1

Laser Light

Laser is an acronym for Light Amplification by Stimulated Emission of Radiation. The term light is used in a broad sense to include radiation at frequencies in the infra-red, visible or ultraviolet regions of the electromagnetic wave spectrum. In common parlance the term laser refers more to a device based on this principle than to the principle itself. The term laser action is often used when referring to the process. Lasers are devices that generate coherent light.

The physical principle (stimulated emission) responsible for laser action was introduced by Albert Einstein in 1916. A device called MASER (microwave amplification by stimulated emission of radiation) based on this principle was first operated in the microwave regime. The laser is an extension of this principle to the visible part of the electromagnetic spectrum.

A summary of principal developments in the field of laser follows.

1916: A. Einstein introduces stimulated emission as a fundamental process of light-matter interaction in addition to the already known processes of absorption and spontaneous emission of light.

1924: Richard Tolman discusses "negative absorption" i.e. amplification, and explains that the emitted radiation would be coherent with the input radiation.

928: Rudolph W. Landenburg confirms existence of stimulated emission.

V.A 1940: Fabrikant suggests method for producing population inversion in his PhD thesis. Population inversion is required for maser/laser operation.

1950: Alfred Kastler suggests a method of "optical pumping" for orientation of paramagnetic atoms or nuclei in the ground state. This was an important step on the way to the development of lasers for which Kastler received the 1966 Nobel Prize in Physics.

1951: Edward Purcell and Robert Pound observe inverted populations of states in a nuclear magnetic resonance experiment. Population inversions is a necessary condition for maser and laser action.

—bf 1952: Nikolay Basov and Alexander Prokhorov describe the principle of the maser (Microwave Amplification by Stimulated Emission of Radiation).

1954: C. H. Townes, J. P. Gordon, and H. J. Zeiger realize the first maser utilizing a beam of excited ammonia molecules to produce amplification of microwaves by stimulated emission at a frequency of 24 gigahertz (GHz).

1958: Charles H. Townes and Arthur L. Schawlow introduce concept of the laser.

1959: Gordon Gould introduces the term laser in a paper, "The LASER: Light Amplification by Stimulated Emission of Radiation"

1960: Laser action observed by T. H. Maiman in Ruby [Nature **187**, 493 (1960)]. It is now known to be one of the most difficult laser systems to operate. Sorokin and Stevenson develop first four-level solid-state laser at IBM. Ali Javan, William Bennett, and Donald Herriott at Bell Labs develop first helium neon (He:Ne) gas laser.

1961: Elias Snitzer reports the operation of a neodymium glass laser, currently the prime candidate as a laser source for fusion. In the first medical use of the laser, Charles Campbell and Charles Koester destroy a retinal tumor with the ruby laser. In the first example of efficient nonlinear optics, P. A. Franken, A. E. Hill, C. W. Peters and G. Weinreich demonstrate generation of second harmonic light by passing the pulses from a ruby laser through a quartz crystal, transforming red light into green.

1962: Scientists at Bell Labs report the first yttrium aluminum garnet (YAG) laser, which continues to dominate material processing applications. Scientists at General Electric, IBM, and MIT Lincoln Laboratory develop a gallium arsenide laser that converts electrical energy directly into infrared light. F. J. McClung and R. W. Hellwarth develop laser Q-switching technique to produces laser pulses of short duration and high peak powers. Four groups in the US (M. I. Nathan et al., R. N. Hall et al, T. M. Quist et al, N. Holonyak and S. F. Bevacqua) nearly simultaneously make first semiconductor diode lasers, which operate pulsed at liquid-nitrogen temperature. Semiconductor diode lasers are the first important step in the development of optical communication, optical storage, optical pumping of solid-state lasers and many other applications.

1963: L. E. Hargrove, R. L. Fork, and M. A. Pollack report the first mode-locked operation of a laser in a helium-neon laser with an acousto-optic modulator. Mode locking is the basis for the femtosecond pulsed laser. Herbert Kroemer and the team of Rudolf Kazarinov and Zhores Alferov independently propose ideas to build semiconductor lasers from heterostructure devices, which lead to their receiving the 2000 Nobel Prize in Physics. C. K. N. Patel develops first carbon dioxide laser at Bell Labs.

1964: C. H. Townes, N. G. Basov and A. M. Prokhorov awarded the Nobel prize for their fundamental work in Quantum Electronics; Townes for demonstrating the ammonia (NH_3) maser and subsequent work in masers and lasers and Basov and Prokhorov for contributing to the development masers and lasers. William B. Bridges develops first noble gas ion laser. J. E. Geusic, and H. M. Marcos, and L. G. Van Uitert develop neodymium-doped yttrium

aluminum garnet (Nd: YAG) laser. This is the most widely used solid state laser; from cutting and welding to medical applications and nonlinear optics. C. J. Koester and E. Snitzer develop neodymium-doped fiber amplification. Fiber amplifiers are used in communication and for high power lasers. Arno Penzias and Robert Wilson use maser amplifier to observe 3K cosmic background radiation proving the existence of the Big Bang. They are awarded the Nobel Prize in Physics in 1978.

1965: George C. Pimentel and Jerome V. V. Kasper demonstrate the first chemical laser. With output currently reaching megawatt levels, chemical lasers get their energy from chemical reactions and are some of the most powerful lasers in the world. James Russell invents the laser compact disk (CD player). Anthony J.DeMaria, D. A. Stetser, and H. A. Heynau report the first generation of picosecond laser pulses using a neodymium glass laser and a saturable absorber.

1966: Peter Sorokin and John R. Lankard built the first widely tunable organic dye laser, now used in ultrafast science and spectroscopy. Charles K. Kao and George Hockham of Standard Telecommunications Laboratories in England publish landmark paper demonstrating that optical fiber can transmit laser signals and reduce loss if the glass strands are pure enough. Alfred Kastler is awarded the Nobel Prize in Physics "for the discovery and development of optical methods for studying Hertzian resonances in atoms".

1968: NASA launches the first satellite equipped with a laser.

1969: Led on Earth by American physicist Carroll Alley and using retroreflectors placed on the moon by Neil Armstrong and Buzz Aldrin, NASA's Lunar Laser Ranging experiments begin. Using these mirrors, scientists on Earth bounce lasers off the moon, measuring its orbital motions, and in the process determining fundamental gravitational and relativistic constants with extraordinary precision. D. J. Spencer, T. A. Jacobs, H. Mirels, and R. W. F. Gross develop the first continuous-wave chemical laser. High power chemical lasers generate megawatts of power, leading to proposals for laser weapons. The pulsed dye laser is invented.

1970: Nikolai Basov, V. A. Danilychev, and Yu. M. Popov of Lebedev Physical Institute in Moscow develop the excimer lasers, which are important in photolithography and laser eye surgery. Zhores Alferov's group at the Ioffe Physical Institute and Mort Panish and Izuo Hayashi at Bell Labs produce the first continuous-wave room-temperature semiconductor lasers, paving the way toward commercialization of fiber optics communications. The world's first laser-driven lighthouse opens in Australia (Point Danger). Robert Maurer, Peter Schultz and Donald Keck at Corning Glass Work prepare the first batch of optical fiber hundreds of yards long capable of carrying optical signal over it. J. Beaulieu invents transversely excited atmospheric (TEA) pressure CO2 laser useful for the machining industry. O. G. Peterson, S. A. Tuccio,

and B. B. Snavely develop CW dye laser leading to a revolution in spectroscopy and ultrafast science. Arthur Ashkin demonstrates the use of laser beams to manipulate microparticles pioneering the field of optical tweezing and trapping, leading to important advances in physics and biology.

1971: Dennis Gabor was awarded the Nobel prize in physics for his invention and development of the holographic method.

1974: The first product logged in a grocery store by a barcode scanner. E. P. Ippen and C. V. Shank develop the sub-picosecond mode-locked CW dye laser, establishing ultrafast optical science.

1975: Laser Diode Labs develops first commercial continuous-wave semiconductor laser operating at room temperature. Continuous-wave operation allows the transmission of telephone conversations.

1976: John Madey and group at Stanford University demonstrate the first free electron laser (FEL). Instead of a gain medium, FELs use a beam of electrons accelerated to near light speed, then passed through a series of alternating magnetic fields. The forced undulating motion results in the release of a coherent photon beams with widest tunable frequency range of any laser type due to the tunable magnetic field.

1977: General Telephone and Electronics send first live telephone traffic through fiber optics, 6 Mbit/s in Long Beach CA.

1981: N. Bloembergen and Arthur Schawlow were awarded the Nobel Prize in physics for their contributions to masers, nonlinear optics and spectroscopy.

1982: Kanti Jain publishes the first paper on excimer laser lithography used extensively today to make microchips for the computer and electronics industry. P. F. Moulton develops titanium-sapphire laser, which has nearly replaced the dye laser for tunable and ultrafast laser applications.

1985: Steven Chu, Claude Cohen-Tannoudji, and William D. Phillips develop methods to cool and trap atoms with laser light. Their research helps to study fundamental phenomena and measure important physical quantities with unprecedented precision. They are awarded the Nobel Prize in Physics in 1997. Grard Mourou and Donna Strickland demonstrate chirped pulse amplification or CPA. They used gratings to lengthen laser pulses before amplification, and then again after the amplification to shorten them to their original length. This permits much higher powers without damaging the amplifying material itself. CPA was later used to create ultrashort, very high-intensity (petawatt) laser pulses.

1987: Ophthalmologist Steven Trokel performs the first laser eye surgery using an excimer laser. Emmanuel Desurvire, David Payne, and P.J. Mears demonstrate optical amplifiers that are built into the fiber-optic cable itself.

1988: Samuel Blum, Rangaswamy Srinivasan, and James Wynne observed the effect of the ultraviolet excimer laser on biological materials. Further investigations revealed that the laser made clean, precise cuts ideal for delicate surgeries. First transatlantic fiber cable is laid with glass so transparent that amplifiers are only needed about every 40 miles. Double clad fiber laser developed by E. Snitzer, H. Po, F. Hakimi, R. Tumminelli, and B. C. McCollum. These high power solid-state lasers are used for machining.

1989: Norman F. Ramsey was awarded the Nobel Prize for the invention of the separated oscillatory fields method and its use in the hydrogen maser and other atomic clocks and Hans G. Dehmelt and Wolfgang Paul for the development of the ion trap technique.

1992: Eric Betzig, Ray Wolfe, Mike Gyorgy, Jay Trautman, and Pat Flynn develop a magneto-optic data storage technique that can squeeze 45 billion bits of data into a square-inch of disk space.

1994: First proposed in 1971 by Rudy Kazarinov and Robert Suris, the first quantum cascade laser was demonstrated by Jerome Faist, Federico Capasso, Deborah Sivco, Carlo Sirtori, Albert Hutchinson, and Alfred Cho of Bell Labs.

1996: S. Nakamura and coworkers develop GaN (Gallium nitride) and InGaN (Indium gallium nitride) semiconductor lasers.

1997: Steven Chu, Claude Cohen-Tannoudji and William D. Phillips awarded the NObel Prize for the development of methods to cool and trap atoms with laser light. Researchers at MIT create the first atom laser.

2000: Zhores I. Alferov and Herbert Kroemer are awarded the Nobel Prize in Physics "for basic work on information and communication technology" and "for developing semiconductor heterostructures used in high-speed- and opto-electronics". John Hall and Theodor Hansch develop optical frequency comb technique used in research as well as in precision metrology and time measurement. This work leads to their receiving the 2005 Nobel Prize in Physics.

2001: Eric A. Cornell, Wolfgang Ketterle and Carl E. Wieman were awarded the Noble Prize for the achievement of Bose-Einstein condensation in dilute gases of alkali atoms, and for early fundamental studies of the properties of the condensates.

2005: J. L. Hall and T. W. Hänsch were awarded the Nobel prize for their contributions to the development of laser-based precision spectroscopy, including the optical frequency comb technique and to Roy Glauber for his contribution to the quantum theory of optical coherence. INTEL creates a chip containing eight continuous Raman lasers by using fairly standard silicon processes rather than the somewhat expensive materials and processes required for making lasers today.

2009: Charles Kao was awarded the Nobel Prize for physics for his work in fiber optics along with Willard S. Boyle and George E. Smith of Bell Labs who developed the CCD (charge coupled device) which made digital photography possible. The SLAC Linac Coherent Light Source produces the first ever coherent, hard X-ray beam. The ultrafast laser pulses are powerful enough to make images of single molecules or atoms in motion.

There are three essential elements of a laser:

1. A gain or amplifying medium consisting of atoms, molecules, ions, or charged carriers along with a pumping mechanism to excite these species to their higher quantum mechanical energy states. The energy stored in the excitation can be emitted as light spontaneously or stimulated by pre-existing light leading to an amplification of light energy.

2. A suitable arrangement of optical elements (lenses, mirrors, prisms, etc.) or some other mechanism to allow multiple passage of light through the gain medium (feedback).

3. A loss mechanism to extract light energy from the device. In addition to this "desirable" or essential loss, nonessential but unavoidable losses of light energy due to absorption, diffraction, scattering, and transmission through mirrors and other optical elelments are also present.

These three elements come in a great variety of forms and shapes and provide the basis for classifying lasers.

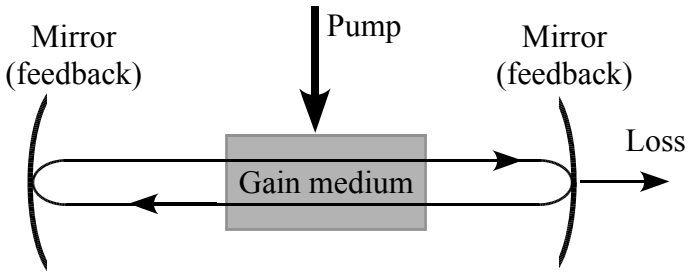


FIGURE 1.1

Basic elements of a laser.

1.1 Survey of Laser Elements

A wide variety of laser gain media and pumps are used to generate radiation ranging from far infra-red (far-IR) to soft X-rays.

1.1.1 Gain Media

Important laser gain media and wavelengths include,

HCN far-IR laser (311, 337, 545, 676, 744 μm)

H_2O far-IR laser (28,48, 120 μm)

CO_2 laser (9.6-10.6 μm)

CO laser (5.1-6.5 μm)

HF chemical laser (2.7-3.0 μm)

Nd:YAG laser (1.06 μm)

He:Ne laser (1.15 μm, 633 nm)

Ga-As semiconductor laser (870 nm)

Ruby laser (694 nm)

Rhodamine 6G dye laser (560-640 nm)

Argon-ion laser (488-514 nm)

Pulsed N_2 discharge laser (337 nm)

Pulsed H_2 discharge laser (160 nm)

1.1.1.1 Laser Pump

Depending on the gain medium, many different types of pump mechanisms are used to supply energy to the gain media. These include

Gas discharge including dc, radio-frequency, and pulsed electrical discharges involving both direct electron excitation and two-stage collision pumping.

Optical pumping using flash lamps, arc lamps, semiconductor LEDs (light emitting diodes), other lasers and even direct sunlight.

Chemical reactions including chemical mixing, photolysis, and combustion.

Direct electrical pumping includes high-voltage electron beams directed into high-pressure gas cells and direct current injection into semiconductor injection lasers.

Nuclear pumping of gases by nuclear fission fragments when a gas laser tube is placed in close proximity of a nuclear reactor.

Supersonic expansion of gases, usually pre-heated by chemical reaction or electrical discharge, through supersonic expansion nozzles, to create the so-called gas-dynamic lasers.

Plasma pumping in hot dense plasmas, created by plasma pinches, focused high-power laser pulses, or electrical pulses. There are reports of X-ray laser action in some laser materials pumped by the explosion of a nuclear bomb.

1.1.2 Optical feedback

The principle mechanism for providing feedback in lasers involves optical cavities or resonators. Resonators store electromagnetic energy. At microwave frequencies these resonators are closed metallic boxes but at optical frequencies they can be open. The need for open resonators arises because we want only a few modes interacting with the atoms to grow[1]. The presence of boundaries in resonators (boundary conditions) allows only certain field configurations (frequency and spatial variation) to exist inside the resonators. These allowed field configurations are called modes of the resonator. For open resonators modes corresponding to propagation away from the resonator axis will have very large losses because any light emitted into such modes will be quickly lost. Only a group of paraxial modes where energy is localized near the axis will experience build up. Even these modes will experience losses due to diffraction, absorption, and transmission by the end mirrors. As a result mode definition in the sense of stationary field configurations can not be used. However, modes with quasi-stationary (long lived) patterns do exist. These modes experience very little loss of light energy in one cavity round trip so that any energy emitted into these modes will remain in the cavity for a long time.

1.1.3 Losses

Mechanisms that lead to a loss of light energy stored in the cavity include
Diffraction by optical elements inside the resonator,
Absorption inside the gain medium and mirrors, coating,
Scattering, and
Transmission through the mirrors.

Although the loss is detrimental to achieving laser action, not all loss is undesirable. In fact the loss of light energy stored in the cavity via mirror transmission is what emerges as the beam of light that has made the laser into such a useful tool that hardly any aspect of our life has remain untouched by it.

[1] At optical frequencies the density (# per unit frequency band per unit volume of the resonator) of modes near a frequency ν is very large $8\pi\nu^2/c^3$. If the atoms interacts with modes in a band $\delta\nu$ (typically 10^9 Hz or more), their number $8\pi\nu^2\delta\nu/c^3$ is enormous. Such a situation is not conducive to the growth of any mode amplitude to a significant level.

