 \cos\theta_c = \sqrt{n_2^2 - n_3^2} \simeq \sqrt{2n_2(n_2 - n_3)}.$$

The quantity $n_1 \sin\theta_m$, where ϑ_m is the maximum angle of incidence that can satisfy this relation is known as the *numerical aperture* or NA of the fibre. If we adopted a mode, as opposed to a ray picture, we would obtain essentially the same relation even for single mode waveguides. For an optical fibre the numerical aperture defines a cone as well as the minimum focal length of the lens one can use to perform the coupling. If one focuses too tightly the coupling efficiency actually decreases. If the fibre is in air then the numerical aperture is simply a measure of the cone angle. For a slab waveguide, one can attach a similar meaning to the numerical aperture, albeit in one dimension. As an example, an optical waveguide for which $n_2 = 1.5$ and $n_3 - n_2 = 0.01$ has a numerical aperture of 0.17, which, for $n_1 = 1.0$ corresponds to a cone angle of 10°. The smaller the cone angle employed the smaller the number of modes which are excited in the waveguide.

Equation 11.2.3 and 11.2.4 can be combined to determine the condition for obtaining single mode behavior. For a slab waveguide this relation is

$$\frac{2t}{\lambda_0} NA < 1.$$

For an optical fibre the analogous relation is

$$\frac{2\pi r}{\lambda_0} NA < 2.4$$

where r is the radius of the fibre core.

Similarly, it can be shown that in the ray, or geometrical optics limit, the total number of modes in a slab waveguide is

$$N = \frac{2t}{\lambda_0 \sqrt{n_2^2 - n_3^2}}$$

with a similar relation for an optical fibre with t replaced with $2r$.

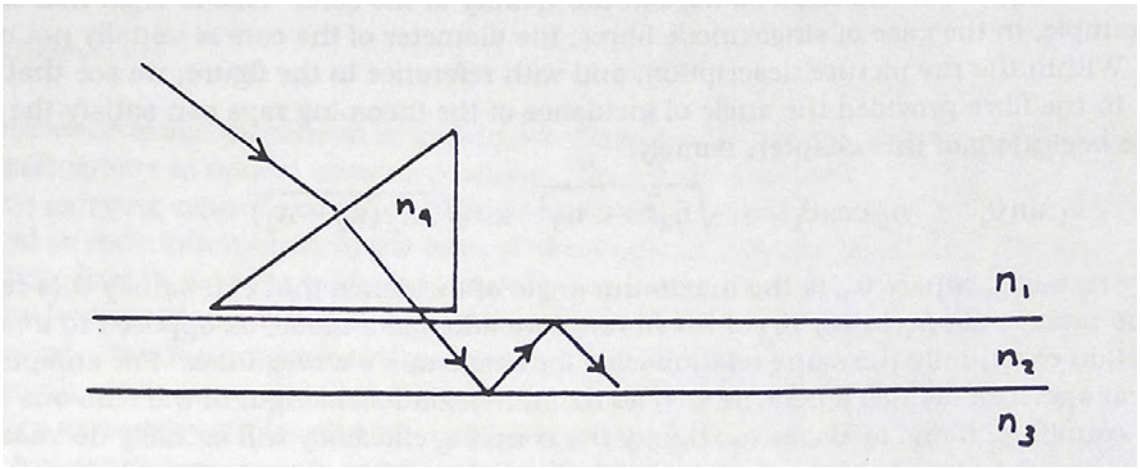


FIGURE 11.4.2. Prism coupler for coupling into a slab waveguide.

The problem of coupling light into a slab waveguide is usually easier to solve for the planar waveguide than it is for the fibre guide particularly if the medium corresponding to n1 is air. Besides the end-fire coupling technique one can employ either a grating coupler or a prism coupler to couple light into the waveguide. Both these techniques alter the wave vector of the light which is to be coupled into the waveguide, so that the phase velocity of the incident light can match that of a particular mode of the guide. Because the values of β which give rise to modes are much larger than that of the k value of light propagating in air, one in general can't match the phase velocities of the two waves and so net destructive interference occurs in the attempt to couple light into the waveguide. This is just another way of stating why one only achieves transmission or reflection of incident light from a dielectric waveguide. To achieve phase matching of the two types of waves in the direction of propagation of the waveguide mode, one would have to alter the wave vector or momentum of the incident light to match that of the waveguide mode.

The prism coupler is illustrated in Figure 11.4.2.

In this arrangement one uses a high index prism of refractive index n_4 placed within the region of the evanescent tail of the waveguide mode one wishes to excite. Usually this requires placing the prism within λ of the surface, which doesn't sound too difficult until one remembers that both the waveguide and the prism bottom surface have to be flat to within this specification over a large area. When the incidence angle is chosen so as to exceed the critical angle for the prism, light from the evanescent tail on the bottom of the prism can couple into the evanescent tail of the particular waveguide mode and a small amount of energy can be transferred, thereby allowing frustrated total internal reflection. Of course, this coupling works both ways so that light from the mode can also be coupled out of the waveguide mode in the same fashion. For coupling into the waveguide, one would choose the spot of total internal reflection near the edge of the prism as shown so that little light can be transferred back to the prism. Coupling efficiency associated with this process is known to reach 80%.

Light may also be coupled in and out of a planar waveguide by using a grating coupler in which a grating, usually a phase grating is etched onto the surface of the waveguide. This is illustrated in Figure 11.4.3.

For the angle of incidence, ν_1 and refraction, ν_1, shown, the grating equation tells us that

$$p\lambda_0 = \Lambda \sin \nu_1 - n_2 \Lambda \sin \nu_2$$

where Λ is the grating spacing and p is an index. The factor n_2 appears in the right hand side because the wavelength of light in the second medium is λ/n_2. If we multiply this equation by _ and rearrange some of the terms we have

$$k \sin \nu_1 = k n_2 \sin \nu_2 + p \frac{2\pi\lambda_0}{\Lambda}.$$

But $kn_2 \sin \nu_2$ is simply the z-component of the wave vector of light in the second medium. If we can choose, Λ, p, or ν_1 so that

$$kn_2 \sin \nu_2 = \beta_m$$

we would be able to phase match into a waveguide mode. As with the prism coupler, the grating coupler can couple light out of as well as into the waveguide so the length of the grating has to be chosen appropriately.

References

M. Young, *Optics and Lasers*, Springer Verlag, New York, 1981.

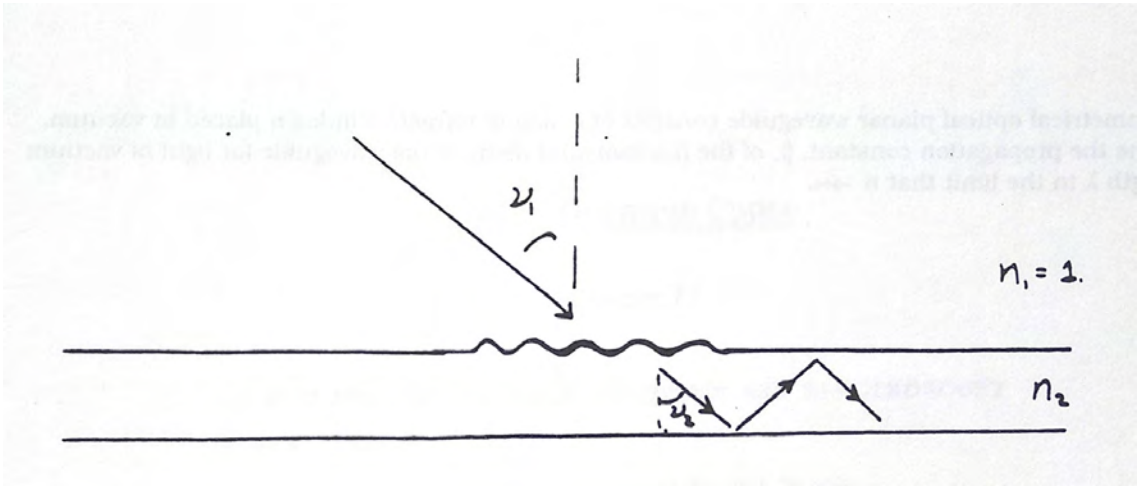


FIGURE 11.4.3. Illustration of the grating coupler geometry.

A. Yariv, *Quantum Electronics*, John Wiley, New York, 1972.

Problems

1. Determine the number of modes which can propagate in a slab waveguide for which n_1=1., n_2=1.5 and n_3 =1.4 if t = 4μm and λ =0.5 μm.

2. Show that if $\mathcal{E}_m(x)$ and $\mathcal{E}_n(x)$ are the electric field distributions associated with two different modes in an optical waveguide, the modes are orthogonal in the sense that

$$\int_{-\infty}^{\infty} \mathcal{E}_m^*(x)\mathcal{E}_n(x)dx = 0$$

3. A symmetrical optical planar waveguide consists of a slab of refractive index n placed in vacuum. Determine the propagation constant, β, of the fundamental mode of the waveguide for light of vacuum wavelength λ in the limit that $n \to \infty$.

Part 4

Quantum Optics

CHAPTER 12

Introduction to Quantum Optics and Spectroscopy

Spectroscopists are Colourful People
– a spectroscopist

In this chapter we provide a brief overview of the interaction of radiation with matter from a semiclassical perspective. That is, as before, we treat the optical fields classically but treat matter quantum mechanically. In so doing we discuss the principal differences between the characteristics of light emitted from atoms, molecules and solids and how the level of excitation of a source influences the intensity and spectrum of light emitted. One of the key points in this chapter centers around a discussion of the three principal ways in which a light beam can exchange energy with matter, namely through spontaneous emission, absorption and stimulated emission. In the following chapter we build on the properties of stimulated emission in a discussion of one of the most useful light sources developed to date, the laser.

12.1. Black body Radiation and the Onset of Quantum Optics

The emission characteristics of a body in thermal equilibrium with its surroundings at an absolute temperature T have been known for over a century, but were not understood until first explained by Planck in 1900. In the ideal case where the emissivity of the body is independent of frequency, and having a value of unity, this radiation has come to be known as black body radiation, since a perfect emitter must be a perfect absorber in thermal equilibrium (hence black to reflected light, though it may glow itself). Planck's ideas are often said to mark the genesis of quantum mechanics. The key point in the theory of Planck is that radiation at a particular frequency is emitted and absorbed in quanta of energy with the energy of a quantum, Q_ω, proportional to the light frequency, ω, so that

$$Q_\omega = \hbar\omega.$$

Here $\hbar = h/2\pi$, with h ($= 6.6 \times 10^{-34}$ J-s) being known as Planck's constant. The overall energy in a monochromatic beam of frequency ω is then given by

$$E_\omega = NQ_\omega.$$

In macroscopic situations, involving Joules of energy, $\sim 10^{19}$ quanta are involved and the "graininess" of the energy would not be noticed. This corresponds to the classical limit of optics which we have been using up to this chapter and for which Maxwell's equations are valid in describing the light field characteristics.

Planck was able to deduce a simple expression for the (energy)/(unit frequency interval)/(unit volume) at a particular frequency, ω, using two simple ideas. For a black body cavity he argued that this energy density, or $\varrho_{BB}(\omega)$, would simply be given by the total number of modes/unit frequency/unit volume times the average energy in a cavity mode at a frequency ω in thermal equilibrium. He then applied thermal equilibrium arguments, and, in particular, the Boltzmann distribution to deduce the thermal average of the energy in a mode. This average energy, for a cavity temperature of T is

$$\langle E_\omega \rangle = \frac{\hbar\omega}{\exp(\frac{\hbar\omega}{k_B T}) - 1}$$

where k_B ($= 1.34 \times 10^{-34}$ Joule/K) is known as Boltzmann's constant. Therefore for a mode density, which is easily calculated to be (see most elementary books on Quantum Mechanics)

$$\sigma_{BB}(\omega) = \frac{n^3\omega^2}{\pi^2 c^3}$$

the energy density in the black body radiation is

$$\varrho_{BB}(\omega) = \sigma_{BB}(\omega)\,\langle E_\omega \rangle = \frac{n^3\omega^2}{\pi^2 c^3}\,\frac{\hbar\omega}{\exp(\frac{\hbar\omega}{k_B T}) - 1}.$$

Typical black body spectra are shown in Figure 12.1.1.

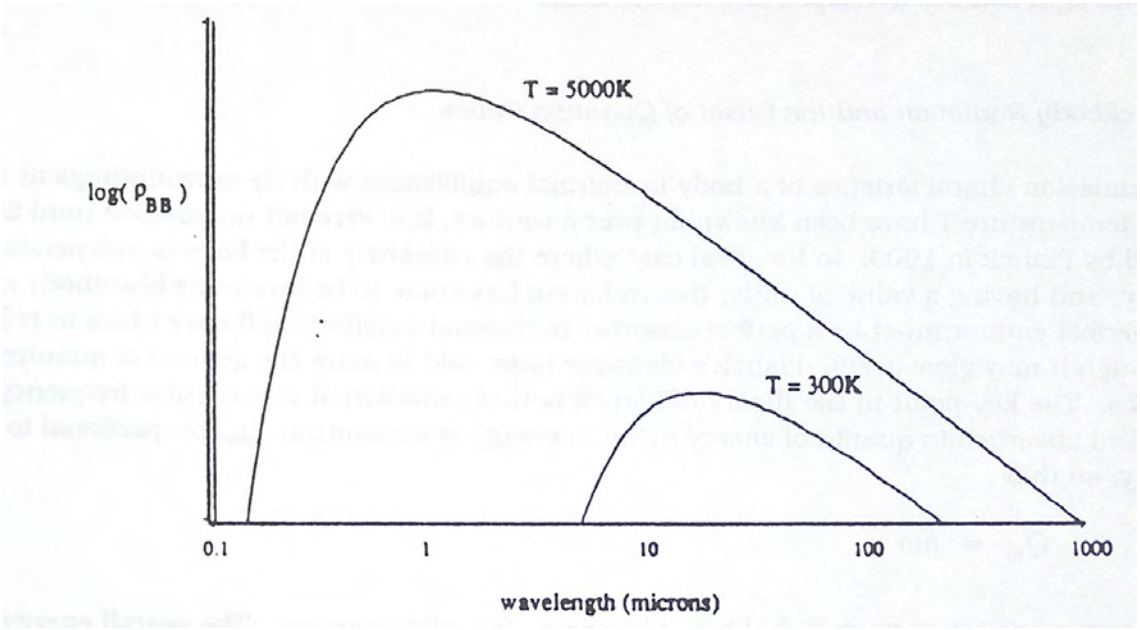


FIGURE 12.1.1. Black body spectra corresponding to T= 300 K and T= 5000 K.

As the temperature increases, the wavelength of the peak emission moves to smaller values, while the overall emission intensity changes. The wavelength corresponding to the peak in the emission spectrum is found from the formula for ρ_{BB} to be

$$\lambda_{max} = 2900/T$$

for T in Kelvin and λ_{max} in microns. This is known as the *Wien displacement law*. From Figure 12.1.1 one notes that λ_{max} for the sun (surface temperature of $\approx$ 5000K) is $\approx$ 0.5 µm (green light) while a human being ($T \approx 300$ K) emits radiation predominantly near 10 µm. The total intensity ($watts/m^2$) emitted by a black body at temperature T is given by the *Stefan-Boltzmann law*

$$I = c\int_0^\infty \rho_{BB}(\omega)d\omega = Z_{SB}T^4$$

where Z_{SB} ($= 5.67 \times 10^{-8}\, W/m^2K^4$) is known as the *Stefan-Boltzmann constant*.

Before proceeding further we should note that one of the main characteristics associated with radiation from a thermal equilibrium source is the breadth of the spectrum of radiation which increases with increasing temperature. For example, in the case of emission from a black body source at a temperature of 5000 K, one sees that significant radiation is emitted from $\lambda = 0.3$ to $\lambda = 3\mu m$, *i.e.* over a decade of wavelength.

In 1917, Einstein was able to use Planck's formula to construct a microscopic model for radiation emission and absorption events for individual atomic oscillators. In so doing he was able to derive relationships between probabilities for atomic emission and absorption events almost ten years before quantum mechanics was developed by Schrödinger, Heisenberg, etc. into the form in which it is used today. By 1917 Einstein was well aware of Bohr's model of the atom which indicated that the electronic energy in an atom was quantized into discrete levels and postulated that absorption or emission of radiation involved transitions of the electron between two levels. Einstein took these postulates one step further. Einstein argued on the basis of equilibrium thermodynamics and the principle of detailed balance that in the walls of a blackbody cavity each pair of energy levels associated with a radiative transition was in thermal equilibrium with the black body radiation field. This is indicated in Figure 12.1.2.

Consider then a pair of energy levels designated by labels i and j, and let g_i and g_j be the degeneracy of these levels, with N_i and N_j their equilibrium populations per unit volume. If these particular levels are associated with the absorption and emission of radiation of frequency ω then the principle of detailed balance says that the net flow of energy between the atoms and the field at this frequency is zero. We must therefore have that the number of emission events must exactly equal the number of absorption events over a sufficiently long period of time. Einstein

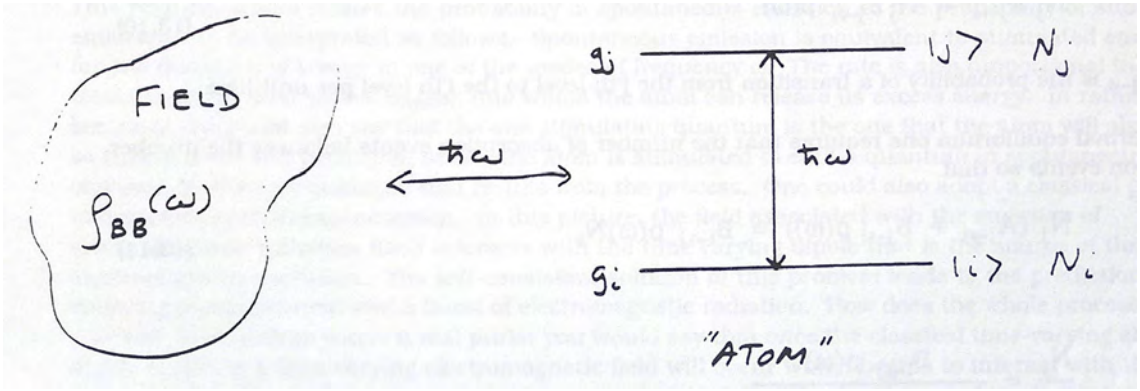


FIGURE 12.1.2. Detailed balance between a black body radiation field and the atomic energy levels.

postulated (quite reasonably) that there are three ways in which the atoms and the black body radiation field can exchange energy at frequency ω. These are:

1) *Absorption*

In this process the rate of absorption events, in terms of number of transitions per unit time per unit volume, would be given by the population of atoms per unit volume in the lower level (N_i) times the probability of a transition per unit time, to the j'th level ($W_{i\to j}$). This is just a matter of definition. What Einstein argued was that $W_{i\to j}$ was directly proportional to the density of available radiation that could be absorbed at frequency ω, $\rho(\omega)$, times an intrinsic atomic parameter, $B_{i\to j}$, which is the probability per unit time and per unit energy density at ω that a transition occurs to the j^{th} level. This leads to the equation

$$N_i W_{i\to j} = N_i \rho(\omega) B_{i\to j}.$$

2) Spontaneous emission

Here, atoms in the excited state, j, drop to a lower level of their own accord by emitting a light quantum. This process does not involve the radiation field. The number of such events per unit time and unit volume is given by

$$N_j W_{j\to i}$$

where $W_{j\to i}$, the probability of a transition per unit time is determined solely by an atomic parameter, $A_{j\to i}$.

