普通高等教育“十二五”规划教材

A Concise Course in University Physics

Second Edition Volume 2

大学物理简明教程（英文版）

第二版

下 册

王安安 伏云昌 主编

科 学 出 版 社

(O-5018.0101)

A Concise Course in University Physics

高等教育出版中心　数理出版分社
联系电话：010-64015178
E-mail：mph@mail.sciencep.com

www.sciencep.com
ISBN 978-7-03-036554-5
定 价：59.00 元（上下册）

A CONCISE COURSE IN
UNIVERSITY PHYSICS
Second Edition Volume 2

大学物理简明教程（英文版）
第二版 下册

王安安 伏云昌 主编
伏云昌 王安安 陈劲波 编写

科 学 出 版 社

北 京

内 容 简 介

本书是按照《理工科类大学物理课程教学基本要求(2010 年版)》修订的，涵盖了所有 A 类的内容，选择了部分 B 类的内容，如非惯性系、质心、气体的范德瓦尔斯方程、玻尔兹曼分布、几何光学、固体能带论和激光简介等. 为方便上、下两学期的教学安排，本次改版将原三册改编为上、下两册. 全书共 19 章：上册为力学和电磁学；下册为热学、振动与波动、光学和近代物理. 本书配有双语课件光盘.

本书可供理工科非物理专业 112～128 学时双语教学使用，也可供在某一部分内容进行双语教学试点选用，还可供对英文物理感兴趣的广大读者自学或作参考书用.

图书在版编目(CIP)数据

大学物理简明教程：全 2 册＝A concise course in university physics：英文/王安安，伏云昌主编. —2 版. —北京：科学出版社，2013. 1
普通高等教育“十二五”规划教材
ISBN 978-7-03-036554-5

Ⅰ. ①大…　Ⅱ. ①王…　②伏…　Ⅲ. ①物理学-高等学校-教材-英文
Ⅳ. ①O4

中国版本图书馆 CIP 数据核字(2013)第 017611 号

责任编辑：昌　盛/ 责任校对：宋玲玲
责任印制：阎　磊/ 封面设计：迷底书装

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com
新科印刷有限公司 印刷
科学出版社发行　各地新华书店经销
*
1998 年 6 月重庆大学出版社第一版
2013 年 1 月第　二　版　　开本：787×1092 1/16
2013 年 1 月第一次印刷　　印张：34 1/2
字数：784 000

定价：59.00 元(上下册)

第二版前言

根据《理工科类大学物理课程教学基本要求(2010年版)》对教学过程的基本要求第5点“双语教学——在保证教学效果的前提下,有条件的学校可开展物理课程的双语教学,以提高学生查阅外文资料和科技外语交流的能力”,为满足坚持双语教学的师生对英文物理教材的需要,我们对*A Concise Course in University Physics*(《大学物理简明教程(英文版)》,1998~2000年出版)进行了部分修订.按2010年版教学基本要求,第二版涵盖了所有A类的内容,保留原有的并新选了部分B类内容,如增加了非惯性系和惯性力、质心和质心运动定理、热力学第零定律、范德瓦尔斯方程及几何光学等内容.为方便目前工科物理两学期教学内容的安排,将第一版的三册改编为上、下两册,供非物理专业112~128学时双语物理教学之用.第二版保留了第一版的主要特色,力求系统完整、理论叙述准确、文字简明易懂,以达到教师好用、学生易学的目的.因篇幅所限,第二版省去了各章总结和三个阅读材料.

全书19章编写分工如下:

上册:力学(第1~5章)由王安安编写;电磁学(第6~9章)主要由樊则宾编写.

下册:热学(第10,11章)由王安安编写;振动与波动部分第12、14章及第15章中的几何光学由伏云昌编写,第15章中的波动光学由陈劲波编写,第13章由王安安和陈劲波编写;近代物理(第16~19章)由伏云昌编写.王安安和伏云昌负责全书的统稿、修改和定稿工作.为方便本书读者进行双语多媒体课堂教学,伏云昌编制了配套教学课件.

北京大学陆果教授对第二版全书进行了认真的审定,并提出了宝贵的意见和建议,谨此致以诚挚谢意!编者也对昆明理工大学校领导多年来对我们工作的支持表示衷心的感谢!

双语教学贵在坚持,我们编写及使用英文物理教材的初衷未改,通过双语教学提高学生的综合素质始终是我们的目标,编者愿与致力于这项事业的同行共同努力,持之以恒,为物理教学改革尽绵薄之力!

因时间仓促、水平有限,书中难免有不当之处,恳请同行与读者提出宝贵意见.

编　者

2011年12月于昆明

第一版前言摘要

我们正处在一个高新技术飞速发展、科技信息量激增、知识更新加快、国际交流日益广泛的时代. 我国的进一步改革开放，社会主义市场经济的建立都要求高校毕业生有更强的适应能力，在人才市场上，有效强的外语应用能力、交流型、综合型的毕业生供不应求. 在这种形势下，我们的高等教育正向着重视素质教育的方向转变，而素质与能力是密切相关的，素质的培养要以一定的知识和能力为基础，其中包括独立获取知识的能力. 毋庸置疑，直接用外语为工具获取知识、进行交流的能力是人才素质的一个重要方面.

然而，由于历史原因、文化背景、经济基础、外语教育模式和各类师资外语水平等诸多因素的影响，我们在应用外语进行教学方面的基础性工作十分薄弱. 在普通高校本科生教育中，教材和教学过程基本上只使用中文这一单一语种，实际上已经制约了学生应用外语（主要是英语）获取知识能力的发展. 为了改变这种现状，跟上时代的步伐，试用英文教材，使用英语进行教学的改革便应运而生了.

本教材的编写是编者主持的“试用英文物理教材”教改试点工作的继续，也是编者大学物理教学经验的总结. 从国外引进的教材，虽有诸多优点，但在系统上与我国的大学物理教学基本要求不完全对应，为了满足师生对英文物理教材的需要，编写一套根据我国工科物理教学基本要求，顺应工科物理教学改革的形势，反映编者在多年物理教学实践中总结出来的教学方法与经验，与我们的学生在一、二年级的英文水平相适应的英文“简明物理学教程”的计划就提到日程上来了，这就是我组织编写这套教材的初衷. 这套英文物理教材是1996年经国家教委批准列入正式出版计划的. 本教材可供普通高等工科院校本科生物理课130～140学时使用.

……

全书21章，具体编写分工如下：

第一册：力学（第1～5章），分子运动论和热力学基础（第6、7章）由王安安编写.

第二册：电学及稳恒电流的磁场（第8～10章）由吴光敏编写，磁介质和电磁感应（第11、12章）由樊则宾编写，麦克斯韦方程组（第13章）由王安安编写.

第三册：机械振动（第14章）由伏云昌编写，机械波（第15章）由王安安和陈劲波编写. 电磁振荡与电磁波（第16章）由吴光敏编写，波动光学部分（第17章）中干涉与衍射由陈劲波编写，光的偏振由樊则宾编写. 近代物理部分（第18～21章）由伏云昌编写.

第一册绝大部分插图由刘富华用计算机绘制，其余插图由李俊昌教授绘制，封面也由李俊昌设计，谨此致以诚挚谢意.

编写大学英文物理教材是一种大胆的尝试，由于编者水平有限，错误疏漏之处在所难免，希望同行和读者批评指正. 我们相信这本教材的出版将对物理教学的现代化和物理教学与国际接轨作出有益的贡献.

编　者

1997年5月于昆明

Contents

Part Three Thermodynamics

Part Four Oscillation and Waves

Part Five The Modern Physics

Part Three

Thermodynamics

Chapter10

The Kinetic Theory of Gases

We come to observations on heat phenomena, such as the temperature of a gas system, the pressure exerted by the gas on the walls of the container, the heat transformation and the work done by a gas system, which we will discuss in detail in this course.

Unlike in mechanics, the object in this part is an intrinsically complex system of a vast number of particles, rather than the simple system of a few particles or objects. There are two distinct methods to study heat phenomena. The first, called the macroscopic method or classical thermodynamics, is established on the basis of macroscopic variables measured as pressure, volume and temperature, deals with the properties of bulk matter, takes no account of the fact that all matter is made up of atoms and molecules.

The second, called microscopic method or kinetic theory, is based on atomic model of matter. The basic assumption of kinetic theory is that the measurable properties of matter like temperature, pressure and volume of a gas system reflect the combined actions of countless numbers atoms and molecules. Kinetic theory attempts to relate the microscopic properties of atoms or molecules which are not directly measurable, such as the mass, velocity, momentum and kinetic energy of a molecule, to the measurable macroscopic parameters of the system by means of the statistic method—investigating the average behavior of the microscopic parameters that characterizes the individual molecule. We shall focus our attention on this method in this chapter.

10.1 The Zeroth Law of Thermodynamics

10.1.1 The thermal equilibrium

Temperature is a central concept in thermodynamics. In order to define it strictly not depending on our subjective sense of touch in which objects feel hot or cold, we must understand the concept about thermal equilibrium first. When we put object A and object B together in an isolate system without any disturber from environment (refer to section 3-6), and let them in thermal contact so that the thermal energy exchanges from one to another. The experimental facts show that the two objects will eventually reach a stable state at which the thermal energy exchange stops. Then the two objects are defined to be in thermal equilibrium.

Furthermore, if we want to determine whether or not system A and system B are in thermal equilibrium but without being contact with each other, usually we can employ a third system C. Suppose that system A and system C are in thermal equilibrium; while system B and system C are in thermal equilibrium respectively, experiments indicate that system A and system B are in thermal equilibrium too. Even if we let them in thermal contact there would be no thermal energy exchange between them at all.

10. 1. 2 The zeroth law of thermodynamics, Definition of temperature

The experimental facts mentioned above are summed up in the zeroth law of thermodynamics: If system A and system B are each in thermal equilibrium with a third system T, then they are in thermal equilibrium with each other. It is called thermal equilibrium law also. According to this law, the objects in thermal equilibrium have a common property so that we define this property as temperature. When two objects are in thermal equilibrium, their temperatures are equal and vice versa.

The zeroth law of thermodynamics allows us to make use of a thermometer, say, to employ the third system T mentioned above to measure the temperature of other system, the only thing we need to do is to calibrate it.

10. 2 State Parameters, Equilibrium State, and Ideal Gas Law

10. 2. 1 State parameters of gas

In mechanics, to describe the states of a mechanical system, the position and velocity were needed. In thermodynamics, the states of a thermodynamic system are described by pressure p, volume V, and temperature T. The p, V, T are called state parameters of the system.

Because molecules are constantly in random motion, the container is filled with gases at any time, so, the volume of gas actually is the volume of the container and in SI unit, its unit is m^3. The pressure p is the force exerted by a gas on per unit area of the wall of the container, and it must surely be related to the steady drumbeat of molecules on the walls of the container. The SI unit of pressure is pascal (Pa), 1 Pa= 1 $\mathrm{N/m}^2$. Meanwhile, the following units are often used

$$1\ \mathrm{cmHg} = 10\ \mathrm{mmHg} = 1.333\times10^3\ \mathrm{Pa}$$

$$1\ \mathrm{atm} = 76\ \mathrm{cmHg} = 1.013\times10^5\ \mathrm{Pa}$$

$$1\ \mathrm{kgf/cm}^2 = 9.8\times10^4\ \mathrm{Pa}$$

We have made definition of temperature in the last section, based on thermal equilibrium, or the zeroth law of thermodynamics. The unit of temperature is Kelvin called the thermodynamic scale or absolute temperature, labeled as K, defined as the ratio of the temperature of triple-point of water to 273. 15. It is that the temperature of triple-point of water, 273. 15 K is defined as a constant temperature in the system of thermodynamic scale. The Celsius temperature scale is also often used, labeled as ℃, and it employs a degree of the same magnitude as that of the Kelvin scale but its zero point is shifted by 273. 15 degrees. Thus if t denotes the

Celsius temperature, and T denotes the Kelvin temperature (i. e. absolute temperature), we have

$$T/\mathrm{K}=t/^{\circ}\mathrm{C}+273.15$$

The pressure, volume and temperature of a given system, all can be determined by measurements, so that we call them as macroscopic properties of that system.

10. 2. 2 Equilibrium state and equilibrium process

When the temperature and pressure are the same at all points in a system, the system is said to be in an equilibrium state. Hence if heat is added at some point to a system in an equilibrium state, we must wait until the process of heat transfer within the system has brought about a new uniform temperature before the system is again in equilibrium state.