A coupling of gain, loss, and feedback mechanisms via the electromagnetic field makes laser action possible and imparts to the light generated by lasers certain extreme characteristics. The long lived modes interact with the gain medium and extract energy from it. The gain medium can either emit light spontaneously into any of the modes - long or short lived - or it can be stimulated to emit into a particular mode by the light energy stored in that mode. Stimulated emission of light into a particular mode increases as the light energy of the mode increases. Clearly the light energy in long lived modes is more likely to be amplified provided they can overcome their energy losses. Thus it is the competition between gain due to stimulated emission and loss due to diffraction, transmission, absorption, etc. that determines whether light amplification by stimulated emission of radiation can take place or not. If gain exceeds loss, laser action can occur.

An understanding of lasers will require us to study an interacting matter and light system inside a resonator. It is an open system as energy can be added to or extracted from the system. This may seem like a formidable problem. We will see, however, that this seemingly complex problem can be dealt with quite effectively.

We will first study atoms and light separately by ignoring their interaction and then couple the two. The atoms are described by Schrödinger equation. Light, being an electromagnetic wave phenomenon, is described by Maxwell's equations. Its propagation as rays or waves, diffraction, and interference, all follow from Maxwell's equations. These equations admit wave-like solutions. Plane waves are the best known of these solutions. For lasers we require beam like wave solutions which, like plane waves, have a pre-dominant direction of propagation but they have finite extent in directions perpendicular to the direction of propagation. In view of the remarks in the preceding paragraph one might suspect that these solutions are the right type to fit the boundary conditions imposed by open optical resonators. We will see that this is indeed the case.

1.2 Laser Light Characteristics

The light output from a laser is electromagnetic radiation and is not fundamentally different from the light emitted by other sources of electromagnetic radiation. There are, however, several important differences in detail between laser light and the light emitted by thermal sources. The output beam produced by lasers have much more in common with the output of conventional low-frequency electronic oscillators than they do with any kind of thermal light sources. We will briefly review laser beam characteristics that distinguish them from other light sources. The numbers stated below for lasers should not be taken as the final word. They change as progress is made in improving the performance of lasers.

1.2.1 Monochromaticity

The light emitted by a laser has high degree of spectral purity. This means it has relatively well-defined frequency or wavelength so that the output signal from an ideal laser is very nearly a constant amplitude sinusoidal wave. Two factors contribute to the spectral purity of laser light. First, light emission from atoms occurs in a narrow range of frequencies around atomic transition frequency $\nu_o = (E_2 - E_1)/h$. This range of frequencies defines the atomic linewidth. Consequently only EM waves with frequencies close to the atomic transition frequency can interact strongly with the atoms and be amplified. Second, the laser cavity forms a resonant structure.

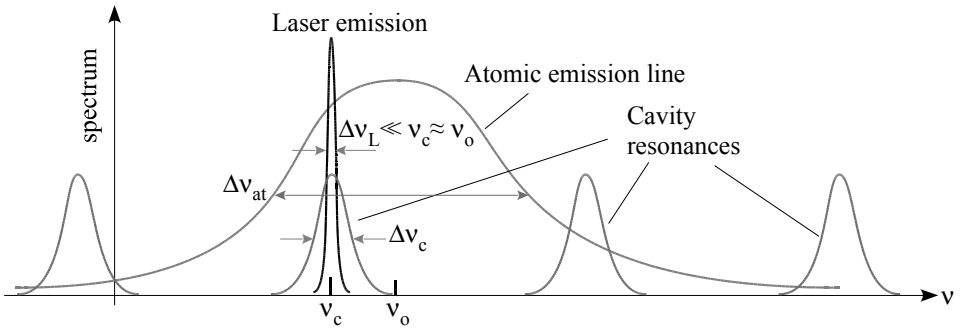


FIGURE 1.2
Relation between atomic emission, cavity resonance and laser light spectral profiles. Linewidth $\Delta\nu$ is a measure of the range of frequencies (usually full width at half maximum or FWHM) for significant response.

This means light feedback inside the cavity can occur only for a small range of frequencies close to certain characteristic frequencies called resonance frequencies ν_c of the cavity. This narrow range of frequencies around a cavity resonance frequency define the cavity linewidth. Thus the amplification of light due to stimulated emission occurs at cavity resonance frequencies that fall within an atomic linewidth of the atomic transition frequency. This amplification leads to a further narrowing of the laser linewidth by a factor that can be as large as 10^{10}. The primary reason for the spectral purity of laser light is the optical resonator (optical cavity) and not the atomic transition. The laser cavity permits continuous oscillations only at certain discrete frequencies (cavity resonance frequencies). The atoms serve primarily to provide gain at these resonance frequencies.

The spectral width of a *single-frequency* laser ranges from $\Delta\nu_L = 10^6$ Hz for moderately stabilized lasers to $\Delta\nu_L = 10$ Hz or less in well stabilized lasers. Since visible light has frequencies of order $\nu = 5 \times 10^{14}$ Hz, the spectral purity of laser light is $\Delta\nu_L/\nu_L = 2 \times 10^{-13}$. The ratio $Q = \frac{\nu_L}{\Delta\nu_L}$ is a measure of the quality factor (Q-factor) of the laser oscillator. Such large values of Q are difficult to achieve in mechanical or electronic oscillators. Thermal sources such as the sun

and incandescent solids generally emit a broadband spectrum of light. There are, however, some thermal sources such as discharge lamps, that emit only a few spectral lines or narrow bands of wavelengths, but the spectral widths of the light emitted by even the best such sources are still limited by the linewidths of the atomic transitions in the discharge which range from $10^8 - 10^{11}$ Hz. Table (1.1) shows a comparison of the spectral purity of light emitted by different sources of light [FWHM=Full Width at Half Maximum].

TABLE 1.1
A comparison of spectral purity of different light sources

Light Source	Peak Wavelength	FWHM ($\Delta\lambda$)	FWHM ($\Delta\nu$)
He:Ne Laser	633 nm	10^{-8} nm	7.5×10^3 Hz
Cadmium low pressure lamp	644 nm	10^{-3} nm	9.4×10^8 Hz
Sodium discharge lamp	590 nm	0.1 nm	9×10^{10} Hz
Blackbody radiator at 5800 K	500 nm	600 nm	10^{14} Hz

The ultimate limit on laser spectral purity is set by quantum noise fluctuations due to spontaneous emission from the atoms in the gain medium. This limit, however, can be reached with great difficulty only on the very best and highly stabilized lasers.

1.2.2 Coherence

The amplitude and the phase of the sine wave from a laser oscillator will in fact change slowly over space and time. Coherence refers to the property of a light wave that its phase and amplitude at one space-time point may be correlated to its phase and amplitude at some other space and time point. There are two types of coherences:

Temporal coherence refers to strong correlations between the amplitude and/or phase of the signal at different times. Laser light has high degree of temporal coherence. The amplitude and phase of the output sine wave at any one instance are strongly correlated with their values at some other instances over a time interval τ_c that spans millions of optical cycles. For stationary beams, the time interval τ_c, called coherence time, is of the order

$$\tau_c \approx \frac{1}{2\pi\Delta\nu} \tag{1.1}$$

The distance traversed by light in one coherence time τ_c is called the (longitudinal) coherence length

$$\ell_c = c\tau_c\,. \tag{1.2}$$

Coherence time τ_c may be thought of as the interval over which the amplitude and phase of the sine wave may be considered constant. Longitudinal coherence ℓ_c may

then be considered as the average length of perfect sine waves emitted by the laser. For lasers ℓ_c can range from 300 m to 10^4 m or even larger.

Source	τ_c (s)	ℓ_c (m)
He:Ne laser	$10^{-4} - 10^2$	$3 \times 10^4 - 3 \times 10^{10}$
Sodium lamp	10^{-10}	3×10^{-2}

Spatial coherence

refers to correlation between laser field at different points in a plane transverse to the direction of wave propagation. At the output of a laser oscillator, laser light has almost perfect transverse spatial coherence. For thermal sources transeverse spatial coherence does not extend over distances much large than a few wavelengths. Transverse spatial coherence of laser light is also a consequence of the presence of a laser cavity.

1.2.3 Directionality or Collimation

Thermal sources emit light in random directions over a broad wavelength range. We can capture some fraction of this radiation and collimate it with a lens or mirror as in a searchlight or flashlight. The resulting degree of collimation (amount of radiation emitted per unit solid angle) is still much smaller than that for a laser. Consequently, thermal beams spread very rapidly with propagation.

A single-transverse mode laser oscillator, on the other hand, can produce a beam that can propagate for sizable distances with very little diffraction spread. The angular divergence (half angle) in radians of a laser beam in the far zone (far from beam waist) is given by

$$\theta \approx \frac{\lambda}{\pi w_o} \tag{1.3}$$

where w_o is the radius of laser beam spot at its waist (location where the beam has narrowest transverse size). The distance over which this beam stays approximately collimated before diffraction spreading significantly increases is given by

$$b \approx \frac{2\pi w_o^2}{\lambda}\,. \tag{1.4}$$

This is a direct consequence of the fact that laser beam comes from a resonant cavity where only the rays propagating close to the cavity axis can pass through the gain medium multiple times and thus grow in energy.

For laser light of wavelength $\lambda = 1.06 \times 10^{-3}$ mm, $w_0 = 3$ mm,

$$\theta = \frac{\lambda}{\pi w_o} = \frac{1.06 \times 10^{-6}}{\pi 3 \times 10^{-3}} = 1.1 \times 10^{-4} \text{ rad} = 0.006^o\,.$$

For a small Helium Neon (He:Ne) laser emitting at $\lambda = 633$ nm, the beam waist might be $w_0 \approx 0.5$ mm. This corresponds to an angular divergence of

$$\theta \approx \frac{0.633 \times 10^{-6}}{\pi 0.5 \times 10^{-3}} = 4 \times 10^{-4} \text{ radians} = 0.02^o$$

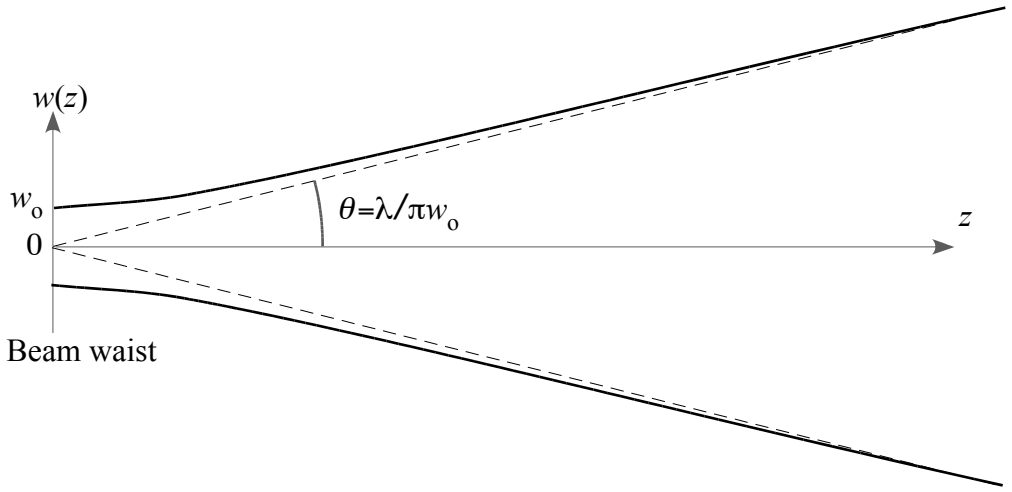


FIGURE 1.3
Divergence of a laser beam from its waist.

and a collimation distance of $b = 2.3$ m. For an Argon-ion (Ar-ion) laser at 514 nm, the waist might be 5 mm corresponding to angular divergence of $\theta = 3 \times 10^{-5}$ and a collimation distance of $b = 310$ m.

Comparing these values to a normal flashlight for which the divergence is about 25^o or a searchlight that has a typical divergence angle of 10^o, the high directionality of laser light is obvious.

1.2.4 Laser Beam Focusing

A laser beam can also be focused by a lens to a small spot only a few laser wavelengths in diameter. The diameter of the focused spot is given by the formula

$$w \approx \frac{\lambda}{\pi w_o} \times f \tag{1.5}$$

where f is the lens focal length. If the laser beam fills the lens aperture, the ratio f/w_0 is simply the f-number of the lens. For best lenses this number is of order one. It follows that a laser beam can be focused to spots which are only a few wavelengths in diameter.

The directionality of laser beam is also a consequence of the presence of a resonator.

1.2.5 Brightness

Brightness B of a source is defined as the power efflux (power emitted per unit area of the emitting surface per unit solid angle). Its units are W/m^2·sr. Spectral B_ν brightness is power emitted per unit area of the source per unit solid angle per unit bandwidth. Its units are W/m^2·sr·Hz. The idea of brightness can be understood

by considering a source that emits through a surface area $\mathcal{S}$. Each area element can emit light into a solid angle 2π steradian. Then, if a surface element $\delta\mathcal{S}$ emits power δP into a solid angle $\delta\Omega$, the brightness of the source is given by

$$B = \frac{\delta^2 P}{\delta\mathcal{S}\delta\Omega}\,, \qquad [B] = \mathrm{W/m^2{\cdot}sr}\,.$$

The power emitted by a black body at temperature T is given by Stefan-Boltzmann

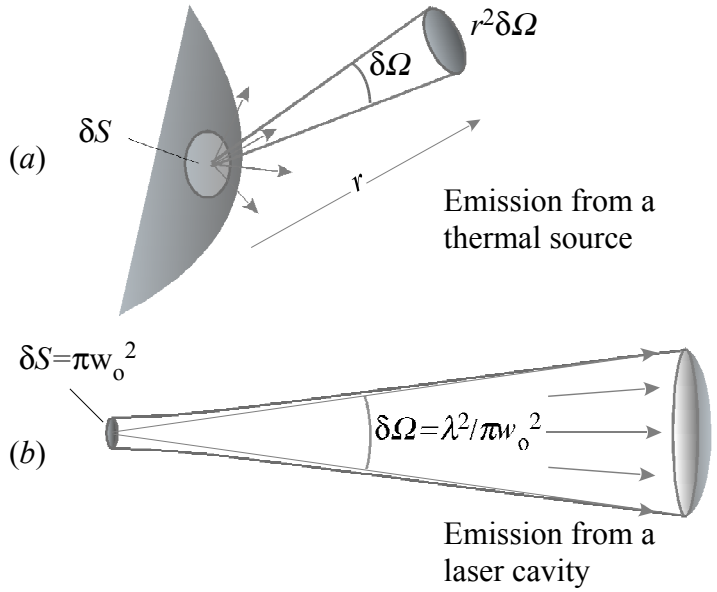


FIGURE 1.4
(a) Emission from the surface of a thermal source. (b) Emission from a laser cavity.

law to be

$$I = \sigma T^4\,, \qquad [I] = \mathrm{W/m^2}$$

For the sun with T=6000 K this gives an efflux of

$$I_{\mathrm{sun}} = 5.6705 \times 10^{-8}(6 \times 10^3)^4 = 7 \times 10^7\,\mathrm{W/m^2}$$

A small He:Ne laser of modest power $P = 1$ mW and beam waist of 0.5 mm will produce an efflux

$$I_{\mathrm{laser}} = \frac{P}{\pi w_o^2} = \frac{10^{-3}}{\pi(0.5 \times 10^{-3})^2} = 1.3 \times 10^3\,\mathrm{W/m^2}$$

This looks even more impressive when we take into account the directionality. The efflux from the sun is emitted isotropically into a solid angle of 2π steradian (sr)

whereas, because of the directionality of the laser beam, the laser emits its power into a solid angle $\pi\theta^2 = \pi(\lambda/\pi w_o)^2 = \lambda^2/\pi w_o^2$. This leads to brightness

$$B = \frac{P}{\pi w_0^2 \pi\theta^2} = \frac{P}{\lambda^2} = \frac{10^{-3}}{(0.6328\times 10^{-6})^2} = 2.5\times 10^8\,\mathrm{W/m^2{\cdot}sr}\,.$$

This exceeds the brightness of the sun. This comparison looks even more impressive if we examine the spectral brightness. The sun radiates like a blackbody over wide spectral bandwidth. Let us compare the spectral brightness of the sun near the peak (yellow) of visible spectrum. The spectral energy density [energy per unit volume per unit bandwidth (J/m^3·Hz)] of a black-body radiator is given by

$$\rho(\nu) = \frac{8\pi\nu^2}{c^3}\frac{h\nu}{e^{\beta h\nu}-1}\,,\quad \beta = \frac{1}{k_{_B}T}\,. \tag{1.6}$$

The spectral intensity then is $\frac{1}{2}c\rho(\nu)$ (W/m^2·Hz), where the factor of half accounts for the fact that only half of the radiation is propagating outward toward the radiating surface of the the black body. Since each surface element radiates isotropically (into and out of the source) only the radiation into the external solid angle $\delta\Omega = 2\pi$ escapes the black body. Hence the spectral brightness of a black body radiator is given by

$$B_\nu = \frac{1}{2}\frac{\rho(\nu)c}{\delta\Omega} = \frac{1}{2}\left[\frac{8\pi\nu^2}{c^3}\frac{h\nu}{e^{\beta h\nu}-1}\right]\frac{c}{2\pi} = \frac{2\nu^2}{c^2}\left[\frac{h\nu}{e^{\beta h\nu}-1}\right]\,.$$

For the sun, which radiates as a black body at a temperature of $T \approx 6000$ K, spectral brightness at a wavelength in the yellow region [$h\nu = 2.5$ eV] we have

$$\begin{aligned}
k_{_B}T &= k_{_B}\cdot 300\frac{T}{300} = \frac{1}{40}\times 20 \approx \frac{1}{2}\,\mathrm{eV}\\
\frac{h\nu}{k_{_B}T} &\approx 5\,,\quad e^{\beta h\nu} = e^5 \approx 150\\
B_\nu &= \frac{2}{(633\times 10^{-9})^2}\frac{2.5\times 1.6\times 10^{-19}}{150-1} \approx 2\times 10^{-8}\,\mathrm{W/m^2\text{-}sr\text{-}Hz}\\
&= 2\times 10^{-12}\,\mathrm{W/cm^2{\cdot}sr{\cdot}Hz}\,.
\end{aligned}$$

For a l mW He:Ne laser ($\lambda = 633$ nm) of spectral width $\Delta\nu = 10^4$ Hz and spot size 0.5 mm, the spectral brightness will be

$$\begin{aligned}
B_\nu &= \frac{P}{\lambda^2\Delta\nu} = \frac{1\times 10^{-3}}{(0.633\times 10^{-6})^2\times 10^4} \approx 2.5\times 10^5\,\mathrm{W/m^2{\cdot}sr{\cdot}Hz}\\
&= 25\,\mathrm{W/cm^2{\cdot}sr{\cdot}Hz}
\end{aligned}$$

For a Neodymium glass (Nd-glass) laser with a power of $P = 10^4$ MW, $\lambda = 1.06\,\mu$, and bandwidth limited by pulse duration of 30 ps, we have

$$\begin{aligned}
\Delta\nu &= \frac{1}{2\pi\tau_p} = \frac{1}{2\pi\times 3\times 10^{-11}} \approx 5\times 10^9\,\mathrm{Hz}\\
B_\nu &= \frac{1\times 10^4\times 10^6}{(10^{-6})^2\times 5\times 10^9} \approx 2\times 10^{12}\,\mathrm{W/m^2.sr.Hz}\\
&= 2\times 10^8\,\mathrm{W/cm^2.sr.Hz}
\end{aligned}$$

1.2.6 Laser Performance Records

1. Wide power range: Continuous wave (CW) powers of up to hundreds of kilowatts are available from certain IR chemical lasers. In pulsed mode peak powers in excess of 10^{13} watts, which exceeds the total electrical power generated in the U.S. for very short times (pico-seconds) are available.