3) *Stimulated emission*

The previous two processes were well established and, given empirical evidence, one could easily argue for the functional forms of the transition rates. Stimulated emission was a previously unheard of process postulated by Einstein. He stated that in the presence of a radiation field at frequency ω, atoms in the j'th level could be "stimulated " to give up a quantum of energy at a rate proportional to an intrinsic atomic rate, Bj→i, times the energy density $\rho(\omega)$. The number of such events per unit time per unit volume is given by

$$N_j W_{j\to i} = N_j B_{j\to i} \rho(\omega)$$

where $W_{j\to i}$ is the probability of a transition from the j'th level to the i'th level per unit time.

In thermal equilibrium one requires that the number of absorption events balances the number of emission events so that

$$N_j \left(A_{j\to i} + B_{j\to i}\rho(\omega)\right) = B_{i\to j}\rho(\omega) N_i$$

or

$$\frac{N_j}{N_i} = \frac{B_{i\to j}\rho(\omega) N_i}{A_{j\to i} + B_{j\to i}\varrho(\omega)}.$$

In thermal equilibrium the ratio of the populations in the i'th and j'th level is given by the Boltzmann factor so that

$$\frac{N_j}{N_i} = \frac{g_j}{g_i} \exp\left(-\frac{\hbar\omega}{k_B T}\right).$$

This additional condition immediately dictates that the energy density of the radiation field in thermal equilibrium with the atoms is

$$\rho(\omega) = \frac{A_{j\to i}}{B_{j\to i}} \frac{1}{\frac{B_{i\to j}}{B_{j\to i}}\frac{g_i}{g_j} \exp\left(-\frac{\hbar\omega}{k_B T}\right) - 1}.$$

When we compare this expression for $\rho(\omega)$ with that for $\rho_{BB}(\omega)$, we reach the following two results

$$1) \quad \frac{B_{i\to j}}{B_{j\to i}}\frac{g_i}{g_j} = 1 \text{ or } g_i B_{i\to j} = g_j B_{j\to i}$$

$$A_{j\to i} = \frac{n^3\hbar\omega^3}{\pi^2 c^3} B_{j\to i}.$$

If Einstein had not considered stimulated emission as one of the elementary processes involved in the interaction of radiation with matter, *i.e.* if $B_{j\to i} \equiv 0$, the form of the Planck formula would not be obtained. Apart from degeneracy factors, which basically only complicate the above expressions, the first result states that all other things being equal the probability of stimulated emission is equal to the probability of (stimulated) absorption. The only reason, apart from degeneracy factors, that matter in thermal equilibrium absorbs more light than it is forced to emit is that the population of atoms in the lower level is higher than the population in the upper level as the Boltzmann factor dictates for $0 \leq T < \infty$.

We can achieve some insight into the meaning of the second result derived by Einstein if we rewrite it in the form

$$A_{j\to i} = \hbar\omega\sigma_{BB}(\omega)B_{j\to i}.$$

This relation, which relates the probability of spontaneous emission to the probability of stimulated emission can be interpreted as follows. Spontaneous emission is equivalent to stimulated emission for one quantum of energy in one of the modes of frequency ω. The rate is also proportional to the total density of final states, $\sigma_{BB}(\omega)$, into which the atom can release its excess energy. In rather simple language one could also say that the one stimulating quantum is the one that the atom also emits, so that in some self-consistent sense, the atom is stimulated to emit a quantum in spontaneous emission by the very quantum that results from the process. One could also adopt a classical picture to describe spontaneous emission. In this picture, the field associated with the emission of electromagnetic radiation itself interacts with the time varying dipole that is the source of the electromagnetic radiation. The self-consistent solution of this problem leads to the prediction of a decaying dipole moment and a burst of electromagnetic radiation. How does the whole process get started? Well, unless you're a real purist you would say that once the classical time-varying electric dipole is set-up a time varying electromagnetic field occurs that begins to interact with the dipole. The full quantum picture is more complicated and would only divert us from the points to be made here.

Einstein was also able to derive additional properties of stimulated emission on the basis of detailed balance arguments. Assuming that the matter that formed the black body was a gas, Einstein argued that all the quantum absorption and emission events must maintain the Maxwellian velocity distribution of the gas. Apart from the energy exchange between the radiation field and the atoms, one has to consider the momentum exchange, which for each elementary process at frequency ω is according to chapter 2,

$$|\vec{p}| = \frac{nE_\omega}{c} = \frac{n\hbar\omega}{c}.$$

Momentum considerations led Einstein to show that in stimulated absorption or emission processes, the only thing that changes in the radiation field is the population of one of the modes of the system at frequency ω. To be more precise, if there exists a mode with a large population of quanta so that it becomes meaningful to talk about the electric field amplitude, wave vector and phase of the plane wave associated with this mode, in stimulated emission and absorption, only the amplitude of the field changes. T*he directionality and phase of the field are exactly the same as before and no other mode of the system, even one at the same frequency is affected.* It can further be argued, if one delved into these matters in greater detail, that stimulated emission and absorption processes are time-reversal pairs. Spontaneous emission, the odd man out (odd person out?), differs from these two in two fundamental ways. First of all, in a given elementary event, if it is meaningful to speak of such, one can't anticipate before-hand the direction of propagation or time of emission of the quantum. In a macroscopic sense, the fields involved have random phase and the direction of propagation of the E-M waves is random. In a time-averaged sense, the radiation from excited atoms, which give up energy through spontaneous emission, is emitted isotropically. The frequency of the elementary quanta are also not well-defined because of the finite emission time of the quantum or lifetime of an excited state so that

$$\triangle\omega\triangle t \geq 2\pi.$$

In the next chapter we point out how stimulated emission makes possible the unique coherent light source known as the laser. For the moment let's examine the fundamental characteristics associated with non-equilibrium light sources and the fundamental reasons why the absorption and emission probabilities are related.

12.2. Two Level Atoms

In this section we examine the problem of absorption and emission of radiation in atomic systems from a modern quantum-mechanical point of view. So as to avoid much repetition we expand the meaning of the word atom to include molecules and solids as well. The fundamental result we wish to achieve is the dependence of the absorption and emission rates on the atomic parameters.

The energy level spectrum of an atom is determined from the time-independent Schrödinger equation with a Hamiltonian $\mathbb{H}_0$. In the presence of an electric field one can determine the new energy level spacings and the transitions between energy levels by solving the Schrödinger equation with the new Hamiltonian,

$$\mathbb{H} = \mathbb{H}_0 + V$$

where the perturbation Hamiltonian associated with the electric field is (in the electric dipole approximation)

$$V = -e\vec{E}(\vec{r}, t) \cdot \vec{r}$$

For our purposes we take the electric field to be associated with a plane wave so that the spatial and temporal dependence of the field is governed by

$$\cos(\omega t - \overrightarrow{k} \cdot \vec{r}).$$

For most situations of interest to us, the wavelength, λ_0, is large compared to the size of the atom so that we make negligible error in replace $\vec{r}$ in the Hamiltonian V with $\vec{R}$, the nuclear position. Only in the case of X- or γ- radiation would this be of concern. This approximation is the *electric-dipole approximation* for the Hamiltonian V, since it in essence allows us to neglect the higher order multipoles of the atomic charge distribution. With this approximation, we have that the perturbation Hamiltonian can be written as

$$V = \frac{e\vec{E}_0 \cdot \vec{r}}{2}\left(e^{i\omega t} + e^{-i\omega t}\right) = \mathrm{V}\left(\mathrm{e}^{\mathrm{i}\omega \mathrm{t}} + \mathrm{e}^{-\mathrm{i}\omega \mathrm{t}}\right)$$

where E_0 is the amplitude of the electric field.

Consider a particular pair of energy levels of an atom labelled i and j as above such that the energy separation between them is

$$E_j - E_i = \hbar\omega_0.$$

For a particular pair of energy levels we can deduce the perturbation induced transition rate between these two levels from *Fermi's Golden rule* which states, that for a harmonic perturbation V, the transition probability per unit time, per unit frequency, between the two levels is

$$dW_{i\to j} = 2\pi\hbar^{-1}\left|\mathrm{V_{ij}}\right|^2 \delta\left(\hbar\left(\omega - \omega_0\right)\right)$$

where

$$\left|\mathrm{V_{ij}}\right|^2 = \frac{e^2}{4}\left|\left\langle i|\vec{E}_0 \cdot \vec{r}|j\right\rangle\right|^2$$

is the square modulus of the matrix element of the perturbation between the i'th and j'th states. If we choose the $\vec{E}_0$ field to be linearly polarized along the x-direction, then

$$V = \frac{eE_0 x}{2}$$

and we have

$$dW_{i\to j} = \frac{\pi e^2\hbar^{-2}E_0^2\left|\mathrm{x_{ij}}\right|^2}{2}\delta\left(\omega - \omega_0\right).$$

In reality the energy levels, apart from the ground state level, are never infinitely sharp and lifetime considerations due to radiative transitions, if nothing else, leads to a spread or uncertainty in the energy levels. The transition rate between the i'th and j'th level is therefore to be determined by integrating the differential transition rate $dW_{i\to j}$ over the density of final states (number of states per unit frequency interval). The density of final states is given by the product of the intrinsic degeneracy of the j'th level, g_j and the line shape function, $g(\omega_0^{'})$, which reflects the energy uncertainty in the transition. This function is peaked at ω_0 but has a non-zero width. By definition of the degeneracy factor, we require that the line shape function be normalized so that the total number of final states in the absence of degeneracy be unity, *i.e.*

$$\int_{-\infty}^{\infty} g(\omega_0^{'})d\omega_0^{'} = 1.$$

The overall density of states is thus $g_j g(\omega_0^{'})$.

The transition rate between the i'th and j'th energy levels induced by a monochromatic field at frequency ω is then

$$W_{i\to j} = \frac{\pi e^2 \hbar^{-2} E_0^2 \left|\mathrm{x_{ij}}\right|^2}{2} g_j \int_{-\infty}^{\infty} g(\omega_0^{'}) \delta\left(\omega - \omega_0^{'}\right) d\omega_0^{'}. \tag{12.2.1}$$

If the time averaged intensity (watts/m^2) of the monochromatic field is given by

$$I^{\omega} = \frac{1}{2} cn\epsilon_0 E_0^2$$

where n is the refractive index of the surrounding medium at frequency ω, then we have for the transition rate

$$W_{i\to j} = \frac{\pi e^2 \hbar^{-2} \left|\mathrm{x_{ij}}\right|^2}{cn\epsilon_0} g_j g(\omega) I^{\omega} = \frac{g_j}{g_i} W_{j\to i}.$$

Note that this expression is only valid under the assumptions used to derive it, especially assumptions regarding the monochromaticity of the source and the linearly polarized state of the $\vec{E}$ field.

The most important atomic parameter which determines the transition rate is the matrix element x_{ij} (with $|x_{ij}|^2$ being proportional to the classical oscillator strength). If this matrix element is zero, for symmetry reasons or otherwise, the pair of energy levels i and j is never associated with a radiative transition. Note too that, in the presence of a monochromatic field the only two levels of the atom that enter into the determination of the transition rate are the two energy levels which are in resonance with the field. Apart from the degeneracy of these levels, only one such pair usually exists for a given atom. In this sense, if the frequency ω is close to one of the resonance frequencies, ω_0, of the atom one can ignore all but two energy levels of the atom, hence the name "two level atom". The matter of degeneracy of these levels requires some additional comment. It may be that the matrix elementsx_{ij} depend on the actual states corresponding to the energy levels E_i and E_j, and not just the energies. That is, a given j'th energy level may be six-fold degenerate but only four of those states may have wave functions that lead to non-zero matrix elements of x between them and the particular i'th state. Also the four non-zero matrix elements may not all be the same. Without belabouring this point unduly at this level, in those cases we interpret g_j to be the number of radiatively coupled states and x_{ij} to be the average matrix element between those states and state i.

Since the matrix element of x determines the transition probability between different energy levels, it must be related to the empirical Einstein A and B coefficients. To derive this relationship we must recall that a black body radiation distribution involves a broad spectrum of light of different polarization states. In a frequency interval dω we can arrive at an effective differential field strength-squared, dE_0^2, from a knowledge of the energy density $\varrho_{BB}(\omega)$. This is given by

$$dE_0^2 = \frac{2}{\epsilon_0 n^2} \rho_{BB}(\omega) d\omega.$$

But the field is, in general, oriented in a random direction and certainly is not always along the x-direction. For an isotropic atom, the random orientation of the field relative to the x-direction can be accounted for by replacing $|x_{ij}|^2$ by $1/3|x_{ij}|^2$, as one can easily verify. In the presence of a black body radiation field we therefore have that the transition rate between the i'th and j'th energy levels is then found from equation 12.2.1 to be

$$W_{i\to j} = \frac{\pi e^2 \hbar^{-2} \left|\mathrm{x_{ij}}\right|^2}{3n^2\epsilon_0} g_j \int_{-\infty}^{\infty} g(\omega) \rho_{BB}(\omega) d\omega.$$

Because the black body spectrum is much broader than the line shape function, the line-shape function acts like a δ-function centered at ω_0 and so we have

$$W_{i\to j} = \frac{\pi e^2 \hbar^{-2} \left|\mathrm{x_{ij}}\right|^2}{3n^2\epsilon_0} g_j \rho_{BB}(\omega_o).$$

When we compare this with the defining equation for $B_{i\to j}$, we see that

$$B_{i\to j} = \frac{\pi e^2 \hbar^{-2} \left|\mathrm{x_{ij}}\right|^2}{3n^2\epsilon_0} g_j.$$

From the relation between the A and B coefficients we can also derive the $A_{i\to j}$ coefficient to be

$$A_{j\to i} = \frac{e^2 \hbar^{-1} n \omega_o^3 \left|\mathrm{x_{ij}}\right|^2}{3\pi\epsilon_0 c^3} g_i$$

where, in the expression relating the A and B coefficient we have replaced c by c/n if the blackbody cavity is filled with a medium of refractive index n. The Einstein A coefficient has the dimensions of $([T]^{-1})$ and indeed corresponds to the rate constant for spontaneous emission. We can therefore write

$$A_{j\to i} = \frac{1}{t_{sp}} = \Gamma = 2\gamma$$

where t_{sp} is the spontaneous emission lifetime and γ is the damping rate of the atomic oscillator which we identified in the classical sense in chapter 1.

Because t_{sp} is much easier to deal with, and certainly much easier to measure directly, than the matrix element of x, it is more convenient to write the transition rate from the i'th to j'th level in the presence of a linearly polarized, monochromatic field in terms of tsp. We then have that

$$W_{i\to j} = \frac{3\pi^2 c^2 g_j}{g_i n^2 \omega^3 t_{sp}} \hbar^{-1} I^{\omega} g(\omega).$$

Note that the apparent ω^{-3} dependence is not robust since $t_{sp} \propto \omega^3$. This latter effect is quite important however and indicates that the larger the frequency of the transition the shorter the spontaneous emission time. In the visible region of the spectrum the typical spontaneous emission time is 10 nanoseconds, whereas in the ultraviolet region of the spectrum the lifetime drops to picoseconds! The shorter lifetime reflects the larger density of final radiation states ($\sigma(\omega) \propto \omega^2$) and quantum energy ($\propto \omega$) involved in the transition.

12.3. Optical Sources

One of the principal characteristics of black body sources is that they can absorb and emit radiation at essentially any frequency. Of course, this is an ideal, and no single or even multiply constituent object can have the continuous absorption spectrum to achieve this, although some, like the sun, come close. The emission spectrum of most objects, whether in thermal equilibrium or not can be highly structured as a function of frequency, mainly because of the different density of radiation states, $\sigma(\omega)$. Another feature of a black body distribution is the requirement of equilibrium between the source and surroundings. In measuring that radiation some of the radiation is lost from the distribution and equilibrium is partially destroyed.

For these and other reasons, most realistic light sources are the result of some non-equilibrium process and we now turn our attention to the light distribution emerging from objects in these cases.

With the results of the past section in mind, it becomes clear that no matter whether or not the source is in thermal equilibrium, the overall rate at which the source emits light of a particular frequency (ω) emanating from a source in terms of watts/m^3, is proportional to the product of the density of atomic oscillators emitting a frequency ω (*i.e.* number per unit volume per unit frequency interval) times the rate at which light energy at frequency ω is being emitted. This is equivalent to the power spectrum (watts per unit frequency per unit volume) of a source, which for an equilibrium situation would be related to the absorption spectrum. The nature of the spectrum depends implicitly on a number of factors.

The density of atomic oscillators emitting light at a given frequency depends on the energy level distribution of the source. In the simplest case of a two level atom, if transitions between these two levels is radiatively allowed, the power spectrum would consist of a single peak, with a width dictated by lifetime effects. This two-level atom in a sense corresponds to the classical or Lorentz atom discussed in chapter 1, where only a single resonance frequency was considered. For a more realistic atom where there are infinitely many energy levels, the spectrum in general consists of a series of peaks, each corresponding to a transition between two energy levels. For atoms or molecules in a gas these peaks are usually discrete and the power spectrum appears as in Figure 12.3.1.

The transitions may in general involve transitions between electronic, vibrational or rotational energy levels. These discrete transitions are usually associated with bound states of the system, with the corresponding wave functions being local in character. The strengths of the various peaks is related to the relative population of the different energy levels and the "oscillator" strengths which empirically are related to the Einstein "A" coefficients of the levels. The populations are determined by the excitation or pumping scheme. Often, broadband optical pumping, collisional excitation or electronic activation in a high voltage discharge tube is used.

One can also have electronic transitions in solids, in particular, semiconductors and insulators. Here, because the electrons are not bound to particular atoms and are "free" to roam around the solid the energy levels develop into virtually continuous bands. The emission spectrum in this case consists of a broad emission band as indicated in figure 12.3.2.