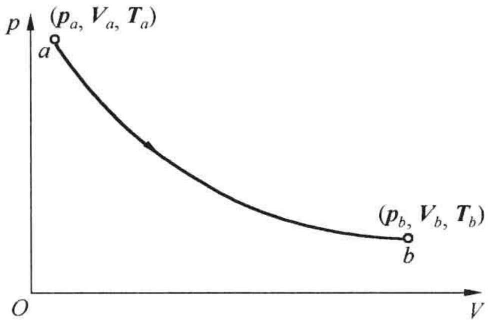


Fig. 10-1 The thermal equilibrium states and process are represented on the p-V diagram

An equilibrium state of a thermodynamic system can be described by its pressure p, volume V and temperature T; for example, (p_a, V_a, T_a,) describes state a, (p_b, V_b, T_b) describes state b. An equilibrium state can be represented by a dot on the pressure-volume diagram (briefly as p -V diagram) shown in Fig. 10-1.

The operation of changing the system from its initial state to its final state is called a thermodynamic process. During such processes, energy may be transferred into (or out of) the system from the environment. We assume that such transfers are carried out very slowly, so that the system remains approximately in thermodynamic equilibrium at all stages. This process is called thermal equilibrium process. A thermal equilibrium process is represented by a smooth curve on the p-V diagram. In Fig. 10-1, curve ab shows an equilibrium process from state a to state b.

10. 2. 3 The ideal gas law

Experiments show that there is a simple relationship among the macroscopic parameters that characterize a gas in equilibrium state. If we take one mole samples of various gases (oxygen, hydrogen, nitrogen or any others), confine them in containers of identical volume, and hold them at the same temperature, we find that their measured pressure are nearly—though not exactly—the same. If we repeat the measurements at lower gas densities, we find that these small differences in the measurements tend to disappear. Further experiments indicate that, at low enough densities, all real gases tend to obey the relation

$$\frac{pV}{T}=\text{constant} \tag{10-1}$$

Eq. (10-1) holds when the sum mass of the system remains unchanged. At normal temperature and pressure, real gases obey this law quite well; but if a real gas is compressed to an excessively high density, then its behavior will deviate from this law. An ideal gas is a gas that obeys Eq. (10-1) exactly. Although there is no such thing in nature as a truly ideal gas, all gasses approach the ideal state at low enough density, that is, under conditions in which molecules are far enough apart. Thus, **ideal gas is an ideal model** that allows us to gain useful insights into the limiting behavior of real gas.

Suppose p_0, V_0 and T_0 are the parameters of a gas under the standard condition, Eq. (10-1) can be rewritten as

$$\frac{pV}{T}=\frac{p_0V_0}{T_0}$$

where $p_0=1.013\times10^5\,\mathrm{Pa}$, $T_0=273.15\,\mathrm{K}$. If M is the sum mass of the system, and μ is its mass per mole, the number of mole is then M/μ. Because the volume of one mole is identical for all gases, i. e., $V_{\mathrm{mol}}=22.4\times10^{-3}\,\mathrm{m^3/mol}$, the volume of a gas of mass M is therefore $V_0=V_{\mathrm{mol}}(M/\mu)$. Substituting those into Eq. (10-1), we have

$$\frac{pV}{T}=\frac{M}{\mu}\frac{p_0V_{\mathrm{mol}}}{T_0}$$

Because p_0V_{mol}/T_0 is a constant for all gases, called mole gas constant, labeled as R, so that, the equation becomes

$$pV=\frac{M}{\mu}RT \tag{10-2}$$

which is called ideal gas law or state equation of ideal gas, and it holds for equilibrium states. The numerical value of R depends on the units in which p, V, M and T are expressed:

(1) In SI system, the unit of pressure is Pa, $1\,\mathrm{Pa}=1\,\mathrm{N/m^2}$, the unit of volume is $\mathrm{m^3}$, and the temperature is in K, the value of R is therefore

$$R=\frac{p_0V_{\mathrm{mol}}}{T_0}=\frac{1.013\times10^5\,\mathrm{N/m^2}\times22.4\times10^{-3}\,\mathrm{m^3/mol}}{273.15\,\mathrm{K}}=8.31\,\mathrm{J/(mol\cdot K)}$$

(2) Note that heat has the same unit as energy, sometimes we use cal as the unit of energy, 1 J=0.24cal, so that

$$R=8.31\,\mathrm{J/(mol\cdot K)}\times0.24\,\mathrm{cal/J}=2\,\mathrm{cal/(mol\cdot K)}$$

(3) If volume is commonly expressed in liters (L), pressure in atmospheres and temperature in K, thus

$$R=\frac{1\,\mathrm{atm}\times22.4\,\mathrm{L/mol}}{273.15\,\mathrm{K}}=0.082\,\mathrm{atm\cdot L/(mol\cdot K)}$$

Example 10-1 The volume of an oxygen tank is 50 L. As oxygen is withdrawn from the tank, the reading of a pressure gauge drops from $2.17\times10^6\,\mathrm{Pa}$ to $7.9\times10^5\,\mathrm{Pa}$ and the temperature of the gas remaining in the tank drops from 30℃ to 10℃.

(1) How many kilograms of oxygen were withdrawn?

(2) How many kilograms of oxygen were in the tank originally?

(3) What volume would be occupied by the oxygen withdrawn from the tank at a pressure of 1atm and a temperature of 20℃?

Solution (1) Let us express volume in cubic meters, and temperature in Kelvins. Thus 50 L=0.05 $\mathrm{m^3}$, 30 L=0.3 $\mathrm{m^3}$, 30℃=303 K, and 10℃=283 K. According to Eq. (10-2) the initial number of moles is

$$n_1 = \frac{p_1 V}{RT_1} = \frac{2.17 \times 10^6\ \text{Pa} \times 0.05\ \text{m}^3}{8.31\ \text{J/(mol} \cdot \text{K)} \times 303\ \text{K}} = 43.1\ \text{mol}$$

The original mass is therefore

$$m_1 = 43.1\ \text{mol} \times 32\ \text{g/mol} = 1379\ \text{g} = 1.379\ \text{kg}$$

(2) The number of moles remaining in the tank is

$$n_2 = \frac{p_2 V}{RT_2} = \frac{7.9 \times 10^5\ \text{Pa} \times 0.05\ \text{m}^3}{8.31\ \text{J/(mol} \cdot \text{K)} \times 283\ \text{K}} = 16.8\ \text{mol}$$

And the mass remaining is

$$m_2 = 16.8\ \text{mol} \times 32\ \text{g/mol} = 538\ \text{g} = 0.538\ \text{kg}$$

The mass withdrawn is, therefore

$$\Delta m = m_1 - m_2 = 1.379\ \text{kg} - 0.538\ \text{kg} = 0.841\ \text{kg}$$

(3) The number of moles withdrawn is

$$\Delta n = 43.1\ \text{mol} - 16.8\ \text{mol} = 26.3\ \text{mol}$$

And the volume occupied by the oxygen withdrawn would be

$$V = \frac{nRT}{p} = \frac{26.3\ \text{mol} \times 8.31\ \text{J} \cdot \text{mol/K} \times 293\ \text{K}}{1.01 \times 10^5\ \text{Pa}} = 0.634\ \text{m}^3$$

10.3 Essential Concepts of the Kinetic Theory of Gases

The hypothesis supported by thousands of physical and chemical observations is that all matter is composed of tiny particles called molecules. The following essential concepts of kinetic theory of gases are given on the basis of the experimental observations.

10.3.1 All matters consist of a very large number of molecules, molecules are separated

Experimental evidence for assuming that the gas contains many molecules is related to the determination of Avogadro's number:

A mole of any pure substance contains a definite number of identical molecules. The number of molecules in a mole is called Avogadro's constant, denoted by N_0,

$$N_0 = 6.022 \times 10^{23}\ \text{mol}^{-1}$$

The enormously large value of the Avogadro's constant suggests how tiny and how numerous atoms must be in a mole of air. Yet, if these molecules were spread uniformly over the surface of the earth, there would be about 120 000 of them in every square centimeter.

The fact that all matter is compressible indicates that there are separations between molecules. For instance, air can be pumped into the tire of the bicycle or car; the gas enclosed in a cylinder can be compressed and its volume can be decreased to a small part, say, one seventh of the original volume, etc. The fact that a drop of red ink diffuses into the water until uniformly mixture is also an evidence that the molecules are separated.

In other parts of physics and chemistry it is important to consider the structure of the molecule, but it is not necessary at this point. The smallest molecules are of the order of 10^{-10} m in size; the largest is at least 10 000 times of this order.

10.3.2 Molecules are constantly in random motion

Experimental evidence for the random motion of molecules is based on the observation of what we call Brownian motion. In 1827, the English botanist Robert Brown used a microscope to observe the motion of pollen grains suspended in water. The pollen appeared to dance about in an erratic fashion. At first, Brown thought that the pollen was alive and that its motion was some sort of dance! Subsequent studies of liquid suspensions of various inanimate particles convinced observers that the liquid itself was responsible for the erratic motion. Fig. 10-2 shows an example of such Brownian motion.

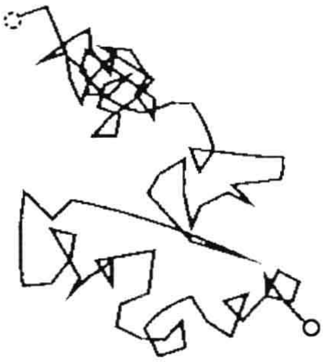


Fig. 10-2 The trace of the Brownian motion of a pollen grain

Eventually, the irregular motion of the suspended particles was explained in detail by assuming that the liquid was composed of molecules in random motion. The suspended particles are continuously bombarded on all sides by the molecules of the fluid. The numbers of molecules striking opposite sides of the particle in any short time interval, being determined by chance, will not be exactly equal. Because of these fluctuations, a randomly directed unbalanced force will act on the suspended particle, accounting for its "Brownian dance".

It is evident from experiments that the "random motion" is related to the temperature: as temperature increases, molecule motion becomes more vigorous. So, the random motion of molecules is also called heat motion.

10.3.3 There is interaction between molecules

One essential characteristic of a molecule is the force that exists between it and a neighboring molecule. There is, of course, a force of gravitational attraction between every pair of molecules, but it turns out that this is negligible in comparison with the forces to be considered now. The forces that hold the molecules of a liquid (or solid) together are chiefly of electrical origin and do not follow a simple inverse-square law.

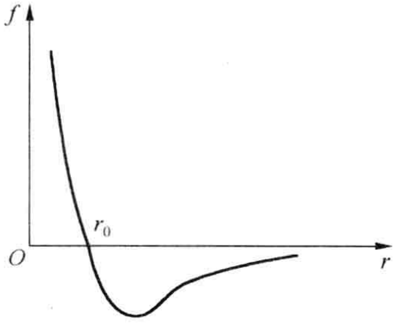


Fig. 10-3 The interaction force between molecules varies with the distance between them

The interaction force is related to the separation between molecules. When the separation of the molecules is large, as in a gas, the force is extremely small and attractive. The attractive force increases as the gas is compressed and its molecules are brought closer together. But since a tremendous pressure is needed to compress a liquid (i. e., to force its molecules closer together than their normal spacing in the liquid state), we conclude that at separations only slightly less than the dimension of a molecule the force is repulsive and relatively large. Thus the force must vary with separation in somewhat the fashion as shown in Fig. 10-3. At large separations the force is small and attractive. As the molecules are brought closer together, the force of attraction becomes larger, passes through a maximum, and then decreases to zero at a separation r_0. When the distance between the molecules is less than r_0, the force becomes repulsive. The order of the distance r_0 is about 10^{-10} m, and when the distance between molecules is larger than 10^{-9} m, the interaction force becomes so small that it can actually be neglected, so, the force between molecules is short-range force. Note that in Fig. 10-3, the attractive force is represented as negative while the repulsive force as positive.

10. 4 The Representation of Pressure for Ideal Gas

The first kinetic theory problem here is to find the connection between the macroscopicparameter of a gas—pressure and the microscopic parameter, the speed of the molecules of the gas. We shall use the statistical method by considering the average motion of the vast number of molecules in the system.

10. 4. 1 The microscopic model of an ideal gas

In order to derive the pressure equation, we must introduce a microscopic model of an ideal gas, including the following assumptions:

(1) The size of molecules may be thought of as so small that it can be considered negligible in comparison with their average distance.

(2) Molecules exert no force to each others except for the instantaneous impulsive force during the collisions with the wall of the container holding the gas and the collisions with each others.