2. Extreme frequency stability: The short term frequency stability of a highly stabilized laser can be as good as one part in10^{13}. A He:Ne laser operating at 3.39 μm stabilized against a methane absorption line has absolute reproducibility of 1 part in 10^{10}. Frequency stabilities of $\Delta\nu = 10^{-2}$ Hz have been achieved.

3. Wide tunability: Most common lasers are limited to sharply defined discrete frequencies. However, widely tunable sources of coherent radiation including dye lasers, titanium-doped sapphire (Ti-sapphire) lasers and optical parametric oscillators (OPOs) can provide tunable radiation up to bandwidths of order $\Delta\lambda \approx 300$ nm. This corresponds to a frequency bandwidth of $\Delta\nu \approx 10^{13} - 10^{14}$ Hz.

4. Ultra-short pulses : Mode-locked laser pulses shorter than 1 ps (picosecond) are routine. Mode-locked compressed dye laser pulses of only a few femtosecond long (FWHM) (few optical cycles) have been produced.

5. Very efficient: Power conversion efficiency is defined to be the ratio of optical power radiated by the laser to the power supplied to operate the laser. Power conversion efficiencies of lasers range from 0.001 to 0.1% for gas lasers, 1 to 2 % for solid state lasers and 50-70% for carbon dioxide (CO_2) and semiconductor lasers.

1.3 Electromagnetic Waves in Homogeneous Media

We will use the complex analytic representation of the fields so that the real physical fields are given by

$$F(\boldsymbol{r}, t) = \mathcal{R}e\mathcal{F}(\boldsymbol{r}, t)\,, \tag{1.7}$$

where $F(\boldsymbol{r}, t)$ represents any of the components of the fields. In a homogeneous transparent isotropic medium characterized by dielectric permittivity ϵ and magnetic permeability μ the constitutive relations are simple proportionalities $\boldsymbol{\mathcal{D}}(\boldsymbol{r}, t) = \epsilon\boldsymbol{\mathcal{E}}(\boldsymbol{r}, t)$ and $\boldsymbol{\mathcal{B}}(\boldsymbol{r}, t) = \mu\boldsymbol{\mathcal{H}}(\boldsymbol{r}, t)$. These relations hold for arbitrarily rapid variation of the fields as long as the most significant part of the field spectrum lies in the transparency range of the medium (ϵ, μ real).

In the presence of atoms or molecules that interact strongly with the field the constitutive relations are modified to read

$$\boldsymbol{\mathcal{D}}(\boldsymbol{r},t) = \epsilon\boldsymbol{\mathcal{E}}(\boldsymbol{r},t) + \boldsymbol{\mathcal{P}}_{at}(\boldsymbol{r},t)\,, \tag{1.8a}$$

$$\boldsymbol{\mathcal{B}}(\boldsymbol{r},t) = \mu\boldsymbol{\mathcal{H}}(\boldsymbol{r},t)\,, \tag{1.8b}$$

$$\boldsymbol{\mathcal{J}}(\boldsymbol{r},t) = \sigma\boldsymbol{\mathcal{E}}(\boldsymbol{r},t)\,, \tag{1.8c}$$

where $\boldsymbol{\mathcal{J}}$ represents the conduction current density and σ is the conductivity of the medium representing dissipation of electromagnetic energy. In writing these relations we have assumed that the magnetic response of the atoms is negligible compared to their electrical response (magnetization $\boldsymbol{\mathcal{M}}_{at} = 0$), which holds for most atomic media. We have also assumed that the fields are quasi-monochromatic[2]. Maxwell's equations then read

$$\boldsymbol{\nabla}\cdot\boldsymbol{\mathcal{D}}(\boldsymbol{r},t) = 0\,, \tag{1.9a}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{\mathcal{B}}(\boldsymbol{r},t) = 0\,, \tag{1.9b}$$

$$\boldsymbol{\nabla}\times\boldsymbol{\mathcal{E}}(\boldsymbol{r},t) = -\frac{\partial\boldsymbol{\mathcal{B}}(\boldsymbol{r},t)}{\partial t}\,, \tag{1.9c}$$

$$\boldsymbol{\nabla}\times\boldsymbol{\mathcal{B}}(\boldsymbol{r},t) = \mu\boldsymbol{\mathcal{J}}(\boldsymbol{r},t) + \mu\frac{\partial\boldsymbol{\mathcal{D}}(\boldsymbol{r},t)}{\partial t}\,. \tag{1.9d}$$

These are coupled first-order partial differential equations. By eliminating the magnetic (electric) field we can obtain a closed equation for the electric (magnetic) field. For example, on taking the curl of Eq. (1.9c) we obtain

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{\mathcal{E}}) = -\frac{\partial}{\partial t}\boldsymbol{\nabla}\times\boldsymbol{\mathcal{B}}\,. \tag{1.10a}$$

Using the identity $\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{\mathcal{E}}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{\mathcal{E}}) - \nabla^2\boldsymbol{\mathcal{E}}$ together with $\boldsymbol{\nabla}\cdot\boldsymbol{\mathcal{E}} = 0$ and eliminating $\boldsymbol{\mathcal{B}}$ with the help of Eq. (1.9d) and the constitutive relations, we find that the electric field satisfies the equation

$$\nabla^2\boldsymbol{\mathcal{E}} - \mu\sigma\frac{\partial\boldsymbol{\mathcal{E}}}{\partial t} - \mu\epsilon\frac{\partial^2\boldsymbol{\mathcal{E}}}{\partial t^2} = \mu_0\frac{\partial^2\boldsymbol{\mathcal{P}}_{at}}{\partial t^2}\,. \tag{1.10}$$

This is driven damped vector wave equation. The second term represents the damping of the wave amplitude due to loss of electromagnetic energy into heat. The term on the right hand side represents the source term involving atomic polarization which is the source of the electromagnetic wave. For now consider lossless ($\sigma = 0$) and source free region ($\boldsymbol{\mathcal{P}}_{at} = 0$), where the equation satisfied by the field becomes

$$\left[\nabla^2 - \mu\epsilon\frac{\partial^2}{\partial t^2}\right]\boldsymbol{\mathcal{E}}(\boldsymbol{r},t) = 0. \tag{1.11}$$

[2]For quasi-monochromatic fields we can write $\boldsymbol{\mathcal{F}}(\boldsymbol{r},t) = \boldsymbol{\mathcal{F}}_o(\boldsymbol{r},t)e^{-i\omega t}$, where the *envelope* $\boldsymbol{\mathcal{F}}_o(\boldsymbol{r},t)$ changes negligibly in times of order $2\pi/\omega$. In terms of Fourier decomposition of the fields, we can say that the field contains frequencies in a narrow band $\Delta\nu$ centered around the carrier frequency ν ($\Delta\nu \ll \nu$). Parameters $\epsilon and \mu$ can then be taken to represent the values of dielectric permittivity and magnetic permeability at the carrier frequency ω.

It is easily checked that in this case, the magnetic field $\boldsymbol{\mathcal{B}}$ also satisfies the same equation.

The quantity $1/\mu\epsilon$ has the dimensions of square of a speed. This speed v is given by

$$v = \sqrt{\frac{1}{\mu\epsilon}} = \sqrt{\frac{1}{\mu_0\epsilon_0}}\sqrt{\frac{\mu_0\epsilon_0}{\mu\epsilon}} = \frac{c}{n}\,, \tag{1.12}$$

where c and n are, respectively, the speed of light in free space and the refractive index of the medium,

$$c = 2.99792458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}. \tag{1.13}$$

$$n = \sqrt{\frac{\mu\epsilon}{\mu_0\epsilon_0}}\,. \tag{1.14}$$

Equation (1.11) is then readily identified as homogeneous (source free) wave equation with v as the wave speed.

1.4 Solutions of the Wave Equation

To gain some insight into these new wave-like solutions of Maxwell's equations, consider the scalar wave equation for fields that depend only on a single spatial variable

$$\left(\frac{\partial^2}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\mathcal{F}(z,t) = 0\,. \tag{1.15}$$

where $\mathcal{F}$ stands for any of the three Cartesian components of the field. By means of the change of variables $\xi = t - z/v$ and $\eta = t + z/v$, the derivatives in the wave equation can be transformed to

$$\frac{\partial^2}{\partial z^2} = \left(-\frac{1}{v}\frac{\partial}{\partial\xi} + \frac{1}{v}\frac{\partial}{\partial\eta}\right)\left(-\frac{1}{v}\frac{\partial}{\partial\xi} + \frac{1}{v}\frac{\partial}{\partial\eta}\right)\,, \tag{1.16}$$

$$\frac{1}{v^2}\frac{\partial^2}{\partial t^2} = \frac{1}{v^2}\left(\frac{\partial}{\partial\xi} + \frac{\partial}{\partial\eta}\right)\left(\frac{\partial}{\partial\xi} + \frac{\partial}{\partial\eta}\right)\,. \tag{1.17}$$

In terms of ξ and η the wave equation becomes

$$\frac{1}{v^2}\frac{\partial^2}{\partial\xi\partial\eta}\mathcal{G}(\xi,\eta) = 0\,, \tag{1.18}$$

where $\mathcal{G}(\xi,\eta)$ is the function $\mathcal{F}(z,t)$ expressed in terms of ξ and η. The solution to this equation are of the form

$$\mathcal{G}(\xi,\eta) = K_1 f(\xi) + K_2 g(\eta)\,, \tag{1.19}$$

where K_1 and K_2 are some constants and $f(\xi)$ and $g(\eta)$ are the two independent solutions of Eq. (1.19). Transforming back to z and t we find the solutions to the wave equation are of the form

$$\mathcal{F}(z,t) = C_1 f(t - z/c) + C_2 g(t + z/c)\,, \tag{1.20}$$

where C_1 and C_2 are constants. We see that $\mathcal{F}(z,t)$ is not an arbitrary function of z and t but in it the space and time variables occur in the combination $t - z/c$ or $t + z/c$. Examples of acceptable waveforms are periodic functions such as $e^{-i\omega(t\pm z/v)}$ or functions of finite duration such as $e^{-a(t\pm z/v)^2}$, where ω and a are constants having dimensions of frequency and frequency squared, respectively.

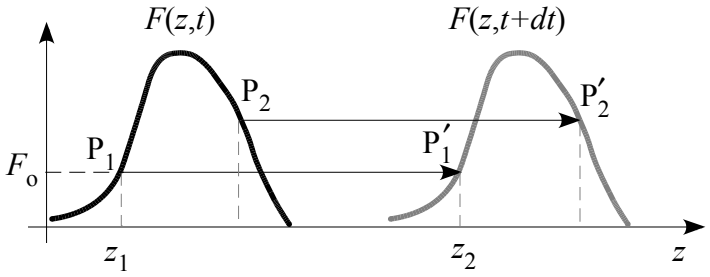


FIGURE 1.5
Pulse profile at times t and $t + dt$ for $\mathcal{F}(z,t) = f(t - z/v)$.

The function $f(t - z/v)$ represents a disturbance propagating in the $+z$ direction with speed v and $g(t+z/v)$ represents a disturbance propagating in the $-z$ direction with speed v. To see this consider a real valued disturbance $F(z,t) = f(t - z/c)$ in the form of a pulse consisting only of the first term. A plot of $F(z,t)$ as function of z when t is held fixed represents pulse shape or profile. Figure (1.5) shows a profile of this pulse at times t. To see what happens to this pulse at a later time, consider a point, such as P_1, where the pulse amplitude has a value F_0. At a later time $t + dt$, the pulse will have this same value F_0 at point z_2 such that $F(z_2, t + dt) = F_0 = f(z_1, t)$. For this to happen the argument of f at time t and $t + dt$ must be the same. This means $t - z_1/v = t + dt - z_2/v$ or $z_2 = z_1 + vdt$. Hence at time t, the disturbance has value F_0 at position z_1 and at time $t + dt$ it has the same value at point $z_2 = z_1 + vdt$. If dt is an infinitesimal interval, z_2 and z_1 will differ by an infinitesimal amount dz so that $dz = vdt$. Similar considerations for other points on the pulse profile at time t show that at time $t + dt$ the pulse has moved to the right, without change of shape, by an amount $dz = vdt$. It is clear that the disturbance propagates with speed $v = dz/dt$ in the positive z-direction. Similar considerations show that $F(z,t) = g(t+z/v)$ represents a wave propagating in the $-z$ direction with constant speed v. We have considered a simple case where wave propagates without change of shape. Propagation with change of pulse shape is possible when the speed of the wave depends on frequency.

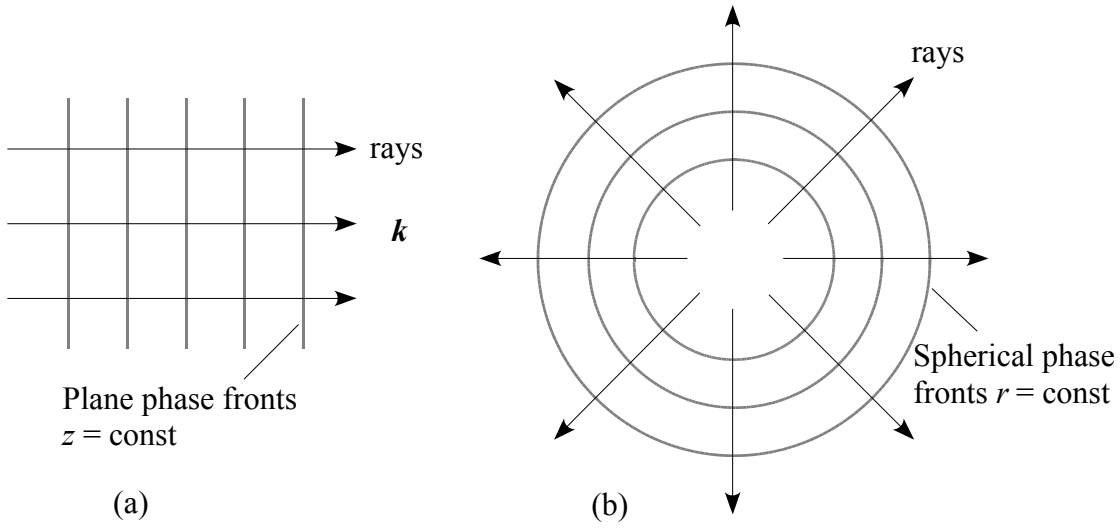


FIGURE 1.6
Wavefronts and rays for (a) plane and (b) spherical waves.