The electronic transitions in this case involve transitions between states in the conduction band and those in the valence band, with the details of the spectrum, in terms of width and shape, being determined by the density of

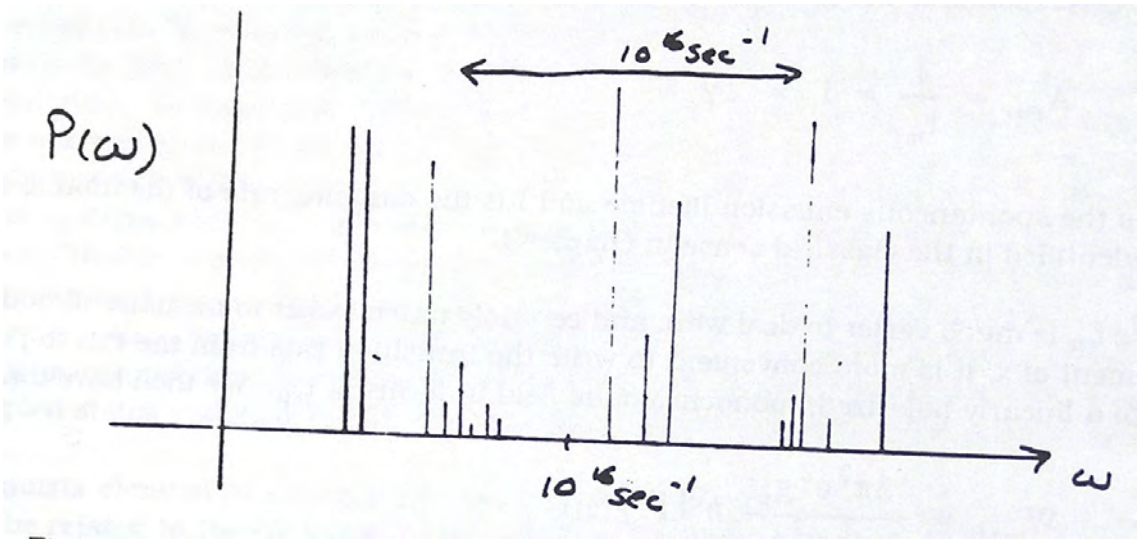


FIGURE 12.3.1. Typical power spectrum from an atomic or molecular gas.

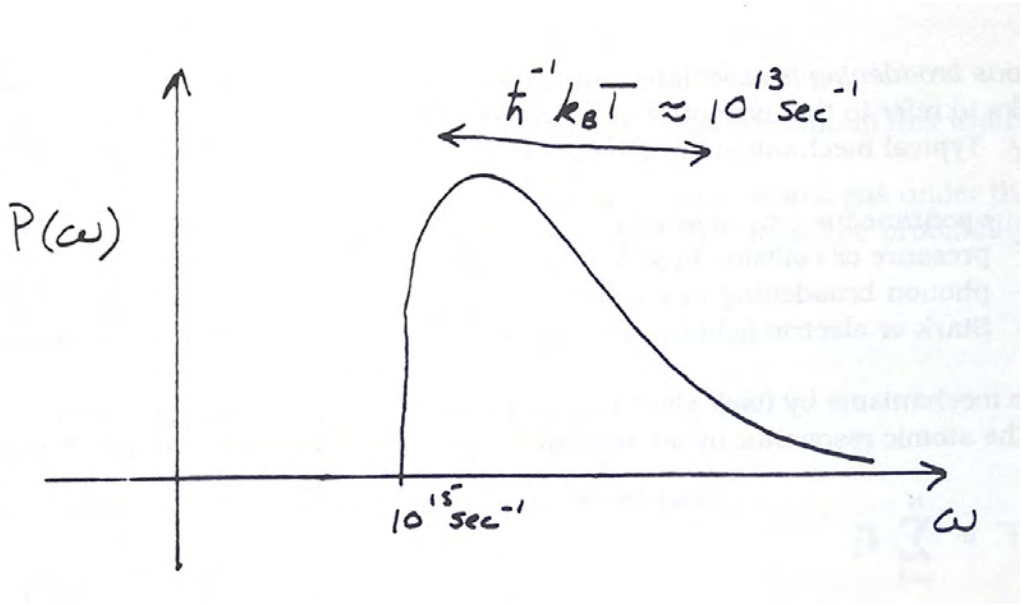


FIGURE 12.3.2. Emission spectrum from a semiconductor.

electronic states and the populations of these states. For a Fermi-Dirac distribution of electrons among the various levels, the width of the emission spectrum is determined by the width of the "Maxwell-Boltzmann" tails since transitions can only occur between filled and unfilled electronic levels. Since the width of the Maxwell-Boltzmann tail is given approximately by k_BT, the width of the spectrum is directly proportional to temperature.

Semiconductors are often used in solid-state light sources such as light-emitting diodes (LEDs) because of their small size and high efficiency for converting electrical energy into visible or near-visible light. Since the current which can be made to flow through a semiconductor (thereby yielding the non-equilibrium between conduction and valence band populations) can be turned on and off in times as short as one nanosecond, light emitting-diodes are used extensively in optical communications and various electronic readout units.

We return to the details of spectra later; here we just wish to emphasize that in the case of gases, involving weakly interacting constituents, the spectrum is usually discrete in nature with the widths of the peaks being typically $10^9 s^{-1}$ while in the case of solid state sources the spectrum usually has a single peak with a width of 0.1 eV or about 100 nm ($\approx 10^{15} s^{-1}$ in the visible region).

12.4. Line shape Functions

Besides the matrix element x_{ij}, the only other atomic parameter which determines the rate of electric-dipole transitions is the line-shape function, $g(\omega)$. If $\omega = \omega_0$ the rate is a maximum while if $\omega \gg \omega_0$, or $\omega \ll \omega_0$ the rate drops to a negligible value. In chapter 1 where we introduced a classical model of the atom we developed a simple harmonic oscillator model for the resonance response with the (amplitude of the) oscillator being damped with an effective rate constant γ. This led to a Lorentzian line-shape function with full width at half maximum of γ.

In general the line shape function, which determines the frequency dependent response of the atom-field interaction near a particular resonance, is determined by a number of factors such as collisions, radiative damping, dc electric fields, etc. These various effects which contribute to the broadening of the resonance are divided into two broad classes—homogeneous and inhomogeneous broadening.

Homogeneous broadening is associated with those mechanisms which effect all atoms in the same way. I like to refer to this type of broadening as democratic broadening in that all the atoms are treated equally. Typical mechanisms which lead to homogeneous broadening are:

i) spontaneous emission or lifetime broadening

ii) pressure or collision broadening in a gas

iii) phonon broadening in a solid

iv) Stark or electric field broadening, in a uniform field (e.g., not in a plasma)

If each of these mechanisms by itself gives rise to a damping of the atomic oscillator or equivalently a broadening of the atomic resonance by an amount Γ_i (i =1, 2, 3....) then the overall damping rate is

$$\Gamma = \sum_{i=1}^{N} \Gamma_i.$$

The line shape function is a Lorentzian line shape function with an overall full width at half maximum of $\triangle\omega = \Gamma$.

The damping rate Γ implies that some physical quantity has a lifetime Γ^{-1}. Which quantity is this? Since measurement in quantum mechanics implies results of an ensemble average, or a large number of individual atomic measurements, the damping rate is the average damping rate associated with a large number of atomic oscillators. For example the spontaneous emission lifetime is an average of the lifetime of an ensemble of oscillators. It is meaningless to discuss the lifetime of an excited state of an individual atom because of the probabilistic aspects of quantum mechanics in general and the emission event in particular. The decay of the macroscopic field emerging from the collection of atoms is the result of the decay of the macroscopic dipole or polarization density of the medium. This macroscopic or average dipole can disappear if all the individual dipoles decay or if the individual dipoles become dephased relative to each other so that the average vector dipole decays to zero. This can occur without the disappearance of the individual dipoles. In general, and for obvious reasons, the individual dipoles can dephase relative to each other without disappearing themselves (as they would, for example, if an atomic transition takes place). The damping rate γ is a damping rate for the macroscopic polarization density, regardless of the mechanism. In general this rate is larger than, but includes, the damping associated with lifetime effects of the individual atoms. This is made clearer when we discuss the second major class of broadening mechanism—inhomogeneous broadening.

Inhomogeneous broadening is associated with mechanisms which effect different atoms in different ways. This "non-democratic" form of broadening is due to such effects as:

1) *Non-equivalent environments of the atoms.* For example, for atoms located in glassy or amorphous solids, the local environment of each atom is different and each atom feels a different local electric field. The presence of these random perturbations on the different atoms in the solid leads to a distribution of resonance frequencies. Therefore, even in the total absence of any homogeneous broadening mechanisms, the resonance frequencies are distributed because of random non-secular perturbation effects. Because of the random aspects of the problem, the line shape function in the absence of any homogeneous broadening is a Gaussian of the form

$$g(\omega) = \frac{1}{\sigma_\omega \sqrt{\pi}} \exp\left(-\frac{(\omega - \omega_0)^2}{\sigma_\omega^2}\right)$$

where the full width at half maximum of the function is given by

$$\triangle\omega = 2\sigma_\omega \sqrt{\ln 2}.$$

In some cases, such as for Nd ions in a glass host (an important laser medium) this width is $10^{13} s^{-1}$.

2) *Doppler broadening in a gas.* The velocity distribution of atoms in a gas under thermal equilibrium conditions is given by the Maxwell distribution. In terms of the probability $f(\vec{v}\,)$ of finding a molecule with a velocity $\vec{v}$, this is

$$f(\vec{v}) = \left(\frac{M}{2\pi k_B T}\right)^{3/2} \exp\left(-\frac{1}{2}\frac{Mv^2}{k_B T}\right)$$

where, since each molecule has some velocity, we must have $\int\int\int_{-\infty}^{\infty} f(\vec{v})d^3v = 1$

If, in their own frame of reference the atoms emit radiation at a frequency ω_0, and if the component of their velocity relative to a detector or a source is v_x the observed frequency is shifted because of the Doppler effect. The observed frequency is given by

$$\omega = \omega_0 + \frac{v_x}{c}\omega_0.$$

Hence, even in the absence of homogeneous broadening the gas behaves like a collection of atoms at different resonance frequencies. The distribution of resonance frequencies can be found by converting the velocity distribution into a frequency distribution. With

$$v_x = \frac{(\omega - \omega_0)c}{\omega_0}$$

and

$$v^2 = v_x^2 + v_y^2 + v_z^2$$

we find from the normalization condition for $f(\vec{v})$, after integrating over the y and z- components of $\vec{v}$, that the line shape function is given by

$$g(\omega) = \frac{c}{\omega_0}\left(\frac{M}{2\pi k_B T}\right)^{3/2} \exp\left(-\frac{1}{2}\frac{Mv^2}{k_B T}\frac{c^2}{\omega_0^2}(\omega - \omega_0)^2\right)$$

where the full width at half maximum of the line shape function is given by

$$\triangle\omega_{FWHM} = 2\omega_0\sqrt{2k_B T\frac{\ln 2}{Mc^2}}$$

.For CO_2 molecules at room temperature (300 K) this width is typically $10^{10}s^{-1}$.

This discussion of the line shape functions for inhomogeneous broadening has led to Gaussian line-shapes in each case. This is characteristic of the random processes involved. Note that the principal difference between homogeneous and inhomogeneous broadening is that in the former case, the broadening mechanism is associated with some induced dephasing of the macroscopic dipole due to some active process which may or may not involve transitions within the atoms, while in the inhomogeneous case the broadening is merely a reflection of the distribution of resonance frequencies. Although it is clear that in a macroscopic sample the different resonance frequencies lead to an initially in-phase set of dipoles falling out of phase with each other in a time of $\Delta\omega^{-1}$, there is no active atomic process associated with this characteristic time.

In general, the overall line shape function is a convolution of a homogeneous and an inhomogeneous line shape function—a messy proposition indeed. Fortunately in many systems one of the types is dominant and one can simply ignore the other.

12.5. Macroscopic Aspects of the Interaction of a Monochromatic Wave with Two-level Atoms

To proceed from a microscopic picture of the interaction of light with a two-level atom to a macroscopic picture where we consider a collection of such atoms let us consider a uniform density of non-interacting two-level atoms as indicated in Figure 12.5.1.

For an incident monochromatic plane wave of intensity I^ω propagating along the z-direction, the net power generated per unit volume in the beam is given by the excess of stimulated emission events over absorption events times the amount of energy involved in a transition. That is,

$$\frac{dP^\omega}{dV} = \left[N_j W_{j\to i} - N_i W_{i\to j}\right]\hbar\omega = \frac{(N_j - \frac{g_j}{g_i}N_i)3\pi^2c^2}{n^2\omega^2 t_{sp}}g(\omega)I^\omega$$

where we have neglected the spontaneous emission term that makes a negligible contribution to the intensity in a single mode.

The power generated per unit volume is the same as the change in intensity per unit length so that we have

$$\frac{dP^\omega}{dV} = \frac{dI^\omega}{dz} = \frac{(N_j - \frac{g_j}{g_i}N_i)3\pi^2c^2}{n^2\omega^2 t_{sp}}g(\omega)I^\omega.$$

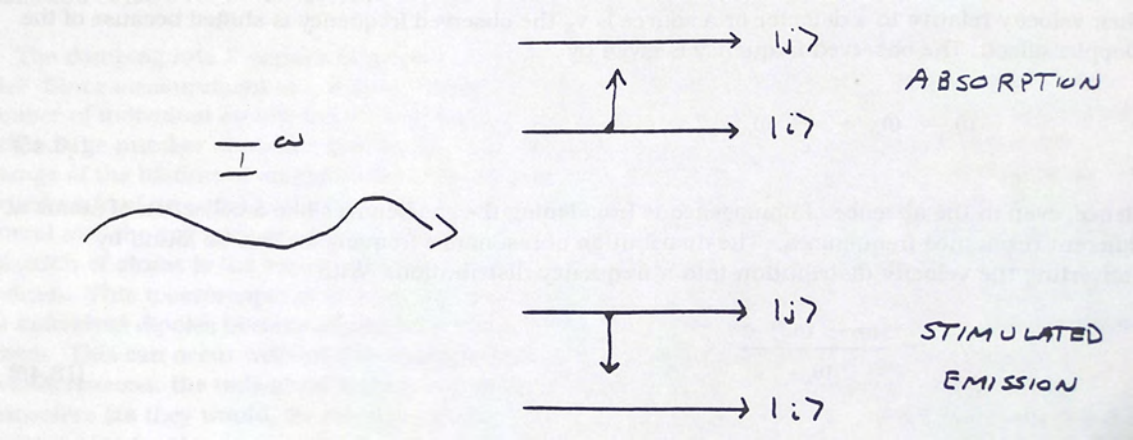


FIGURE 12.5.1. Interaction of a monochromatic field with two level atoms.

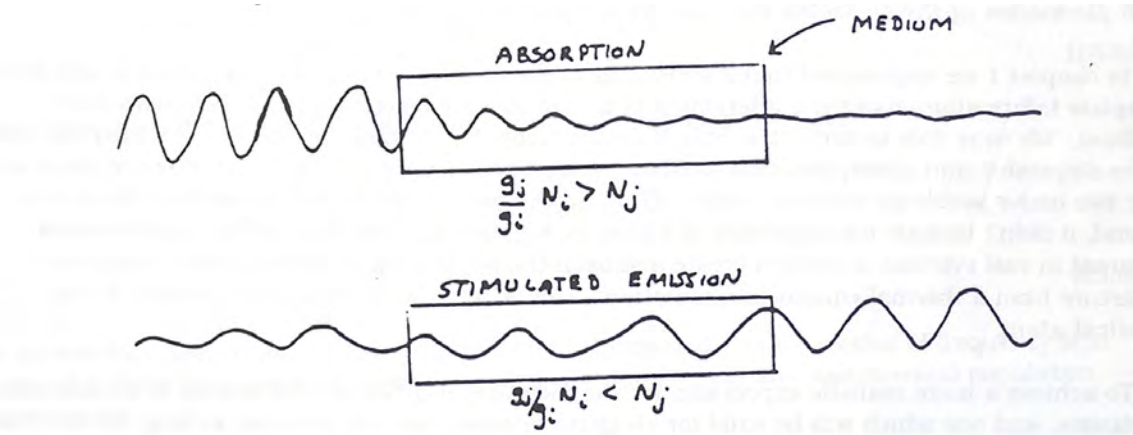


FIGURE 12.5.2. Amplification and Absorption of a monochromatic beam in a two level medium

The solution to this equation for I^ω is easily found to be

$$I^\omega(z) = I^\omega(0)e^{\gamma(\omega)z}$$

where

$$\gamma(\omega) = \frac{(N_j - \frac{g_j}{g_i}N_i)3\pi^2c^2}{n^2\omega^2 t_{sp}}g(\omega).$$

We refer to he parameter $\gamma(\omega)$, which has the dimensions of $[L]^{-1}$, as the *gain coefficient*. It has a positive value if the quantity in brackets is greater than zero. This corresponds to a situation where the upper level population density is greater than the lower level, if we don't concern ourselves with degeneracy effects. This is referred to as a population inversion. In such a case the rate of stimulated emission is greater than the rate of absorption and the intensity of the beam increases upon passage through the medium. On the other hand, if the population density of the lower level is greater than that of the upper level, $\gamma(\omega)$ is negative and the incident beam experiences net absorption. Indeed, $\gamma(\omega)$ in this case is the quantum mechanical analogue of the classical absorption formula we derived earlier. The net gain and net loss situations are illustrated in Figure 12.5.2.

To develop a feeling for the magnitude of the gain coefficient let us consider a typical solid state laser medium, $Al_2O_3 : Cr^{3+}$—otherwise known as ruby. Typically the density of Cr^{3+} ions in ruby is approximately $10^{19}cm^{-3}$. Under intense optical pumping, one can typically achieve a population inversion of

$$N_j - \frac{g_j}{g_i}N_i = 5 \times 10^{17}cm^{-3}$$

between the two levels with an energy difference corresponding to the wavelength $\lambda = 0.6943$ μm. Note that this population difference is only about 5% of the total atomic Cr^{3+} density, so that the population of the upper level is

only fractionally larger than that of the lower level. For ruby we also have that

$$g(\omega_0) \approx \frac{1}{\triangle\omega} \approx 3 \times 10^{-12} s$$

$$t_{sp} = 3 \times 10^{-3} s$$

$$\omega_0 \simeq 2.5 \times 10^{15} s^{-1}$$

$$n(\omega_0) = 1.5$$

so that

$$\gamma(\omega_0) = 1 cm^{-1}$$

a small value indeed. And this is the gain coefficient for a solid. A typical value for a gas would be lower still because of the much smaller atomic density. If we start with a very weak beam, we would be able to extract very little of the energy stored in the upper level of the medium. In the next chapter we will see that feedback is important if we wish to extract large amounts of energy from the medium. That is, we have to allow a beam to pass through the medium several times in order to extract significant amounts of energy.

Before proceeding to the next chapter which has to do with laser radiation, let us complete the quantum mechanical picture of the interaction of a monochromatic beam with a two level medium by deriving the dielectric constant in the vicinity of the resonance.

12.6. Derivation of the Dielectric Function for a Collection of Two-level Atoms

In chapter 1 we emphasized that a knowledge of the dielectric function of a medium would give one complete information about the interaction of a macroscopic electromagnetic field with that medium. We were able to derive a simple classical model for the dielectric function, which had many of the dispersion and absorption characteristics one observed in real systems. In essence there were only two major problems with the model. First, all the parameters were phenomenological, and second, it didn't include the possibility of stimulated emission. This latter effect only becomes apparent in real systems if one can create a substantial population inversion in real atoms—a departure from a thermal equilibrium situation. This is impossible within the context of the classical atom.

