PC2232 Physics for Electrical Engineers
AY2014/15 Semester II
Part I: Modern Optics

Mankei Tsang

E-mail address: mankei@nus.edu.sg

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING, NATIONAL UNIVERSITY OF SINGAPORE, 4 ENGINEERING DRIVE 3, SINGAPORE 117583

DEPARTMENT OF PHYSICS, NATIONAL UNIVERSITY OF SINGAPORE, 2 SCIENCE DRIVE 3, SINGAPORE 117551

PLEASE NOTE.

- Important equations are $\boxed{boxed}$. If you understand what they mean and where they come from you will probably do well. Simply memorizing them will not help much.
- * denotes advanced topics. They won't be the focus of exams but it is good to read them to gain a deeper understanding.
- There are **Exercises** throughout these lecture notes. They are probably more difficult than the tutorials and the exams but can be used to test and enhance your understanding of the discussions around them. **Answers** are given if they are relevant to future discussions. **Questions** are similar but more conceptual problems.
- Please **do not share** any lecture material, including these notes, **outside the class**; they may contain copyrighted material not suitable for circulation.
- Notations:
  - We follow the engineering convention and use j for the imaginary number. Note that physics literature often uses i instead. They mean the same thing.
  - Vectors are denoted by $\boldsymbol{bold\ italic}$ face.
  - Unit vectors are denoted by a hat $\hat{\boldsymbol{e}}$ on a vector.
  - Time average is denoted by an $\overline{overline}$.
  - Some symbols are recycled and may have different meanings in different contexts. For example, V may mean volume in one context and voltage in another, θ is used many times to denote an angle, ϕ is commonly used to denote an arbitrary phase, n usually means refractive index but may also be used to mean an integer index in a sum, while $\hat{\boldsymbol{n}}$ is usually a normal unit vector.
  - There may be similar-looking symbols such as v (speed) and ν (greek letter "nu," used to denote frequency in Hertz).

CHAPTER 1

Review of Electromagnetism and Vector Calculus

Optics is the science of **light**, which can come from the sun, fluorescent lamps, your computer or smartphone LCD displays, light-emitting diodes (LED), laser pointers, or laser diodes in your optical drives (CD, DVD, Blu-ray, etc.). As discovered by Maxwell (http://en.wikipedia.org/wiki/James_Clerk_Maxwell), light is **electromagnetic waves**, which obey the four **Maxwell's equations**. Before we go into the equations, let's review the basic physical quantities in electromagnetism.

1.1. Charge and charge density

The first quantity is **charge** possessed by charged particles, such as electrons and protons. The unit of charge in MKS units (also called SI units, http://en.wikipedia.org/wiki/MKS_system_of_units) is **Coulomb** (C), and each electron has a charge given by $-e$, where

$$e \approx 1.602 \times 10^{-19} \text{ C}. \tag{1.1}$$

(do not confuse this with the natural number used in an exponential). For historical reasons, the electron charge is negative. (http://en.wikipedia.org/wiki/Elementary_charge)

Charges in general are distributed in three-dimensional space and may vary in time. We denote the position vector in Cartesian coordinates by

$$\boldsymbol{r} = x\hat{\boldsymbol{x}} + y\hat{\boldsymbol{y}} + z\hat{\boldsymbol{z}}, \tag{1.2}$$

where $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$, $\hat{\boldsymbol{z}}$ are unit vectors in the three dimensions, and x, y, z are the three components of the vector along each direction, as shown in Fig. 1.1. We shall use ***bold italic*** face to denote a vector and the hat $\hat{}$ to denote a unit vector.

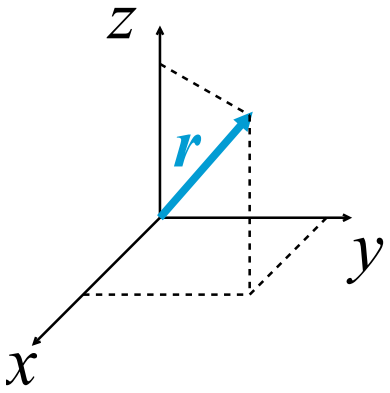


FIGURE 1.1. Position vector $\boldsymbol{r}$ in Cartesian coordinates.

To describe the variation of charges in space at each time, we use another quantity called **charge density** $\rho(\boldsymbol{r}, t)$:

$$\rho(\boldsymbol{r}, t) = \text{charge density (C/m}^3\text{)}. \tag{1.3}$$

Note that its unit is charge (Coulomb) per volume. To understand its meaning, consider a very small box with lengths dx, dy, and dz near position $\boldsymbol{r}$, as shown in Fig. 1.2. The volume of the box is $dV = dxdydz$, and if there is charge dQ inside the box, the charge density is

$$\rho(\boldsymbol{r}, t) = \frac{dQ}{dV}. \tag{1.4}$$

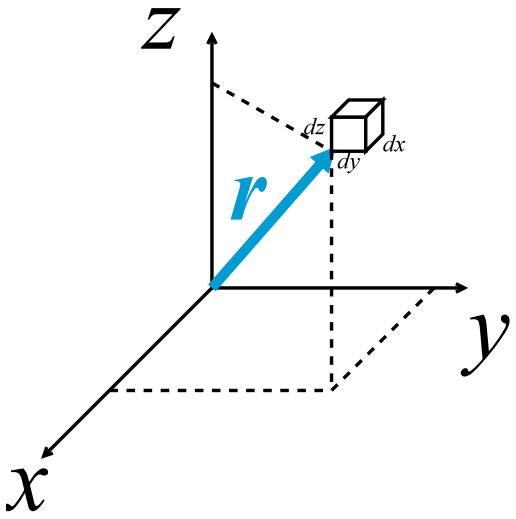


FIGURE 1.2. if there is charge dQ in a very small box at position $\boldsymbol{r}$ with volume $dV = dxdydz$, the charge density $\rho(\boldsymbol{r}, t)$ is dQ/dV.

Now let's consider a volume in three-dimensional space, as shown in Fig. 1.3, and the charge density inside this volume is $\rho(\boldsymbol{r}, t)$. Think of this volume as consisting of many small boxes, each box containing a charge given by

$$dQ(\boldsymbol{r}, t) = \rho(\boldsymbol{r}, t)dV = \rho(\boldsymbol{r}, t)dxdydz. \tag{1.5}$$

The total charge inside this volume is then the sum of all these charges. We take the continuous limit, assume each box is **infinitesimally** small, and instead of a sum we do the **volume integral**:

$$Q \text{ inside the volume} = \int_{\boldsymbol{r} \text{ inside the volume}} dQ(\boldsymbol{r}, t) = \int_{\boldsymbol{r} \text{ inside the volume}} \rho(\boldsymbol{r}, t)dV. \tag{1.6}$$

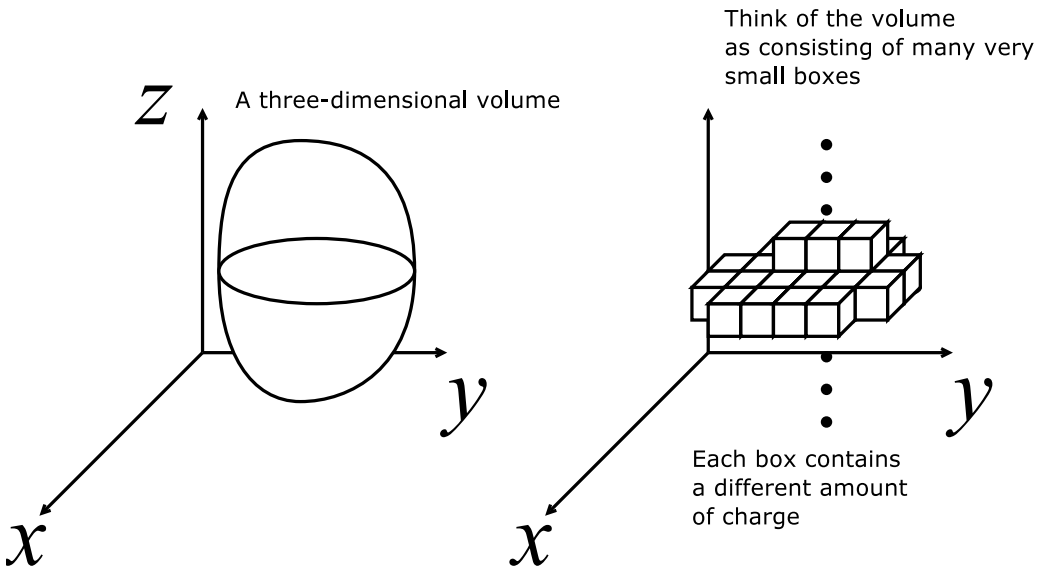


FIGURE 1.3. Think of a volume as consisting of many very small boxes. Each box contains a charge denoted by $dQ(\boldsymbol{r}, t)$, and the total charge is the sum of all the charges inside the volume.

1.2. Current and current density

The next quantity of interest is the **current** I, the unit is **Ampere**, which is equal to Coulomb per second. It is the amount of charge that passes through a surface per second. We shall focus on the **current density** $\boldsymbol{J}$:

$$\boldsymbol{J}(\boldsymbol{r}, t) = \text{current density (A/m}^2\text{)}, \tag{1.7}$$

which is the current passing through a surface per unit area. Note that it is a vector quantity, as depicted in Fig. 1.4, as charges can move in different velocities. In Cartesian coordinates, it can be written as

$$\boldsymbol{J} = J_x\hat{\boldsymbol{x}} + J_y\hat{\boldsymbol{y}} + J_z\hat{\boldsymbol{z}}. \tag{1.8}$$

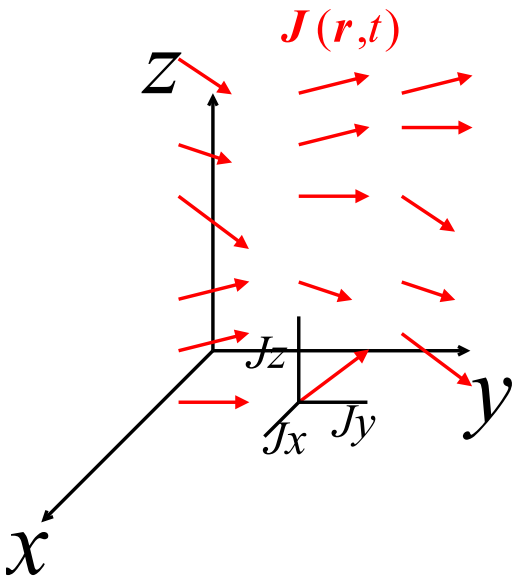


FIGURE 1.4. The current density $\boldsymbol{J}(\boldsymbol{r}, t)$ is a vector field. It is a vector at each point in space and also varies in time.

To understand the concept of current density, consider the small surface with area $dA = dxdz$ in Fig. 1.5. Suppose that the charge density ρ is constant near the surface and the charges there are moving at a velocity in the y direction. The velocity is assumed to be $\boldsymbol{v} = v\hat{\boldsymbol{y}}$, and the current density turns out to be given by

$$\boldsymbol{J}(\boldsymbol{r}, t) = \rho\boldsymbol{v} = \rho v\hat{\boldsymbol{y}}. \tag{1.9}$$

To see this, consider a surface perpendicular to the velocity. At time t, consider a volume of charge $dQ = \rho dxdydz$ just before the surface, and suppose that, after a certain time interval dt, the volume has completely passed the surface. The distance that the charge has to travel is dy, and the time interval is then $dt = dy/v$. The current, or charge passing through the surface per second, is

$$dI = \frac{dQ}{dt} = \frac{\rho dxdydz}{dt} = \rho v dxdz = JdA, \tag{1.10}$$

where $J = dI/dA$ is the magnitude of $\boldsymbol{J}$.

For a slightly less trivial example, consider a uniform charge density with the same velocity, but the surface is now at an angle θ, as shown in Fig. 1.6. The surface area is now

$$dA = \frac{dxdz}{\cos\theta}, \tag{1.11}$$

because one of the lengths of the surface is now longer and given by $dz/\cos\theta$. What is the current?

A way to think about it is to consider a parallelogram of charge in Fig. 1.6. Its volume is still $dxdydz$, and since we already assumed ρ is constant everywhere, the charge in this volume is still $dQ = \rho dxdydz$. The time it takes for the volume to pass through the surface is still $dt = dy/v$, and we finally obtain

$$dI = \frac{dQ}{dt} = \frac{\rho dxdydz}{dt} = \rho v dxdz = \rho v\cos\theta dA. \tag{1.12}$$

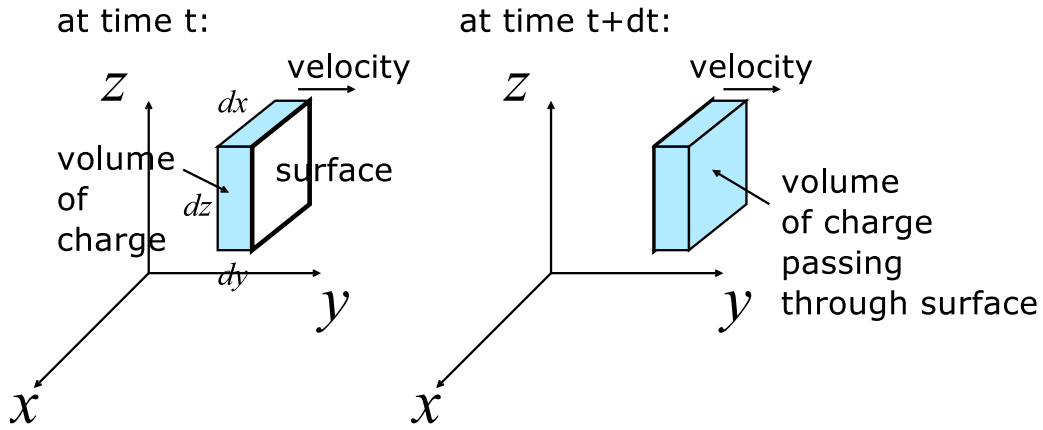


FIGURE 1.5. Current passing through a surface when the surface is perpendicular to the velocity.

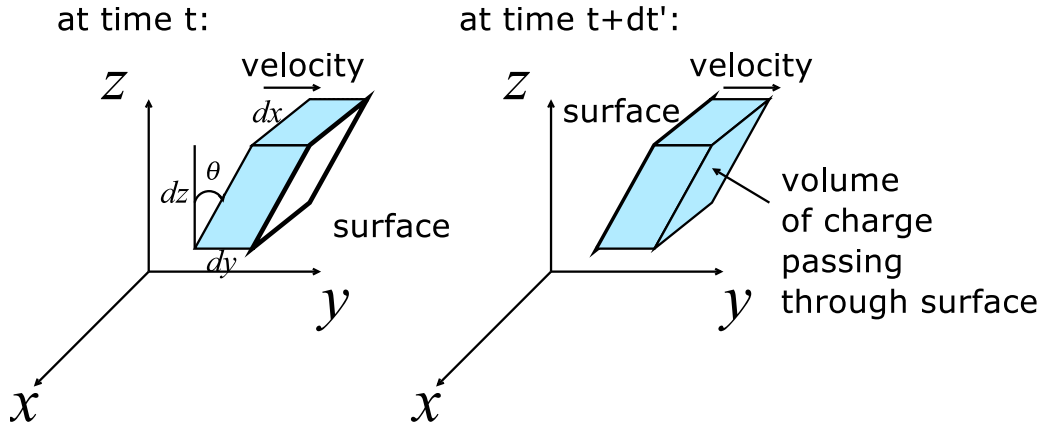


FIGURE 1.6. Current passing through a surface when the surface is at an angle θ with respect to the velocity.

The important point to note here is that current is now a **fraction** of the current density magnitude times the surface area, and the fraction is determined by the angle between the velocity and the surface.

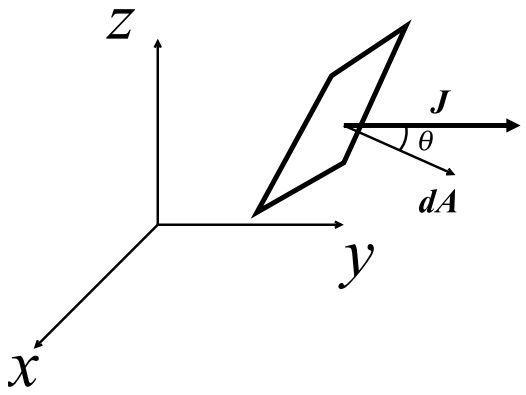


FIGURE 1.7. An infintesimal surface is associated with a magnitude given by its surface area dA and a direction normal to the surface.

In general, we can assign a direction to a surface in terms of its **normal vector**. The surface area differential $d\boldsymbol{A}$ then has a magnitude given by its area and a direction that is normal to the surface, as shown in Fig. 1.7. For example, you should convince yourself that the angle in Fig. 1.6 is the same as the angle between the normal

vector for the surface and the y direction. The current passing through this surface in general is

$$dI = \boldsymbol{J} \cdot d\boldsymbol{A} = J \cos\theta dA, \tag{1.13}$$

where $\cdot$ is the dot product and θ is the angle between the two vectors.

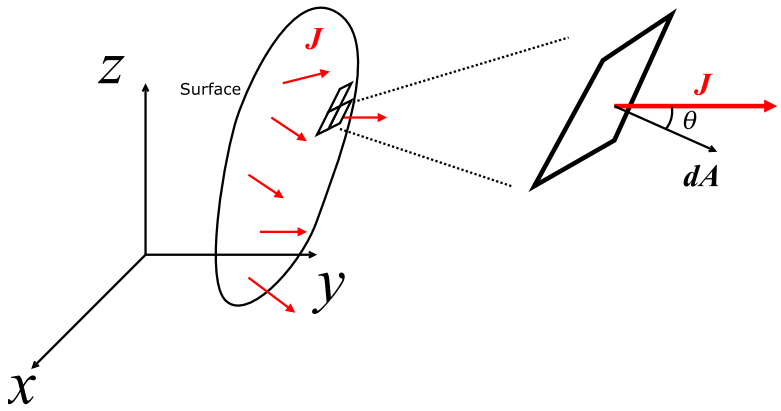


FIGURE 1.8. Think of a surface as consisting of many very small squares, and the current through a surface is the sum of currents through all the squares.

It should now be obvious that the current through any surface is the sum of all currents through all the infinitesimal surfaces, as shown in Fig. 1.8, and we can use a **surface integral** to compute it:

$$\text{current through a surface} = \int_{\boldsymbol{r} \text{ on surface}} dI(\boldsymbol{r}, t) = \int_{\boldsymbol{r} \text{ on surface}} \boldsymbol{J}(\boldsymbol{r}, t) \cdot d\boldsymbol{A}. \tag{1.14}$$

1.3. Electric field and Gauss's law

Charges and currents generate electromagnetic fields around them. First consider the electric field $\boldsymbol{E}(\boldsymbol{r}, t)$:

$$\boxed{\boldsymbol{E}(\boldsymbol{r}, t) = \text{Electric field (V/m).}} \tag{1.15}$$

Its unit is voltage (Volt) per length (meter). It is also a vector field. In Cartesian coordinates, we can write

$$\boldsymbol{E} = E_x \hat{\boldsymbol{x}} + E_y \hat{\boldsymbol{y}} + E_z \hat{\boldsymbol{z}}. \tag{1.16}$$

The electric field obeys **Gauss's law** with respect to the charge density $\rho(\boldsymbol{r}, t)$:

$$\boxed{\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}, \quad \text{(Gauss's law)}} \tag{1.17}$$

where $\nabla\cdot$ is called the divergence and ∇ is called the vector differential operator. In Cartesian coordinates, we can write is as

$$\nabla \equiv \frac{\partial}{\partial x}\hat{\boldsymbol{x}} + \frac{\partial}{\partial y}\hat{\boldsymbol{y}} + \frac{\partial}{\partial z}\hat{\boldsymbol{z}}. \tag{1.18}$$

We can treat ∇ like a vector, except that it must be in front of an actual vector. For example,

$$\nabla \cdot \boldsymbol{E} = \left(\frac{\partial}{\partial x}\hat{\boldsymbol{x}} + \frac{\partial}{\partial y}\hat{\boldsymbol{y}} + \frac{\partial}{\partial z}\hat{\boldsymbol{z}}\right) \cdot \left(E_x \hat{\boldsymbol{x}} + E_y \hat{\boldsymbol{y}} + E_z \hat{\boldsymbol{z}}\right),$$

note that the unit vectors are all orthogonal to one another, so their dot products are zero unless they are the same ($\hat{\boldsymbol{x}} \cdot \hat{\boldsymbol{x}} = \hat{\boldsymbol{y}} \cdot \hat{\boldsymbol{y}} = \hat{\boldsymbol{z}} \cdot \hat{\boldsymbol{z}} = 1$). This gives us

$$\left(\frac{\partial}{\partial x}\hat{\boldsymbol{x}} + \frac{\partial}{\partial y}\hat{\boldsymbol{y}} + \frac{\partial}{\partial z}\hat{\boldsymbol{z}}\right) \cdot \left(E_x \hat{\boldsymbol{x}} + E_y \hat{\boldsymbol{y}} + E_z \hat{\boldsymbol{z}}\right) = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}. \tag{1.19}$$

Gauss's law says that the divergence of the electric field is proportional to the charge density. The divergence theorem (http://en.wikipedia.org/wiki/Divergence_theorem) says that the volume integral of a divergence is given by the surface integral:

$$\int_{\text{volume}} \nabla \cdot \boldsymbol{E} dV = \oint_{\text{surface around volume}} \boldsymbol{E} \cdot d\boldsymbol{A}, \tag{1.20}$$

where $d\boldsymbol{A}$ is assumed to point out of the volume, and we know that

$$\int_{\text{volume}} \rho(\boldsymbol{r}, t) dV = Q \text{ inside the volume}, \tag{1.21}$$

so the integral form of Gauss's law is

$$\oint \boldsymbol{E} \cdot d\boldsymbol{A} = \frac{Q}{\epsilon_0}. \tag{1.22}$$

The surface integral of a vector field is called the flux, and $\int \boldsymbol{E} \cdot d\boldsymbol{A}$ is the electric flux. Gauss's law means that the net electric flux out of a volume is proportional to the charge inside the volume. The electric field must stick out like a sea urchin if there is positive charge inside, but pointing inwards if there is negative charge, as shown in Fig. 1.9.

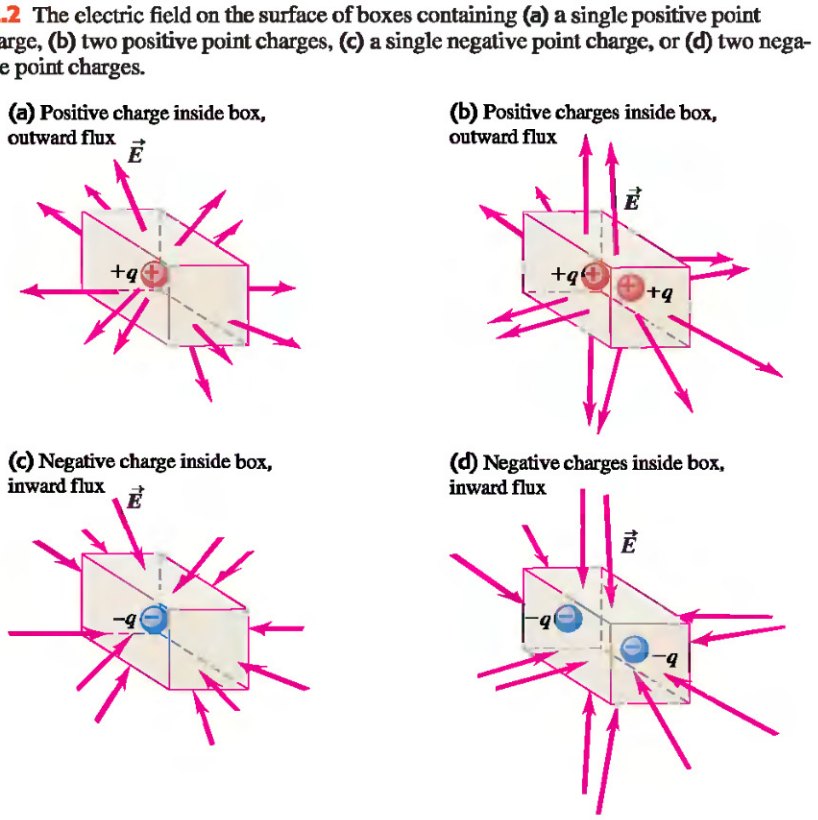


FIGURE 1.9. From Chap. 22, Ref. [1].

For example, a solution of electric fields from a point charge with charge Q and position $\boldsymbol{r} = 0$ is

$$\boldsymbol{E} = \frac{Q}{4\pi\epsilon_0 r^2}\hat{\boldsymbol{r}}, \tag{1.23}$$

where $r = \sqrt{x^2 + y^2 + z^2} = |\boldsymbol{r}|$ is the radial position and $\hat{\boldsymbol{r}}$ is the radial direction.[1] You can verify that this is correct by noting that the electric field is constant with respect to the radial position r, the surface area of a sphere with center at the origin is $4\pi r^2$, and the enclosed charge of the sphere is always Q no matter how small r is, since we are assuming a point charge.

ϵ_0 is a constant that relates the charge to the electric field and called the **free-space permittivity**:

$$\boxed{\epsilon_0 = \text{free-space permittivity} \approx 8.854 \times 10^{-12} \text{ F/m}.} \tag{1.24}$$

F stands for Farad, the unit for capacitance, which is charge per voltage or Coulomb per Volt.

[1] *Note, however, that this is just one solution. We have not shown that this is the only solution, and in fact there can be more solutions that consist of this electrostatic solution plus some electromagnetic waves.

- **Exercise**: From Gauss's law, check that ϵ_0 indeed has the unit F/m.

1.4. Magnetic field and Gauss's law for magnetism

The **magnetic field** is

$$\boxed{\boldsymbol{H}(\boldsymbol{r},t) = \text{Magnetic field (A/m).}} \tag{1.25}$$

The unit is Ampere per meter. Here we adopt the engineering convention and assume that $\boldsymbol{H}$ is the fundamental magnetic field, but note that physicists prefer to multiply it by a constant and call

$$\boldsymbol{B} = \mu_0 \boldsymbol{H} \quad \text{(T)}, \tag{1.26}$$

with the unit Tesla, the fundamental magnetic field. One can use either of these in practical calculations with no difference in the result as long as one is consistent. μ_0 is called the **free-space permeability**:

$$\boxed{\mu_0 = \text{free-space permeability} = 4\pi \times 10^{-7}\ \text{Tm/A},} \tag{1.27}$$

and we will see it again later.

As far as we know, there is no magnetic charge, and the magnetic field obeys the **Gauss's law for magnetism**:

$$\boxed{\nabla \cdot \boldsymbol{H} = 0, \quad \text{(Gauss's law for magnetism)}} \tag{1.28}$$

which is the same as the electric-field version except that the magnetic charge is zero. The integral form is

$$\oint \boldsymbol{H} \cdot d\boldsymbol{A} = 0. \tag{1.29}$$

The surface integral of the magnetic field $\int \boldsymbol{H} \cdot d\boldsymbol{A}$ is called the magnetic flux. Gauss's law for magnetism means that the total magnetic flux in or out of a closed volume must be zero.

1.5. Faraday's law

Besides the presence of a charge, a changing magnetic field can also lead to an electric field. This rule is governed by **Faraday's law**:

$$\boxed{\nabla \times \boldsymbol{E} = -\mu_0 \frac{\partial \boldsymbol{H}}{\partial t}, \quad \text{(Faraday's law)}} \tag{1.30}$$

The right-hand side is proportional to the rate of change of the magnetic field $\boldsymbol{H}$ in time, and the left-hand side is the **curl** of the electric field. It can be thought of as the cross product between ∇ and $\boldsymbol{E}$; in Cartesian coordinates it can be written as a determinant:

$$\nabla \times \boldsymbol{E} = \begin{vmatrix} \hat{\boldsymbol{x}} & \hat{\boldsymbol{y}} & \hat{\boldsymbol{z}} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ E_x & E_y & E_z \end{vmatrix} = \left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right)\hat{\boldsymbol{x}} + \left(\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x}\right)\hat{\boldsymbol{y}} + \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right)\hat{\boldsymbol{z}}. \tag{1.31}$$

The Stokes's theorem (http://en.wikipedia.org/wiki/Stokes_theorem) says that the surface integral of a curl is the line integral:

$$\int (\nabla \times \boldsymbol{E}) \cdot d\boldsymbol{A} = \oint_{\text{line around surface}} \boldsymbol{E} \cdot d\boldsymbol{l}, \tag{1.32}$$

where the right-hand rule is assumed to determine the direction of $d\boldsymbol{l}$ with respect to $d\boldsymbol{A}$. The integral form of Faraday's law is

$$\oint \boldsymbol{E} \cdot d\boldsymbol{l} = -\mu_0 \frac{\partial}{\partial t} \int \boldsymbol{H} \cdot d\boldsymbol{A}. \tag{1.33}$$

Again, $\int \boldsymbol{H} \cdot d\boldsymbol{A}$ is the magnetic flux. Qualitatively speaking, this says that a change of the magnetic field in time will induce an electric field that goes around it, as sketched on the left of Fig. 1.10.

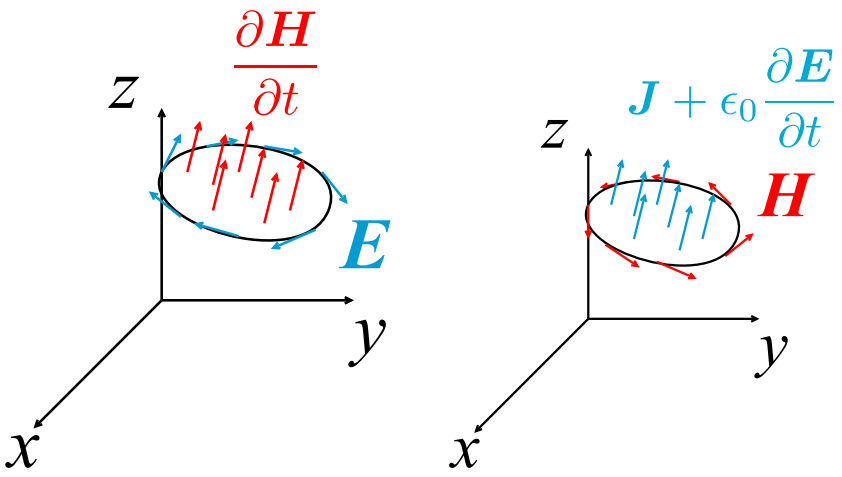


FIGURE 1.10. Left: Meaning of Faraday's law: A change of magnetic field $\boldsymbol{H}$ in time induces an electric field $\boldsymbol{E}$ that goes around it. Right: Meaning of modified Ampere's law: A current density $\boldsymbol{J}$, or a change of the electric field $\boldsymbol{E}$ in time, induces a magnetic field $\boldsymbol{H}$ that goes around them.

1.6. Modified Ampere's law

Like yin and yang, a changing electric field can also induce a magnetic field, as governed by the **modified Ampere's law**,

$$\boxed{\nabla \times \boldsymbol{H} = \boldsymbol{J} + \epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t}, \quad \text{(Modified Ampere's law)}} \tag{1.34}$$

where $\boldsymbol{J}$ is the current density we have studied in Sec. 1.2. Ampere's law is $\nabla \times \boldsymbol{H} = \boldsymbol{J}$, and the extra term $\epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t}$ was added by Maxwell. This extra term is also called the displacement current, as it acts like a current and creates a magnetic field around it, as sketched on the right of Fig. 1.10.

The integral form of modified Ampere's law is

$$\oint \boldsymbol{H} \cdot d\boldsymbol{l} = I + \epsilon_0 \frac{\partial}{\partial t} \int \boldsymbol{E} \cdot d\boldsymbol{A}. \tag{1.35}$$

At this stage, you should be familiar with what each term means: $\oint \boldsymbol{H} \cdot d\boldsymbol{l}$ is the line integral of $\boldsymbol{H}$ around a surface, I is the current that passes through the surface, and $\int \boldsymbol{E} \cdot d\boldsymbol{A}$ is the electric flux across the surface.

1.7. *Charge conservation

If we take the divergence of modified Ampere's law given by Eq. (1.34) and noting that the divergence of the curl is zero, we get

$$0 = \nabla \cdot \boldsymbol{J} + \epsilon_0 \nabla \cdot \frac{\partial \boldsymbol{E}}{\partial t}, \tag{1.36}$$

We can always interchange the order of time and space partial derivatives,

$$\nabla \cdot \frac{\partial \boldsymbol{E}}{\partial t} = \frac{\partial}{\partial t} \left(\nabla \cdot \boldsymbol{E} \right), \tag{1.37}$$

and from Gauss's law,

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}. \tag{1.38}$$

Combining the three equations in this section together,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \boldsymbol{J} = 0. \tag{1.39}$$

It relates the charge density ρ to the current density $\boldsymbol{J}$. It is called the charge continuity equation. To understand what it means, let's take the volume integral,

$$\frac{\partial}{\partial t}\int \rho dV = \frac{\partial Q}{\partial t} = -\int \nabla\cdot\boldsymbol{J}dV = -\oint \boldsymbol{J}\cdot d\boldsymbol{A}. \tag{1.40}$$

It simply means that the rate of change of charge inside a volume is equal to the net current flowing into the volume ($\oint \boldsymbol{J}\cdot d\boldsymbol{A}$ is the net current out of the volume, so the negative of that is the current into the volume). It is nothing but a statement of **charge conservation**; which wouldn't be possible if we didn't use the modified form of Ampere's law.

1.8. Maxwell's equations in free space

Let us collect all four of the Maxwell's equations here, and focus on the scenario of **free space**, where there is no charge or current, just electromagnetic fields:

$$\boxed{\nabla\cdot\boldsymbol{E} = 0, \quad \text{(Gauss's law)}} \tag{1.41}$$

$$\boxed{\nabla\cdot\boldsymbol{H} = 0, \quad \text{(Gauss's law for magnetism)}} \tag{1.42}$$

$$\boxed{\nabla\times\boldsymbol{E} = -\mu_0\frac{\partial \boldsymbol{H}}{\partial t}, \quad \text{(Faraday's law)}} \tag{1.43}$$

$$\boxed{\nabla\times\boldsymbol{H} = \epsilon_0\frac{\partial \boldsymbol{E}}{\partial t}. \quad \text{(Modified Ampere's law)}} \tag{1.44}$$

The charge density and current density are called electromagnetic **sources**, and these Maxwell's equations are sometimes called source-free Maxwell's equations. The important point to note here is that a changing magnetic field induces an electric field around it, while a changing electric field induces a magnetic field around it. **This is the origin of electromagnetic waves: even if there is no charge or current, the electric and magnetic fields can sustain each other if they oscillate in time.**

1.9. Linearity of Maxwell's equations

A very important property of the Maxwell's equations is that they are **linear**. Consider the source-free version for simplicity.

(1) If $(\boldsymbol{E},\boldsymbol{H})$ has been found to be a solution, then $(a\boldsymbol{E}, a\boldsymbol{H})$, where a is any constant, is also a solution. To show this, we can plug $a\boldsymbol{E}$ and $a\boldsymbol{H}$ into the Maxwell's equations:

$$\nabla\cdot(a\boldsymbol{E}) = a(\nabla\cdot\boldsymbol{E}) = 0, \tag{1.45}$$

$$\nabla\cdot(a\boldsymbol{H}) = a(\nabla\cdot\boldsymbol{H}) = 0, \tag{1.46}$$

$$\nabla\times(a\boldsymbol{E}) = a(\nabla\times\boldsymbol{E}) = -a\mu_0\frac{\partial \boldsymbol{H}}{\partial t} = -\mu_0\frac{\partial(a\boldsymbol{H})}{\partial t}, \tag{1.47}$$

$$\nabla\times(a\boldsymbol{H}) = a(\nabla\times\boldsymbol{H}) = a\epsilon_0\frac{\partial \boldsymbol{E}}{\partial t} = \epsilon_0\frac{\partial(a\boldsymbol{E})}{\partial t}, \tag{1.48}$$

and find that, since $(\boldsymbol{E},\boldsymbol{H})$ satisfies the equations, $(a\boldsymbol{E}, a\boldsymbol{H})$ also satisfies the equations.

(2) If $(\boldsymbol{E}_1,\boldsymbol{H}_1)$ is a solution and $(\boldsymbol{E}_2,\boldsymbol{H}_2)$ is another solution, $(\boldsymbol{E}_1+\boldsymbol{E}_2, \boldsymbol{H}_1+\boldsymbol{H}_2)$ is also a solution.
- **Exercise**: Verify this.

It follows that, if $(\boldsymbol{E}_1,\boldsymbol{H}_1),(\boldsymbol{E}_2,\boldsymbol{H}_2),\ldots,(\boldsymbol{E}_N,\boldsymbol{H}_N)$ is a set of solutions, so is

$$\left(\sum_{n=1}^{N} a_n\boldsymbol{E}_n, \sum_{n=1}^{N} a_n\boldsymbol{H}_n\right) \tag{1.49}$$

This is called a **superposition** of EM fields. Linearity is extremely convenient when we search for EM-field solutions that satisfy desired boundary or initial conditions.

- **Exercise**: Prove that the Maxwell's equations, including the sources, are linear with respect to $(\boldsymbol{E},\boldsymbol{H},\rho,\boldsymbol{J})$.

CHAPTER 2

Electromagnetic Waves in One Space Dimension

2.1. Wave equation in one space dimension

To study the wave solutions of Maxwell's equations in free space, we will first consider a simple scenario by assuming that the fields do not change in x or y, such that

$$\textbf{Assumption}: \quad \frac{\partial \boldsymbol{E}}{\partial x} = 0, \qquad \frac{\partial \boldsymbol{E}}{\partial y} = 0, \qquad \frac{\partial \boldsymbol{H}}{\partial x} = 0, \qquad \frac{\partial \boldsymbol{H}}{\partial y} = 0. \tag{2.1}$$

Note that the fields are still vector fields, they just do not vary in the x and y direction;

$$\boldsymbol{E}(x, y, z, t) = \boldsymbol{E}(x', y', z, t), \quad \boldsymbol{H}(x, y, z, t) = \boldsymbol{H}(x', y', z, t) \quad \text{for any } (x, y) \text{ and } (x', y'). \tag{2.2}$$

This means that they are functions of z and t only:[1]

$$\boldsymbol{E}(\boldsymbol{r}, t) = \boldsymbol{E}(z, t), \qquad \boldsymbol{H}(\boldsymbol{r}, t) = \boldsymbol{H}(z, t). \tag{2.3}$$

An example is shown in Fig. 2.1.

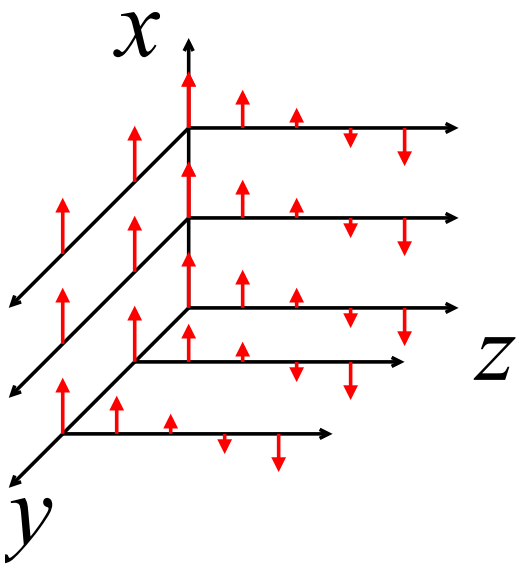


FIGURE 2.1. An example of a vector field that doesn't change in x or y and varies only along z.

The Gauss's law is

$$\nabla \cdot \boldsymbol{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0, \tag{2.4}$$

and since we assume that $\boldsymbol{E}$ is a function of (z, t) only, $\frac{\partial E_x}{\partial x} = 0$, $\frac{\partial E_y}{\partial y} = 0$, and we are left with

$$\frac{\partial E_z}{\partial z} = 0. \tag{2.5}$$

[1] It is a common convention in optics to assume waves propagating in the z direction. In *University Physics* [1] and the early lecture slides, however, the propagation is often assumed to be in the x direction. Here I will stick with variation along z to be consistent with later material. The fundamental physics is of course the same.

Hence, if the electric field varies in z only, the z component of $\boldsymbol{E}(z,t)$ does not change with respect to z. Thus we can assume it is a function of t only:

$$E_z(\boldsymbol{r},t) = E_z(z,t) = E_z(t). \tag{2.6}$$

We have exactly the same situation for the magnetic field from $\nabla \cdot \boldsymbol{H} = 0$, so

$$H_z(\boldsymbol{r},t) = H_z(z,t) = H_z(t). \tag{2.7}$$

Now let's look at Faraday's law. Our assumption simplifies the curl in Eq. (1.31) substantially:

$$\nabla \times \boldsymbol{E} = -\frac{\partial E_y}{\partial z}\hat{\boldsymbol{x}} + \frac{\partial E_x}{\partial z}\hat{\boldsymbol{y}}. \tag{2.8}$$

Note that this has no component along z. The right-hand side is

$$-\mu_0 \frac{\partial \boldsymbol{H}}{\partial t} = -\mu_0 \left(\frac{\partial H_x}{\partial t}\hat{\boldsymbol{x}} + \frac{\partial H_y}{\partial t}\hat{\boldsymbol{y}} + \frac{\partial H_y}{\partial t}\hat{\boldsymbol{z}} \right). \tag{2.9}$$

Matching each vector component in Eqs. (2.8) and (2.9),

$$\frac{\partial E_y}{\partial z} = \mu_0 \frac{\partial H_x}{\partial t}, \tag{2.10}$$

$$\frac{\partial E_x}{\partial z} = -\mu_0 \frac{\partial H_y}{\partial t}, \tag{2.11}$$

$$\frac{\partial H_z}{\partial t} = 0. \tag{2.12}$$

The space derivative of E_y is coupled to the time derivative of H_x, and space derivative of E_x is coupled to the time derivative of H_y. We see that H_z doesn't change in time as well, so it must be a constant everywhere and anytime.

We can do the same for the modified Ampere's law:

$$\nabla \times \boldsymbol{H} = \epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t} \rightarrow -\frac{\partial H_y}{\partial z}\hat{\boldsymbol{x}} + \frac{\partial H_x}{\partial z}\hat{\boldsymbol{y}} = \epsilon_0 \left(\frac{\partial E_x}{\partial t}\hat{\boldsymbol{x}} + \frac{\partial E_y}{\partial t}\hat{\boldsymbol{y}} + \frac{\partial E_y}{\partial t}\hat{\boldsymbol{z}} \right), \tag{2.13}$$

leading to

$$-\frac{\partial H_y}{\partial z} = \epsilon_0 \frac{\partial E_x}{\partial t}, \tag{2.14}$$

$$\frac{\partial H_x}{\partial z} = \epsilon_0 \frac{\partial E_y}{\partial t}, \tag{2.15}$$

$$\frac{\partial E_z}{\partial t} = 0. \tag{2.16}$$

These equations show that the space derivative of H_y is coupled to the time derivative of E_x, the space derivative of H_x is coupled to the time derivative of E_y, and E_z, like H_z, is also a constant.

Let's take stock of what we have obtained so far:

- We have assumed no variation in x and y, so the fields are functions of z and t only.
- E_y is coupled to H_x through Eqs. (2.10) and (2.15); they don't depend on the other field components.
- E_x is coupled to H_y through Eqs. (2.11) and (2.14); they don't depend on the other field components.
- E_z and H_z must be constants (do not change in z and t) and not coupled to the other field components.

Let's focus on E_x and H_y, which obey Eqs. (2.11) and (2.14). Take the time derivative of Eq. (2.14):

$$-\frac{\partial}{\partial t}\frac{\partial H_y}{\partial z} = \epsilon_0 \frac{\partial^2 E_x}{\partial t^2}. \tag{2.17}$$

We can interchange the order of the partial derivatives on the left-hand side,

$$-\frac{\partial}{\partial t}\frac{\partial H_y}{\partial z} = -\frac{\partial}{\partial z}\frac{\partial H_y}{\partial t}. \tag{2.18}$$

Now we can use Eq. (2.11) to find out $\frac{\partial H_y}{\partial t}$:

$$\frac{\partial H_y}{\partial t} = -\frac{1}{\mu_0}\frac{\partial E_x}{\partial z}, \qquad -\frac{\partial}{\partial t}\frac{\partial H_y}{\partial z} = -\frac{\partial}{\partial z}\frac{\partial H_y}{\partial t} = \frac{1}{\mu_0}\frac{\partial^2 E_x}{\partial z^2}. \tag{2.19}$$

Substituting this back into Eq. (2.17), we obtain

$$\boxed{\frac{\partial^2 E_x}{\partial z^2} = \mu_0 \epsilon_0 \frac{\partial^2 E_x}{\partial t^2}.} \tag{2.20}$$

It involves just one field component E_x and some second-order partial derivatives. Eq. (2.20) is called a **wave equation** because it admits a wave solution:

$$E_x(z,t) = f(z - vt), \tag{2.21}$$

for any single-variable function f. Since the fields do not change with respect to the plane perpendicular to the propagation direction z, these waves are called **plane waves**. To show that the plane wave is a solution, note that, since f is a single-variable function,

$$\frac{\partial f(z-vt)}{\partial z} = \frac{df(X)}{dX}\Big|_{X=z-vt}, \qquad \frac{\partial f(z-vt)}{\partial t} = -v\frac{df(X)}{dX}\Big|_{X=z-vt}. \tag{2.22}$$

The velocity v of the electromagnetic wave is then

$$v^2 = \frac{1}{\mu_0 \epsilon_0}, \qquad \boxed{|v| = c \equiv \frac{1}{\sqrt{\mu_0 \epsilon_0}} = 299,792,458 \text{ m/s} \approx 3 \times 10^8 \text{ m/s}.} \tag{2.23}$$

c is the **speed of light** in free space. It is defined exactly as 299,792,458 m/s; the meter is defined in terms of this value and the second is defined using something else (atomic clock).