In general, a wave propagating in the direction specified by the unit vector $\boldsymbol{\kappa} = \boldsymbol{k}/|\boldsymbol{k}|$ is given by

$$\mathcal{F}(\boldsymbol{r},t) = C_1 f(t - \boldsymbol{\kappa}\cdot\boldsymbol{r}/v) + C_2 g(t + \boldsymbol{\kappa}\cdot\boldsymbol{r}/v) \tag{1.21}$$

The term plane wave is used for such solutions because the surfaces of constant wave amplitude, called wavefronts, are planes. For example, the wavefronts for $\mathcal{F}(\boldsymbol{r},t) = Cf(t - \boldsymbol{\kappa}\cdot\boldsymbol{r}/v)$ are given by $t - \boldsymbol{\kappa}\cdot\boldsymbol{r}/v = const$, which for different values of the constant defines a family of planes perpendicular to $\boldsymbol{\kappa}$. A wavefront corresponding to a definite value $t - \boldsymbol{\kappa}\cdot\boldsymbol{r}/v = C_0$ moves with velocity given by $d\boldsymbol{r}/dt = v\boldsymbol{\kappa}$. If an energy density is associated with the modulus squared of $\mathcal{F}(\boldsymbol{r},t) = Cf(t - \boldsymbol{\kappa}\cdot\boldsymbol{r})$, transport of energy occurs along trajectories called rays, which for the waves considered here are straight lines parallel to $\boldsymbol{\kappa}$. For example,

$$\mathcal{F}(z,t) = Ae^{-i[\omega(t-z/v)-\phi_o]}, \tag{1.22}$$

and

$$\mathcal{F}(\boldsymbol{r},t) = Ae^{-i[\omega(t-\boldsymbol{\kappa}\cdot\boldsymbol{r}/v)-\phi_o]}, \tag{1.23}$$

where A is a constant amplitude and ϕ_o is some constant phase angle, represent single frequency plane waves propagating, respectively, in $+z$ and $\boldsymbol{\kappa}$ directions. It is possible to have other types of scalar solutions. For example solutions of the form $\mathcal{F}(\boldsymbol{r},t) = \mathcal{F}(r,t)$ exist which satisfy

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\mathcal{F}(r,t) = 0. \tag{1.24}$$

Since $\mathcal{F}$ is assumed to depend only on r and t, we can use the identity $\nabla^2\mathcal{F}(r) =$

$(1/r)(\partial^2/\partial r^2)r\mathcal{F}(r)$ in source free regions. Then the wave equation reduces to

$$\left(\frac{\partial^2}{\partial r^2} - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right) r\mathcal{F}(r,t) = 0 \tag{1.25}$$

Noting the similarity of this equation with the one-dimensional wave equation (1.15), its solution can be written down at once as $r\mathcal{F}(r,t) = C_1 f(t - r/v) + C_2 g(t + r/v)$. This can be rewritten in the form

$$\mathcal{F}(r,t) = C_1 \frac{f(t - r/v)}{r} + C_2 \frac{g(t + r/v)}{r}\,. \tag{1.26}$$

Here the first term represents a spherically diverging wave and the second term represents a spherically converging wave. Note that for $\mathcal{F}(r,t) = C_1 f(t - r/v)/r$ the surfaces of constant wave amplitude (wavefronts) are spheres centered at the origin and energy transport occurs along radial lines diverging from the origin. The wave amplitude falls off as $1/r$ so that the energy of the wave as it propagates remains constant. More complex scalar spherical wave solutions which behave like outgoing or incoming waves far from the origin are

$$\mathcal{F}(\boldsymbol{r},t) = \begin{cases} \text{const} \times\ h_\ell^{(1)}(kr) Y_\ell^m(\theta,\varphi)\, e^{-i\omega t}\,, & \text{outgoing wave} \\ \text{const} \times\ h_\ell^{(2)}(kr) Y_\ell^m(\theta,\varphi)\, e^{-i\omega t}\,, & \text{incoming wave} \end{cases} \tag{1.27}$$

where $h_\ell^{(1)}(kr)$ and $h_\ell^{(2)}(kr)$ are spherical Hankel functions of the first and second kind. In cylindrical coordinates we have two-dimensional cylindrical waves

$$\mathcal{F}(\rho,t) = C_1 \frac{f(t - \rho/v)}{\sqrt{\rho}} + C_2 \frac{g(t + \rho/v)}{\sqrt{\rho}}, \tag{1.28}$$

where the wavefronts are (surfaces of constant wave amplitude) are cylinders coaxial with the z-axis and rays are radial lines diverging from the axis. Other solutions that involve Hankel functions and behave like outgoing or incoming cylindrical waves far from the z-axis also exist. We have mentioned only some of the simplest traveling wave solutions of the scalar wave equation. Many other solutions with more complex wavefronts representing standing or traveling waves are possible.

Vector waves predicted by Maxwell's equations can be constructed from the solutions of the scalar wave equation. If the vector wave field has a fixed direction $\boldsymbol{e}$ in space, then plane wave solutions of the form $\boldsymbol{\mathcal{F}}(\boldsymbol{r},t) = \boldsymbol{e} f(t - \boldsymbol{\kappa}\cdot\boldsymbol{r}/c)$ exist. Thus a monochromatic vector plane wave propagating in the direction of vector $\boldsymbol{k}$ constructed from the solutions of the scalar wave function has the form

$$\boldsymbol{\mathcal{F}}(\boldsymbol{r},t) = \boldsymbol{\mathcal{F}}_o e^{i(\boldsymbol{k}\cdot\boldsymbol{r} - \omega t + \phi_o)}\,. \tag{1.29}$$

Thus plane wave-like electric and magnetic fields satisfying Maxwell's equations have the form of this equation. Maxwell's equation places further restrictions on the amplitudes and relative orientations of wave vector $\boldsymbol{k}$ and the electric and magnetic field vectors. Thus a plane wave solution of Maxwell's equation is given by

$$\boldsymbol{\mathcal{E}}(\boldsymbol{r},t) = \boldsymbol{\mathcal{E}}_o e^{i(\boldsymbol{k}\cdot\boldsymbol{r} - \omega t + \phi_o)}\,, \qquad \boldsymbol{k}\cdot\boldsymbol{\mathcal{E}} = 0\,, \tag{1.30a}$$

$$\boldsymbol{\mathcal{B}}(\boldsymbol{r},t) = \frac{\boldsymbol{k}\times\boldsymbol{\mathcal{E}}}{\omega} = \frac{\boldsymbol{k}\times\boldsymbol{\mathcal{E}}_o}{\omega} e^{i(\boldsymbol{k}\cdot\boldsymbol{r} - \omega t + \phi_o)}\,, \qquad \boldsymbol{k}\cdot\boldsymbol{\mathcal{B}} = 0\,. \tag{1.30b}$$

These equations imply that in a transparent medium, electric, magnetic and propagation vectors form a right handed triad of vectors. Furthermore, the time-averaged electric and magnetic energy densities $\bar{u}_e$ and $\bar{u}_m$ are equal and contribute equally to the overall energy density $\bar{u}_{em}$ associated with the wave:

$$\overline{u}_e = \frac{1}{2}\mathcal{R}e\left[\frac{1}{2}\epsilon\boldsymbol{\mathcal{E}}\cdot\boldsymbol{\mathcal{E}}^*\right] = \frac{1}{4}\epsilon\,|\boldsymbol{\mathcal{E}}_o|^2\,, \tag{1.31a}$$

$$\overline{u}_m = \frac{1}{2}\mathcal{R}e\left[\frac{1}{2\mu}\boldsymbol{\mathcal{B}}\cdot\boldsymbol{\mathcal{B}}^*\right] = \frac{1}{4\mu}\frac{k^2}{\omega^2}|\boldsymbol{\mathcal{E}}_o|^2 = \frac{1}{4}\epsilon\,|\boldsymbol{\mathcal{E}}_o|^2 = \overline{u}_e\,, \tag{1.31b}$$

$$\overline{u}_{em} \equiv \overline{u}_e + \overline{u}_m = \frac{1}{2}\epsilon\,|\boldsymbol{\mathcal{E}}_o|^2\,. \tag{1.31c}$$

Time-averaged Poynting vector (energy flux density vector) that describes power flow in the wave is given by

$$\begin{aligned}\overline{\boldsymbol{S}} &= \frac{1}{2}\mathcal{R}e\left[\frac{\boldsymbol{\mathcal{E}}\times\boldsymbol{\mathcal{B}}^*}{\mu}\right] = \frac{1}{2}\mathcal{R}e\left[\frac{\boldsymbol{\mathcal{E}}\times(\boldsymbol{k}\times\boldsymbol{\mathcal{E}}^*)}{\omega\mu}\right]\\ &= \frac{1}{2}\epsilon|\boldsymbol{\mathcal{E}}_o|^2 v\,\boldsymbol{\kappa} \equiv I\,\boldsymbol{\kappa}\,.\end{aligned} \tag{1.31d}$$

The quantity $I \equiv |\overline{\boldsymbol{S}}| = \frac{1}{2}\epsilon|\boldsymbol{\mathcal{E}}_o|^2 v$, referred to as the wave intensity in physics (irradiance in radiometry), has units of J/s/m^2 (W/m^2). Recalling the relation between ray direction and energy flow in a wave, it is clear that the direction of Poynting vector is the ray direction in the wave. For a nonmagnetic medium ($\mu = \mu_o$) the expression for wave intensity reduces to $I = \frac{1}{2}\epsilon_o n|\boldsymbol{\mathcal{E}}_o|^2 c$ with the refractive index given by $n = \sqrt{\epsilon/\epsilon_o}$.

Vector spherical or cylindrical wave solutions are more complex even in the simplest case. For example, the simplest vector spherical wave allowed by Maxwell's equations has the form

$$\boldsymbol{\mathcal{F}}(\boldsymbol{r},t) = \text{const}\times\left[\frac{e^{-i\omega(t-r/v)}}{kr} - i\frac{e^{-i\omega(t-r/v)}}{(kr)^2}\right]\sin\theta\,\boldsymbol{e}_\varphi\,,\quad k=\omega/v\,, \tag{1.32}$$

where ω is a frequency. We encounter this and other types of vector spherical waves in the context of scattering and radiation problems in electrodynamics.

Thus Maxwell's equations admit solutions that also satisfy the wave equation. However, not all solutions of the wave equation are admissible as solutions of Maxwell's equations; only those that also satisfy the constraints imposed by the Maxwell's equations describe electromagnetic waves. This must be kept in mind even though there are situations where vector character of the field can be ignored.

1.5 Solutions in a Cavity: Mode Density

Let us recall that the amplification of a propagating light signal requires its repeated passage through a collection of excited atoms (gain medium) with which it interacts

strongly. We also know that the atoms will interact strongly with signal frequencies in a small range centered at one of their transition frequencies. We may take the FWHM of atomic line to be a measure of this range of frequencies. However, an excited population of atoms, although necessary, is not sufficient to produce significant amplification and build up of optical signals.Optical resonators that provide feedback (multiple passage through the gain medium) are an essential element of lasers. Optical resonators are needed to

(i) Store and build up light energy at the frequency of interest since the rate of stimulated emission is proportional to light energy density at a particular frequency.

(ii) Act as filters (spatial and frequency) responding selectively to field with prescribed spatial variation and frequency. Spatial filtering is responsible for the collimation properties of amplified signals and frequency filtering is responsible for their narrow bandwidth.

The ability of a resonator to perform these two tasks is measured by a figure of merit, the quality factor Q. Let us examine the field configurations and frequencies that an optical resonator will support. These field configurations are referred to as modes of the resonator.

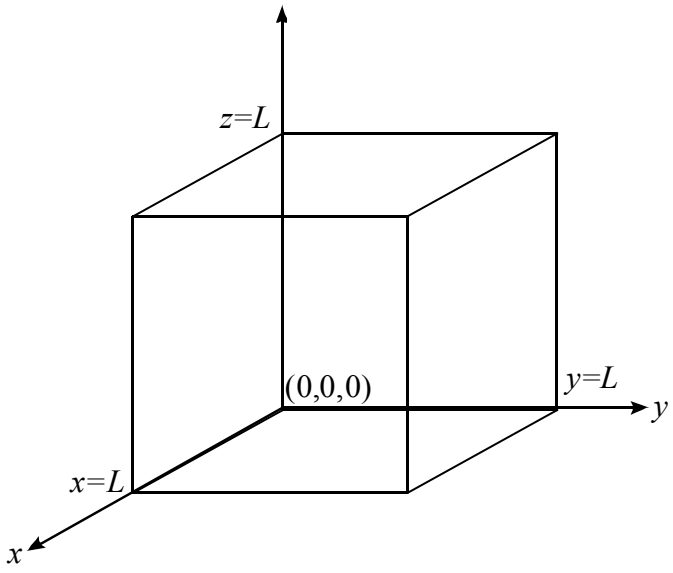


FIGURE 1.7
A rectangular cavity with conducting walls will support a discrete set of modes to allow the electric and magnetic field vectors to satisfy certain boundary conditions at the walls.

For a rectangular cavity with conducting walls and each side of length L, the electric field satisfies the wave equation and the boundary condition that the tangential component of the filed vanishes at the walls $x = 0$, $x = L$, $y = 0$, $y = L$, $z = 0$ and $z = L$. This electric field is given by $\boldsymbol{\mathcal{E}} = \mathcal{E}_0 \boldsymbol{u}_{n_1 n_2 n_3}(\boldsymbol{r}) e^{-i\omega_{n_1 n_2 n_3} t}$ where

$$u_x = A_{1x} \cos\left(\frac{n_1 \pi x}{L}\right) \sin\left(\frac{n_2 \pi y}{L}\right) \sin\left(\frac{n_3 \pi z}{L}\right) , \tag{1.33}$$

$$u_y = A_{2x} \sin\left(\frac{n_1 \pi x}{L}\right) \cos\left(\frac{n_2 \pi y}{L}\right) \sin\left(\frac{n_3 \pi z}{L}\right) , \tag{1.34}$$

$$u_z = A_{3x} \sin\left(\frac{n_1 \pi x}{L}\right) \sin\left(\frac{n_2 \pi y}{L}\right) \cos\left(\frac{n_3 \pi z}{L}\right) , \tag{1.35}$$

where the mode wave vector, wave vector length and frequency are given by

$$\boldsymbol{k}_{n_1 n_2 n_3} = \frac{\pi}{L}(n_1, n_2, n_3) \ , \quad k_{n_1 n_2 n_3} = \frac{\pi}{L}\sqrt{n_1^2 + n_2^2 + n_3^2}, \tag{1.36}$$

and

$$\omega_{n_1 n_2 n_3} \equiv 2\pi \nu_{n_1 n_2 n_3} = \frac{c k_{n_1 n_2 n_3}}{n} , \tag{1.37}$$

respectively. Since the electric field must satisfy $\boldsymbol{\nabla} \cdot \boldsymbol{\mathcal{E}} = 0$ inside the cavity, the mode indices (n_1, n_2, n_3) and the field amplitudes (A_1, A_2, A_3) satisfy the constraint

$$\frac{\pi}{L}\left[n_1 A_{1x} + n_2 A_{2x} + n_3 A_{3x}\right] = 0 .$$

The magnetic field associated with this electric field can be calculated using Maxwell's equations.

The mode functions satisfy the usual orthogonality relation

$$\int_V \boldsymbol{u}_{n_1 n_2 n_3} \cdot \boldsymbol{u}^*_{n_1 n_2 n_3} d^3 r = \delta_{n_1 n_1'} \delta_{n_2 n_2'} \delta_{n_3 n_3'} .$$

These modes can be represented by discrete points in a three-dimensional space spanned by (n_1, n_2, n_3) axes as shown in Fig. 1.7. Each point is associated with two modes corresponding to two independent polarizations for each wave vector.

With this mode structure, the number of modes in the frequency interval ν and $\nu + d\nu$ is then the number of points inside a spherical shell of radius ν and thickness $d\nu$

$$\begin{aligned} \frac{dN}{d\nu} &= \frac{1}{8} \frac{\text{volume of spherical shell}}{\text{volume associated with one mode}} \times 2 = \frac{1}{8} \frac{4\pi\nu^2 d\nu}{(c^3/8n^3L^3)} \times 2 \\ &= \frac{8\pi\nu^2 n^3 d\nu V}{c^3} . \end{aligned} \tag{1.38}$$

The factor $\frac{1}{8}$ in the first line of this equation accounts for the fact that mode indices are positive integers. Hence only the points (n_1, n_2, n_3) in the positive octant of the sphere in n_1, n_2, n_3 space count toward the number of modes. The factor of 2 at the end takes into account the two polarization degree of freedom for light for each set wave number $k_{n_1 n_2 n_3}$.

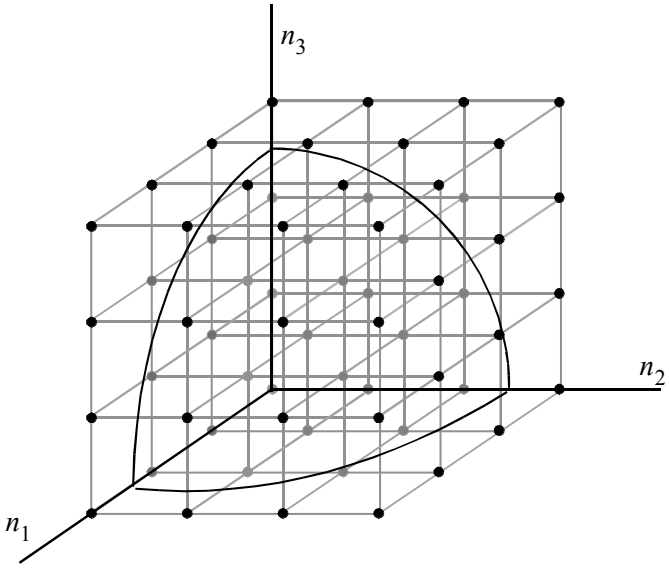


FIGURE 1.8
Each mode is characterized by three positive integers n_1, n_2 and n_3 can be represented as a point in the positive octant of (n_1, n_2, n_3) space. Each point is counted twice corresponding to two polarizations of the field.

If such a resonator is used at optical frequency with an inverted gain medium inside the resonator, the number of resonator modes p falling under the gain bandwidth of laser transition will be

$$p = \frac{8\pi\nu^2}{c^3} V \delta\nu_{at} = 8\pi \left(\frac{V}{\lambda^3}\right) \frac{\delta\nu_{at}}{\nu_o}\,, \tag{1.39}$$

where we have put the refractive index $n = 1$. For a frequency $\nu = 5 \times 10^{14}$ Hz ($\lambda = 600$ nm), $\delta\nu_{at} = 1.5 \times 10^9$ Hz and a cavity of volume V=1 cm^3, we find the number of modes interacting with the atom is $p = 3.5 \times 10^8$. For a closed resonator all of these modes will have access to atomic gain. They will have similar losses and feedback so that oscillation would occur at a very large number of frequencies. Such a behavior would be highly undesirable because it would result in light from the laser being emitted in a wide spectral range and in all directions (and hence no collimation).