(3) Molecules are in constantly random motion, they collide with one another and also with the walls of the container frequently, and these collisions are perfectly elastic.

(4) The motion of an individual molecule obeys Newton's laws of motion.

We can make a conclusion about this microscopic model that the molecule of an ideal gas is like infinitesimal elastic particle having no interaction with the others.

It is necessary to introduce **a statistical assumption for the great number of molecules in random motion: Under equilibrium state, molecules have equal possibilities (opportunities) to move in all different possible directions.** There is no any preference of direction of motion to the others, that is, the numbers of molecules moving in all different directions are equal; otherwise, the molecules would be concentrated at a certain part of the container, this is opposite to the assumption for the system in equilibrium state. When a gas is in equilibrium state, its density is all the same everywhere in the container. The statistical meaning of this assumption is that, as a consequence of the random motion, the average values of the squares of the molecule's velocity components are equal, that is

$$\overline{v_x^2} = \overline{v_y^2} = \overline{v_z^2} \tag{10-3}$$

10. 4. 2 Derivation of pressure equation of ideal gas

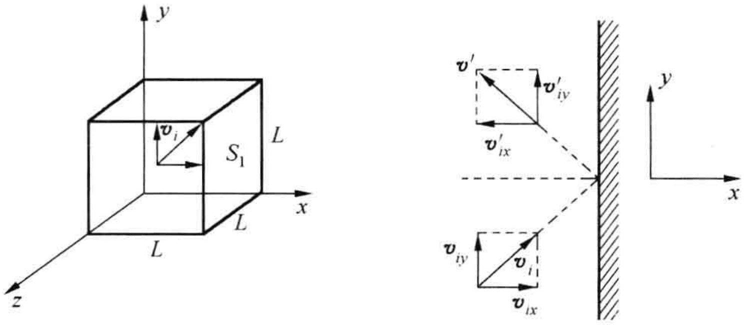
Now, we are ready to discuss the pressure of a gas against the walls of its container, which is due to the impacts of the molecules on the walls during the collision. Suppose there are N molecules in a cube box of size L^3 as shown in Fig. 10-4(a). The molecules with varying speeds in the box are moving in all directions, bumping into each other and bumping from the walls of the box. We ignore (for the time being) collisions of the molecules with each other and consider only collisions with the walls.

1. The impulse exerted by one molecule during each collision

Fig. 10-4(a) shows a typical molecule, say the ith molecule with velocity $\boldsymbol{v}_i$. We can resolve this velocity into components v_{ix}, v_{iy} and v_{iz}, parallel to edges of the box assumed as a cube of size L. Since the collisions with the wall are elastic, after collision, the molecule will rebound from the wall S_1 with the x component of its velocity reversed from v_{ix} to $-v_{ix}$, the other two components being unchanged as shown in Fig. 10-4(b); the change in the molecule's momentum will then be

$$-mv_{ix}-(mv_{ix})=-2mv_{ix}$$

in which m is the mass of a single molecule. According to momentum-impulse theorem and Newton's third law, we see that the impulse exerted to the molecule by the wall S_1 in $-x$ direction is $-2mv_{ix}$, so that, the impulse exerted to the wall S_1 by the molecule during each collision is therefore equal to $2mv_{ix}$ in $+x$ direction.



(a) (b)

Fig. 10-4

(a) A molecule of mass m at velocity v_i is about to collide with the wall S_1 at $x=L$.

(b) The x component of momentum changes from mv_{ix} to $mv'_{ix}=-mv_{ix}$

2. The rate of collision by a single molecule with the wall

A single molecule will hit the wall S_1 ($x=L$) repeatedly, because the collisions with other walls of the box do not affect v_x. The time taken between collisions being the travel time from S_1 to the opposite wall ($x=0$) and back to S_1 again, is $\Delta t=2L/v_{ix}$, thus the rate of collision by this single molecule with the wall S_1 is $1/\Delta t=v_{ix}/2L$.

3. The average force exerted by a single molecule to the wall

The impulse exerted by this molecule to the wall S_1 in a unit time is therefore

$$f_{ix}=2mv_{ix}\cdot\frac{v_{ix}}{2L}=\frac{mv_{ix}^2}{L}$$

From the definition of impulse $dI_x=f_x dt$, we see that it is $mv_{ix}{}^2/L$ that is the average force exerted by a single molecule to the wall S_1 at $x=L$.

4. The pressure on the wall by the gas

The force exerted by a single molecule is not observable, and the gauge pressure is related to the combined effect of many such collision forces, which is the total force F_x on the wall. To find this force, we must add up the contributions of all the other molecules that strike this face, that is

$$F_x = \frac{mv_{1x}^2}{L} + \frac{mv_{2x}^2}{L} + \frac{mv_{3x}^2}{L} + \cdots$$

Dividing the total force by the area of wall ($=L^2$) gives the pressure on that wall, which is the quantity we seek, thus

$$p = \frac{F_x}{L^2} = m\left(\frac{v_{1x}^2 + v_{2x}^2 + v_{3x}^2 + \cdots}{L^3}\right) \tag{10-4}$$

Because there are N molecules in the box, Eq. (10-4) can be rewritten as

$$p = \frac{Nm}{L^3}\left(\frac{v_{1x}^2 + v_{2x}^2 + v_{3x}^2 + \cdots}{N}\right) \tag{10-5}$$

The quantity in brackets is the average value of the square of the x component of the molecule's speed, symbolized by $\overline{v_x^2}$. Then Eq. (10-5) becomes

$$p = \frac{Nm}{L^3}\overline{v_x^2} \tag{10-6}$$

Because N is the total number of molecules in the gas and L^3 is the volume of the container, so that, the number of molecules in per unit volume is then

$$n = \frac{N}{L^3}$$

Eq. (10-6) becomes

$$p = nm\overline{v_x^2} \tag{10-7}$$

For any molecule, $v^2 = v_x^2 + v_y^2 + v_z^2$, so we have

$$\frac{\sum v_i^2}{N} = \frac{\sum v_{ix}^2}{N} + \frac{\sum v_{iy}^2}{N} + \frac{\sum v_{iz}^2}{N}$$

that is

$$\overline{v^2} = \overline{v_x^2} + \overline{v_y^2} + \overline{v_z^2} \tag{10-8}$$

Combining Eq. (10-8) with Eq. (10-3), $\overline{v_x^2} = \overline{v_y^2} = \overline{v_z^2}$, the statistic assumption about the ideal gas, leads to

$$\overline{v_x^2} = \overline{v_y^2} = \overline{v_z^2} = \frac{1}{3}\overline{v^2} \tag{10-9}$$

The molecule's translational kinetic energy at any instant is $\frac{1}{2}mv^2$, and the average translational kinetic energy over all molecules in the gas is $\frac{1}{2}m\overline{v^2}$, or $\bar{\varepsilon}_t = \frac{1}{2}m\overline{v^2}$, so Eq. (10-7) can be rewritten as

$$p = \frac{1}{3}nm\overline{v^2} \tag{10-10}$$

or

$$p = \frac{2}{3}n\left(\frac{1}{2}m\overline{v^2}\right) = \frac{2}{3}n\varepsilon_t \tag{10-11}$$

which is the representation of the pressure in kinetic theory of ideal gas. It is very much in the spirit of kinetic theory. It tells us how the pressure of the gas (a purely macroscopic quantity) depends on the average translational kinetic energy of the molecules (a purely microscopic quantity). The pressure is the statistical average effect of the collisions on the wall of the container by the huge number of molecules.

The steps we took to derive the pressure equation show a typical statistical method to establish the relationship between a macroscopic quantity and the average of a microscopic quantity. Eq. (10-11) indicates that pressure, as a macroscopic parameter of gas, is a statistic average quantity, because $\frac{1}{2}m\overline{v^2}$ and n in this equation have definite meaning only for a great number of molecules, otherwise they would be fluctuating and uncertain. Therefore, pressure has definite meaning only for a system that consists of a vast number of molecules.

10.5 The Average Translational Kinetic Energy and Temperature

Let us now to find the relationship between the temperature, a macroscopic parameter of a gas and the average translational kinetic energy, a microscopic parameter of the molecules in the gas.

From Eq. (10-2), the ideal gas law is

$$pV = \frac{M}{\mu}RT$$

where M is the total mass of the gas, μ is the mass per mole of that gas. If N is the total number of molecules in the gas, N_0 represents the number of molecules per mole (Avogadro's Constant), and m is the mass of a single molecule, thus $M=Nm$ and $\mu = N_0 m$. Substituting them into ideal gas law leads to

$$p = \frac{N}{V}\frac{R}{N_0}T \tag{10-12}$$

Because R and N_0 are two constants, R/N_0 is also a constant and called Boltzmann constant, labeled by k. Sometimes it is called the gas constant for a single molecule (rather than for a mole), and

$$k = \frac{R}{N_0} = \frac{8.31\,\mathrm{J/(mol \cdot K)}}{6.022 \times 10^{23}\,\mathrm{mol^{-1}}} = 1.38 \times 10^{-23}\,\mathrm{J/K}$$

Note that $N/V = n$ is the number of molecules per unit volume (that is the density of number of molecules), Eq. (10-12) can be written as

$$p = nkT \tag{10-13}$$

Comparing Eq. (10-13) with Eq. (10-11), we have

$$\frac{1}{2}m\overline{v^2} = \frac{3}{2}kT \tag{10-14a}$$

or

$$\overline{\varepsilon_t} = \frac{3}{2}kT \tag{10-14b}$$

The macroscopic parameter, temperature T is connected with the microscopic parameter, the average translational kinetic energy $\overline{\varepsilon_t}$ by Eq. (10-14). Like the equation of pressure, it is another essential relation in kinetic theory of ideal gas. Eq. (10-14) tells us the statistic meaning of the temperature in molecular view, which is perhaps an unexpected thing:

> **At a given temperature T, all gas molecules no matter what their masses are—have the same average translational kinetic energy, namely $\frac{1}{2}m\overline{v^2}=\frac{3}{2}kT$.**
>
> **When we measure the temperature of a gas, we are measuring the average translational kinetic energy of its molecules.**

The average translational kinetic energy per molecule, therefore, depends only on the temperature and not on the pressure, volume, or species of molecules. The higher the temperature is, the larger the average kinetic energy is, and the more vigorous the random motion of the molecules in the gas is. Therefore, the temperature is also the measurement of the level of the molecules' motion. On the other hand, by the same argument as that for pressure, temperature is also a statistic average quantity. It has definite meaning only when the system consists of a great number of molecules, so, it does not make any sense to say temperature for a single or a few molecules.

There is one more thing that may be worth to note, that is, actually, nothing in above derivation has specified that we are dealing with molecules. They could be any objects, for example, the tiny pollen grains in Brown's experiment. We see here the basis for the Brownian motion. A pollen grain, suspended in water and in thermal equilibrium, behaves like a very large molecule and has the same translational kinetic energy as do the water molecules that surround it. Because of its very much larger mass, of course, the pollen grain has a correspondingly smaller speed, small enough to make its motion observable.

Example 10-2 Calculate, in the standard state:

(1) What is the average translational kinetic energy of the oxygen molecules and the hydrogen molecules?

(2) How many molecules are there in 1 m^3 gas?

Solution Because the average translational kinetic energy depends only on the temperature but not on the nature of the gas, so, the answer is the same for two gasses:

(1) $\overline{\varepsilon_t}=\frac{3}{2}kT=\frac{3}{2}\times 1.38\times 10^{-23}\ \text{J/K}\times 273\ \text{K}=5.65\times 10^{-21}\ \text{J}$

(2) From equation of $p=nkT$, in which n is the number of molecules in 1 m^3 gas, we have

$$n=\frac{p}{kT}=\frac{1.013\times 10^{5}\ \text{N/m}^2}{1.38\times 10^{-23}\ \text{J/K}\times 273\ \text{K}}=2.69\times 10^{25}\ \text{m}^{-3}$$

Example 10-3 $\sqrt{\overline{v^2}}$ is defined as the root-mean-square speed. Find $\sqrt{\overline{v^2}}$ for oxygen molecules and for hydrogen molecules at 273 K.