To achieve a more realistic expression for the dielectric function in the vicinity of a particular resonance, and one that is valid for all circumstances, we write the dielectric function as the sum of two parts. These reflect the contributions from all the other resonances in the system (a background contribution) and the contribution from the particular resonance itself, *i.e.,*

$$\hat{\epsilon} = \hat{\epsilon}_B + \epsilon_0 \chi$$

where the B labels the non-resonant or background contribution and $\epsilon_0\chi$ labels the contribution from the resonance itself with χ being the electric susceptibility. The complex refractive index is defined by

$$\hat{n}^2 = \frac{\hat{\epsilon}}{\epsilon_0} = n_B^2 + \chi$$

where the non-resonant contribution to the refractive index will be taken to be real, a result which is valid if we are far from other resonances in the system. We then have that

$$\hat{n} = n_B \left\{ 1 + \frac{\chi}{n_B^2} \right\}^{1/2}$$

which, in the limit of a small density of atomic oscillators or small χ, gives us

$$\hat{n} = n_B + \frac{1}{2}\frac{\chi}{n_B} = n + i\kappa.$$

Taking real and imaginary parts we have

$$n = n_B + \frac{1}{2}\frac{Re(\chi)}{n_B}$$

$$\kappa = \frac{1}{2}\frac{Im(\chi)}{n_B}.$$

But we have already shown that the gain coefficient is the negative of the absorption coefficient so that

$$\gamma(\omega) = -\alpha(\omega) = -\frac{4\pi\kappa}{\lambda} = -\frac{k}{n_B} Im(\chi)$$

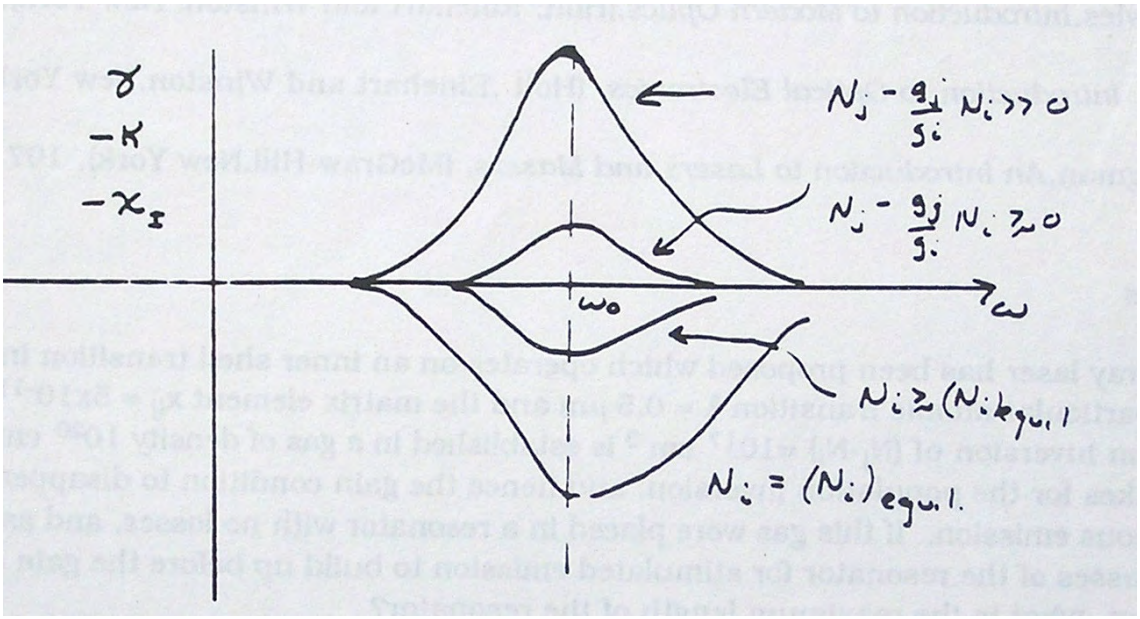


FIGURE 12.6.1. Schematic plot of the variation of $\gamma(\omega)$ -$\varkappa$ and -Im(χ) with frequency.

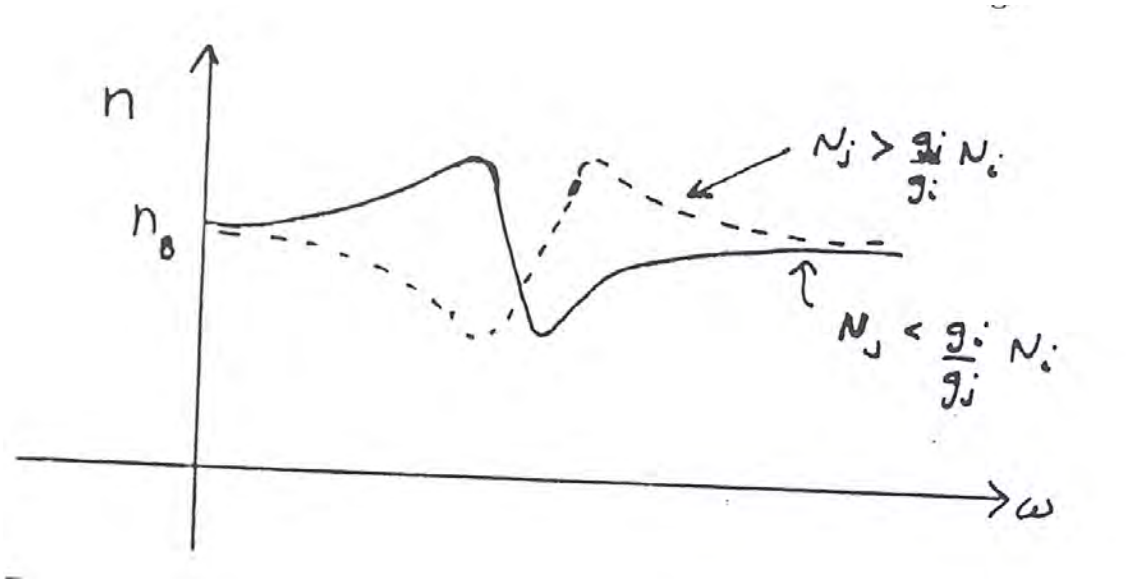


FIGURE 12.6.2. Plot of the refractive index near resonance for thermal and non-thermal distributions of population.

Using the quantum mechanical expression of the gain coefficient we have

$$Im(\chi) = -\frac{(N_j - \frac{g_j}{g_i}N_i)3\pi^2c^3}{n_B\omega^3 t_{sp} g_i} g(\omega).$$

The parameters $\gamma(\omega)$, $-Im(\chi)$ and -κ are shown plotted schematically as a function of frequency ω in Figure 12.6.1 for a Lorentzian line shape function and for thermal and non-thermal population distributions.

The corresponding expression for the $Re(\chi)$ can be found by using the Kramers-Kronig relation which relates the real and imaginary parts of the susceptibility. One finds that

$$Re(\chi) = \frac{2(N_j - \frac{g_j}{g_i}N_i)3\pi^2c^3}{n_B\omega^3 t_{sp} g_i} g(\omega)\frac{\omega_0 - \omega}{\triangle\omega} = 2\frac{\omega_0 - \omega}{\triangle\omega} Im(\chi)$$

Figure 12.6.2 shows a graph of the refractive index near resonance including the background contribution.

Note, that a change in sign of the population inversion not only implies a change in the sign of the absorption coefficient but also changes the sign of the resonant contribution to the refractive index as well.

References

G.R. Fowles, *Introduction to Modern Optics*, Holt, Rinehart and Winston, New York, 1963.

A. Yariv, *Introduction to Optical Electronics*, Holt ,Rinehart and Winston,New York, 1989.

A.E. Siegman, *An Introduction to Lasers and Masers*, McGraw-Hill,New York, 1971.

Problems

1. An X-ray laser has been proposed which operates on an inner shell transition in atomic cesium. For the particular atomic transition $\lambda = 0.5$ µm and the matrix element $x_{ij} = 5 \times 10^{-11} m$. If an initial population inversion of $(N_j - N_i) = 10^{17} cm^{-3}$is established in a gas of density $10^{20} cm^{-3}$, determine the time it takes for the population inversion, and hence the gain condition to disappear because of spontaneous emission. If this gas were placed in a resonator with no losses, and assuming one would like 10 passes of the resonator for stimulated emission to build up before the gain condition disappears, what is the maximum length of the resonator?

2. An atomic transition is both homogeneously and inhomogeneously broadened. If $g_1(\omega)$ and $g_2(\omega)$ are the respective independent line shape functions associated with each, derive an expression for the overall line shape function.

3. Compare the quantum and classical expressions for the susceptibility functions and comment on their numerical and functional differences.

CHAPTER 13

The Laser

Optics is Light Work
anon

In this chapter we present some of the salient features associated with one of the most useful sources of radiation, the laser. In the previous chapter we learned that if stimulated emission processes dominated absorption processes in an active medium, then an incident beam with a frequency near resonance can experience gain. The gain coefficient, even for energy level populations far from equilibrium, is never very large so some form of feedback is required. Typically the active medium is located in a Fabry-Perot resonator which allows standing waves to be set up in the form of Hermite-Gaussian beams as we have seen. If the reflectivity of one of the mirrors is less than unity, one can extract energy from the resonator. We derive the conditions that are necessary to achieve steady state or cw (continuous-wave) emission. We then examine general pumping schemes which are used to achieve the population inversion before considering typical laser systems from the gaseous, solid-state(insulator), solid-state(semiconductor) and liquid types.

13.1. Threshold Condition for CW Laser Oscillation

A typical laser system consists of an active medium which produces amplification of light together with a surrounding resonator structure which provides optical feedback for maximum energy extraction. Indeed the word laser is an acronym for light amplification by stimulated emission of radiation. This is illustrated in Figure 13.1.1.

The active medium may fill all or part of the cavity. For simplicity we only consider the former case here. We won't worry about how such a laser gets started, although typically this is done through spontaneous emission into one or more of the cavity modes. What we wish to do is to derive the condition for continuous laser oscillation as a function of the atomic and resonator parameters. The problem in general is quite complex, and at this stage we can only give a simplified picture. The thinking, however, proceeds as follows.

If we wish to form a self-sustaining oscillation in the resonator or laser cavity while extracting light from the system, the active medium has to be able to supply energy at the rate at which it is leaking out of the cavity, through one of the mirrors or other loss mechanisms such as scattering inside the medium or diffraction past the mirrors. Just at the threshold for operation we must have that the gain per unit length of the cavity must balance the loss per unit length averaged over the cavity. Under this condition we have that the amplitude of the light field in the cavity remains constant.

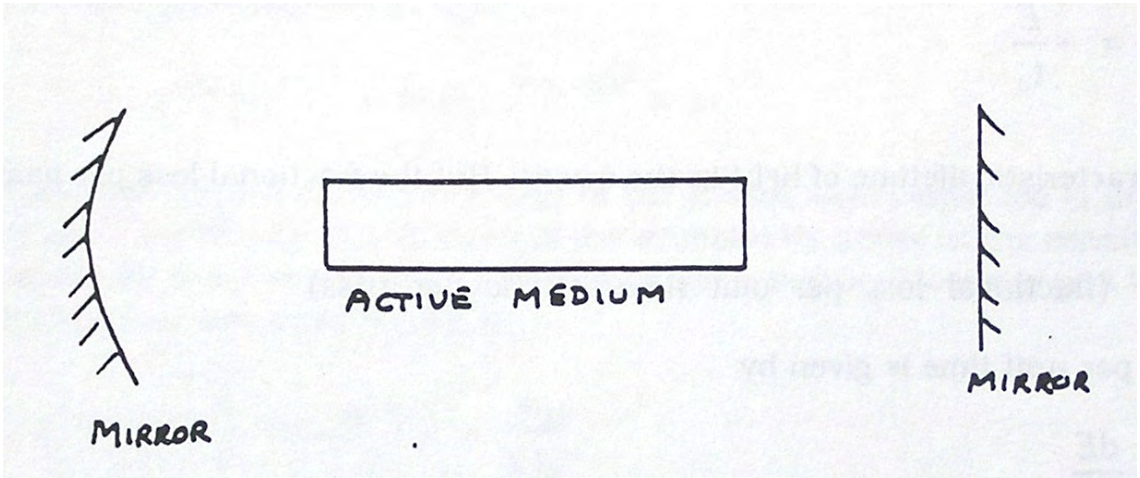


FIGURE 13.1.1. A typical laser oscillator.

Since a resonant cavity only supports modes at a certain frequency, phase considerations also are important in defining self-sustaining modes. As a matter of fact, there is a "competition" between the active medium and the resonator in determining at which frequency or frequencies the laser operates. The cavity, as mentioned, only supports modes at frequencies defined by the resonator geometry. On the other hand the active medium offers the most gain at line center where ω =ωo. Which aspect wins? Well, as in most conflicts a compromise is reached and the laser doesn't operate at either of the frequencies offered by the individual constituents.

13.2. Threshold Operation of a Laser— Amplitude Condition

Let us begin by determining what the relationship between the laser parameters must be to allow the laser to achieve self-sustaining oscillation at threshold where the gain per unit length balances the loss per unit length. The gain per unit length is simply given in the previous chapter, so it only remains to discuss the loss parameters for the cavity.

In general there are two types of loss. There is a distributed loss which occurs as a result of scattering of light from imperfections in the active medium: this gives rise to a loss per unit length of ς. The second loss mechanism is related to the loss of light from the cavity through the mirrors. This depends on the mirror energy reflectivities, R_1 and R (not to be confused with their radii of curvature).

To determine the effective loss per unit length, let us consider a cavity filled with light energy at time $t = 0$. If E is the instantaneous energy in the cavity at any time then due to scattering or leakage effects, this energy decays over time according to

$$\frac{dE}{dt} = -\frac{E}{t_c}$$

where t_c is the characteristic lifetime of light in the cavity. But the fractional loss per pass in the cavity, $\mathcal{L}$, is simply

$$\mathcal{L} = (\text{fractional loss per unit time}) \times (\text{time per pass}).$$

The fractional loss per unit time is given by

$$-\frac{1}{E}\frac{dE}{dt}$$

while the time per pass is just the length of the cavity divided by the speed of light in the cavity, that is,

$$n_B\frac{L}{c}.$$

It follows that

$$\mathcal{L} = -\frac{1}{E}\frac{dE}{dt}\frac{n_B L}{c}$$

and we then have

$$t_c = n_B\frac{L}{c\mathcal{L}}$$

After a round trip, an optical beam has its initial amplitude reduced through losses by the factor

$$e^{-2\mathcal{L}} = R_1R_2e^{-2\zeta L}$$

where ζis the (for example) scattering loss per unit length, giving for $\mathcal{L}$

$$\mathcal{L} = \zeta L - \ln\sqrt{R_1R_2}$$

and finally giving the cavity lifetime as

$$t_c = \frac{n_B}{c\left[\zeta - L^{-1}\ln\sqrt{R_1R_2}\right]}$$

For a 1 m long cavity, negligible ς and $R_1 = 1$. with $0.1 < R_2 < 0.99$, the cavity lifetime lies in the range of microseconds to nanoseconds. The gain per round trip in the cavity energy at threshold is governed by the factor

$$e^{2\gamma_t L}$$

where the subscript t indicates the gain coefficient at threshold of operation. At threshold the gain per round trip must balance the loss per round trip so that

$$\left(R_1R_2e^{-2\zeta L}\right)e^{2\gamma_t L} = R_1R_2e^{2(\gamma_t-\zeta)L} = 1$$

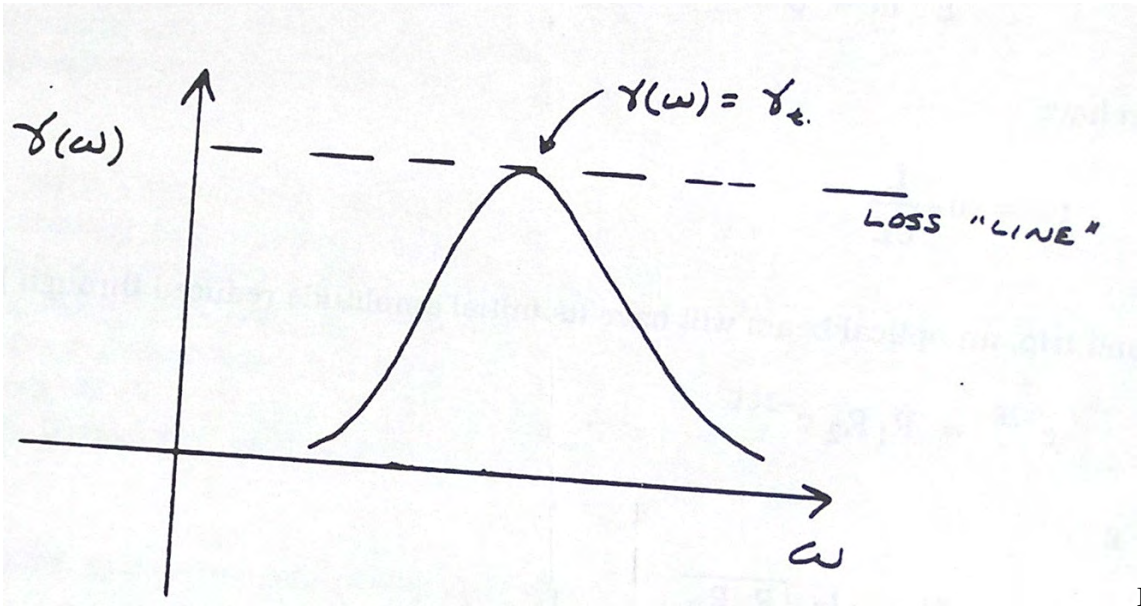


FIGURE 13.2.1. Relat
between gain and loss at threshold.

This equation guarantees that the stored energy or the electric field amplitude in the cavity is a constant. By taking logarithms of both sides of the equation we arrive at the condition that the minimum or threshold gain per unit length is equal to the distributed loss per unit length, plus end losses averaged out over the cavity length, or

$$\gamma_t = \zeta - \frac{1}{L}\ln\left(\sqrt{R_1 R_2}\right) = \frac{n_B}{ct_c} \tag{13.2.1}$$

The actual gain per unit length $\gamma(\omega_a)$ at an operating frequency ω_a must at least equal this threshold value γ_t. Using the relation between the gain coefficient and population inversion developed in the previous chapter, we have that the population inversion at threshold, ΔN_t, is given by

$$\triangle N_t = (N_j - \frac{g_j}{g_i}N_i) = \frac{n_B^2\omega_a^2 t_{sp}}{g(\omega_a)3\pi^2c^2}(\zeta - \frac{1}{L}\ln\left(\sqrt{R_1R_2}\right)$$

Alternatively, one can view this formula as determining what the cavity losses have to be to utilize a given population inversion as the basis for a laser. Equation 13.2.1 can be represented graphically as shown in Figure 13.2.1

The graph shows the gain per unit length and loss per unit length plotted together. The gain curve basically reflects the frequency dependence of the line shape function, which as we have seen typically has a FWHM of $\sim 10^9 s^{-1}$. The loss per unit length on this frequency scale is essentially independent of frequency since one can't obtain or make mirrors which have significant reflectivity variations over such a narrow range of frequency. Similarly the distributed loss coefficient isn't very sensitive to frequency. For this reason the loss curve is usually referred to as the loss line. To achieve threshold operation in a laser, the gain curve has to intercept the loss line at least at one point. If they do not touch laser action cannot be achieved.