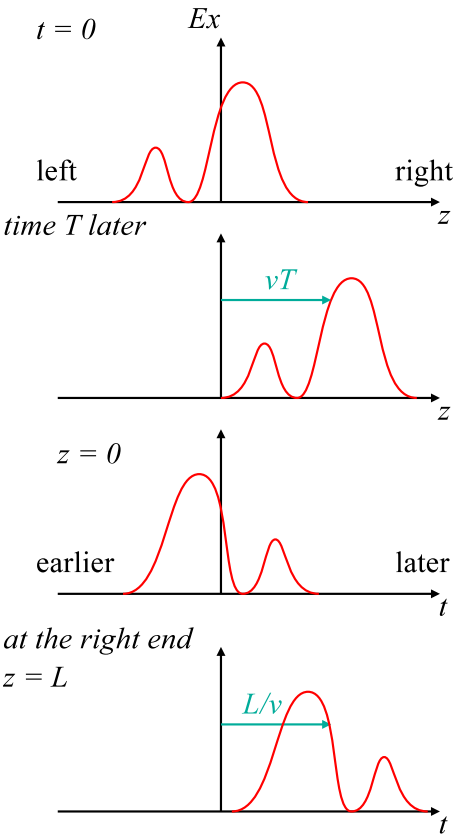
- **Exercise**: Perform time derivative of Eq. (2.11) and show that H_y obeys the same wave equation.
- **Exercise**: Show that E_y and H_x also obey the same wave equation.
- **Exercise**: Verify that $1/\sqrt{\mu_0 \epsilon_0}$ has the unit of velocity.

v can be positive or negative; in fact, the wave equation admits a **superposition** of waves that propagate in either direction:

$$E_x = f(z - ct) + g(z + ct), \tag{2.24}$$

where f and g are arbitrary functions and $g(z + ct)$ propagates in the negative z direction. This is because Maxwell's equations, and the wave equation by extension, are **linear**: if there are two solutions, the addition of the solutions is also a solution.

2.2. Propagation of plane wave



Assume that $E_x(z,t) = f(z - vt)$, and v is positive. At $t = 0$,

$$E_x(z,0) = f(z). \tag{2.25}$$

Then after some time T, the field distribution is **shifted to the right** :

$$E_x(z,T) = f(z - vT). \tag{2.26}$$

The velocity is v.
Now let's fix $z = 0$ and observe the field as it changes in time:

$$E_x(0,t) = f(-vt). \tag{2.27}$$

Note that the temporal shape is flipped with respect to the spatial shape; for a right-propagating wave, the *right* part arrives *earlier*. At some distance L to the right, the field changes in time as

$$E_x(L,t) = f(L - vt) = f\left(-v\left(t - \frac{L}{v}\right)\right). \tag{2.28}$$

The **time delay** is L/v. The faster the speed v, the smaller the delay.

Fig. 2.2 shows what a wave looks like in a 2D plot with respect to (z,t): A Java applet of wave propagation can be found at `http://phet.colorado.edu/` (PhET/sims/radio-waves/radio-waves_all.jar).

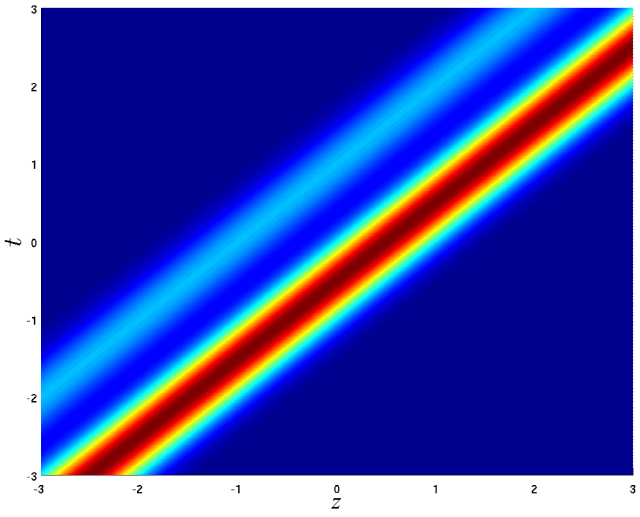


FIGURE 2.2. A color plot of $f(z - vt)$, v positive.

2.3. Monochromatic plane wave

An extremely important wave solution is the monochromatic (meaning single color or single frequency) plane wave:[2]

$$E_x(z,t) = E_{\max}\cos(kz - \omega t + \phi), \tag{2.29}$$

where ϕ is an arbitrary phase. The spatial period

$$\boxed{\lambda = \frac{2\pi}{k}} \tag{2.30}$$

is called the **wavelength**. k is called the **wavenumber** (with unit m^{-1}), which characterizes how rapidly the field changes in space. The temporal frequency in Hertz (Hz) is

$$\boxed{\nu = \frac{\omega}{2\pi}.} \tag{2.31}$$

Note that this is a greek letter "nu", not v. ω is the angular frequency (rad/s). Speed of light is thus related to the frequency and wavelength by

$$\boxed{c = \frac{\omega}{k} = \nu\lambda.} \tag{2.32}$$

This relation tells us how the variation of the wave in space, as specified by k (or λ), is related to the variation in time, as specified by ω (or ν), through the speed c. We will deal with these quantities again and again in the coming discussions.

How about H_y? We have already noted that E_x is coupled to H_y, and we can assume that all the other field components are all zero. Going back to Eqs. (2.11), we find

$$\frac{\partial E_x}{\partial z} = -kE_{\max}\sin(kz - \omega t + \phi) = -\mu_0\frac{\partial H_y}{\partial t}. \tag{2.33}$$

To obtain H_y, we can integrate this in time:

$$H_y(z,t) = \frac{kE_{\max}}{\mu_0}\int dt\,\sin(kz - \omega t + \phi) \tag{2.34}$$

$$= \frac{kE_{\max}}{\mu_0\omega}\cos(kz - \omega t + \phi) + C. \tag{2.35}$$

This constant C doesn't depend on time and represents a possible magnetostatic component; we shall assume it to be zero. Hence

$$H_y(z,t) = \frac{E_{\max}}{Z_0}\cos(kz - \omega t + \phi), \tag{2.36}$$

$$Z_0 = \frac{\mu_0\omega}{k} = \mu_0 c = \frac{\mu_0}{\sqrt{\mu_0\epsilon_0}} = \sqrt{\frac{\mu_0}{\epsilon_0}}, \tag{2.37}$$

where we have used the facts $k = \omega/c$ in Eq. (2.32) and $c = 1/\sqrt{\mu_0\epsilon_0}$ in Eq. (2.23). Z_0 is the **free-space impedance**:

$$\boxed{Z_0 = 119.9169832\pi\ \text{Ohm} \approx 120\pi\ \text{Ohm}\ (\Omega).} \tag{2.38}$$

- **Exercise**: Verify that Z_0 has the unit of Ohm, the unit of resistance, or Volt/Ampere.

The magnetic field oscillates at the same phase as the electric field and is proportional to the electric-field amplitude, as shown in Fig. 2.3.

[2] also called sinusoidal plane wave in the lecture slides.

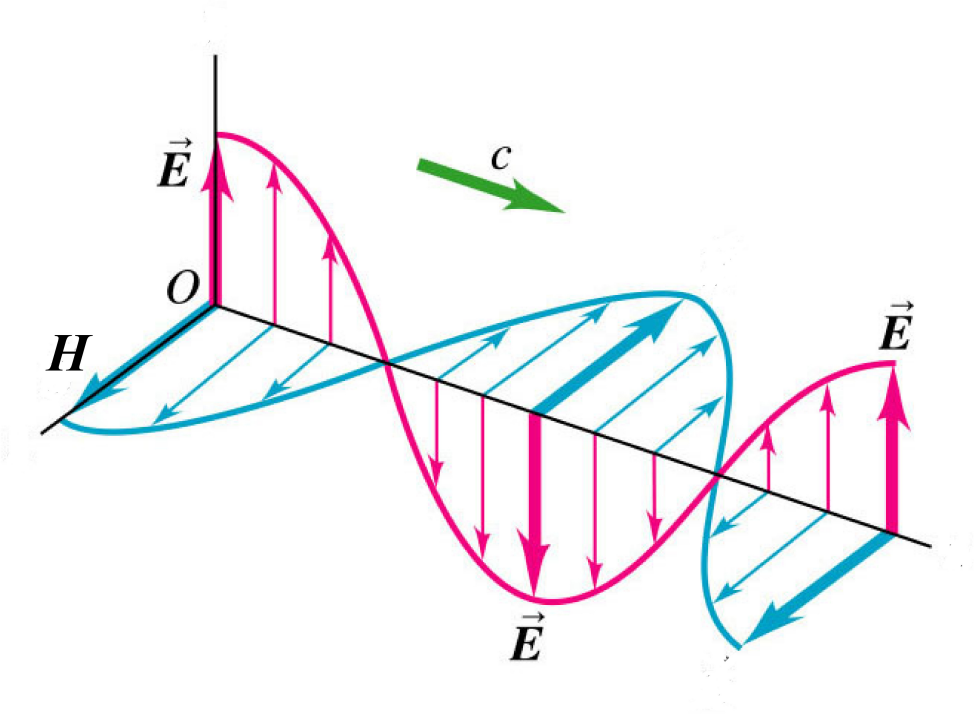


FIGURE 2.3. A snapshot of an electromagnetic monochromatic plane wave.

2.4. Electromagnetic spectrum

Optics usually refer to wavelengths and frequencies near the visible ($\lambda = 400 - 700$ nm). A rough order-of-magnitude estimate is

$$\lambda \sim 1\ \mu\text{m}\ (10^{-6}\ \text{m}), \qquad \nu \sim 300\ \text{THz}\ (3 \times 10^{14}\ \text{Hz}) \tag{2.39}$$

This frequency is **extremely high**. Compare it, for example, to your microwave oven (2.4 GHz), CPU clock speed (GHz), radio frequency (kHz–MHz), audio frequency (20 Hz–20 kHz), AC power source (60 Hz), heartbeat (Hz), etc., as shown in Fig. 2.4.

An EM wave at optical frequencies is often labeled by **its wavelength in free space** c/ν. For light in a medium, the speed of light and the wavelength may change, but the frequency ω is usually fixed according to input waves/boundary conditions. In common optics convention, the frequency of the wave is still labeled by its free-space wavelength value with respect to ω.

Light from the sun consists of a continuous spectrum of frequencies and appears white to our eyes. The photoreceptor cells on the retina of our eyes contain three types of cones that respond differently to different wavelengths (http://en.wikipedia.org/wiki/Cone_cell). One can therefore simulate the sensation of white light by stimulating these cones using three distinct wavelengths, such as those in a white light-emitting diode (LED).

2.5. Complex form of monochromatic plane wave solution

As we shall see later, it is often extremely convenient to use complex exponentials instead of sines and cosines to represent monochromatic plane waves. Recall that, for a real variable X,

$$\exp(jX) = \cos X + j \sin X, \tag{2.40}$$

$$\cos X = \frac{1}{2}\left(e^{jX} + e^{-jX}\right) = \text{Re}\left(e^{jX}\right), \tag{2.41}$$

$$\sin X = \frac{1}{2j}\left(e^{jX} - e^{-jX}\right) = \text{Im}\left(e^{jX}\right), \tag{2.42}$$

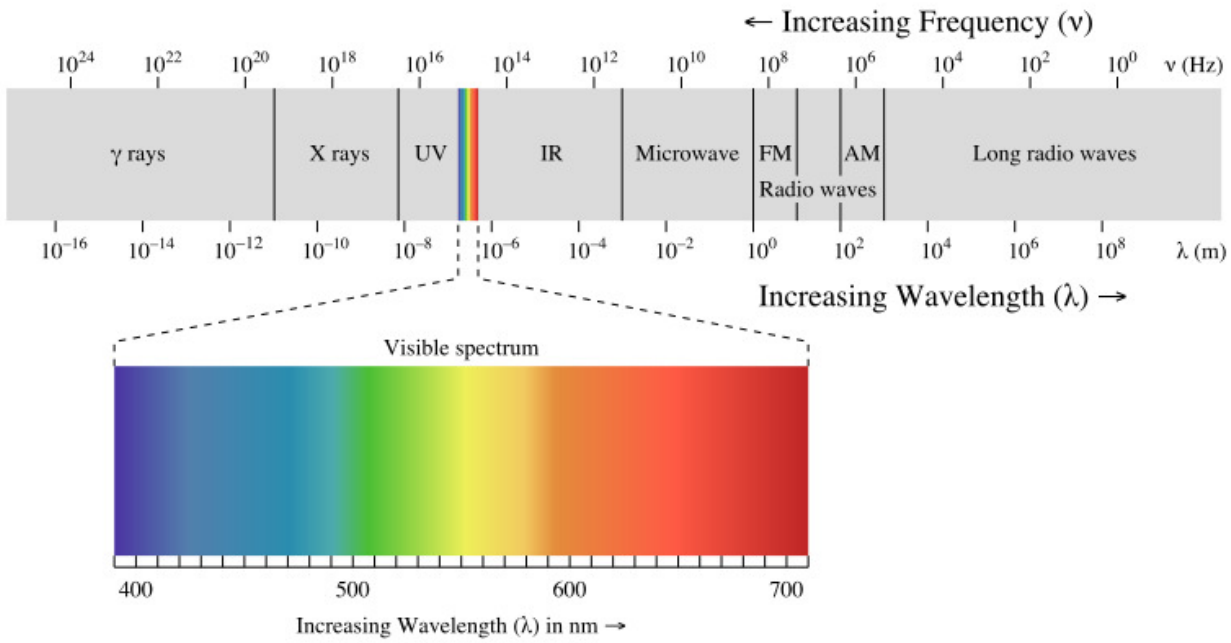


FIGURE 2.4. Electromagnetic spectrum.

where j is the imaginary number with $j^2 = -1$, Re denotes the real part, and Im denotes the imaginary part. We can then express Eq. (2.29) as[3]

$$\boldsymbol{E}(z,t) = \mathrm{Re}\left[\hat{\boldsymbol{x}}\tilde{E}\exp(jkz - j\omega t)\right], \qquad \tilde{E} = E_{\max}\exp(j\phi), \tag{2.43}$$

where $\tilde{E}$ is the **complex amplitude** of the wave, encoding not just the maximum electric field $E_{\max}$ but also the phase ϕ. Similarly, Eq. (2.36) can be written as

$$\boldsymbol{H} = \mathrm{Re}\left[\hat{\boldsymbol{y}}\frac{\tilde{E}}{Z_0}\exp(jkz - j\omega t)\right]. \tag{2.44}$$

Although the electromagnetic fields must be real in practice, $\tilde{E}\exp(jkz - j\omega t)$ is a solution of the Maxwell's equations in its own right, as

$$\tilde{E}\exp(jkz - j\omega t) = E_{\max}\cos(kz - \omega t + \phi) + jE_{\max}\sin(kz - \omega t + \phi), \tag{2.45}$$

and both the cosine and sine parts are solutions. We will often focus on the complex solution and assume implicitly that the real part should be taken at the end of the calculations to obtain the real-life solution. This is okay because the real part can be written as

$$\mathrm{Re}\left[\tilde{E}\exp(jkz - j\omega t)\right] = \frac{1}{2}\left[\tilde{E}\exp(jkz - j\omega t) + \tilde{E}^*\exp(-jkz + j\omega t)\right], \tag{2.46}$$

where the second term is the **complex conjugate** of the first. If a complex solution for $(\boldsymbol{E}, \boldsymbol{H})$ of the Maxwell's equations has been found, its complex conjugate $(\boldsymbol{E}^*, \boldsymbol{H}^*)$ is also a solution. It follows that the real part of the complex solution, being the superposition of the complex solution and its conjugate, must also be a solution.

2.6. Polarizations

Let us focus on the electric field of a plane wave. The wave $\hat{\boldsymbol{x}}\tilde{E}\exp(jkz - j\omega t)$ in Eq. (2.43) we have been studying so far is called **linearly polarized** because the electric field oscillates in one direction only. Another example of linear polarization is $\hat{\boldsymbol{y}}\tilde{E}\exp(jkz - j\omega t)$, as shown in Fig. 2.5.

[3] In some engineering texts the complex exponential is assumed to be $\exp(-jkz + j\omega t)$, but since the real parts of $\exp(jkz - j\omega t)$ and $\exp(-jkz + j\omega t)$ are the same, they both represent a wave propagating in the positive z direction. As long as one sticks with one form and be consistent, the final result will be the same.

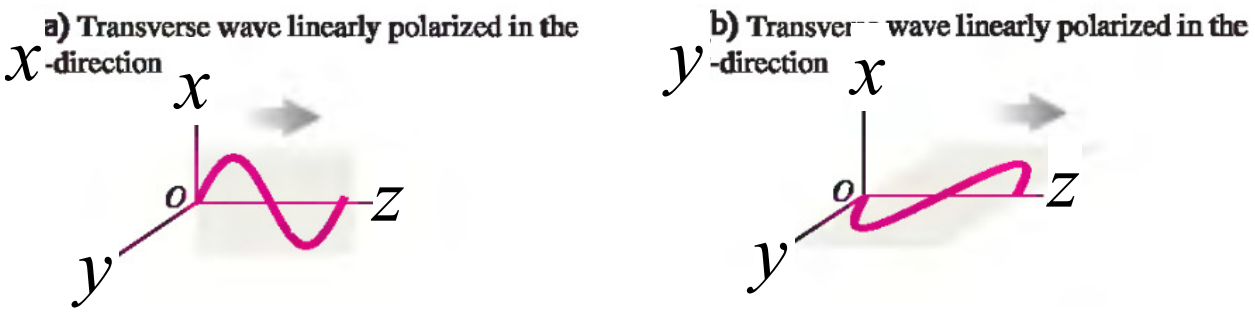


FIGURE 2.5. Two examples of linear polarizations, where the electric field oscillates in $\hat{\boldsymbol{x}}$ or $\hat{\boldsymbol{y}}$ direction. (from [1])

In general, if we add these two linearly-polarized waves in the following way, where θ is another parameter:

$$\textbf{Linear polarization}: \quad \boldsymbol{E} = (\hat{\boldsymbol{x}}\cos\theta + \hat{\boldsymbol{y}}\sin\theta)\,\tilde{E}\exp(jkz - j\omega t), \tag{2.47}$$

it is also a valid solution by linearity. Focusing on $z = 0$, the field is simply oscillating in the $(\hat{\boldsymbol{x}}\cos\theta + \hat{\boldsymbol{y}}\sin\theta)$ direction. This is also linear polarization. See http://en.wikipedia.org/wiki/Polarization_%28waves%29.

2.7. *Circular polarization

What about

$$\boldsymbol{E} = \frac{(\hat{\boldsymbol{x}} - j\hat{\boldsymbol{y}})}{\sqrt{2}}\tilde{E}\exp(jkz - j\omega t), \qquad \tilde{E} = E_{\max}e^{j\phi}? \tag{2.48}$$

The factor

$$-j = \exp\left(-j\frac{\pi}{2}\right) \tag{2.49}$$

means that there is a **phase difference** between the $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ phasor components. Looking at the field at $z = 0$ and assume $\phi = 0$ for simplicity,

$$\textbf{Left circular polarization}: \boldsymbol{E}(z=0) = \frac{(\hat{\boldsymbol{x}} - j\hat{\boldsymbol{y}})}{\sqrt{2}}\tilde{E}\exp(-j\omega t), \tag{2.50}$$

$$\mathrm{Re}\,\boldsymbol{E}(z=0) = \frac{E_{\max}}{\sqrt{2}}\left[\hat{\boldsymbol{x}}\cos(\omega t) - \hat{\boldsymbol{y}}\sin(\omega t)\right]. \tag{2.51}$$

The real part of $\hat{\boldsymbol{x}}$ component oscillates like $\cos(\omega t)$ and the real part of $\hat{\boldsymbol{y}}$ component oscillates like $-\sin(\omega t)$. The real electric field is thus tracing out **a circle**. This is called left **circular polarization** (left-hand rule with respect to propagation direction $\hat{\boldsymbol{z}}$). Another possibility is right circular polarization:

$$\textbf{Right circular polarization}: \quad \boldsymbol{E}(z=0) = \frac{(\hat{\boldsymbol{x}} + j\hat{\boldsymbol{y}})}{\sqrt{2}}\tilde{E}\exp(-j\omega t), \tag{2.52}$$

$$\mathrm{Re}\,\boldsymbol{E}(z=0) = \frac{E_{\max}}{\sqrt{2}}\left[\hat{\boldsymbol{x}}\cos(\omega t) + \hat{\boldsymbol{y}}\sin(\omega t)\right]. \tag{2.53}$$

In general, if the amplitudes of the x and y components are not equal, the real electric field traces out an ellipse instead, and the polarization is called **elliptical polarization**.

2.8. Polarizer

The polarization of light from the sun and fluorescent lamps is **unpolarized**, meaning that the polarization is random. A polarizing filter, or polarizer, **absorbs** the electric-field component along its polarizing axis, but lets the other component passes through. Suppose that the polarizer is at $z = 0$ and the input field is linearly polarized:

$$\boldsymbol{E}_{\text{in}} = \tilde{E}\,(\hat{\boldsymbol{x}}\cos\theta + \hat{\boldsymbol{y}}\sin\theta)\exp(jkz - j\omega t). \tag{2.54}$$

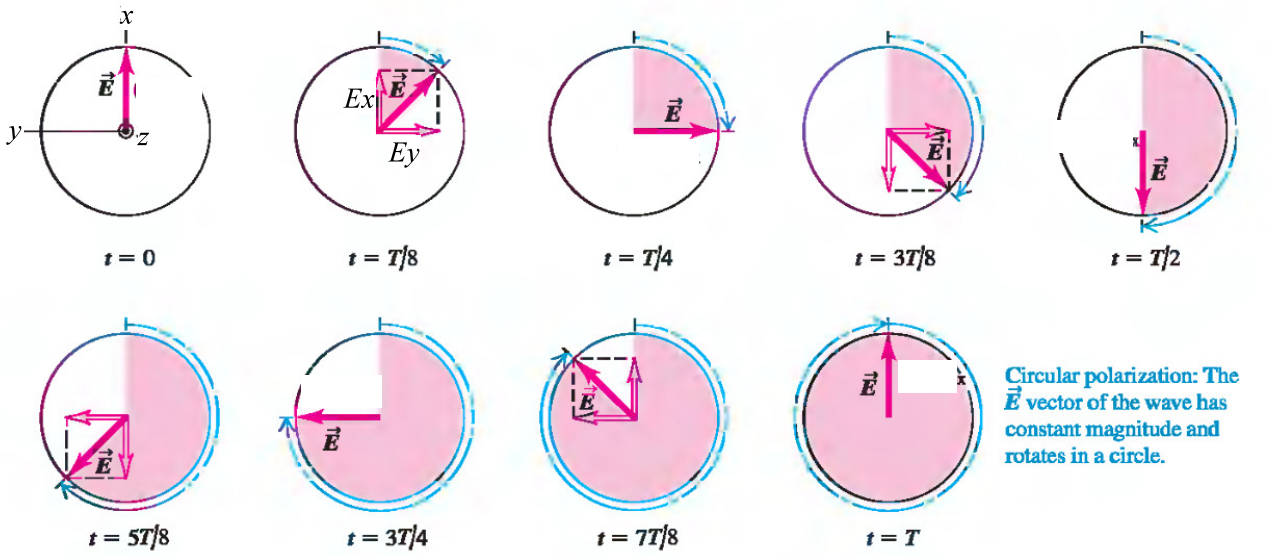


FIGURE 2.6. The electric field of circularly polarized wave at $z = 0$ traces out a circle in time. This is left circular polarization as the rotation and the propagation direction obeys the left-hand rule. (from [1])

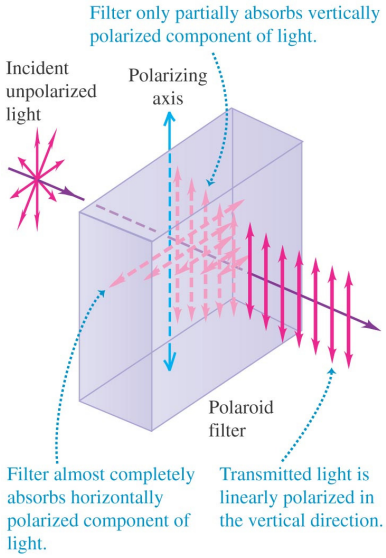


FIGURE 2.7. A polarizer absorbs light polarized along a certain direction but transmits light polarized along its polarizing axis. (from [1])

Suppose that the polarizing axis is $\hat{\boldsymbol{x}}$. The ideal behavior of a polarizer is to leave only the $\hat{\boldsymbol{x}}$ component at the output,

$$\boldsymbol{E}_{\text{out}} = \hat{\boldsymbol{x}}\tilde{E}\cos\theta\exp(jkz - j\omega t), \tag{2.55}$$

while absorbing the $\hat{\boldsymbol{y}}$ component completely. Many sunglasses also use polarizers to attenuate light.

Consider two polarizers in Fig. 2.8, with the two polarizing axes making an angle θ between them. The second polarizer is called the analyzer. Suppose the first polarizing axis is $\cos\theta\hat{\boldsymbol{x}} + \sin\theta\hat{\boldsymbol{y}}$, and the second polarizing axis is $\hat{\boldsymbol{x}}$. The final transmitted field is the $\cos\phi$ component, and the intensity obeys Malus's law: the output intensity $\propto |\tilde{E}\cos\theta|^2$ is a factor of $\cos^2\theta$ smaller than the input intensity $\propto |\tilde{E}|^2$ before the analyzer.

An important application of polarizers is **liquid-crystal display (LCD)**: http://www.youtube.com/watch?v=k7xGQKpQAWw, used in most modern-day computer displays such as LCD monitors, laptops, smartphones, and tablets. For the LCD to work, it requires not only a polarizer but also a controllable birefringent medium, to be discussed later.

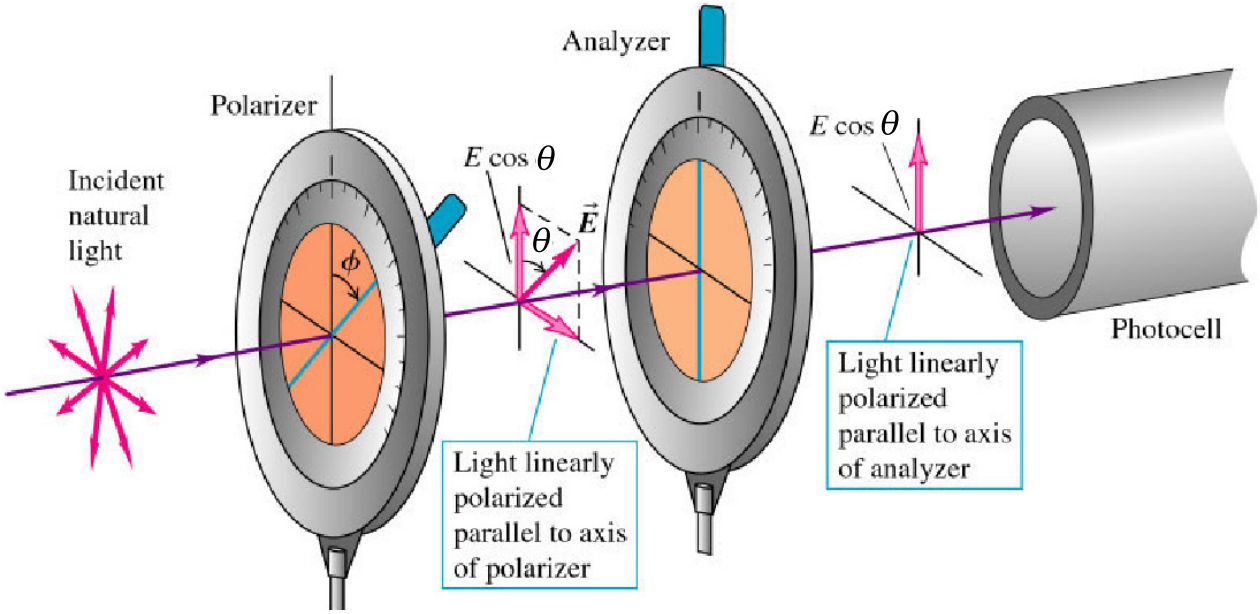


FIGURE 2.8. Two polarizers with polarizing axes making an angle θ with respect to one another. (from [1])

CHAPTER 3

Waves in Three Space Dimensions

We shall now jump into the general case of waves in three-dimensional space (plus one dimension of time). In reality, there is nothing special about the z direction, and we need to account for situations where waves can propagate in any direction.

3.1. Wave equation

Assume source-free Maxwell's equations. Our derivation of a general wave equation starts from Faraday's law given by Eq. (1.43). Let's take the curl on both sides,

$$\nabla \times (\nabla \times \boldsymbol{E}) = -\mu_0 \nabla \times \frac{\partial \boldsymbol{H}}{\partial t} = -\mu_0 \frac{\partial}{\partial t} (\nabla \times \boldsymbol{H}) . \tag{3.1}$$

Now that we have $\nabla \times \boldsymbol{H}$ on the right-hand side, we can relate it to $\boldsymbol{E}$ through Ampere's law:

$$-\mu_0 \frac{\partial}{\partial t} (\nabla \times \boldsymbol{H}) = -\mu_0 \frac{\partial}{\partial t} \left(\epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t} \right) = -\mu_0 \epsilon_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}, \tag{3.2}$$

which can be plugged back into Eq. (3.1) to give

$$\nabla \times (\nabla \times \boldsymbol{E}) = -\mu_0 \epsilon_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}. \tag{3.3}$$

The left-hand side can be further simplified if we use the following vector identity (http://en.wikipedia.org/wiki/Vector_calculus_identities#Curl_of_the_curl):

$$\nabla \times (\nabla \times \boldsymbol{E}) = \nabla(\nabla \cdot \boldsymbol{E}) - \nabla^2 \boldsymbol{E}, \tag{3.4}$$

where

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{3.5}$$

is called the Laplacian. We recognize that $\nabla \cdot \boldsymbol{E} = 0$ from Gauss's law (no charge). Putting this back in Eq. (3.3), the final result is

$$\boxed{\nabla^2 \boldsymbol{E} = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) \boldsymbol{E} = \mu_0 \epsilon_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}.} \tag{3.6}$$

It looks similar to the (1+1)D case; if we assume that $\frac{\partial \boldsymbol{E}}{\partial y} = 0$, $\frac{\partial \boldsymbol{E}}{\partial z} = 0$, and focus on the E_x component, we would get back Eq. (2.20).

3.2. Monochromatic plane wave

A monochromatic plane-wave solution of Eq. (3.6) is

$$\boxed{\boldsymbol{E} = \tilde{\boldsymbol{E}} \exp(j\boldsymbol{k} \cdot \boldsymbol{r} - j\omega t),} \tag{3.7}$$

where $\tilde{\boldsymbol{E}}$ is a vectoral complex amplitude that doesn't depend on space or time and the dot product is

$$\boldsymbol{k} \cdot \boldsymbol{r} = k_x x + k_y y + k_z z, \tag{3.8}$$

leading to

$$\begin{aligned} \exp(j\boldsymbol{k} \cdot \boldsymbol{r} - j\omega t) &= \exp(jk_x x + jk_y y + jk_z z - j\omega t) \\ &= \exp(jk_x x) \exp(jk_y y) \exp(jk_z z) \exp(-j\omega t). \end{aligned} \tag{3.9}$$

As I warned you before, we shall now focus on the complex solution; it is understood implicitly that the real-life solution is its real part. $\boldsymbol{k}$ is called the **wavevector**, which denotes the propagation direction of the plane wave. To see this, let's write $\boldsymbol{k}\cdot\boldsymbol{r}$ in terms of the unit vector $\hat{\boldsymbol{k}}$ and length k of $\hat{\boldsymbol{k}}$:

$$\boldsymbol{k}\cdot\boldsymbol{r} = k(\hat{\boldsymbol{k}}\cdot\boldsymbol{r}), \tag{3.10}$$

where $\hat{\boldsymbol{k}}\cdot\boldsymbol{r}$ is the projected length of $\boldsymbol{r}$ along the direction of $\boldsymbol{k}$ as shown in Fig. 3.1. Think of $\hat{\boldsymbol{k}}\cdot\boldsymbol{r}$ as a **coordinate**, like x, y, or z, except that it is the length along the direction of $\boldsymbol{k}$ and the $\boldsymbol{k}$ vector is its axis. Moreover, if we consider any direction perpendicular to $\boldsymbol{k}$, we can define a perpendicular component of $\boldsymbol{r}$ as

$$\boldsymbol{r}_\perp \equiv \boldsymbol{r} - \left(\hat{\boldsymbol{k}}\cdot\boldsymbol{r}\right)\hat{\boldsymbol{k}}, \tag{3.11}$$

and

$$\boldsymbol{k}\cdot\boldsymbol{r} = k\left(\hat{\boldsymbol{k}}\cdot\boldsymbol{r}\right) \text{ for any } \boldsymbol{r}_\perp, \tag{3.12}$$

meaning that $\exp(j\boldsymbol{k}\cdot\boldsymbol{r})$ is always equal to its value at $\boldsymbol{r}_\perp = 0$ and constant along any direction perpendicular to $\hat{\boldsymbol{k}}$. For example, if $\boldsymbol{k} = k_z\hat{\boldsymbol{z}}$, $\hat{\boldsymbol{k}}\cdot\boldsymbol{r} = z$, $\boldsymbol{k}\cdot\boldsymbol{r} = kz$, $\exp(jkz)$ is periodic along z and constant along x and y.

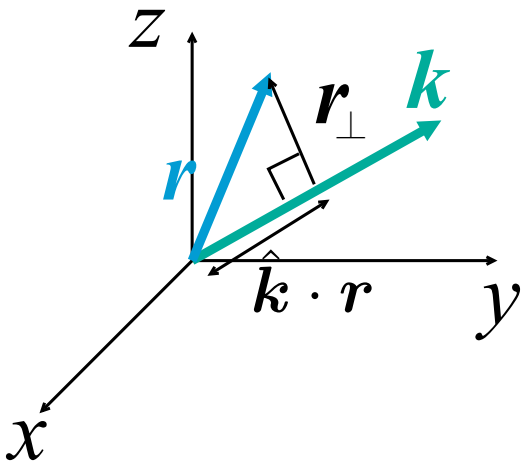


FIGURE 3.1. The complex exponential $\exp(j\boldsymbol{k}\cdot\boldsymbol{r})$ as a function of position $\boldsymbol{r}$ varies only along the direction of $\boldsymbol{k}$ but is constant along any direction $\boldsymbol{r}_\perp$ that is perpendicular to $\boldsymbol{k}$.

The spatial period of the wave along the direction of $\boldsymbol{k}$ is again the wavelength, and it is related to wavenumber, or the magnitude of the wavevector, as follows

$$\boxed{\lambda = \frac{2\pi}{k},} \qquad k = \sqrt{k_x^2 + k_y^2 + k_z^2}. \tag{3.13}$$

To verify that Eq. (3.7) is a solution of the wave equation given by Eq. (3.6), it is very useful to remember that

$$\frac{\partial}{\partial t}\exp(-j\omega t) = -j\omega\exp(-j\omega t), \tag{3.14}$$

$$\frac{\partial}{\partial x}\exp(jk_x x) = jk_x\exp(jk_x x), \tag{3.15}$$

$$\frac{\partial}{\partial y}\exp(jk_y y) = jk_y\exp(jk_y y), \tag{3.16}$$

$$\frac{\partial}{\partial z}\exp(jk_z z) = jk_z\exp(jk_z z). \tag{3.17}$$

These mean that, whenever I apply a derivative to a complex exponential, I can replace the derivative by a multiplicative factor such as $-j\omega$ or jk_x. Now the left-hand side of Eq. (3.6) becomes

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\tilde{\boldsymbol{E}}\exp(j\boldsymbol{k}\cdot\boldsymbol{r} - j\omega t) = \left[(jk_x)^2 + (jk_y)^2 + (jk_z)^2\right]\tilde{\boldsymbol{E}}\exp(j\boldsymbol{k}\cdot\boldsymbol{r} - j\omega t) \tag{3.18}$$

$$= -\left(k_x^2 + k_y^2 + k_z^2\right)\tilde{\boldsymbol{E}}\exp(j\boldsymbol{k}\cdot\boldsymbol{r} - j\omega t), \tag{3.19}$$

and the right-hand side is

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\tilde{\boldsymbol{E}}\exp(j\boldsymbol{k}\cdot\boldsymbol{r}-j\omega t)=\frac{1}{c^2}(-j\omega)^2\tilde{\boldsymbol{E}}\exp(j\boldsymbol{k}\cdot\boldsymbol{r}-j\omega t)=-\frac{\omega^2}{c^2}\tilde{\boldsymbol{E}}\exp(j\boldsymbol{k}\cdot\boldsymbol{r}-j\omega t). \tag{3.20}$$

Putting them together, and assuming that $\tilde{\boldsymbol{E}}$ is nonzero, we are left with the so-called **dispersion relation** between the magnitude of $\boldsymbol{k}$ and the frequency ω:

$$\boxed{k^2=k_x^2+k_y^2+k_z^2=\frac{\omega^2}{c^2}.} \tag{3.21}$$

It is a relation between how the wave varies in space, as specified by $\boldsymbol{k}$, and how the wave varies in time, as specified by ω, through the speed of light c. In particular, it says that no component of $\boldsymbol{k}$ can exceed ω/c (if all components are real). Given the results in Chap. 2, this should not be surprising; instead of a wave propagating in the z direction in Chap. 2, here we simply have a wave propagating in a different direction, and we should still have the relation $k=\omega/c$.

We can represent Eq. (3.21) as a sphere with radius ω/c in a space with axes k_x, k_y, and k_z called $\boldsymbol{k}$ space, as shown in Fig. 3.2.

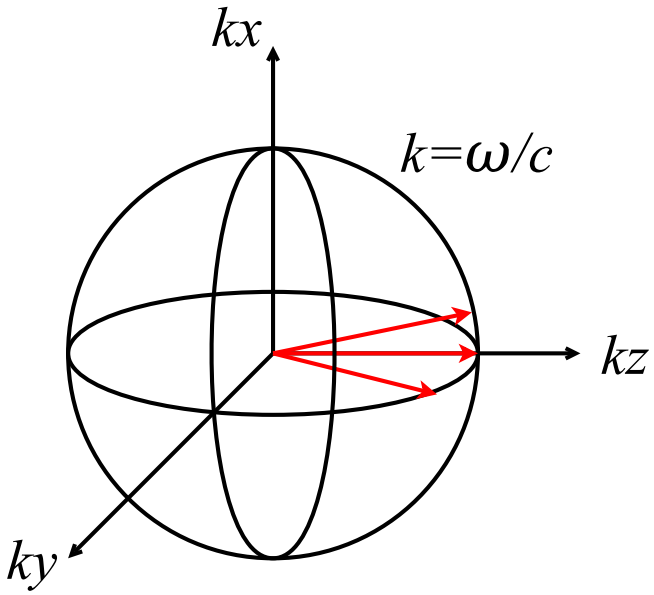


FIGURE 3.2. A sphere in $\boldsymbol{k}$ space representing the dispersion relation in Eq. (3.21). Note that this space is different from the real space with coordinates (x,y,z). Given ω, (k_x,k_y,k_z) must be on the sphere.

3.3. Transverse waves

In Chap. 2, we have found that, if the wave propagates in the z direction, E_z and H_z have to be constant and uncoupled from the other field components. These constant field components along the propagation direction can then be assumed to be zero without affecting the solutions for the other components. We can generalize this observation by plugging Eq. (3.7) into Gauss's law. The key is to observe that, since derivatives are replaced by multiplications when applied to complex exponentials, the ∇ operator can be replaced by $j\boldsymbol{k}$ here:

$$\nabla\cdot\boldsymbol{E}=j\boldsymbol{k}\cdot\tilde{\boldsymbol{E}}\exp(j\boldsymbol{k}\cdot\boldsymbol{r}-j\omega t)=0. \tag{3.22}$$

For this to hold at all space and time,

$$\boldsymbol{k}\cdot\tilde{\boldsymbol{E}}=0, \tag{3.23}$$

which means that the wave vector $\boldsymbol{k}$ and the vector amplitude $\tilde{\boldsymbol{E}}$ must be **perpendicular**. Similarly, if we assume the magnetic-field solution of

$$\boldsymbol{H}=\tilde{\boldsymbol{H}}\exp(j\boldsymbol{k}\cdot\boldsymbol{r}-j\omega t), \tag{3.24}$$

Gauss's law of magnetism would give

$$\boldsymbol{k} \cdot \tilde{\boldsymbol{H}} = 0, \tag{3.25}$$

meaning that the magnetic field is also perpendicular to the propagation direction. Thus electromagnetic waves are **transverse waves**; the fields oscillate in directions perpendicular to the propagation direction.

The vector amplitudes $\tilde{\boldsymbol{E}}$ and $\tilde{\boldsymbol{H}}$ are also related. Let's substitute Eqs. (3.7) and (3.24) into Faraday's law:

$$\nabla \times \boldsymbol{E} = -\mu_0 \frac{\partial \boldsymbol{H}}{\partial t} \rightarrow \boldsymbol{k} \times \tilde{\boldsymbol{E}} = \mu_0 \omega \tilde{\boldsymbol{H}}. \tag{3.26}$$

We have converted the curl into a cross product between the wavevector and a vector complex amplitude. With $\tilde{\boldsymbol{E}}$ perpendicular to $\boldsymbol{k}$ according to Eq. (3.23), Eq. (3.26) means that $\tilde{\boldsymbol{H}}$ is not only perpendicular to $\boldsymbol{k}$ according to Eq. (3.25) but also perpendicular to $\tilde{\boldsymbol{E}}$, as shown in Fig. 3.3. Furthermore, if we write $\boldsymbol{k}$ in terms of its magnitude k and unit vector $\hat{\boldsymbol{k}}$,

$$\boxed{\tilde{\boldsymbol{H}} = \frac{k}{\mu_0 \omega} \hat{\boldsymbol{k}} \times \tilde{\boldsymbol{E}} = \frac{1}{Z_0} \hat{\boldsymbol{k}} \times \tilde{\boldsymbol{E}},} \tag{3.27}$$

which tells us the magnetic-field amplitude as a function of the electric-field amplitude and the free-space impedance Z_0. This relation between $\tilde{\boldsymbol{E}}$ and $\tilde{\boldsymbol{H}}$ is a generalization of the relation in Eq. (2.36) for the (1+1)D case.

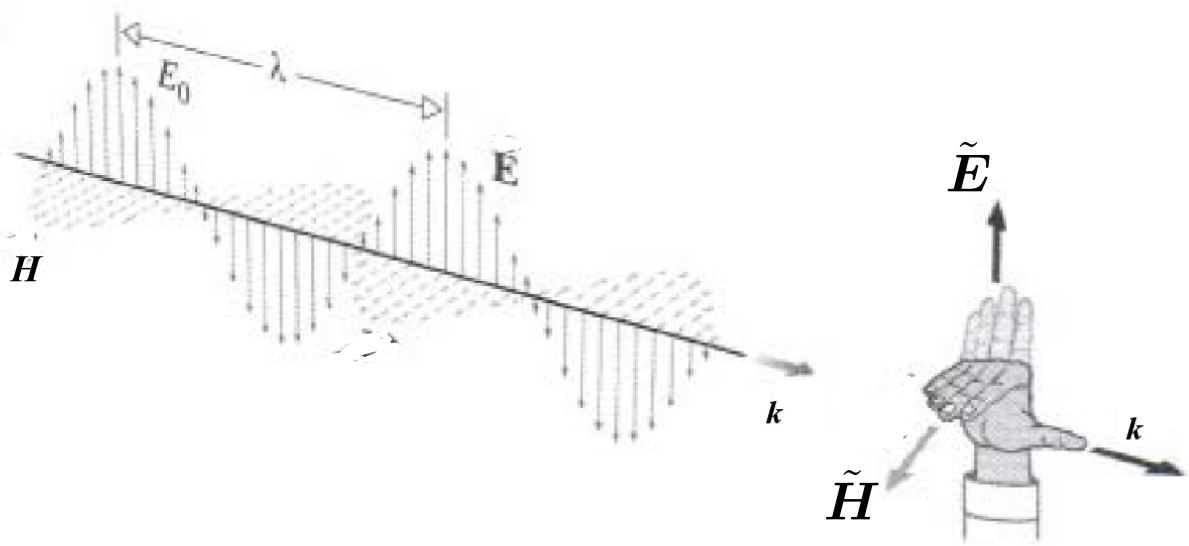


FIGURE 3.3. An electromagnetic monochromatic plane wave. The three vectors must be perpendicular to one another and in the directions specified by the right-hand rule in the figure.

- **Exercise**: Suppose that $\boldsymbol{k}$ is in the (x, z) plane and given by

$$\boldsymbol{k} = k \sin\theta \hat{\boldsymbol{x}} + k \cos\theta \hat{\boldsymbol{z}}, \tag{3.28}$$

where θ is the angle between $\boldsymbol{k}$ and $\hat{\boldsymbol{z}}$, and the electric field is in the y direction:

$$\boldsymbol{E} = \tilde{E} \hat{\boldsymbol{y}} \exp(j\boldsymbol{k} \cdot \boldsymbol{r} - j\omega t), \tag{3.29}$$

as shown in Fig. 3.4. What is $\boldsymbol{H}$?
Answer: The unit vector along $\boldsymbol{k}$ is

$$\hat{\boldsymbol{k}} = \sin\theta \hat{\boldsymbol{x}} + \cos\theta \hat{\boldsymbol{z}}, \tag{3.30}$$

so $\tilde{\boldsymbol{H}}$ is

$$\tilde{\boldsymbol{H}} = \frac{1}{Z_0} \hat{\boldsymbol{k}} \times \hat{\boldsymbol{y}} \tilde{E} = \frac{\tilde{E}}{Z_0} \left(-\cos\theta \hat{\boldsymbol{x}} + \sin\theta \hat{\boldsymbol{z}}\right), \tag{3.31}$$

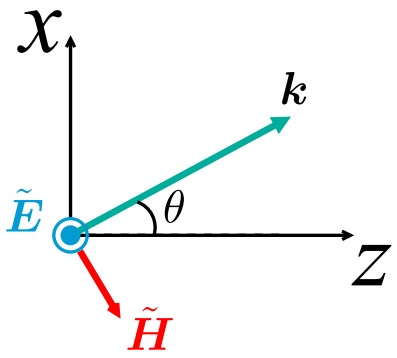


FIGURE 3.4. In this example, the wavevector is $\boldsymbol{k} = k(\sin\theta\hat{\boldsymbol{x}} + \cos\theta\hat{\boldsymbol{z}})$, and $\tilde{\boldsymbol{E}}$ is in the $\hat{\boldsymbol{y}}$ direction. What is $\boldsymbol{H}$?

where we've used $\hat{\boldsymbol{x}} \times \hat{\boldsymbol{y}} = \hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{z}} \times \hat{\boldsymbol{y}} = -\hat{\boldsymbol{x}}$. Hence

$$\boldsymbol{H} = \frac{\tilde{E}}{Z_0}\left(-\cos\theta\hat{\boldsymbol{x}} + \sin\theta\hat{\boldsymbol{z}}\right)\exp(j\boldsymbol{k}\cdot\boldsymbol{r} - j\omega t). \tag{3.32}$$

CHAPTER 4

Energy

4.1. Energy density

The EM **energy density** in free space is[1]

$$\boxed{u(\boldsymbol{r},t) = \frac{1}{2}\epsilon_0 \boldsymbol{E}(\boldsymbol{r},t)\cdot\boldsymbol{E}(\boldsymbol{r},t) + \frac{1}{2}\mu_0 \boldsymbol{H}(\boldsymbol{r},t)\cdot\boldsymbol{H}(\boldsymbol{r},t) \quad (\text{Joule/m}^3).} \tag{4.1}$$

The unit is energy per volume. The energy density is analogous to charge density ρ that we studied in Sec. 1.1, except that here we are considering the density of energy in the electromagnetic fields, not charge. It is a **quadratic** function of the electromagnetic fields. We should be careful when calculating it for complex EM fields; **always take the real part first** before taking the products.