Solution From equation of $\frac{1}{2}m\overline{v^2}=\frac{3}{2}kT$, we have the root-mean-square speed

$$\sqrt{\overline{v^2}}=\sqrt{\frac{3kT}{m}}=\sqrt{\frac{3RT}{mN_0}}=\sqrt{\frac{3RT}{\mu}} \tag{10-14c}$$

which depends not only on T but also on m, the mass of a molecule of the particular gas.

$$\text{For oxygen } \sqrt{\overline{v^2}} = \sqrt{\frac{3 \times 8.31 \times \text{J/(mol} \cdot \text{K)} \times 273\ \text{K}}{32 \times 10^{-3}\,\text{kg}}} = 461\ \text{m/s}$$

$$\text{For hydrogen } \sqrt{\overline{v^2}} = \sqrt{\frac{3 \times 8.31 \times \text{J/(mol} \cdot \text{K)} \times 273\ \text{K}}{2 \times 10^{-3}\,\text{kg}}} = 1845\ \text{m/s}$$

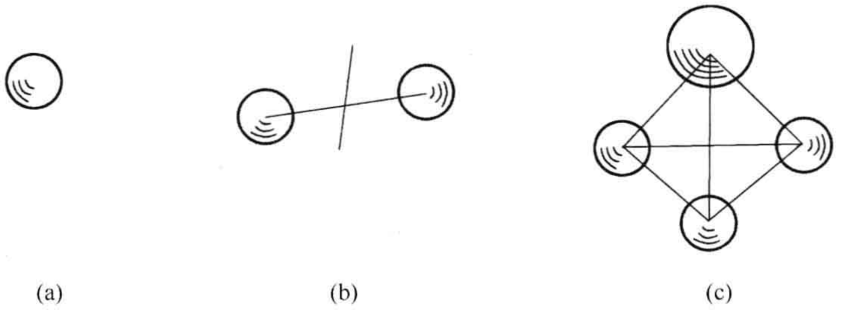
The answers of these examples give us the basic idea about the order of some microscopic parameters of gases.

10.6 The Equipartition Theorem of Energy

10.6.1 Modification of the model of molecules

When we discuss the translational kinetic energy, we have assumed that the molecules behave like point-like particles. But if we consider the total kinetic energy of a molecule, we must take the possibility that molecules with more than one atom can store energy in other forms into account. Therefore, the model of ideal gas needs to be modified according to the number of atoms in the particular molecule.

If we ignore (for the time being) the interaction between any two atoms of a molecule, we can give the simplified kinetic theory models for monatomic molecule, diatomic molecule, and polyatomic molecule as shown respectively in Fig. 10-5(a), (b) and (c).



(a) (b) (c)

Fig. 10-5 (a) Particle model of monatomic molecule; (b) Rigid model of diatomic molecule; (c) Rigid model of polyatomic molecule

On the basis of these models, we can believe that monatomic molecules, such as helium (He) and argon (Ar) which are essentially point-like (Fig. 10-5(a)) and have only a very small rotational inertia about any axis-can store energy only in their translational motion. Diatomic molecules, such as oxygen (O_2), hydrogen (H_2) and nitrogen (N_2), can be regarded as two point-like particles rigidly connected together (like a dumbbell, see Fig. 10-5(b)). The molecule may rotate about either of the two axes through the center of mass that the line joining atoms or perpendicular to the line joining atoms. Polyatomic molecules, such as ammonia (NH_3), methane (CH_4) and water vapor (H_2O), have more than two atoms which are regarded as connected together rigidly too (like a rigid body, see Fig. 10-5(c)), and their rotational motions are even more complicated than those of diatomic ones. If the latter two kinds of molecules collide with other molecules or with the wall of the container, they will usually start rotating about their center of mass. We therefore expect that, on the average, they should be able to store substantial additional amounts of energy by rotating.

How do we take these possibilities quantitatively into account to find the total kinetic energy of a molecule? Let us firstly introduce the concept of degree of freedom.

10. 6. 2 The degrees of freedom of molecule

The number of degrees of freedom is defined as the independent coordinates introduced to determine the position of a moving body in space.

For translational motion, there are three degrees of freedom, corresponding to the three independent axes, say x, y and z. A monatomic molecule can only do translational motion, so the number of degrees of freedom for a monatomic molecule is three and labeled by t to specify the translational degrees of freedom, that is, $t = 3$. Besides, diatomic and polyatomic molecules can under go translational motion also. They all have three translational degrees of freedom.

For rotational motion, a monatomic molecule has no degree of freedom. A diatomic molecule of rigid dumbbell model has two rotational degrees of freedom corresponding to the two independent angular coordinates, say α and β to determine the orientation of the line connecting the nuclei of its two constituent atoms. Here we have considered the relation of $\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1$ connecting three angular coordinates α, β, and γ to orientate a straight line in the space. The number of rotational degrees of freedom is labeled as r. So $r = 2$ for a diatomic molecule. For a polyatomic molecule of rigid body model, besides α and β which determine the orientation of a given straight line fixed in the molecule, one more independent coordinate, φ is needed to describe the angular position of the molecule rotating about that line, so that the number of rotational degrees of freedom, $r = 3$ for a polyatomic molecule.

We can make a conclusion about the total numbers of degrees of freedom, labeled as i: $i = 3$ for monatomic molecules, $i = 5$ for diatomic ones and $i = 6$ for polyatomic ones, which are listed in Table 10-1.

Table 10-1 Degrees of freedom for various molecules

Molecule Examples	Number of Degrees of Freedom		
	Translational	Rotational	Total
	(t)	(r)	(i)
Monatomic He, Ar, Kr etc.	3	0	3
Diatomic O_2, H_2, N_2 etc.	3	2	5
Polyatomic NH_3, CH_4, H_2O etc.	3	3	6

10. 6. 3 The equipartition theorem of energy

We have proved that the average translational kinetic energy is

$$\overline{\varepsilon_t} = \frac{1}{2}m\overline{v^2} = \frac{3}{2}kT$$

From Eq. (10-3)

$$\overline{v_x^2} = \overline{v_y^2} = \overline{v_z^2} = \frac{1}{3}\overline{v^2}$$

we have

$$\frac{1}{2}m\overline{v_x^2}=\frac{1}{2}m\overline{v_y^2}=\frac{1}{2}m\overline{v_z^2}=\frac{1}{3}\left(\frac{1}{2}m\overline{v^2}\right)=\frac{1}{2}kT \tag{10-15}$$

in which, $\frac{1}{2}m\overline{v_x^2}$, $\frac{1}{2}m\overline{v_y^2}$ and $\frac{1}{2}m\overline{v_z^2}$ are corresponding to the three translational degrees of freedom respectively. So, Eq. (10-15) means that on the average, each translational degree of freedom has equal kinetic energy of $\frac{1}{2}kT$ per molecule associate with it.

Furthermore, this distribution of translational energy can be extended to the other form of motion, because during every time of collisions, the kinetic energy can be transferred from one molecule to another, from one translational degree of freedom to another, and also can be transferred from a translational degree of freedom to a rotational degree of freedom, and so on. Due to the too much frequent and random collisions in the thermal motion of so great number of molecules, we can believe that, from the point of statistic view, all possible transformations of energy have an even chance, not any degree of freedom be preferable to the others. Consequently, it turns out that Eq. (10-15) is true not only for the components of translational motion, but also for rotational motion. The general conclusion is known as the equipartition theorem of energy:

Each degree of freedom of a molecule is associated with—on an average—a kinetic energy of $\frac{1}{2}kT$.

Put in the other words, each translational or rotational component of the random thermal motion has an average kinetic energy of $\frac{1}{2}kT$ for one molecule. According to this theorem, if the total number of degrees of freedom for a molecule is i, its average total kinetic energy is then

$$\overline{\varepsilon_k}=\frac{i}{2}kT \tag{10-16}$$

Substituting the quantity of i in Table 10-1 into Eq. (10-16), we obtain

For a monatomic molecule $\overline{\varepsilon_k}=\frac{3}{2}kT$

For a rigid diatomic molecule $\overline{\varepsilon_k}=\frac{5}{2}kT$

For a rigid polyatomic molecule $\overline{\varepsilon_k}=3kT$

Note that in the above general argument we have ignored the vibration motions of the atoms in the molecules with more than one atom. The inter-atomic forces do not really hold these atoms rigidly; rather, the forces act somewhat like springs and then produce a restricted back-and-forth vibration of the atoms about their equilibrium positions. The precise derivation about the vibration energy of a molecule lies beyond the realm of classical physics; it lies in the realm of quantum physics where it is established that rotations of atoms about their own axis and vibrations of atoms in a molecule do not occur unless the temperature is rather high, say 400℃ or more. As we shall see in the discussion about the heat capacities of ideal gas (in chapter 11), the energies calculated from Eq. (10-16) actually agree pretty well with experiments, provided that we do not exceed this temperature limit. We shall make use of the equipartition theorem in the rest of thermodynamics of this course.

10.6.4 Internal energy of ideal gas

Internal energy is a macroscopic parameter introduced to describe the energy of gas. It is the sum of the total kinetic energy and interaction potential energy of all molecules in the gas. For ideal gas, ignoring the interaction between the molecules means that there is no potential energy in the gas; therefore, the internal energy of ideal gas is simply the sum of the kinetic energies of all molecules. From Eq. (10-16), the internal energy of one mole ideal gas is obtained as

$$E_0 = N_0\left(\frac{i}{2}kT\right) = \frac{i}{2}RT \tag{10-17}$$

For the ideal gas with mole mass of μ and sum mass of M, the internal energy equals then

$$E = \frac{M}{\mu}\frac{i}{2}RT \tag{10-18}$$

where M/μ is the number of moles in the system.

Eq. (10-18) shows that the internal energy of a given ideal gas depends only on the thermodynamic temperature T, and not on the pressure and volume of the gas. Moreover, as long as the change in temperature is the same during any thermal processes, the amount of the change in internal energy is simply the same, no matter what the particular process is. That is the internal energy is a function of the state of the system, just as p, V and T are .

For a real gas, the interaction force between molecules could not be ignored any more. Besides the kinetic energies, there are also potential energies between molecules, which relate to the separations of molecules, that is, relate to the volume of the gas. Therefore, the internal energy of real gas is the function of both temperature and volume of the gas.

Example 10-4 Calculate the average rotational kinetic energy $\overline{\varepsilon_r}$, total kinetic energy $\overline{\varepsilon_k}$ of a molecule, and the internal energy of one mol gas at temperature $T = 300$ K for (1) argon; (2) nitrogen; (3) carbon dioxide.

Solution (1) For Ar or any other monatomic molecules $r = 0$, $i = t = 3$, we have

$$\overline{\varepsilon_r} = \frac{r}{2}kT = 0$$

$$\overline{\varepsilon_k} = \frac{i}{2}kT = \frac{3}{2}\times 1.38\times 10^{-23}\,\text{J/K}\times 300\text{K} = 6.21\times 10^{-21}\,\text{J}$$

$$E = \frac{M}{\mu}\frac{i}{2}RT = 1\times\frac{3}{2}\times 8.31\times \text{J/mol}\times 300\text{K} = 3.74\times 10^{3}\,\text{J}$$

(2) For N_2 or any other diatomic molecules $r = 2$, $i = t + r = 5$, we obtain

$$\overline{\varepsilon_r} = \frac{r}{2}kT = 1.38\times 10^{-23}\,\text{J/K}\times 300\text{K} = 4.14\times 10^{-21}\,\text{J}$$

$$\overline{\varepsilon_k} = \frac{i}{2}kT = \frac{5}{2}\times 1.38\times 10^{-23}\,\text{J/K}\times 300\text{K} = 10.4\times 10^{-21}\,\text{J}$$

$$E = \frac{M}{\mu}\frac{i}{2}RT = 1\times\frac{5}{2}\times 8.31\text{J/mol}\times 300\text{K} = 6.23\times 10^{3}\,\text{J}$$

(3) For CO_2 or any other polyatomic molecules $r = 3$, $i = t + r = 6$, we have

$$\overline{\varepsilon_r} = \frac{r}{2}kT = \frac{3}{2}kT = 6.21 \times 10^{-21}\,\mathrm{J}$$

$$\overline{\varepsilon_k} = \frac{i}{2}kT = 3kT = 12.4 \times 10^{-21}\,\mathrm{J}$$

$$E = \frac{M}{\mu}\frac{i}{2}RT = 1 \times 3 \times 8.31\,\mathrm{J/mol} \times 300\,\mathrm{K} = 7.48 \times 10^{3}\,\mathrm{J}$$

10.7 The Mean Free Path

In the discussion of equipartition of energy, we have seen that the collisions between molecules play a crucial role in transformation of energy (and also momentum). It is by the frequent collisions that an equilibrium state can be established and maintained in a gas. Actually, collisions play even more important role in other thermal phenomena.