To get a feel for some of the laser parameters, let us consider operating a He-Ne laser at 0.632 μm at line center where we can use

$$g(\omega_0) \equiv \frac{1}{\triangle\omega}.$$

For the active medium (neon) we have that $\triangle\omega \approx 10^{10}s^{-1}$ and $t_{sp} = 10^{-7}s$. Note that $\Delta\omega >> t_{sp}^{-1}$. This reflects the fact that the main broadening mechanism in this gas laser is Doppler broadening, an inhomogeneous broadening mechanism. Now, for a laser cavity with the following parameters:

$$R_1 = 1\, R_2 = 0.97\, \varsigma = 0\, L = 0.1m\, n_B = 1$$

we have that $\Delta N_t \approx 10^9 cm^{-3}$ a rather modest population inversion. If this laser were to operate at a pressure of 1 mTorr, the atomic density would be $\approx 10^{13} cm^{-3}$. Hence, only a small fraction of all the atoms are in an excited state in laser operation.

In a solid state laser such as the ruby laser, where t_{sp} with a value of 1 ms is so long, the required population inversion is at least 10^6 times higher.

The formula for the threshold condition is not explicitly dependent on the transverse mode structure. Recall however, that for modes higher than the TEM_{00} mode, more of the energy of the mode is contained off-axis. Given the finite transverse extent of the mirrors or the gain medium, sooner or later the condition for self-sustained oscillations is violated. The fact that ς and γ may depend on transverse co-ordinates only complicates the simplistic condition given here. Each mode has to be considered on its own merits, in terms of the ζ and γ that it experiences. Nonetheless, it should be obvious that by controlling γ, or ς one can determine the mode structure of the laser. One simple way of doing this is by inserting an aperture in the cavity perpendicular to the axis of the laser. By controlling the diameter of the aperture one can change the operation from single transverse mode, TEM_{00}, behaviour to multimode behaviour, provided the various modes can experience sufficient gain.

Finally it should be noted that the derived threshold condition says nothing about what the power output of the laser is. That is determined by the rate at which power is delivered to the atoms and by the fraction of the power which is delivered into the modes of the cavity which satisfy the threshold condition (a much more difficult problem to analyze). Indeed, it is also true that, provided steady state operation is maintained, the population inversion and any other laser parameter is unaltered. No matter how fast energy is delivered to the atoms to excite them, the population inversion is fixed. If this were not so, then the gain would be more or less than the loss, and the amplitude of the laser output would rise or fall. This would violate the steady state condition. Hence, under steady state operation, the gain per unit length is always equal to the threshold gain per unit length.

13.3. Threshold Operation of a Laser—Frequency Condition

Before we determine the frequencies at which the laser operates, let us take a second look at the Fabry-Perot resonator. In an earlier chapter we saw that the maximum transmission of a Fabry-Perot interferometer or cavity occurs for wavelengths, spacings, etc. that satisfy the standing wave condition. When the standing wave condition is satisfied, the amount of stored energy in the resonator is a maximum for a given incident energy. For other frequencies, the transmission response and stored energy of the cavity is reduced. In many ways the Fabry-Perot cavity behaves like a set of macroscopic oscillators with the ability to absorb (and re-transmit) certain frequencies while allowing other frequencies to pass. The lifetimes, t_c which in general are dependent on ω through R_1, R_2 and ς, correspond to the spontaneous emission times of these macroscopic oscillators. Such non-zero lifetimes, correspond to broadened resonances in the response functions and as we saw earlier in dealing with the Fabry-Perot interferometers, the transmission response functions are Lorentzians with FWHM determined by the mirror reflectivities. Here, for an asymmetrical Fabry-Perot with distributed losses, treated in the time domain from an energy storage point of view, we also see that the non-zero lifetime implies a frequency broadening with FWHM for each resonance of $\approx t_c^{-1}$. The fact that the two derived formulas for the FWHM of each of the resonances do not agree is related to different definitions for the "response" function, one related to transmission and the other to energy storage. It is a difference that we do not dwell on here. The point simply is that a "leaky" Fabry-Perot cavity does not possess a spectrum of δ-function but rather a series of broadened resonance peaks and the frequency width of these peaks is $\approx t_c^{-1}$.

In many ways the laser can be viewed as a coupled oscillator system. We have energy from atomic oscillators being fed into a geometrical or cavity oscillator. If the resonance frequencies of the oscillators match, then efficient energy transfer takes place. If both the atomic and cavity oscillators had infinitely sharp resonances then overlap of the spectra and laser action would not be possible. Of course lifetime effects alone immediately tell us that the atomic resonances won't be sharp and since we want some energy to come out of the laser ($R_2 \neq 1$) we won't have sharp cavity resonances either. Overlap is therefore possible. The only question is: what now determines the frequency of the light coming from the laser? We analyze the problem with respect to single mode, TEM_{00}, laser operation. The extension to multimode behaviour is straightforward, albeit messy.

In dealing with Gaussian beams we arrived at a condition for standing wave behaviour in a cavity and we developed an expression for the allowed frequencies of the longitudinal modes, ωq, in terms of the cavity parameters and the refractive index n. As mentioned, but ignored at that time, was the possible frequency dependence of the refractive index. In the case of laser operation involving an active medium we must explicitly take into account the frequency of the refractive index since we are operating in the vicinity of an atomic resonance where, as we saw earlier,

$$n = n_B + \frac{1}{2}\frac{Re(\chi)}{n_B} = n_B\left(1 + \frac{1}{2n_B^2}Re(\chi)\right).$$

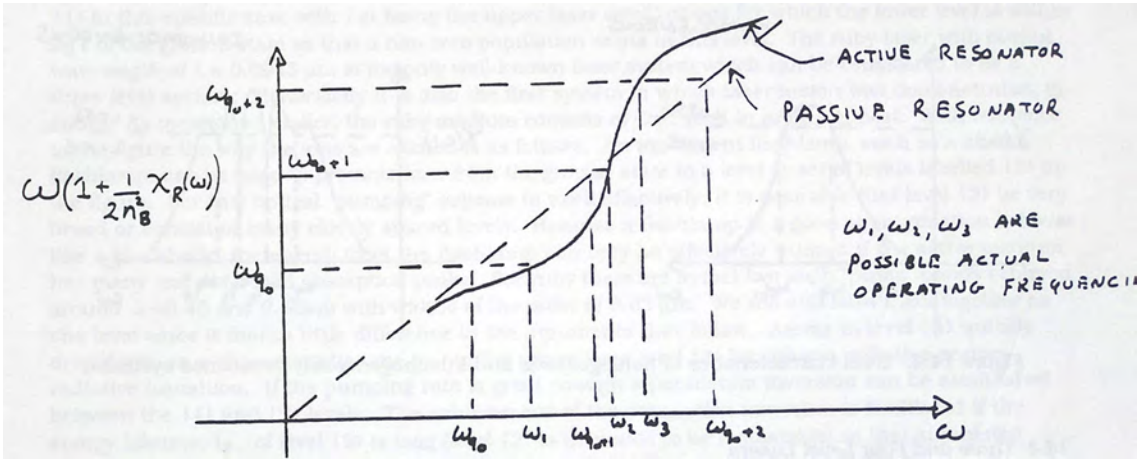


FIGURE 13.3.1. Determination of active cavity resonance frequencies in a laser.

If now ω is the actual frequency at which the laser operates, the round trip phase condition dictates that

$$\omega n_B \left(1 + \frac{1}{2n_B^2} Re(\chi)\right) = \omega_q \tag{13.3.1}$$

with ω_q being the passive cavity resonance frequency with a medium of constant refractive index n_B.

But for a Lorentzian line shape we have

$$Re(\chi) = \frac{2(\omega_0 - \omega)}{\triangle\omega} Im(\chi)$$

where $\triangle\omega$ is the full width at half maximum of the atomic resonance. We also saw that

$$\gamma(\omega) = -\frac{k}{n_B} Im(\chi)$$

It follows that equation 13.3.1 can be transformed into

$$\omega \left[1 - \frac{(\omega_0 - \omega)}{\triangle\omega} \frac{\gamma(\omega)}{k n_B}\right] \tag{13.3.2}$$

Now for most lasers, where the cavity resonances are much sharper than the atomic resonances, and hence more selective, we can anticipate that the actual operating frequency is closer to an ωq than the atomic resonance frequency ω_0. Hence, we can replace $\gamma(\omega))$ by $\gamma(\omega_q)$ and ω_q/k by c. Rearranging equation 13.3.2 we have

$$\omega = \omega_q - (\omega - \omega_0)\frac{c\gamma(\omega_q)}{n_B \triangle\omega} = \omega_q - (\omega - \omega_0)\frac{\triangle\omega_c}{\triangle\omega}$$

Solving for the actual operating frequency ω we find

$$\omega = \frac{\omega_q - \omega_0 \frac{\triangle\omega_c}{\triangle\omega}}{1 + \frac{\triangle\omega_c}{\triangle\omega}}$$

where $\Delta\omega$ ($= t_c^{-1}$) is the FWHM of the cavity response functions. From this it can be seen that the actual operating frequency is a weighted average of the atomic and cavity resonance frequencies. If the cavity is more selective than the atomic medium ($\Delta\omega_c \ll \Delta\omega$) then the operating frequencies is the same as the passive cavity frequencies. On the other hand, if we have a "lossy cavity" with $\Delta\omega \ll \Delta\omega_c$ the laser operates near the atomic resonance frequency ω_0, since without the constraints of a cavity an active medium has the highest gain at $\omega = \omega_0$.

Figure 13.3.1 illustrates graphically how the operating frequencies are determined from the condition that the round trip phase change of an active mode must be a multiple of 2π. Note that the effect of the active medium with a frequency dependent refractive index, is to "push" or "pull" the operating frequency towards or away from the passive cavity resonances.

For the He-Ne laser considered above one finds that $\triangle\omega = 10^7 s^{-1}$ and the spectrum of actual operating frequencies differs from the spectrum of cavity resonances by about $10^6 s^{-1}$. This is not a large shift and only of concern to those interested in high resolution spectroscopy.

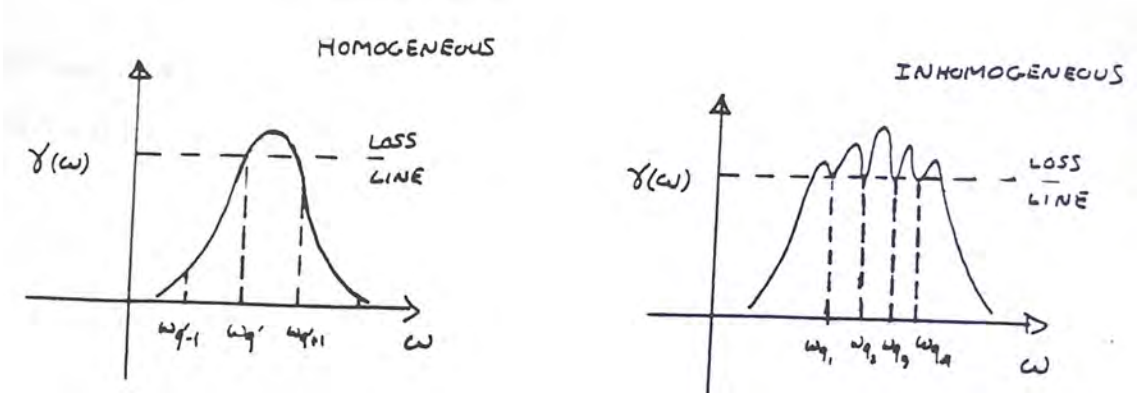


FIGURE 13.3.2. Gain characteristics of homogeneous and inhomogeneously broadened systems.

Finally, some words are necessary with respect to the difference in operation of homogeneously and inhomogeneously broadened lasers. These differences are based on the fact that for homogeneously broadened systems, all the atoms experience the same effects, including the interaction with any monochromatic beams whereas inhomogeneously broadened systems can be viewed as being composed of different "classes" of atoms. A homogeneously broadened system in steady state can only operate at one frequency for a TEM_{00} mode.

This can be seen in the following way. As the steady-state pumping intensity is increased, the gain curve as a whole rises to meet the loss-line. The curve continues to rise until some cavity mode q' with frequency ω'_q succeeds as the first cavity mode to produce gain equal to the losses, and lasing is supported. At this point, stimulated emission competes with pumping to clamp the gain profile: if we try to pump with higher power, the system finds a new steady-state, one in which the laser power is increased, producing increased stimulated emission, and depleting the population inversion more rapidly against the increased pumping of the population inversion. This effect depends on the lasing atoms all having the same gain profile, so that the gain/loss balancing condition for the mode applies to all the atoms contributing to the inversion, *i.e.*, that the system is homogeneous. No other TEM_{00} mode can then reach threshold, and single-frequency single-transverse-mode behaviour results. This is illustrated in Figure 13.3.2.

For the inhomogeneously broadened systems multi-frequency, single transverse mode behaviour is possible. Once the gain curve reaches the loss line so that a mode at a particular frequency can operate, increased activation of atoms still clamps the gain curve in the vicinity of the particular frequency. However, atoms which are more than a few homogeneous line widths away from the active class don't interact with the beam present in the cavity and so increased activation further increases their population inversion until a new class of atoms reach threshold. This can be continued until a large number of modes can be made to oscillate independent of each other. The number of such modes depends on the pumping rate but can be as high as the inhomogeneous line width divided by the homogeneous line width, since this equals the number of independent classes. In many ways the inhomogeneously broadened systems can be viewed as many independent lasers in one cavity. The gain characteristics of this type of system is shown in Figure 13.3.2

13.4. Three and Four Level Lasers

The previous section has assumed the existence of a population inversion between two levels. This section briefly addresses the problem of how this is achieved in practice and some of the atomic parameters that would be desirable to bring this about. The discussion is highly qualitative since the intent is only to give a flavour for the considerations and not to give a blue-print for laser construction.

Everything we have said up to now concerning laser operation has been based on the interaction of a monochromatic or quasi-monochromatic beam with a two-level atom. However, a true two-level system would never be able to give laser action. The problem lies in the achievement of the population inversion. It may be possible that some form of electronic activation (as in a gas discharge tube) could be found to maintain a population inversion but this would require careful energy matching to the one upper level. Similarly, one might contemplate optical pumping of the system by absorption of incoherent light, say from a flashlamp, to increase the population of the upper level. However, in this scheme, as soon as the population of the upper level becomes equal to that of the lower level, no further net absorption takes place and a population inversion cannot be achieved. It follows that, at least with respect to the problem of achieving a population inversion, at least one other energy level has to exist. Hence, a realistic laser must be at least a three level system. Such a system is shown in Figure 13.4.1.

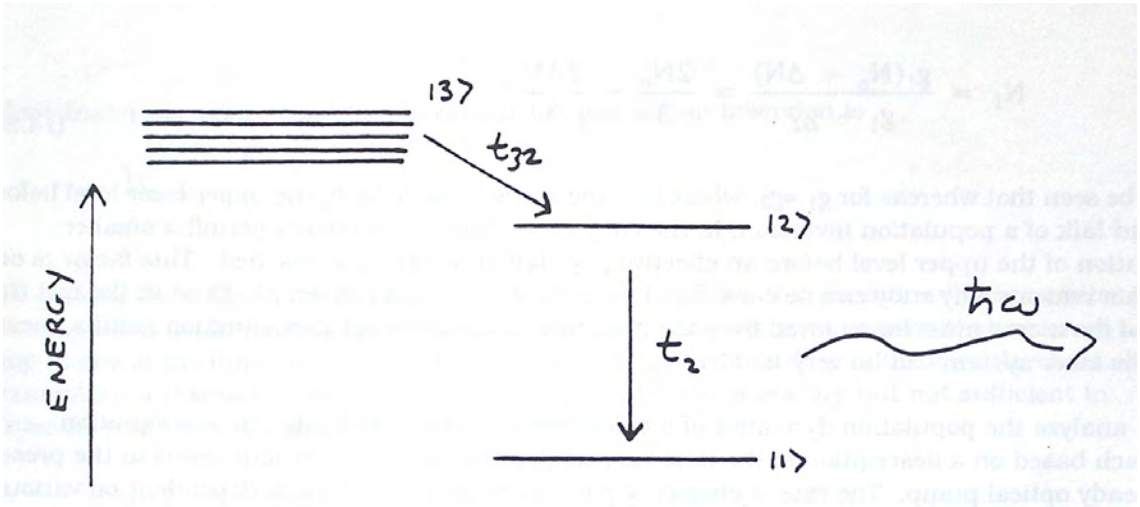


FIGURE 13.4.1. Energy level scheme for a three level laser.

A three-level laser is one for which the atomic ground state is the lower laser level (designated by $|1\rangle$ in this specific case with $|2\rangle$ being the upper laser level)) or one for which the lower level is within k_BT of the ground state so that a non-zero population exists in this level. The ruby laser with output wavelength of $\lambda = 0.6943$ μm is the only well-known laser system which can be considered to be a three level system. Historically it is also the first system in which laser action was demonstrated, in 1960). As mentioned earlier, the ruby medium consists of Cr^{3+} ions in an Al_2O_3 host. With reference to the figure the way the ions are excited is as follows. An incoherent flashlamp, such as a xenon flashlamp, can be used to promote ions from the ground state to a level or set of levels labelled $|3\rangle$ in the figure. For this optical "pumping" scheme to work effectively, it is desirable that level $|3\rangle$ be very broad or consist of many closely spaced levels. Because a flashlamp to a good approximation behaves like a black body, the energy from the flashlamp can only be effectively utilized if the active medium has many and/or broad absorption peaks. For ruby there are in fact two such "pump" bands centered around λ =0.45 and 0.55μm with widths of the order of 0.03 μm. We still label them together as one level since it makes little difference to the arguments that follow. Atoms in level $|3\rangle$ quickly drop down, in a characteristic time t_{32}, to the upper laser level $|2\rangle$ by either a radiative or non-radiative transition. If the pumping rate is great enough a population inversion can be established between the $|1\rangle$ and $|2\rangle$ levels. The achievement of the population inversion is facilitated if the energy lifetime, t_2, of level $|2\rangle$ is long (level $|2\rangle$ is then said to be metastable) so that stimulated emission can build up before all the stored energy is lost due to spontaneous emission.