To obtain the EM energy in a volume, picture Fig. 1.3, except that each infinitesimal box contains energy $u(\boldsymbol{r},t)dxdydz$. Summing over all the energy in the boxes and passing to the continuous limit, we obtain a volume integral:

$$\text{Energy in a volume} = \int_{\text{volume}} u(\boldsymbol{r},t)dV, \tag{4.2}$$

just like the integral of a charge density to obtain the total charge within.

Consider the real monochromatic plane wave solution given by Eq. (2.29) and (2.36):

$$\boldsymbol{E} = E_{\max}\hat{\boldsymbol{x}}\cos(kz-\omega t+\phi), \qquad \boldsymbol{H} = \frac{E_{\max}}{Z_0}\hat{\boldsymbol{y}}\cos(kz-\omega t+\phi). \tag{4.3}$$

The electric field has an energy density

$$\frac{1}{2}\epsilon_0\boldsymbol{E}\cdot\boldsymbol{E} = \frac{1}{2}\epsilon_0 E_{\max}^2\cos^2(kz-\omega t+\phi), \tag{4.4}$$

and the magnetic field has an energy density given by

$$\frac{1}{2}\mu_0\boldsymbol{H}\cdot\boldsymbol{H} = \frac{1}{2}\frac{\mu_0}{Z_0^2}E_{\max}^2\cos^2(kz-\omega t+\phi) = \frac{1}{2}\epsilon_0 E_{\max}^2\cos^2(kz-\omega t+\phi), \tag{4.5}$$

(remember that $Z_0 = \sqrt{\mu_0/\epsilon_0}$, so $\mu_0/Z_0^2 = \epsilon_0$). Note that the electric-field energy density and the magnetic-field energy density are equal in an electromagnetic wave. The total energy density becomes

$$u(\boldsymbol{r},t) = \epsilon_0 E_{\max}^2\cos^2(kz-\omega t+\phi). \tag{4.6}$$

At a given z, the energy density oscillates between 0 and $\epsilon_0 E_{\max}^2$.

In optics, ω can be very high, so our detectors (eyes, cameras, etc.) cannot usually respond so quickly to the instantaneous energy density. Instead, the usual measured quantity is the **time-averaged** energy density, where $u(\boldsymbol{r},t)$ is integrated over a very long time relative to the temporal period $2\pi/\omega$ and averaged. At this point, it is helpful to note that the average of $\cos^2$ is $1/2$:

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T dt\cos^2(kz-\omega t+\phi) = \frac{1}{2}, \tag{4.7}$$

so the **time-averaged** energy density, denoted by the bar¯, of an electromagnetic monochromatic plane wave is

$$\bar{u}(\boldsymbol{r}) \equiv \lim_{T\to\infty}\frac{1}{T}\int_0^T dt u(\boldsymbol{r},t) = \frac{1}{2}\epsilon_0 E_{\max}^2, \tag{4.8}$$

which is constant everywhere.

[1] *There is a very deep reason in classical Hamiltonian mechanics behind why this expression is the right one for the EM energy density, but we shall not bother with it here.

- **Exercise**: Verify Eq. (4.7).
- **Question**: What is the total energy of this monochromatic wave solution?
  Answer: Since the time-averaged energy density $\bar{u}$ is constant, the energy inside any box with volume V is $\bar{u}V$. If the volume is infinite, the time-averaged energy is also infinite! This tells us that, in practice, the plane wave solution is **unphysical**; we will never have this solution on its own in real life. Fortunately, we can still form physical solutions with finite energies by adding many plane waves together (as is done in *Fourier Optics* [2]). In many cases, the plane wave is still a very good **approximation** of physical waves, such as laser beams, provided that the propagation is not too far.

4.2. Power density

The electromagnetic power density is

$$\boxed{\boldsymbol{S}(\boldsymbol{r},t) = \boldsymbol{E} \times \boldsymbol{H} \text{ (W/m}^2\text{)},} \tag{4.9}$$

which is called the **Poynting vector**. The Poynting vector is analogous to the current density $\boldsymbol{J}(\boldsymbol{r},t)$ that we studied in Sec. 1.2, except that here we are considering the electromagnetic power (energy per second, or Watt) rather than current (charge per second).

To convince you that this is the right expression for power density, consider the plane wave solution given by Eqs. (4.3), and picture Fig. 1.5, but instead of a volume of charge, here we consider a volume of energy $u(\boldsymbol{r},t)dxdydz$ with energy density given by Eq. (4.6). The velocity of the wave is $\boldsymbol{v} = c\hat{\boldsymbol{z}}$, so the power dP across the surface with area $dxdz$ is

$$dP = \frac{u(\boldsymbol{r},t)dxdydz}{dt} = u(\boldsymbol{r},t)cdxdz, \tag{4.10}$$

and the power per unit area becomes

$$S = \frac{dP}{dxdz} = u(\boldsymbol{r},t)c = \epsilon_0 c E_{\max}^2 \cos^2(kz-\omega t+\phi) = \frac{E_{\max}^2}{Z_0}\cos^2(kz-\omega t+\phi). \tag{4.11}$$

Lo and behold, this is the same as the magnitude of the Poynting vector:

$$\boldsymbol{S} = [E_{\max}\hat{\boldsymbol{x}}\cos(kz-\omega t+\phi)] \times \left[\frac{E_{\max}}{Z_0}\hat{\boldsymbol{y}}\cos(kz-\omega t+\phi)\right] \tag{4.12}$$

$$= \frac{E_{\max}^2}{Z_0}\cos^2(kz-\omega t+\phi)\hat{\boldsymbol{z}}. \tag{4.13}$$

Again, because ω is very high for optics and our detectors (eyes, cameras, etc.) cannot respond to such high frequencies, we are often more interested in the time-averaged quantity:

$$\bar{\boldsymbol{S}}(\boldsymbol{r}) = \lim_{T\to\infty}\frac{1}{T}\int_0^T \boldsymbol{S}(\boldsymbol{r},t). \tag{4.14}$$

For the plane wave example,

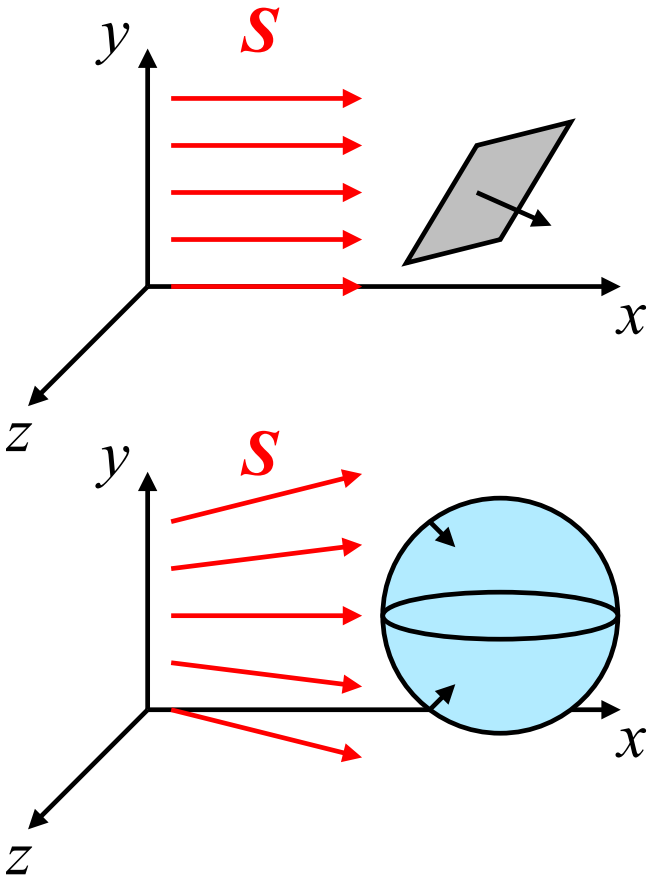
$$\bar{\boldsymbol{S}} = \frac{E_{\max}^2}{2Z_0}\hat{\boldsymbol{z}}. \tag{4.15}$$

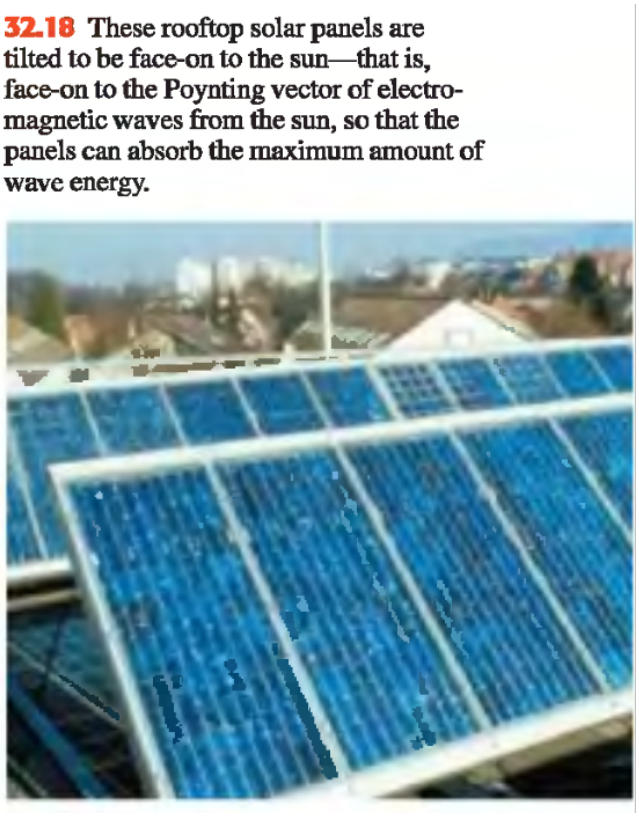
For a plane wave propagating in any direction, Fig. 3.3 should convince you that $\boldsymbol{E}\times\boldsymbol{H}$ is always parallel to $\boldsymbol{k}$, the wave propagation direction, as one might expect.

As we discussed in Sec. 1.2 for the case of current density, the power across a surface is different when the power density is not perpendicular to the surface. The result is

$$\boxed{\text{EM power across a surface} = \int_{\text{surface}} \boldsymbol{S}\cdot d\boldsymbol{A}.} \tag{4.16}$$

This dot product is quite important for **solar energy**, as the solar power that can be absorbed by a solar panel is maximized if the surface is **perpendicular** to the Poynting vector.





32.18 These rooftop solar panels are tilted to be face-on to the sun—that is, face-on to the Poynting vector of electromagnetic waves from the sun, so that the panels can absorb the maximum amount of wave energy.

FIGURE 4.1. Left figure: some examples of Poynting vector and surfaces. The total across a surface is the surface integral of the Poynting vector. Right figure is from Chap. 32 of [1].

4.3. Poynting vector for complex fields

We shall now calculate the Poynting vector if we assume complex fields with the following time dependence:

$$\boldsymbol{E} = \boldsymbol{\mathcal{E}}(\boldsymbol{r}) \exp(-j\omega t), \qquad \boldsymbol{H} = \boldsymbol{\mathcal{H}}(\boldsymbol{r}) \exp(-j\omega t). \tag{4.17}$$

Such fields are **monochromatic** as they oscillate at a single frequency ω in time, but they may not be a plane wave, as $\boldsymbol{\mathcal{E}}(\boldsymbol{r})$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{r})$ may have a more nontrivial spatial dependence. Since $\boldsymbol{S}$ is a **quadratic** function of the fields, it is important that we take the real part first:

$$\boldsymbol{S} = \operatorname{Re} \boldsymbol{E} \times \operatorname{Re} \boldsymbol{H} \tag{4.18}$$

$$= \frac{1}{2} (\boldsymbol{E} + \boldsymbol{E}^*) \times \frac{1}{2} (\boldsymbol{H} + \boldsymbol{H}^*) \tag{4.19}$$

$$= \frac{1}{4} (\boldsymbol{E} \times \boldsymbol{H} + \boldsymbol{E}^* \times \boldsymbol{H}^* + \boldsymbol{E} \times \boldsymbol{H}^* + \boldsymbol{E}^* \times \boldsymbol{H}) . \tag{4.20}$$

Note that

$$\boldsymbol{E} \times \boldsymbol{H} = \boldsymbol{\mathcal{E}}(\boldsymbol{r}) \times \boldsymbol{\mathcal{H}}(\boldsymbol{r}) e^{-2j\omega t} \propto e^{-2j\omega t}, \qquad \boldsymbol{E}^* \times \boldsymbol{H}^* \propto e^{2j\omega t}, \tag{4.21}$$

whereas

$$\boldsymbol{E} \times \boldsymbol{H}^* = \boldsymbol{\mathcal{E}}(\boldsymbol{r}) \times \boldsymbol{\mathcal{H}}^*(\boldsymbol{r}) \tag{4.22}$$

and $\boldsymbol{E}^* \times \boldsymbol{H} = (\boldsymbol{E} \times \boldsymbol{H}^*)^* = \boldsymbol{\mathcal{E}}(\boldsymbol{r}) \times \boldsymbol{\mathcal{H}}^*(\boldsymbol{r})$ are constant in time. When we average over time,

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T dt e^{\pm 2j\omega t} = 0. \tag{4.23}$$

This means that, for the **time-averaged** Poynting vector, the first two terms in Eq. (4.20) average to zero, and we are left with

$$\boxed{\bar{\boldsymbol{S}}(\boldsymbol{r}) = \frac{1}{4}\left(\boldsymbol{E}\times\boldsymbol{H}^* + \boldsymbol{E}^*\times\boldsymbol{H}\right) = \frac{1}{2}\,\mathrm{Re}\left(\boldsymbol{E}\times\boldsymbol{H}^*\right)}$$

$$\boxed{\text{for complex fields } \boldsymbol{E}\propto e^{-j\omega t},\, \boldsymbol{H}\propto e^{-j\omega t}.} \tag{4.24}$$

For example, for a plane wave propagating in the z direction, a complex form of the solution is shown in Eqs. (2.43) and (2.44), leading to

$$\boldsymbol{\mathcal{E}}(\boldsymbol{r}) = \hat{\boldsymbol{x}}\tilde{E}\exp(jkz) \qquad \boldsymbol{\mathcal{H}}(\boldsymbol{r}) = \hat{\boldsymbol{y}}\frac{\tilde{E}}{Z_0}\exp(jkz), \qquad \tilde{E} = E_{\max}e^{j\phi}, \tag{4.25}$$

and the time-averaged Poynting vector is

$$\bar{\boldsymbol{S}} = \hat{\boldsymbol{z}}\frac{|\tilde{E}|^2}{2Z_0}, \tag{4.26}$$

which is the same as Eq. (4.15). The important point to note here is that the **intensity**, which is defined as the magnitude of $\bar{\boldsymbol{S}}$,

$$\boxed{I \equiv |\bar{\boldsymbol{S}}|,} \tag{4.27}$$

is proportional to $|\tilde{E}|^2$.

- **Exercise**: Show that, for a plane wave propagating in a general $\boldsymbol{k}$ direction (see Chap. 3), the time-averaged Poynting vector is

$$\boxed{\bar{\boldsymbol{S}} = \hat{\boldsymbol{k}}\frac{|\tilde{E}|^2}{2Z_0},} \tag{4.28}$$

  where $\hat{\boldsymbol{k}}$ is the unit vector for $\boldsymbol{k}$.
- **Exercise**: Show that the time-averaged energy density is

$$\bar{u} = \frac{1}{4}\epsilon_0\,\mathrm{Re}\left(\boldsymbol{\mathcal{E}}\cdot\boldsymbol{\mathcal{E}}^*\right) + \frac{1}{4}\mu_0\,\mathrm{Re}\left(\boldsymbol{\mathcal{H}}\cdot\boldsymbol{\mathcal{H}}^*\right). \tag{4.29}$$

4.4. *Energy conservation (advanced topic but important)

To double-check that $\boldsymbol{S}$ is indeed the EM power density, it should obey an energy-conservation relation like the relation between $\boldsymbol{J}$ and ρ for charges in Sec. 1.7. To derive a continuity equation for EM energy, Let's take the divergence of $\boldsymbol{S}$:

$$\nabla\cdot\boldsymbol{S} = \nabla\cdot\left(\boldsymbol{E}\times\boldsymbol{H}\right). \tag{4.30}$$

A vector-calculus identity (http://en.wikipedia.org/wiki/Vector_calculus_identities) says

$$\nabla\cdot\left(\boldsymbol{E}\times\boldsymbol{H}\right) = \boldsymbol{H}\cdot\left(\nabla\times\boldsymbol{E}\right) - \boldsymbol{E}\cdot\left(\nabla\times\boldsymbol{H}\right). \tag{4.31}$$

Now substitute in Faraday's law for $\nabla\times\boldsymbol{E}$ and modified Ampere's law for $\nabla\times\boldsymbol{H}$ (source-free):

$$\nabla\cdot\left(\boldsymbol{E}\times\boldsymbol{H}\right) = \boldsymbol{H}\cdot\left(-\mu_0\frac{\partial\boldsymbol{H}}{\partial t}\right) - \boldsymbol{E}\cdot\left(\epsilon_0\frac{\partial\boldsymbol{E}}{\partial t}\right), \tag{4.32}$$

which is equal to the negative of

$$\frac{\partial u}{\partial t} = \frac{1}{2}\epsilon_0\left(\frac{\partial\boldsymbol{E}}{\partial t}\cdot\boldsymbol{E} + \boldsymbol{E}\cdot\frac{\partial\boldsymbol{E}}{\partial t}\right) + \frac{1}{2}\mu_0\left(\frac{\partial\boldsymbol{H}}{\partial t}\cdot\boldsymbol{H} + \boldsymbol{H}\cdot\frac{\partial\boldsymbol{H}}{\partial t}\right) \tag{4.33}$$

$$= \epsilon_0\boldsymbol{E}\cdot\frac{\partial\boldsymbol{E}}{\partial t} + \mu_0\boldsymbol{H}\cdot\frac{\partial\boldsymbol{H}}{\partial t}. \tag{4.34}$$

Hence

$$\frac{\partial u}{\partial t} + \nabla\cdot\boldsymbol{S} = 0. \tag{4.35}$$

This is a continuity equation like the one for charge in Eq. (1.39) and is a statement of **energy conservation**. Integrating in volume, the energy inside is

$$U = \int u dV, \tag{4.36}$$

and the continuity equation becomes

$$\boxed{\frac{\partial U}{\partial t} = -\oint \boldsymbol{S} \cdot d\boldsymbol{A},} \tag{4.37}$$

which says that the rate of change of energy in a volume, or the total **power** into the volume, is $-\oint \boldsymbol{S} \cdot d\boldsymbol{A}$. This confirms that $\oint \boldsymbol{S} \cdot d\boldsymbol{A}$ is the power flowing out of the volume and $\boldsymbol{S}$ is indeed the power density.

- **Exercise**: Suppose that the (time-averaged) Poynting vector is constant with magnitude I and pointing in the $\hat{\boldsymbol{z}}$ direction, as shown in Fig. 4.2. Suppose that an object inside the cylinder completely absorbs the incoming energy, such that the Poynting vector is zero on the other side of the cylinder. What is the absorbed power?
  Answer: If the cylinder has radius R on its circular surface, so that the area of the circle is πR^2, the absorbed power is

$$\frac{\partial U}{\partial t} = I\pi R^2. \tag{4.38}$$

  The shape of the object inside the cylinder does not matter, I don't need to compute the surface integral for the object itself, I just need to appeal to energy conservation to know that the power absorbed must be equal to the power going in the cylinder.

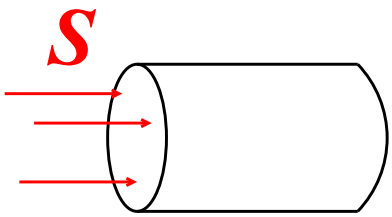


FIGURE 4.2. Power flow into a cylinder.

CHAPTER 5

*Momentum

5.1. *Momentum density

Energy conservation and **momentum conservation** are among the most important principles in physics. When EM waves interact with matter (e.g., absorption, reflection, refraction, emission), the total energy and the total momentum of the whole system must be conserved.

What is momentum? For Newtonian massive objects,

$$\boldsymbol{p} \approx m\boldsymbol{v}. \tag{5.1}$$

For relativistic massive objects (http://en.wikipedia.org/wiki/Momentum#Four-vector_formulation)

$$\boldsymbol{p} = \frac{m\boldsymbol{v}}{\sqrt{1 - v^2/c^2}}, \tag{5.2}$$

where m is the rest mass. For monochromatic wave, we might expect that the direction of momentum should be along the direction of power flow $\boldsymbol{S}$, but we don't really know the magnitude, as EM waves have zero mass and speed of light $v = c$, so Eq. (5.2) blows up both in the numerator and the denominator.

We have to rely on the relativistic energy-momentum relation (http://en.wikipedia.org/wiki/Energy-momentum_relation):

$$U^2 = m^2c^4 + p^2c^2. \tag{5.3}$$

With $m = 0$, we know that the magnitude of $\boldsymbol{p}$ must be

$$p = \frac{U}{c}, \tag{5.4}$$

meaning that the **EM momentum magnitude is proportional to the energy** (compare with the relation with mass or Newtonian $U = p^2/2m$). The momentum density is thus $u(\boldsymbol{r}, t)/c$. We already saw in Sec. 4.2 that $Sdxdz = uc$ for the plane-wave example. In general, the EM momentum density (momentum per unit area) turns out to be equal to

$$\boldsymbol{g}(\boldsymbol{r}, t) = \frac{\boldsymbol{S}(\boldsymbol{r}, t)}{c^2}. \tag{5.5}$$

See http://www.feynmanlectures.caltech.edu/II_27.html#Ch27-S6 for a more detailed explanation.

5.2. *Radiation pressure

If an object completely **absorbs** the monochromatic wave propagating in z direction given by Eq. (4.3), the wave momentum must be transferred to the object. Suppose that the object has a flat surface in the (x, y) plane with area A; picture Fig. 1.5, except that the volume contains energy and the energy is absorbed by the surface. At time $t + dt$, the volume of EM wave is absorbed, and the volume is $Ady = Acdt$. For infinitesimal time we can assume dy and dt to be very small and assume u to be constant near the surface, so the energy in this volume is simply

$$dU = uAdy = uAcdt, \tag{5.6}$$

and the momentum is

$$dp = \frac{dU}{c} = uAdt. \tag{5.7}$$

The object must then acquire a momentum given by $dp = \frac{u}{c}Acdt = uAdt$ along the z direction. By Newton's second law, the rate of change of momentum is force:

$$F = \frac{dp}{dt} = Au = A\epsilon_0 E_{\max}^2 \cos^2(kx - \omega t + \phi) \text{ (N)}, \tag{5.8}$$

which has the unit of Newton. The **radiation pressure** (force/unit area) is then

$$P_{\text{rad}} = \frac{F}{A} = \epsilon_0 E_{\max}^2 \cos^2(kx - \omega t + \phi). \tag{5.9}$$

Averaged over time,

$$\bar{P}_{\text{rad}} = \frac{1}{2}\epsilon_0 E_{\max}^2. \tag{5.10}$$

If the wave is completely **reflected** by a perpendicular mirror, the wave reverses its direction, that is, the wave momentum change is -2$p\hat{\boldsymbol{x}}$. The radiation pressure is then

$$\bar{P}_{\text{rad,reflect}} = \epsilon_0 E_{\max}^2, \tag{5.11}$$

which is twice the value for absorption.

Radiation pressure is actually miniscule and mostly negligible in everyday life, but has some specialized applications (optical tweezing, solar sail, Casimir force, etc.).

CHAPTER 6

Waves in Dielectric Media

6.1. Simple dielectric model

We shall consider a more complicated scenario, where a medium produces charge and current densities in response to the electromagnetic fields. Instead of modeling the sources and their dependence on the fields explicitly, it is often more convenient to define a new field called the **displacement field** $\boldsymbol{D}$ that changes as a function of $\boldsymbol{E}$, and simply model the medium response by describing how $\boldsymbol{D}$ varies with $\boldsymbol{E}$. Similarly, we can assume that $\boldsymbol{B}$ may have a more complicated relationship with $\boldsymbol{H}$, although $\boldsymbol{B} = \mu_0 \boldsymbol{H}$ is usually an excellent approximation for optics. In terms of $\boldsymbol{D}$ and $\boldsymbol{B}$, the **Maxwell's equations in a dielectric medium** now read

$$\nabla \cdot \boldsymbol{D} = 0, \tag{6.1}$$

$$\nabla \cdot \boldsymbol{B} = 0, \tag{6.2}$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{6.3}$$

$$\nabla \times \boldsymbol{H} = \frac{\partial \boldsymbol{D}}{\partial t}, \tag{6.4}$$

where we assume no **free charge** and no **free current**, meaning that the effect of all the actual charges and currents in the medium is already modeled by $\boldsymbol{D}$ and $\boldsymbol{B}$ and there is no additional sources that we need to worry about. In free space, obviously $\boldsymbol{D} = \epsilon_0 \boldsymbol{E}$, $\boldsymbol{B} = \mu_0 \boldsymbol{H}$.

The simplest model of a dielectric is to simply replace ϵ_0 by the **permittivity of the medium**:

$$\boxed{\boldsymbol{D} = \epsilon \boldsymbol{E},} \tag{6.5}$$

where we assume ϵ to be real for now. All our previous discussions remain valid if we replace ϵ_0 with ϵ. **Linearity** still works. The speed of light in the dielectric becomes

$$\boxed{v = \frac{1}{\sqrt{\mu_0 \epsilon}} = \frac{c}{n} = \frac{\omega}{k},} \tag{6.6}$$

where n is the **refractive index**:

$$\boxed{n = \sqrt{\frac{\epsilon}{\epsilon_0}},} \tag{6.7}$$

which is a dimensionless number that characterizes how much the speed slows down in the medium. Fig. 6.1 shows some typical values of n.

The wavelength in the medium is still the spatial period, or

$$\lambda = \frac{2\pi}{k}, \tag{6.8}$$

but since $k = \omega n / c$ now, for a given ω,

$$\boxed{\lambda = \frac{2\pi c}{\omega n} = \frac{\lambda_0}{n},} \tag{6.9}$$

where it is smaller than the free-space value λ_0, now denoted by the subscript 0,

$$\lambda_0 \equiv \frac{2\pi c}{\omega} \tag{6.10}$$

by a factor of n. In most EM problems involving many media, the frequency ω is given and determined by the frequency of the external source or boundary condition. Because of the linear-time-invariant property of the Maxwell's equations, **a single-frequency source will lead to single-frequency fields everywhere**, while the

Table 33.1 Index of Refraction for Yellow Sodium Light $\lambda_0 = 589$ nm

Substance	Index of Refraction, n
Solids	
Ice (H_2O)	1.309
Fluorite (CaF_2)	1.434
Polystyrene	1.49
Rock salt (NaCl)	1.544
Quartz (SiO_2)	1.544
Zircon ($ZrO_2 \cdot SiO_2$)	1.923
Diamond (C)	2.417
Fabulite ($SrTiO_3$)	2.409
Rutile (TiO_2)	2.62
Glasses (typical values)	
Crown	1.52
Light flint	1.58
Medium flint	1.62
Dense flint	1.66
Lanthanum flint	1.80
Liquids at 20°C	
Methanol (CH_3OH)	1.329
Water (H_2O)	1.333
Ethanol (C_2H_5OH)	1.36
Carbon tetrachloride (CCl_4)	1.460
Turpentine	1.472
Glycerine	1.473
Benzene	1.501
Carbon disulfide (CS_2)	1.628

FIGURE 6.1. Some typical values of refractive indices from Chap. 33, Ref. [1].

spatial properties such as the wavelengths λ and the wavevectors $\boldsymbol{k}$ differ depending on the refractive indices. The dispersion relation becomes

$$k^2 = k_x^2 + k_y^2 + k_z^2 = \left(\frac{\omega n}{c}\right)^2, \tag{6.11}$$

which is also a sphere in $\boldsymbol{k}$ space like Fig. 3.2, except that the radius is multiplied by n, as shown in Fig. 6.2. It implies that no component of $\boldsymbol{k}$ can exceed $\omega n/c$ if all components are real.

As the impedance of the medium is also modified, the magnetic-field amplitude for a plane wave is

$$H_{\max} = \frac{E_{\max}}{Z}, \qquad \boxed{Z = \sqrt{\frac{\mu_0}{\epsilon}} = \sqrt{\frac{\mu_0}{\epsilon_0}}\sqrt{\frac{\epsilon_0}{\epsilon}} = \frac{Z_0}{n}.} \tag{6.12}$$

Moreover, the energy density is now

$$u = \frac{1}{2}\epsilon \boldsymbol{E}\cdot\boldsymbol{E} + \frac{1}{2}\mu_0 \boldsymbol{H}\cdot\boldsymbol{H}, \tag{6.13}$$

but the Poynting vector remains the same expression:

$$\boldsymbol{S} = \boldsymbol{E}\times\boldsymbol{H}. \tag{6.14}$$

- **Exercise**: Calculate the time-averaged Poynting vector for a monochromatic plane wave propagating in the $\boldsymbol{k}$ direction using Sec. 3.2 and Sec. 4.3 in a medium with refractive index n.
  Answer:

$$\bar{\boldsymbol{S}} = \frac{|\tilde{E}|^2}{2Z}\hat{\boldsymbol{k}}, \qquad Z = \sqrt{\frac{\mu_0}{\epsilon}} = \frac{Z_0}{n}. \tag{6.15}$$

- **Exercise**: Derive the continuity equation for u and $\boldsymbol{S}$ in a dielectric with permittivity ϵ.

Note that the constant ϵ is a very simplistic model of dielectric media. In reality,

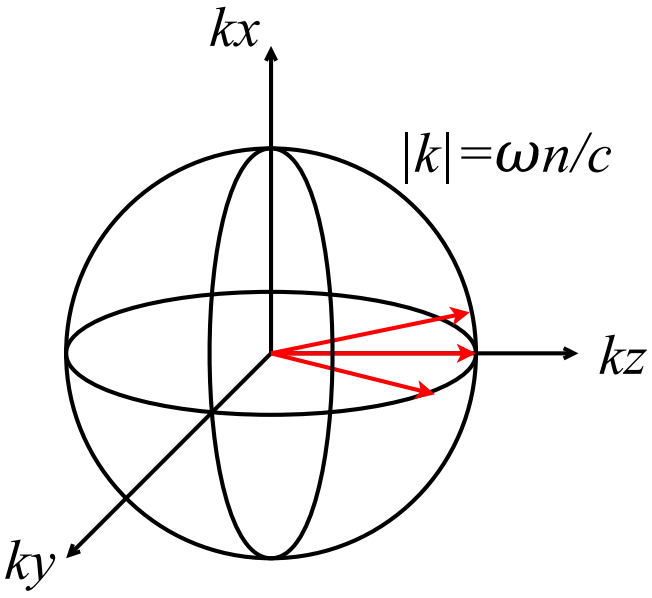


FIGURE 6.2. The dispersion relation in $\boldsymbol{k}$ space for a dielectric medium with refractive index n.

- ϵ and n usually depend on frequency and are complex, leading to dispersion and loss, which are important effects for optical communications,
- ϵ may even be a matrix instead of a scalar; as such, the speed of light may depend on the electric-field direction.
- ϵ may depend on $\boldsymbol{E}$ if the field magnitude is very high. In that case the Maxwell's equations are no longer linear, and we would be dealing with *nonlinear optics*, which is beyond the scope of this module.
- The permeability may also change from μ_0, although this is not usual in optics.

6.2. Boundary conditions between two media

Suppose there are two media with different refractive indices and there is no free charge or free current along the interface. Maxwell's equations given by Eqs. (6.1)–(6.4) still hold. Again we will assume

$$\boldsymbol{B} = \mu_0 \boldsymbol{H}, \qquad \boldsymbol{D} = \epsilon \boldsymbol{E}, \tag{6.16}$$

where ϵ depends on the medium.

- **Exercise**: Show that **linearity** still works when ϵ is a function of position $\boldsymbol{r}$.

Suppose that we may have some EM fields on one side, such as an incident laser beam, and some EM fields on the other side. To solve for the relations between the EM fields on both sides, it is important that we know the **boundary conditions** between the two media. The derivation can be found in any electromagnetics or optics textbook; see, for example, Griffiths, *Introduction to Electrodynamics* [3]. The derivation involves expressing Eqs. (6.1)–(6.4) in integral form and considering the fields in an infinitesimal volume or loop near the surface. I will not go through the derivation in detail here, but simply tell you briefly how it works.

Suppose that $\hat{\boldsymbol{n}}$ is the normal vector of the interface. The first condition involves the displacement field $\boldsymbol{D}$. It says that that the component of $\boldsymbol{D}$ in medium 1 projected on the normal vector $\hat{\boldsymbol{n}}$, given by $\hat{\boldsymbol{n}} \cdot \boldsymbol{D}_1$, has to be equal to its counterpart $\hat{\boldsymbol{n}} \cdot \boldsymbol{D}_2$ in the second medium. It comes from the fact that we can imagine a small cylinder around the surface, as shown in Fig. 6.3, and from Gauss's law, $\oint \boldsymbol{D} \cdot d\boldsymbol{A} = 0$, or the net flux with respect to $\boldsymbol{D}$ has to be zero. If we assume that the cylinder has infinitesimal height $h \to 0$, and dA is small enough for $\hat{\boldsymbol{n}} \cdot \boldsymbol{D}_1$ and $\hat{\boldsymbol{n}} \cdot \boldsymbol{D}_2$ there to be constant, the flux consists only the $\boldsymbol{D}_1 \cdot \hat{\boldsymbol{n}} dA$ on the top surface and $-\boldsymbol{D}_2 \cdot \hat{\boldsymbol{n}} dA$ on the bottom surface:

$$\oint \boldsymbol{D} \cdot d\boldsymbol{A} = \boldsymbol{D}_1 \cdot \hat{\boldsymbol{n}} dA - \boldsymbol{D}_2 \cdot \hat{\boldsymbol{n}} dA. \tag{6.17}$$

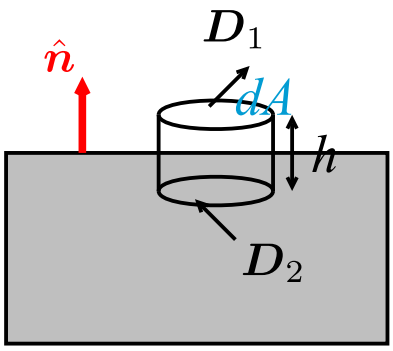


FIGURE 6.3. Derivation of the boundary condition for the field components normal to the interface. dA is the area of the flat surfaces on the cylinder and h is the height of the cylinder. The integral form of Gauss's laws means that the flux $\oint \boldsymbol{D} \cdot d\boldsymbol{A}$ is zero, and for h going to zero the closed surface integral consists only of $\boldsymbol{D}_1 \cdot \hat{\boldsymbol{n}} dA - \boldsymbol{D}_2 \cdot \hat{\boldsymbol{n}} dA$. In other words, $\boldsymbol{D}_1 \cdot \hat{\boldsymbol{n}} = \boldsymbol{D}_2 \cdot \hat{\boldsymbol{n}}$.

Hence

$$\hat{\boldsymbol{n}} \cdot \boldsymbol{D}_1 = \hat{\boldsymbol{n}} \cdot \boldsymbol{D}_2. \tag{6.18}$$

This must be satisfied everywhere along the interface, as the cylinder can be anywhere and infinitesimally small. The derivation is the same for $\hat{\boldsymbol{n}} \cdot \boldsymbol{B}_1$ and $\hat{\boldsymbol{n}} \cdot \boldsymbol{B}_2$ with $\oint \boldsymbol{B} \cdot d\boldsymbol{A} = 0$, resulting in

$$\hat{\boldsymbol{n}} \cdot \boldsymbol{B}_1 = \hat{\boldsymbol{n}} \cdot \boldsymbol{B}_2. \tag{6.19}$$

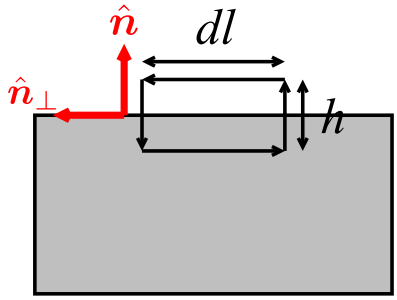


FIGURE 6.4. Derivation of the boundary condition for the parallel field components.

For another set of boundary conditions, we consider a loop as sketched in Fig. 6.4. The integral form of Faraday's law is

$$\oint \boldsymbol{E} \cdot d\boldsymbol{l} = -\frac{\partial}{\partial t} \int \boldsymbol{B} \cdot d\boldsymbol{A}. \tag{6.20}$$

If we make $h \to 0$, the area of the loop and therefore the magnetic flux $\int \boldsymbol{B} \cdot d\boldsymbol{A}$ go to zero, while the line integral consists only of the parallel component of the electric field on one side and the counterpart on the either side. For small enough dl,

$$\oint \boldsymbol{E} \cdot d\boldsymbol{l} = \boldsymbol{E}_1 \cdot \hat{\boldsymbol{n}}_\perp dl - \boldsymbol{E}_2 \cdot \hat{\boldsymbol{n}}_\perp dl, \tag{6.21}$$

where $\hat{\boldsymbol{n}}_\perp$ is a vector perpendicular to the normal vector ($\hat{\boldsymbol{n}}_\perp \cdot \hat{\boldsymbol{n}} = 0$) and therefore parallel to the surface. Since the magnetic flux and therefore the line integral are zero as $h \to 0$, we are left with equal electric fields when they are projected onto $\hat{\boldsymbol{n}}_\perp$:

$$\hat{\boldsymbol{n}}_\perp \cdot \boldsymbol{E}_1 = \hat{\boldsymbol{n}}_\perp \cdot \boldsymbol{E}_2. \tag{6.22}$$

The argument is the same for the components of $\boldsymbol{H}$ along the interface:

$$\hat{\boldsymbol{n}}_\perp \cdot \boldsymbol{H}_1 = \hat{\boldsymbol{n}}_\perp \cdot \boldsymbol{H}_2. \tag{6.23}$$

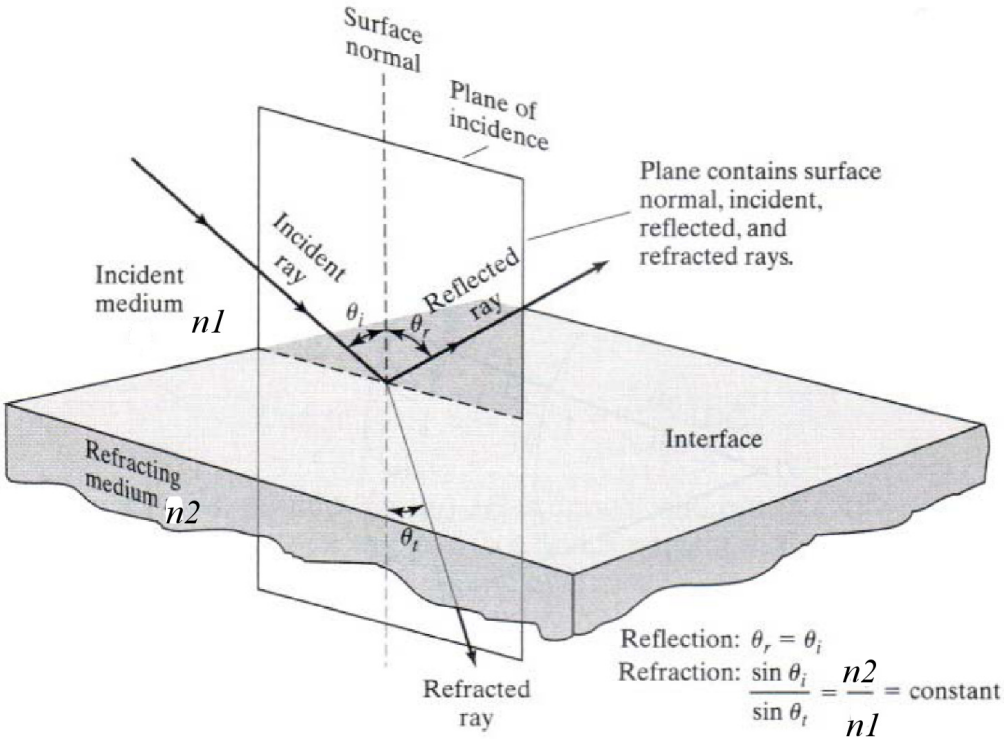


FIGURE 6.5. Geometry of reflection and refraction of plane waves (from [3])

6.3. Reflection and refraction between two media

Suppose that the first medium has refractive index n_1 and the second medium has refractive index n_2 with a flat interface, as shown in Fig. 6.5. To simplify the mathematics, we can assume that the z axis is pointing down, as in Fig. 6.6, the interface is in the (x, y) plane, medium 1 is in $z < 0$, and medium 2 is in $z > 0$, such that the normal vector is $\hat{\boldsymbol{z}}$. An **incident** monochromatic plane wave from medium 1 is

$$\boldsymbol{E}_i = \tilde{\boldsymbol{E}}_i \exp(j\boldsymbol{k}_i \cdot \boldsymbol{r} - j\omega t), \quad z < 0, \tag{6.24}$$

where $\boldsymbol{k}_i$ points in the incident ray direction. We can further align our coordinate system such that $\boldsymbol{k}_i$ is given by

$$\boldsymbol{k}_i = k_i \left(\sin\theta\hat{\boldsymbol{x}} + \cos\theta\hat{\boldsymbol{z}}\right), \tag{6.25}$$

with no y component in $\boldsymbol{k}_i$. θ is the incident angle, that is, the angle $\boldsymbol{k}_i$ makes with the normal vector of the interface, as shown in Fig. 6.6.

$\boldsymbol{E}_i$ by itself should satisfy the Maxwell's equations in medium 1, such that

- k_i, the magnitude of $\boldsymbol{k}_i$, should satisfy the dispersion relation

$$k_i = \frac{\omega n_1}{c}. \tag{6.26}$$

- $\tilde{\boldsymbol{E}}_i$ should be perpendicular to $\boldsymbol{k}_i$.
- The magnetic-field vector amplitude is perpendicular to both $\tilde{\boldsymbol{E}}_i$ and $\boldsymbol{k}_i$ (again consult Fig. 3.3) and related to the electric-field magnitude by the impedance of medium 1

$$Z_1 = \frac{Z_0}{n_1}. \tag{6.27}$$

When medium 2 is taken into account, however, one must include additional waves in both media to satisfy the boundary conditions in Sec. 6.2. I will not go through the full derivation here, which can again be found in, e.g., Griffiths [3], but I will just tell you what the final solution looks like. It involves a **reflected** plane wave in medium 1 and a **refracted** (also called **transmitted**) plane wave in medium 2.

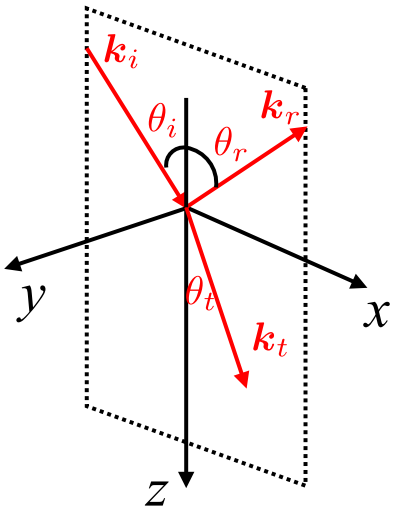


FIGURE 6.6. Geometry of the incident wavevector $\boldsymbol{k}_i$, reflected wavevector $\boldsymbol{k}_r$, and refracted (or transmitted) wavevector $\boldsymbol{k}_t$.

The reflected wave is given by

$$\boldsymbol{E}_r = \tilde{\boldsymbol{E}}_r \exp(j\boldsymbol{k}_r \cdot \boldsymbol{r} - j\omega t). \tag{6.28}$$

It is a wave in medium 1 but with a different wavevector $\boldsymbol{k}_r$ and different amplitude $\tilde{\boldsymbol{E}}_r$. By itself, it should also satisfy the Maxwell's equations in medium 1, including the dispersion relation

$$k_r = \frac{\omega n_1}{c}, \tag{6.29}$$

the transverse wave property, and the ratio between electric and magnetic amplitudes. **The frequency ω is the same as the incident value**. The total field in medium 1 is the addition of $\boldsymbol{E}_i$ and $\boldsymbol{E}_r$:

$$\boldsymbol{E} = \tilde{\boldsymbol{E}}_i \exp(j\boldsymbol{k}_i \cdot \boldsymbol{r} - j\omega t) + \tilde{\boldsymbol{E}}_r \exp(j\boldsymbol{k}_r \cdot \boldsymbol{r} - j\omega t) \quad \text{in medium 1 } (z < 0). \tag{6.30}$$

Linearity is at work here: the incident wave and the reflected wave each satisfies the Maxwell's equations in medium 1, although we need a superposition of them to satisfy the boundary conditions.

- **Exercise**: What is the time-averaged Poynting vector of this solution in medium 1?

The tranmitted wave is given by

$$\boldsymbol{E}_t = \tilde{\boldsymbol{E}}_t \exp(j\boldsymbol{k}_t \cdot \boldsymbol{r} - j\omega t) \quad \text{in medium 2 } (z > 0). \tag{6.31}$$

Now this must satisfy the Maxwell's equations in medium 2. In particular,

$$k_t = \frac{\omega n_2}{c}. \tag{6.32}$$

Again, **it should have the same frequency ω as the incident value**; all three waves have the same frequency because the waves all meet at the interface and the boundary conditions must be satisfied everywhere along the interface and at all times.

To see this, suppose for simplicity that the electric fields all point in the $\hat{\boldsymbol{y}}$ direction, which is parallel to the interface. The electric fields then have to be the same on both sides, leading to

$$\tilde{E}_i e^{j\boldsymbol{k}_i \cdot \boldsymbol{r} - j\omega t} + \tilde{E}_r e^{j\boldsymbol{k}_r \cdot \boldsymbol{r} - j\omega t} = \tilde{E}_t e^{j\boldsymbol{k}_t \cdot \boldsymbol{r} - j\omega t} \text{ at } z = 0. \tag{6.33}$$

If one of the waves varies in time in a different way, say, at a different frequency, the boundary condition won't be able to hold at all times and at all points on the interface.