You can "feel" the effect of collisions in a gas in this daily experience: if someone open the cap of a bottle of perfume in a corner of a room, you are apart from it just 10 meters, but you do not smell the perfume until a couple of minutes later. Considering that on the average, the molecular speed is about $10^2 \sim 10^3\,\mathrm{m/s}$ at room temperature (see Example. 10-2), you may wonder why the perfume molecules take so long time to cross such short distance with so large speed? The reason is that a molecule encounters many too many collisions in just one second, changing both direction and magnitude of velocity abruptly as it collides with other molecules, so that the path of the molecule consists of a series of zigzags as the path from point A to point B shown in Fig. 10-6. Although the other molecules are shown in stationary positions, they are also moving in a similar fashion. The distance between one collision and the next called free path, is a straight line of erratic length.

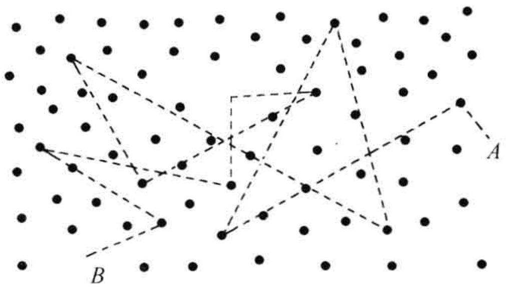


Fig. 10-6 The zigzag path of a molecule traveling through a gas

To describe this random motion quantitatively, we introduce a useful parameter, the mean free path, designated by $\bar{\lambda}$. The mean free path is the average distance a molecule travels between one collision and the next. While another parameter, the average collision rate of a molecule, labeled by $\bar{Z}$, is the average number of collisions per unit time a molecule suffers as it moves through the gas. Suppose that the average speed of molecules is represented by $\bar{v}$, thus we have the following relation

$$\bar{v} = \bar{\lambda}\,\bar{Z} \tag{10-19}$$

where $\bar{Z}$ is also the average number of segments of the zigzag passed per unit time by a molecule, $\bar{\lambda}$ is the average length of segments, so that $\bar{Z}\,\bar{\lambda}$ represents the sum of distances traveled by a molecule per unit time, i. e. the average speed $\bar{v}$.

To find the formula of collision rate of the molecules of a gas, we adopt a crude model in which the molecules are hard elastic spheres of some given diameter d being the minimum distance between two molecules' centers. This means that two molecules exert no mutual forces unless their centers come to within a distance d, so d is also called the effective diameter of a molecule. Let us now focus on a single molecule, suppose that it moves with the average speed $\bar{v}$ and that all the other molecules are at rest. Later we shall relax this assumption.

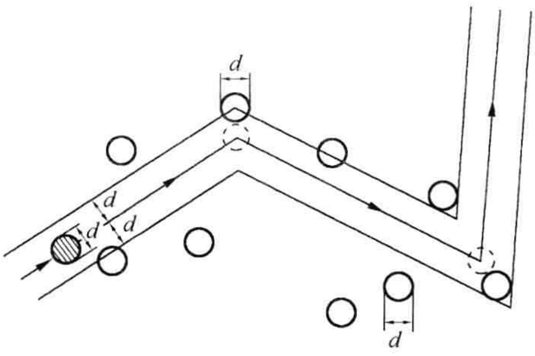


Fig. 10-7 Imaginary cylinder of diameter d surrounding the path of a molecule

Imagine that the molecule moves through the gas, sweeping out a jaggedly broken cylinder whose cross-section area is πd^2 (called as target area), and engaging in a collision whenever the center of another molecule lies within this cylinder as shown in Fig. 10-7. The molecule travels a distance $\bar{v}$ in one second, the stretched-out length of our zigzag cylinder is $\bar{v} \times 1$ and the volume of the cylinder is $(\pi d^2)(\bar{v} \times 1)$. The number of collisions that occur in one second is then equal to the number of molecules whose centers lie within this cylinder, that is the collision rate $\bar{Z}$,

$$\bar{Z} = n\pi d^2 \bar{v} \tag{10-20}$$

where n is the molecular density, the number of molecules per unit volume. This simple result is approximate because it based on the assumption that all the molecules except one are at rest. In fact, all molecules are moving. When this is taken properly into account, Eq. (10-20) is modified as

$$\bar{Z} = \sqrt{2}\pi d^2 \bar{v} n \tag{10-21}$$

where $\sqrt{2}\bar{v}$ has proved to be the average relative speed $\overline{v_r}$ in statistic physics. Combining Eq. (10-19) and Eq. (10-21), we have

$$\bar{\lambda} = \frac{\bar{v}}{\bar{Z}} = \frac{1}{\sqrt{2}\pi d^2 n} \tag{10-22}$$

which means that $\bar{\lambda}$ is conversely proportional to d^2 and n and independent of $\bar{v}$. Using equation $p = nkT$, we have

$$\bar{\lambda} = \frac{kT}{\sqrt{2}\pi d^2 p} \tag{10-23}$$

It indicates that the mean free path $\bar{\lambda}$ is conversely proportional to pressure p when the temperature remains unchanged.

The data listed in Table 10-2, Table 10-3 and the calculation in Example. 10-5 will help you to get some ideal and to imagine the picture about the random motion of molecules in a gas.

Table 10-2 The $\bar{\lambda}$ and d of some gases under the standard condition

Gas	$\bar{\lambda}$ /m	d /m	Gas	$\bar{\lambda}$ /m	d /m
H_2	1.123×10^{-7}	2.3×10^{-10}	O_2	0.647×10^{-7}	2.9×10^{-10}
N_2	0.599×10^{-7}	3.1×10^{-10}	Air	6.9×10^{-8}	3.5×10^{-10}

Table 10-3 The $\bar{\lambda}$ of air at 273 K under different pressures

Pressure/mmHg	$\bar{\lambda}$/m	Pressure/mmHg	$\bar{\lambda}$/m
760	7×10^{-8}	10^{-4}	5×10^{-1}
1	5×10^{-5}	10^{-8}	50
10^{-2}	5×10^{-3}		

1 mmHg=1.33322×10^{2} Pa.

Example 10-5 Calculate the mean free path $\bar{\lambda}$ and the average collision rate $\bar{Z}$ of air under the standard condition. Suppose that its effective diameter $d=3.5\times10^{-10}$ m and its average speed $\bar{v}=448$ m /s at that condition.

Solution (1) Putting $T=273$ K, $p=1.01\times10^{5}$ N/m^{2}, $d=3.5\times10^{-10}$ m into Eq. (10-23), we have

$$\bar{\lambda}=\frac{kT}{\sqrt{2}\pi d^{2}p}=\frac{1.38\times10^{-23}\times273}{1.41\times3.14\times(3.5\times10^{-10})^{2}\times1.01\times10^{5}}=6.9\times10^{-8}(\mathrm{m})$$

This shows that the mean free path is about 200 times of effective diameter of air under the standard condition.

(2) From Eq. (10-19), we obtain

$$\bar{Z}=\frac{\bar{v}}{\bar{\lambda}}=\frac{448}{6.9\times10^{-8}}=6.5\times10^{9}(\mathrm{s}^{-1})$$

which means that, on the average, under the standard condition a molecule collides with other molecules for multiplying 65 by a hundred million times in just one second. Considering that there is about 10^{25} molecules in 1 m^{3} (Example. 10-2), can you imagine how random the thermal motion in a gas is?

10.8 The Maxwell Speed Distribution

10.8.1 The concept of speed distribution

Although the root-mean-square speed can give us a general idea of the molecular speed at a given temperature, like the calculation in Example 10-3, the magnitude of molecular speed in a gas varies from zero to some quite large value, and the velocity can be in any possible direction. Moreover, the velocity of a molecule changes continuously due to the frequent and erratic collisions. From this point of view, for an individual molecule, both magnitude and direction of velocity are completely random, so it is impossible to determine exactly the molecular speed at any instant. On the other hand, for a given gas system, in its entirety, the distribution of molecular speeds at a given temperature obeys a certain statistic law called the Maxwell distribution of molecular speed, which deduced by the Scottish physicist James Clerk Maxwell in about 1860s.

What is the speed distribution among molecules? For a given sample of gas, suppose that the whole range of the values of molecular speed is divided into many intervals with equal width, say the intervals of 0～100 m/s, 100～200 m/s, 200～300 m/s 300～400 m/s and so on, thus at a given equilibrium state, the fraction (that is, the percentage) of molecules whose speeds lie in every interval is certainly given by means of statistic method. For example, it is found that at 273 K, the fraction of oxygen molecules whose speeds lie in the interval of 0～100 m/s is 1.4%, the fraction in 100～200 m/s is 8.1%, the fraction in 200～300 m/s is 16.5%, etc..

10.8.2 The Maxwell distribution of speed

The distribution of speeds in a gas is described mathematically by the distribution function. Suppose N is the total number of molecules and the number of molecules that have speeds in the range v to $v+\mathrm{d}v$ is $\mathrm{d}N$, thus $\mathrm{d}N/N$ represents the fraction of molecules that have speeds in this small range. Experiments indicate that $\mathrm{d}N/N$ is proportional to the width of interval $\mathrm{d}v$, that is

$$\frac{\mathrm{d}N}{N} = f(v)\,\mathrm{d}v \tag{10-24}$$

where $f(v) = \dfrac{\mathrm{d}N}{N\mathrm{d}v}$ is called distribution function which equals the fraction of molecules having the speeds in the unit speed interval adjacent to v, and specifies the number of molecules per unit speed interval.

Maxwell first solved the problem of finding the speed distribution function of gas molecules, which is

$$f(v) = 4\pi\left(\frac{m}{2\pi kT}\right)^{3/2} \mathrm{e}^{-\frac{mv^2}{2kT}} v^2 \tag{10-25}$$

Here v is the molecular speed, T is the gas temperature, m is the mass of a molecule, and k is Boltzmann constant. The Maxwell speed distribution law then is defined by the product $f(v)\,\mathrm{d}v$, that is

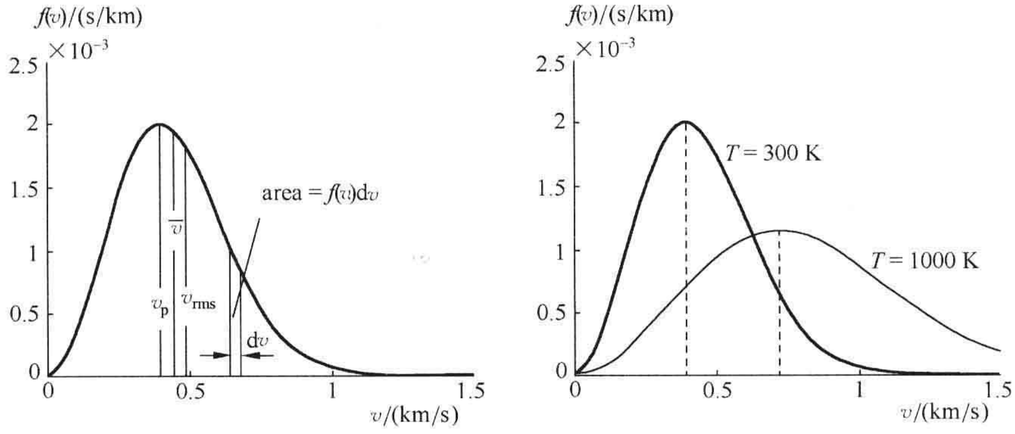
$$\frac{\mathrm{d}N}{N} = f(v)\mathrm{d}v = 4\pi\left(\frac{m}{2\pi kT}\right)^{3/2} \mathrm{e}^{-\frac{mv^2}{2kT}} v^2\,\mathrm{d}v \tag{10-26}$$

which is the fraction of molecules whose speeds lie in the range v to $v+\mathrm{d}v$. Eq. (10-26) holds for all systems in the equilibrium states, if ignoring the interactions between molecules.

Fig. 10-8(a) shows the speed distribution function curve for oxygen molecules at $T=300$ K. This kind of curve is called the Maxwell speed distribution curve. The area of a strip whose height is $f(v)$ and width is $\mathrm{d}v$, is $f(v)\mathrm{d}v = \mathrm{d}N/N$, a dimensionless quantity, which equals the fraction of molecules having speeds distributed in a small range v to $v+\mathrm{d}v$. Similarly, the area under the curve between v_1 and v_2 equals then the fraction of molecules distributed within the range v_1 to v_2. Therefore the total area under the distribution curve corresponds to the fraction of molecules whose speeds lie between zero and infinity. Because all molecules fall into this category so that the value of this total area is unity. Since the total area is the integration $\int_0^\infty f(v)\mathrm{d}v$, so

$$\int_0^\infty f(v)\mathrm{d}v = 1 \tag{10-27}$$

The distribution function $f(v)$ must satisfy Eq. (10-27) which is called the normalizing condition.