In the case of the ruby laser $g_1 = 4$ and $g_2 = 2$ so that $N_2 = N_1/2$ for the gain to equal absorption. To obtain laser action the pumping rate would have to be increased beyond the rate required to reach this state so that a net gain can compensate for a net loss in the cavity. If ΔN is the threshold population inversion then

$$N_2 - \frac{1}{2}N_1 = \triangle N.$$

However if N_0 is the atomic number density of Cr^{3+} ions we also have that

$$N_1 + N_2 = N_0.$$

It follows that at threshold the level populations are given by

$$N_2 = \frac{g_1\triangle N + g_2 N_0}{g_1 + g_2} = \frac{N_0}{3} + \frac{2\triangle N}{3}$$

$$N_1 = \frac{g_1(N_0 - \triangle N)}{g_1 + g_2} = \frac{2N_0}{3} - \frac{2\triangle N}{3}.$$

It can be seen that whereas for $g_1 = g_2$, where half the atoms have to be in the upper laser level before one can talk of a population inversion, in the ruby case, degeneracy factors permit a smaller population of the upper level before an effective population inversion is reached. This factor is one of the main reasons why ruby can be considered a candidate for a laser material. Even so the fact that a third of the atoms must be removed from the ground state before a net gain situation results, means that the laser system is very inefficient.

To analyze the population dynamics of a three level system we adopt a rate equation approach based on a description of the time varying populations of the various levels in the presence of a steady optical pump. The rate of change of population in a given level is dependent on various constant energy relaxation rates of the levels of the atom. In the presence of the optical pump, the rate of change of population density in level $|3\rangle$ is given by

$$\frac{dN_3}{dt} = W_p(N_1 - N_3) - \frac{N_3}{t_{32}}$$

where W_p is the transition probability per unit time between levels $|1\rangle$ and $|3\rangle$ due to the optical pump, and, of course, t_{32} is the decay rate from the pump band to the upper laser level. For level $|2\rangle$ we have

$$\frac{dN_2}{dt} = \frac{N_3}{t_{32}} - \frac{N_2}{t_2}$$

where t_2 is the energy lifetime due to all effects of level $|2\rangle$. We can also set up a rate equation for level 1 as well or we can deduce its population at any time if we know N_2 and N_3 from the fact that

$$N_1 + N_2 + N_3 = N_0$$

with N_0 being the total atomic density. The steady state solutions can be found by setting the time derivatives equal to zero and solving for the population densities. This gives

$$N_3 = \frac{W_p t_{32}}{1 + W_p t_{32}} N_1$$

$$N_2 = \frac{t_3}{t_{32}} N_3$$

$$N_1 = N_0 - N_2 - N_3$$

Neglecting degeneracy effects, the population inversion on the $|2\rangle \rightarrow |1\rangle$ laser transition can be found to be

$$\triangle N = N_2 - N_1 = \frac{(1 - \frac{t_{32}}{t_2})W_p t_2 - 1}{(1 + \frac{2t_{32}}{t_2})W_p t_2 + 1}$$

It is clear that a necessary condition to establish a population inversion is

$$\frac{t_{32}}{t_2} < 1$$

and the smaller the ratio on the left hand side the better. Without this condition the numerator in the previous equation could never be positive. This condition is equivalent to saying that we must be able to dump atoms in the upper laser level at a rate faster than they can drop to the lower level, an obvious requirement for a population inversion. This condition is also necessary but not sufficient to guarantee an inversion. We must also have that

$$W_p t_2 > 1 \qquad \text{or} \qquad \mathrm{W_p} > \frac{1}{\mathrm{t_p}}$$

which says that the rate of pumping into excited states must be greater than the rate of decay from the upper laser level. This also would be necessary for population inversion.

If the population inversion can be maintained in the presence of stimulated emission (which we didn't include in the above analysis) into one mode, then in a steady state situation we would have that the power delivered into this mode or the power exiting from the laser would be

$$P = \triangle N_t V W_{2\rightarrow 1} \hbar\omega$$

where $W_{2\rightarrow 1}$ is the stimulated emission rate, V is the effective volume of the mode in the laser cavity and we have assumed the output to be purely monochromatic. The need to maintain at least $N_0/2$ in the upper laser level calls for a minimum expenditure of power in the laser in the form of spontaneous emission or other energy relaxation effects not associated with stimulated emission. This power is given by

$$P' = \frac{N_0}{2} \hbar\omega \frac{V}{t_2}.$$

For $t_2 = t_{sp}$ the emitted power is in the form of fluorescence and the power P' is referred to as the critical fluorescence power This is the power which, from the point of view of laser operation, is purely wasted but is the power that must be supplied to the system to maintain the system just below threshold of laser action. Higher power must be supplied to achieve threshold. In a ruby laser the critical fluorescence power is approximately 10 kW, which, given that most ruby rods are roughly the size of a cigarette, means that most ruby lasers are operated in pulsed mode

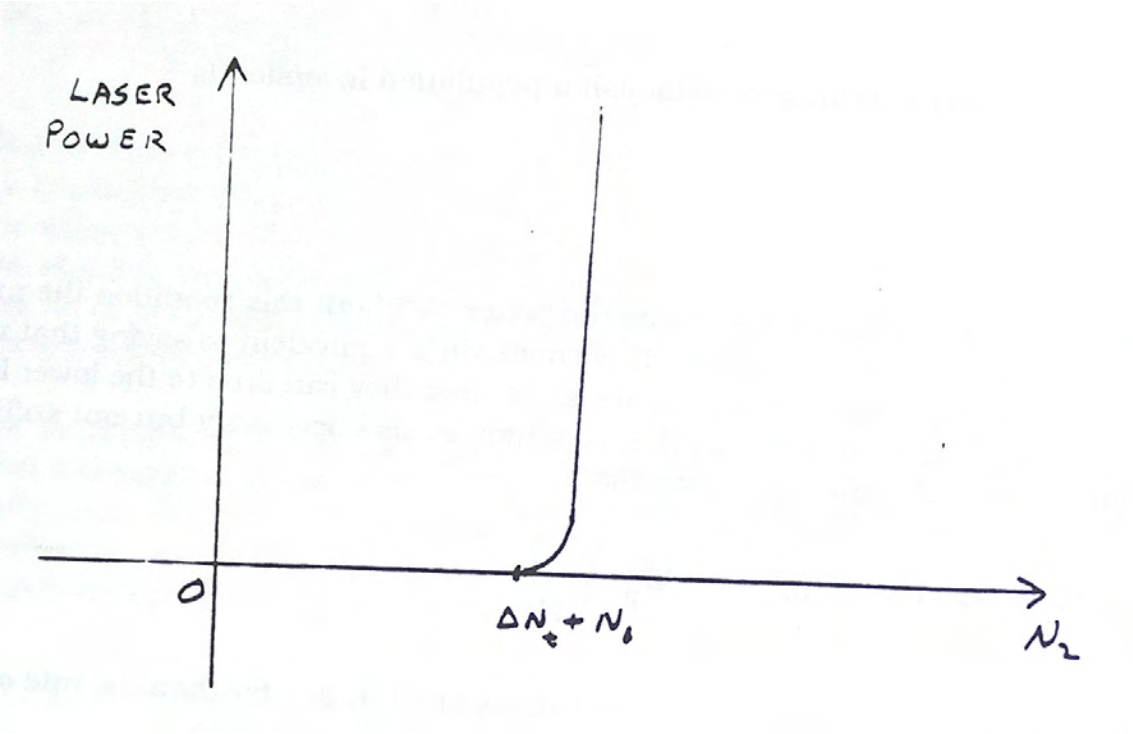


FIGURE 13.4.2. Output power of a three level laser as a function of population of the upper laser level.

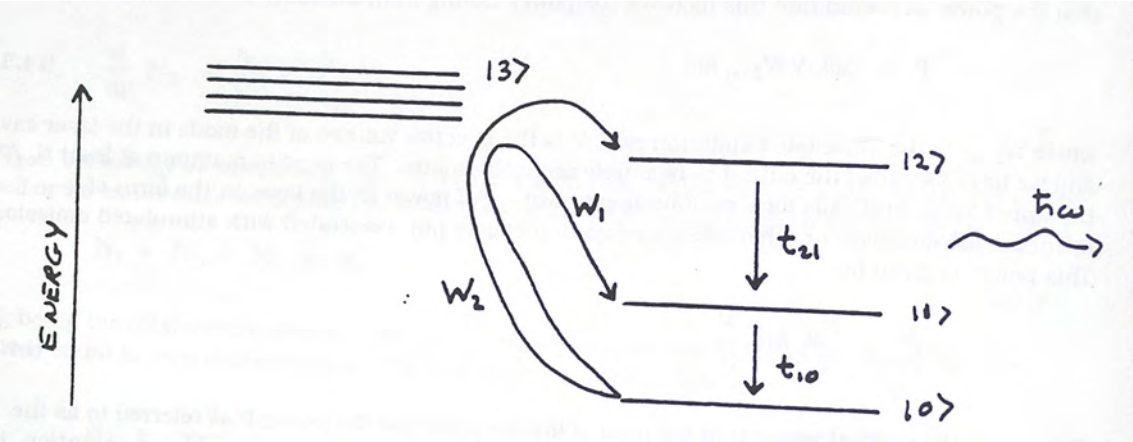


FIGURE 13.4.3. Schematic energy level diagram of a four level laser system.

since it is difficult to dissipate this energy continuously. Figure 13.4.2 shows the output power in the laser mode as a function of the population of the upper level in the case of $g_2 = g_1$.

N_t labels the threshold population, which, as we have seen, depends on the cavity geometry. Note that little, if any, power is emitted into the laser mode until threshold is reached. At that point, because of the population inversion and feedback from the cavity, the power in the laser mode becomes a very steep function of the excess population above threshold.

The obvious problems associated with three level lasers, particularly their low efficiency led researchers to consider schemes in which the lower laser level was not the ground state. This then calls for a four-level laser as illustrated in Figure 13.4.3.

If the lower laser level, $|1\rangle$, is energetically removed from the ground state by an amount $>> k_BT$, then even a small population of the upper laser level, $|2\rangle$, leads to a population inversion, and the possibility of a much more efficient system. It turns out that most lasers in use today are 4-level systems.

The analysis of the population dynamics of such systems is similar to that of the three level systems using the rate equation approach. To simplify part of the analysis we assume effective pumping rates from the ground state to the upper and lower laser levels. These are denoted by W_2 and W_1respectively. Of course the pump bands are used to create the initial atomic excitation but we assume that the atoms in these bands drop into one of the laser levels very quickly, thereby allowing us to talk of effective pumping rates into these levels. We also introduce stimulated emission in the analysis of the 4-level system since these systems are much more important than the 3-level systems.

In terms of the various lifetimes identified in the figure we have that the rate equations which describe the populations of the two laser levels are given by

$$\frac{dN_2}{dt} = -\frac{N_2}{t_{21}} - W_{21}\left(N_2 - \frac{g_2}{g_1}\right)N_1 + W_2$$

$$\frac{dN_1}{dt} = -\frac{N_1}{t_{10}} + \frac{N_2}{t_{21}} + W_{21}\left(N_2 - \frac{g_2}{g_1}\right)N_1 + W_1.$$

Where W_{21} is the transition probability per unit time between levels 1 and 2. In steady state, where all the time derivatives are zero we have that the population inversion is given by

$$\triangle N = (N_2 - \frac{g_2}{g_1}N_1) = \frac{W_2 t_{21} - (W_1 + W_2)t_{10}g_2 g_1^{-1}}{1 + W_{21}t_{21}}.$$

A necessary condition for achievement of population inversion is that

$$W_2 t_{21} > (W_1 + W_2)t_{10}g_2 g_1^{-1}$$

which in practice means that the lifetime of the upper laser level must be larger than that of the lower laser level. The effectiveness of the pump is reduced by the non-zero pump rate W_1 and the non-zero lifetime t_{10}. The effective pumping rate is therefore

$$R = W_2 - \frac{(W_1 + W_2)}{t_{21}}t_{10}g_2 g_1^{-1}$$

and the population inversion can be written as

$$\triangle N = \frac{R}{t_{21}^{-1} + W_{21}}.$$

Below the threshold of oscillation of the laser $W_{21} \approx 0$ since the intensity of the laser mode is effectively zero. In this case one simply has that

$$\triangle N \propto R$$

and the population inversion increases linearly with the pump intensity. This situation persists as we raise the pump intensity until

$$R = R_{21} = \frac{\triangle N_t}{t_{21}}$$

at which point we have reached the threshold population inversion. Further increase of ΔN is impossible as this would violate the steady state assumption. Hence, increasing the pump rate increases the power output in the laser mode but maintains ΔN clamped. This is possible of course if W_{21} is allowed to increase, which it does, so that

$$W_{21} = \frac{R}{\triangle N_t} = \frac{R}{\triangle N_t} - \frac{1}{t_{21}} \quad \text{for} \quad t_{21} \ll t_{10}$$

and the power in the laser mode is given by

$$P = \triangle N_t V W_{21}\hbar\omega.$$

This, of course, is the same expression as was derived for the three level laser except that for a 4-level laser $\Delta N_t << N_0/2$. By combining the last two equations we have

$$P = P_S(\frac{R}{R_t} - 1)$$

where

$$P_S = \frac{\triangle N_t}{t_{sp}}\hbar\omega V$$

and we have assumed $t_{21} \equiv t_{sp}$. The power P_S is the critical fluorescence for the four level system. In a typical cigarette-size system it has a value of 100 W which is a factor of a hundred less than that required for the 3-level system.

13.5. Effects of Spontaneous Emission

We have been careful up to now to ignore spontaneous emission in discussing the laser. In general its influence on laser operation is three-fold:

1) Spontaneous emission allows the laser oscillation to build up from "noise". Once the optical pump is turned on and the upper laser level has an excess population, spontaneous emission can occur. Although spontaneous emission is isotropic in character, some of the quanta emitted by the atoms are emitted into a cavity mode, and in particular the mode which has the lowest threshold. Once the population inversion reaches threshold, these spontaneously emitted quanta can participate in stimulated emission and build into a macroscopic beam. Thus the laser field builds up from randomly generated quanta or noise.

2) Spontaneous emission robs the laser of power in the laser mode. Once an excess population occurs and even after stimulated emission has built up a laser field, spontaneous emission always occurs, but only a small fraction of these quanta are emitted into the laser mode. The rest are simply lost in other propagation directions.

3) Because there is no correlation between spontaneous and stimulated emission events, spontaneous emission tends to add a small amount of phase noise to the coherent beam which otherwise would build up. The degree of phase purity or the line width of the emitted laser beam is fundamentally limited by spontaneous emission. Since spontaneously emitted quanta have a line width equal to the natural line width whereas stimulated emission quanta have no line width, the line width of the laser line is given approximately by

$$\triangle\omega_{laser} = \triangle\omega \times \frac{\text{number of spontaneously emitted quanta}}{\text{number of stimulated emitted quanta}}.$$

The line width of the laser beam can be estimated as follows. At or near threshold the number of quanta being emitted spontaneously is approximately equal to the number being emitted through stimulated emission. However, we have to determine what fraction of the spontaneous quanta are being emitted into the laser mode. For a wavelength of λ = 1μm and a natural line width of $\Delta\omega = 10^{10}s^{-1}$, the number of available modes per unit volume is

$$\sigma(\omega)\triangle\omega = \frac{\omega^2}{\pi^2c^3}\triangle\omega = \frac{4}{\lambda^2 c}\triangle\omega = 10^{14}m^{-3}.$$

For a laser cavity with a volume of $0.01m^3$ this would imply a total number of available modes for spontaneous emission of 10^{12}. Hence, in this case if the laser is operating in a single frequency single transverse mode, we would have

$$\triangle\omega_{laser} = 10^{-12}\triangle\omega = 10^{-2}s^{-1}.$$

The laser line is indeed very narrow. Our calculation is only meant to be accurate to within an order of magnitude or two but it gives the idea. If the laser is made to operate far above threshold so that more power exists in the laser beam, the line width can be further reduced.

In practice the line width of most lasers is not determined by spontaneous emission. Most lasers do not operate in a single frequency, single transverse mode and so the line width is determined by the frequency spread of the number of oscillating modes. Also, jitter in the laser mirrors due to vibrations, causes a modulation in the cavity length and this of course modulates the mode frequencies. In practice, for most continuously operating lasers, if one can get them to operate with a line width of 10^6s^{-1} one is doing well.

13.6. Pulsed lasers

(This site under construction). Relaxation oscillations, burst mode, Q-switching, gain switching mode-locking, pulse compression, attosecond pulse generation.

References

A.Yariv, *Quantum Electronics*, John Wiley, New York 1975.

A. Yariv, *Introduction to Optical Electronics*, Holt, Rinehart, Winston, New York 1976.

A.E. Siegman, *An Introduction to Lasers and Masers*, McGraw Hill, New York, 1971.

Problems

1. Calculate the critical fluorescence power P_s of the He-Ne laser operating at 0.63μm. Assume V= 2 cm^3, L = 1% per pass, and $\triangle\omega = 10^{10}s^{-1}$.

2. Calculate the critical inversion density ΔN_t of the He-Ne laser described in the first problem.

3. The longitudinal mode spacing of a semiconductor laser resonator is not uniform if the medium is dispersive in the region of oscillation. Show that the spacing, in wavelength, of adjacent longitudinal modes is

$$\triangle\lambda = \frac{\lambda^2}{2nL\left[1 - \frac{\lambda}{n}\frac{dn}{d\lambda}\right]}$$

where n is the refractive index and L is the resonator length.

4. The coupled rate equations for the cavity photon number (ϕ) and the upper-level atomic population (N_2) in an idealized single-mode laser, given one preferred lowest-loss cavity mode and very fast relaxation from the lower atomic level, are

$$\frac{d\phi}{dt} = K(\phi + 1)N_2 - \phi/\tau_c$$
$$\frac{dN_2}{dt} = -K\phi N_2 - \frac{N_2}{\tau_2} + R_p$$

where K, τ_c, τ_2 and R_p are constants.

a) Explain the significance of each of the terms in the equations. b) Obtain an expression for the steady state photon number in terms of the stated parameters and the pump rate normalized with respect to threshold (*i.e.*, $r = R_p/R_t$)) For $K\tau_2 = 10^{-10}$ evaluate the steady state photon numbers at $r = 1 + 10^{-5}$ and $r = 1 - 10^{-5}$ to obtain an indication of the sharpness of the laser threshold.

5. A He-Ne laser has a Doppler-broadened gain profile with Doppler width (FWHM) of $10^{10}s^{-1}$. If the gain constant at line centre is $3 \times 10^{-3}\,cm^{-1}$ and the only loss from the cavity is through a 98% mirror, what lengths of the cavity are allowed so that no more than two axial modes of the cavity are oscillating?

CHAPTER 14

Specific Laser Systems

Lead, kindly light...
J.H. Newman

In spite of the early difficulties associated with the invention of the first laser system, the ruby laser, it has since been shown that numerous materials can form the active medium for a laser. Active media include substances from the gas, solid and liquid phases of matter with insulators, semiconductors and metals all being made into lasers. A variety of different pumping schemes have also been developed including electronic discharge in a high voltage tube, flashlamp pumping, and chemical reactions.