We shall now go through the other relations among Eq. (6.30) and (6.31) one by one.

6.4. Snell's law

Another consequence of Eq. (6.33) is that the three plane waves must also oscillate in space along the interface in the same way:

$$\boldsymbol{k}_i \cdot \boldsymbol{r} = \boldsymbol{k}_r \cdot \boldsymbol{r} = \boldsymbol{k}_t \cdot \boldsymbol{r} \text{ at } z = 0, \tag{6.34}$$

which means that $\boldsymbol{k} \cdot (x\hat{\boldsymbol{x}} + y\hat{\boldsymbol{y}})$ must be the same for all at all x and y. In other words, the wavevectors parallel to the interface must be equal. Let the x component of $\boldsymbol{k}_i$ be k_{ix}, y component of $\boldsymbol{k}_i$ be k_{iy}, etc. Then

$$k_{ix} = k_{rx} = k_{tx}, \qquad k_{iy} = k_{ry} = k_{ty}. \tag{6.35}$$

These relations among the wavevector components parallel to the interface turn out to hold for any polarization.

Since we have already assumed the y component of $\boldsymbol{k}_i$ to be zero in Eq. (6.25), all wavevectors end up with no y component and pointing just in the x and z direction, as shown in Fig. 6.6. We can now assume

$$\boldsymbol{k}_i = \frac{\omega n_1}{c} \left(\sin\theta_i \hat{\boldsymbol{x}} + \cos\theta_i \hat{\boldsymbol{z}} \right), \tag{6.36}$$

which is the same as Eqs. (6.25) and (6.26) as before,

$$\boldsymbol{k}_r = \frac{\omega n_1}{c} \left(\sin\theta_r \hat{\boldsymbol{x}} - \cos\theta_r \hat{\boldsymbol{z}} \right), \tag{6.37}$$

where θ_r is the reflected angle, $k_r = \omega n_1/c$ since it is a wave in medium 1, and the reflected wave is assumed to propagate away from the interface, hence the minus sign in front of $\cos\theta_r$, and

$$\boldsymbol{k}_t = \frac{\omega n_2}{c} \left(\sin\theta_t \hat{\boldsymbol{x}} + \cos\theta_t \hat{\boldsymbol{z}} \right). \tag{6.38}$$

These vectors are plotted in Fig. 6.7 in $\boldsymbol{k}$ space for $n_2 > n_1$.

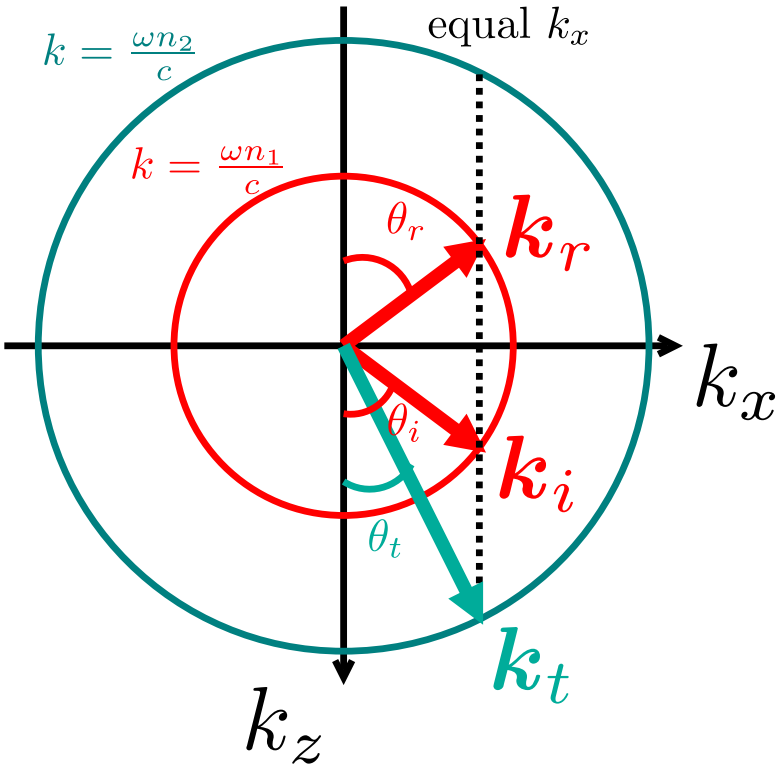


FIGURE 6.7. $\boldsymbol{k}$-space diagram for $n_2 > n_1$. The incident and reflected wavevectors $\boldsymbol{k}_i$ and $\boldsymbol{k}_r$ are in medium 1, so they must be on the circle $k = \sqrt{k_x^2 + k_y^2} = \omega n_1/c$ in $\boldsymbol{k}$ space. Similarly, assuming $n_2 > n_1$, the transmitted wavevector $\boldsymbol{k}_t$ must be on the circle $k = \omega n_2/c$. Equal k_x components mean that the angles are related. In particular, the reflected angle θ_r must be equal to the incident angle θ_i, and the refracted angle θ_t obeys Snell's law.

Equal x components of the wavevectors in Eq. (6.35) thus mean that the three angles are related:

$$\frac{\omega n_1}{c}\sin\theta_i = \frac{\omega n_1}{c}\sin\theta_r = \frac{\omega n_2}{c}\sin\theta_t, \tag{6.39}$$

$$n_1\sin\theta_i = n_1\sin\theta_r = n_2\sin\theta_t. \tag{6.40}$$

This is **Snell's law**, which states that the reflected angle is equal to the given incident angle:

$$\boxed{\theta_r = \theta_i,} \tag{6.41}$$

and the refracted angle depends also on the refractive indices

$$\boxed{\frac{\sin\theta_t}{\sin\theta_i} = \frac{n_1}{n_2}.} \tag{6.42}$$

For example, if the wave is going from a lower-index medium into a higher-index medium, $n_1 < n_2$, and θ_t is smaller than θ_i, or vice versa. A very nice demonstration of Snell's law can be found at `http://phet.colorado.edu` (PhET/sims/bending-light).

6.5. Total internal reflection and evanescent waves

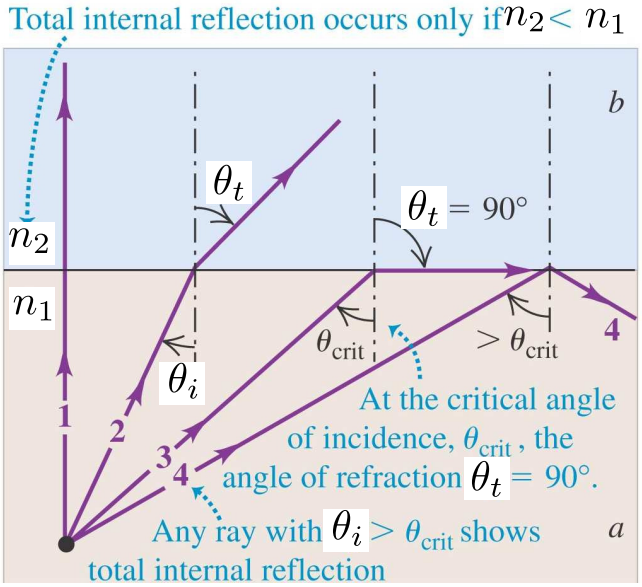


FIGURE 6.8. Total internal reflection (from Ref. [1]).

If $n_1 > n_2$ instead, as shown in Fig. 6.8, $\sin\theta_t/\sin\theta_i = n_1/n_2 > 1$, the refracted angle θ_t is larger than the incident θ_i, and for large enough θ_i we can have a situation where

$$\sin\theta_t = \frac{n_1}{n_2}\sin\theta_i > 1. \tag{6.43}$$

When this happens, there is no real solution for θ_t!

To see what went wrong here, recall that the x component of the incident wavevector is

$$k_{ix} = \frac{\omega n_1}{c}\sin\theta_i. \tag{6.44}$$

When $\frac{n_1}{n_2}\sin\theta_i > 1$, $n_1\sin\theta_i > n_2$, and

$$k_{ix} = \frac{\omega n_1}{c}\sin\theta_i > \frac{\omega n_2}{c}, \tag{6.45}$$

which exceeds $k_t = \omega n_2/c$, the wavenumber of the lower-index medium 2, as shown in Fig. 6.9.

Let $\boldsymbol{k}_t = k_{tx}\hat{\boldsymbol{x}} + k_{tz}\hat{\boldsymbol{z}}$. If both components are real, they cannot exceed the total length of $\boldsymbol{k}_t$, since

$$k_t^2 = k_{tx}^2 + k_{tz}^2 = \left(\frac{\omega n_2}{c}\right)^2, \tag{6.46}$$

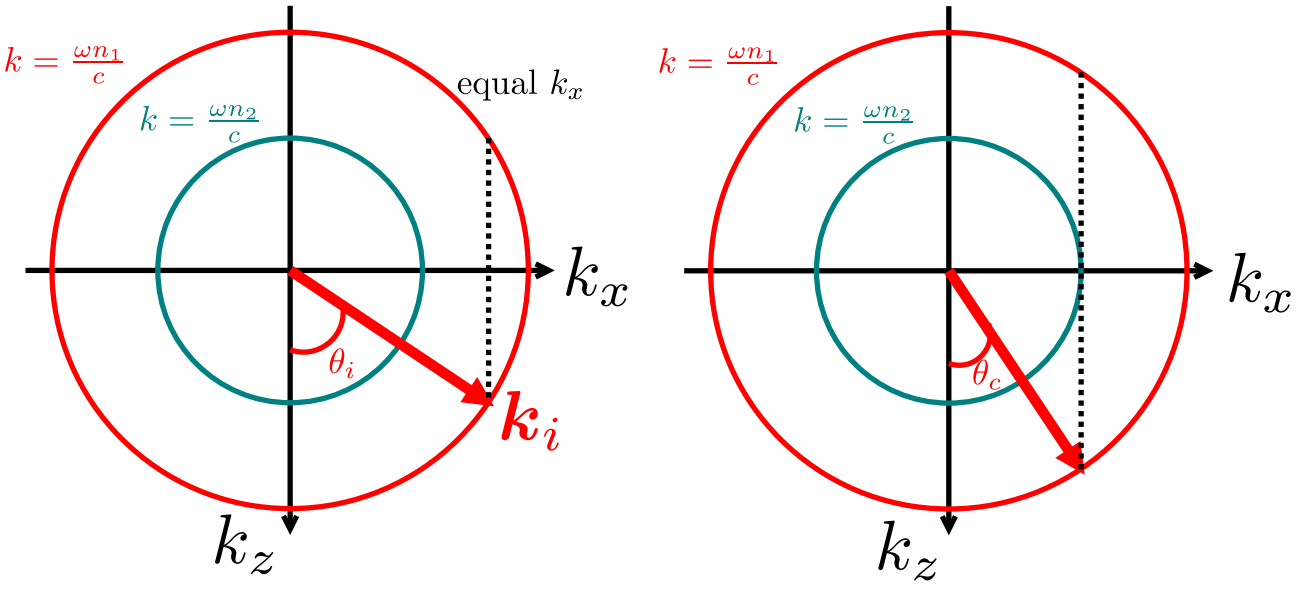


FIGURE 6.9. $\boldsymbol{k}$-space diagram of $\boldsymbol{k}_i$ and $\boldsymbol{k}_t$. Left: since $n_2 < n_1$, no real solution exists when k_{xi} is larger than the radius $\omega n_2/c$ of the smaller circle for medium 2. Right: the critical angle θ_c is defined as the largest angle at which there is still a real solution for k_{tz}. When $\theta_i > \theta_c$, no real solution for k_{tz} exists and total internal reflection occurs.

as the smaller circle in Fig. 6.9 shows. When k_{tx}, which must match k_{ix}, exceeds $\omega n_2/c$, it means that

$$k_{tz}^2 = k_t^2 - k_{tx}^2 < 0, \tag{6.47}$$

and we have no choice but to make k_{tz}, the z component of the transmitted wavevector $\boldsymbol{k}_t$, **imaginary**:

$$k_{tz} = \sqrt{k_t^2 - k_{tx}^2} = j\sqrt{k_{tx}^2 - k_t^2}, \tag{6.48}$$

Recall that the wave depends on k_{tz} via $\exp(jk_{tz}z)$, and if k_{tz} is imaginary, the wave must decay **exponentially** away from z as

$$\exp(jk_{tz}z) = \exp(-\sqrt{k_{tx}^2 - k_t^2}z). \tag{6.49}$$

This is called an **evanescent wave**;[1] it contains energy that sits near the interface but does not radiate power away from the interface (on average in time). Fig. 6.10 plots the exponential decay function to remind you what it looks like.

When the transmitted wave is evanescent, it can be shown that no power is transmitted in the $z \to \infty$ direction on average, and any incident power must be reflected completely. This is known as **total internal reflection**. This happens when the incident angle θ_i is larger than a **critical angle** θ_c defined by

$$\boxed{n_1 \sin\theta_c = n_2.} \tag{6.50}$$

such that,

$$\boxed{\text{when } \theta_i > \theta_c, \quad \sin\theta_t = \frac{n_1}{n_2}\sin\theta_i > 1,} \tag{6.51}$$

as shown in Fig. 6.9, and the transmitted wave must be evanescent.

[1] *You may have noted that $k_{tz} = -j\sqrt{k_{tx}^2 - k_t^2}$ can also satisfy the dispersion relation, making the wave **grow** exponentially away from the interface. While nothing seems to rule this possibility out in our analysis, it is important to note that, by assuming $\exp(-j\omega t)$, we are performing a **steady-state** analysis where the fields are assumed to oscillate from the beginning of the universe to the end of time $t \in (-\infty, \infty)$. In reality, the experiment must start from some **initial condition** (when the lasers are turned on, for example), and it is the decaying wave that is the physical steady-state limit, while the growing wave solution contains unphysical infinite energy and cannot be reached by any reasonable initial condition. Similar considerations are actually involved in choosing the sign of k_{zt} in general.

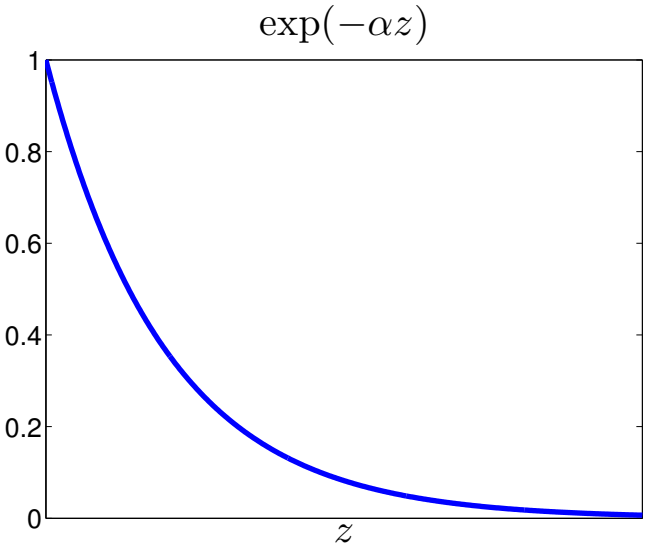


FIGURE 6.10. Plot of an exponential decay function.

It is not difficult to demonstrate the phenomenon of refraction and total internal reflection using a laser pointer and a water bowl, as shown in Fig. 6.11. The key is to note that we can approximate each laser beam by a plane wave with a wavevector along the direction of the beam.

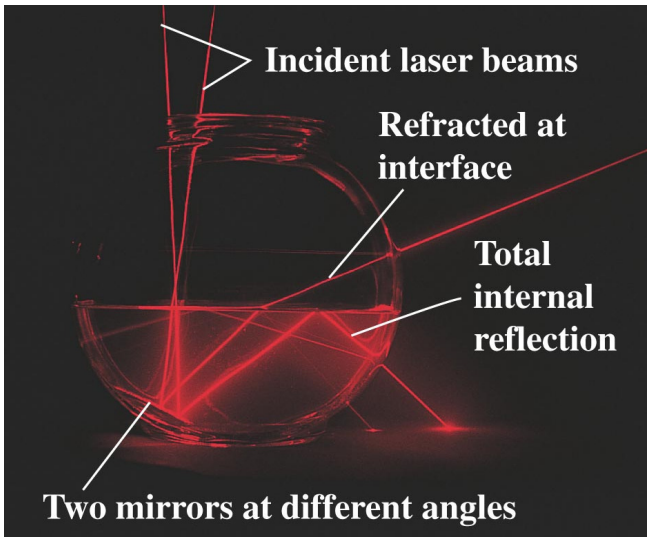


FIGURE 6.11. Reflections in a water bowl. Each laser beam can be well approximated by a plane wave with propagation direction coinciding with the plane-wave wavevectors. (from PC2232 AY13/14).

6.6. Application: Optical fibers

An **optical fiber** consists of a tube of dielectric **core** surrounded by a cladding with lower refractive index, as shown in Fig. 6.12. Light can be trapped inside the core due to total internal reflection and propagate along the fiber.[2] The use of total internal reflection for optical confinement, rather than metal mirrors, is important for reducing losses because the former allows an optical fiber to be made only of low-loss dielectrics, whereas metals are much more lossy.

[2]*Note that the interface between the core and the cladding in a fiber is cylindrical, and plane-wave solutions are not accurate if the fiber core is very thin (relative to the wavelength). The concept of total internal reflection still applies to **cylindrical waves** however.

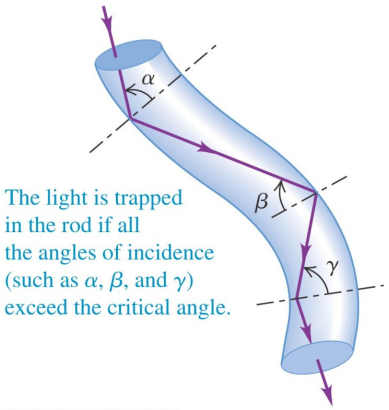


FIGURE 6.12. An optical fiber consists of a tube of dielectric core surrounded by a cladding with slightly lower refractive index. Light can be trapped inside by total internal reflection and propagate along the fiber. From [1].

Optical fibers play an essential role in **optical communications**. Optical fibers made of high-purity silica have very low loss across a very large bandwidth. Optical signals can be carried over transoceanic distances without suffering from much loss.

6.7. TE and TM polarizations

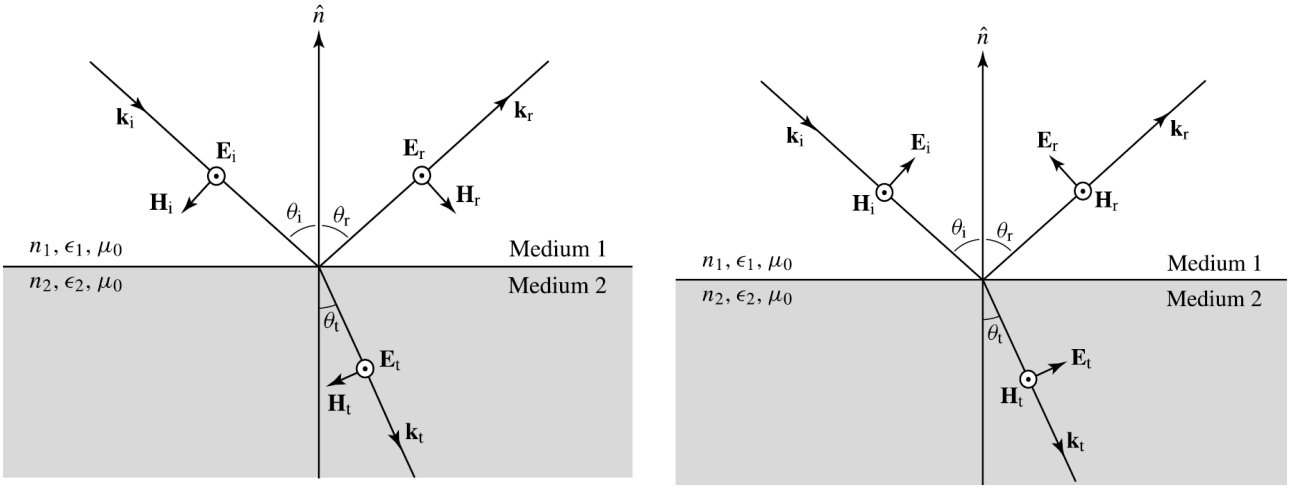


FIGURE 6.13. Two cases of reflection and refraction for the TE polarization (left) and TM polarization (right). From [3].

Given an incident wavevector $\boldsymbol{k}_i$, two situations occur depending on the vectoral direction of the incident field $\tilde{\boldsymbol{E}}_i$, as shown in Fig. 6.13. In the left figure, where $\tilde{\boldsymbol{E}}_i$ is parallel to the interface, the incident wave is called **TE (transverse-electric) polarized**, or s polarized, in which case the electric fields of the incident and refracted waves are all parallel to the interface due to boundary conditions.

In the right figure of Fig. 6.13, where the magnetic field of the incident wave is parallel to the interface, the incident wave is called **TM (transverse-magnetic) polarized**, or p polarized. The magnetic fields of the incident and refracted waves are all parallel to the interface due to boundary conditions. Note that both polarizations satisfy the transverse wave property for all three waves, while any other polarization for the same $\boldsymbol{k}_i$ can be written as a superposition of the TE and TM polarizations.

6.8. Reflection and transmission coefficients

The reflected wave amplitude and the refracted wave amplitude can be computed by **matching the boundary conditions at** $z = 0$. For TE or s polarization, we suppose

$$\tilde{\boldsymbol{E}}_i = \hat{\boldsymbol{y}}\tilde{E}_i, \qquad \tilde{\boldsymbol{E}}_r = \hat{\boldsymbol{y}}\tilde{E}_r, \qquad \tilde{\boldsymbol{E}}_t = \hat{\boldsymbol{y}}\tilde{E}_t. \tag{6.52}$$

Define the reflection coefficient r_s and transmission coefficient t_s by

$$\boxed{\tilde{E}_r = r_s\tilde{E}_i,} \qquad \boxed{\tilde{E}_t = t_s\tilde{E}_i} \tag{6.53}$$

for the TE polarization.[3] After some algebra (see Griffiths [3]), the reflection and transmission coefficients are

$$r_s \equiv \frac{\tilde{E}_r}{\tilde{E}_i} = \frac{n_1\cos\theta_i - n_2\cos\theta_t}{n_1\cos\theta_i + n_2\cos\theta_t}, \qquad t_s \equiv \frac{\tilde{E}_t}{\tilde{E}_i} = \frac{2n_1\cos\theta_i}{n_1\cos\theta_i + n_2\cos\theta_t}. \tag{6.54}$$

Note that θ_t depends on θ_i through Snell's law. The important point here is that $\tilde{E}_r$ and $\tilde{E}_t$ are **linear** with respect to the input $\tilde{E}_i$, and the proportionality coefficients depend on the polarization, the refractive indices, and the angles.

When total internal reflection occurs,

$$\cos\theta_t = \sqrt{1-\sin^2\theta_t} = \sqrt{1-\frac{n_1^2}{n_2^2}\sin^2\theta_i} = j\sqrt{\frac{n_1^2}{n_2^2}\sin^2\theta_i - 1}, \tag{6.55}$$

is imaginary and the sign is chosen to make the evanescent wave decay exponentially away from z as

$$\exp(jk_{tz}z) = \exp\left(jk_t\cos\theta_t z\right). \tag{6.56}$$

For TM polarization or p polarization, the amplitudes are defined according to the unit vectors in the right figure of Fig. 6.13, namely,

$$\tilde{E}_i = \hat{\boldsymbol{e}}_i\tilde{E}_i, \tag{6.57}$$

where $\hat{\boldsymbol{e}}_i$ is the unit vector for $\tilde{\boldsymbol{E}}_i$ in the right figure of 6.13. Similarly,

$$\tilde{E}_r = \hat{\boldsymbol{e}}_r\tilde{E}_r, \qquad \tilde{E}_t = \hat{\boldsymbol{e}}_t\tilde{E}_t. \tag{6.58}$$

These unit vectors are fixed by the incident angle because we have assumed that $\boldsymbol{H}$ is parallel to the interface for TM and the electric fields must be perpendicular to both $\boldsymbol{H}$ and the wavevectors.

The reflected and transmitted amplitudes are still proportional to the incident value, although the coefficients are different:

$$\boxed{\tilde{E}_r = r_p\tilde{E}_i,} \qquad \boxed{\tilde{E}_t = t_p\tilde{E}_i.} \tag{6.59}$$

The result is

$$r_p \equiv \frac{\tilde{E}_r}{\tilde{E}_i} = \frac{n_2\cos\theta_i - n_1\cos\theta_t}{n_2\cos\theta_i + n_1\cos\theta_t}, \qquad t_p \equiv \frac{\tilde{E}_t}{\tilde{E}_i} = \frac{2n_1\cos\theta_i}{n_2\cos\theta_i + n_1\cos\theta_t}. \tag{6.60}$$

- **Exercise**: Compare the TE and TM cases for $\theta_i = 0$ (normal incidence). Convince yourself that they correspond to the same scenario and the results are consistent.
- **Exercise**: What are the limits of Snell's law and the reflection/transmission coefficients as $n_2/n_1 \to \infty$? How about $n_1/n_2 \to \infty$?
- **Exercise**: Brewster angle: For TM polarization, find the incident angle θ_i such that the reflection is zero.
- **Exercise**: What is $|r_s|^2$ and $|r_p|^2$ when total internal reflection occurs? What does this say about power conservation and the power carried by the evanescent wave?

So far we have studied the problem for one specific incident wavevector. Due to the linearity property, multiple incident waves with different incident wavevectors can be studied by solving for the solutions for each individual wave before summing the solutions back together.

[3] Note that here r_s and t_s are not related to position vector $\boldsymbol{r}$ and time t, I am just recycling the symbols here.

CHAPTER 7

Input-Output Analysis

In a lot of optics problems, we are interested in how electromagnetic waves coming out of a device as **outputs** depend on the waves going in as **inputs**. Reflection and refraction is an example: given the incident wave as input, we have found how the reflected wave and the transmitted wave behaves as two outputs from the interface. We would like to generalize that situation and consider the relations between multiple input waves and multiple output waves in optical devices.

7.1. Delay line

Let us begin our discussion by considering the simplest device: propagation in a medium with refractive index n as a delay line. We will simply consider propagation along the z axis and linear polarization $\hat{\boldsymbol{x}}$, but allow waves to propagate in opposite directions, as shown in Fig. 7.1.

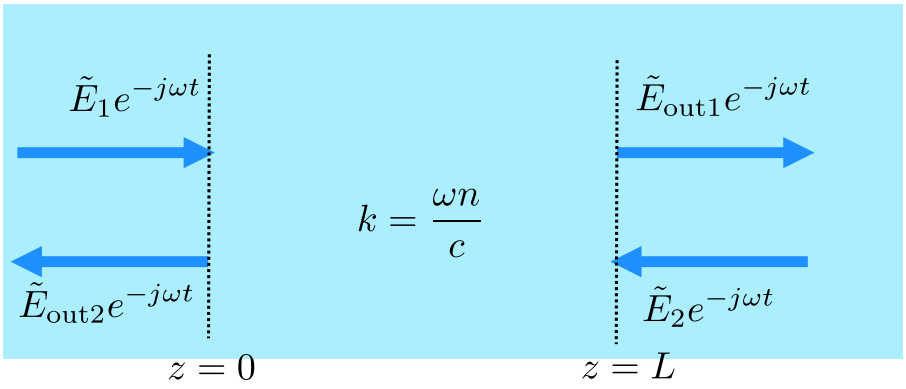


FIGURE 7.1. Input-output analysis of monochromatic plane waves in a delay line.

First consider the wave propagating to the right:

$$\boldsymbol{E}_R = \hat{\boldsymbol{x}} \tilde{E}_R \exp(jkz - j\omega t), \tag{7.1}$$

where the wavenumber is as usual:

$$k = \frac{\omega n}{c}. \tag{7.2}$$

We will take the amplitude at $z = 0$ to be an **input amplitude** $\tilde{E}_1$, such that

$$\boldsymbol{E}_R(z = 0) = \hat{\boldsymbol{x}} \tilde{E}_1 e^{-j\omega t}, \tag{7.3}$$

and take the field at $z = L$ to be an **output amplitude** $\tilde{E}_{\text{out1}}$:

$$\boldsymbol{E}_R(z = L) = \hat{\boldsymbol{x}} \tilde{E}_{\text{out1}} e^{-j\omega t}. \tag{7.4}$$

Comparing Eq. (7.1) at $z = 0$ and (7.3), we find that the wave amplitude $\tilde{E}_R$ is equal to the input amplitude $\tilde{E}_1$. Comparing Eq. (7.1) at $z = L$ and (7.4), it follows that the output amplitude $\tilde{E}_{\text{out1}}$ is $\tilde{E}_R \exp(jkL)$, or

$$\tilde{E}_{\text{out1}} = \tilde{E}_1 \exp(jkL). \tag{7.5}$$

What this means is that the field acquires a **phase delay** after propagating a distance of L. The delay is proportional to L and the wavenumber $k = \omega n/c$.

For a wave propagating to the left, let us assume

$$\boldsymbol{E}_L = \hat{\boldsymbol{x}} \tilde{E}_L \exp(-jkz - j\omega t). \tag{7.6}$$

Now assume that the second input amplitude $\tilde{E}_2$ is at $z = L$:

$$\boldsymbol{E}_L(z = L) = \hat{\boldsymbol{x}}\tilde{E}_2 \exp(-j\omega t), \tag{7.7}$$

and the second output amplitude $\tilde{E}_{\text{out2}}$ is at $z = 0$:

$$\boldsymbol{E}_L(z = 0) = \hat{\boldsymbol{x}}\tilde{E}_{\text{out2}} \exp(-j\omega t). \tag{7.8}$$

We find that the input $\tilde{E}_2$ is now $\tilde{E}_L e^{-jkL}$, which has an additional phase factor with respect to the wave amplitude $\tilde{E}_L$, while the output $\tilde{E}_{\text{out2}}$ is $\tilde{E}_L$. In other words,

$$\tilde{E}_{\text{out2}} = \tilde{E}_L = \tilde{E}_2 \exp(jkL), \tag{7.9}$$

that is, the wave would also acquire a phase delay $\exp(jkL)$ with respect to the input when propagating to the left, which is hardly surprising.

7.2. Scattering matrix

In general, we can have any superposition of waves propagating to the right ($\boldsymbol{E}_R$) and to the left ($\boldsymbol{E}_L$). Although they overlap in space in our simple theory, we can assume that they model counter-propagating laser beams that propagating at a very small angle with respect to the z axis, in which case the two beams will eventually separate for a long distance and we can impose the inputs independently and also measure the outputs independently, as depicted in Fig. 7.2.

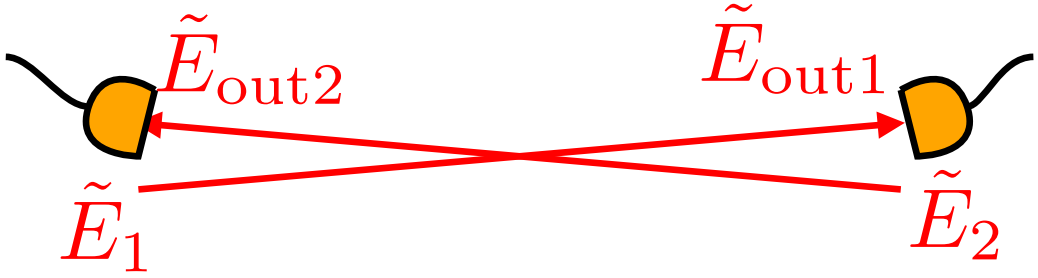


FIGURE 7.2. Two counter-propagating optical beams. We can use the simple delay line model to model the measured output amplitudes as a function of the input amplitudes; if the angle between the beams is very small, propagation on one axis can be assumed.

When we have two arbitrary input waves with amplitudes $\tilde{E}_1$ and $\tilde{E}_2$, we have a **superposition** of waves in opposite directions in a delay line, and the output amplitudes to be measured will be

$$\begin{pmatrix} \tilde{E}_{\text{out1}} \\ \tilde{E}_{\text{out2}} \end{pmatrix} = \begin{pmatrix} e^{jkL} & 0 \\ 0 & e^{jkL} \end{pmatrix} \begin{pmatrix} \tilde{E}_1 \\ \tilde{E}_2 \end{pmatrix}, \tag{7.10}$$

in **matrix form**.

In general, we call the matrix that relates an input vector to an output vector a **scattering matrix**. Explicitly, a scattering matrix for two inputs and two outputs is written as

$$\boldsymbol{s} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}, \tag{7.11}$$

the **input and output vectors** are written as

$$\tilde{\boldsymbol{E}} = \begin{pmatrix} \tilde{E}_1 \\ \tilde{E}_2 \end{pmatrix}, \qquad \tilde{\boldsymbol{E}}_{\text{out}} = \begin{pmatrix} \tilde{E}_{\text{out1}} \\ \tilde{E}_{\text{out2}} \end{pmatrix}, \tag{7.12}$$

such that the multiplication of the input vector by the scattering matrix gives the output vector:

$$\boxed{\tilde{\boldsymbol{E}}_{\text{out}} = \begin{pmatrix} \tilde{E}_{\text{out1}} \\ \tilde{E}_{\text{out2}} \end{pmatrix} = \boldsymbol{s}\tilde{\boldsymbol{E}} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \begin{pmatrix} \tilde{E}_1 \\ \tilde{E}_2 \end{pmatrix}.} \tag{7.13}$$

Explicitly, each element of the output vector is given by

$$\tilde{E}_{\text{out}\,a} = \sum_b s_{ab}\tilde{E}_b, \quad a = 1, 2, b = 1, 2, \tag{7.14}$$

which is the definition of matrix multiplication.

For a one-dimensional delay line, the scattering matrix is given by the simple form in Eq. (7.10). We shall now study more complicated optical devices that lead to coupling between waves.

7.3. Partial mirror

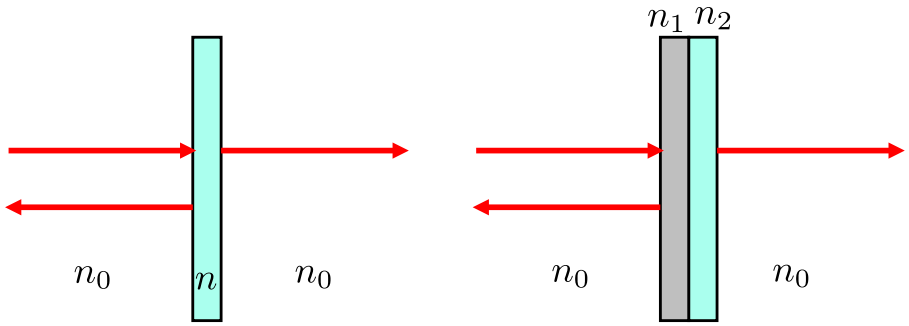


FIGURE 7.3. Two examples of partial mirrors. Normal incidence is assumed.

A partial mirror consists of a stack of dielectrics; some examples are shown in Fig. 7.3. The simplest case is simply a dielectric slab, but usually partial mirrors are made of a very thin layer of silver coated on glass. For simplicity, let us consider normal incidence of plane waves.

If there is an incident wave propagating to the right, in general there will be a **reflected** wave and a **transmitted** wave. Assume that the mirror material is from $z = 0$ to $z = L$. Similar to Sec. 6.3, assume that the incident wave is

$$\boldsymbol{E}_i = \hat{\boldsymbol{x}}\tilde{E}_i \exp(jk_0 z - j\omega t), \quad z < 0, \tag{7.15}$$

the reflected wave is

$$\boldsymbol{E}_r = \hat{\boldsymbol{x}}\tilde{E}_r \exp(-jk_0 z - j\omega t), \quad z < 0, \tag{7.16}$$

and the transmitted wave is

$$\boldsymbol{E}_t = \hat{\boldsymbol{x}}\tilde{E}_t \exp(jk_0 z - j\omega t), \quad z > L. \tag{7.17}$$

where $k_0 = \omega n_0/c$ is the wavenumber of the surrounding medium. Define an input amplitude $\tilde{E}_1$ by

$$\boldsymbol{E}_i(z = 0) = \hat{\boldsymbol{x}}\tilde{E}_1 e^{-j\omega t}, \tag{7.18}$$

and output amplitudes $\tilde{E}_{\text{out1}}$ and $\tilde{E}_{\text{out2}}$ by

$$\boldsymbol{E}_r(z = 0) = \hat{\boldsymbol{x}}\tilde{E}_{\text{out1}} e^{-j\omega t}, \qquad \boldsymbol{E}_t(z = L) = \hat{\boldsymbol{x}}\tilde{E}_{\text{out2}} e^{-j\omega t}, \tag{7.19}$$

as shown in Fig. 7.4. The exact calculation of **reflection** and **transmission coefficients** involves matching all the boundary conditions at the different interfaces and is quite difficult for us at this stage. Instead, for now we will simply assume that they are given to us:

$$\tilde{E}_{\text{out1}} = s_{11}\tilde{E}_1, \qquad \tilde{E}_{\text{out2}} = s_{21}\tilde{E}_1, \tag{7.20}$$

where s_{11} is the reflection coefficient, and s_{21} is the transmission coefficient. Note that these coefficients are usually **complex**.

We now consider what happens if there is an incident wave from the right to the left instead, as shown in Fig. 7.5. Now let us define this second incident wave as

$$\tilde{\boldsymbol{E}}_{i2} = \hat{\boldsymbol{x}}\tilde{E}_{i2} \exp(-jk_0 z - j\omega t), \quad z > L, \tag{7.21}$$

and the input amplitude $\tilde{E}_2$ as

$$\tilde{\boldsymbol{E}}_{i2}(z = L) = \hat{\boldsymbol{x}}\tilde{E}_2 e^{-j\omega t}. \tag{7.22}$$

The key here is that the reflected wave is the same as the $\boldsymbol{E}_t$ we defined earlier, and the transmitted wave is the same as $\boldsymbol{E}_r$. The two output amplitudes are now related to $\tilde{E}_2$ in a different way:

$$\tilde{E}_{\text{out1}} = s_{12}\tilde{E}_2, \qquad \tilde{E}_{\text{out2}} = s_{22}\tilde{E}_2. \tag{7.23}$$

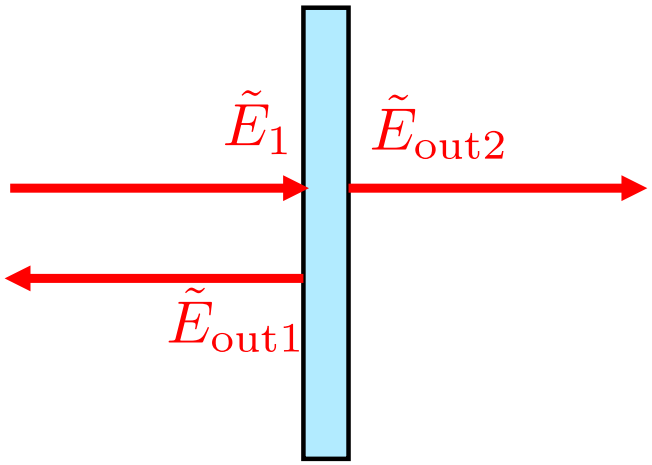


FIGURE 7.4. For one incident wave from the left to the right for a partial mirror, there are one input amplitude and two output amplitudes.

s_{12} is now the transmission coefficient for this second scenario, and s_{22} is the reflection coefficient for the other side.

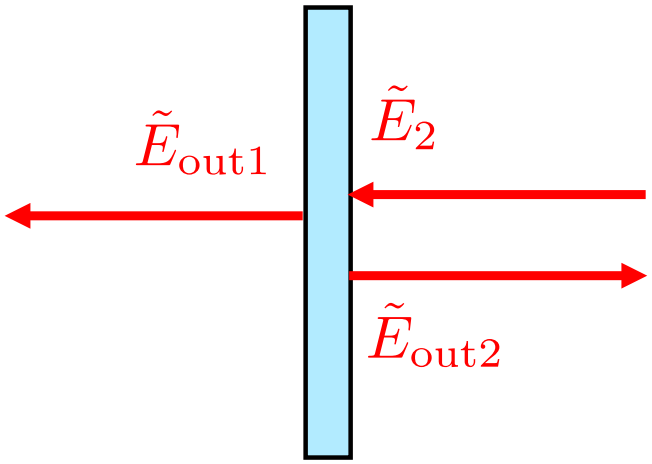


FIGURE 7.5. For one incident wave from the right to the left, the two output amplitudes for a partial mirror are the same as the ones for a left-incident input, although now their roles are reversed: $\tilde{E}_{\text{out1}}$ becomes the transmitted amplitude and $\tilde{E}_{\text{out2}}$ becomes the reflected amplitude.

*For a simple dielectric slab on the left of Fig. 7.3, we expect the two reflection coefficients s_{11} and s_{22} to be equal and the two transmission coefficients s_{21} and s_{12} to be equal, as the second scenario is just the inverted version of the first, but for the more complicated setup on the right of Fig. 7.3, there is no such simple relation, as the coefficients should be different depending on whether the incident wave meets the n_1 medium first or the n_2 medium first.

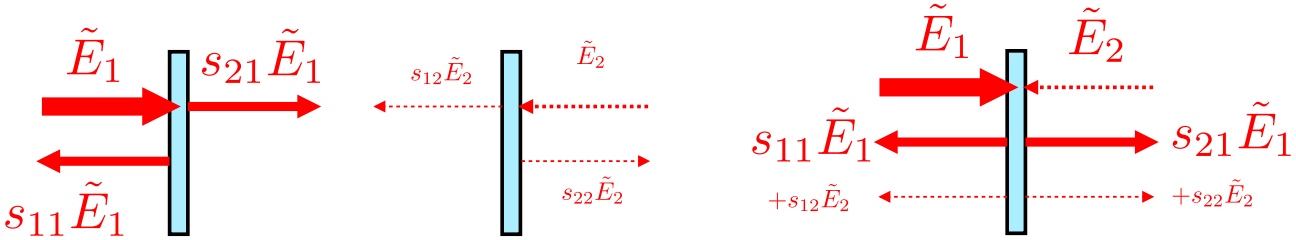


FIGURE 7.6. Superposition of the two single-input cases (left and center figures) leads to the input-output relation for two inputs and two outputs (right figure).

We now arrive at a crucial point in this section: What happens when there are incident waves from both sides of the partial mirror? Once again we appeal to **linearity**: the final solution is just the **superposition** of the two cases depicted by Figs. 7.4 and 7.5, as shown in Fig. 7.6. Mathematically, $\tilde{E}_{\text{out1}}$ and $\tilde{E}_{\text{out2}}$ should be the addition of Eq. (7.20) due to one input $\tilde{E}_1$ and Eq. (7.23) due to the second input $\tilde{E}_2$. The final result is

$$\underbrace{\tilde{E}_{\text{out1}}}_{\text{output to the left}} = \underbrace{s_{11}\tilde{E}_1}_{\text{reflection of first input}} + \underbrace{s_{12}\tilde{E}_2}_{\text{transmission of second input}} , \tag{7.24}$$

$$\underbrace{\tilde{E}_{\text{out2}}}_{\text{output to the right}} = \underbrace{s_{21}\tilde{E}_1}_{\text{transmission of first input}} + \underbrace{s_{22}\tilde{E}_2}_{\text{reflection of second input}} . \tag{7.25}$$

In matrix form, the relations also obey Eq. (7.13). Unlike the delay line described in Sec. 7.1, here all four elements of the scattering matrix may be nonzero. They are also **complex** in general.

- **Question**: How does a one-way mirror, shown in Fig. 7.7, work?
  Answer: A one-way mirror is nothing but a partial mirror, with one room much brighter than the other. Suppose that $\tilde{E}_1$ in room 1 has a very high magnitude compared with $\tilde{E}_2$ in room 2. Then the reflected waves in room 1 can become very strong compared with the transmitted wave in room 1, and one mostly see the reflections in the bright room. In the dark room, the transmitted component from the bright room is very strong relative to the reflected component, so one mostly sees the transmitted light from the bright room as well.

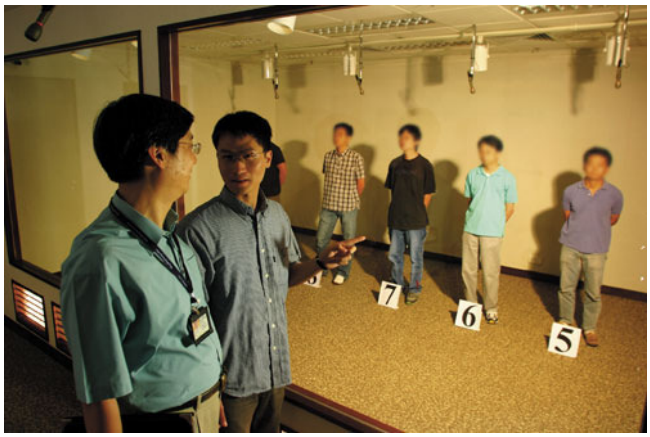


FIGURE 7.7. One-way mirror (Image from http://archive.news.softpedia.com/news/How-One-way-Mirrors-Work-85259.shtml)

7.4. Beamsplitter

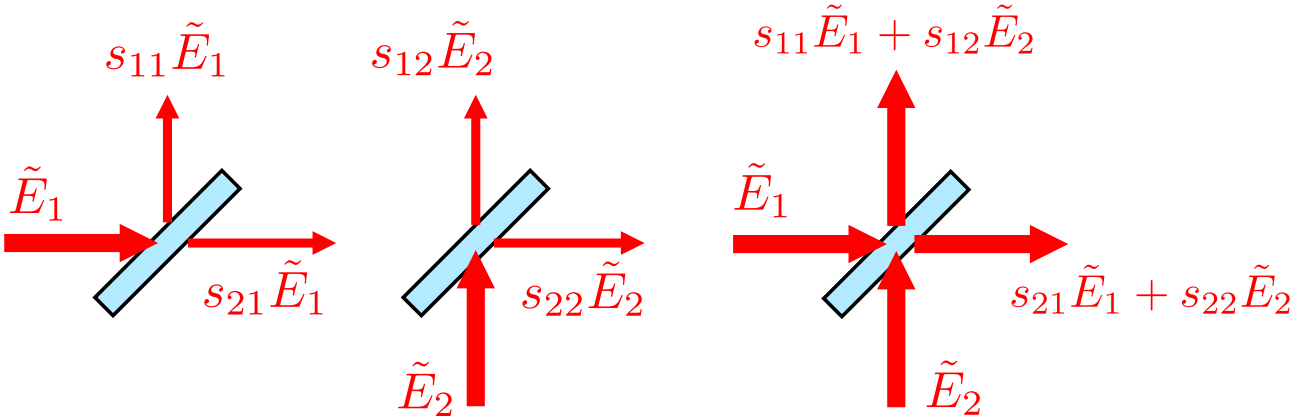


FIGURE 7.8. Superposition of the waves for the single-input scenarios (left and center figures) leads to the two-input two-output relations for a beamsplitter (right figure).