(a) (b)

Fig. 10-8

(a) The Maxwell speed distribution curve for oxygen at $T=300$ K

(b) The curve for $T=300$ K is compared with the curve for $T=1000$ K

As we can see from the curve, the molecular speeds spread over broad range. The distribution function fades away as $v \rightarrow 0$ and $v \rightarrow \infty$. This indicates that there are few molecules near zero speed, and few molecules of very large speed. The fraction of molecules having intermediate speeds is larger, that is, there are many molecules distributing in the range of intermediate speeds. The peak of the speed distribution curve corresponds to the most probable speed, labeled as v_p, at which $f(v)$ in Eq. (10-25) has its maximum value. This means that when the molecular speed range $(0 \rightarrow \infty)$ is divided into many equal width intervals, thus the fraction of molecules distributing in the interval centered on v_p, has the maximum value, at the given temperature for the given sample gas. On the other words, the possibility for the molecules to have the speed close to v_p is most.

Keep in mind that when dealing with the distribution of speeds, or any other kind of physical quantity with a continuous range of variation, it is not reasonable to ask how many molecules have a speed v. The only reasonable question is how many molecules have speeds in some specified interval of speeds.

Comparing the curves at the two different temperatures for the same gas in Fig. 10-8(b), we see that the most probable speed increases with temperature increasing and the number of molecules having smaller speeds decreases while the number of molecules having larger speeds increases with temperature. This agrees with the fact that molecules move faster at the higher temperature. Note that the area under each curve should be the same, both equal to a value of unity, so the curve corresponding to higher temperature becomes flatter. The experimental measurements to determine the distribution of molecular speeds are in perfect agreement with the Maxwell distribution.

10.8.3 Three kinds of speed

1. The most probable speed v_p

We can find the most probable speed by requiring that $\frac{df(v)}{dv}=0$ and solving for v. Doing so yields

$$v_p = \sqrt{\frac{2kT}{m}} = \sqrt{\frac{2RT}{\mu}} \tag{10-28}$$

From Eq. (10-28), you can find the effect of molecular mass m on v_p, therefore on the distribution curve.

2. The average speed $\bar{v}$

Suppose dN is the number of molecules whose speeds lie in the interval $v \rightarrow v+dv$, we can make product vdN and add up (or integrate) all these products over the entire range of speeds. Then dividing the sum by the total number N of molecules, we get $\int \frac{v\,dN}{N}$ defined as the average speed $\bar{v}$. Note that $\frac{dN}{N} = f(v)\,dv$, so that

$$\bar{v} = \int \frac{vdN}{N} = \int_0^{\infty} v\,f(v)\,dv \tag{10-29}$$

Substituting for $f(v)$ from Eq. (10-25), after evaluating the integration, leads to

$$\bar{v} = \sqrt{\frac{8kT}{\pi m}} = \sqrt{\frac{8RT}{\pi\mu}} \tag{10-30}$$

3. The root-mean-square speed $\sqrt{\overline{v^2}}$

By means of the same method as in (2) above except that we multiply v^2 by dN, after another integration, we have

$$\overline{v^2} = \int \frac{v^2 dN}{N} = \int_0^{\infty} v^2 f(v)\,dv = \frac{3kT}{m} \tag{10-31}$$

The root-mean-speed, simplified as v_{rms}, is the square root of this quantity, or

$$v_{rms} = \sqrt{\overline{v^2}} = \sqrt{\frac{3kT}{m}} = \sqrt{\frac{3RT}{\mu}} \tag{10-32}$$

Eq. (10-32) is identical to Eq. (10-14c), which we derived earlier in Example. 10-2.

For the same sample gas at the same temperature, we have

$$v_p : \bar{v} : \sqrt{\overline{v^2}} = 1.41 : 1.60 : 1.73$$

that is

$$v_p < \bar{v} < \sqrt{\overline{v^2}}$$

We can make use of each of these speeds in different problems. To calculate the probability of the molecular distribution we use v_p, and to calculate the mean free path we use $\bar{v}$. For calculating mean kinetic energy, thus v_{rms} can be used.

Example 10-6 A container is filled with oxygen gas maintained at 300 K. What fraction of the molecules has speeds in the range 400～420 m/s?

Solution From Eq. (10-26)

$$\frac{dN}{N} = 4\pi \left(\frac{m}{2\pi kT}\right)^{3/2} e^{-\frac{mv^2}{2kT}} v^2\,dv$$

and

$$v_p = \sqrt{\frac{2RT}{\mu}} = \sqrt{\frac{2 \times 8.31 \times 300}{32 \times 10^{-3}}} = 394.7 \approx 395\text{m/s}$$

Eq. (10-26) can be rewritten as

$$\frac{\mathrm{d}N}{N}=\frac{4}{\sqrt{\pi}}\mathrm{e}^{-(\frac{v}{v_\mathrm{p}})^2}\cdot\left(\frac{v}{v_\mathrm{p}}\right)^2\cdot\frac{\mathrm{d}v}{v_\mathrm{p}}$$

When $v=400$ m/s, $\mathrm{d}v=20$ m/s and $v_\mathrm{p}=395$ m/s, we have

$$\frac{\mathrm{d}N}{N}=\frac{4}{\sqrt{\pi}}\mathrm{e}^{-(\frac{400}{395})^2}\times\left(\frac{400}{395}\right)^2\times\frac{20}{395}=0.042=4.2\%$$

10.9 The Boltzmann Distribution

10.9.1 The Maxwell distribution of velocity

Eq. (10-26) gives us only the distribution of molecular speed (i. e. the magnitude of velocity). The probable question is that if we take both direction and magnitude of molecular velocity into account how the distribution of molecules turns out? Now we introduce it as below:

$$\frac{\mathrm{d}N}{N}=\left(\frac{m}{2\pi kT}\right)^{3/2}\mathrm{e}^{-m(v_x^2+v_y^2+v_x^2)/2kT}\,\mathrm{d}v_x\mathrm{d}v_y\mathrm{d}v_z$$

or

$$\frac{\mathrm{d}N}{N}=\left(\frac{m}{2\pi kT}\right)^{3/2}\mathrm{e}^{-mv^2/2kT}\,\mathrm{d}v_x\mathrm{d}v_y\mathrm{d}v_z \tag{10-33}$$

which is the fraction (or ratio) of molecules whose velocities lie in the range $\boldsymbol{v}$ to $\boldsymbol{v}+\mathrm{d}\boldsymbol{v}$, (i. e. the range v_x to $v_x+\mathrm{d}v_x$, v_y to $v_y+\mathrm{d}v_y$, v_z to $v_z+\mathrm{d}v_z$). It remains for the systems at equilibrium states, under the assumption of ignoring the intermolecular actions. Eq. (10-33) is known as Maxwell distribution of molecular velocity. Actually, Eq. (10-26) is deduced from Eq. (10-33) in statistic physics.

10.9.2 The Boltzmann distribution of energy

Here we note the interesting fact that the term $mv^2/2$ in exponent of Eq. (10-33) is just the translational kinetic energy of a molecule or the total energy of a free particle moving without any effect of external force. So, we may put it in this way: the fraction of molecules (or free particles) decreases exponentially with the increase of its kinetic energy, or the probability of finding molecules in a given energy range varies exponentially with the negative value of the kinetic energy in that range, divided by kT.

Boltzmann extended the distribution given by Eq. (10-33) to the general cases of molecular system in some potential field where the molecules are acted on by external forces such as gravitational force or electrostatic force and so on. Therefore, what we noticed in a special case turns out to be turn in general by means of using total energy $\varepsilon=\varepsilon_\mathrm{k}+\varepsilon_\mathrm{p}$ instead of $\varepsilon_\mathrm{k}=mv^2/2$ in Eq. (10-33) and considering not only the velocity range but also the position range, because the potential energy ε_p usually relates to the position of molecules in space, say (x, y, z). Thus, when a gaseous system is in equilibrium state, suppose that there is some kind of force acting on the molecules in the gas, corresponding to a potential energy ε_p. Then the number of molecules whose positions lie in the range x to $x+\mathrm{d}x$, y to $y+\mathrm{d}y$, z to $z+\mathrm{d}z$, and velocities in the range of v_x to $v_x+\mathrm{d}v_x$, v_y to $v_y+\mathrm{d}v_y$, v_z to $v_z+\mathrm{d}v_z$, is determined by

$$\mathrm{d}N = n_0\left(\frac{m}{2\pi kT}\right)^{3/2} \mathrm{e}^{-(\varepsilon_\mathrm{k}+\varepsilon_\mathrm{p})/kT}\,\mathrm{d}v_x\mathrm{d}v_y\mathrm{d}v_z\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{10-34}$$

Here the quantity n_0 is the total number of molecules per unit volume, with all possible velocities, at the position where the potential energy is zero. Eq. (10-34) is called the **Boltzmann distribution of energy, briefly, Boltzmann distribution.** It shows that the number of molecules within a given spacing range varies exponentially with the negative of the total energy of a molecule divided by kT, or, the probability of finding molecules in a given spacing range is proportional to the exponential function $\mathrm{e}^{-\varepsilon/kT}$ which is called Boltzmann factor. Boltzmann distribution tells us that for all those equal ranges, the larger the molecular energy ε is, the less the number of molecules $\mathrm{d}N$, found in the range about that energy.

Boltzmann distribution is a general principle in statistic physics; it holds for any system of particles, such as the atoms and molecules in gases, liquids, solids and for Brownian grains.

We note that in Boltzmann distribution Eq. (10-34), the energy is a continuous variable in classical mechanics, but for subatomic particles governed by quantum mechanics, energy is a series discrete variable corresponding to a discrete value of energy, so the Boltzmann distribution becomes

$$N_\mathrm{i} = c\,\mathrm{e}^{-\varepsilon_\mathrm{i}/kT} \tag{10-35}$$

where N_i is the number of particles on the energy level ε_i and c is a constant.

10.9.3 The distribution of particle height in gravitational field

As an application of Boltzmann distribution, let us consider a gaseous system in a gravitational field. Suppose the z axis is chosen up forward, thus the potential energy of a particle is $\varepsilon_\mathrm{p} = mgz$, the total energy is $\varepsilon = \frac{1}{2}mv^2 + mgz$ and Eq. (10-34) becomes

$$\mathrm{d}N = n_0\left(\frac{m}{2\pi kT}\right)^{3/2} \mathrm{e}^{-\left(\frac{1}{2}mv^2+mgz\right)/kT}\,\mathrm{d}v_x\mathrm{d}v_y\mathrm{d}v_z\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{10-36}$$

Considering all possible values of velocity, we can get the number of molecules at height z in the volume element $\mathrm{d}V = \mathrm{d}x\mathrm{d}y\mathrm{d}z$ (i. e. in the range x to $x+\mathrm{d}x$, y to $y+\mathrm{d}y$, z to $z+\mathrm{d}z$) by integrating Eq. (10-36) over the entire range of velocity

$$\mathrm{d}N' = n_0\iiint_{-\infty}^{+\infty}\left(\frac{m}{2\pi kT}\right)^{3/2} \mathrm{e}^{-\varepsilon_\mathrm{k}/kT}\,\mathrm{d}v_x\mathrm{d}v_y\mathrm{d}v_z \cdot \mathrm{e}^{-mgz/kT}\,\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{10-37}$$

in which $\left(\frac{m}{2\pi kT}\right)^{3/2} \mathrm{e}^{-\varepsilon_\mathrm{k}/kT}$ is the distribution function of velocity, and should satisfy the normalizing condition, i. e.

$$\iiint_{-\infty}^{+\infty}\left(\frac{m}{2\pi kT}\right)^{3/2} \mathrm{e}^{-\varepsilon_\mathrm{k}/kT}\,\mathrm{d}v_x\mathrm{d}v_y\mathrm{d}v_z = 1 \tag{10-38}$$

So that, Eq. (10-37) can be rewritten as

$$\mathrm{d}N' = n_0\,\mathrm{e}^{-mgz/kT}\,\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{10-39}$$

This is the number of molecules in the volume element $\mathrm{d}V = \mathrm{d}x\mathrm{d}y\mathrm{d}z$ at the position of height

z. Therefore, the number of molecules found in a unit volume at height z, or the density of molecular number at height z is given by

$$n = \frac{\mathrm{d}N'}{\mathrm{d}V} = n_0 \mathrm{e}^{-mgz/kT} \tag{10-40}$$

Here n_0 is obviously the density of molecules at $z = 0$. Eq. (10-40) indicates that the density of molecules goes down exponentially with height. Fig. 10-9(a) shows how the temperature affects the distribution. The other parameter that affects the density distribution is the molecules mass m. If we have different kinds of molecules with different masses, they go down with different exponentials. The ones which were heavier would decrease with altitude faster than the light ones. Therefore we would expect that because oxygen is heavier than hydrogen, as we go higher and higher in the atmosphere we live in, the proportion of hydrogen would increase while oxygen would decrease as shown in Fig. 10-9(b).