How do we reduce the number of modes? One possibility (suggested by p being proportional to cavity volume V) is to make a small cavity. Suppose we want $p = 1$ at λ=600 nm. This will require a cavity of volume

$$V = p\frac{\lambda^3}{8\pi}\frac{\nu_o}{\delta\nu_{at}} = 1 \times \frac{(0.6\,\mu\text{m})^3}{8\pi}\frac{5 \times 10^{14}}{1 \cdot 5 \times 10^9} = 14\ (\mu\text{m})^3\ ! \tag{1.40}$$

Such a cavity, although not impossible nowadays, is not practical because we need some room to place the amplifying medium. Moreover, even if we could accommodate the medium, the gain itself will be very small.

Problems discussed above can be overcome to a large extent by employing open resonators. In open resonators only those modes that correspond to a superposition of waves traveling very nearly parallel to the resonator axis will have low enough losses for fields to build up. Energy in all other modes will be lost in a few traversals. These modes will have a very low Q. Open resonators were first suggested by Schawlow and Townes and Prokhorov.[3] In open resonators there are no conducting side walls. The active medium is placed between two mirrors carefully aligned. In such resonators photons traveling along the axis are trapped. Those photons that travel at an angle eventually escape. One can improve things a bit by using curved mirrors so that due to the focusing action of mirrors only photons making small angles with the axis are trapped. For such photons we can write

$$\frac{k_1}{k} = \frac{n_1\pi/L}{(\pi/L)\sqrt{n_1^2+n_2^2+n_3^2}}\,, \tag{1.41a}$$

$$\frac{k_2}{k} = \frac{(n_2\pi/L)}{(\pi/L)\sqrt{n_1^2+n_2^2+n_3^2}}\,, \tag{1.41b}$$

$$\frac{k_3}{k} = \frac{(n_3\pi/L)}{(\pi/L)\sqrt{n_1^2+n_2^2+n_3^2}}\,. \tag{1.41c}$$

With $k_1, k_2 \ll k, k_3$ and using the relation $k = \frac{2\pi}{\lambda} = \frac{\pi}{L}\sqrt{n_1^2+n_2^2+n_3^2}$, we can write $\frac{k_1}{k} = \frac{n_1\lambda}{2L}$, $k_2 = \frac{n_2\lambda}{2L}$, $k_3 = \frac{n_3\lambda}{2L}$, $n_1, n_2 \ll n_3$. Hence for each value of n_3 we have a small group of modes that have a frequency close to $\nu_{n_3} = kv = \frac{kc}{n} \approx n_3\frac{c}{2nL}$ that will be amplified. We call $\nu_{n_3} = \frac{n_3 c}{2nL}$ a resonance frequency. These resonance frequencies are separated from each other by

$$\Delta\nu = \nu_{n_3+1} - \nu_{n_3} = \frac{c}{2nL}\,.$$

These groups of frequencies are referred to as quasi-modes of the resonator because for open cavities true modes (stationary modes) are not defined. It is clear that the light coming out from open resonators will have beam-like quality. This is because only $k_1, k_2 \ll k_3$ modes will be populated. There is another change that occurs due to finite mirror apertures; even paraxial modes suffer some energy loss due to diffraction as every time a wave hits the end mirrors, energy will be lost due to their finite size. Because of decay of field energy one cannot use the concept of true modes (stationary configurations) for open resonators. However quasi-modes (field configurations lasting millions of optical cycles) do exist, which have transverse extent that falls off rapidly with distance from the axis. These fields have space-time structure

$$\boldsymbol{E}(\boldsymbol{r},t) = E_o\boldsymbol{U}(\boldsymbol{r})e^{-i\omega t}e^{-t/2\tau_c}\,, \tag{1.42}$$

where E_o is the field amplitude at time $t = 0$ and τ_c is a characteristic time scale on which the amplitude of the field decays. It is clear that in the absence of amplification, quasi-mode field amplitude decays to zero in time. In the rest of this

[3] A. L. Schawlow and C. H. Townes, Phys. Rev. 112, 1940 (1958); A. M. Prokhorov, Sov. Phys. JETP 1, 1140(1958).

course when we talk about modes of open resonators, we will be referring to these quasi-modes.

In order to see what kind of quasi-modes are possible in open resonators, we use the complementary description of wave propagation in terms of rays, which as we have seen in the preceding section, are geometrical curves, orthogonal to phase fronts, along which electromagnetic energy is transported. In essence, we will be looking for ray trajectories that stay confined in open resonator structures. Such rays correspond to phase fronts that are stable spatial structures.

1.6 Ray Optics

In ray optics[4], optical energy is regarded as being transported along certain curves called light rays, which as we have seen in Sec. 1.4, are geometrical curves orthogonal to phase fronts along which light energy is transported. Since Poynting vector describes the local energy flow in an electromagnetic wave, rays represent the local direction of Poynting vector.

We will show that for small wavelength the field locally has the same general character as that of a plane wave and the laws of reflection and refraction established for plane waves incident upon a plane boundary remain valid under more general conditions. Hence if a light ray falls on a sharp boundary (for example the surface of a lens) it is split into a reflected ray and transmitted ray obeying the laws of reflection and refraction. The preceding remarks imply that, when wavelength is small enough, optical phenomena may be deduced from geometrical considerations by determining the path of the light rays.

In this limit, wavelike monochromatic solutions of Maxwell's equations are of the form

$$\boldsymbol{\mathcal{E}}(\boldsymbol{r},t) = \boldsymbol{e}(\boldsymbol{r})\, e^{-i(\omega t - k_o \Psi(\boldsymbol{r}))}\,, \qquad \boldsymbol{\mathcal{B}}(\boldsymbol{r},t) = \boldsymbol{b}(\boldsymbol{r})\, e^{-i(\omega t - k_o \Psi(\boldsymbol{r}))}\,. \tag{1.43}$$

Here the envelope functions $\boldsymbol{e}(\boldsymbol{r})$ and $\boldsymbol{b}(\boldsymbol{r})$ are some slowly varying[5] functions of position and $\Psi(\boldsymbol{r})$ is a real scalar function which remains to be determined. Substituting these into Maxwell's equations for transparent media (μ and ϵ real),

$$\begin{array}{ll} \boldsymbol{\nabla}\cdot\epsilon\boldsymbol{\mathcal{E}} = 0 & \boldsymbol{\nabla}\cdot\boldsymbol{\mathcal{B}} = 0 \\ \boldsymbol{\nabla}\times\boldsymbol{\mathcal{E}} = i\omega\boldsymbol{\mathcal{B}} & \boldsymbol{\nabla}\times\frac{\boldsymbol{\mathcal{B}}}{\mu} = -i\omega\epsilon\boldsymbol{\mathcal{E}} \end{array} \tag{1.44}$$

[4]The branch of optics characterized by the neglect of wavelength λ in comparison to the characteristic length ℓ of the problem - for example, the length scale on which the refractive index changes significantly or the radius of curvature of interfaces between media - is known as geometrical optics. In this approximation the laws of optics may be formulated in the language of geometry.

[5]Fractional change in their values is negligible over distances of the order of a few wavelengths

we find, with $\Phi(\boldsymbol{r},t) = \omega t - k_0\Psi(\boldsymbol{r})$,

$$\begin{aligned}
\epsilon\left[\boldsymbol{e}\cdot\boldsymbol{\nabla}\ln\epsilon + \boldsymbol{\nabla}\cdot\boldsymbol{e} + ik_0\boldsymbol{e}\cdot\boldsymbol{\nabla}\Psi(\boldsymbol{r})\right]e^{-i\Phi(\boldsymbol{r},t)} &= 0\\
\left[\boldsymbol{\nabla}\cdot\boldsymbol{b} + ik_0\boldsymbol{b}\cdot\boldsymbol{\nabla}\Psi(\boldsymbol{r})\right]e^{-i\Phi(\boldsymbol{r},t)} &= 0\\
\left[\boldsymbol{\nabla}\times\boldsymbol{e} + ik_0\boldsymbol{\nabla}\Psi(\boldsymbol{r})\times\boldsymbol{e}\right]e^{-i\Phi(\boldsymbol{r},t)} &= i\omega\boldsymbol{b}\,e^{-i\Phi(\boldsymbol{r},t)}\\
\frac{1}{\mu}\left[-\boldsymbol{\nabla}\ln\mu\times\boldsymbol{b} + \boldsymbol{\nabla}\times\boldsymbol{b} + ik_0\boldsymbol{\nabla}\Psi(\boldsymbol{r})\times\boldsymbol{b}\right]e^{-i\Phi(\boldsymbol{r},t)} &= -i\omega\epsilon\boldsymbol{e}\,e^{-i\Phi(\boldsymbol{r},t)}
\end{aligned} \tag{1.45}$$

Rearranging these we find

$$\boldsymbol{e}\cdot\boldsymbol{\nabla}\Psi(\boldsymbol{r}) = \frac{1}{ik_0}\left[\boldsymbol{e}\cdot\boldsymbol{\nabla}\ln\epsilon + \boldsymbol{\nabla}\cdot\boldsymbol{e}\right] \tag{1.46a}$$

$$\boldsymbol{b}\cdot\boldsymbol{\nabla}\Psi(\boldsymbol{r}) = \frac{1}{ik_0}\left[\boldsymbol{\nabla}\cdot\boldsymbol{b}\right] \tag{1.46b}$$

$$\boldsymbol{\nabla}\Psi(\boldsymbol{r})\times\boldsymbol{e} - c\boldsymbol{b} = -\frac{1}{ik_0}\left[\boldsymbol{\nabla}\times\boldsymbol{e}\right] \tag{1.46c}$$

$$\boldsymbol{\nabla}\Psi(\boldsymbol{r})\times\boldsymbol{b} + c\mu\epsilon\boldsymbol{e} = -\frac{1}{ik_0}\left[-(\boldsymbol{\nabla}\ln\mu)\times\boldsymbol{b} + \boldsymbol{\nabla}\times\boldsymbol{b}\right] \tag{1.46d}$$

If the changes in μ, ϵ and the envelopes $\boldsymbol{e}$ and $\boldsymbol{b}$ over distances of the order of a few wavelengths are small we can ignore the terms on the right hand side of each of these equations,

$$\boldsymbol{e}(\boldsymbol{r})\cdot\boldsymbol{\nabla}\Psi(\boldsymbol{r}) = 0 \tag{1.47a}$$

$$\boldsymbol{b}(\boldsymbol{r})\cdot\boldsymbol{\nabla}\Psi(\boldsymbol{r}) = 0 \tag{1.47b}$$

$$\boldsymbol{\nabla}\Psi(\boldsymbol{r})\times\boldsymbol{e}(\boldsymbol{r}) - c\boldsymbol{b}(\boldsymbol{r}) = 0 \tag{1.47c}$$

$$\boldsymbol{\nabla}\Psi(\boldsymbol{r})\times\boldsymbol{b}(\boldsymbol{r}) + c\mu\epsilon\boldsymbol{e}(\boldsymbol{r}) = 0 \tag{1.47d}$$

Note that the first two equations follow from the last two by taking scalar product with $\boldsymbol{\nabla}\Psi(\boldsymbol{r})$. It follows from Eqs.(1.47) that $\boldsymbol{e}(\boldsymbol{r})$, $\boldsymbol{b}(\boldsymbol{r})$, and $\boldsymbol{\nabla}\Psi(\boldsymbol{r})$ form a right handed triad of mutually orthogonal vectors at each point $\boldsymbol{r}$. Furthermore, since the vector $\boldsymbol{\nabla}\Psi(\boldsymbol{r})$ is perpendicular to the surface $\Psi(\boldsymbol{r})$=const, vectors $\boldsymbol{e}(\boldsymbol{r})$ and $\boldsymbol{b}(\boldsymbol{r})$ are tangential to the surface $\Psi(\boldsymbol{r})$=const. This surface may be called the geometrical optics wave surface or the geometrical wavefront.

By eliminating $\boldsymbol{e}$ or $\boldsymbol{b}$ from Eqs. (1.37c) and (1.37d), we find that the condition for nontrivial solutions of Eqs.(1.47) is

$$[\boldsymbol{\nabla}\Psi(\boldsymbol{r})]^2 = \mu\epsilon c^2 = n^2(\boldsymbol{r})\,, \tag{1.48}$$

where $n(\boldsymbol{r})$ is the refractive index of the medium

$$n(\boldsymbol{r}) = \sqrt{\frac{\mu(\boldsymbol{r})\epsilon(\boldsymbol{r})}{\mu_0\epsilon_0}}\,. \tag{1.49}$$

In general, $n(\boldsymbol{r})$ is a function of position because ϵ and μ are functions of position. Equation (1.48), known as the eikonal (from German Eikonal, which is from Greek

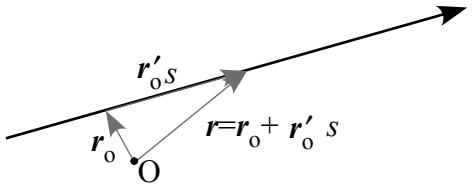


FIGURE 1.9
(a) Geometrical wavefronts and the direction of the unit vector $\boldsymbol{s}$. (b) Rays are directed (pointing in the direction of energy flow) trajectories perpendicular to phasefronts.

$\epsilon\iota\kappa\omega\nu$, image) 0equation, is the basic equation of geometrical optics. The function Ψ is known as the eikonal. It follows from the eikonal equation that the vector

$$\boldsymbol{s} = \frac{\boldsymbol{\nabla}\Psi(\boldsymbol{r})}{n(\boldsymbol{r})} \tag{1.50}$$

has unit magnitude and is perpendicular to the surface $\Psi(\boldsymbol{r})$ =const. The surfaces $\Psi(\boldsymbol{r})$ =const are called the geometrical wavefronts

With the help of Eqs. (1.37c) and (1.37d) we can express the electric and magnetic field vectors as

$$\boldsymbol{b}(\boldsymbol{r}) = \frac{\boldsymbol{s}\times\boldsymbol{e}(\boldsymbol{r})}{v}\,, \tag{1.51a}$$

$$\boldsymbol{e}(\boldsymbol{r}) = -v\boldsymbol{s}\times\boldsymbol{b}(\boldsymbol{r})\,, \tag{1.51b}$$

where $v = c/n$ is the wave speed in the medium. The time averaged electric and magnetic energy densities are then

$$\overline{u}_e = \frac{1}{4}\epsilon|\boldsymbol{e}|^2 \tag{1.52a}$$

$$\overline{u}_m = \frac{1}{4}\frac{|\boldsymbol{b}|^2}{\mu} = \frac{1}{4\mu}\frac{\boldsymbol{e}\cdot\boldsymbol{e}^*}{v^2} = \frac{1}{4}\epsilon\frac{|e|^2}{\mu\epsilon v^2} = \frac{1}{4}\epsilon|e|^2 = \overline{u}_e \tag{1.52b}$$

Thus in the limit of geometrical optics, the time averaged electric and magnetic energy densities associated with a monochromatic wave in a transparent medium are equal.

The time averaged Poynting vector (energy flux density vector) is given by

$$\overline{\boldsymbol{S}} = \frac{1}{2\mu}\mathcal{R}e\,[\boldsymbol{e}\times\boldsymbol{b}^*] = \frac{\boldsymbol{e}\times[n\boldsymbol{s}\times\boldsymbol{e}^*]}{2\mu c} = \frac{1}{2}\Big(\frac{c}{n}\Big)\,\epsilon\boldsymbol{e}\cdot\boldsymbol{e}^*\boldsymbol{s} = v\overline{u}_{em}\boldsymbol{s}\,. \tag{1.53}$$

The average Poynting vector (energy flux vector) is in the direction of normal to the geometrical wavefront $\Psi(\boldsymbol{r})$ =const. Its magnitude is equal to the product of the average energy density and speed $v = c/n$ of the wave in the medium. It follows from Eqs. (1.28)-(1.33) that the fields in the geometrical optics limit have the same local character as a plane wave.