Lasers are often thought by the public to be death rays or "photon torpedoes". As with most myths, their basis is in misunderstanding. Most lasers, particularly the ones operating continuously, produce less than 1 Watt of optical power, much less than an average light bulb. However, the optical radiation can be spectrally pure (has a high degree of temporal coherence), be spatially pure (has spatial coherence or a uniform phase front) and be highly directed. This opens up many applications. These qualities are achieved usually at great expense of energy. One myth that can immediately be put to bed is that the laser is a great source of energy. In fact, it is a lousy source of raw optical energy! A quick calculation points this out in the case of a flashlamp powered laser. The overall efficiency of a laser is the product of the following efficiency factors. Typical values are enclosed in brackets:

1) Efficiency of conversion of wall-plug electrical energy to energy stored in the capacitors. (0.99)

2) Efficiency of converting stored electrical energy into flashlamp energy. (0.90)

3) Efficiency of absorption of flashlamp energy by pump bands in atoms. (0.05)

4) Atomic quantum efficiency related to fact that pump bands are at higher energy than energy associated with useful emission of quanta. (0.7) 5) Efficiency of stimulated versus spontaneous emission in removing energy from atoms. (0.5).

The product of these various factors gives an overall efficiency of $\approx 1\%$, which is to be compared with gasoline or even steam engines which run at ten times this value. In certain laser systems (in particular those which avoid flashlamps!) the efficiency can be close to 20% as in the case of certain semiconductor or CO_2lasers.

In this section we discuss a variety of different types of lasers pointing out their main characteristics and applications. The intent is to indicate the variety of different types of lasers that can be produced. We begin the discussion by looking at two doped insulator lasers, the ruby laser and the Nd:YAG laser and then consider gas, semiconductor and dye lasers.

14.1. Ruby Laser

We mention this laser here since it was the first operational laser and was initially demonstraed by Ted Maiman of Hughes Laboratories in 1960. At present, apart from use in holography(see below) it is not used extensively. The ruby laser active medium is a small cigarette shaped pink rod which consists of approximately 0.05% of Cr^{3+} in Al_2O_3 ($N_0 \approx 5 \times 10^{19} cm^{-3}$). The main laser wavelength is 0.6943µm which occurs in the red region of the spectrum between a doubly degenerate excited state of Cr^{3+} and a quadruply degenerate state which is, at room temperature, well within k_BT of the ground state. Because the lower laser level essentially coincides with the ground state, ruby is a three-level laser system. Ruby was the first medium made into a laser, perhaps because of all that was known about this material in connection with work in microwave oscillators. We mentioned earlier that there are two main pump bands associated with this medium, one in the blue region and one in the green region of the spectrum. This is why ruby looks pink or red in transmission. The rod can have both of its ends polished with one end coated to yield a high reflecting surface while the other is coated to give a partial reflecting coating. The ruby rod as such would form a laser cavity. Usually, however, the ends are not coated and external mirrors define the actual cavity. The discharge of a capacitor provides energy to the flashlamp, which is usually wound around the rod in a helical fashion to provide uniform pumping of the rod. In pulsed operation the capacitor stores as much as 5 kJ of energy with approximately 10 J emerging in the beam along the rod axis. The main culprit in the poor efficiency of the

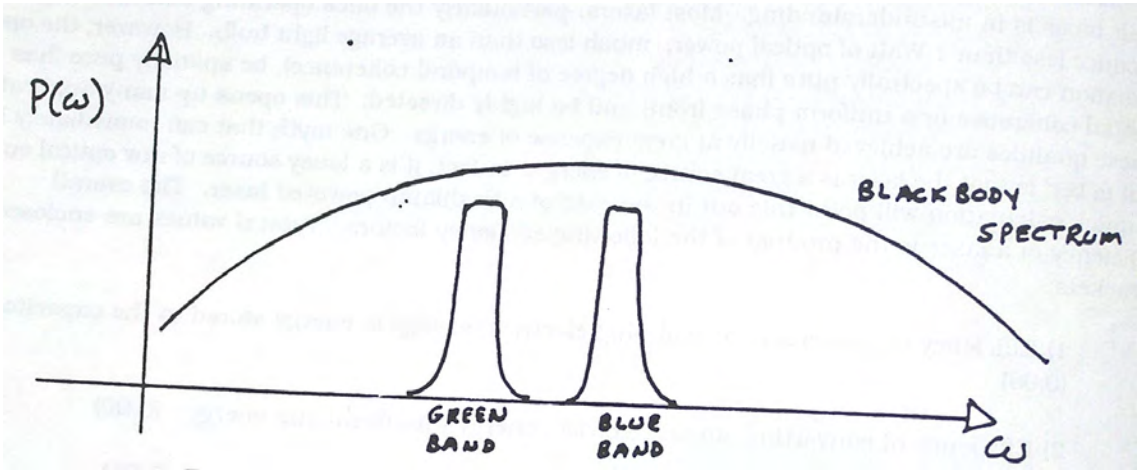


FIGURE 14.1.1. Overlap of the blackbody emission spectrum of a flashlamp and the pump bands of ruby.

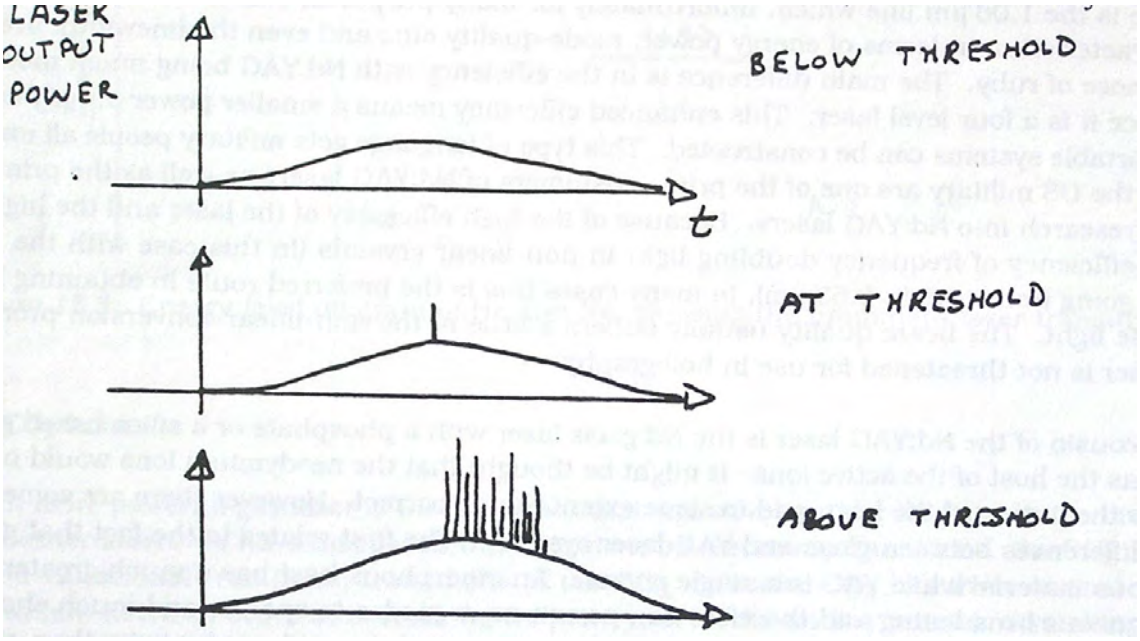


FIGURE 14.1.2. Output emission of the ruby laser under different pumping conditions.

ruby laser—as it is with most flashlamp pumped systems—is the poor overlap between the pump bands and the flashlamp output. Figure 14.1.1 illustrates the situation with respect to the bands in ruby.

Because the flashlamp effectively behaves like a black-body emitter with a temperature of ≈5000 K, its emission range, in terms of wavelength, is usually more than a decade larger than the corresponding range covered by the absorption bands. In what is known as conventional mode operation the flashlamp is excited by a 100 µs long electrical pulse. The output characteristics of the laser far below threshold, at threshold, and above threshold pumping conditions are shown in Figure 14.1.2.

At, and below, threshold pumping the output emission is basically that of the spontaneous emission into the laser spatial mode, with the rise and fall of the intensity being associated with the temporal characteristics of the flashlamp. Above threshold, an irregular series of spikes is observed in the output emission, near the peak of the spontaneous emission profile. However, these spikes have an intensity several orders of magnitude larger than that associated with the spontaneous emission and indeed are related to stimulated emission. The spiking behaviour is related to the ratio of the build up rate for stimulated emission divided by the pump rate into the upper laser level. In the case here this ratio is large. The flashlamp can provide enough population inversion so that the gain in the cavity exceeds the loss, stimulated emission is produced and within hundreds of round-trip times of the stimulated

beam, the population inversion is depleted and stimulated emission ceases. An output pulse approximately 1 µs long is produced. However the flashlamp is still on and over the next several µs will again produce a net population inversion and the laser will pulse again, *etc.* This cycle is repeated hundreds of time until the strength of the flashlamp intensity is not sufficient to yield a net gain situation in the cavity. In this burst mode the ruby laser produces a "jagged" output pulse with an overall duration of approximately 100 µs. The oscillatory behaviour of the output intensity is referred to as relaxation oscillations. For 10J of output energy the average power during the pulse is about 100kW. It is possible to produce shorter, more intense pulses from ruby and other types of lasers using Q-switching and mode-locking. With these techniques laser pulses as short as 30 picoseconds can be achieved.

When ruby is operated as a room temperature laser, the laser line is inhomogeneously broadened and capable of operating in as many as 1000 different longitudinal modes simultaneously within a gain profile with a width of $3 \times 110^{11} s^{-1}$. At much lower temperatures the laser is homogeneously broadened and produces a spectrally narrow beam.

Ruby is practically the only high power visible laser. Although in the early days of lasers it was regarded as a work-horse laser and many new scientific developments were achieved, because of its poor efficiency it is used today only in specialized applications requiring light at its particular wavelength of operation. Since, the ruby rod is essentially a single crystal, and large single crystals are difficult to grow, it is not possible to scale the size of ruby lasers to larger dimensions with associated higher power. Industry and the military have therefore long lost interest in ruby lasers.

There is however one very important use for ruby lasers which remains—pulsed holography. Because it is a visible wavelength laser capable of producing good beam quality, ruby lasers have seen widespread use in pulsed holography. With Q-switched pulses as short as 10 nanoseconds, one can take a hologram of most animate and mechanically moving objects. Such objects would not be able to move more than a wavelength over the duration of such a short pulse.

14.2. The Nd:YAG laser

A much more useful laser than the ruby laser, and one which was developed in about 1963 is the Nd:YAG laser. YAG is a synthetic, transparent host with the chemical formula $Y_3Al_5O_{12}$ and sometimes is referred to by its full name, yttrium aluminum garnet. Neodymium is a rare earth ion which serves as the active laser medium. Neodymium has a number of atomic transitions which have been successfully employed in laser operation (e.g., 0.96, 1.06 and 1.35 µm) as part of a four-level laser system with the lower laser level removed from the ground state. The most popular and most efficient line is the 1.06 µm line which, unfortunately for many purposes, occurs in the infrared. The output characteristics, in terms of energy power, mode-quality etc., and even the linewidth, are very similar to those of ruby. The main difference is in the efficiency with Nd:YAG being much more efficient since it is a four level laser. This enhanced efficiency means a smaller power supply can be used and portable systems can be constructed. This type of language gets military people all excited and indeed the US military are one of the prime customers of Nd:YAG lasers as well as the prime funders for research into Nd:YAG lasers. Because of the high efficiency of the laser and the high conversion efficiency of frequency doubling light in non-linear crystals (in this case with the wavelength going from 1.06 to 0.53 µm), in many cases this is the preferred route in obtaining high power visible light. The beam quality usually suffers a little in the non-linear conversion process, so the ruby laser is not threatened for use in holography.

A close cousin of the Nd:YAG laser is the Nd:glass laser with a phosphate or a silica based glass being used as the host of the active ions. It might be thought that the neodymium ions would not be influence by the nature of the host, and to some extent that is correct. However there are some important differences between glass and YAG laser systems. The first relates to the fact that glass is an amorphous material while YAG is a single crystal. An amorphous host has a much greater extent of inhomogeneous broadening and therefore can permit more modes to operate and much shorter pulses to emerge. Secondly, amorphous materials have a poorer thermal conductivity than single crystal materials and so cannot dissipate heat energy as rapidly. While YAG lasers can be pulsed as often as 20 times per second, some large glass laser systems can only be fired once every half-hour! Finally, Nd:glass lasers are scalable while the single crystal YAG lasers are not. It is therefore possible to build very large glass laser oscillator/amplifier systems which can deliver 100's of kJ of energy in pulses as short as 1 nanosecond. These lasers are currently under study for laser fusion systems.

14.3. The He-Ne laser

The He-Ne laser was the first gas laser developed and was done so by A. Javan at Bell Laboratories in 1961. A typical laser consists of Ne at a pressure of 1 Torr and He at a pressure of 0.1 Torr in a thin bore tube, of the order

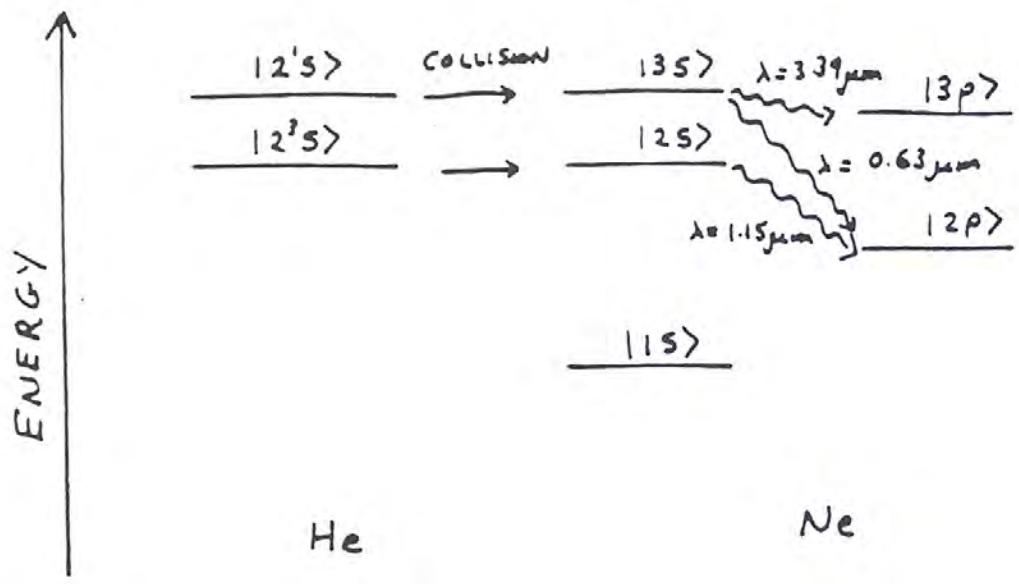


FIGURE 14.3.1. Energy level diagram of He and Ne, showing the prominent laser transitions.

of 10's of centimeters to meters long. A dc or rf discharge occurs in the tube through electrodes located near the ends. Initially it is the He gas that is excited and transfers its energy to the Ne which is not as easily excited. The Ne can then behave as a four-level laser system with efficient radiative transitions occurring at 0.63, 1.15 and 3.34 μm as indicated in Figure 14.4.2.

Although the first He-Ne laser was made to operate at 3.34 μm most of the He-Ne lasers in existence operate on the "red" line at 0.63μm. Because the laser is a gas laser, the density of active atoms and the output power is low. A very powerful He-Ne laser would have an output power of 50 mW. Because the gain of the laser is small, the mirrors have very high reflectivity and the circulating power inside the laser can be as much as 100 times the emitted power. Because of the low gain, nearly all He-Ne lasers operate as cw lasers.

14.4. The CO_2 laser

A much more powerful gas laser is the CO_2 laser that was developed by K. Patel in the mid-1960's. Unlike the other lasers we have discussed which operate on electronic transitions, the CO_2 laser operates on vibrational-rotational transitions of the CO_2 molecule. CO_2 has three normal modes of vibration which, listed in order of decreasing frequency, are: the anti-symmetric stretching, the symmetric stretching, and the bending modes. These modes are indicated schematically in 14.4.1. The vibrational energy in these modes is a function of their vibrational quantum number, vi, and can be represented as

$$E(v_1, v_2, v_3) = \sum_{i=1}^{3} \hbar\Omega_i(v_i + \frac{1}{2})$$

where Ω_i is the frequency of the i'th mode.

With the frequency of the various modes understood one can specify the vibration energy of the molecules solely in terms of the vibrational quantum numbers, *i.e.* (v_1, v_2, v_3). The convention is that the first digit is associated with the symmetric stretching mode, the second with the bending mode and the third with the anti-symmetric stretching mode. For each of these vibrational energy levels the molecule can possess rotational energy specified by the rotational quantum number, J, and given by

$$E_{rot} = BJ(J+1).$$

Typically the rotational energy level spacing of most molecular systems is smaller than that of the vibrational level spacing. This certainly is the case for CO_2. The composite energy level diagram of the rotational and vibrational modes of CO_2 is depicted in figure 14.4.2.

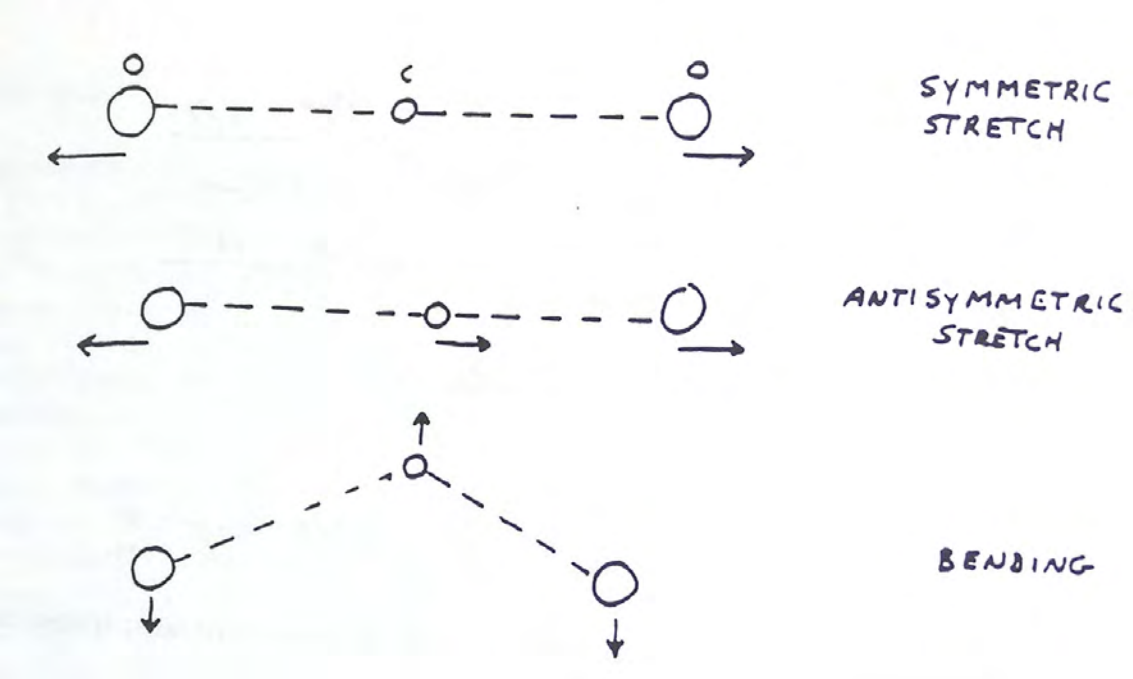


FIGURE 14.4.1. Normal modes of vibration of the CO_2 molecule.