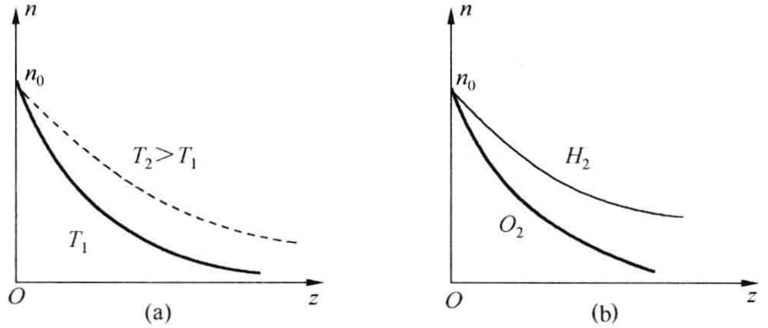


Fig. 10-9 (a) Boltzmann distribution curve and the effect of temperature on the density. (b) The effect of molecular mass on the density

Example 10-7 Evaluate the height where barometer reading drops to 0.88 atm. Suppose that the mole mass of air is about 29×10^{-3}kg and $t = 0$℃.

Solution Combining Eq. (10-39) and Eq. (10-13), $p=nkT$, we have

$$p = nkT = n_0 kT \mathrm{e}^{-mgz/kT} = p_0 \mathrm{e}^{-mgz/kT} \tag{10-41}$$

Eq. (10-41) is known as the isothermal pressure formula. The height z is then

$$z = \frac{kT}{mg}\ln\frac{p_0}{p} = \frac{RT}{\mu g}\ln\frac{p_0}{p} \tag{10-42}$$

in which p_0 is the pressure at $z = 0$, i. e. $p_0 = 1.013\times10^5$ Pa. Eq. (10-42) is used to evaluate the height for aviation and climb. So that

$$z = \frac{8.31\times273}{29\times10^{-3}\times9.8}\ln\frac{1.0}{0.88} \approx 1020(\mathrm{m})$$

Because it is not an isothermal atmosphere and the mole mass of air is also not a constant with different height, so the evaluation is just an approximation. Eq. (10-42) can be used to evaluate the height from the barometer reading only when the difference in height is not too much.

*10.10 Van der Waals Equation of State

The last topic we are going to discuss is that the corrections for the ideal gas law, Eq. (10-2)

$$pV = \frac{M}{\mu}RT \quad \text{or} \quad pV = RT \tag{10-43}$$

for one mole gas. This law was obtained from experiments under the conditions of low enough density that is at normal pressures and normal temperatures, thus molecules are far enough apart so that the volumes of molecules themselves and the attractive forces between them can be ignored. But, for real gas, if it is compressed to an excessively high density, then the behavior of the molecules will deviate from the ideal gas law. Because that the volume of the molecules themselves and the attractive forces between them can not be neglected, if approximate corrections for these two omissions, the gas law will hold for real gases.

First, take the volume of molecules into account. The space in which molecules can move about is somewhat less then V, the volume of the container in Eq. (10-43), so that replacing V by $(V-b)$, Eq. (10-43) becomes

$$p(V-b) = RT \tag{10-44}$$

in which b is an empirical constant for one mole given gas measured by experiments.

Second, consider the effect of attractive forces between molecules. It is proved that the molecules moving toward the wall of container are slowdown by a net force pulling them back into the gas, so that those molecules will exert less impact force during collision with the wall and therefore less pressure on the wall then if there is no attractive force between them. The pressure reduced by an amount of p' that is proportional to the density of the molecules in the layer near the wall, which collide with the wall of container, and also proportional to the density inside layers which exert the inward forces. So the reduced amount p' can be written as

$$p' \propto n^2 \propto \frac{1}{V^2} \quad \text{or} \quad p' = \frac{a}{V^2}$$

The real pressure will be

$$p = \frac{RT}{V-b} - p' = \frac{RT}{V-b} - \frac{a}{V^2}$$

Rearrange it, Eq. (10-44) becomes

$$\left(p + \frac{a}{V^2}\right)(V-b) = RT \tag{10-45}$$

This is the **Van der Waals equation** for one mole of real gas, in which a is a correction constant for one mole given gas. It is obviously when pressure is normal and V becomes large enough so that the correction amounts a/V^2 and b become insignificant, Eq. (10-45) reduces back to the ideal gas law. The correction constants a and b are determined by experiments, being different for different gases. For example, $a = 2.48\times10^4\ \text{Pa}\cdot\text{L}^2/\text{mol}^2$, $b = 2.67\times10^{-2}\ \text{L/mol}$ for Hydrogen; $a = 8.39\times10^4\ \text{Pa}\cdot\text{L}^2/\text{mol}^2$, $b = 3.05\times10^{-2}\ \text{L/mol}$ for Nitrogen.

Eq. (10-45) holds for one mole real gas. If the mass of gas is M, the mol mass is μ, the mole number is then M/μ, under the same pressure and same temperature, the volume of the gas that is the volume of the container, becomes $V' = (M/\mu)V$ where V being the volume of one mol that gas. Substitute $V = (\mu/M)V'$ into Eq. (10-45), which can be rewritten as

$$\left(p + \frac{M^2}{\mu^2}\frac{a}{V'^2}\right)\left(\frac{\mu}{M}V' - b\right) = RT$$

Because V being used to represent the volume of the container holding the gas, so replace V' by V and rearrange above equation, the **Van der Waals equation** becomes

$$\left(p+\frac{M^2}{\mu^2}\frac{a}{V^2}\right)\left(V-\frac{M}{\mu}b\right)=\frac{M}{\mu}RT \tag{10-46}$$

This equation approaches the behavior of real gas quite well. For example, when the pressure increases up to 1.013×10^8 Pa under constant temperature for a Nitrogen gas, the deviation of the result of the left side by using Van der Waals equation compare with the result by experiment is 2.0%; while the deviation by using ideal gas law is over 100%!

Questions

10-1 What is the difference between the concepts of average speed in kinetic theory and that in mechanics?

10-2 Explain from statistic point of view.

(1) The average velocity of the molecules in a gas must be zero if the gas as a whole and the container is not in translational motion.

(2) How can it be that the average velocity is not zero?

(3) Why $\overline{v_x}=\overline{v_y}=\overline{v_z}$ equals zero while $\overline{v_x^2}=\overline{v_y^2}=\overline{v_z^2}$ is not zero?

10-3 Justify the fact that the pressure of a gas depends on the square of the speed of its particles by explaining the dependence of pressure on the collision frequency and the momentum transfer of the particles.

10-4 In the derivation of the pressure formula, in which steps, did we use (1) the statistic method (2) the Newton's law?

10-5 Explain from the kinetic theory of view for a given gas:

(1) the pressure increases as the volume decreases with the temperature unchanged;

(2) the pressure increases as the temperature rises with the volume unchanged.

10-6 Explain why we say that the temperature is a statistic concept. Can we talk about the temperature of a single or few molecules?

10-7 What is the meaning of the following terms from energy point of view?

(1) $\frac{1}{2}kT$; (2) $\frac{3}{2}kT$; (3) $\frac{i}{2}kT$; (4) $\frac{i}{2}RT$; (5) $\frac{M}{\mu}\frac{i}{2}RT$.

10-8 (1) Do half the molecules in a gas in thermal equilibrium have speeds greater than $v_{\rm p}$? than $\bar{v}$? than $v_{\rm rms}$?

(2) Which speed $v_{\rm p}$, $\bar{v}$ or $v_{\rm rms}$ corresponds to a molecule having average kinetic energy?

10-9 Explain the meaning of the following terms:

(1) $f(v)$; (2) $f(v)\mathrm{d}v$; (3) $Nf(v)\mathrm{d}v$; (4) $\int_{v_1}^{v_2}Nf(v)\mathrm{d}v$.

10-10 The fraction of molecules within a given range Δv of the most probable speed decrease as the temperature of a gas rises. Explain.

10-11 A gas is in a vessel with certain volume. If the temperature of the gas becomes twice of its original value, what are the changes of $\bar{Z}$ and $\bar{\lambda}$?

10-12 (1) Why close some people get ill caused from the lack of oxygen when they just move to Tibet from the plain area?

(2) Why is food hard to be well-cooked by steaming in Tibet?

Problems

10-1 10 grams of oxygen are subjected to a pressure of 3.04×10^5 Pa at 10℃. Heating at a constant pressure expanded the oxygen to 10L. Find:

(1) the volume and the density of the gas before expansion;

(2) the temperature and density of the gas after expansion.

10-2 A cylinder containing a gas at 27℃ is divided into two parts of equal volume, each of 100 cm^3, and at equal pressure, by a piston of cross-section area 15 cm^2. The gas in one part is raised in temperature to 100℃; the other volume is maintained at the original temperature. The piston and walls are perfect insulators. How far will the piston move during the change in temperature?

10-3 A mercury barometer tube extends 89.4 cm above a free mercury surface and has air bubble in the region above the mercury column shown in Fig. 10-10. The height of the column is 74.5 cm at 25℃ when the reading on a true barometer is 76.0 cm. On a day when the temperature is 11℃ it reads 75.2 cm. What is the true atmospheric pressure?

10-4 A vessel A with a capacity of $V_1=3$L contains gas at a pressure of $p_{10}=2$atm and a vessel B with a capacity of $V_2=4$L contains the same gas at a pressure of $p_{20}=1.013\times10^5$ Pa. The temperature is the same in both vessels. What pressure will the gas be under if vessel A and B are connected by a tube?

Fig. 10-10 For problem 10-3

10-5 The mass of the H_2 molecule is 3.3×10^{-24} g. If 10^{23} hydrogen molecules per second strike 2.0 cm^2 of wall at an angle of 55°with the normal when moving with a speed of 1.0×10^5 cm/s, what pressure do they exert on the wall?

10-6 Assuming that a gas beam with a speed v strikes vertically against a wall in a vacuum chamber, find the pressure on the wall produced by the gas beam. Suppose the collision is elastic, the density of the molecular number is n, and the single molecular mass is m.

10-7 Suppose that the number of molecules per unit volume in an ideal gas is n, the gas is in a spherical container with a diameter of $2R$, and the mass of a molecule is m. Deduce the pressure formula of kinetic theory (hint: find the change in momentum of a molecule during one collision first; suppose it is moving toward the wall of the container, with a speed v_i at an angle θ_i with the normal as Fig. 10-11 shows).

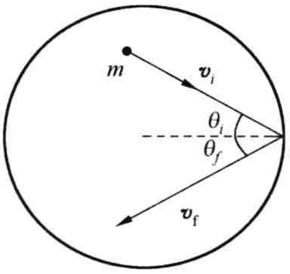


Fig. 10-11 For problem 10-7

10-8 One mole gas of CO_2 is in a container with volume of 10 liters, and the pressure is 2.35×10^5 N/m^2. Find the root-mean-square speed (i. e. rms speed) for the molecules.

10-9 The finest vacuum reached by modem vacuum pumps is 1×10^{-14} Pa. Find the number of molecules per m^3 in this "vacuum" at 273 K.

10-10 At 1 atm the density of a gas is 1 kg/m^3.

(1) Find the rms speed of molecule;

(2) If the temperature is 390 K, find the mole mass of the gas and identify it.

10-11 A vessel contains oxygen gas at 27℃ and 1 atm. Find:

(1) the number of molecules per unit volume;

(2) the density of the gas;

(3) the molecular mass of oxygen;

(4) the average distance between molecules;

(5) the average translational kinetic energy of a molecule;

(6) the change in potential energy as a molecule descends 0.3 m (the height of the cubic vessel), and compare it with the value of $\overline{\varepsilon_t}$.

10-12 Calculate the average translational kinetic energies and rms speeds for a oxygen molecule and a hydrogen molecule at 27℃, respectively.

10-13 Supposing there is one mole oxygen at 300 K, find:

(1) the translational kinetic energy;

(2) the rotational kinetic energy for the gas. How about the change of the total kinetic energy when $\Delta T = 1\ \mathrm{K}$?