Let us now consider oblique incidence for a partial mirror. In this case the device is also called a beamsplitter. The physics is almost the same as the partial mirror case, and each output is the overlap of a transmitted component and a reflected component. The input-output relations are also given by Eq. (7.13), except that the scattering matrix depends on the incident angle in general. In practice, it is much easier to use a beamsplitter at oblique incidence to separate the input and output beams physically for separate measurements.

7.5. Power conservation

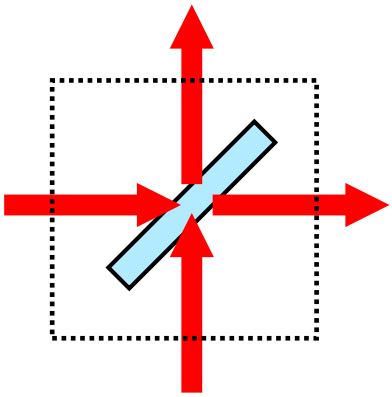


FIGURE 7.9. If the beamsplitter is lossless, the total optical power should be conserved, and the net power going in and out of the box should be zero.

Imagine a box with equal lengths around a lossless beamsplitter, as shown in Fig. 7.9. A lossless beamsplitter means that it does not absorb (or provide) any power, and the net optical power that goes in and out of the box must be zero by power conservation.

To simplify the problem, forget about the spatial overlap of the plane waves at the moment and imagine that they are laser beams that go in and out of the box. The average input power due to the first input is the intensity I times the area of a surface of the box A, or

$$\bar{P}_1 \approx I_1 A = \frac{|\tilde{E}_1|^2 A}{2Z_0}; \tag{7.26}$$

(see Sec. 4.3 for intensity of a plane wave, Z_0 is the impedance of the surrounding medium). Similarly, the input power due to the second input is $|\tilde{E}_2|^2 A/(2Z_0)$, the output power due to the first output is $|\tilde{E}_{\text{out1}}|^2 A/(2Z_0)$, and the output power due to the second output is $|\tilde{E}_{\text{out2}}|^2 A/(2Z_0)$. The waves are all propagating in the same medium, so the impedance is the same.

By power conservation, the input power must be equal to the output power for a lossless beamsplitter; this means that

$$\underbrace{\frac{|\tilde{E}_1|^2 A}{2Z_0}}_{\text{power from the left}} + \underbrace{\frac{|\tilde{E}_2|^2 A}{2Z_0}}_{\text{power from the bottom}} = \underbrace{\frac{|\tilde{E}_{\text{out1}}|^2 A}{2Z_0}}_{\text{power to the right}} + \underbrace{\frac{|\tilde{E}_{\text{out2}}|^2 A}{2Z_0}}_{\text{power to the top}}, \tag{7.27}$$

$$\boxed{|\tilde{E}_1|^2 + |\tilde{E}_2|^2 = |\tilde{E}_{\text{out1}}|^2 + |\tilde{E}_{\text{out2}}|^2.} \tag{7.28}$$

This **quadratic** relation between the input and output amplitudes in general holds for any lossless optical device when the input and output waves all propagate in media with the **same refractive index**.

7.6. Unitarity

In matrix form, the left-hand side of Eq. (7.28) can be written as

$$|\tilde{E}_1|^2 + |\tilde{E}_2|^2 = \begin{pmatrix} \tilde{E}_1^* & \tilde{E}_2^* \end{pmatrix} \begin{pmatrix} \tilde{E}_1 \\ \tilde{E}_2 \end{pmatrix}. \tag{7.29}$$

Similarly,

$$|\tilde{E}_{\text{out1}}|^2 + |\tilde{E}_{\text{out2}}|^2 = \begin{pmatrix} \tilde{E}^*_{\text{out1}} & \tilde{E}^*_{\text{out2}} \end{pmatrix} \begin{pmatrix} \tilde{E}_{\text{out1}} \\ \tilde{E}_{\text{out2}} \end{pmatrix}. \tag{7.30}$$

We shall define the **conjugate transpose**, denoted by the dagger †, of a vector or a matrix as the following:

$$\tilde{\boldsymbol{E}}^\dagger = \begin{pmatrix} \tilde{E}^*_1 & \tilde{E}^*_2 \end{pmatrix}, \tag{7.31}$$

which is the **transpose** of a matrix with **complex-conjugate elements**. Then Eq. (7.28) can be rewritten as

$$\tilde{\boldsymbol{E}}^\dagger \tilde{\boldsymbol{E}} = \tilde{\boldsymbol{E}}^\dagger_{\text{out}} \tilde{\boldsymbol{E}}_{\text{out}}. \tag{7.32}$$

From Eq. (7.13), we know that $\tilde{\boldsymbol{E}}_{\text{out}} = \boldsymbol{s}\tilde{\boldsymbol{E}}$. What about $\tilde{\boldsymbol{E}}^\dagger_{\text{out}}$?

$$\tilde{\boldsymbol{E}}^\dagger_{\text{out}} = \begin{pmatrix} \tilde{E}^*_{\text{out1}} & \tilde{E}^*_{\text{out2}} \end{pmatrix} = \begin{pmatrix} \tilde{E}^*_1 & \tilde{E}^*_2 \end{pmatrix} \begin{pmatrix} s^*_{11} & s^*_{21} \\ s^*_{12} & s^*_{22} \end{pmatrix} = \tilde{\boldsymbol{E}}^\dagger \boldsymbol{s}^\dagger. \tag{7.33}$$

This property is inherited from the matrix transpose, since $(AB)^\top = B^\top A^\top$. Eq. (7.32) becomes

$$\tilde{\boldsymbol{E}}^\dagger \tilde{\boldsymbol{E}} = \tilde{\boldsymbol{E}}^\dagger \boldsymbol{s}^\dagger \boldsymbol{s} \tilde{\boldsymbol{E}}. \tag{7.34}$$

Writing $\tilde{\boldsymbol{E}}^\dagger \tilde{\boldsymbol{E}} = \tilde{\boldsymbol{E}}^\dagger \boldsymbol{I} \tilde{\boldsymbol{E}}$, where $\boldsymbol{I}$ is the **identity matrix**:

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{7.35}$$

and grouping terms in Eq. (7.34) on one side,

$$\tilde{\boldsymbol{E}}^\dagger \left(\boldsymbol{s}^\dagger \boldsymbol{s} - \boldsymbol{I} \right) \tilde{\boldsymbol{E}} = \boldsymbol{0}. \tag{7.36}$$

There are two solutions: either $\tilde{\boldsymbol{E}} = 0$, which says that the inputs must be zero and is not very interesting, or

$$\boxed{\boldsymbol{s}^\dagger \boldsymbol{s} = \boldsymbol{I}.} \tag{7.37}$$

Explicitly,

$$\begin{pmatrix} s^*_{11} & s^*_{21} \\ s^*_{12} & s^*_{22} \end{pmatrix} \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{7.38}$$

Eq. (7.37) specifies a relation among the scattering matrix elements that must be obeyed by any lossless (power-conserving) optical device. Mathematically, Eq. (7.37) is the definition of a **unitary matrix**; it says that the inverse of a unitary matrix is equal to the complex transpose.

7.7. 50/50 beamsplitter

Let us study a more specific example called 50/50 beamsplitters. If there is just one input, a 50/50 beamsplitter will split the input power equally between the two outputs. To make the math simpler, I will assume the following scattering matrix for the beamsplitter:

$$\begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}. \tag{7.39}$$

Note that the matrix elements can be complex in general, but here I assume that they are all real for simplicity.

- **Exercise**: check that Eq. (7.39) is a unitary matrix. If the minus sign is not there for s_{22}, is the matrix still unitary?

To confirm that this represents a scattering matrix for a 50/50 beamsplitter, let us check the output powers when there is just one input and $\tilde{E}_2 = 0$, as shown in Fig. 7.10:

$$\tilde{E}_{\text{out1}} = s_{11}\tilde{E}_1 = \frac{1}{\sqrt{2}}\tilde{E}_1, \qquad \left|\tilde{E}_{\text{out1}}\right|^2 = \frac{1}{2}\left|\tilde{E}_1\right|^2, \tag{7.40}$$

$$\tilde{E}_{\text{out2}} = s_{12}\tilde{E}_2 = \frac{1}{\sqrt{2}}\tilde{E}_2, \qquad \left|\tilde{E}_{\text{out2}}\right|^2 = \frac{1}{2}\left|\tilde{E}_2\right|^2. \tag{7.41}$$

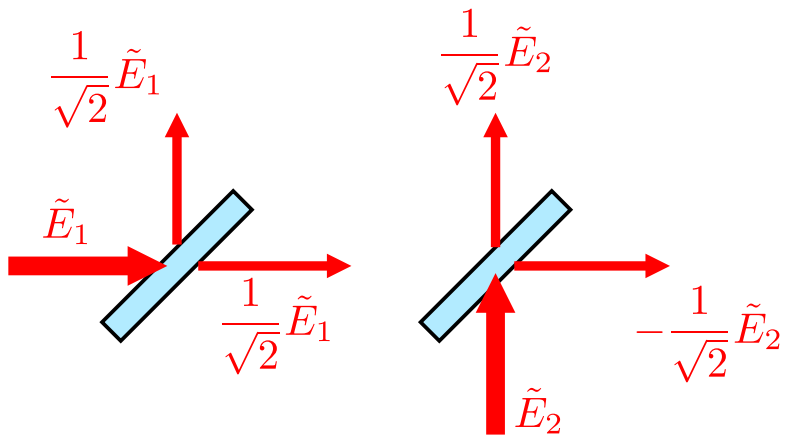


FIGURE 7.10. A 50/50 beamsplitter with one input.

Similarly, if $\tilde{E}_1 = 0$, there is just one input into the second port,

$$\tilde{E}_{\text{out1}} = s_{12}\tilde{E}_2 = \frac{1}{\sqrt{2}}\tilde{E}_2, \qquad \left|\tilde{E}_{\text{out1}}\right|^2 = \frac{1}{2}\left|\tilde{E}_2\right|^2, \tag{7.42}$$

$$\tilde{E}_{\text{out2}} = s_{22}\tilde{E}_2 = -\frac{1}{\sqrt{2}}\tilde{E}_2, \qquad \left|\tilde{E}_{\text{out2}}\right|^2 = \frac{1}{2}\left|\tilde{E}_2\right|^2, \tag{7.43}$$

and indeed the input power is equally split, if there is just one input.

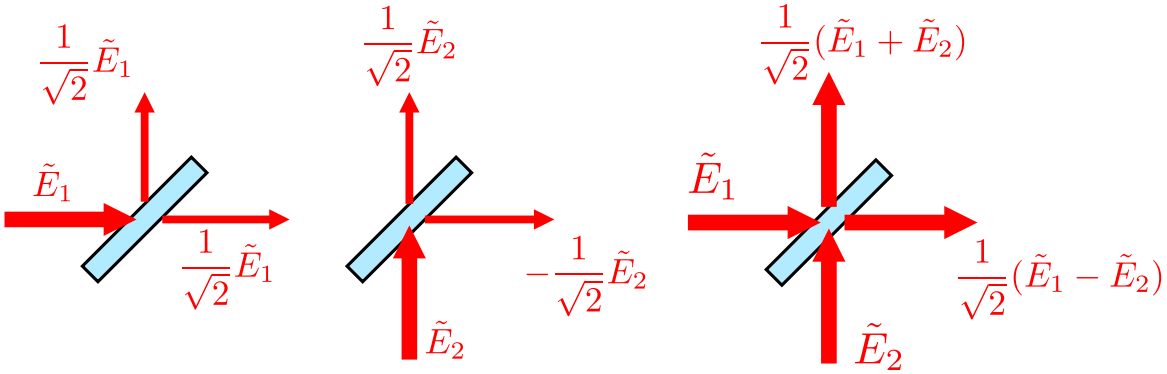


FIGURE 7.11. The relations for two inputs and two outputs of the 50/50 beam splitter (right), as a superposition of the solutions for the one-input cases (left and center).

With two inputs, we can add the two field solutions, as shown in Fig. 7.11. For example, if $\tilde{E}_2 = \tilde{E}_1$, the inputs have the same magnitude and phase. the outputs are

$$\tilde{E}_{\text{out1}} = \frac{1}{\sqrt{2}}\left(\tilde{E}_1 + \tilde{E}_1\right) = \sqrt{2}\tilde{E}_1, \qquad \tilde{E}_{\text{out2}} = \frac{1}{\sqrt{2}}\left(\tilde{E}_1 - \tilde{E}_1\right) = 0, \tag{7.44}$$

there is a **constructive interference** at the first output and **destructive interference** at the second output. The total input power is $\propto |\tilde{E}_1|^2 + |\tilde{E}_2|^2 = 2|\tilde{E}_1|^2$, and the total output power $\propto 2|\tilde{E}_1|^2$ all goes to the first output.

For another example, suppose that $\tilde{E}_2 = -\tilde{E}_1$, such that the inputs have the same magnitude but 180° out of phase.

$$\tilde{E}_{\text{out1}} = \frac{1}{\sqrt{2}}\left(\tilde{E}_1 - \tilde{E}_1\right) = 0, \qquad \tilde{E}_{\text{out2}} = \frac{1}{\sqrt{2}}\left(\tilde{E}_1 + \tilde{E}_1\right) = \sqrt{2}\tilde{E}_1, \tag{7.45}$$

now all power goes into the second output instead.

In general, the output powers depend on the **relative phase** between the two inputs. Suppose that $\tilde{E}_2 = e^{j\phi}\tilde{E}_1$, such that the two inputs have equal magnitude but a relative phase of ϕ. The outputs become

$$\tilde{E}_{\text{out1}} = \frac{1}{\sqrt{2}}\left(\tilde{E}_1 + e^{j\phi}\tilde{E}_1\right), \qquad \tilde{E}_{\text{out2}} = \frac{1}{\sqrt{2}}\left(\tilde{E}_1 - e^{j\phi}\tilde{E}_1\right). \tag{7.46}$$

In terms of power,

$$\left|\tilde{E}_{\text{out1}}\right|^2 = \frac{\left|\tilde{E}_1\right|^2}{2}\left|1+e^{j\phi}\right|^2 = 2|\tilde{E}_1|^2\cos^2\frac{\phi}{2}, \tag{7.47}$$

$$\left|\tilde{E}_{\text{out2}}\right|^2 = \frac{\left|\tilde{E}_2\right|^2}{2}\left|1-e^{j\phi}\right|^2 = 2|\tilde{E}_1|^2\sin^2\frac{\phi}{2}, \tag{7.48}$$

which shows that the output powers depend on the **relative phase** between the inputs. Specifically, for $\phi = 0$, all power goes to the first output, and for $\phi = \pi$, all power goes to the second output. Note that the occurence of constructive interference at the first output at $\phi = 0$ (and destructive interference at the first output at $\phi = \pi$) is specific to the scattering matrix we have assumed. For a 50/50 beamsplitter with a different scattering matrix, constructive and destructive interferences at the outputs can occur at different values of ϕ.

- **Exercise**: Verify that $|1+e^{j\phi}|^2 = 4\cos^2(\phi/2)$ and $|1-e^{j\phi}|^2 = 4\sin^2(\phi/2)$.
  Answer: The easiest way is to write

$$|1+e^{j\phi}|^2 = \left(1+e^{j\phi}\right)\left(1+e^{j\phi}\right)^* \tag{7.49}$$

$$= \left(1+e^{j\phi}\right)\left(1+e^{-j\phi}\right) \tag{7.50}$$

$$= 1+e^{j\phi}+e^{-j\phi}+1 \tag{7.51}$$

$$= 2+2\cos\phi = 4\cos^2\frac{\phi}{2}, \tag{7.52}$$

  where I used the half-angle formula $\cos^2(\phi/2) = (1+\cos\phi)/2$. $|1-e^{j\phi}|^2$ is similar except that I need to use $\sin^2(\phi/2) = (1-\cos\phi)/2$.
- **Exercise**: The scattering matrix of a certain 50/50 beam splitter is given by

$$\begin{pmatrix} 1/\sqrt{2} & j/\sqrt{2} \\ j/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}. \tag{7.53}$$

  Check that it is unitary. Confirm that the output powers are equally split if there is just one input. Find $|\tilde{E}_{\text{out1}}|^2$ and $|\tilde{E}_{\text{out2}}|^2$ if the input amplitudes have equal magnitude but relative phase given by ϕ.

CHAPTER 8

Mach-Zehnder and Michelson Interferometers

8.1. Mach-Zehnder interferometer

We will now combine the delay lines discussed in Sec. 7.1 and two identical 50/50 beamsplitters as described in Sec. 7.7 to form a two-input two-output **interferometer** called a Mach-Zehnder interferometer, as shown in Fig. 8.1.

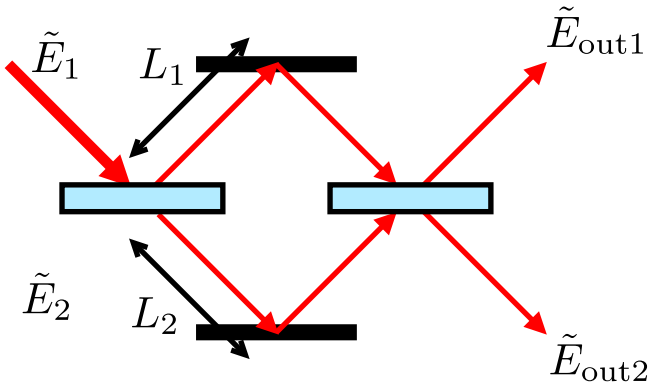


FIGURE 8.1. A Mach-Zehnder interferometer with two beamsplitters and two delay lines as arms.

We will go through the input-output analysis step-by-step and find out how the final output amplitudes depend on the input. For simplicity, assume that there is just one input $\tilde{E}_1$, and the second input is zero $\tilde{E}_2 = 0$. The scattering matrix for both beamsplitters are assumed to be Eq. (7.39). For the first beamsplitter, the input is split into two equal outputs, as shown in Fig. 8.2.

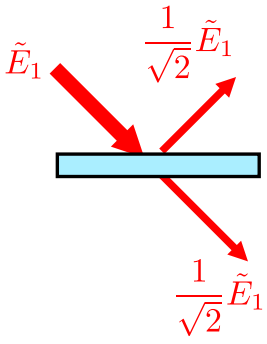


FIGURE 8.2. First beamsplitter of a Mach-Zehnder interferometer, assuming just one input.

The second step is the two delay lines in the two arms of the interferometer, shown in Fig. 8.3. We take the outputs of the first beamsplitter as the inputs to the delay lines. We assume for simplicity that the two mirrors in the arms are perfect mirrors with reflection coefficient 1. Each wave then picks up a phase factor proportional to the length of the arm.

Finally, we take the outputs of the delay lines and feed them into the inputs of the second beamsplitter, as shown in Fig. 8.4. The two inputs now have equal magnitude but different phases.

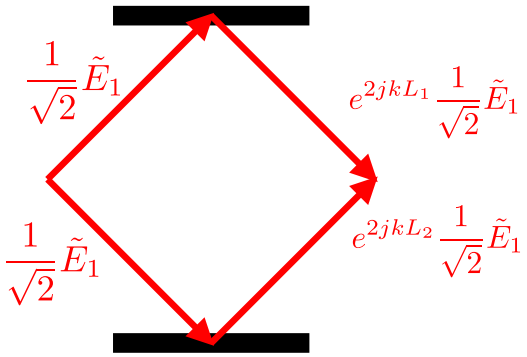


FIGURE 8.3. Propagation of waves in the two arms. Each wave picks up a phase factor depending on the lengths of the arms. Each mirror is assumed to be perfect with reflection coefficient 1.

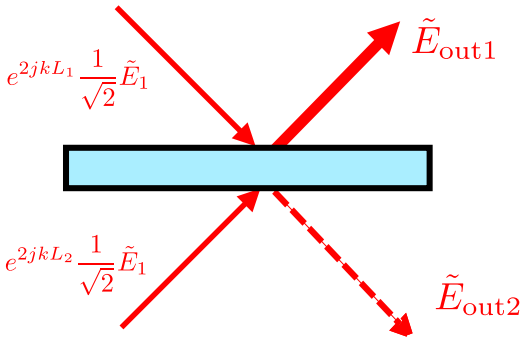


FIGURE 8.4. The second beamsplitter. Now we have two inputs with equal magnitude but different phases as inputs.

The final outputs are

$$\tilde{E}_{\text{out1}} = \left(s_{11}e^{2jkL_1}s_{11} + s_{12}e^{2jkL_2}s_{21}\right)\tilde{E}_1 = \frac{1}{2}\left(e^{2jkL_1} + e^{2jkL_2}\right)\tilde{E}_1, \tag{8.1}$$

$$\tilde{E}_{\text{out2}} = \left(s_{21}e^{2jkL_1}s_{11} + s_{22}e^{2jkL_2}s_{21}\right)\tilde{E}_1 = \frac{1}{2}\left(e^{2jkL_1} - e^{2jkL_2}\right)\tilde{E}_1. \tag{8.2}$$

Note that the minus sign comes from the s_{22} component for the second beamsplitter.

*Mathematically, what we have done here is simply the successive multiplication of scattering matrices of the components:

$$\begin{pmatrix} \tilde{E}_{\text{out1}} \\ \tilde{E}_{\text{out2}} \end{pmatrix} = \underbrace{\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}}_{\text{second beamsplitter}} \underbrace{\begin{pmatrix} e^{2jkL_1} & 0 \\ 0 & e^{2jkL_2} \end{pmatrix}}_{\text{delay lines}} \underbrace{\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}}_{\text{first beamsplitter}} \begin{pmatrix} \tilde{E}_1 \\ \tilde{E}_2 \end{pmatrix}. \tag{8.3}$$

We can do the same for other types of beamsplitters, delay lines, and more complicated optical networks. With N sequential components with scattering matrices $\{\boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, \ldots, \boldsymbol{s}^{(N)}\}$, the output is simply

$$\tilde{\boldsymbol{E}}_{\text{out}} = \boldsymbol{s}^{(N)} \ldots \boldsymbol{s}^{(2)}\boldsymbol{s}^{(1)}\tilde{\boldsymbol{E}}. \tag{8.4}$$

This approach also works for any number of inputs and outputs as long as the components are sequential.

- **Exercise**: If all the scattering matrices in Eq. (8.4) are unitary, prove that the final scattering matrix $\boldsymbol{s}^{(N)} \ldots \boldsymbol{s}^{(2)}\boldsymbol{s}^{(1)}$ is also unitary. What does it mean in terms of power conservation?

8.2. Optical switch

Suppose that the two arms have equal length $L_1 = L_2 = L$[1]. This means that the two inputs to the second beamsplitter have equal magnitude and phase, leading to

$$\tilde{E}_{\text{out1}} = e^{2jkL}\tilde{E}_1, \qquad \tilde{E}_{\text{out2}} = 0. \tag{8.5}$$

We have another example of **constructive interference** in the first output and **destructive interference** in the second output, similar to what we obtained in Sec. 7.7. All power goes into the first output.

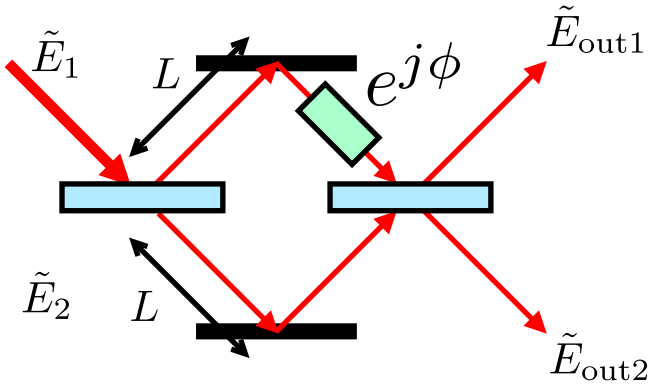


FIGURE 8.5. A Mach-Zehnder interferometer with an additional phase shift ϕ in the upper arm.

Suppose that we can introduce an **additional phase shift** ϕ to the upper arm, as shown in Fig. 8.5. This can be done by changing the length, the refractive index, or adding a phase modulator. Now there is a relative phase ϕ between the two inputs of the second beamsplitter, and the final outputs become

$$\tilde{E}_{\text{out1}} = \frac{1}{2}\left(e^{2jkL+j\phi} + e^{2jkL}\right)\tilde{E}_1, \qquad \tilde{E}_{\text{out2}} = \frac{1}{2}\left(e^{2jkL+j\phi} - e^{2jkL}\right)\tilde{E}_1. \tag{8.6}$$

Taking the common factor e^{2jkL} out of the bracket, and considering the power $\propto |\tilde{E}|^2$,

$$\boxed{\left|\tilde{E}_{\text{out1}}\right|^2 = \frac{1}{4}\left|e^{j\phi}+1\right|^2\left|\tilde{E}_1\right|^2 = \left|\tilde{E}_1\right|^2\cos^2\frac{\phi}{2},} \tag{8.7}$$

$$\boxed{\left|\tilde{E}_{\text{out2}}\right|^2 = \frac{1}{4}\left|e^{j\phi}-1\right|^2\left|\tilde{E}_1\right|^2 = \left|\tilde{E}_1\right|^2\sin^2\frac{\phi}{2}.} \tag{8.8}$$

For $\phi = 0$, all power goes to the first port, but the outputs can be controlled by changing ϕ. When $\phi = \pi$, all power goes to the second port instead.

The Mach-Zehnder interferometer is useful because it converts **phase modulation** to **intensity modulation**. It can be used as an optical intensity modulator by controlling ϕ, as an optical switch that routes an optical signal to either output port, or as a sensor that infers the unknown phase shift ϕ by measuring the output intensities.

- **Exercise**: Suppose that the beamsplitter scattering matrix is now given by Eq. (7.53). Calculate the output amplitudes and $|\tilde{E}_{\text{out1}}|^2$ and $|\tilde{E}_{\text{out2}}|^2$.
- **Exercise**: Calculate the output amplitudes if there are two arbitrary input amplitudes. Write the final input-output relation in matrix form. Confirm that the final scattering matrix for the Mach-Zehnder interferometer is unitary.

8.3. Partial-wave analysis

You might have noticed from Eq. (8.1) that the output amplitude $\tilde{E}_{\text{out1}}$ is the superposition of two terms times the input amplitude: one is $s_{11}e^{2jkL_1}s_{11}$ and the other is $s_{12}e^{2jkL_2}s_{21}$, and these two terms can be considered as the **complex weights** of two possible **paths** from the input to the output shown in Fig. 8.6. Each path, by itself, is not a solution of the wave equations and cannot exist on its own, but if we sum all the complex

[1] note that this L is no longer the thickness of a partial mirror; I am just recycling the symbol here.

weights that correspond to all the possible paths, we can still arrive at the final valid solution. For this reason, each path is called a partial wave.

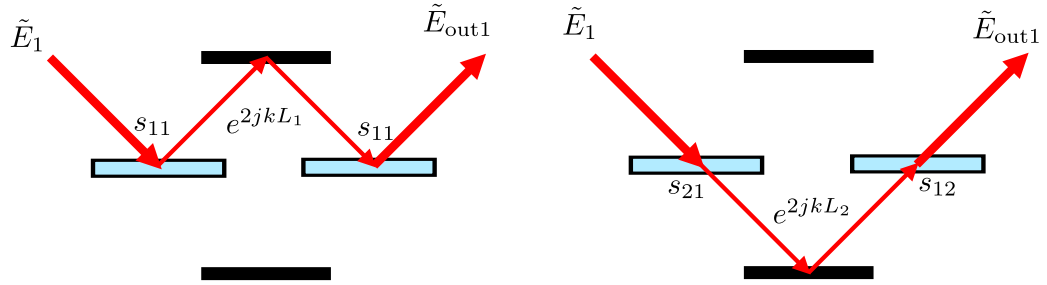


FIGURE 8.6. The two paths of how the input $\tilde{E}_1$ affects the output $\tilde{E}_2$. The first path on the left consists of reflection (s_{11}), propagation (e^{2jkL_1}), and reflection again (s_{11}), while the second path on the right consists of transmission (s_{21}), propagation (e^{2jkL_2}), and transmission (s_{12}).

Mathematically, consider the matrix form of the input-output analysis given by Eq. (8.4) for the Mach-Zehnder interferometer:

$$\tilde{\boldsymbol{E}}_{\text{out}} = \boldsymbol{s}^{(3)}\boldsymbol{s}^{(2)}\boldsymbol{s}^{(1)}\tilde{\boldsymbol{E}}, \tag{8.9}$$

where $\boldsymbol{s}^{(1)}$ is the scattering matrix of the first beamsplitter, $\boldsymbol{s}^{(2)}$ is that for the two delay lines, and $\boldsymbol{s}^{(3)}$ is that of the second beamsplitter. Let us write out the elements with the indices explicitly:

$$\tilde{E}_{\text{out}\,a} = \sum_b s^{(3)}_{ab} \sum_c s^{(2)}_{bc} \sum_d s^{(1)}_{cd} \tilde{E}_d. \tag{8.10}$$

You can verify this using the definition of matrix multiplication (Eq. (7.14)). Instead of doing the matrix multiplication at every step, Suppose that we defer the summations to the end of the calculation:

$$\tilde{E}_{\text{out}\,a} = \sum_{b,c,d} s^{(3)}_{ab} s^{(2)}_{bc} s^{(1)}_{cd} \tilde{E}_d. \tag{8.11}$$

Each of the terms $s^{(3)}_{ab} s^{(2)}_{bc} s^{(1)}_{cd}$ corresponds to the weight of a path of how an input $\tilde{E}_d$ affects the final output $\tilde{E}_{\text{out}\,a}$ through the intermediate inputs and outputs denoted by the indices c and b. The final result is to sum over the weights of all the possible paths from the inputs to an output. Fig. 8.6 is thus a simple example of this partial-wave analysis; the summing of weights comes only at the end of the calculation.

- **Exercise**: Using the partial-wave analysis, calculate the output $\tilde{E}_{\text{out2}}$ if there is just one input $\tilde{E}_1$. Confirm that it is equal to Eq. (8.2).
- **Exercise**: Using the partial-wave analysis, calculate the outputs if there is just one input $\tilde{E}_2$.

8.4. Michelson interferometer

A Michelson interferometer, shown in Fig. 8.7, is simply a folded Mach-Zehnder interferometer. The input-output relations, assuming the same beamsplitter scattering matrix given by Eq. (7.39), are the same as those for the Mach-Zehnder, as demonstrated by the input-output analysis in Fig. 8.8:

$$\tilde{E}_{\text{out1}} = \frac{1}{2}\left(e^{2jkL_1} + e^{2jkL_2}\right)\tilde{E}_1, \qquad \tilde{E}_{\text{out2}} = \frac{1}{2}\left(e^{2jkL_1} - e^{2jkL_2}\right)\tilde{E}_1. \tag{8.12}$$

The folded geometry makes it easier to change the relative lengths of the two arms. The phase difference between the two paths is $\phi = 2kL_1 - 2kL_2$. In terms of the intensities,

$$\left|\tilde{E}_{\text{out1}}\right|^2 = \cos^2\left[k(L_1 - L_2)\right]\left|\tilde{E}_1\right|^2, \qquad \left|\tilde{E}_{\text{out2}}\right|^2 = \sin^2\left[k(L_1 - L_2)\right]\left|\tilde{E}_1\right|^2. \tag{8.13}$$

Historically the Michelson interferometer was used to measure the speed of light in different directions with respect to the motion of the earth. The measured constant speed of light motivated Einstein's discovery of special relativity. The Michelson interferometer is still being used in important scientific experiments, e.g., gravitational-wave detection. For more details, see http://en.wikipedia.org/wiki/Michelson_interferometer.

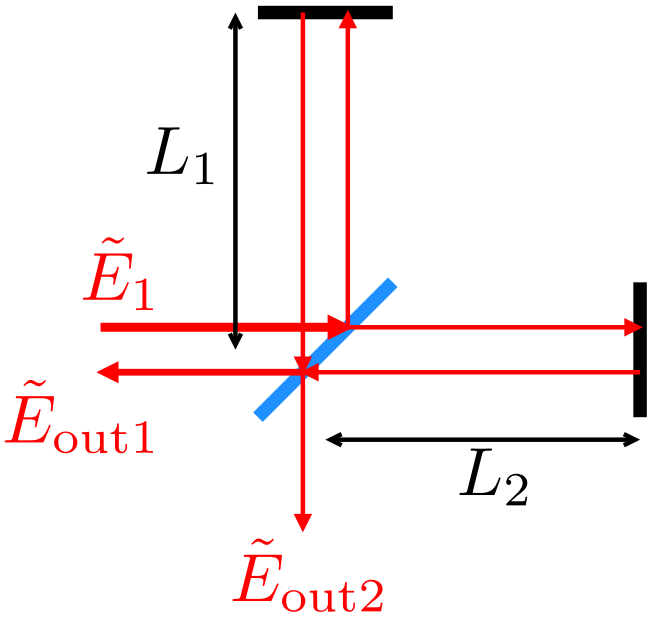


FIGURE 8.7. A Michelson interferometer.

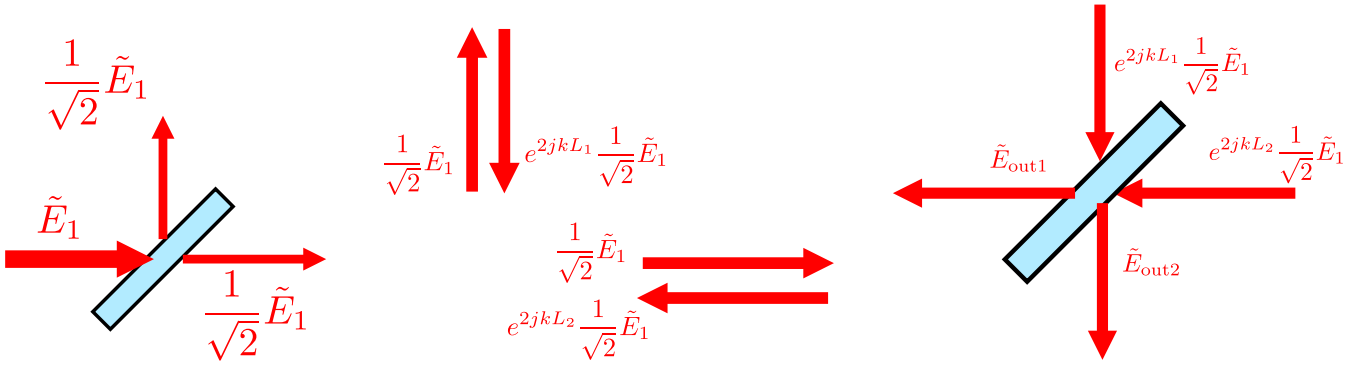


FIGURE 8.8. Input-output analysis of the Michelson interferometer. Left: first scattering off the beam splitter. Center: Propagation in the two arms. Right: Second scattering off the beam splitter.

CHAPTER 9

Fabry-Pérot Interferometer

For the Mach-Zehnder and Michelson interferometers, the optical components are sequential, meaning that there is no **feedback** involved: the outputs of one component always go to the inputs of a different component such that the inputs will never be affected by themselves. Here we will study a different type of interferometer called the Fabry-Pérot interferometer (also called Fabry-Pérot etalon or Fabry-Pérot **cavity**), which involves feedback and **multiple reflections** between two interfaces, and the mathematics and physical results will turn out to look quite different.

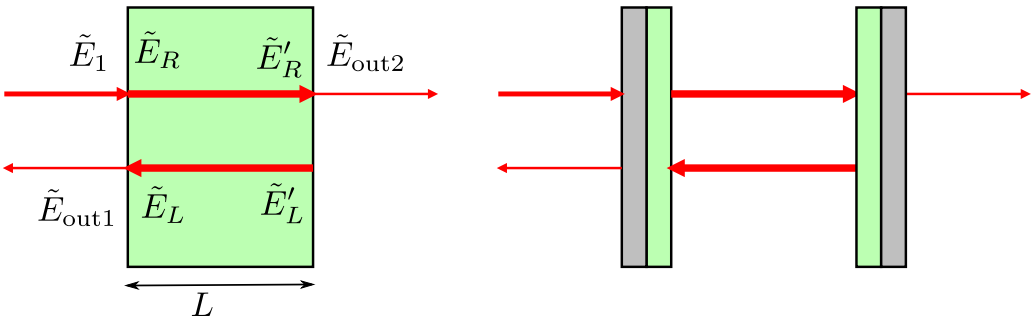


FIGURE 9.1. Two examples of the Fabry-Pérot interferometer with normal incidence. Counter-propagating waves between the two parallel interfaces lead to multiple reflections.

Two different examples of the Fabry-Pérot interferometer are shown in Fig. 9.1. The simplest case is simply a dielectric slab, but unlike Sec. 7.3, we will now explicitly calculate the final transmission and reflection coefficients from the material properties. A different setup is on the right of Fig. 9.1, which consists of two parallel partial mirrors. We will assume normal incidence for simplicity; the analysis for oblique incidence is a bit more complicated but follows the same procedure.

9.1. Closed cavity with perfect mirrors

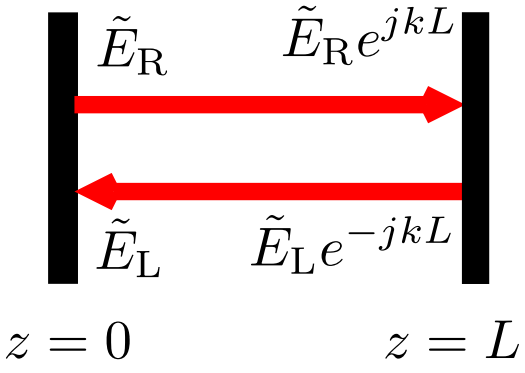


FIGURE 9.2. A closed Fabry-Pérot cavity with perfect mirrors.

As a warm-up exercise, let us first consider a Fabry-Pérot cavity with two perfectly conducting mirrors and a dielectric with index n in-between, as shown in Fig. 9.2. There is no external input or output, and the fields inside the cavity are

$$\boldsymbol{E} = \hat{\boldsymbol{x}} \left[\tilde{E}_{\text{R}} \exp(jkz - j\omega t) + \tilde{E}_{\text{L}} \exp(-jkz - j\omega t) \right], \tag{9.1}$$

where $k = \omega n/c$ and we assume one linear polarization for simplicity. $\tilde{E}_{\mathrm{R}}$ is the complex amplitude for the right-propagating wave, and $\tilde{E}_{\mathrm{L}}$ is that for the left-propagating wave. We know that a perfect conductor has infinite conductance ($\boldsymbol{J} = \sigma \boldsymbol{E}$, $\sigma \to \infty$) and therefore zero electric field:

$$\boldsymbol{E}(z = 0) = \boldsymbol{E}(z = L) = 0. \tag{9.2}$$

Given these boundary conditions, we would like to solve for possible electric fields inside the cavity. Matching the boundary conditions at $z = 0$ and $z = L$:

$$\tilde{E}_{\mathrm{R}} + \tilde{E}_{\mathrm{L}} = 0, \quad \text{at } z = 0, \tag{9.3}$$

$$\tilde{E}_{\mathrm{R}} e^{jkL} + \tilde{E}_{\mathrm{L}} e^{-jkL} = 0, \quad \text{at } z = L, \tag{9.4}$$

which lead to

$$\tilde{E}_{\mathrm{R}} = s_{22} \tilde{E}_{\mathrm{L}}, \qquad s_{22} = -1, \tag{9.5}$$

$$\tilde{E}_{\mathrm{L}} e^{-jkL} = s_{22} \tilde{E}_{\mathrm{R}} e^{jkL}, \qquad s_{22} = -1. \tag{9.6}$$

If we think of these as input-output relations, Eq. (9.5) says that $\tilde{E}_L$ as the input is reflected by the left mirror with a reflection coefficient $s_{22} = -1$ to become the output $\tilde{E}_R$, while Eq. (9.6) says that the input at the right mirror is now the right-propagating wave, and after acquiring a phase delay e^{jkL} via propagation, it is reflected with a reflection coefficient $s_{22} = -1$. The output for the right mirror is $\tilde{E}_{\mathrm{L}} e^{-jkL}$, because it is the input to the delay line between the mirrors with output defined as $\tilde{E}_{\mathrm{L}}$. We see the phenomenon of **feedback** here: the input and output at one mirror switch their roles at another mirror, causing an input to become its own input.

Combining Eqs. (9.5) and (9.6),

$$\tilde{E}_{\mathrm{L}} = \tilde{E}_{\mathrm{L}} s_{22}^2 e^{2jkL}. \tag{9.7}$$

Qualitatively speaking, this condition says that the wave must be equal to itself after completing one **round trip**, as the **round-trip factor**

$$G = s_{22}^2 e^{2jkL} \tag{9.8}$$

corresponds to the weights that the wave picks up in one round trip in the cavity, consisting of reflection by the left mirror (s_{22}), propagation to the right (e^{jkL}), reflection by the right mirror (s_{22}), and propagation to the left (e^{jkL}).

To satisfy this consistency requirement, either $\tilde{E}_{\mathrm{L}} = 0$ and $\tilde{E}_{\mathrm{R}} = 0$ (no wave inside), or

$$\boxed{s_{22}^2 e^{2jkL} = e^{2jkL} = 1.} \tag{9.9}$$

The boundary conditions hence require the round-trip factor to be equal to 1. For this to happen, $2kL$ must be integer-multiples of 2π, which is a requirement on k and therefore the frequency $2\pi\nu = \omega = ck/n$ that can exist inside the cavity,

$$\boxed{kL = \pi q, \quad L = \frac{q\lambda}{2}, \quad q = \text{integer},} \tag{9.10}$$

$$\boxed{\omega = \frac{q\pi c}{nL},} \tag{9.11}$$

$$\boxed{\nu = \frac{\omega}{2\pi} = \frac{qc}{2nL}.} \tag{9.12}$$

These discrete frequencies are called **resonant frequencies** of the cavity. They are **equally spaced** for the optical cavity, with the separation in Hertz given by

$$\boxed{\Delta\nu \equiv \frac{c}{2nL},} \tag{9.13}$$

which is called the **free-spectral range**. The electric field becomes

$$\boldsymbol{E} \propto \hat{\boldsymbol{x}} \sin(kz) \exp(-j\omega t), \tag{9.14}$$

which is a **standing wave**, and only discrete numbers of half-waves can fit inside the cavity; the first few modes are shown in Fig. 9.3. A discussion of standing waves in a microwave cavity can be found at https://www.youtube.com/watch?v=kp33ZprO0Ck.

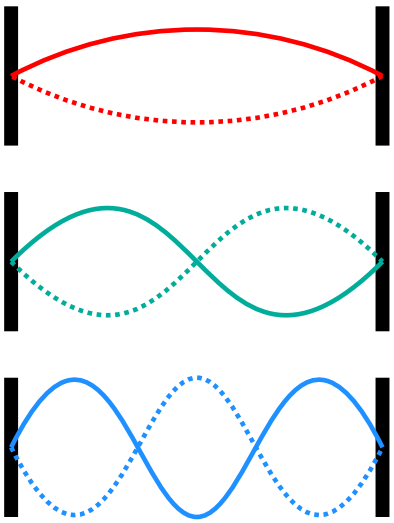


FIGURE 9.3. The field amplitudes of the first few standing-wave modes with $k = \pi/L$, $k = 2\pi/L$, and $k = 3\pi/L$, with solid curves corresponding to the amplitudes at one time and dash curves corresponding to the amplitudes at another time. ω is proportional to k and also becomes higher for higher-order modes.

9.2. Input-output analysis of a dielectric slab

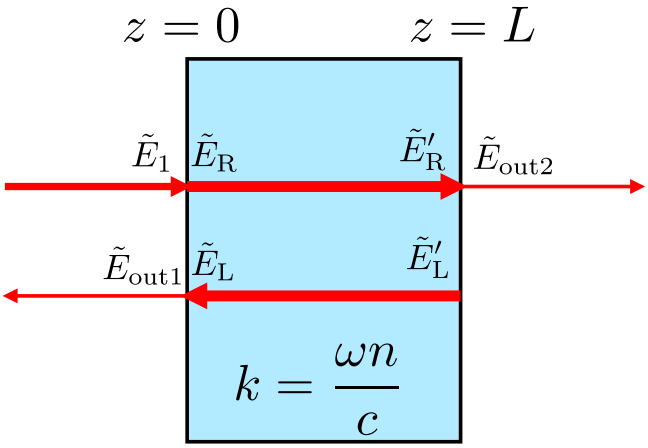


FIGURE 9.4. A dielectric slab as an example of a Fabry-Pérot interferometer.

We will now consider the dielectric slab in Fig. 9.4 with refractive index n and free space ($n_0 = 1$) on either side. First, the scattering matrix for the left interface is

$$\begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} = \begin{pmatrix} \frac{1-n}{1+n} & \frac{2n}{1+n} \\ \frac{2}{1+n} & \frac{n-1}{1+n} \end{pmatrix}, \tag{9.15}$$

which can be obtained from the reflection and transmission coefficients for TE polarization in Eq. (6.54).[1] In terms of the n_1 and n_2 defined in the reflection/refraction problem (Fig. 6.13), note that

$$s_{11} = \frac{n_1 - n_2}{n_1 + n_2} = \frac{1-n}{1+n}, \qquad s_{21} = \frac{2n_1}{n_1 + n_2} = \frac{2}{1+n} \tag{9.18}$$

are the reflection and transmission coefficients for an incident wave ($\tilde{E}_1$) from $n_1 = 1$ free space to a $n_2 = n$ dielectric, whereas s_{22} and s_{12} are the coefficients for an incident wave ($\tilde{E}_\mathrm{L}$) from the $n_1 = n$ dielectric to the $n_2 = 1$ free space:

$$s_{22} = \frac{n_1 - n_2}{n_1 + n_2} = \frac{n-1}{n+1}, \qquad s_{12} = \frac{2n_1}{n_1 + n_2} = \frac{2n}{n+1}. \tag{9.19}$$

The input-output relation becomes

$$\begin{pmatrix} \tilde{E}_{\mathrm{out}1} \\ \tilde{E}_\mathrm{R} \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \begin{pmatrix} \tilde{E}_1 \\ \tilde{E}_\mathrm{L} \end{pmatrix}. \tag{9.20}$$

Second, propagation inside the the dielectric with wavenumber $k = \omega n/c$ leads to

$$\begin{pmatrix} \tilde{E}'_\mathrm{R} \\ \tilde{E}_\mathrm{L} \end{pmatrix} = \begin{pmatrix} e^{jkL} & 0 \\ 0 & e^{jkL} \end{pmatrix} \begin{pmatrix} \tilde{E}_\mathrm{R} \\ \tilde{E}'_\mathrm{L} \end{pmatrix}. \tag{9.21}$$

In terms of the delay line, $\tilde{E}_\mathrm{L}$ is the output and $\tilde{E}'_\mathrm{L}$ is the input, so the input has a **negative phase delay** with respect to the output.