1.7 Ray Propagation

We can now define the geometrical light rays as the orthogonal trajectories to the geometrical wavefronts $\Psi(\boldsymbol{r}) = \text{const}$. We regard them as directed curves whose direction coincides everywhere with the direction of the average Poynting vector. We may then say that in geometrical optics light energy is transported along the light rays. The differential equation obeyed by the ray is easily derived as follows. Let $\boldsymbol{r}(s)$ denote the position vector of a point P on a ray, considered as a function of the arc length s along the ray measured from some fixed point on it. Then the unit vector $d\boldsymbol{r}/ds = ds$ is tangential to the ray in the direction of energy flow. Using the relation of $\boldsymbol{s}$ to $\boldsymbol{\nabla}\Psi(\boldsymbol{r})$, the equation for the ray can be written as

$$\frac{d\boldsymbol{r}}{ds} = \boldsymbol{s} \equiv \frac{\boldsymbol{\nabla}\Psi(\boldsymbol{r})}{n} \qquad \Longrightarrow \qquad n\frac{d\boldsymbol{r}}{ds} = \boldsymbol{\nabla}\Psi(\boldsymbol{r})\,. \tag{1.54}$$

This equation is purely formal as it specifies rays in terms of $\Psi(\boldsymbol{r})$ which must be determined from eikonal equation. We can derive another differential equation for the rays directly in terms of the refractive index $n(\boldsymbol{r})$ which is much more useful. Differentiating the equation for the ray with respect to arc length we obtain

$$\frac{d}{ds}n\frac{d\boldsymbol{r}}{ds} = \frac{d}{ds}\boldsymbol{\nabla}\Psi\,. \tag{1.55}$$

We can express the right hand sides in terms of $n(\boldsymbol{r})$ as

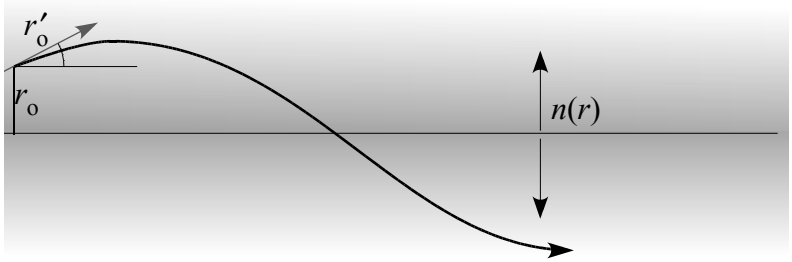


FIGURE 1.10
Rays are curves along which energy of a wave is transported.

$$\begin{aligned}\frac{d}{ds}\boldsymbol{\nabla}\Psi &= (\boldsymbol{s}\cdot\boldsymbol{\nabla})\boldsymbol{\nabla}\Psi \\ &= \left(\frac{\boldsymbol{\nabla}\Psi}{n}\cdot\boldsymbol{\nabla}\right)\boldsymbol{\nabla}\Psi = \frac{1}{2n}\boldsymbol{\nabla}(\boldsymbol{\nabla}\Psi)^2 = \frac{1}{2n}\boldsymbol{\nabla}(n^2) = \boldsymbol{\nabla}n\,.\end{aligned} \tag{1.56}$$

Using these results we find the equation of a ray in terms of the variation of the refractive index

$$\frac{d}{ds}n\frac{d\boldsymbol{r}}{ds} = \boldsymbol{\nabla}n\,. \tag{1.57}$$

We can gain some insight into what this equation says by writing this in yet another form. We first note that $\boldsymbol{s} = \boldsymbol{\nabla}\Psi/n$ [Eq. (1.50)] is a unit vector so that $\boldsymbol{s} \cdot \boldsymbol{s} = 1$. By differentiating this with respect to s we find $\boldsymbol{s} \cdot (d\boldsymbol{s}/ds) = 0$ which means that $d\boldsymbol{s}/ds$ is a vector perpendicular to $\boldsymbol{s}$. In fact from differential geometry

$$\frac{d\boldsymbol{s}}{ds} = \frac{\boldsymbol{\nu}}{R} = \frac{d^2\boldsymbol{r}}{ds^2}\,, \tag{1.58}$$

where R is the radius of curvature of the trajectory and $\boldsymbol{\nu}$ is a unit vector along the principal normal. Using these results we find the equation of a ray can also be written as

$$\frac{dn}{ds}\boldsymbol{s} + n\frac{\boldsymbol{\nu}}{R} = \boldsymbol{\nabla}n\,. \tag{1.59}$$

Rewriting this equation as

$$\frac{\boldsymbol{\nu}}{R} = \frac{1}{n}\left[\boldsymbol{\nabla}n - \boldsymbol{s}\frac{dn}{ds}\right]\,, \tag{1.60}$$

and taking the scalar product with $\boldsymbol{\nu}$ we find

$$\frac{1}{R} = \frac{\boldsymbol{\nu}}{n} \cdot \boldsymbol{\nabla}n = \boldsymbol{\nu} \cdot \boldsymbol{\nabla}\ln n \tag{1.61}$$

This equation says that a ray is bent toward the region of higher refractive index.

1.7.1 Homogeneous Medium

As an application of this equation we consider a homogeneous medium $n = \text{const}$. In this case the equation for the ray becomes

$$\frac{d^2\boldsymbol{r}}{ds^2} = 0\,. \tag{1.62}$$

On integrating this equation we obtain

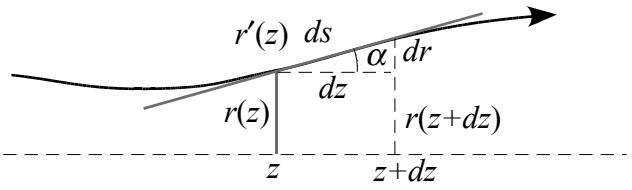


FIGURE 1.11
In a homogeneous medium rays are straight lines.

$$\begin{aligned} \boldsymbol{r}\,'(s) &= \boldsymbol{r}\,'_o \\ \boldsymbol{r}(s) &= \boldsymbol{r}\,'_o s + \boldsymbol{r}_o\,. \end{aligned} \tag{1.63}$$

where $\boldsymbol{r}\,'_0$ is the initial slope of the ray at point $\boldsymbol{r}_0$. This is the equation of a straight line.

1.7.2 Rays in a Duct

As another example we consider an axially symmetric medium with quadratic index variation in directions perpendicular to the axis of symmetry (taken to be the z-axis) $n = n_0 - \frac{1}{2}n_2(x^2 + y^2)$ with $n_2 > 0$. Then in the paraxial approximation $ds = dz$, the equation for paraxial rays becomes

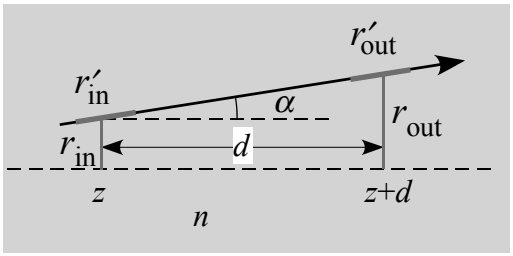


FIGURE 1.12
Rays in a duct with radially decreasing refractive index.

$$\frac{d^2r}{dz^2} + \frac{n_2}{n_0}r = \frac{d^2r}{dz^2} + \alpha^2 r = 0\,, \quad \alpha = \sqrt{n_2/n_0}. \tag{1.64}$$

with solution

$$r(z) = r_0\cos(\alpha z) + \left(\frac{r_0'}{\alpha}\right)\sin(\alpha z) \tag{1.65}$$

where r_o is the ray displacement from the z axis and r_o' is the initial slope of the ray at $z = 0$. Thus the ray oscillates up and down about the z-axis. We note also that a family of parallel rays (rays with the same slope but different displacement) periodically converge in planes $z = (4n + 1)\pi/2\alpha$ at points at a height $r_f = (r_o'/\alpha)$ from the z-axis and emerge from these planes with their original slope. These planes are separated from each other by $z_f = 2\pi/\alpha$.

Exercise 1:

Show that a section of the quadratic medium of length ℓ surrounded by a medium of refractive index n_m acts as a lens. [Hint: Show that a family of parallel rays entering at $z = 0$ with different displacement converge after emerging at $z = \ell$ to a common focus at a distance $f = \frac{1}{n_m\alpha}\cot\alpha\ell$.]

Exercise 2:

Find the equation for paraxial rays when $n = n_0 + \frac{1}{2}n_2(x^2 + y^2)$.
The answer is

$$r(z) = r_0\cosh\alpha z + \left(\frac{r_0'}{\alpha}\right)\sinh\alpha z\,, \quad \alpha = \sqrt{n_2/n_0}\,.$$

1.7.2.1 Laws of Reflection and Refraction

In a two dimensional region ($Y-Z$ plane) where the refractive index depends on y only, the ray equation becomes

$$\frac{d}{ds}n\frac{dz}{ds}=0 \quad \Rightarrow \quad n\frac{dz}{ds}=\text{const}\,. \tag{1.66}$$

Using $\frac{dz}{ds}=\cos\alpha=\sin\theta$ [Fig. 1.13], we can write this as $n\cos\alpha=$ constant. For

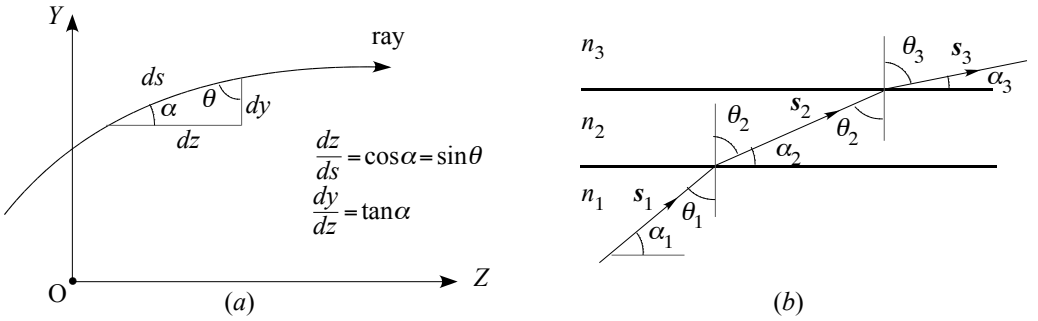


FIGURE 1.13
(a) A ray in a continuous inhomogeneous medium; (b) Ray path in a layered (piecewise homogeneous) medium.

a layered medium this leads to $n_1\sin\alpha_1=n_2\sin\alpha_2=n_3\sin\alpha_3=\cdots$. Expressing angle α in terms of the angle that the ray makes with the normal to the interface, we have

$$n_1\sin\theta_1=n_2\sin\theta_2=n_3\sin\theta_3=\cdots\,. \tag{1.67}$$

This is precisely the Snell's law for a layered medium.

The laws of refraction and reflection at a surface across which the refractive index changes abruptly can be derived more rigorously using the fact that $n\boldsymbol{s}=\boldsymbol{\nabla}\Psi$, which implies $\boldsymbol{\nabla}\times n\boldsymbol{s}=0$. To see this we replace the surface of discontinuity by a thin transition layer in which n changes rapidly but smoothly from its value n_1 on one side of the surface of discontinuity to its value n_2 on the other side [Fig. 1.14]. Then integrating the normal component of $\boldsymbol{\nabla}\times n\boldsymbol{s}$ over the open surface of a small rectangle $P_1Q_1Q_2P_2P_1$ straddling the interface between two media [Fig. 1.14], using Stokes' theorem to convert the surface integral into a line integral and letting the thickness of the rectangle δh (transition layer thickness) shrink to zero, we obtain

$$[n_1\boldsymbol{s}_1\cdot\boldsymbol{t}_1+n_2\boldsymbol{s}_2\cdot\boldsymbol{t}_2]\delta\ell=0\,, \tag{1.68}$$

where $\delta\ell$ is the line element in which the rectangle intersects the surface. Expressing the unit vectors $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ in terms the unit tangent vector $\boldsymbol{t}$ along the surface [Fig. 1.14] we obtain $\boldsymbol{a}\cdot[\boldsymbol{n}_{12}\times(n_2\boldsymbol{s}_2-n_1\boldsymbol{s}_1)]=0$. Since the orientation of the rectangle and

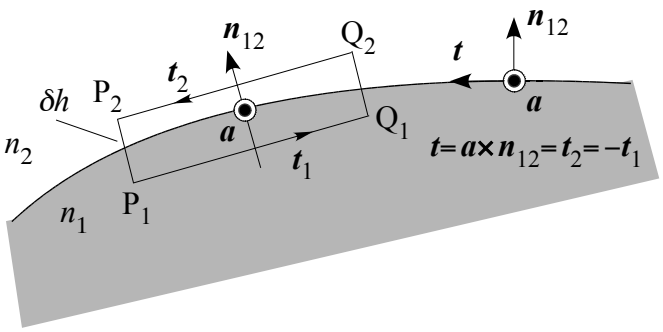


FIGURE 1.14
Ray transformation across a surface of discontinuity of refractive index.

therefore the unit vector $\boldsymbol{a}$ is arbitrary, we conclude that the tangential component of the ray vector is continuous across the surface of discontinuity.

$$\boldsymbol{n}_{12} \times (n_2 \boldsymbol{s}_2 - n_1 \boldsymbol{s}_1) = 0 \quad \Longrightarrow \quad \boldsymbol{n}_{12} \times n_2 \boldsymbol{s}_2 = \boldsymbol{n}_{12} \times n_1 \boldsymbol{s}_1 \tag{1.69}$$

where $\boldsymbol{n}_{12}$ is a unit normal to the interface directed from medium 1 to 2. This equation implies that the normal to the boundary $\boldsymbol{n}_{12}$ and the ray vectors $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are coplanar and the angles the ray vectors makes with the normal to the boundary are related by [Fig. 1.15]

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \,. \tag{1.70}$$

Here θ_1 and θ_2 are the angles that the rays in the two media make with $\boldsymbol{n}_{12}$. A similar procedure for the reflected ray leads to the laws of reflection [Fig. 1.15(b)].

These laws are usually derived for plane waves incident on a refracting plane surface. Here we find that they are valid for more general waves and refracting surface provided that the wavelength is sufficiently small. This means the radius of curvature of the incident wave and of the interface must be large compared to the wavelength of the incident light.

Thus in the limit of short wavelength Maxwell's equations lead to a description of light propagation in terms of rays which are geometric curves along which light energy is transported. We also see that in this limit wave nature of light is masked and phenomena such as diffraction are neglected. It may seem a drastic simplification but it is very useful in instrument optics and we will see that laws of ray optics are very useful even for understanding diffraction effects. In what follows we will limit our discussion to paraxial rays.

1.7.3 Paraxial Rays

In problems involving instrument optics or laser resonators we are interested in rays that stay close to the optical axis, usually taken to be the $z-$ axis, of the system. Such rays are called paraxial rays. They make small angles with the optical axis such

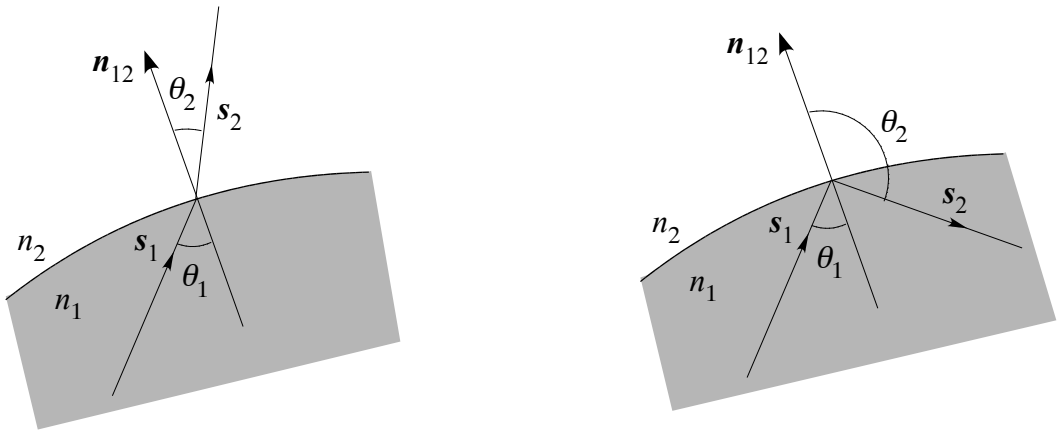


FIGURE 1.15
Laws of (a) refraction and (b) reflection. Note that for reflection $n_2 = n_1$ so that Eq. (1.70) gives $\sin\theta_2 = \sin\theta_1$, which implies $\theta_2 = \pi - \theta_1$, which is the law of reflection.

that the sine and tangent of the angle can be approximated by the angle (expressed in radians) itself,

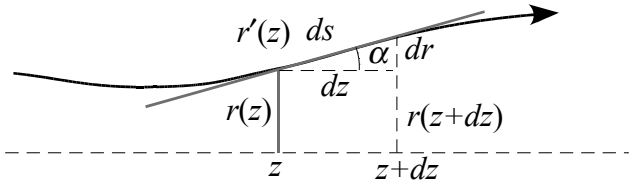


FIGURE 1.16
A paraxial ray makes small angles with the optical axis (usually the $z-$axis) and stays close to it during propagation.

$$\tan\alpha \approx \alpha \approx \sin\alpha\,. \tag{1.71}$$

This approximation is good to within 3% for angles less than about 18^o ($0.31 \approx 1/\pi$ radian). For such rays the arc length $ds \approx dz$ and the equation for the ray becomes

$$\frac{d}{dz} n \frac{d\boldsymbol{r}}{dz} = \boldsymbol{\nabla} n \tag{1.72}$$

For media with cylindrical symmetry we can write $\boldsymbol{r} = \boldsymbol{e}_\perp r + \boldsymbol{e}_z z$, where r is the lateral displacement of the ray from the z axis, we find the equation of a ray becomes

$$\frac{d}{dz} n \frac{dr}{dz} = \boldsymbol{e}_\perp \cdot \boldsymbol{\nabla} n\,. \tag{1.73}$$

This equation determines a ray given the lateral displacement r_o and slope $r'_o = [dr/dz]_{z=z_o}$ of a ray at some fixed point z_o. In what follows we will use the abbreviation $r' = dr/dz$ to denote the slope.

Example 1. In a homogeneous medium n =const so that the equation of a ray becomes

$$\frac{d}{dz} n \frac{dr}{dz} = 0\,. \tag{1.74}$$

Integrating this equation from z_o to z, we find the ray is a straight line given by

$$\begin{aligned} r'(z) &= r'_o\,, \\ r(z) &= r_o + r'_o(z - z_o)\,, \end{aligned} \tag{1.75}$$

where r_o and r'_o are, respectively, the ray displacement and slope at $z = z_o$.

If n changes, as for example when a ray is incident from one homogeneous medium (refractive index n_1) with slope r'_1 and displacement r_1 onto another homogeneous medium (refractive index n_2), an integration of the ray equation across the interface ($z = 0$) into the second medium gives

$$\begin{aligned} r'_2(z) &= \frac{n_1}{n_2}\, r'_1\,, \\ r_2(z) &= r_1 + \frac{n_1}{n_2}\, r'_1 z\,. \end{aligned} \tag{1.76}$$

The first of these equations is Snell's law in the paraxial approximation. The second equation gives the ray displacement in the second medium in terms of the ray displacement r_1 and slope $r'_2 = (n_1/n_2)r'_1$ at the boundary just inside the second medium.

The study of ray propagation is important in its own right in instrument optics. It turns out that although the treatment of light as a collection of geometrical optics rays ignores diffraction, laws of paraxial ray propagation turn out to be very useful in understanding the full diffractive propagation of light in optical resonators and laser beams.