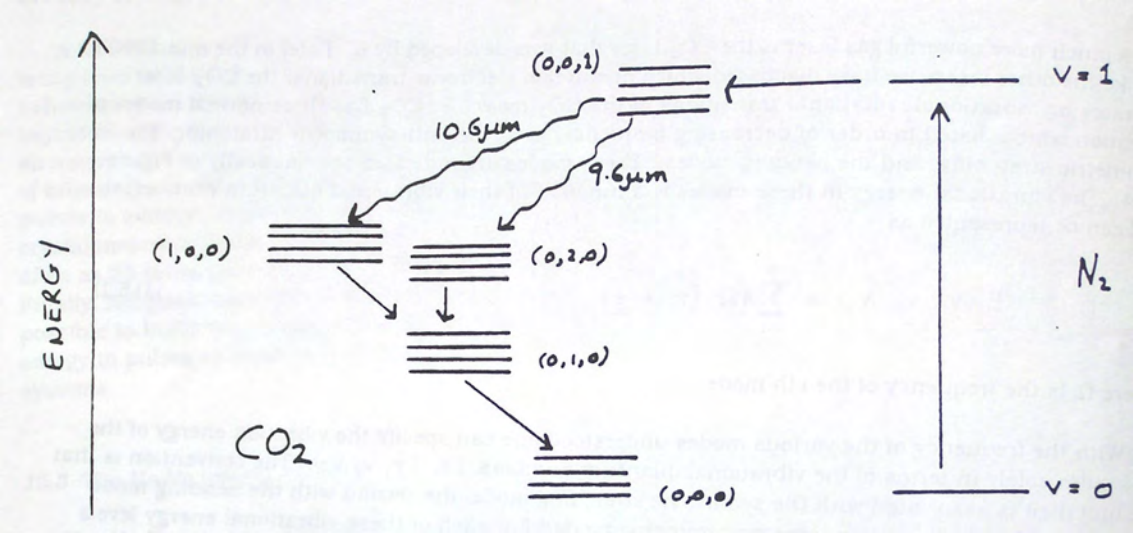


FIGURE 14.4.2. Energy level diagram of CO_2 indicating possible laser lines.

If molecules can be created in sufficiently high rotational-vibrational states then a population inversion exists relative to lower states. CO_2 is difficult to excite directly, however, through an electronic discharge. On the other hand, N2 is easily excited and is resonant with the (0,0,1) level of CO2. Once the excitation is transferred to this energy level a number of possible stimulated emission transitions can occur. The two most prominent ones occur to the (1,0,0), in the vicinity of 10.6 μm, and the (0,2,0) level, near 15.6 μm. There are of course a number of possible transitions because of the multitude of initial and final rotational states. The only selection rule is that $\Delta J = \pm 1$ for a rotational transition when occurring with a vibrational transition. Although the degeneracy of rotational states increases with J (it's given by 2J+1) the Boltzman factor restricts the population of high energy rotational levels. At room temperature the most populated state occurs for J = 21, and this rotational state gives rise to the largest gain with $\Delta J = 1$. The overall relative gain spectrum for the CO_2 laser is indicated in Figure 14.4.3. At high pressures the

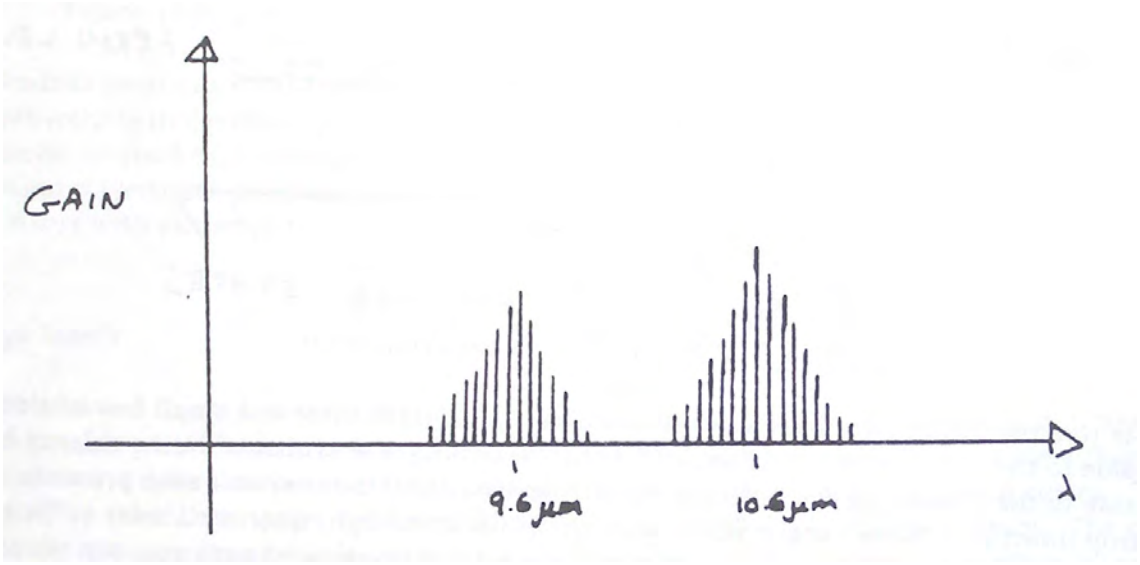


FIGURE 14.4.3. Relative gain spectrum for the CO_2 laser

various laser lines can become broadened and above one atmosphere of pressure the various lines experience overlap, so that the entire wavelength region from 9-11μm can yield gain.

The early CO_2 lasers were not very powerful and detailed research revealed that there was a bottleneck in the de-excitation process with molecules in the (0,1,0) state being long-lived. Such molecules could not be re-excited and robbed the laser of gain. To aid in the de-excitation of these molecules He gas is typically added to the laser. As it turns out, of the three gases present in the CO_2 laser, CO_2 is the least plentiful, constituting typically only 10%, with the balance being almost equally divided between N2 and He.

When the CO_2 laser was invented in 1965 it was made to operate only as a cw, low power laser with a power of milliwatts when only CO_2 was present. With the addition of N2 and He the power was boosted to several watts. Towards the end of the 1960's several schemes were proposed to allow high powers to be achieved in pulsed mode and in large volume lasers. Indeed, because the laser is a gas laser it is highly scalable. The problem with going to higher pressures or larger volumes is that it is difficult to get an electrical discharge to flow through the gas, let alone a uniform electrical discharge. In 1968 a group of scientists at the Defense Research establishment at Valcartier, Quebec, proposed using a transverse discharge rather than the longitudinal discharge that had been used previously. With this simple alteration in geometry the CO_2 laser, as a TEA (transverse electrical atmospheric pressure) laser is able to produce large amounts of laser energy with high efficiency ($\approx 10\%$). Indeed the CO_2 laser is being actively used in many industrial processes involving, cutting welding, drilling, etc.

14.5. The Semiconductor Laser

One of the smallest and lowest power lasers is the semiconductor laser which was developed at IBM in 1962. It had been known for many years that the luminous efficiency of certain semiconductors was very high but until the first laser was invented no one had thought to make a laser out of them. The first semiconductor laser was based on the compound GaAs, but since those early days many different kinds of semiconductors and semiconductor alloys have been fabricated into lasers spanning the wavelength region from 0.8 to 10 μm.

Like some of the earlier lasers we have considered, the semiconductor laser is based on electronic transitions. However, in semiconductors the energy levels are not discrete but form bands. The band which contains nearly all the electrons at room temperature is the valence band, while the first excited band is called the conduction band. There are many different ways in which a semiconductor can be pumped to promote electrons from the valence band to the conduction band to establish a population inversion within a restricted energy range. These include optical pumping and electron beam pumping. However, the most useful types of semiconductor lasers employ direct pumping with the passage of an electrical current through a p-n junction. A typical p–n junction is indicated in Figure 14.5.1. The n-type material has a number of impurity donors which have one more electron than that required for bonding requirements. This electron becomes thermally activated and occupies a state in the conduction band. The Fermi level which is a measure of the energy below which states are occupied is therefore close to the conduction band edge. In the p-type material, acceptor impurities are present which have a deficiency of electrons for bonding requirements

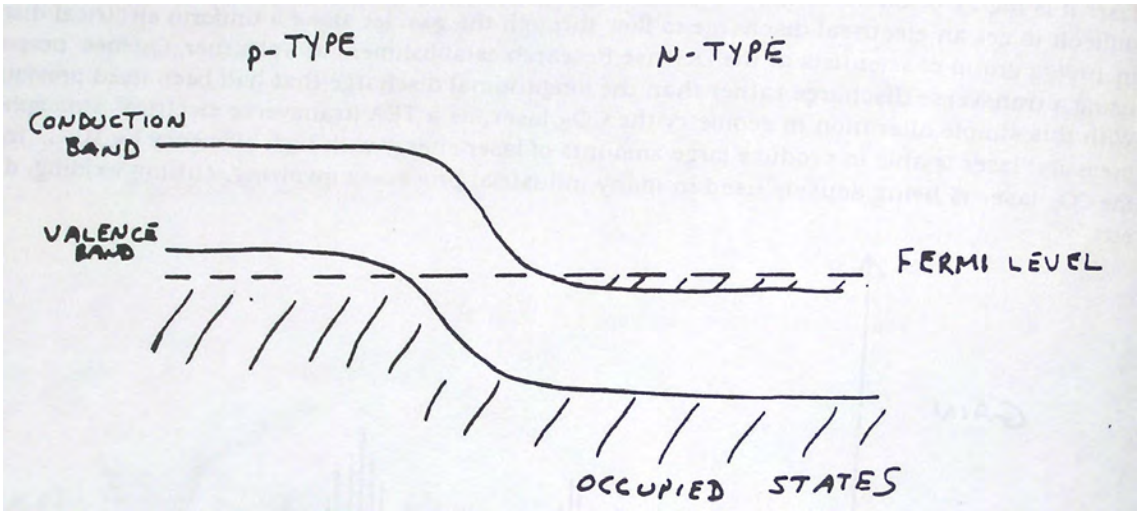


FIGURE 14.5.1. P-N junction in a semiconductor.

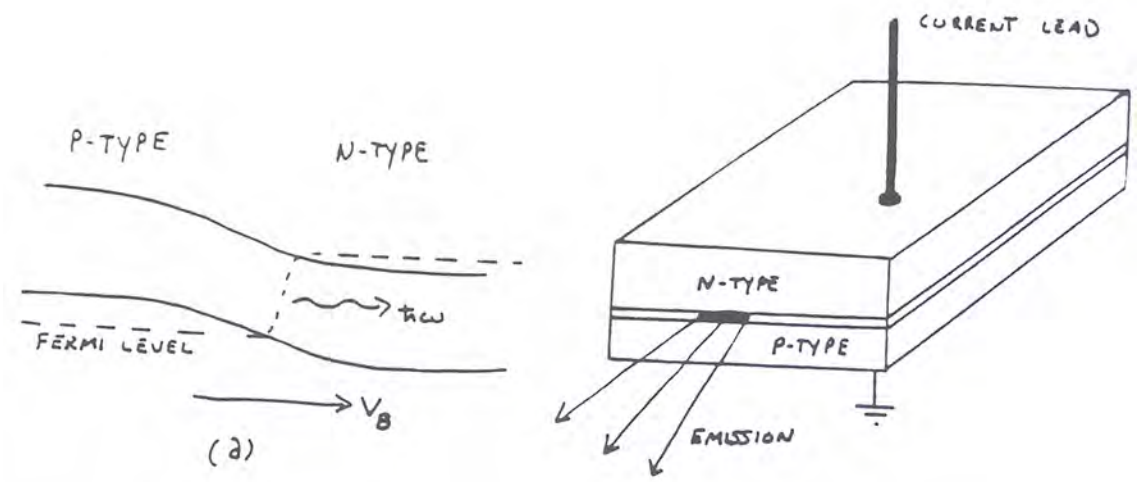


FIGURE 14.5.2. P-N junction under bias: a)energy level diagram b)actual junction.

and therefore remove electrons from the valence band. The Fermi level for the p-type materials is therefore near the valence band edge as indicated in the figure. When the two materials are brought together charges move between the two materials to lead to a neutral layer in the interface region. As a result of the charge movement from one material to another, the electron bands of the p-type material are raised relative to the bands in the n-type material. Stated differently, when the two materials come to equilibrium so that there is no net particle flow, the Fermi level in the two materials become equal as indicated.

When the junction is forward biased as indicated in Figure 14.5.2, there is a small flow of electrons from the n-side to the p-side of the junction and a corresponding flow of "holes" in the valence band from the p-side to the n-side. In the junction region, electrons find themselves in the presence of holes and drop down into these vacant states with the emission of light quanta. Under sufficiently high bias or current flow a population inversion is achieved and stimulated emission can occur. The typical junction occurs as a sandwich layer between the p- and n-type slabs as indicated in 14.5.2.

Light is emitted in the plane of the junction with the cleaved sides of the semiconductor crystal serving as end mirrors, with or without an optical coating depending on the semiconductor material. Because the typical current density required to reach threshold is of the order of $100\text{Å}/\text{cm}^2$, the lateral extent of the junction is usually confined to be less than a few microns so as to prevent significant heating. At the same time the thickness of the junction is of the order of 1 μm. The beam which emerges from the active region of the junction therefore experiences considerable diffraction with the light spreading into a cone with angular dimensions of 20° x40° typically. The cavity length of

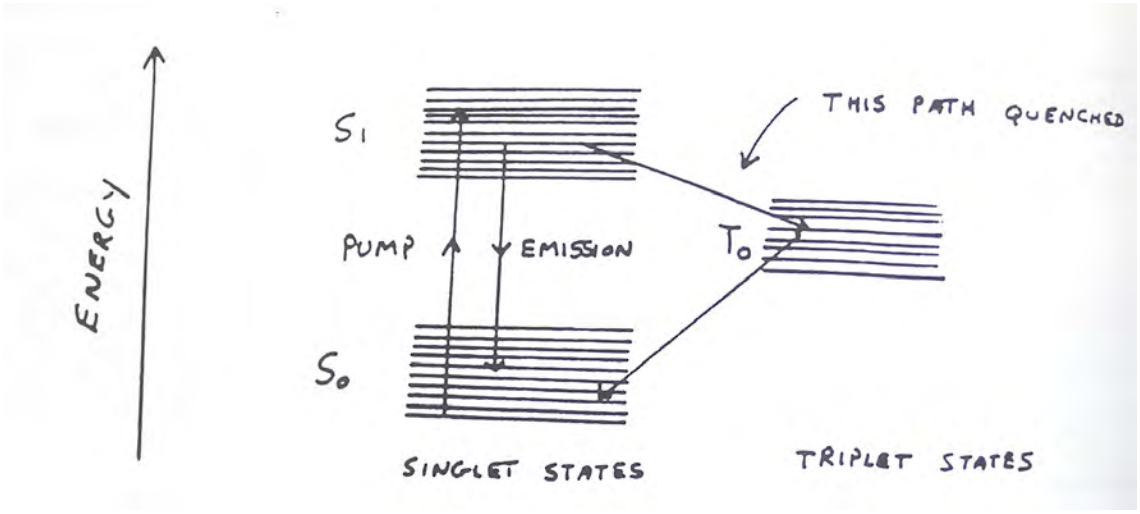


FIGURE 14.6.1. Energy level diagram of a dye molecule.

the laser is typically 100μm, so semiconductor lasers are incredibly small lasers. However, because of their direct electrical pumping mechanism, they are very efficient (typically greater than 20%) and can emit optical powers of the order of milliwatts.

Because of their direct pumping mechanism and their small size semiconductor diode lasers are used extensively in optical communications since they can be modulated electronically on a time scale of a nanosecond. This allows for information transfer rates of gigabytes per second. In these applications the output of the laser is coupled directly into the end of an optical fiber and the diffraction effects of the laser become unimportant.

Nowadays most of the research in diode lasers (and they represent one of the most active areas of laser research) is in developing new types of semiconductor alloys based on ternary and quaternary compounds, to yield high efficiency lasers operating at wavelengths corresponding to the minimum attenuation wavelength of optical fibers around 1.3 or 1.55 μm. Other research relates to developing pulsed lasers with pulses in the picosecond range.

14.6. Dye Lasers

Dye lasers are liquid lasers based on an active medium which contains a solution of highly fluorescent macromolecules in solvents like methyl alcohol or water. The types of dyes that are used are similar to the dyes that are used in the manufacture of clothing and fluorescent rulers. The dyes are labelled by such descriptive names as Rhodamine 6G, Coumarin and Cresyl Violet. The main point that we wish to make about dye lasers here, given the extensive discussion of the other laser systems is that dye lasers are tunable visible light sources. This tunability arises from the many degrees of freedom associated with the electronic, rotational, and vibrational co-ordinates of these molecules which are organic molecules with molecular weights in excess of 500. A typical energy level diagram for a dye molecule in solution is indicated in Figure 14.6.1. For each of the electronic levels indicated by S_0, S_1, etc. there are numerous vibrational and rotational levels. As a result when a dye molecule is pumped by a flashlamp or by another dye laser there are a large number of possible upper states an electron can go to. In this band of upper states the electrons find themselves in a population inversion relative to numerous unoccupied lower states, and indeed can proceed via spontaneous and eventually stimulated emission to a spectrum of lower states. Because of the large number of lower states, the emission spectrum can be as much as 200 Å wide.

It is the large fluorescent bandwidth that makes dye lasers so useful. By inserting a tuning element in the laser cavity such as a Fabry-Perot filter or, better yet, replacing the back mirror with a diffraction grating which behaves like a wavelength tunable mirror, one can select a particular wavelength of operation. With the hundreds of dyes that are available it is now possible to continuously cover the entire visible spectrum as well as portions of the infrared.

Dye lasers can be operated in both pulsed and continuous mode. Their output power is comparable to but less than that of the solid state insulator lasers such as ruby and YAG lasers discussed earlier.

14.7. Titanium Sapphire laser

(This site under construction). Vibronic laser, high bandwidth.

References

A.Yariv, *Quantum Electronics*, John Wiley, New York 1975.
A. Yariv, *Introduction to Optical Electronics*, Holt, Rinehart, Winston, New York 1976.
A.E. Siegman, *An Introduction to Lasers and Masers*, McGraw Hill, New York, 1971.

Index