Finally, at the right interface,

$$\begin{pmatrix} e^{-jkL}\tilde{E}_\mathrm{L} \\ \tilde{E}_{\mathrm{out}2} \end{pmatrix} = \begin{pmatrix} s_{22} & s_{21} \\ s_{12} & s_{11} \end{pmatrix} \begin{pmatrix} e^{jkL}\tilde{E}_\mathrm{R} \\ \tilde{E}_2 \end{pmatrix}. \tag{9.22}$$

There are a few important details to note about Eq. (9.22):

- We have substituted $\tilde{E}'_R$ with $e^{jkL}\tilde{E}_R$ and $\tilde{E}'_L$ with $e^{-jkL}\tilde{E}_L$ according to Eq. (9.21).
- We have used the same scattering matrix elements as those for the left interface, except that now $\tilde{E}'_\mathrm{R}$ is an input from the dielectric to free space, and $\tilde{E}_2$ is an input from free space to the dielectric, so the positions of the elements in the matrix are all switched.
- $\tilde{E}_2$ is the input from the right of the slab, but we shall assume that it is zero for simplicity, as shown in Fig. 9.4.

I will write Eqs. (9.20) and (9.22) explicitly with $\tilde{E}_2 = 0$ again, side-by-side, corresponding to Fig. 9.5:

$$\begin{aligned} \tilde{E}_{\mathrm{out}1} &= s_{11}\tilde{E}_1 + s_{12}\tilde{E}_\mathrm{L}, & e^{-jkL}\tilde{E}_\mathrm{L} &= s_{22}e^{jkL}\tilde{E}_\mathrm{R}, \\ \tilde{E}_\mathrm{R} &= s_{21}\tilde{E}_1 + s_{22}\tilde{E}_\mathrm{L}, & \tilde{E}_{\mathrm{out}2} &= s_{12}e^{jkL}\tilde{E}_\mathrm{R}. \end{aligned} \tag{9.23}$$

The important point to note here is that $\tilde{E}_\mathrm{L}$ is an input and $\tilde{E}_\mathrm{R}$ is an output on the left, but their roles are switched on the right: $\tilde{E}_\mathrm{R}$ is now an input and $\tilde{E}_\mathrm{L}$ is an output. This is **feedback**: we feed an output of the left interface back to the input through the right interface, causing **multiple reflections**.

Let's focus on the transmitted amplitude $\tilde{E}_{\mathrm{out}2}$ first, with $\tilde{E}_2 = 0$, and find out how it depends on the input $\tilde{E}_1$. From Eqs. (9.23), this output depends on $\tilde{E}_\mathrm{R}$ on the right interface as follows:

$$\tilde{E}_{\mathrm{out}2} = s_{12}e^{jkL}\tilde{E}_\mathrm{R}. \tag{9.24}$$

[1] Note that this scattering matrix is **not unitary**, even though **power is conserved**. This is because the intensity in a dielectric is

$$I = \frac{|\tilde{E}|^2}{2Z} = \frac{n|\tilde{E}|^2}{2Z_0}, \tag{9.16}$$

which is proportional to n, and power conservation would lead to

$$n|\tilde{E}_L|^2 + |\tilde{E}_1|^2 = n|\tilde{E}_R|^2 + |\tilde{E}_{\mathrm{out}1}|^2. \tag{9.17}$$

This would not lead to a mathematically unitary scattering matrix. Power conservation leads to a unitary matrix only if all the input and output waves are propagating in media with the same refractive index.

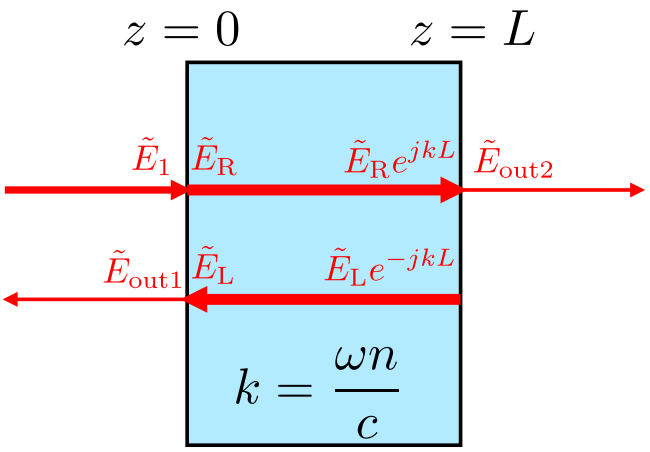


FIGURE 9.5. Same as Fig. 9.4, but with the propagation phase delays taken into account implicitly.

The simplest way to proceed is to **reverse** the top-right input-output relation in Eqs. (9.23) and write $\tilde{E}_\mathrm{R}$ with respect to $\tilde{E}_\mathrm{L}$:

$$\tilde{E}_\mathrm{R} = \frac{1}{s_{22} e^{2jkL}} \tilde{E}_\mathrm{L}. \tag{9.25}$$

Substituting this into the bottom-left equation of Eqs. (9.23),

$$\frac{1}{s_{22} e^{2jkL}} \tilde{E}_\mathrm{L} = s_{21} \tilde{E}_1 + s_{22} \tilde{E}_\mathrm{L}, \tag{9.26}$$

or

$$\tilde{E}_\mathrm{L} = \frac{s_{21}}{1/(s_{22} e^{2jkL}) - s_{22}} \tilde{E}_1. \tag{9.27}$$

Now I can substitute this back to Eq. (9.25):

$$\tilde{E}_\mathrm{R} = \frac{s_{21}}{1 - s_{22}^2 e^{2jkL}} \tilde{E}_1, \tag{9.28}$$

and then back to Eq. (9.24):

$$\boxed{\tilde{E}_\mathrm{out2} = \frac{s_{12} s_{21} e^{jkL}}{1 - s_{22}^2 e^{2jkL}} \tilde{E}_1.} \tag{9.29}$$

This is the transmission for a Fabry-Pérot interferometer. You might have seen something like this in **transmission line theory**. Note that some scattering coefficients end up in the denominator, quite unlike the outputs of a sequential interferometer, and the round-trip factor $s_{22}^2 e^{2jkL}$ will play a crucial role in our discussion.

To express $\tilde{E}_\mathrm{out1}$, the reflected wave, as a function of $\tilde{E}_1$, we can simply substitute Eq. (9.27) into the top-left of Eq. (9.23), resulting in

$$\boxed{\tilde{E}_\mathrm{out1} = \left(s_{11} + \frac{s_{12} s_{21} s_{22} e^{2jkL}}{1 - s_{22}^2 e^{2jkL}} \right) \tilde{E}_1.} \tag{9.30}$$

This is the reflection for a Fabry-Pérot interferometer.

- **Exercise**: Suppose that we have also found how $\tilde{E}_\mathrm{out1}$ and $\tilde{E}_\mathrm{out2}$ depend on arbitrary inputs $\tilde{E}_1$ and $\tilde{E}_2$ for the Fabry-Pérot interferometer. Is the corresponding scattering matrix unitary?

9.3. Partial-wave analysis

There is an alternative and clumsier way of analyzing the problem, if I was not smart enough to reverse the input-output relation like what I did in Eq. (9.25) and strictly follow how an input affects an output. This method will nonetheless provide a different perspective on the physics of a feedback system and is not unlike the simple partial-wave analysis Sec. 8.3 for Mach-Zehnder interferometer.

Again we start from Eq. (9.24) and write $\tilde{E}_{\rm R}$ in terms of the inputs of the left interface using bottom-left of Eq. (9.23):

$$\begin{aligned}\tilde{E}_{\rm out2} &= s_{12}e^{jkL}\tilde{E}_{\rm R} \\ &= s_{12}e^{jkL}\left(s_{21}\tilde{E}_1 + s_{22}\tilde{E}_{\rm L}\right).\end{aligned} \tag{9.31}$$

We have a term that depends on the input $\tilde{E}_1$, which is what we want, but there is also a term $\tilde{E}_{\rm L}$, so let's express it in terms of the inputs on the right interface using top-right of Eq. (9.23):

$$\begin{aligned}\tilde{E}_{\rm out2} &= s_{12}e^{jkL}\tilde{E}_{\rm R} \\ &= s_{12}e^{jkL}\left(s_{21}\tilde{E}_1 + s_{22}\tilde{E}_{\rm L}\right) \\ &= s_{12}e^{jkL}\left[s_{21}\tilde{E}_1 + s_{22}\left(s_{22}e^{2jkL}\tilde{E}_{\rm R}\right)\right]\end{aligned} \tag{9.32}$$

Let's do this one more time to express $\tilde{E}_{\rm R}$ in terms of $\tilde{E}_{\rm L}$:

$$\begin{aligned}\tilde{E}_{\rm out2} &= s_{12}e^{jkL}\tilde{E}_{\rm R} \\ &= s_{12}e^{jkL}\left(s_{21}\tilde{E}_1 + s_{22}\tilde{E}_{\rm L}\right) \\ &= s_{12}e^{jkL}\left[s_{21}\tilde{E}_1 + s_{22}\left(s_{22}e^{2jkL}\tilde{E}_{\rm R}\right)\right] \\ &= s_{12}e^{jkL}\left\{s_{21}\tilde{E}_1 + s_{22}\left[s_{22}e^{2jkL}\left(s_{21}\tilde{E}_1 + s_{22}\tilde{E}_{\rm L}\right)\right]\right\}\end{aligned} \tag{9.33}$$

We get another term with respect to the input $\tilde{E}_1$, but yet again a new term $\propto \tilde{E}_{\rm L}$ appears. At this point you should recognize that this forward way of analysis will go on indefinitely and never end, and we will end up with an **infinite series**:

$$\tilde{E}_{\rm out2} = \left[s_{12}e^{jkL}s_{21} + s_{12}\left(s_{22}^2e^{2jkL}\right)e^{jkL}s_{21} + s_{12}\left(s_{22}^2e^{2jkL}\right)^2 e^{jkL}s_{21} + \ldots\right]\tilde{E}_1. \tag{9.34}$$

Each of these terms can be represented by a **partial wave** depicted in Fig. 9.6 and corresponds to the **complex weight** of a possible **path** from the input $\tilde{E}_1$ to the output $\tilde{E}_{\rm out2}$.

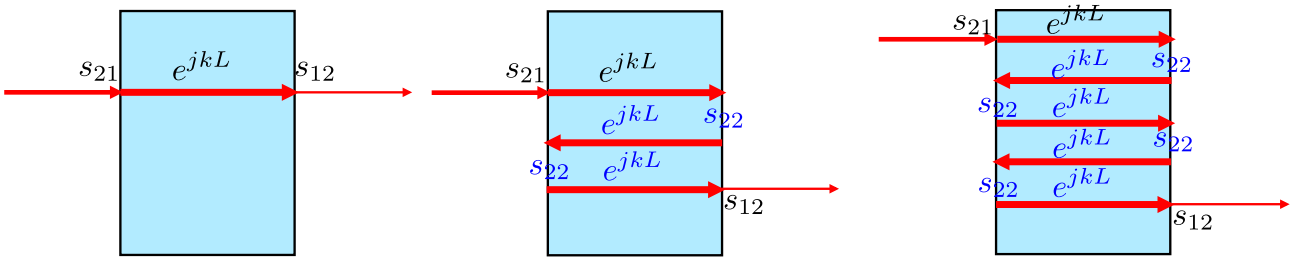


FIGURE 9.6. Each partial wave corresponds to a possible path from the input to the output. The left figure is the lowest-order term with a weight given by $s_{12}e^{jkL}s_{21}$, the center figure has a weight $s_{12}e^{jkL}s_{22}e^{jkL}s_{22}e^{jkL}s_{21}$, etc.

Unlike the Mach-Zehnder interferometer, where there are just two possible paths from an input to an output, here there are **infinite paths**, due to **infinite multiple reflections**. Moreover, the weights of the paths are related: The zeroth-order path on the left of Fig. 9.6 just goes straight through the slab, the first-order path in the center of Fig. 9.6 does one **round trip** before coming out, while the second-order path does two **round trips**. Higher-order paths are similar and differ from the lower-order paths only by the number of round trips performed. The infinite series is hence

$$\tilde{E}_{\rm out2} = s_{12}s_{21}e^{jkL}\left[1 + \left(s_{22}^2e^{2jkL}\right) + \left(s_{22}^2e^{2jkL}\right)^2 + \ldots\right] \tag{9.35}$$

$$= \left[s_{12}s_{21}e^{jkL}\sum_{m=0}^{\infty}\left(s_{22}^2e^{2jkL}\right)^m\right]\tilde{E}_1. \tag{9.36}$$

Now we can use an identity for infinite geometric series:

$$\sum_{m=0}^{\infty} G^m = \begin{cases} \frac{1}{1-G}, & |G| < 1, \\ \infty, & |G| \geq 1, \end{cases} \tag{9.37}$$

to obtain the final answer

$$\tilde{E}_{\text{out2}} = \frac{s_{12}s_{21}e^{jkL}}{1 - s_{22}^2 e^{2jkL}} \tilde{E}_1, \tag{9.38}$$

which is the same as Eq. (9.29). The **round-trip factor** G corresponds to the weight for one round trip inside the cavity:

$$\boxed{G = s_{22}^2 e^{2jkL}.} \tag{9.39}$$

Fortunately for us here, $|G| = |s_{22}| < 1$, as s_{22} is the reflection cofficient inside the cavity, and the infinite series converges.

The transmission depends crucially on the magnitude of G and the reflection coefficient s_{22} inside the cavity. For example, if the reflections at the interfaces are very weak, such that $|s_{22}| \ll 1$ $|G| = |s_{22}^2 e^{2jkL}| \ll 1$, and $|1 - G| \approx 1$, the higher-order partial waves have negligible magnitudes relative to the zeroth-order term, and we obtain

$$\tilde{E}_{\text{out2}} \approx s_{12}s_{21}e^{jkL}\tilde{E}_1, \tag{9.40}$$

which says that the transmission essentially does not involve round trips inside the cavity if the reflections off the interfaces are very weak.

- **Exercise**: Expand $1/(1 - G)$ in Taylor series to show that it is equal to the infinite geometric series. Note that the Taylor series does not converge for $|G| \geq 1$.
- **Exercise**: Using the partial-wave analysis, show that the reflection relation of a Fabry-Pérot interferometer is given by Eq. (9.30). Fig. 9.7 should help you.

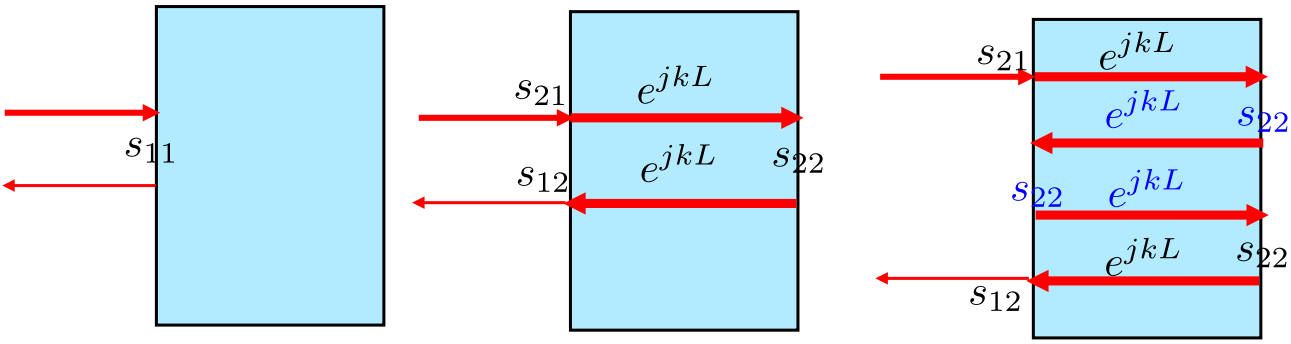


FIGURE 9.7. The first three partial waves for the reflection by a Fabry-Pérot interferometer. Note that the partial wave at the center has not completed a round trip, while the partial wave on the right has completed one round trip.

9.4. Transmission and reflection spectra

We would now like to study the frequency dependence of the **power transmission coefficient** from Eq. (9.29):

$$\boxed{T \equiv \left| \frac{\tilde{E}_{\text{out2}}}{\tilde{E}_1} \right|^2 = \frac{|s_{12}s_{21}|^2}{\left|1 - s_{22}^2 e^{2jkL}\right|^2},} \tag{9.41}$$

Recall that the $\boldsymbol{s}$ elements, given by Eq. (9.15), do not depend on frequency, and the only frequency dependence is the e^{2jkL} term in the denominator. When is the transmission **maximum**?

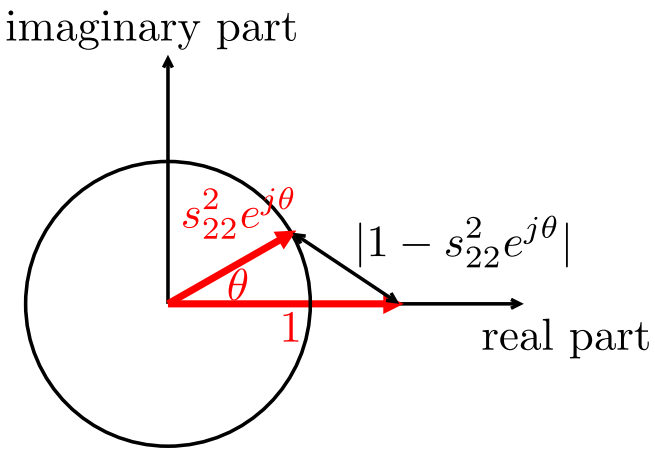


FIGURE 9.8. Depiction of $|1 - s_{22}^2 e^{j\theta}|$ in the complex plane. The circle has a radius s_{22}^2, and the point $s_{22}^2 e^{j\theta}$ is on the circle and has a phase θ. The distance between the two points 1 and $s_{22}^2 e^{j\theta}$ is minimum when θ is integer-multiples of 2π.

Let's study the term $|1 - s_{22}^2 e^{2jkL}|$ in the denominator, as drawn in the complex plane in Fig. 9.8. It corresponds to the distance between the two complex points 1 and $s_{22}^2 e^{2jkL}$, and the distance is minimum when the angle $\theta = 2kL$ is integer-multiples of 2π, or

$$2kL = 2\pi q, \quad q = \text{integer}. \tag{9.42}$$

Since $k = \omega n/c = 2\pi\nu n/c$, the frequencies at which the denominator is minimum and the transmission is maximum are

$$\omega = \frac{q\pi c}{nL}, \quad \nu = \frac{qc}{2nL}, \quad q = \text{integer}. \tag{9.43}$$

The maximum-transmission frequencies coincide with the **resonant frequencies** of a closed cavity given by Eq. (9.12) and are **equally spaced** by the **free spectral range** $\Delta\nu$ given by Eq. (9.13).

Another way of understanding the resonant transmission is to consider the round-trip factor $G = s_{22}^2 e^{2jkL}$. The partial waves described in Sec. 9.3 can produce constructive interference if they all have the same phase, and since their weights all differ from one another by multiples of G, they have the same phase when the phase of G is multiples of 2π.

Since the input and output waves of the Fabry-Pérot interferometer are in media with the same refractive index, we expect $|\tilde{E}_1|^2 + |\tilde{E}_2|^2 = |\tilde{E}_{\text{out1}}|^2 + |\tilde{E}_{\text{out2}}|^2$ from Eq. (7.28) to hold as a consequence of power conservation. With $\tilde{E}_2 = 0$, and defining the **power reflection coefficient** as

$$\boxed{R \equiv \frac{|\tilde{E}_{\text{out1}}|^2}{|\tilde{E}_1|^2},} \tag{9.44}$$

we can expect

$$T + R = \frac{|\tilde{E}_{\text{out2}}|^2}{|\tilde{E}_1|^2} + \frac{|\tilde{E}_{\text{out1}}|^2}{|\tilde{E}_1|^2} = 1 \tag{9.45}$$

for a lossless Fabry-Pérot interferometer. Fig. 9.9 plots the transmission and reflection spectra against normalized frequency $\nu/\Delta\nu = 2nL\nu/c$. Transmission is maximum when ν is multiples of $\Delta\nu$, and it turns out to be 100% at those resonant frequencies for a lossless dielectric slab. Reflection $R = 1 - T$ is then naturally minimum at those frequencies.

- **Exercise**: Confirm that the transmission is 100% for a lossless dielectric slab at maximum transmission using Eq. (9.41).

For a large n, the reflection cofficients at each interface (s_{11} and s_{22}) have high magnitudes, and if there is just one interface we expect most light will be reflected. With **two interfaces**, however, now the input wave can be resonant with the waves inside the cavity and produce maximum transmission at resonant frequencies. This counterintuitive phenomenon of **resonant transmission with highly reflective mirrors** is a hallmark of wave physics.

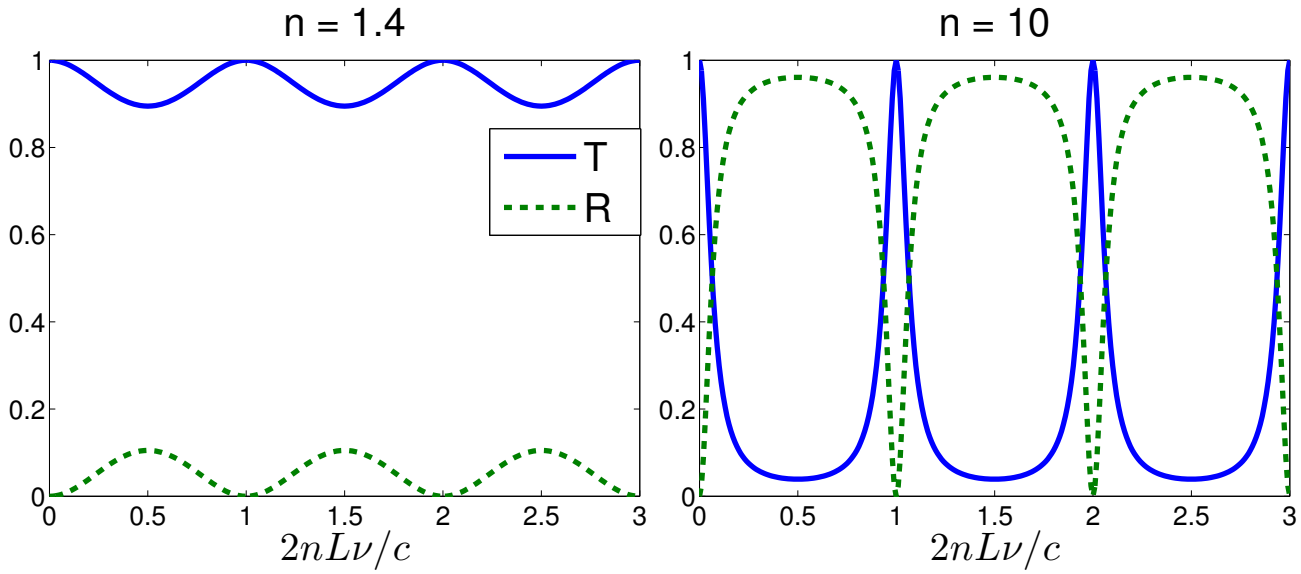


FIGURE 9.9. Plots of the power transmission coefficient T and reflection coefficient R with respect to normalized frequency $\nu/\Delta\nu = 2nL\nu/c$ for two different indices n. Transmission is maximum and reflection is minimum when ν is multiples of $\Delta\nu$ (resonant frequencies).

For thin dielectric films such as oil on water or soap bubbles, the effect of multiple reflections at the interfaces is to transmit some frequencies and reflect others. The transmission and reflection also depend on frequency and incident angle, so this is why they appear colorful.

FIGURE 9.10. Left: oil film on water; right: soap bubble. From [1].

In practice, Fabry-Pérot interferometers with highly reflecting partial mirrors (right of Fig. 9.1) are often used as **color filters** in optical experiments.

9.5. Anti-reflective coating

Anti-reflective coatings can be found on many optical components, such as glasses, camera lens, televisions, and computer displays. The purpose is to maximize transmission and minimize reflection by choosing the refractive index and thickness of the coating appropriately, as shown in Fig. 9.11.

- **Question**: If you look at your glasses or a camera lens at an angle, such as Fig. 9.12, why do they have a red/blue tint?
  Answer: Anti-reflective coating is typically optimized for **green color** (it has the highest intensity in sunlight spectrum) and **normal incidence**. The coating does not work as well for red and blue colors at oblique angles and therefore reflects more of them.

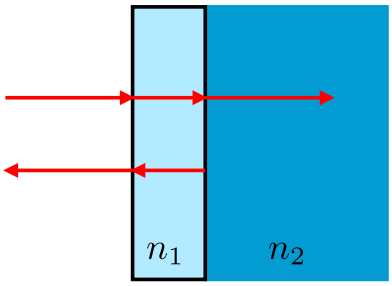


FIGURE 9.11. An anti-reflective coating with index n_1 is used in-between two media to maximize transmission into the n_2 medium.

FIGURE 9.12. Red/blueish tint of a camera lens. (Image from o5.com/how-to-get-rid-of-mold-on-your-camera-lens/)

9.6. Laser cavity

A Fabry-Pérot optical cavity is an important component of a laser. A typical laser cavity consists of one partially transmitting mirror and a perfectly reflective mirror. There is also an optical gain medium inside to provide power, as shown in Fig. 9.13. Multiple reflections inside the cavity at resonance can enhance the gain when on resonance. We will discuss more of this later.

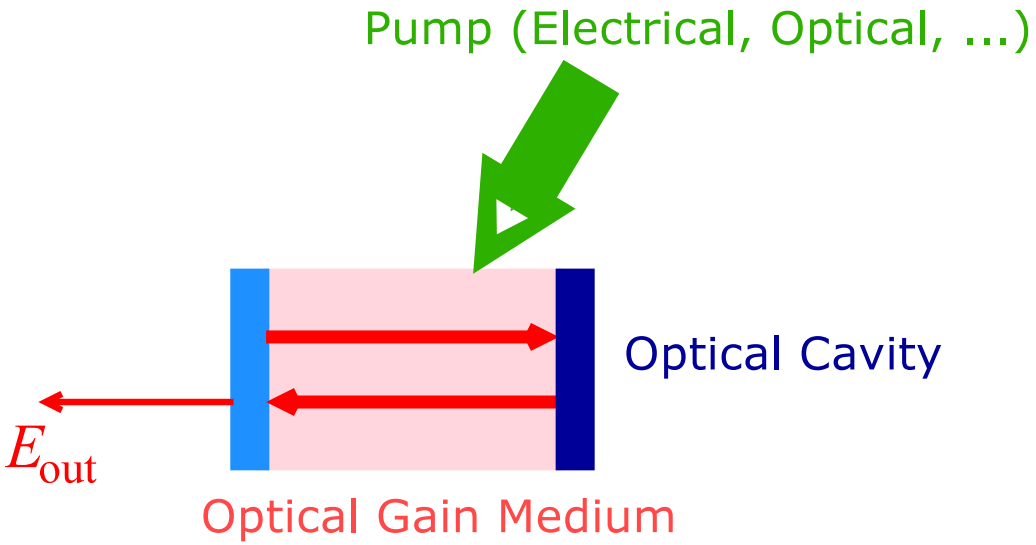


FIGURE 9.13. A typical laser consists of a Fabry-Pérot cavity and an optical gain medium inside to provide optical amplification. The purpose of the cavity is to introduce multiple reflections inside the cavity at resonance to enhance the gain and increase the output.

CHAPTER 10

Interference in Two Space Dimensions

We shall now study waves with **spatially varying intensity**. We know plane wave solutions in free space and spatially homogeneous dielectric media and their intensities are constant everywhere, but when we look at light coming from the pixels of our computer or smartphone displays, they do not look like plane waves but are fields with a spatially varying intensity. Our ultimate goal in the next two chapters is to relate some spatially varying fields at an **input plane**, such as your monitor or smartphone screen, to the fields and intensity at the **output plane**, such as your eyes or a camera. In this section we shall perform some warm-up calculations by considering the superposition of two waves in two dimensional space.

10.1. Interference between two plane waves

To produce a spatially varying intensity, the simplest example is two monochromatic plane waves. This can be observed, for example, using two overlapping laser beams that propagate at an angle with respect to each other. Suppose that one plane wave is propagating at an angle θ with respect to the z axis, as shown in Fig. 3.4, but now we have another plane wave with wavevector $\boldsymbol{k}'$ that is also in the (x, z) plane with the opposite angle, as shown in Fig. 10.1. Assume that they have the same frequency ω and also the same electric-field amplitude $\tilde{\boldsymbol{E}} = \tilde{E}\hat{\boldsymbol{y}}$. What is the resulting intensity?

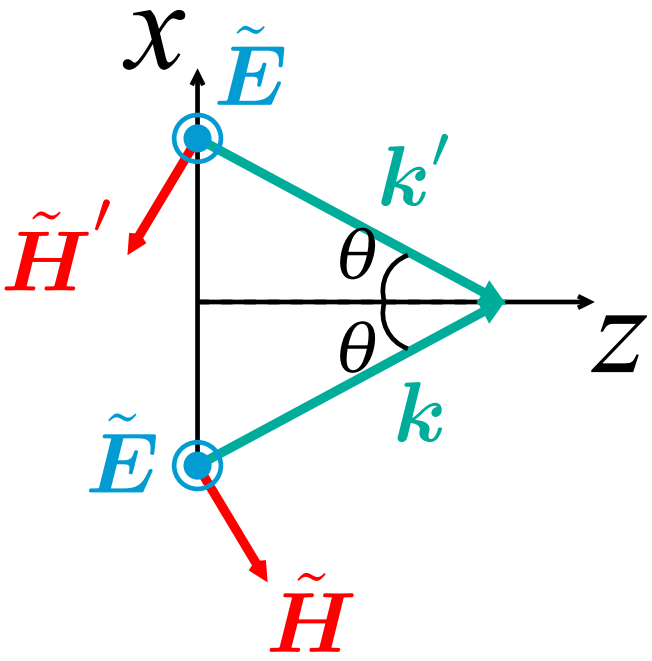


FIGURE 10.1. Two plane waves at an angle. Note that θ is assumed to be the angle that each wavevector makes with the z axis, and the angle between the wavevectors is 2θ.

First of all, we can write the wavevectors $\boldsymbol{k}$ as

$$\boldsymbol{k} = k_x\hat{\boldsymbol{x}} + k_z\hat{\boldsymbol{z}} = k\sin\theta\hat{\boldsymbol{x}} + k\cos\theta\hat{\boldsymbol{z}}, \tag{10.1}$$

$$\boldsymbol{k}' = k'_x\hat{\boldsymbol{x}} + k'_z\hat{\boldsymbol{z}} = -k\sin\theta\hat{\boldsymbol{x}} + k\cos\theta\hat{\boldsymbol{z}}. \tag{10.2}$$

Note that, since they have the same frequency, they have the same wavenumber

$$k = \frac{\omega n}{c}, \tag{10.3}$$

and the two wavevectors happen to have the same $\hat{\boldsymbol{z}}$ component ($k'_z = k_z$) but opposite $\hat{\boldsymbol{x}}$ component ($k'_x = -k_x$).

Second, the total electric field is the superposition of the two plane waves and can be written as

$$\boldsymbol{E} = \tilde{E}\hat{\boldsymbol{y}}\exp(j\boldsymbol{k}\cdot\boldsymbol{r} - j\omega t) + \tilde{E}\hat{\boldsymbol{y}}\exp(j\boldsymbol{k}'\cdot\boldsymbol{r} - j\omega t), \tag{10.4}$$

$$= \tilde{E}\hat{\boldsymbol{y}}\exp(jk_x x + jk_z z - j\omega t) + \tilde{E}\hat{\boldsymbol{y}}\exp(-jk_x x + jk_z z - j\omega t) \tag{10.5}$$

$$= \tilde{E}\hat{\boldsymbol{y}}e^{jk_z z - j\omega t}\left[\exp(jk_x x) + \exp(-jk_x x)\right] \tag{10.6}$$

$$= 2\tilde{E}\hat{\boldsymbol{y}}e^{jk_z z - j\omega t}\cos(k_x x). \tag{10.7}$$

With the same polarization $\hat{\boldsymbol{y}}$, amplitude $\tilde{E}$, frequency ω, and k_z, the common factors that correspond to such properties can all be taken out of the sum, and the opposite k_x components lead to a real $\cos(k_x x)$ term in the complex field.

To calculate the intensity, which is the magnitude of the Poynting vector, we need the magnetic field as well. Eq. (3.27) suggests that the magnetic-field amplitudes for the plane waves are

$$\tilde{\boldsymbol{H}} = \frac{1}{Z}\hat{\boldsymbol{k}}\times\tilde{\boldsymbol{E}} = \frac{\tilde{E}}{Z}(-\cos\theta\hat{\boldsymbol{x}} + \sin\theta\hat{\boldsymbol{z}}), \tag{10.8}$$

$$\tilde{\boldsymbol{H}}' = \frac{1}{Z}\hat{\boldsymbol{k}}'\times\tilde{\boldsymbol{E}} = \frac{\tilde{E}}{Z}(-\cos\theta\hat{\boldsymbol{x}} - \sin\theta\hat{\boldsymbol{z}}), \tag{10.9}$$

where $Z = Z_0/n$ is now the impedance of the medium. Here the magnetic fields have equal x components but opposite z components. The total magnetic field becomes

$$\boldsymbol{H} = \tilde{\boldsymbol{H}}\exp(j\boldsymbol{k}\cdot\boldsymbol{r} - j\omega t) + \tilde{\boldsymbol{H}}'\exp(j\boldsymbol{k}'\cdot\boldsymbol{r} - j\omega t) \tag{10.10}$$

$$= \frac{2\tilde{E}}{Z}e^{jk_z z - j\omega t}\left[-\hat{\boldsymbol{x}}\cos\theta\cos(k_x x) + j\hat{\boldsymbol{z}}\sin\theta\sin(k_x x)\right], \tag{10.11}$$

which is a bit more complicated. Since the fields are monochromatic and oscillate in time as $e^{-j\omega t}$, we can appeal to Eq. (4.24) to calculate the time-averaged Poynting vector. The result is

$$\bar{\boldsymbol{S}} = \frac{1}{2}\operatorname{Re}(\boldsymbol{E}\times\boldsymbol{H}^*) = \frac{2|\tilde{E}|^2}{Z}\hat{\boldsymbol{z}}\cos\theta\cos^2(k_x x), \tag{10.12}$$

$$I \equiv |\bar{\boldsymbol{S}}| = \frac{2|\tilde{E}|^2}{Z}\cos\theta\cos^2(k_x x), \tag{10.13}$$

where the $j\sin(k_x x)$ term in Eq. (10.11) doesn't appear in the final result because it is imaginary in $\boldsymbol{E}\times\boldsymbol{H}^*$ and goes away when taking the real part.

The important point about Eq. (10.13) is that the intensity is proportional to $\cos^2(k_x x)$, which is periodic along x with a period of

$$\boxed{\Lambda \equiv \text{Period of intensity pattern} = \frac{\pi}{k_x} = \frac{\lambda_0}{2n\sin\theta}} \tag{10.14}$$

with respect to the free-space wavelength λ_0, refractive index n, and the angle θ.[1] Note also the factor of 2; although the electric field is proportional to $\cos(k_x x)$ and has a period of $2\pi/k_x$, here we are interested in the intensity, which is $\propto\cos^2(k_x x)$ and has a period of π/k_x, as plotted in Fig. 10.2.

This periodic intensity results from the **interference** between the two plane waves, as the intensity maxima correspond to points where the two plane waves are in phase ($\exp(jk_x x) = \exp(-jk_x x)$) and add up constructively, while the intensity minima correspond to points where the waves are π out of phase ($\exp(jk_x x) = -\exp(-jk_x x)$) and the intensity is zero. The periodic pattern is also called **interference fringes**.

In practice, the fringes can be observed by overlapping two laser beams at an angle and putting a camera in the overlapping region, as shown in Fig. 10.3.

- **Exercise**: Calculate the intensity if the second plane wave has an electric field amplitude $\tilde{\boldsymbol{E}}' = \tilde{\boldsymbol{E}}e^{j\phi}$.
- **Exercise**: Calculate the intensity if the second plane wave has an electric field amplitude $\tilde{\boldsymbol{E}}' = \cos\theta\hat{\boldsymbol{x}} + \sin\theta\hat{\boldsymbol{z}}$, which is perpendicular to $\tilde{\boldsymbol{E}}$.

[1] Do not confuse the period of the intensity pattern Λ with the wavelength λ, the latter of which is the period of a plane wave in the medium.

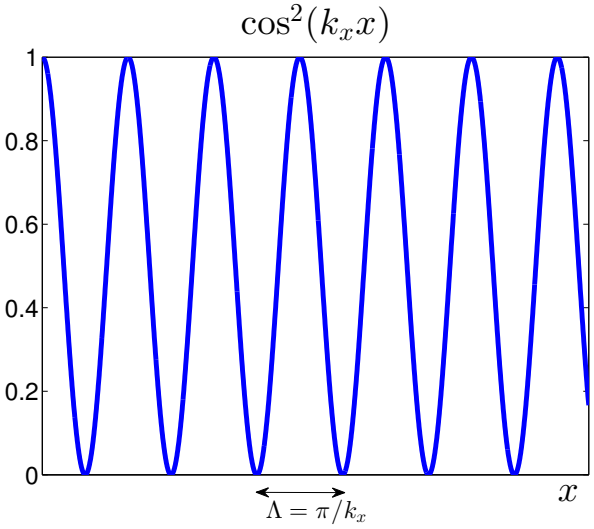


FIGURE 10.2. Plot of $\cos^2(k_x x)$. Note the period.

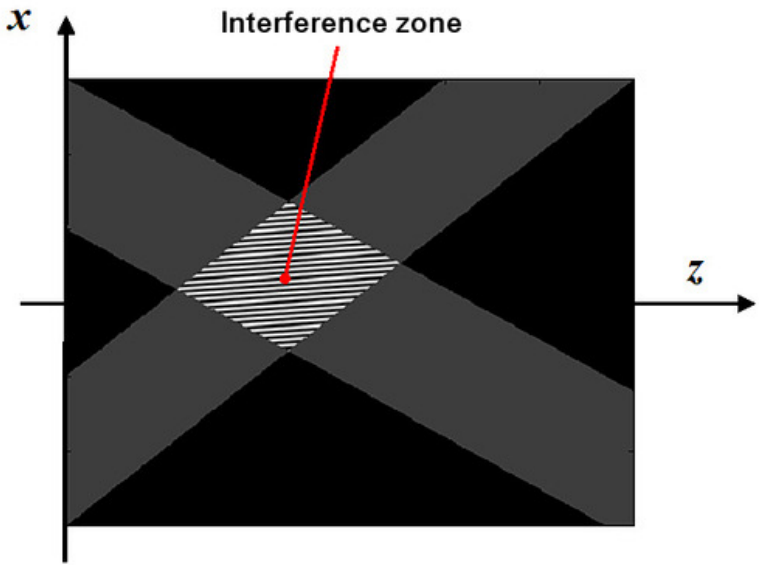


FIGURE 10.3. Two laser beams that overlap in space will result in a periodic intensity pattern depending on the angle between them.

10.2. Application: interferometric lithography

Optical lithography or photolithography, is one of the most important tools in semiconductor fabrication (http://en.wikipedia.org/wiki/Photolithography), integrated circuits (IC) in particular, and a multi-billion-dollar business (see, for example, http://physicsworld.com/cws/article/news/2012/jul/26/intel-invests-in-extreme-ultraviolet-lithography). The basic principle is to shine a spatially varying optical intensity pattern on a chemical film called **photoresist** on top of a semiconductor substrate. The photoresist will undergo chemical changes depending on the intensity and can be chemically treated to produce holes in the photoresist, and further etching through the holes can produce the desired semiconductor structures.

The spatially varying optical intensity obtained by overlapping and interfering multiple plane waves, in which case the technique is called **interferometric lithography**. The simplest pattern is the periodic intensity pattern studied earlier, which can be made by just two plane waves at an angle. The **feature size**, or the period of this pattern Λ, is then given by Eq. (10.14). To make smaller features, Λ should be smaller, meaning that

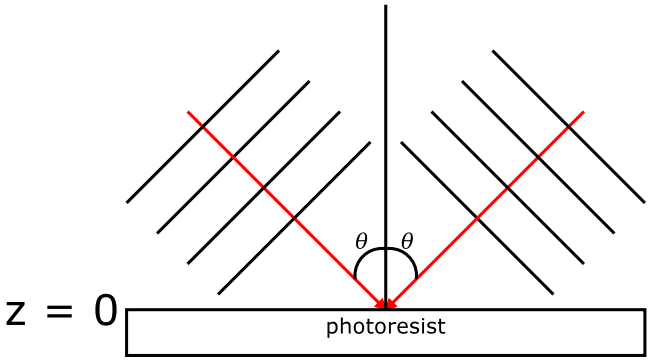


FIGURE 10.4. Two plane waves arriving at an angle at the photoresist can produce a periodic intensity pattern with period given by Eq. (10.14). The red arrows denote wavevectors and the black lines denote wavefronts, which are constant-phase surfaces perpendicular to the wavevectors, to be discussed later.

it is desirable to **make the free-space wavelength λ_0 from the source small**, **the refractive index of the medium n high**, and **the angle θ to be as large as possible**. Current lithographic technology that is used to make your computer chips uses a $\lambda_0 = 193$ nm laser source, and immersion of the optical system in water ($n \approx 1.44$ at $\lambda_0 = 193$ nm). The limit of resolution by the wavelength is one of the biggest bottlenecks to computer technology and Moore's law currently, as it is increasingly difficult to make good laser sources, optical components, and photoresists that work at smaller wavelengths. Extreme-ultraviolet lithography ($\lambda_0 \approx 20$ nm) is supposedly the next-generation technology, but it has not become commercial yet.

10.3. *Fourier optics

*To study more complicated intensity patterns, at this point we have two possible routes. One follows more naturally from our previous discussions and studies what happens when we have **multiple plane waves**. In fact, we can pass to the continuous limit and ask what happens when we have a superposition of infinite plane waves with different wavevectors. This is the approach of **Fourier optics** [2]. It involves **Fourier transform** as the central mathematical tool. To give you a rough idea, instead of having the superposition of a discrete number of plane waves, like

$$\sum_m \tilde{\boldsymbol{E}}_m \exp(j\boldsymbol{k}_m \cdot \boldsymbol{r} - j\omega t), \tag{10.15}$$

in Fourier optics we study something like

$$\int_{-\infty}^{\infty} \frac{dk_x}{2\pi} \int_{-\infty}^{\infty} \frac{dk_y}{2\pi} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \tilde{\boldsymbol{E}}(k_x, k_y, \omega) \exp(j\boldsymbol{k} \cdot \boldsymbol{r} - j\omega t), \tag{10.16}$$

where $k_z = \sqrt{\omega^2 n^2/c^2 - k_x^2 - k_y^2}$. This is an **inverse Fourier transform** and implies that any spatially varying field can be written as a **superposition of plane waves**.

10.4. Spherical wave from a point source

Unfortunately, Fourier transform is a bit too advanced for a level-2 module, and instead we will use a more qualitative and heuristic approach to study the propagation of spatially varying fields. Instead of plane waves, we will start by studying waves from a very small source with size much smaller than the wavelength of the medium. Such a source is called a point source. Later we will add these waves together to study interference and add a continuous superposition of them together to study diffraction.

In three space dimensions, a monochromatic point source will produce a **spherical wave**. The way to solve it is to consider the wave equation given by Eq. (3.6) in spherical coordinates. We shall not bother with the derivation, which can be found in, for example, [4]. I will simply tell you an approximation of the full solution (http://en.wikipedia.org/wiki/Dipole#Dipole_radiation). We make the following approximations:

- The point source is monochromatic (oscillating as $e^{-j\omega t}$) and located at the origin $\boldsymbol{r}_0 = 0$.

- The position $\boldsymbol{r}$ where we measure the intensity is much farther away from the point source than the wavelength, that is, $r \gg \lambda$.
- We consider only position coordinates x and y with distances from the z axis that are much smaller than the propagation distance z ($z > 0$):

$$|x| \ll z, \qquad |y| \ll z, \tag{10.17}$$

that is, we consider only **position vectors that make small angles with the z axis**, as shown in Fig. 10.5. This is known as the **paraxial approximation**.

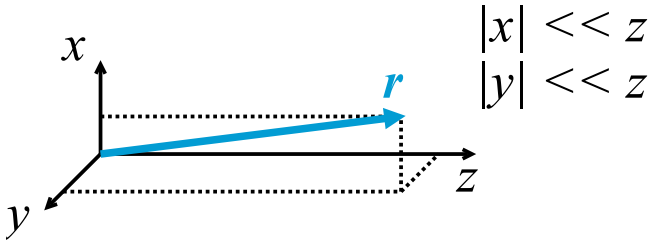


FIGURE 10.5. Paraxial approximation for the position vector $\boldsymbol{r}$.

Under these approximations, an electromagnetic spherical wave solution is

$$\boldsymbol{E} \approx \hat{\boldsymbol{x}} \frac{\tilde{E}}{r} \exp(jkr - j\omega t), \qquad r = \sqrt{x^2 + y^2 + z^2}, \tag{10.18}$$

with a magnetic field given by

$$\boldsymbol{H} \approx \frac{1}{Z} \hat{\boldsymbol{z}} \times \boldsymbol{E}. \tag{10.19}$$

The animation at http://en.wikipedia.org/wiki/Dipole#Dipole_radiation may help visualizing a spherical wave. Note the $\exp(jkr)$ phase factor for a spherical wave has **the same period in all directions from the origin** and is very different from the $\exp(j\boldsymbol{k} \cdot \boldsymbol{r})$ factor for a plane wave, despite their similar appearances. The spherical wave also decays as $1/r$ as we move further away from the source. Furthermore, by making the paraxial approximation we assume that the wave is *almost* like a plane wave close to the z axis, such that the electric field unit vector $\hat{\boldsymbol{x}}$ and the magnetic field unit vector $\hat{\boldsymbol{y}}$ are perpendicular to the z axis.