In paraxial optics a ray is specified by its displacement r from the optical axis and its slope $r' = dr/dz$. Both of these quantities vary with z as the ray propagates through an optical system. If we introduce a column matrix with r and r' its elements by

$$\tilde{r} = \begin{bmatrix} r \\ r' \end{bmatrix}\,, \tag{1.77}$$

we can describe the effect of an optical element on ray parameters (r, r') by a 2×2 matrix of the form

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}\,. \tag{1.78}$$

For most purposes we have to know the transformation properties of three basic elements:

(i) free propgation in a homogeneous medium of length L and refractive index n;

(ii) reflection from a curved surface of radius of curvature R;

(iii) refraction at a curved interface (radius of curvature R) between two media with refractive indices n_1 and n_2 when a ray is incident from medium n_1.

Matrices for most others elements can be derived from these.

1.7.3.1 Propagation in a homogeneous medium

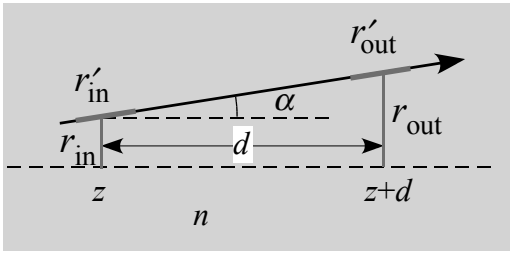


FIGURE 1.17
Propagation in a homogeneous medium of length d.

Consider a ray propagating from plane z to plane $z+d$ in a homogeneous medium of refractive index n, the ray parameters at the input and output faces are related by [Eq. (1.75)]

$$\begin{aligned} r_{\text{out}} &= r_{\text{in}} + d\, r'_{\text{in}}\,, \\ r'_{\text{out}} &= r'_{\text{in}} \end{aligned} \quad \Rightarrow \quad \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \tag{1.79}$$

1.7.3.2 Reflection at a curved surface

From Fig. (1.18) it is clear that the ray displacement remains unchanged in reflection.

$$r_{\text{out}} = r_{\text{in}} \tag{1.80}$$

To find a relation between the input and output ray slopes we use the law of reflection, which leads to

$$\theta_2 = \theta_1 \qquad \text{(law of reflection)}$$

or $$\alpha_{\text{out}} - \phi = \phi - \alpha_{\text{in}}$$

or $$\alpha_{\text{out}} = 2\phi - \alpha_{\text{in}} \tag{1.81}$$

To relate these angles to ray slopes [Fig. 1.18] we note that r'_{in} is positive whereas r'_{out} is negative,

$$\begin{aligned} r'_{\text{in}} &= \alpha_{\text{in}}\,, \\ r'_{\text{out}} &= -\alpha_{\text{out}}\,, \\ \phi &= \frac{r_{\text{in}}}{R}\,, \end{aligned} \tag{1.82}$$

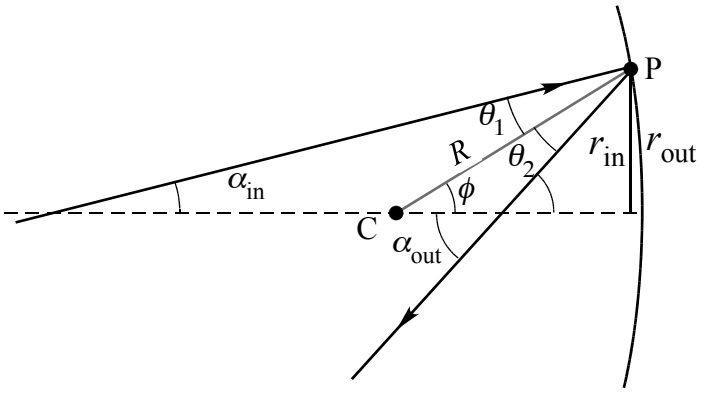


FIGURE 1.18
According to the laws of reflection the angles of incidence and reflection are equal $\theta_2 = \theta_1$. Note that r'_{out} is negative so that $r'_{\text{out}} = -\alpha_{\text{out}}$.

where we consider R to be *positive for a concave mirror (reflecting surface).* Using these, we find that the exit ray slope is given by

$$r'_{\text{out}} = -\frac{2r_{\text{in}}}{R} + r'_{\text{in}} \tag{1.83}$$

With the help of Eqs. (1.80) and (1.83) we obtain the ABCD matrix for a reflecting surface

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\frac{2}{R} & 1 \end{bmatrix} \tag{1.84}$$

1.7.3.3 Refraction at a curved interface.

From Fig. (1.19) we see that the input an output displacements are the same

$$r_{\text{out}} = r_{\text{in}}\,. \tag{1.85}$$

To relate input and output ray slopes we consider the relation between angles,

$$n_1\theta_1 = n_2\theta_2 \quad \text{Snell's law}$$

$$\text{or} \qquad n_1(\alpha_{\text{in}} + \phi) = n_2(\phi + \alpha_{\text{out}})$$

$$\text{or} \qquad \alpha_{\text{out}} = -\frac{(n_2 - n_1)}{n_2}\phi + \frac{n_1}{n_2}\alpha_{\text{in}} \tag{1.86}$$

Now noting $r'_{\text{in}} = \alpha_{\text{in}}$ and $r'_{\text{out}} = \alpha_{\text{out}}$ and $\phi = r_{\text{in}}/R$ where we choose R *positive for a convex refracting surface.* This leads us to

$$r'_{\text{out}} = -\frac{(n_2 - n_1)}{n_2}\frac{r_{\text{in}}}{R} + \frac{n_1}{n_2}r'_{\text{in}} \tag{1.87}$$

From Eqs. (1.85) and (1.87) we find the ABCD matrix

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\dfrac{n_2 - n_1}{n_2}\dfrac{1}{R} & \dfrac{n_1}{n_2} \end{bmatrix} \tag{1.88}$$

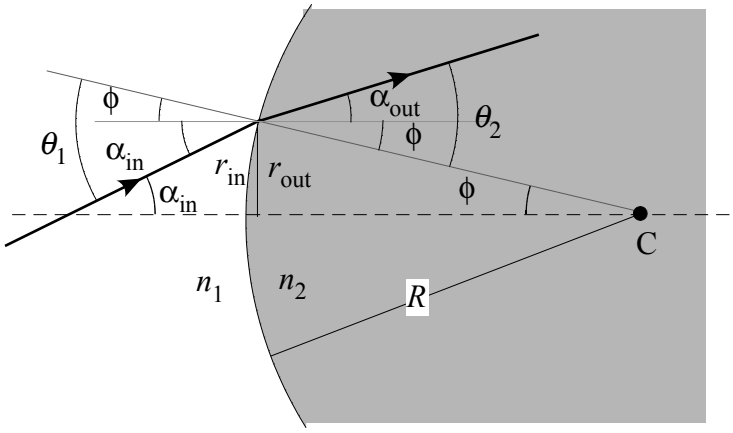


FIGURE 1.19
Refraction at a curved interface between two media. Note that in the figure $r'_{\rm out}$ is positive so that $r'_{\rm out} = \alpha_{\rm out}$.

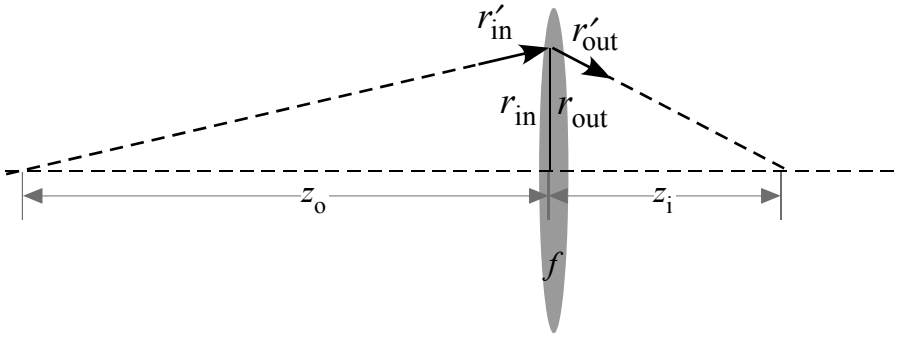


FIGURE 1.20
Relation between input and output ray parameters for a thin lens.

For a thin lens we can then find the ABCD matrix by multiplying the ABCD matrix for each of its surfaces,

$$\begin{bmatrix} 1 & 0 \\ -\dfrac{n_1 - n_2}{n_1}\dfrac{1}{R_2} & \dfrac{n_2}{n_1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\dfrac{n_2 - n_1}{n_2}\dfrac{1}{R_1} & \dfrac{n_1}{n_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\dfrac{n_2 - n_1}{n_1}\left(\dfrac{1}{R_1} - \dfrac{1}{R_2}\right) & 1 \end{bmatrix} \tag{1.89}$$

We can, of course, derive the matrix of a thin lens of focal length f directly with the help of Fig. 1.20. Let $r_{\rm in}$ and $r'_{\rm in}$ denote the displacement and slope of the incident ray just before the lens and $r_{\rm out}$ and $r'_{\rm out}$ their values just after the lens. Then it follows

$$r_{\rm out} = r_{\rm in}\,. \tag{1.90}$$

To determine the slope after the lens we note that the incident ray may be thought of as coming from the axial point z_o and the emergent ray may thought of as proceeding

towards the point z_i. These distances are related by the thin lens formula

$$\frac{1}{z_o} + \frac{1}{z_i} = \frac{1}{f}\,. \tag{1.91}$$

Multiplying both sides by $r_{\rm in}$ and noting that $r_{\rm in}/d_o = r'_{\rm in}$ and $r_{\rm in}/z_i = -r'_{\rm out}$ (the emergent ray has negative slope) and rearranging the terms we find the slope of the emergent ray

$$r'_{\rm out} = r_{\rm in} - \frac{1}{f} r_{\rm in}\,. \tag{1.92}$$

From the preceding examples, it is clear that we can write the relation between the input and output ray parameters in matrix form as

$$\begin{bmatrix} r_{\rm out} \\ r'_{\rm out} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} r_{\rm in} \\ r'_{\rm in} \end{bmatrix} \tag{1.93}$$

This relation represents a transformation of input ray parameters into putput ray parameters. The matrix of transformation, also called the *ABCD* matrix, depends on the nature of the optical element inside the black box. For example the ABCD matrix for propagation over a section of length L in a homogeneous medium is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \tag{1.94}$$

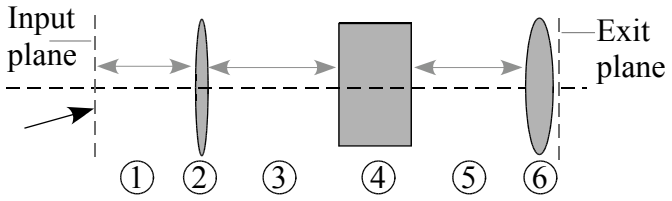


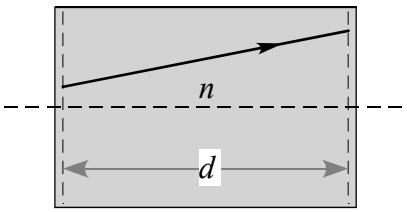
FIGURE 1.21
The overall transformation matrix for ray propagation through an optical system is obtained by multiplying the *ABCD* matrices for each optical element, including free space sections, in correct order [see the text].

ABCD matrices for a number of common optical elements are given in Table 1.3. Once these basic matrices are known we can calculate the overall ABCD matrix for any system of optical elements. Consider, for example, the passage of a ray through a sequence of optical elements shown in Fig. [a thin lens, followed by free space and a dilectric slab(refractive index n)]. By labeling various optical elelments as $1\,, 2\,, 3 \cdots$ in the order in which they are encountered, we can write the ABCD matrix as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = M_6 \cdot M_5 \cdots M_4 \cdot M_3 \cdot M_2 \cdot M_1\,. \tag{1.95}$$

TABLE 1.2
ABCD matrices for paraxial rays for three basic optical elements.

Straight section of length d in a homogeneous medium of refractive index n

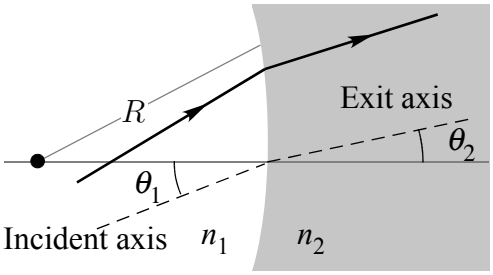


$$\begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}$$

Dielectric interface, radius of curvature R (+ for convex and − for concaved refracting surface), arbitrary angle of incidence

$\Delta n_T = \dfrac{n_2}{\cos\theta_1} - \dfrac{n_1}{\cos\theta_2}$,

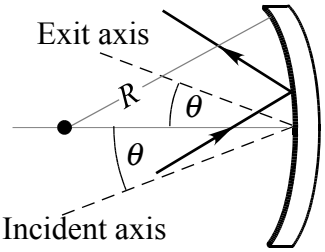
$\Delta n_S = n_2\cos\theta_2 - n_1\cos\theta_1$.



$$\begin{bmatrix} \dfrac{\cos\theta_2}{\cos\theta_1} & 0 \\ -\dfrac{\Delta n_T}{n_2 R} & \dfrac{n_1\cos\theta_1}{n_2\cos\theta_2} \end{bmatrix}$$ plane of incidence (tangential plane)

$$\begin{bmatrix} 1 & 0 \\ -\dfrac{\Delta n_S}{n_2 R} & \dfrac{n_1}{n_2} \end{bmatrix}$$ perpendicular to the plane of incidence (sagittal plane)

Spherical mirror of radius of curvature R (+ for concave mirror and − for convex mirror), arbitrary angle of incidence

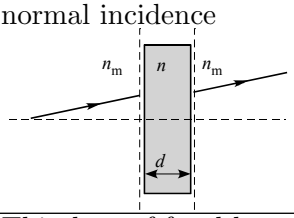


$$\begin{bmatrix} 1 & 0 \\ -\dfrac{2}{R\cos\theta} & 1 \end{bmatrix}$$ plane of incidence (tangential plane)

$$\begin{bmatrix} 1 & 0 \\ -\dfrac{2\cos\theta}{R} & 1 \end{bmatrix}$$ perpendicular to the plane of incidence (sagittal plane)

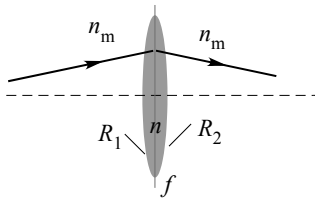
TABLE 1.3
ABCD matrices of some common optical elements.

Plate of thickness d, refractive index n in a medium of refractive index n_m, near normal incidence



$$\begin{bmatrix} 1 & \frac{d}{n} \\ 0 & 1 \end{bmatrix}$$

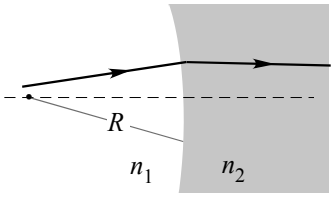
Thin lens of focal length f in a medium of refractive index n_m, normal incidence



$$\begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix}$$

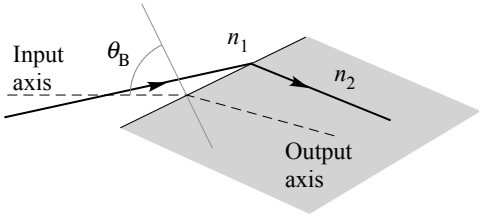
$$\frac{1}{f} = \frac{n - n_m}{n_m}\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

Dielectric interface with radius of curvature R (+ for convex and − for concave interface).



$$\begin{bmatrix} 1 & 0 \\ -\frac{n_2 - n_1}{n_2}\frac{1}{R} & \frac{n_1}{n_2} \end{bmatrix}$$

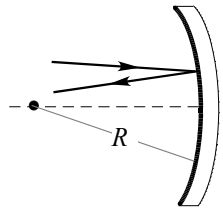
Dielectric interface at Brewster's angle.



$$\begin{bmatrix} \frac{n_2}{n_1} & 0 \\ 0 & \frac{n_1^2}{n_2^2} \end{bmatrix}$$ plane of incidence (tangential plane)

$$\begin{bmatrix} 1 & 0 \\ 0 & \frac{n_1}{n_2} \end{bmatrix}$$ perpendicular to the plane of incidence (sagittal plane)

Spherical mirror of radius of curvature R, normal incidence ($R > 0$ for concave mirror and $R < 0$ for convex mirror)



$$\begin{bmatrix} 1 & 0 \\ -\frac{2}{R} & 1 \end{bmatrix}$$

Note that the matrices are written from right to left in the order in which they are encountered by the ray.

1.7.4 Periodic Focusing System

An interesting and important application of ray matrices comes in the analyses of periodic focusing systems in which the same sequence of elements is repeated many times down a cascaded chain. An optical resonator can be modeled by such an iterated periodic focusing system because propagation through repeated round trips in the resonator is physically equivalent to propagation through repeated sections of a periodic lens guide.