10.5. Wavefronts and rays

A common way of drawing a wave is to draw surfaces of constant phases of the fields, as shown in Fig. 10.6 for a spherical wave. These surfaces are called **wavefronts**. On each wavefront, the fields all have the same phase and thus the same real value. At some later instant of time, a spherical wavefront from a point source will expand to a bigger sphere, where the fields remain to have the same phase but amplitude dropping like $1/r$. We can also draw **rays** that are perpendicular to the wavefronts to denote the propagation direction of the phase. The rays are often parallel to the time-averaged Poynting vector, although this is not true in general. For plane waves, wavefronts and rays are shown on the right of Fig. 10.6.

The rays happen to be parallel to the time-averaged Poynting vector for exact spherical waves, but Eqs. (10.18) and (10.19) suggest that the Poynting vector is in the $\hat{\boldsymbol{z}}$ direction and not exactly in the directions of the rays shown in Fig. 10.6. This is because we have assumed $x \ll z$ and $y \ll z$, and the Poynting vectors and also the rays are almost in the $\hat{\boldsymbol{z}}$ direction.

10.6. Paraxial approximation

Let us consider a fixed positive propagation distance z away from the point source. Assume $y = 0$ for simplicity. The spherical wave as a function of x depends on r, which is of course

$$r = \sqrt{x^2 + z^2}. \tag{10.20}$$

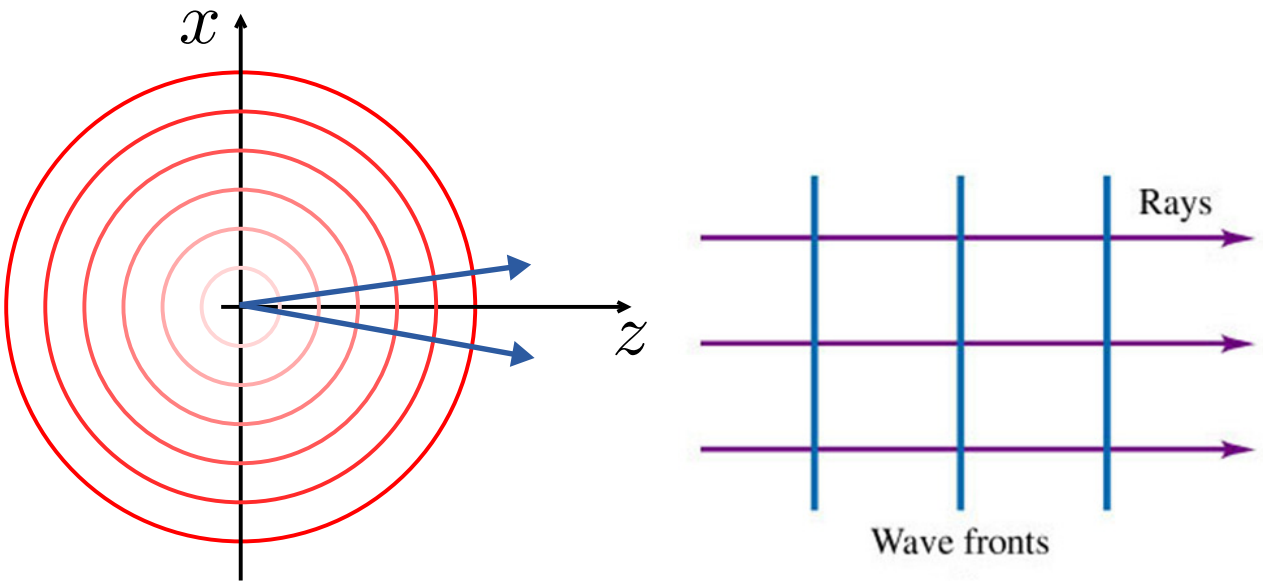


FIGURE 10.6. Left: Spherical wavefronts from a point source at origin at an instant of time. The fields all have the same phase and therefore the same real value on each sphere at an instant of time. These surfaces are called wavefronts. We can also draw rays perpendicular to the wavefronts to denote the propagation direction of the wavefronts. At some later time, each wavefront will expand to a bigger sphere, where the fields remain to have the same phase. Right: for a plane wave, wavefronts are planes and rays are parallel to the wavevector.

Since we already assumed $|x| \ll z$, we can make a further approximation. Let us rewrite r as

$$r = \sqrt{x^2 + z^2} = z\sqrt{1 + \frac{x^2}{z^2}}. \tag{10.21}$$

With $x^2/z^2 \ll 1$, we can approximate the square root up to the first order of the Taylor series ($\sqrt{1+\epsilon} \approx 1+\epsilon/2$, with $\epsilon \ll 1$, see Sec. 10.6.1):

$$\sqrt{1 + \frac{x^2}{z^2}} \approx 1 + \frac{x^2}{2z^2}. \tag{10.22}$$

This means that

$$r = z\sqrt{1 + \frac{x^2}{z^2}} \approx z\left(1 + \frac{x^2}{2z^2}\right) = z + \frac{x^2}{2z}, \tag{10.23}$$

as shown in Fig. 10.7.

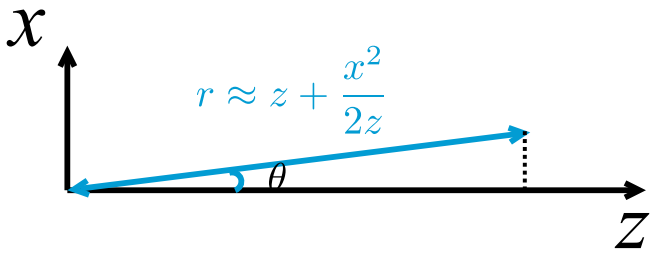


FIGURE 10.7. Paraxial approximation of distance r from the origin to the point (x, z) when $|x| \ll z$. Paraxial approximation also means that the angle θ is small.

10.6.1. Taylor series. The Taylor series (http://en.wikipedia.org/wiki/Taylor_series) is a way of expressing a function $f(\epsilon)$ as an infinite series and can be used to approximate a function near a

point if the higher-order terms are omitted. With respect to $\epsilon = 0$, it can be expressed as

$$f(\epsilon) = f(0) + \frac{df(\epsilon)}{d\epsilon}\Big|_{\epsilon=0}\epsilon + \frac{1}{2}\frac{d^2 f(\epsilon)}{d\epsilon^2}\Big|_{\epsilon=0}\epsilon^2 + \dots \tag{10.24}$$

If the ϵ^2 term is small, all the higher-order terms are usually small, and we can just keep the first two terms:

$$f(\epsilon) \approx f(0) + \frac{df(\epsilon)}{d\epsilon}\Big|_{\epsilon=0}\epsilon. \tag{10.25}$$

For example, in the paraxial approximation we have

$$f(\epsilon) = \sqrt{1+\epsilon}, \tag{10.26}$$

where $\epsilon = x^2/z^2 \ll 1$. We would like to approximate $\sqrt{1+\epsilon}$ near $f(\epsilon = 0) = 1$. The first derivative is

$$\frac{df(\epsilon)}{d\epsilon} = \frac{1}{2\sqrt{1+\epsilon}}. \tag{10.27}$$

Evaluated at $\epsilon = 0$,

$$\frac{df(\epsilon)}{d\epsilon}\Big|_{\epsilon=0} = \frac{1}{2\sqrt{1+\epsilon}}\Big|_{\epsilon=0} = \frac{1}{2}. \tag{10.28}$$

Therefore

$$f(\epsilon) \approx f(0) + \frac{df(\epsilon)}{d\epsilon}\Big|_{\epsilon=0}\epsilon = 1 + \frac{1}{2}\epsilon, \tag{10.29}$$

which is what we used earlier.

Consider now what happens if the point source is shifted up the x axis from the origin, and the new position is

$$\boldsymbol{r}_0 = x_0\hat{\boldsymbol{x}}, \tag{10.30}$$

that is, it is at $x = x_0$, $z = 0$. The field is now

$$\boldsymbol{E} \approx \hat{\boldsymbol{x}}\tilde{E}\frac{\exp\left(jk|\boldsymbol{r} - \boldsymbol{r}_0| - j\omega t\right)}{|\boldsymbol{r} - \boldsymbol{r}_0|}, \tag{10.31}$$

with the magnetic field still given by Eq. (10.19). The wavefronts now look like Fig. 10.8.

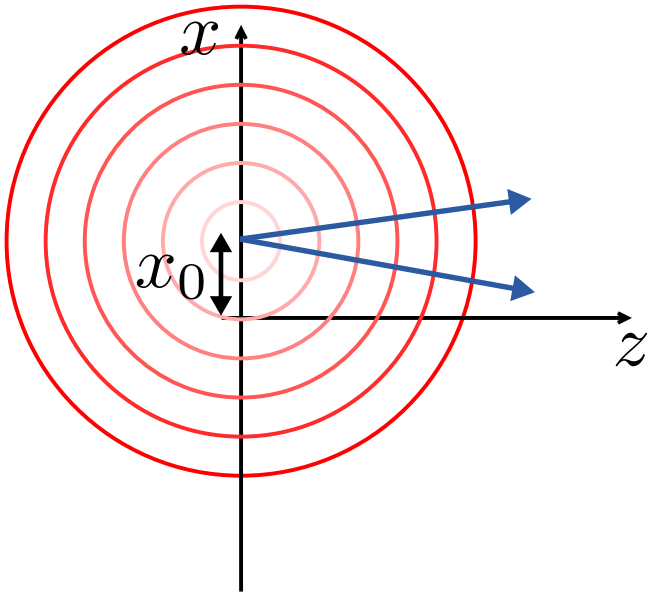


FIGURE 10.8. Spherical wave for a point source shifted from the origin.

The distance between the source and the position at propagation distance z is

$$|\boldsymbol{r} - \boldsymbol{r}_0| = \sqrt{(x - x_0)^2 + z^2}, \tag{10.32}$$

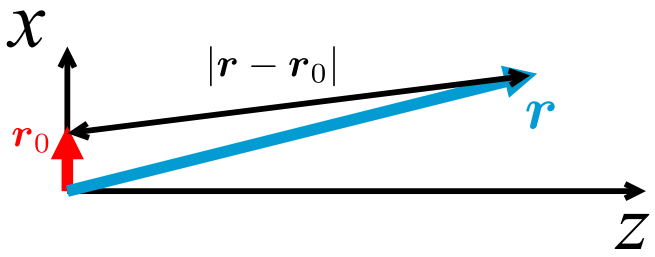


FIGURE 10.9. Distance between a point source at $\boldsymbol{r}_0$ and the position vector $\boldsymbol{r}$.

as shown in Fig. 10.9 (this comes from Pythagoras theorem if you don't remember how to calculate the magnitude of a vector). We shall now make a paraxial approximation for $\boldsymbol{r}-\boldsymbol{r}_0$:

$$|x-x_0| \ll z. \tag{10.33}$$

This enables us to assume

$$|\boldsymbol{r}-\boldsymbol{r}_0| \approx z + \frac{(x-x_0)^2}{2z}. \tag{10.34}$$

- **Exercise**: Make the paraxial approximation for Sec. 10.1 by assuming $\theta \ll 1$ and check that that $\boldsymbol{H} \approx \frac{1}{Z}\hat{\boldsymbol{z}} \times \boldsymbol{E}$.

10.7. Spherical wave interference

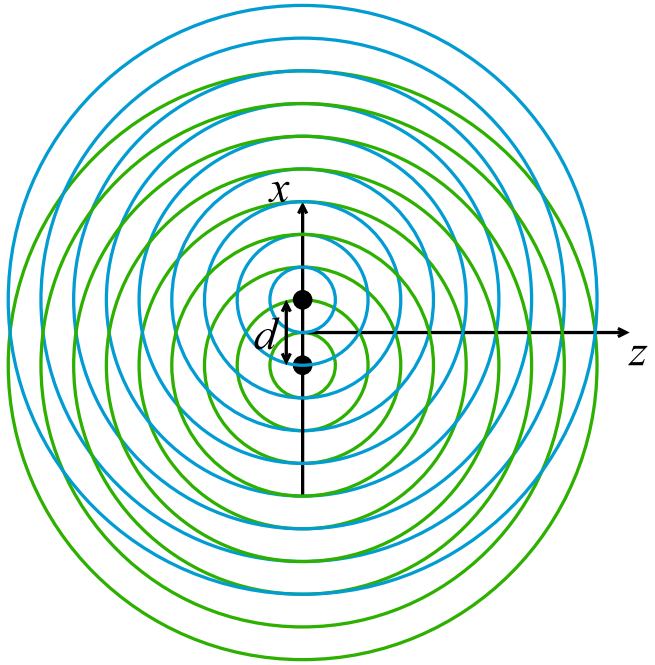


FIGURE 10.10. Interference of two spherical waves from two point sources separated by d. The concentric circles represent wavefronts for the spherical waves.

So far we have done some warm-up calculations about the paraxial approximation. What if there are two spherical waves from two point sources separated by a distance d, as shown in Fig. 10.10? There will be different places where the two waves interfere constructively and where the waves interfere destructively, although the positions will be somewhat different from the case of two plane waves studied in Sec. 10.1.

If the positions of the point sources are

$$\boldsymbol{r}_0 = \frac{d}{2}\hat{\boldsymbol{x}}, \qquad\qquad \boldsymbol{r}_1 = -\frac{d}{2}\hat{\boldsymbol{x}}, \tag{10.35}$$

the electric field is the superposition of spherical waves from the two sources:

$$\boldsymbol{E} \approx \hat{\boldsymbol{x}} \frac{\tilde{E}}{|\boldsymbol{r}-\boldsymbol{r}_0|} \exp(jk|\boldsymbol{r}-\boldsymbol{r}_0| - j\omega t) + \hat{\boldsymbol{x}} \frac{\tilde{E}}{|\boldsymbol{r}-\boldsymbol{r}_1|} \exp(jk|\boldsymbol{r}-\boldsymbol{r}_1| - j\omega t), \tag{10.36}$$

assuming that they have the same amplitude, polarization, and frequency. The distances are shown in Fig. 10.11.

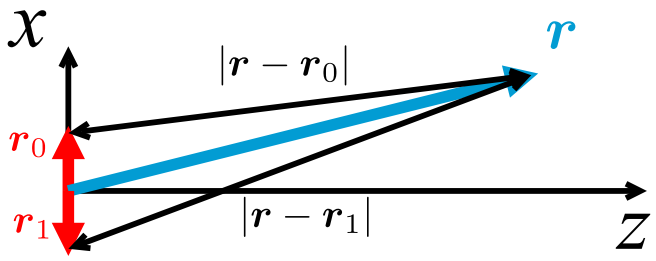


FIGURE 10.11. The field at position $\boldsymbol{r}$ is a superposition of a wave from position $\boldsymbol{r}_0 = (d/2)\hat{\boldsymbol{x}}$ and another wave from position $\boldsymbol{r}_1 = (-d/2)\hat{\boldsymbol{x}}$.

Once again we will consider $y = 0$. To simplify further, we assume that the denominators are proportional to z due to the paraxial approximation:

$$\frac{1}{|\boldsymbol{r}-\boldsymbol{r}_0|} \approx \frac{1}{|\boldsymbol{r}-\boldsymbol{r}_1|} \approx \frac{1}{z}, \tag{10.37}$$

and the x dependence is assumed to be in the phases only. For the phases, we use Eq. (10.34) to write[2]

$$\exp\left(jk|\boldsymbol{r}-\boldsymbol{r}_0|\right) \approx \exp\left\{jk\left[z + \frac{(x-d/2)^2}{2z}\right]\right\}. \tag{10.38}$$

We also make a similar approximation for

$$\exp\left(jk|\boldsymbol{r}-\boldsymbol{r}_1|\right) \approx \exp\left\{jk\left[z + \frac{(x+d/2)^2}{2z}\right]\right\}, \tag{10.39}$$

where we replace x_0 with $d/2$ in Eq. (10.34) for the position of one source and replace x_0 with $-d/2$ for the other source. The electric field then becomes

$$\boldsymbol{E} \approx \left\{\hat{\boldsymbol{x}}\frac{\tilde{E}}{z}e^{jkz}\exp\left[\frac{jk(x-d/2)^2}{2z}\right]e^{-j\omega t} + \hat{\boldsymbol{x}}\frac{\tilde{E}}{z}e^{jkz}\exp\left[\frac{jk(x+d/2)^2}{2z}\right]e^{-j\omega t}\right\}, \tag{10.40}$$

There are a lot of common factors to be taken out, resulting in

$$\boldsymbol{E} \approx \hat{\boldsymbol{x}}\frac{\tilde{E}}{z}e^{jkz-j\omega t}\exp\left\{\frac{jk[x^2+(d/2)^2]}{2z}\right\}\left[\exp\left(\frac{-jkxd}{2z}\right) + \exp\left(\frac{jkxd}{2z}\right)\right] \tag{10.41}$$

$$= \hat{\boldsymbol{x}}\frac{2\tilde{E}}{z}e^{jkz-j\omega t}\exp\left\{\frac{jk[x^2+(d/2)^2]}{2z}\right\}\cos\left(\frac{kxd}{2z}\right). \tag{10.42}$$

The only remaining difference between the two waves is two phase factors with opposite phases. Constructive interference occurs where the two terms are equal:

$$\exp\left(\frac{-jkxd}{2z}\right) = \exp\left(\frac{jkxd}{2z}\right), \tag{10.43}$$

so they add up to produce twice the electric field of each individual spherical wave. This happens when

$$\exp\left(-\frac{jkxd}{z}\right) = 1, \qquad \frac{kxd}{z} = 2\pi m, \quad m = \text{integer}. \tag{10.44}$$

Substituting $k = 2\pi/\lambda$, The points of constructive interference are

$$x = m\frac{\lambda z}{d}, \quad m = \text{integer}. \tag{10.45}$$

[2]There is a better reason in Fourier optics to justify why we just assume z for the denominator in the paraxial approximation but include more terms in the phases, but we won't bother with it here.

Destructive interference occurs where they are π out of phase such that they cancel each other and the electric field is zero:

$$\exp\left(\frac{-jkxd}{2z}\right) = -\exp\left(\frac{jkxd}{2z}\right), \tag{10.46}$$

or

$$\exp\left(-\frac{jkxd}{z}\right) = -1, \qquad \frac{kxd}{z} = \pi + 2\pi m, \quad m = \text{integer}, \qquad x = \left(m + \frac{1}{2}\right)\frac{\lambda z}{d}. \tag{10.47}$$

To calculate what we actually see in our eyes or a camera and observe the interference pattern more clearly, we need to compute the intensity, and that involves the magnetic field. To do this in a simpler way once and for all, observe that Eq. (10.42) is an example of an electric field that looks like

$$\boxed{\boldsymbol{E} \approx \hat{\boldsymbol{x}}\mathcal{E}(x,z)e^{jkz-j\omega t},} \tag{10.48}$$

where I have defined $\mathcal{E}(x,z)$ as the scalar envelope of the electric field over a phase factor $e^{jkz-j\omega t}$. The magnetic field then looks like

$$\boldsymbol{H} \approx \frac{1}{Z}\hat{\boldsymbol{z}} \times \boldsymbol{E} \approx \hat{\boldsymbol{y}}\frac{1}{Z}\mathcal{E}(x,z)e^{jkz-j\omega t}. \tag{10.49}$$

We can use Eq. (4.24) to compute the time-averaged Poynting vector and the intensity:

$$\bar{\boldsymbol{S}} = \frac{1}{2}\operatorname{Re}\left(\boldsymbol{E} \times \boldsymbol{H}^*\right) \approx \frac{\hat{\boldsymbol{z}}}{2Z}|\mathcal{E}(x,z)|^2, \tag{10.50}$$

$$\boxed{I(x,z) \approx \frac{1}{2Z}\left|\mathcal{E}(x,z)\right|^2,} \tag{10.51}$$

which is proportional to the magnitude squared of the envelope. In the following, **whenever we see an electric field of the form in Eq. (10.48), the magnetic field will be given by Eq. (10.49), and we can just use Eq. (10.51) to compute the intensity.** All these are possible because of the paraxial approximation.

Comparing Eq. (10.42) and Eq. (10.48), we can write

$$\mathcal{E}(x,z) = \frac{2\tilde{E}}{z}\exp\left\{\frac{jk[x^2 + (d/2)^2]}{2z}\right\}\cos\left(\frac{kxd}{2z}\right). \tag{10.52}$$

The intensity is hence

$$I(x,z) \approx \frac{1}{2Z}\left|\mathcal{E}(x,z)\right|^2 = \frac{2|\tilde{E}|^2}{Zz^2}\cos^2\left(\frac{kxd}{2z}\right) \propto \cos^2\left(\frac{\pi d}{\lambda z}x\right), \tag{10.53}$$

where I have used $k = 2\pi/\lambda$. The period of the intensity interference fringes is

$$\boxed{\Lambda = \frac{\lambda z}{d},} \tag{10.54}$$

which is **proportional to the wavelength λ and also the propagation distance** z. The more nontrivial feature of the period is that it is **inversely proportional to d, the separation of the two sources**. The larger the separation, the smaller the period, and vice versa. This inverse relation between the source feature size and output intensity feature size will appear time and again when we study diffraction. A nice demo of spherical wave interference is at `http://phet.colorado.edu` (PhET/sims/wave-interference).

The intensity patterns ends up being similar to interference fringes from two plane waves. With hindsight, this is perhaps not so surprising, because a spherical wave far away looks somewhat like a plane wave with wavevector at an angle $\sin\theta \approx \theta \approx \pm d/(2z)$, if we compare the figures in Fig. 10.6 and Fig. 10.8.

10.8. Cylindrical waves from slits

Instead of point sources, we shall now consider a plane wave passing through a narrow slit at position $\boldsymbol{r}_0 = x_0\hat{\boldsymbol{x}}$ with width

$$\Delta x \ll \lambda, \tag{10.55}$$

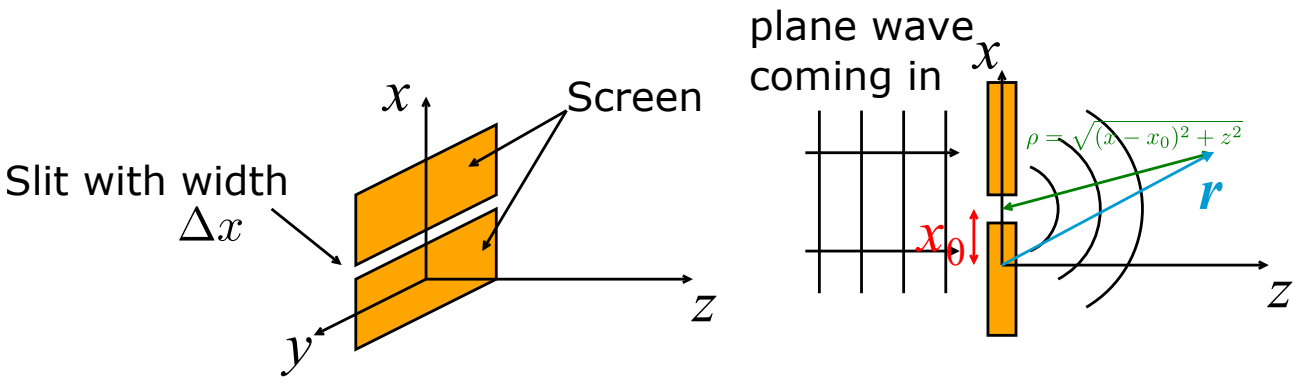


FIGURE 10.12. Cylindrical wave from a slit. The resulting fields are uniform along y.

as shown in Fig. 10.12. Again the derivation of the resulting cylindrical wave is quite involved, but with the help of paraxial approximation it can be shown that

$$\boldsymbol{E} \approx \hat{\boldsymbol{x}} \frac{\tilde{E}\Delta x}{\sqrt{\rho}} \exp(jk\rho - j\omega t), \qquad \rho \equiv \sqrt{(x-x_0)^2+z^2}, \tag{10.56}$$

where ρ is now a type of cylindrical coordinate that does not depend on y and the fields are **uniform across** y, and we won't make the $y = 0$ assumption any more. This **cylindrical wave** decays as $1/\sqrt{\rho}$ and has cylindrical wavefronts. We can calculate the interference pattern if we again assume the paraxial approximation, and it looks almost the same as the spherical case (with $y = 0$).

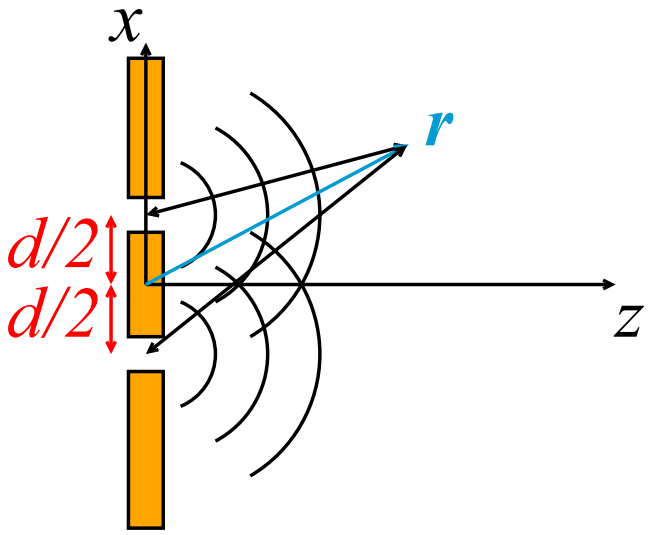


FIGURE 10.13. Geometry of the two-slit interference experiment.

- **Exercise** (Young's double-slit interference, Fig. 10.13): Using the paraxial approximation, calculate the intensity for cylindrical waves from two slits at $z = 0$ separated by a distance d. Assume that the waves have the same amplitude.
  Answer: Suppose that one of the sources is at $x_0 = d/2$, with distance between the source and the position being $\sqrt{(x-x_0)^2+z^2}$. Suppose that the other source is at $x_1 = -d/2$, with distance from this other source to the same distance being $\sqrt{(x-x_1)^2+z^2} = \sqrt{(x+d/2)^2+z^2}$. The total

electric field is

$$\boldsymbol{E} \approx \hat{\boldsymbol{x}} \frac{\tilde{E}\Delta x}{[(x-d/2)+z^2]^{1/4}} \exp\left[jk\sqrt{(x-d/2)^2+z^2} - j\omega t\right] + \hat{\boldsymbol{x}} \frac{\tilde{E}\Delta x}{[(x+d/2)+z^2]^{1/4}} \exp\left[jk\sqrt{(x+d/2)^2+z^2} - j\omega t\right]. \tag{10.57}$$

Similar to the spherical case, we assume that the denominators are

$$\frac{1}{[(x-d/2)+z^2]^{1/4}} \approx \frac{1}{\sqrt{z}}, \tag{10.58}$$

$$\frac{1}{[(x+d/2)+z^2]^{1/4}} \approx \frac{1}{\sqrt{z}}, \tag{10.59}$$

and assume the following for the phases:

$$\sqrt{(x-d/2)^2+z^2} \approx z + \frac{(x-d/2)^2}{2z}, \tag{10.60}$$

$$\sqrt{(x+d/2)^2+z^2} \approx z + \frac{(x+d/2)^2}{2z}. \tag{10.61}$$

The electric field then becomes

$$\boldsymbol{E} \approx \hat{\boldsymbol{x}} \frac{\tilde{E}\Delta x}{\sqrt{z}} e^{jkz-j\omega t} \left\{ \exp\left[\frac{jk(x-d/2)^2}{2z}\right] + \exp\left[\frac{jk(x+d/2)^2}{2z}\right] \right\}. \tag{10.62}$$

This has the form of Eq. (10.48), so we can just write

$$\mathcal{E}(x,z) = \frac{\tilde{E}\Delta x}{\sqrt{z}} \left\{ \exp\left[\frac{jk(x-d/2)^2}{2z}\right] + \exp\left[\frac{jk(x+d/2)^2}{2z}\right] \right\}. \tag{10.63}$$

Taking the common factors out,

$$\mathcal{E}(x,z) = \frac{\tilde{E}\Delta x}{\sqrt{z}} \exp\left\{\frac{jk[x^2+(d/2)^2]}{2z}\right\} \left[\exp\left(\frac{-jkxd}{2z}\right) + \exp\left(\frac{jkxd}{2z}\right)\right] \tag{10.64}$$

$$= \frac{\tilde{E}\Delta x}{\sqrt{z}} \exp\left\{\frac{jk[x^2+(d/2)^2]}{2z}\right\} 2\cos\left(\frac{kxd}{2z}\right). \tag{10.65}$$

Now we can use Eq. (10.51):

$$I(x,z) = \frac{2|\tilde{E}|^2(\Delta x)^2}{Zz} \cos^2\left(\frac{kxd}{2z}\right). \tag{10.66}$$

We can also write $k = 2\pi/\lambda$ such that

$$I(x,z) = \frac{2|\tilde{E}|^2(\Delta x)^2}{Zz} \cos^2\left(\frac{\pi d}{\lambda z}x\right), \tag{10.67}$$

which also has the period given by Eq. (10.54). A nice demo of the interference pattern can be found at `http://phet.colorado.edu/` (PhET/sims/wave-interference).

In general, we can have multiple narrow slits, as shown in Fig. 10.14. Again, assume that all the slits are at $z = 0$, except that their x coordinates are different. Assume that the slits are at $x_0, x_1, \ldots, x_{M-1}$ (total of M slits), and they all have width Δx. The resulting wave will be

$$\boldsymbol{E} \approx \hat{\boldsymbol{x}} \tilde{E}\Delta x \sum_{m=0}^{M-1} \frac{\exp(jk\sqrt{z^2+(x-x_m)^2} - j\omega t)}{[z^2+(x-x_m)^2]^{1/4}}, \tag{10.68}$$

and if we make the paraxial approximation,

$$\frac{(x-x_m)^2}{z^2} \ll 1, \qquad \mathcal{E}(x,z) = \frac{\tilde{E}\Delta x}{\sqrt{z}} \sum_{m=0}^{M-1} \exp\left[\frac{jk(x-x_m)^2}{2z}\right]. \tag{10.69}$$

This formula will be important for our discussion of diffraction.

- **Exercise**: Verify Eq. (10.69) using the paraxial approximation.

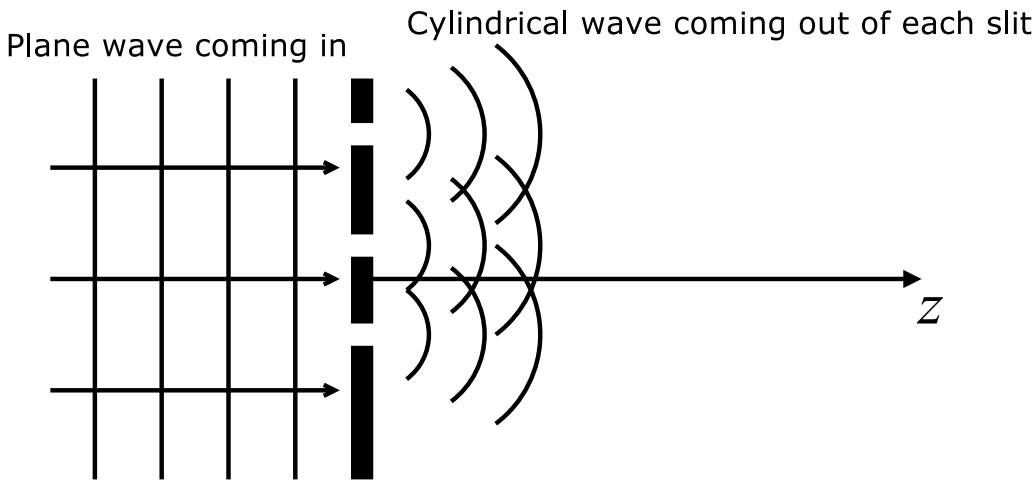


FIGURE 10.14. Multiple cylindrical waves from multiple slits.

CHAPTER 11

Diffraction

11.1. Fresnel diffraction of a wide slit

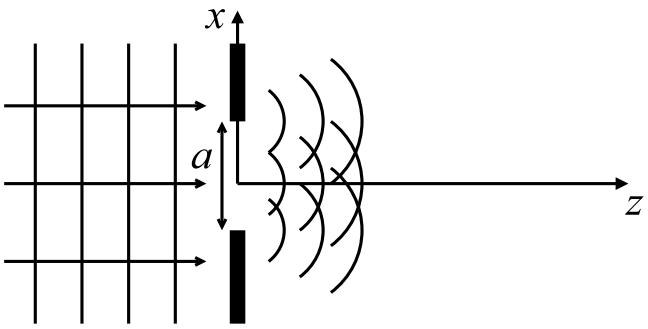


FIGURE 11.1. Diffraction of a wide slit. A plane wave impinges on the screen with a wide slit. To calculate the field $\boldsymbol{E}(x, z, t)$ after the slit, we imagine that the wide slit is made of infinite slits with infinitesimal width dx, and compute an integral over the position of each infinitesimal slit x'.

Instead of a narrow slit studied in Sec. 10.8, consider a wide slit, as shown in Fig. 11.1. The width of the slit is a, and suppose that it is at $z = 0$ and goes from $x = -a/2$ to $x = a/2$. What are the fields after the slit at a propagation distance z?

The key to solving this problem is to imagine the wide slit as **an infinite number of narrow slits with infinitesimal width** dx. We can then take the continous limit of the discrete sum for multiple slits in Eq. (10.69) and arrive at an **integral**:

$$\boxed{\mathcal{E}(x, z) = \frac{\tilde{E}}{\sqrt{z}} \int_{-a/2}^{a/2} dx' \exp\left[\frac{jk(x - x')^2}{2z}\right],} \tag{11.1}$$

where $\mathcal{E}(x, z)$ is defined as the envelope of the electric field in Eq. (10.48) and I have replaced x_m with a continuous variable x', **which denotes the position of each infinitesimal slit**. This integral is known as Fresnel diffraction of a wide slit. It is valid under the **paraxial approximation**, which we used to derive Eq. (10.69) and means that we are concerned only with points near the z axis.

The integral in Eq. (11.1) is still difficult to solve analytically, so we shall make a further approximation in the next section.

11.2. Fraunhofer diffraction

For an optical beam from a wide slit, Fresnel diffraction will make the beam wider and wider as it propagates along z. An example can be seen from a laser pointer; when the laser beam propagates for a very long distance z, the beam size will expand (See, for example, https://www.youtube.com/watch?v=uxB7kTinPgQ for a demonstration). When a beam expands so much, we can approximate its **large-scale features along** x at a certain z by considering a scale of x that is much larger than the slit width a, as depicted in Fig. 11.2.

When this Fraunhofer approximation is valid, the variable x' in the integral Eq. (11.1), which denotes the position of each infinitesimal slit within the wide slit, can be assumed to be much smaller than x. In other

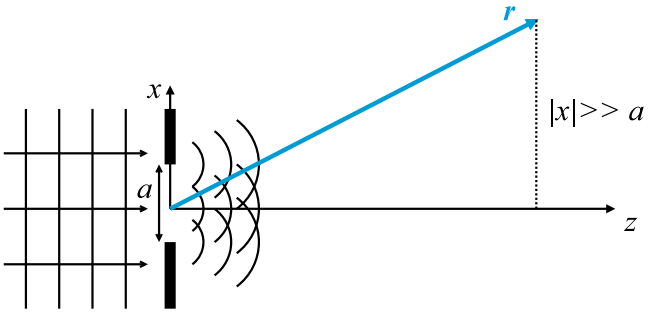


FIGURE 11.2. The Fraunhofer approximation considers features of the fields along x that have a much larger scale than the width of the slit a.

words, we assume

$$|x| \gg |x'|. \tag{11.2}$$

We can then approximate the $(x-x')^2$ term in Eq. (11.1) as

$$(x-x')^2 = x^2 - 2xx' + x'^2 \approx x^2 - 2xx', \tag{11.3}$$

since $|x'^2| \ll |2xx'|$ if $|x'| \ll |x|$. That leads to

$$\mathcal{E}(x,z) = \frac{\tilde{E}}{\sqrt{z}} \int_{-a/2}^{a/2} dx' \exp\left[\frac{jk(x^2-2xx')}{2z}\right]. \tag{11.4}$$

Taking the $\exp[jkx^2/(2z)]$ out of the integral since it doesn't depend on x', we obtain

$$\boxed{\mathcal{E}(x,z) = \frac{\tilde{E}}{\sqrt{z}} \exp\left(\frac{j\pi x^2}{\lambda z}\right) \int_{-a/2}^{a/2} dx' \exp\left(-\frac{j2\pi xx'}{\lambda z}\right),} \tag{11.5}$$

where we have substituted $k = 2\pi/\lambda$. This is the Fraunhofer diffraction of a wide slit. It is valid when the propagation distance z is so large that Eq. (11.2) is valid. To be concrete, let us work out the integral and say more about the $|x| \gg |x'|$ approximation later.

The following integral is with respect to x' and goes from $-a/2$ to $a/2$:

$$\int_{-a/2}^{a/2} dx' \exp\left(-\frac{j2\pi xx'}{\lambda z}\right) = \frac{1}{-j2\pi x/(\lambda z)} \exp\left(-\frac{j2\pi xx'}{\lambda z}\right)\Big|_{x'=-a/2}^{x'=a/2} \tag{11.6}$$

$$= \frac{1}{-j2\pi x/(\lambda z)} \left[\exp\left(-\frac{j\pi xa}{\lambda z}\right) - \exp\left(\frac{j\pi xa}{\lambda z}\right)\right] \tag{11.7}$$

$$= a\frac{\sin[\pi ax/(\lambda z)]}{\pi ax/(\lambda z)}. \tag{11.8}$$

Note that I have written it in such a way that both the numerator and the denominator depend on $\pi ax/(\lambda z)$. This is because I want to define it as a **sinc function**:

$$\operatorname{sinc} X \equiv \frac{\sin(\pi X)}{\pi X}. \tag{11.9}$$

The sinc function is 1 at $X = 0$ (via L'Hopital rule), goes up and down but dies away from the center, as plotted on the left of Fig. 11.3. The first zero of sinc on the left is at $X = -1$ and the first zero on the right is at $X = 1$.

The integral is hence

$$\int_{-a/2}^{a/2} dx' \exp\left(-\frac{j2\pi xx'}{\lambda z}\right) = a \operatorname{sinc}\left(\frac{ax}{\lambda z}\right), \tag{11.10}$$

the electric field envelope is

$$\mathcal{E}(x,z) = \frac{\tilde{E}}{\sqrt{z}} \exp\left(\frac{j\pi x^2}{\lambda z}\right) a \operatorname{sinc}\left(\frac{ax}{\lambda z}\right), \tag{11.11}$$

and the intensity according to Eq. (10.51) is

$$\boxed{I(x,z) \approx \frac{|\tilde{E}|^2 a^2}{2Zz}\,\mathrm{sinc}^2\left(\frac{ax}{\lambda z}\right).} \tag{11.12}$$

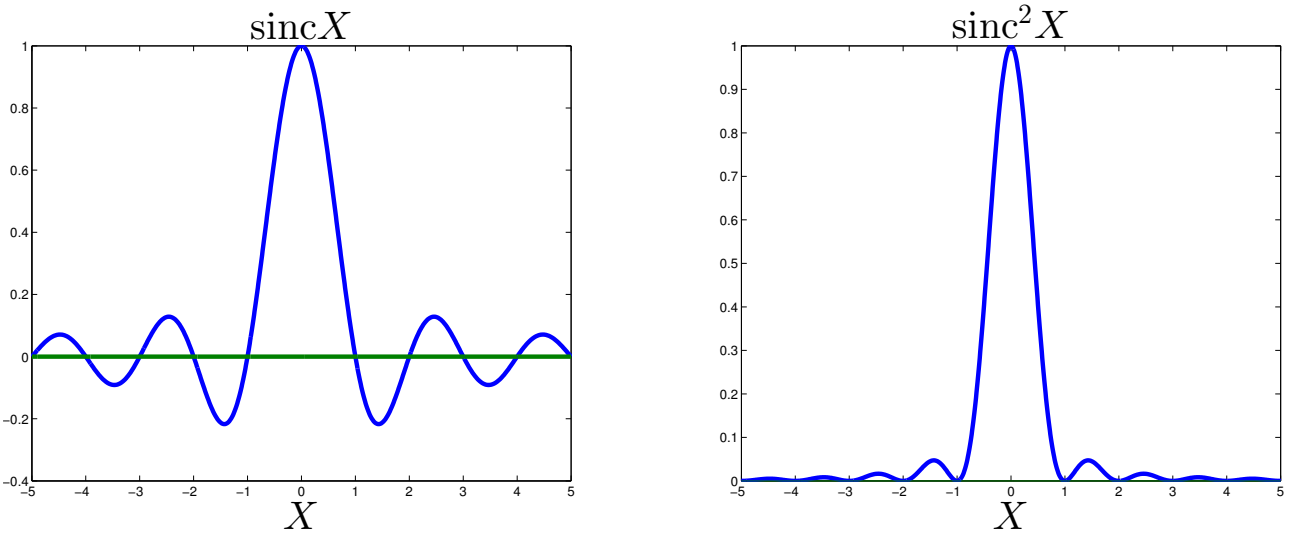


FIGURE 11.3. Plots of the sinc function (left) and the sinc^2 function (right).

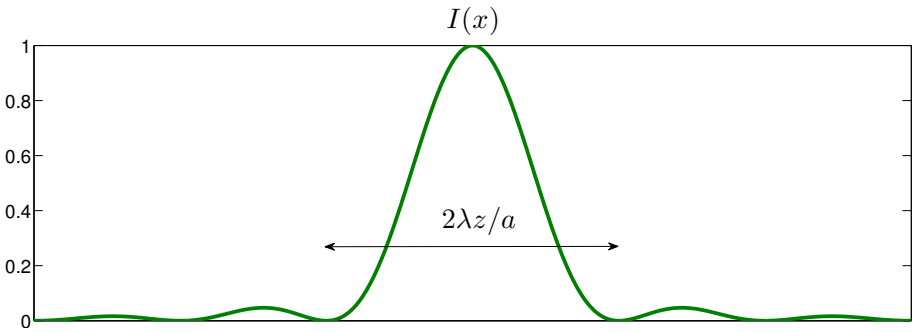


FIGURE 11.4. The Fraunhofer diffraction intensity $\propto \mathrm{sinc}^2[ax/(\lambda z)]$ for a wide slit. For a slit width a, wavelength λ, and propagation distance z, the width between the two first zeroes is $2\lambda z/a$.

The important point here is that the intensity is proportional to $\mathrm{sinc}^2[ax/(\lambda z)]$, which is plotted in Fig. 11.4. The first zero on the left is

$$X = \frac{ax}{\lambda z} = -1, \tag{11.13}$$

the first zero on the right is

$$X = \frac{ax}{\lambda z} = 1, \tag{11.14}$$

and if we define the width of the main peak as the distance between these two zeros, it becomes

$$\boxed{W \equiv \text{Width of main intensity peak} = \frac{2\lambda z}{a}.} \tag{11.15}$$

This width is **proportional to wavelength** λ, **proportional to propagation distance** z, and **inversely proportional to the width of the slit** a.[1] Note again the **inverse relation** between the feature size of the source (a) and the feature size of the diffraction pattern ($2\lambda z/a$). We can observe a similar behavior for spherical and cylindrical wave interference in Eq. (10.54). Fig. 11.5 shows an image of the actually observed intensity pattern.

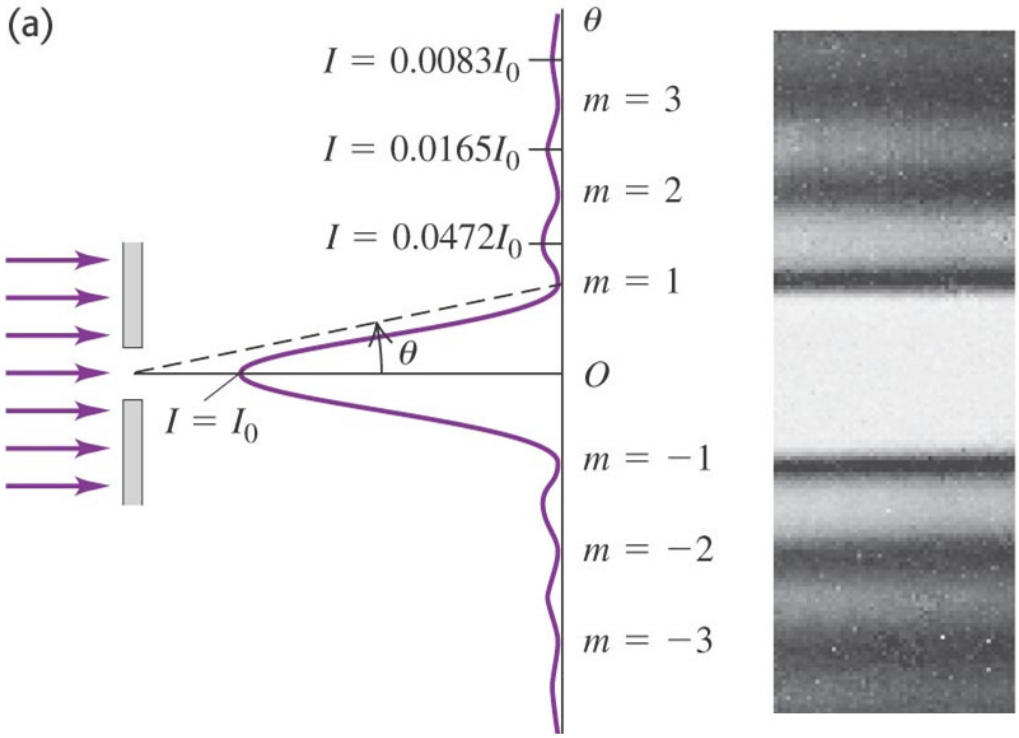


FIGURE 11.5. Fraunhofer diffraction intensity pattern (Image from [1]). Note that $\theta \approx \tan\theta \approx x/z$ because of the paraxial approximation.

- **Exercise**: How does the width of the Fraunhofer diffraction intensity peak change if I change the medium to a medium with higher refractive index, assuming everything else is the same?

11.3. Two wide slits

We are now ready to generalize the case of two narrow slits in Sec. 10.8 to two wide slits separated by a distance d. Consider first the top slit, which goes from $x' = d/2 - a/2$ to $x' = d/2 + a/2$. The Fraunhofer diffraction from it is then

$$\mathcal{E}_1(x,z) = \frac{\tilde{E}}{\sqrt{z}} \int_{d/2-a/2}^{d/2+a/2} dx' \exp\left(-\frac{j2\pi x x'}{\lambda z}\right). \tag{11.17}$$

[1] Now we can go back and verify that the $|x| \gg |x'|$ approximation in Eq. (11.2) is okay. Fraunhofer diffraction says that the field will vary with respect to x like the left of Fig. 11.3. The width of this pattern is much larger than the width of the slit if

$$\frac{2\lambda z}{a} \gg a, \text{ or } z \gg \frac{a^2}{2\lambda}. \tag{11.16}$$

This assumption of propagation distance should be satisfied for Fraunhofer diffraction to be valid. The $|x| \gg |x'|$ approximation doesn't hold for points very close to the center of the main peak, but we normally don't care about those when the peak is so wide.