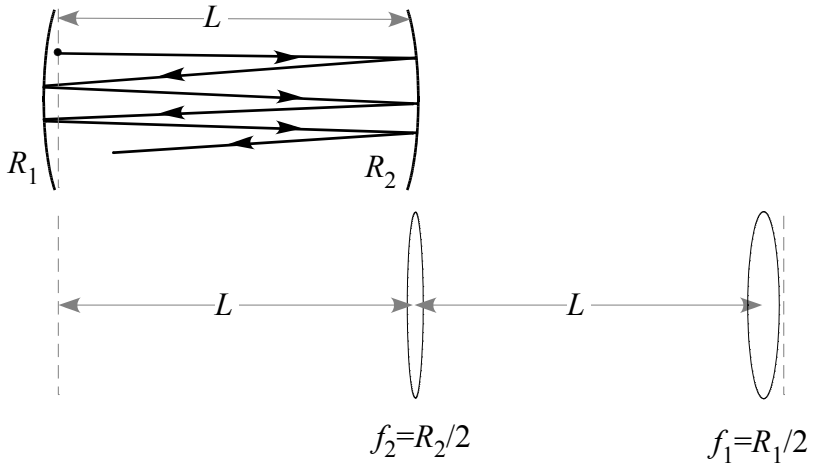


FIGURE 1.22
An optical cavity formed by two spherical mirrors of radii R_1 and R_2 separated by a distance L is equivalent to a periodic lens waveguide whose repeat unit is shown below the resonator.

As an example consider an optical resonator shown in Fig. 1.22 formed by two spherical mirrors of radii of curvature R_1 and R_2 placed a distance L apart. Imagine a ray propagating to the right starting at the left end of the resonator. After a round trip, this ray will have been transformed by a straight section of length L, a spherical mirror of radius of curvature R_1 another section of length L, and finally a spherical mirror of radius of curvature R_2. In each roundtrip, the ray encounters the same transformation. This is equivalent to a lens waveguide where lenses of focal length $f_1 = R_1/2$ and $f_2 = R_2/2$ are placed alternately separated by L. The ABCD matrix M describing the ray transformation in a roundtrip through the resonator is given by $M = M_1 \cdot M_L \cdot M_2 \cdot M_L$, where M_1 and M_2 are the ray matrices of mirrors R_1 and R_2, respectively, and M_L is the ray matrix of a section of length L. Note that the matrices are written from right to left in the order in which the optical element

was encountered:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\dfrac{2}{R_1} & 1 \end{bmatrix} \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\dfrac{2}{R_2} & 1 \end{bmatrix} \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 - \dfrac{2L}{R_2} & 2L\left(1 - \dfrac{L}{R_2}\right) \\ -\dfrac{2}{R_1} - \dfrac{2}{R_2} + \dfrac{4L}{R_1 R_2} & 1 - \dfrac{4L}{R_1} - \dfrac{2L}{R_2} + \dfrac{4L^2}{R_1 R_2} \end{bmatrix}$$

$$\equiv \begin{bmatrix} 2g_2 - 1 & 2Lg_2 \\ \frac{2}{L}(2g_1 g_2 - g_1 - g_2) & 4g_1 g_2 - 2g_2 - 1 \end{bmatrix}, \quad g_i \equiv 1 - \frac{L}{R_i} \tag{1.96}$$

After the nth roundtrip, the ray is given by

$$\widetilde{r}_n = M\,\widetilde{r}_{n-1} = M^2\,\tilde{r}_{n-2} = M^n\,\tilde{r}_o \tag{1.97}$$

A cascaded system such as this is best analyzed by finding the eigenvalues and eigen solutions of the matrix M, where M is the matrix of the repeated section. This means we look for a set of eigen-rays and eigenvalues Λ satisfying

$$M\widetilde{r} = \Lambda\widetilde{r} \qquad \text{or} \qquad \begin{bmatrix} A - \Lambda & B \\ C & D - \Lambda \end{bmatrix} \begin{bmatrix} r \\ r' \end{bmatrix} = 0 \tag{1.98}$$

Nontrivial solutions of this equation require the determinant of $M - \Lambda\,T$ must vanish:

$$det \begin{bmatrix} A - \Lambda & B \\ C & D - \Lambda \end{bmatrix} = \Lambda^2 - (A + D) + (AD - BC) = 0 \tag{1.99}$$

This equation can be simplified by noting that the determinant of matrix M is unity: $\det M \equiv AD - BC = 1$. The easiest way to check this is to recall that the determinant of the product of a set of matrices is the product of the determinants of individual matrices $\det(M) = \det(M_1)\det M_L)\det(M_2)\det(M_L)$ and that the determinant of each of these matrices is unity. We also note that the the coefficient of Λ is the trace of the matrix. By introducing a parameter m by

$$m = \frac{1}{2}\,\mathrm{Tr}\;M = \frac{1}{2}(A + D)\,, \tag{1.100}$$

we can solve Eq. (1.99) for Λ and write the two eigenvalues as

$$\Lambda_a = m + \sqrt{m^2 - 1}\,, \qquad \Lambda_b = m - \sqrt{m^2 - 1}\,. \tag{1.101}$$

We also have

$$\Lambda_a + \Lambda_b = 2m\,, \quad \text{and} \quad \Lambda_a \Lambda_b = 1\,. \tag{1.102}$$

There are two matching eigenrays $\widetilde{r}_a$ and $\widetilde{r}_b$ such that

$$M\widetilde{r}_a = \Lambda_a \widetilde{r}_a, \qquad M\widetilde{r}_b = \Lambda_b \widetilde{r}_b\,. \tag{1.103}$$

This equation means that the propagation of each eigen ray is specified simply by multiplying it by the corresponding eigenvalue.

In terms of these eigen rays, any input ray $\widetilde{r}_0$ can be written as

$$\widetilde{r}_0 = C_a\widetilde{r}_a + C_b\widetilde{r}_b\,, \tag{1.104}$$

where C_a and C_b are some constants which are determined by the intial ray displacement r_0 and slope r_0'. After n roundtrips this ray becomes

$$\widetilde{r}_n = M^n\widetilde{r}_0 = C_a\Lambda_a^n\widetilde{r}_a + C_b\Lambda_b^n\widetilde{r}_b\,. \tag{1.105}$$

Here we have used the fact that the propagation of each eigen ray is specified simply by multiplying it by the corresponding eigenvalue raised to the number of round trips.

We can now see from Eq. (1.105) that if either eigenvalue is greater than unity, the ray displacement will continue to grow with n. Thus whether a periodic array has stable ray trajectories or not depends on the nature of the eigenvalues. We can divide periodic focusing systems into stable and unstable systems depending on the nature of the eigenvalues which in turn depend on the parameter m via Eq.(1.101). If the parameter m^2 is less than unity, we have complex eigenvalues. If, on the other hand, $m^2 > 1$, we have real eigenvalues and at least one of them is greater than unity. Let us consider these cases separately.

Case 1: $m^2 < 1$ or $-\leq m \leq 1$

Since m is real and its magnitude is less than unity we can parametrize m and the eigenvalues as

$$m = \frac{1}{2}(A+D) \equiv \cos\theta\,, \tag{1.106}$$

$$\Lambda_{a,b} = \cos\theta \pm i\sqrt{1-\cos^2\theta} = \cos\theta \pm i\sin\theta = e^{\pm i\theta}\,. \tag{1.107}$$

Then ray propagationin the periodic system takes the form

$$\widetilde{r}_n = C_a e^{in\theta}\widetilde{r}_a + C_b e^{-in\theta}\widetilde{r}_b = \widetilde{c}_0\cos n\theta + \widetilde{s}_0\sin n\theta\,, \tag{1.108}$$

where

$$\widetilde{c}_0 = C_a\widetilde{r}_a + C_b\widetilde{r}_b \quad \text{and} \quad \widetilde{s}_0 = i(C_a\widetilde{r}_a - C_b\widetilde{r}_b)\,. \tag{1.109}$$

The displacment after n round-trips is then bounded function

$$r_n = c_0\cos n\theta + s_0\sin n\theta\,. \tag{1.110}$$

Thus a periodic focusing system is stable if $|m| < 1$. Rays in this system will oscillate back and forth about the axis as shown in Fig. (1.23). Note that r_n only gives the displacement at the successive reference planes; it does not say anything about what happens to the ray inside the repeated section between the reference planes. Note that n, the number of trips through the repeated section, is the variable. For the ray trajectory to be periodic, θ must be a rational fraction (< 1) of 2π. In other words,

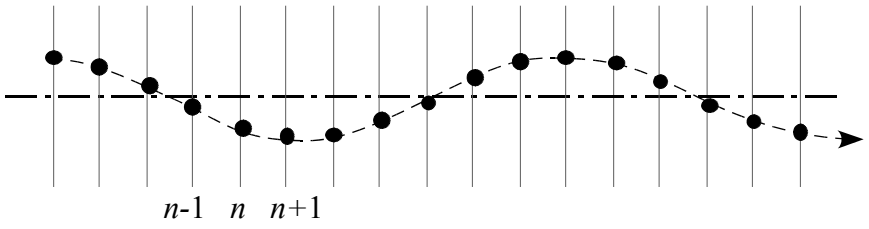


FIGURE 1.23
The dots indicate ray position at successive reference planes. The continuous curve joining the dots is simply a guide to the eye.

the condition $N\theta = 2\pi s$, where N and s ($s < N$) are relative prime positive integers, must be satisfied. N can then be interpreted as the number of trips through the repeated section after which the ray retraces its trajectory.
Case 2: $m^2 > 1\,(m \leq -1 \,\text{or}\, m \geq 1)$.
In this case the eigenvalues of M are real given by

$$\Lambda_a = m + \sqrt{m^2 - 1} \equiv \Lambda\,, \tag{1.111a}$$

$$\Lambda_b = m - \sqrt{m^2 - 1} \equiv \frac{1}{\Lambda}\,. \tag{1.111b}$$

It can be seen that one of the eigenvalues has a magnitude greater than unity. Then after n steps the ray coordinates are given by

$$\widetilde{r}_n = C_a\, e^{n\theta}\widetilde{r}_a + C_b\, e^{-n\theta}\widetilde{r}_b\,, \quad \theta = \ln\Lambda \tag{1.112}$$

The ray displacement

$$r_n = C_a\, e^{n\theta} r_a + C_b\, e^{-n\theta} r_b \tag{1.113}$$

grows with each step. For $\Lambda > 1$ the ray displacement grows monotonically whereas for $\Lambda < -1$ ($\ln\Lambda = \ln|\Lambda| + i\pi$ for $\Lambda < -1$) it grows in an oscillatory fashion as seen in Fig. (1.24). Periodic systems with $m > 1$ are clearly unstable.

1.7.4.1 Two-Mirror Resonator Stability

For a two-mirror cavity, the round trip ABCD matrix is given by Eq. (1.96). The parameter $m = \frac{1}{2}(A + D)$ for a two mirror cavity is then

$$m = 1 - \frac{2L}{R_1} - \frac{2L}{R_2} + \frac{2L^2}{R_1 R_2} = 2\left(1 - \frac{L}{R_1}\right)\left(1 - \frac{L}{R_2}\right) - 1 \tag{1.114}$$

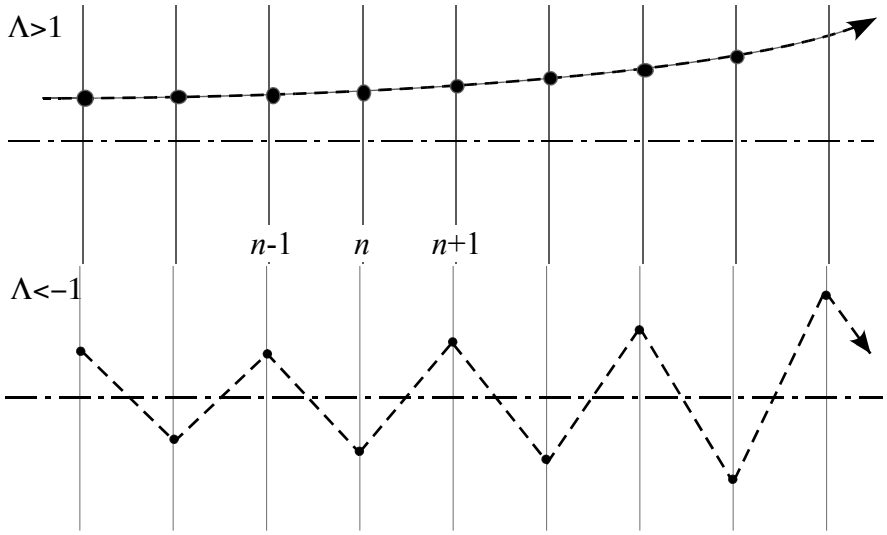


FIGURE 1.24
In an unstable periodic array the ray displacement increases without limit either monotonically or in an oscillatory fashion. Only the dots are meaningful. The continuous curve is simply a guide to the eye.

The condition for stability, $-1 \leq m \leq 1$, then implies

$$\begin{aligned} -1 \leq & 2\left(1-\frac{L}{R_1}\right)\left(1-\frac{L}{R_2}\right)-1 \leq 1 \\ 0 \leq & 2\left(1-\frac{L}{R_1}\right)\left(1-\frac{L}{R_2}\right) \leq 2 \\ 0 \leq & \left(1-\frac{L}{R_1}\right)\left(1-\frac{L}{R_2}\right) \leq 1 \end{aligned} \tag{1.115}$$

On introducing the parameters

$$g_1 = 1-\frac{L}{R_1}\,, \qquad g_2 = 1-\frac{L}{R_2}\,. \tag{1.116}$$

the stability condition can be written more compactly as

$$0 \leq g_1 g_2 \leq 1\,. \tag{1.117}$$

This important relation specifies the range of stability of two-mirror optical resonators. A plot of this inequality is shown in Fig. (1.25).

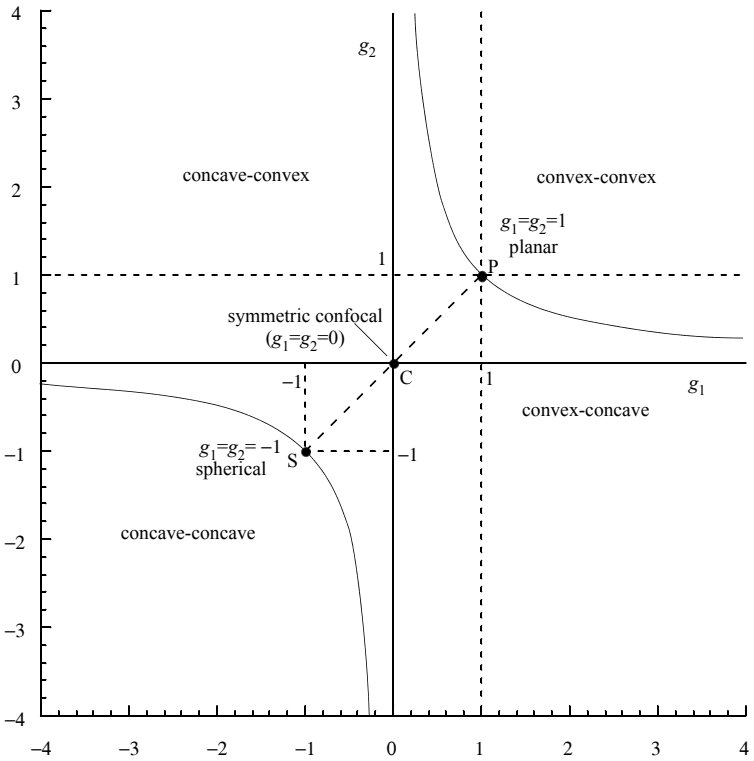


FIGURE 1.25
Stable resonators lie in the region bounded the coordinate axes and the two branches of hyperbola $g_1 g_2 = 1$. Symmetric resonators lie along the line $g_2 = g_1$.

In computing the values of various physical quantities, the values of fundamental constants given in Table 1.4 will be useful. In practice, it is more convenient to remember certain combinations [Table 1.5] of fundamental constants rather than the constants themselves. For example, the energy of a photon can be calculated by using the formula

$$E \equiv h\nu = \frac{2\pi\hbar c}{\lambda} = \frac{1238}{\lambda(\text{in nm})}\,\text{eV}\,.$$

TABLE 1.4
Some fundamental constants

Constant Name	Symbol	Current Value
Speed of light	c	$2.997\,924\,58\times10^{8}$ m/s
Elementary charge	e	$1.602\,177\times10^{-19}$ C
Planck's constant	$\hbar$	$1.054\,571\,628(53)\times10^{-34}$ J·s
Gravitational constant	G	$6.67428(67)\times10^{-11}$ N·m^2/kg^2
Stefan-Boltzmann constant	σ	5.6705×10^{-8} W/m^2·K^4
Boltzmann constant	k_B	$1.380\,7\times10^{-23}$ J/K
Avogadro's number	N_A	$6.022\,137\times10^{23}$ mol^{-1}
Solar mass	$M_s\odot$	1.989×10^{30} kg
Solar radius	R_s	6.96×10^{8} m
Earth mass	M_e	5.976×10^{24} kg
Earth radius	R_e	6.374×10^{6} m
Earth-Sun mean distance	AU	1.496×10^{11} m
Earth-Moon mean distance		3.844×10^{8} m
Ideal gas molar volume at STP		22.414 liter
Air density at STP		1.293 kg/m^3
Triple point of water		273.16 K

TABLE 1.5
Some useful conversion factors

Fine structure constant $\alpha=\dfrac{e^2}{4\pi\epsilon_0\hbar c}$	$1/137.036\approx1/137$
$\hbar c$	197.3271 eV· nm
k_BT	1/40 eV at 293K and 1/39 eV at 300K
Electron rest energy (mass) m_ec^2	0.5110 MeV
Proton rest energy (mass) m_pc^2	938.28 MeV
Neutron rest energy m_nc^2	939.57 MeV
Proton-electron mass ratio m_p/m_e	1836.15
Bohr radius $a_0=\hbar/m_ec\alpha$	0.5292×10^{-10} m
Planck time $\sqrt{\hbar G/c^5}$	5.4×10^{-44} s
Compton wavelength of electron $\hbar/m_ec$	3.8616×10^{-13} m
1 calorie	4.1860 J
1 atm	1.1013×10^{5} Pa
1 degree	1.745×10^{-2} rad
1 parsec	3.09×10^{16} m