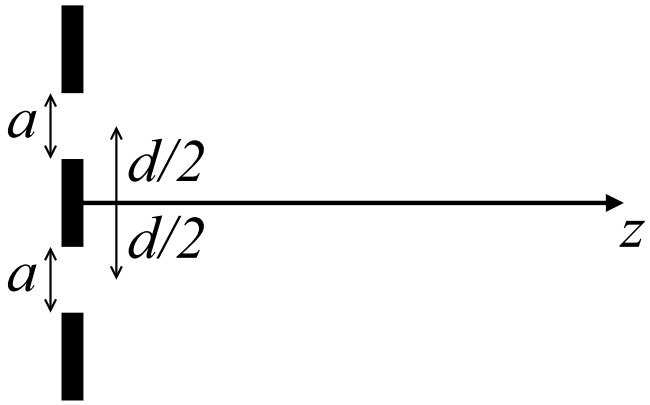


FIGURE 11.6.

The trick to do this is change to a new variable $x'' = x' - d/2$ for the integral, in which case the new limits become $-a/2$ to $a/2$ for x'' and $x' = x'' + d/2$:

$$\mathcal{E}_1(x,z) = \frac{\tilde{E}}{\sqrt{z}} \int_{-a/2}^{a/2} dx'' \exp\left[-\frac{j2\pi x(x''+d/2)}{\lambda z}\right] \tag{11.18}$$

$$= \frac{\tilde{E}}{\sqrt{z}} \exp\left(-\frac{j\pi x d}{\lambda z}\right) \int_{-a/2}^{a/2} dx'' \exp\left(-\frac{j2\pi x x''}{\lambda z}\right) \tag{11.19}$$

$$= \frac{\tilde{E}}{\sqrt{z}} \exp\left(-\frac{j\pi x d}{\lambda z}\right) a\,\mathrm{sinc}\left(\frac{ax}{\lambda z}\right). \tag{11.20}$$

The important point is that, when the slit position is shifted from the center, the diffracted field has an additional phase factor $\exp[-j\pi x d/(\lambda z)]$.

For Fraunhofer diffraction from the bottom slit, the procedure is similar, and we now integrate from $x' = -d/2 - a/2$ to $x' = -d/2 + a/2$:

$$\mathcal{E}_2(x,z) = \frac{\tilde{E}}{\sqrt{z}} \int_{-d/2-a/2}^{-d/2+a/2} dx' \exp\left(-\frac{j2\pi x x'}{\lambda z}\right) \tag{11.21}$$

$$= \frac{\tilde{E}}{\sqrt{z}} \exp\left(\frac{j\pi x d}{\lambda z}\right) a\,\mathrm{sinc}\left(\frac{ax}{\lambda z}\right). \tag{11.22}$$

Now the phase factor has the opposite phase because of the different slit position.

When we superimpose the waves from both slits, the sinc function is a common factor, while the phase factors combine to produce a cosine function:

$$\mathcal{E}(x,z) = \mathcal{E}_1(x,z) + \mathcal{E}_2(x,z) \tag{11.23}$$

$$= \frac{\tilde{E}}{\sqrt{z}} a\,\mathrm{sinc}\left(\frac{ax}{\lambda z}\right)\left[\exp\left(-\frac{j\pi x d}{\lambda z}\right) + \exp\left(\frac{j\pi x d}{\lambda z}\right)\right] \tag{11.24}$$

$$= \frac{\tilde{E}}{\sqrt{z}} a\,\mathrm{sinc}\left(\frac{ax}{\lambda z}\right) 2\cos\left(\frac{\pi x d}{\lambda z}\right). \tag{11.25}$$

The intensity is hence

$$\boxed{I(x,z) \approx \frac{2|\tilde{E}|^2 a^2}{Zz} \underbrace{\mathrm{sinc}^2\left(\frac{ax}{\lambda z}\right)}_{\text{envelope}} \underbrace{\cos^2\left(\frac{\pi d}{\lambda z}x\right)}_{\text{fringes}}}. \tag{11.26}$$

This is the multiplication of a $\cos^2$ pattern with period $\lambda z/d$ that we have seen for two narrow slits in Sec. 10.8 with the sinc^2 pattern with peak width $2\lambda z/a$ that we have seen for one wide slit in Sec. 11.2. Note that the width of each slit a has to be smaller than the distance d between them, so the sinc^2 pattern has a larger width

$2\lambda z/a$ than the period $\lambda z/d$ of $\cos^2$, as shown in Fig. 11.7. The sinc^2 function thus acts as an envelope for the interference fringes.

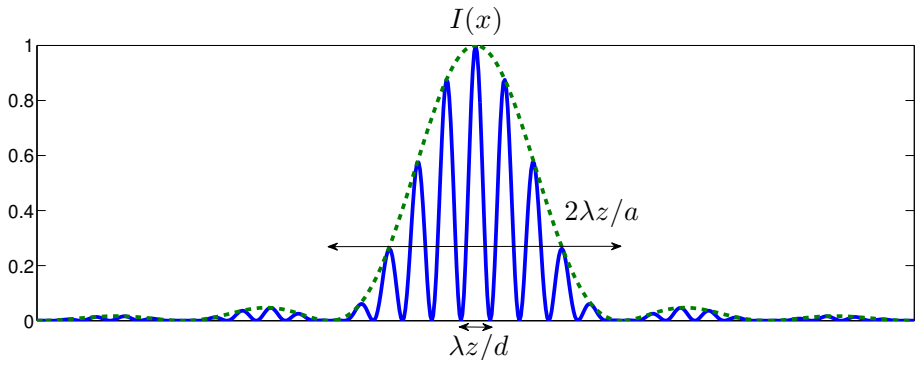


FIGURE 11.7. The Fraunhofer diffraction intensity for two wide slits. The dash line in green is the sinc^2 envelope.

- **Exercise**: Show that, as the width of the slits a becomes very small, the intensity pattern becomes that for two narrow slits. Show also that, as the distance between the two slits d becomes smaller, the intensity pattern becomes that for one wide slit.

11.4. Diffraction grating

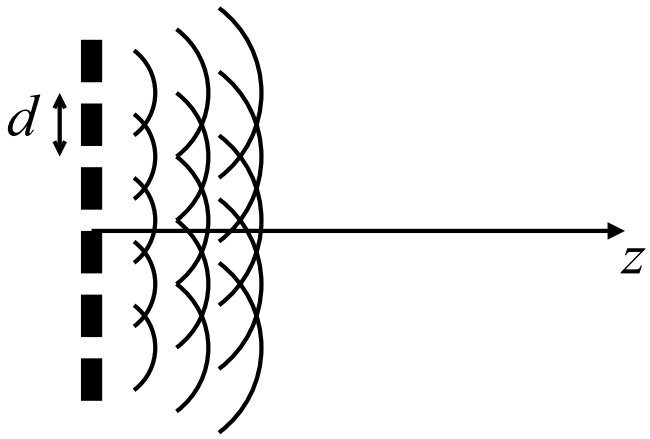


FIGURE 11.8.

Diffraction grating is a useful optical component for separating different frequency components in a light source. It consists of multiple **equally spaced** slits, as shown in Fig. 11.8. Let's assume that the slits are at positions

$$x_m = md, \quad m = -M, -M+1, \ldots, -1, 0, 1, \ldots, M-1, M, \tag{11.27}$$

so there are a total of $2M+1$ slits, and the distance between adjacent slits is d. For simplicity, suppose that their width Δx is much smaller than the wavelength such that we can use the formula given by Eq. (10.69):

$$\mathcal{E}(x,z) = \frac{\tilde{E}\Delta x}{\sqrt{z}} \sum_{m=-M}^{M} \exp\left[\frac{j\pi(x-md)^2}{\lambda z}\right]. \tag{11.28}$$

The differences between this and Eq. (10.69) are that we now have $2M+1$ slits with the index m going from $-M$ to M, the positions of the slits are equally spaced by d, and we have substituted $k = 2\pi/\lambda$. We again make the Fraunhofer approximation to omit the $x_m^2 = (md)^2$ term:

$$(x-md)^2 = x^2 - 2mxd + (md)^2 \approx x^2 - 2mxd, \tag{11.29}$$

arriving at

$$\mathcal{E}(x,z) = \frac{\tilde{E}\Delta x}{\sqrt{z}} \sum_{m=-M}^{M} \exp\left[\frac{j\pi(x^2 - 2mxd)}{\lambda z}\right] \tag{11.30}$$

$$= \frac{\tilde{E}\Delta x}{\sqrt{z}} \exp\left(\frac{j\pi x^2}{\lambda z}\right) \sum_{m=-M}^{M} \exp\left(\frac{-j2\pi mxd}{\lambda z}\right). \tag{11.31}$$

The $\exp[j\pi x^2/(\lambda z)]$ comes out of the sum, like the Fraunhofer diffraction integral, and we are left with a geometric series:

$$\sum_{m=-M}^{M} \exp\left(\frac{-j2\pi mxd}{\lambda z}\right) = \sum_{m=-M}^{M} \left[\exp\left(\frac{-j2\pi xd}{\lambda z}\right)\right]^m. \tag{11.32}$$

To simplify, we can use the formula (http://en.wikipedia.org/wiki/Geometric_series)

$$\sum_{n=0}^{N} G^n = \frac{1 - G^{N+1}}{1 - G}. \tag{11.33}$$

To use this, define a new index $m' = m + M$, such that, in terms of this new index, the sum in Eq. (11.32) now goes from $m' = 0$ to $m' = 2M$ with $m = m' - M$:

$$\sum_{m=-M}^{M} \exp\left(\frac{-j2\pi mxd}{\lambda z}\right) = \sum_{m'=0}^{2M} \exp\left[\frac{-j2\pi(m' - M)xd}{\lambda z}\right] \tag{11.34}$$

$$= \exp\left(\frac{j2\pi Mxd}{\lambda z}\right) \sum_{m'=0}^{2M} \exp\left(\frac{-j2\pi m'xd}{\lambda z}\right). \tag{11.35}$$

Now we associate m' with n and $2M$ with N in Eq. (11.33), leading to

$$\sum_{m=-M}^{M} \exp\left(\frac{-j2\pi mxd}{\lambda z}\right) = \exp\left(\frac{j2\pi Mxd}{\lambda z}\right) \frac{1 - \exp\left(\frac{-j2\pi(2M+1)xd}{\lambda z}\right)}{1 - \exp\left(\frac{-j2\pi xd}{\lambda z}\right)} \tag{11.36}$$

$$= \frac{\exp\left(\frac{j2\pi Mxd}{\lambda z}\right) - \exp\left(\frac{-j2\pi(M+1)xd}{\lambda z}\right)}{1 - \exp\left(\frac{-j2\pi xd}{\lambda z}\right)}. \tag{11.37}$$

To make this look nicer, multiply both the numerator and the denominator by the factor $\exp[j\pi xd/(\lambda z)]$:

$$\sum_{m=-M}^{M} \exp\left(\frac{-j2\pi mxd}{\lambda z}\right) = \frac{\exp[j\pi xd/(\lambda z)]}{\exp[j\pi xd/(\lambda z)]} \frac{\exp\left(\frac{j2\pi Mxd}{\lambda z}\right) - \exp\left(\frac{-j2\pi(M+1)xd}{\lambda z}\right)}{1 - \exp\left(\frac{-j2\pi xd}{\lambda z}\right)} \tag{11.38}$$

$$= \frac{\exp\left[\frac{j\pi(2M+1)xd}{\lambda z}\right] - \exp\left[\frac{-j\pi(2M+1)xd}{\lambda z}\right]}{\exp\left(\frac{j\pi xd}{\lambda z}\right) - \exp\left(\frac{-j\pi xd}{\lambda z}\right)} \tag{11.39}$$

$$= \frac{\sin\left[\frac{\pi(2M+1)xd}{\lambda z}\right]}{\sin\frac{\pi xd}{\lambda z}}. \tag{11.40}$$

This is called a periodic sinc function, which looks like Fig. 11.9. Plugging this back into Eq. (11.31),

$$\mathcal{E}(x,z) = \frac{\tilde{E}\Delta x}{\sqrt{z}} \exp\left(\frac{j\pi x^2}{\lambda z}\right) \frac{\sin\left[\pi(2M+1)xd/(\lambda z)\right]}{\sin[\pi xd/(\lambda z)]}, \tag{11.41}$$

and the intensity is

$$I(x,z) = \frac{|\tilde{E}|^2(\Delta x)^2}{2Zz} \frac{\sin^2\left[\pi(2M+1)xd/(\lambda z)\right]}{\sin^2[\pi xd/(\lambda z)]}. \tag{11.42}$$

The period of this function is

$$\boxed{\Lambda = \frac{\lambda z}{d},} \tag{11.43}$$

which is inversely proportional to the spacing d of the slits, while the width of each big peak (zero to zero) is

$$\boxed{W = \frac{2\lambda z}{(2M+1)d},} \tag{11.44}$$

which is inversely proportional to $(2M+1)d$, the distance from the top slit to the bottom slit and therefore the total size of the grating, as indicated in Fig. 11.10.

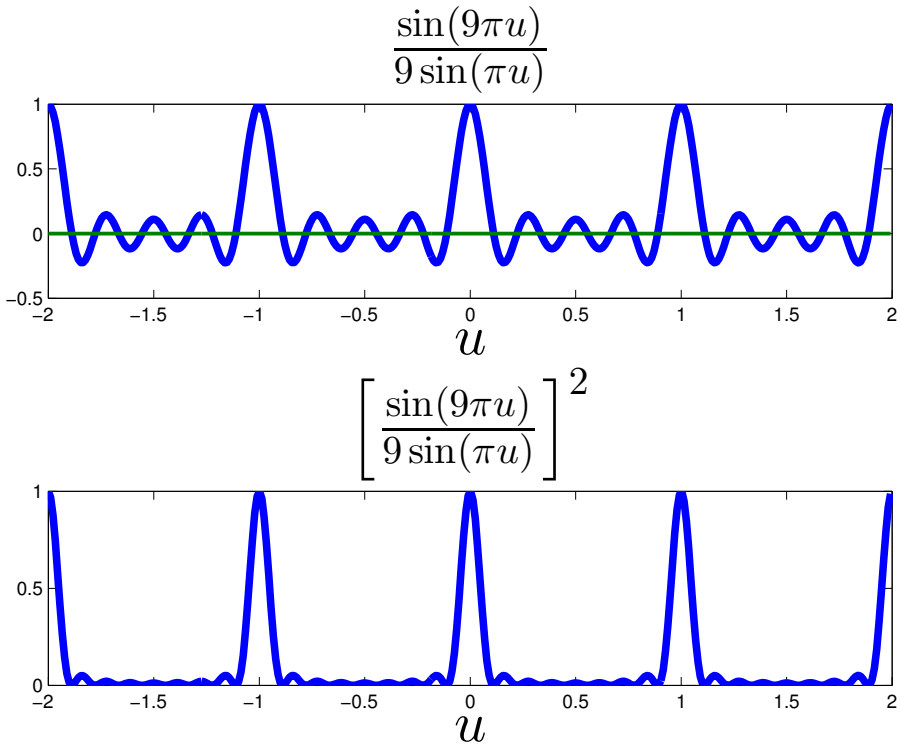


FIGURE 11.9. Left: A normalized periodic sinc function. Right: A normalized periodic sinc^2 function.

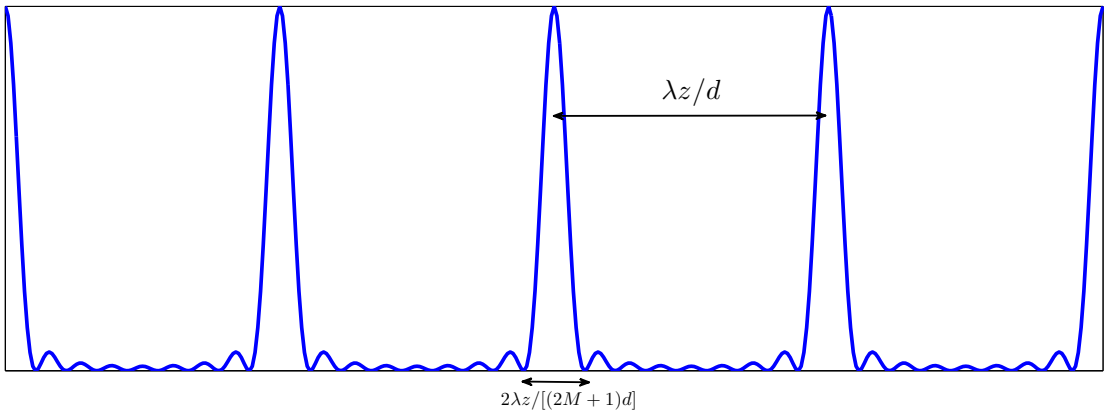


FIGURE 11.10. The period of the Fraunhofer diffraction intensity pattern is $\lambda z/d$, which is inversely proportional to the spacing d of the slits, while the width of each peak is $2\lambda z/[(2M+1)d]$, which is inversely proportional to the total size of the grating.

So far this is for a single frequency ω. Note that the period of the intensity pattern Λ is proportional to the wavelength λ, and if there are multiple frequencies in the light source, each frequency will lead to periodic peaks with period proportional to λ, and the frequency components end up being separated spatially by the diffraction grating. See http://h2physics.org/?cat=49 for example.

- **Exercise**: Look at the image at http://h2physics.org/wp-content/uploads/2009/08/coloredspectrum.jpg for white light passing through a diffraction grating and explain why one would see a white peak at the middle, violet/blue light closer to the center and red light away from the center.

CHAPTER 12

Focusing and Imaging

As we saw in Chap. 11, the diffracted intensity looks nothing like the fields where they come from. The fact that we are able to see is thanks to the lenses in our eyes, which are curved dielectrics that bend and focus light onto our retina. Lenses are also used to focus laser beams onto an optical disc, such as CD, DVD, and Blu-ray, to read the data stored on it. In this section we shall study the basic principle of a focusing lens in wave optics.

12.1. Focusing lens

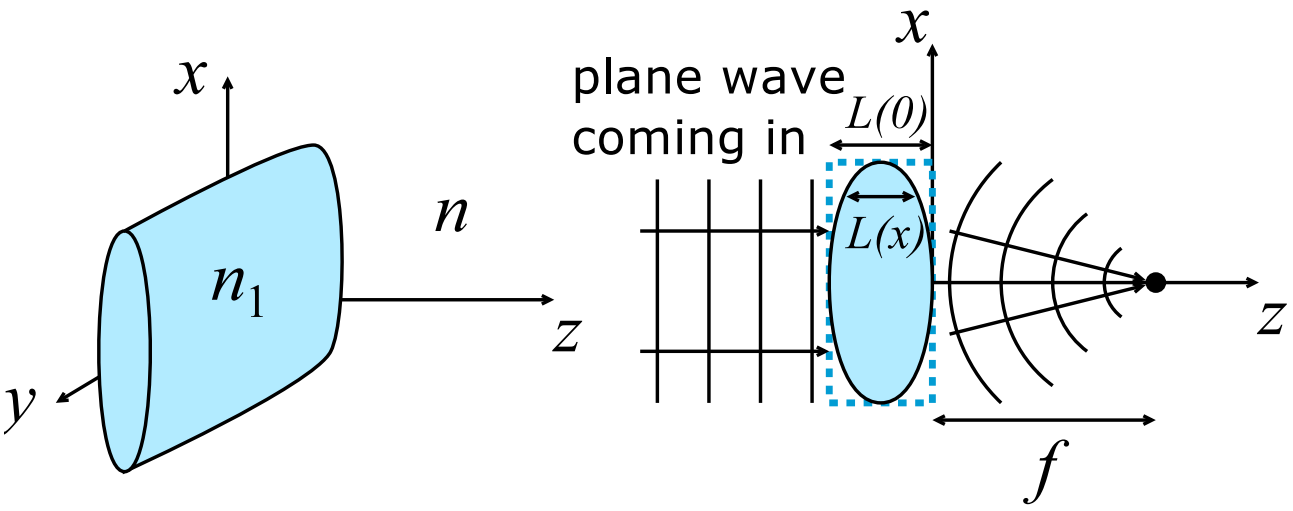


FIGURE 12.1. A cylindrical lens with index n_1 and thickness $L(x)$ as a function of x. The surrounding has an index n, and f is the focal length depending on the curvature in $L(x)$ and the indices n_1 and n.

We shall first consider a cylindrical lens, as shown in Fig. 12.1. It is a dielectric curved along x but uniform along y. Suppose that the surrounding medium has an index n and the lens has a higher refractive index

$$n_1 > n. \tag{12.1}$$

Let $L(x)$ denote the thickness of the lens as a function of x; $L(x)$ is the thickest at $x = 0$ and gets thinner away from the center. Roughly speaking, when a wave propagates past the lens and we ignore multiple reflections, the wave emerging on the other side of the blue box in Fig. 12.1 should acquire a phase delay

$$\underbrace{\exp\left[jk_1L(x)\right]}_{\text{inside lens}}\underbrace{\exp\left\{jk\left[L(0)-L(x)\right]\right\}}_{\text{outside lens}} = e^{jkL(0)}\exp\left[j(k_1-k)L(x)\right], \tag{12.2}$$

where $k_1 = \omega n_1/c$, because at each x it picks up a phase delay $\exp\left[jk_1L(x)\right]$ as it propagates inside the lens and also picks up a total phase delay $\exp\left\{jk\left[L(0)-L(x)\right]\right\}$ propagating in the surrounding medium around the lens inside the blue box. We shall make a parabolic approximation of $L(x)$ near the center of the lens:

$$L(x) \approx L(0) - \frac{1}{2}Cx^2, \tag{12.3}$$

which is a parabola used to approximate the curvature of $L(x)$, as shown in Fig. 12.2, and C is some curvature parameter. This means that the phase delay becomes

$$e^{jkL(0)} \exp\left[j(k_1-k)L(x)\right] \approx e^{jk_1L(0)} \exp\left[-\frac{j(k_1-k)Cx^2}{2}\right]. \tag{12.4}$$

This means that there is more delay at the center and less delay at the sides. If we think about the wavefronts, they will look like the right figure of Fig. 12.1, as the center portion of the wavefront is delayed more than the side portion.

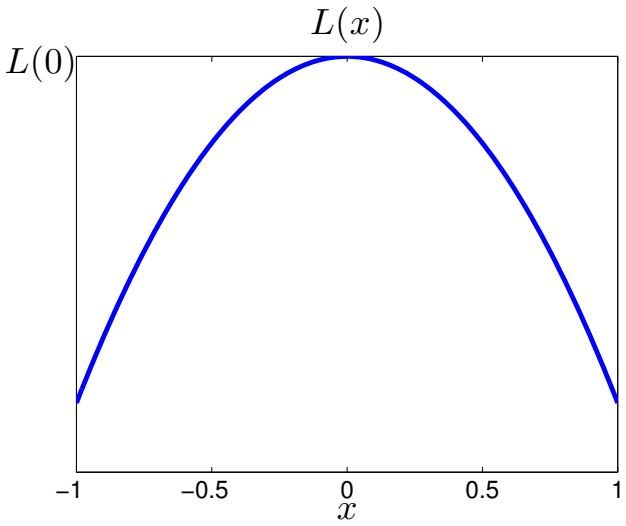


FIGURE 12.2. A parabola $L(0) - Cx^2/2$ approximating the thickness $L(x)$ near $L(x=0)$.

Just after the lens, the field is then given by

$$\mathcal{E}(x', z=0) \approx \tilde{E} \exp\left[-\frac{j(k_1-k)Cx'^2}{2}\right], \tag{12.5}$$

neglecting the constant phase factor $e^{jk_1L(0)}$. For comparison with Sec. 11.1, I also write x' here instead of x. Now we consider each point on the lens along x' as a cylindrical wave source, like what we did for a wide slit in Sec. 11.1, except that each source has this quadratic phase factor depending on its position. Taking the paraxial approximation and integrating the cylindrical waves from each x', we obtain the Fresnel diffraction of this field at a later z given by

$$\mathcal{E}(x,z) \approx \frac{\tilde{E}}{\sqrt{z}} \int_{-\infty}^{\infty} dx' \quad \underbrace{\exp\left[\frac{jk(x-x')^2}{2z}\right]}_{\text{phase factor from cylindrical wave}} \quad \underbrace{\exp\left[-\frac{j(k_1-k)Cx'^2}{2}\right]}_{\text{phase factor from lens}}, \tag{12.6}$$

where we assume that the lens is infinitely big along x' for now so that we are integrating from $x' = -\infty$ to $x' = \infty$.

Looking at this integral, we see a fortuitous simplification if

$$\frac{k}{2z} = \frac{(k_1-k)C}{2}, \tag{12.7}$$

in which case the terms with respect to x'^2 cancel:

$$\exp\left(\frac{jkx'^2}{2z}\right) \exp\left[-\frac{j(k_1-k)Cx'^2}{2}\right] = 1 \quad \text{if } \frac{k}{2z} = \frac{(k_1-k)C}{2}. \tag{12.8}$$

We denote this magical propagation distance as $z = f$:

$$\frac{k}{2f} = \frac{(k_1-k)C}{2}, \qquad f = \frac{k}{(k_1-k)C} = \frac{n}{(n_1-n)C}. \tag{12.9}$$

When this happens, we don't need to assume the Fraunhofer approximation like what we did in Sec. 11.2, the lens cancels the quadratic phase factor inside the Fresnel diffraction integral for us. Thus

$$\mathcal{E}(x, z=f) \approx \frac{\tilde{E}}{\sqrt{f}} \int_{-\infty}^{\infty} dx' \exp\left[\frac{jk(x^2 - 2xx')}{2f}\right] \tag{12.10}$$

$$= \frac{\tilde{E}}{\sqrt{f}} \exp\left(\frac{jkx^2}{2f}\right) \int_{-\infty}^{\infty} dx' \exp\left(\frac{-jkxx'}{f}\right). \tag{12.11}$$

Note that

$$\int_{-\infty}^{\infty} dx' \exp\left(\frac{-jkxx'}{f}\right) = 2\pi\delta\left(\frac{kx}{f}\right), \tag{12.12}$$

where $\delta(X)$ is the **Dirac delta function** (http://en.wikipedia.org/wiki/Dirac_delta_function). It has the following properties:

$$\delta(X) = \begin{cases} 0, & X \neq 0, \\ \infty, & X = 0, \end{cases} \qquad \int_{-\infty}^{\infty} dX\delta(X) = 1, \tag{12.13}$$

$$\delta(x)g(x) = \delta(x)g(0) \text{ for any } g(x), \qquad \delta\left(\frac{kx}{f}\right) = \frac{f}{k}\delta(x). \tag{12.14}$$

We then get

$$\mathcal{E}(x, z=f) \approx \frac{\tilde{E}}{\sqrt{f}} \exp\left(\frac{jkx^2}{2f}\right) \frac{2\pi f}{k}\delta(x) \propto \delta(x). \tag{12.15}$$

The remarkable thing about this is that the field ends up being a delta function along x, so it is **concentrated at a single spot at** $x = 0$ and is zero anywhere else on x. In other words, the quadratic phase factor due to the lens, together with Fresnel diffraction, **focuses** a plane wave to a single spot along x. Note that the fields and the intensity are still uniform along y, meaning that, if we observe the focal intensity on the (x, y) plane, it will look like an infinitesimally thin line at $x = 0$ but uniform along y. We expected this for a cylindrical lens. The magical propagation distance f given by Eq. (12.9) is called the **focal length**. It is not only a function of the curvature parameter C but also the index of the lens n_1 as well as the index of the surrounding n.

The field and also the intensity are infinite at $x = 0$ because we have assumed an infinitely big plane wave coming in and an infinitely big lens. For a more realistic model, let's consider a finite-size lens in the next section instead.

- **Question**: What happens when the curvature parameter C is negative, so the lens is thinner in the middle and thicker on the sides? What happens if C is positive but $n_1 < n$? Do you still get focusing?

12.2. Finite lens aperture

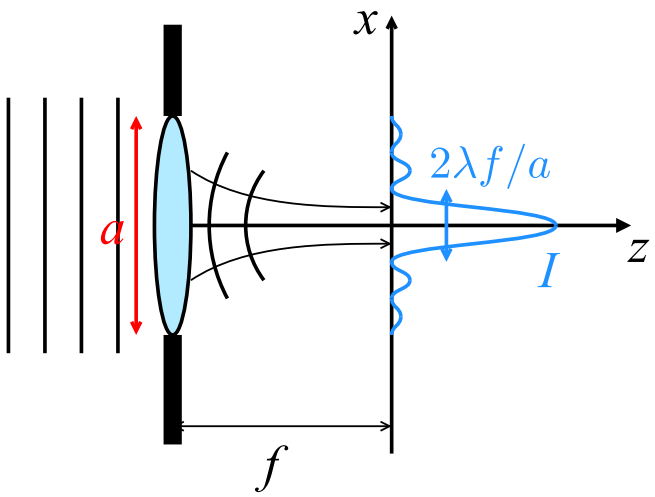


FIGURE 12.3. Focusing by a cylindrical lens with finite aperture size a. The resulting intensity pattern at focal length f has a width $2\lambda f/a$.

Instead of assuming an infinitely big lens and integrating over x' from $-\infty$ to ∞, let's assume that the lens is inside a slit (also called an aperture in this context) with width a. Now the integral in Eq. (12.11) becomes

$$\mathcal{E}(x,z) \approx \frac{\tilde{E}}{\sqrt{f}} \exp\left(\frac{jkx^2}{2f}\right) \int_{-a/2}^{a/2} dx' \exp\left(\frac{-jkxx'}{f}\right), \tag{12.16}$$

where the only change is the integration limits that are now from $x' = -a/2$ to $a/2$ to indicate the finite aperture; the waves are assumed to be blocked outside the aperture. We have done this integral before for Fraunhofer diffraction in Sec. 11.2, except that this time we don't need to assume the Fraunhofer approximation because the lens already cancels the quadratic x'^2 term in the Fresnel diffraction integral for us. The result is

$$\mathcal{E}(x,z) \approx \frac{\tilde{E}}{\sqrt{f}} \exp\left(\frac{j\pi x^2}{\lambda f}\right) a \operatorname{sinc}\left(\frac{ax}{\lambda f}\right), \tag{12.17}$$

and the intensity is

$$I(x,z) \approx \frac{|\tilde{E}|^2 a^2}{2Zf} \operatorname{sinc}^2\left(\frac{ax}{\lambda f}\right). \tag{12.18}$$

Instead of a delta function with an infinitesimal width, the spot size in terms of the zero-to-zero width of the main intensity peak is now

$$\boxed{W = \frac{2\lambda f}{a}.} \tag{12.19}$$

Note that this is **proportional to** λ and **inversely proportional to the aperture size** a. The smaller the aperture size a, the larger the spot size, and vice versa. With the paraxial approximation, we must have $a \ll f$, so this focus spot size W will always be somewhat larger than the wavelength λ in this paraxial regime. **Beyond the paraxial regime, focusing doesn't work as well, so we are usually restricted to the paraxial regime for a lens to work well and have to live with a spot size W somewhat larger than λ.**

- **Exercise**: If I assume f to be fixed, describe three ways to reduce the laser spot size.
  Answer:
  - Increase the size of the lens (while making sure that the paraxial approximation still holds).
  - Increase the frequency of the laser ω and therefore reduce the free-space wavelength λ_0.
  - Use a dielectric medium that has a higher refractive index n to reduce the wavelength in the medium λ_0/n. Note that the lens has to be redesigned for a different surrounding medium for the focal length to be held fixed, because $f = n/[(n_1 - n)C]$ in Eq. (12.9) and if I change n, I also have to change n_1 and/or the curvature parameter C to keep f the same.

12.3. *Spherical lens

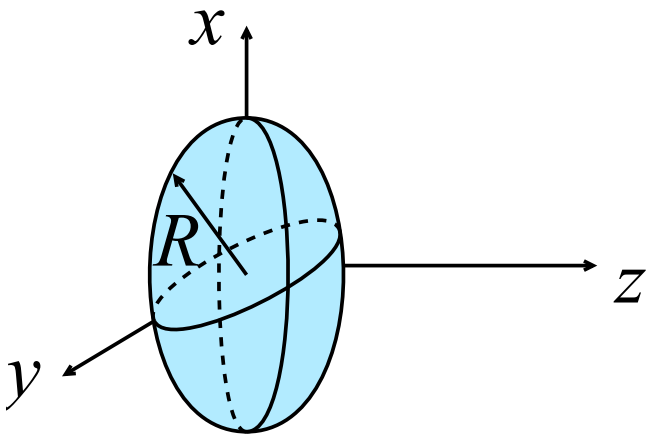


FIGURE 12.4. A spherical lens has a surface that has curvative along x and y. A lens should have a finite size with radius R.

A spherical lens, as plotted in Fig. 12.4, is not so different from a cylindrical lens, except that it has curvature in both x and y. The thickness with the parabolic approximation is now

$$L(x,y) \approx L(0,0) - \frac{1}{2}C\left(x^2+y^2\right), \tag{12.20}$$

and the phase delay introduced by the lens is then

$$\exp(jk_1 L) \approx e^{jk_1 L(0,0)} \exp\left[-\frac{j(k_1-k)C\left(x^2+y^2\right)}{2}\right]. \tag{12.21}$$

Fresnel diffraction should now be expressed in terms of spherical wave from each point on the lens:

$$\mathcal{E}(x,y,z) \approx \frac{\tilde{E}}{z}\int_{\text{surface on lens}} dx'dy' \underbrace{\exp\left[\frac{jk[(x-x')^2+(y-y')^2]}{2z}\right]}_{\text{phase factor from spherical wave}} \underbrace{\exp\left[-\frac{j(k_1-k)C\left(x'^2+y'^2\right)}{2}\right]}_{\text{phase factor from lens}}. \tag{12.22}$$

Again, the phase factor from the lens cancels the x'^2 and y'^2 terms in the Fresnel diffraction integral at the focal length $z = f = n/[(n_1-n)C]$:

$$\mathcal{E}(x,y,f) \approx \frac{\tilde{E}}{f}\exp\left[\frac{jk(x^2+y^2)}{2z}\right]\int_{\text{surface on lens}} dx'dy' \exp\left(\frac{-jkxx'}{z}\right)\exp\left(-\frac{jkyy'}{z}\right). \tag{12.23}$$

If we again assume that the lens is infinitely big, the two-dimensional integral is separable into two one-dimensional integral, and we end up having delta functions $\delta(x)\delta(y)$ centered at $x = 0$ and $y = 0$. The spherical lens hence focuses the light onto a single spot on the two-dimensional (x,y) plane at the focal length $z = f$.

If the lens has a size, such as a circular aperture with radius R shown in Fig. 12.4, then we need to do this two-dimensional integral:

$$\int_{\sqrt{x'^2+y'^2}\le R} dx'dy' \exp\left(\frac{-jkxx'}{z}\right)\exp\left(-\frac{jkyy'}{z}\right) = \pi R^2 \,\text{jinc}(\rho), \tag{12.24}$$

$$\text{jinc}(\rho) \equiv \frac{2J_1(\pi\rho)}{\pi\rho}, \quad \rho \equiv \frac{2R}{\lambda f}\sqrt{x^2+y^2}. \tag{12.25}$$

The jinc function (http://en.wikipedia.org/wiki/Sombrero_function) is defined with respect to the first-order Bessel function J_1 and plotted on the left of Fig. 12.5. The intensity is then proportional to its square, which is plotted on the right of Fig. 12.5. In two dimensions with respect to x and y, the jinc function looks like the right figure of Fig. 12.6. On a camera, the focus spot looks like a disk called **Airy disk**, shown in Fig. 12.7.

The jinc function has its first zero at $\rho \approx 1.22$, so we can define the width of the spot size (zero to zero) as

$$W \approx 1.22\frac{\lambda f}{R}. \tag{12.26}$$

This means that the spot size is **proportional to** λ and **inversely proportional to** R, similar to the behavior of the spot size for the cylindrical lens in the previous section. Again, paraxial approximation means that $f \gg R$, so W must be somewhat larger than λ in the paraxial regime. Again, focusing doesn't work as well beyond the paraxial regime, and we have to live with a spot size W that is somewhat larger than λ.

12.4. Application: optical discs

The basic principle of optical disc storage, such as CD, DVD, and Blu-ray, is to represent data bits on an optical disc as lands and pits on the disc, focus a laser beam onto the disc, as shown in Fig. 12.8, and measure the reflected beam as the disc rotates. The reflected beam is deflected by different amounts as the height of the structure on the disc changes, so the bits can be read by measuring the deflection. The size of the lands and pits must be comparable to the spot size or the laser spot will overlap many bits and the bits cannot be inferred accurately anymore. As we discussed in the previous two sections, the focus spot size has to be somewhat larger than λ, which then limits the storage density. In successive generations of optical discs, lasers with increasingly

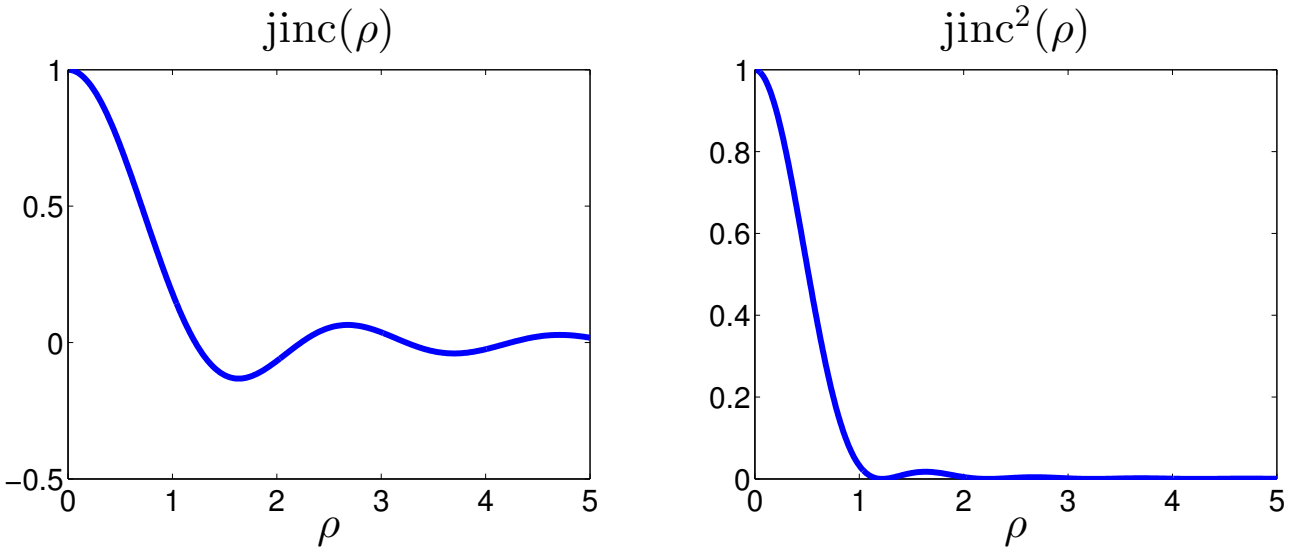


FIGURE 12.5. Left: plot of jinc function. Right: plot of jinc^2.

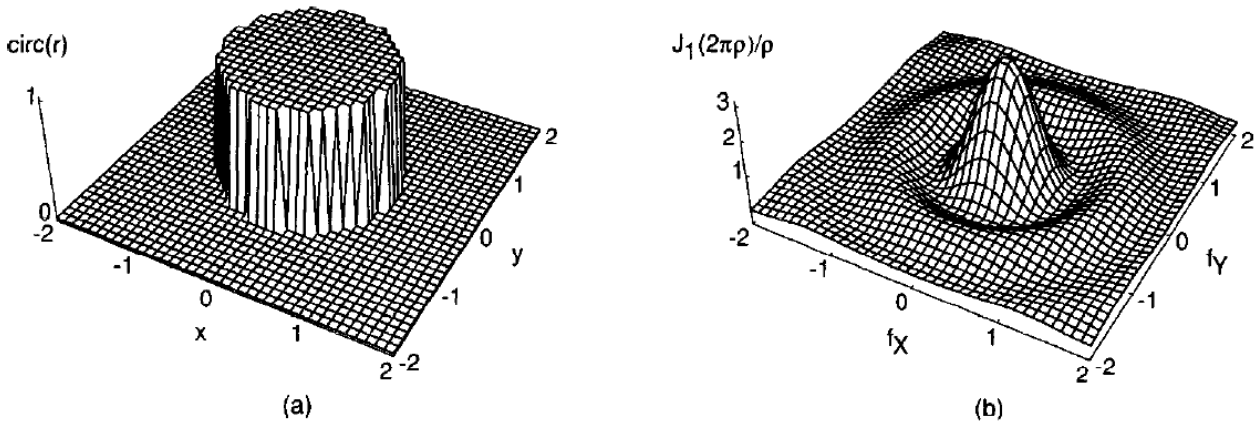


FIGURE 2.3
(a) The circle function and (b) its transform.

FIGURE 12.6. The jinc function mathematically comes from the two-dimensional Fourier transform of the circle function, analogous to the sinc function that comes from the Fourier transform of a rectangle function. From [2].

smaller wavelengths are therefore used to make the spot size smaller, the island and pit size to be smaller, and the data storage density to be higher. See also http://www.youtube.com/watch?v=77qW8I3G_PQ.

12.5. *Single-lens imaging

12.6. *Rayleigh resolution limit

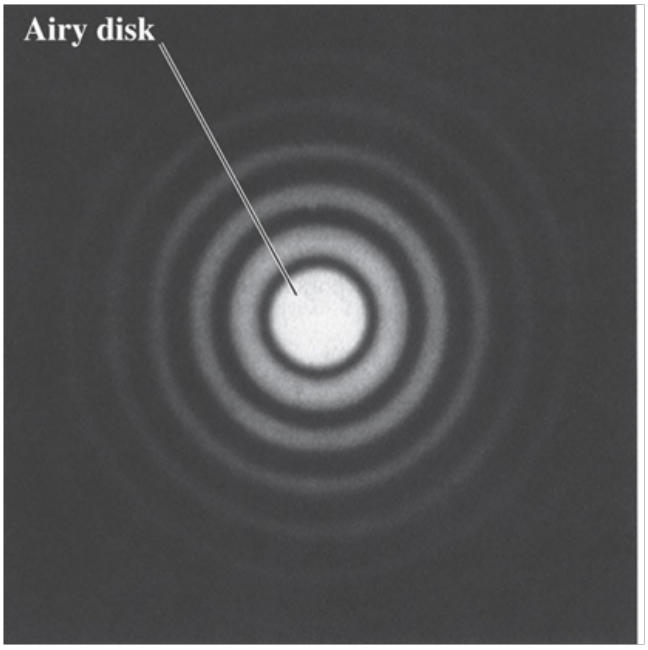


FIGURE 12.7. Camera image of the focus spot from a spherical lens. From [2]?

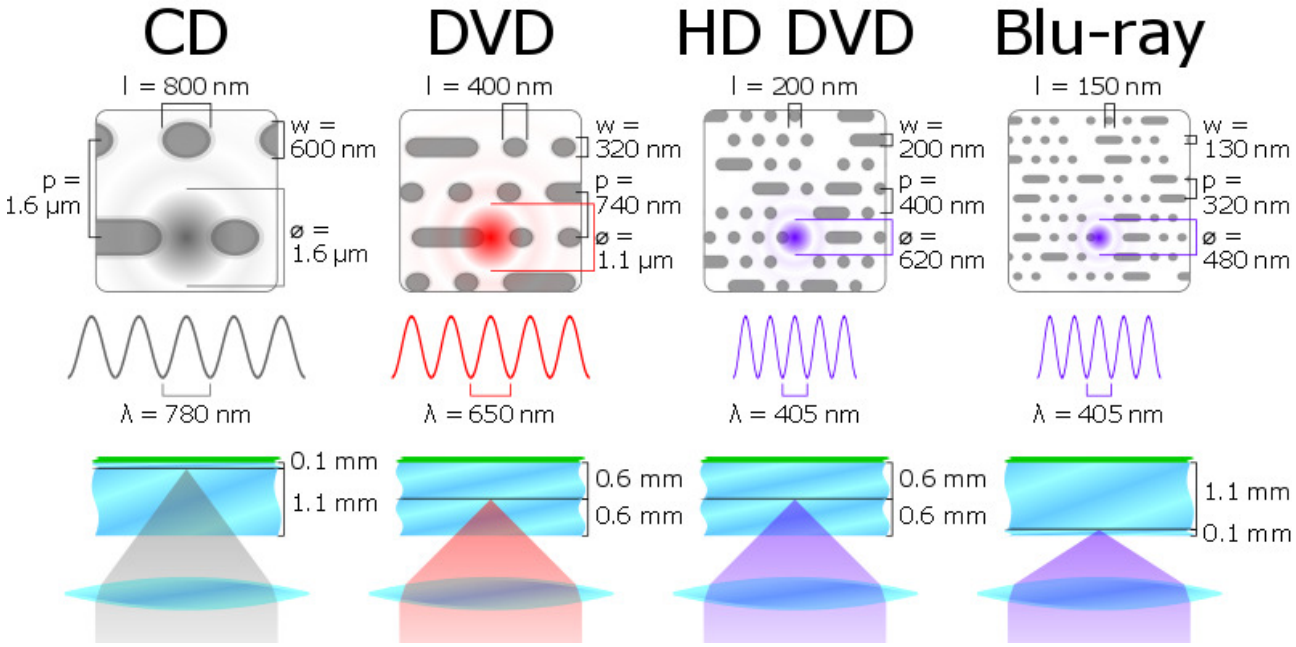


FIGURE 12.8. Different formats of optical discs, the sizes of the islands and pits, and the laser wavelengths in free space. From http://upload.wikimedia.org/wikipedia/commons/a/ad/Comparison_CD_DVD_HDDVD_BD.svg

CHAPTER 13

More Realistic Models of Dielectric Media

13.1. Loss

13.2. Gain

13.3. Dispersion

13.4. Birefringence

13.5. Application: Liquid-crystal display

http://www.youtube.com/watch?v=k7xGQKpQAWw

CHAPTER 14

*Dynamics of Classical Charged Particles

Bibliography

[1] Hugh D. Young and Roger A. Freedman, *University Physics* (Pearson/Addison-Wesley, San Francisco, 2008).
[2] Joseph W. Goodman, *Introduction to Fourier Optics* (McGraw-Hill, New York, 1996).
[3] David J. Griffiths, *Introduction to Electrodynamics* (Prentice-Hall, Upper Saddle River, 1999).
[4] John D. Jackson, *Classical Electrodynamics* (Wiley, Hoboken, 1999).