10-14 Find the temperature at which the average translational kinetic energy of gas molecules equals to the kinetic energy gained when an electron is accelerated by 1.0 V electric potential difference from the rest.

10-15 What is the change in internal energy as the temperature increases 1 K?

(1) For 1 mol of monatomic molecules, 1 mol diatomic molecules and 1 mol polyatomic molecules respectively.

(2) Answer the same question for 1 g of N_2 and 1 g of H_2.

10-16 In the case of nitrogen under standard condition, find:

(1) the average number of collisions encountered by a molecule within one second;

(2) the total number of collisions occurring between the molecules within $1\mathrm{cm}^3$ of nitrogen per second ($d = 3.1\times10^{-10}$ m).

10-17 Determine the mean free path of carbon dioxide gas (CO_2) molecules

(1) at standard condition;

(2) at a temperature of 100℃ and a pressure of 0.1 mmHg($d = 3.2\times10^{-10}$ m).

10-18 An ionization gauge installed in artificial satellite showed that $1\mathrm{cm}^3$ of the atmosphere contained about a thousand million particles of gas at a height of 300 km from the Earth's surface. Find the mean free path of the gas particles at this height (diameter of a particle 2×10^{-10} m).

10-19 There is a 500 students' score distribution of a physics examination in the table:

Score range	<10	10~20	20~30	30~40	40~50	50~60	60~70	70~80	80~90	>90
Numbers	5	9	16	28	73	93	121	88	45	22

Calculate

(1) the average score;

(2) the root mean square score;

(3) the most probable score and the fraction of the number lies in this range.

Using the medium value of a range, represent the score in that range, for instant, 73 students' scores lie in the range around score of 45.

10-20 Prove that the Maxwell distribution of speed can be written as

$$\frac{\mathrm{d}N}{N} = \frac{4}{\sqrt{\pi}}\mathrm{e}^{-x^2}x^2\,\mathrm{d}x$$

where

$$x = \frac{v}{v_\mathrm{p}},\quad v_\mathrm{p} = \sqrt{\frac{2kT}{m}}$$

10-21 What is the fraction of oxygen molecules having the speed from 100 m/s to 110 m/s at 0℃?

10-22 How many times is the number of molecules $\mathrm{d}N_1$ whose speed ranges from v_p to $v_\mathrm{p} + \mathrm{d}v$ large than the number of molecules $\mathrm{d}N_2$ whose speed ranges from $\sqrt{\overline{v^2}}$ to $\sqrt{\overline{v^2}} + \mathrm{d}v$?

10-23 What is the fraction of molecules in a gas, which have the speed within the range v_p to $1.01\ v_p$?

10-24 N molecules are assumed to have a speed distribution as shown in Fig. 10-12.

(1) Write the expression of a using N and v_0;

(2) Find the most probable speed v_p;

(3) Separating the whole range of speed into 3 ranges with equal width of v_0, calculate the fractions of molecules whose speeds lie in the ranges from 0 to v_0, from v_0 to $2v_0$, from $2v_0$ to $3v_0$.

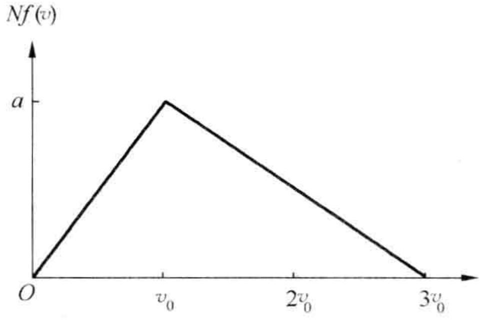


Fig. 10-12 For problem 10-24

10-25 A gas consists of molecules of mass m and at temperature T. Making use of Maxwell speed distribution function, find the corresponding distribution of molecular translational kinetic energy E_k. Determine the most probable value of the kinetic energy E_p. Does E_p correspond to the most probable speed v_p?

10-26 The mean free path of oxygen molecules is 9.5×10^{-8} m at 0℃ and a certain pressure. What is the average number of collisions of the gas molecules per second if the vessel is evacuated to 0.01 of the initial pressure, the temperature is constant.

10-27 According to Eq. (10-25) and Eq. (10-29), calculate the average speed by integrating, and using the integration formula $\int_0^{\infty} v^3 e^{-bv^2}\, dv = \frac{1}{2b^2}$.

10-28 According to Eq. (10-35), calculate the ratio of number of atoms with energy ε_2 to that with energy ε_1 for hydrogen atom at 300 K. Here $\varepsilon_1 = -13.6$ eV, is the ground energy level, and $\varepsilon_2 = -3.4$ eV is the first excited energy level.

10-29 At what altitude is the pressure of air equal to 75% of that at sea level? Assume the temperature to be constant and equal to 0℃.

10-30 There is a high altitude cosmic station on a mountain at a height of 3250 m above the sea level. Find the air pressure at this altitude. Assume the temperature to be constant and equal to 5℃, the mole mass of air is 29×10^{-3} kg/mol and the pressure is 1.013×10^5 Pa at sea level.

10-31 Calculate the temperature for an oxygen gas of 0.05 L under pressure 1.01×10^8 Pa by using Van der Waals equation, and compare the result with that by using ideal gas law, represent the relative deviation by a percentage error. For oxygen gas $a = 1.38\times10^5\ \text{Pa}\cdot\text{L}^2/\text{mol}^2$ and $b = 3.18\times10^{-2}$ L/mol are known.

Chapter11

Fundamentals of Thermodynamics

Thermodynamics is the study of the relationships among the purely macroscopic parameters describing the behavior of a system. The most important category of heat phenomena of thermodynamics concerns the conversion of one form of energy into another, especially the conversion of heat into other forms of energy. These conversions are governed by the two fundamental laws of thermodynamics. The first of these is essentially a general statement of the law of conservation of energy and the second is a statement about the maximum efficiency attainable in the conversion of heat into work and about the directivity of heat processes.

11.1 Internal Energy, Heat, and Work

11.1.1 Internal energy of a system

The concept of internal energy of a system has been introduced in the last part of section 10-5. For a real gas, internal energy includes the total kinetic energy which is the single value function of the temperature of the gas and the total potential energy produced by the inter-molecular forces which are related to the volume of the gas, therefore the internal energy is the function of both temperature and volume of the gas, i. e.

$$E = E(T, V)$$

For ideal gases, the forces between molecules can be ignored, so that, the internal energy is simply the total kinetic energy of molecules represented by Eq. (10-18),

$$E = \frac{M}{\mu}\frac{i}{2}RT$$

which means that the internal energy is the single value function of the temperature of the gas.

On the basis of the concept of internal energy established in kinetic theory of ideal gas, we will study internal energy of the gaseous system and the conversion of one form of energy into the other forms of energy by means of thermodynamic method or the macroscopic method, and establish the relationships among the macroscopic parameters describing the behavior of the gaseous system, such as pressure, temperature, volume and so on. Compared with the microscopic method used in kinetic theory of gases, such a macroscopic, large-scale description is necessarily somewhat crude, since it overlooks all of the small-scale, microscopic details. Moreover, in practical applications, these microscopic details are often irrelevant. For instance, when we investigate the behavior of the combustion gases in the cylinder of an automobile engine, we can get by reasonably well with the macroscopic quantities of the temperature, pressure, density, and heat capacity.

11. 1. 2 Heat and work

Consider some amount of gas that is in a cylinder fitted with a piston. We describe the gas as a system, and the relevant parts of the surrounding as the environment of the system as shown in Fig. 11-1. If we heat the gas while keep the piston in stationary, its state will be changed, and the temperature may increase, so that, the internal energy of it may be increased consequently. In this case, we say that some energy flows or transfers from its environment. So, heat (symbol Q) is actually a form of energy that flows between a system and its environment by virtue of temperature difference that exists between them. On the other hand, energy can also be transferred between a system and its environment by means of work (symbol W), which we always associate with a force moving through a distance as shown in Fig. 11-1. If we push the piston and compress the gas, at the consequence, the internal energy of the system may be changed also. In practice, we change the state of a gaseous system by both of heat transformation and work in the same time.

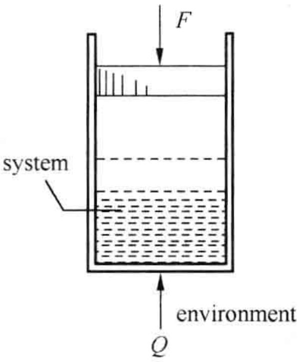


Fig. 11-1 A gaseous system and its environment

Note that heat and work unlike pressure, volume, temperature (and internal energy) are not intrinsic properties of a system. They have meaning only as they describe energy transforms into or out off the system, adding to or subtracting from the system's store of internal energy.

Both heat and work represent energy-in-transit between a system and its environment. There is bound to be a relationship between the heat and work caused respectively the same temperature increment for the same system. It is proved by Joule that 1 cal of heat causes the same temperature increment as that of 4. 18 joule of work does, i. e.

$$1\ \text{cal}=4.18\text{J} \quad \text{or} \quad 1\text{J}=0.24\text{cal}$$

It means that 1 cal of heat is equivalent to 4. 18 J of work in changing the energy of a system. In scientific work, it was decided that, since heat (like work) is a form of energy, the SI unit for heat should be the same as for all other forms of energy, namely, the joule. However, the cal continues to be used in nutrition and in some area of chemistry.

11. 2 The First Law of Thermodynamics

11. 2. 1 The first law of thermodynamics

Suppose that a system (Fig. 11-1) changes from state 1 to state 2 by the work W' done on it and the heat Q transferred meanwhile into it, the corresponding total change in internal energy of the system is ΔE. According to the law of conservation of energy, we have

$$\Delta E = E_2 - E_1 = W' + Q \qquad (11\text{-}1)$$

where E_1 and E_2 represent the internal energy of the gas at state 1 and stare 2 respectively. In manufacture and technique, the most important problem concerned is the work done by the system on the environment. We usually use W to represent the work done by the system, thus, substituting $-W = W'$ into the equation above, we get the relationship among the energies exchanged through a system, i. e.

$$E_2 - E_1 = -W + Q$$

or

$$Q = (E_2 - E_1) + W \qquad (11\text{-}2a)$$

which is known as **the first law of thermodynamics**. It means that the input heat is transferred into two parts of energy, one is served to do work performed by the system on the environment, another is served to increase the internal energy of the system. It is the extension of the conservation of energy principle to include the heat phenomena and thermodynamic system.

The value of Q, W and ΔE could be positive or negative. The sign for Q is positive if heat is added to the system and negative if heat is withdrawn from the system; the sign for W is positive if the work performed by the system, conversely, a negative value of W which means that the work is done on the system; the positive value of ΔE corresponds to a increment of internal energy of the system while the negative value to a decrement of internal energy.

If the thermodynamic system changes by only a differential amount, we can write the first law as

$$\mathrm{d}Q = \mathrm{d}E + \mathrm{d}W \qquad (11\text{-}2b)$$

Note that $\mathrm{d}E$ is the differential of the internal energy independent of the process while the amounts of $\mathrm{d}Q$ and $\mathrm{d}W$ depend on the particular processes that will be seen in the following sections.

The first law sometimes permits us to calculate the unknown amount of heat or work required for a process. To do this, let us take a closer look at the work done by the system.

11. 2. 2 Calculation of the work

Here we look in some detail at how heat and work are exchanged between a system and its environment during an equilibrium process (we investigate only equilibrium processes in the followings). Let us take a gas confined to a cylinder with a movable piston as our system as shown in Fig. 11-2. Suppose that the pressure of the confined gas on the piston is p and the area of the piston is S, the force exerted on the piston by the gas is therefore $F = pS$. When the gas expands slowly, the piston moves a differential distance $\mathrm{d}x$ and the element work performed by the gas equals

$$\mathrm{d}W = F\mathrm{d}x = pS\,\mathrm{d}x = p\,\mathrm{d}V$$

in which $\mathrm{d}V$ is the differential change in volume of the gas. If the system changes from a initial state to a final

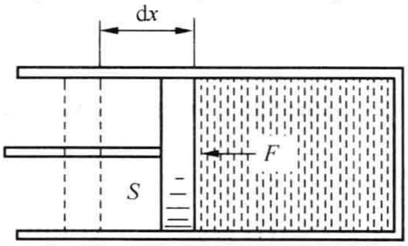


Fig. 11-2 The work done by the system

state and both states are represented by dots on the p-V diagram, the operation of changing the system from the state i to the state f is called a thermodynamic process represented by a smooth curve on p-V diagram under the assumption that the operation is carried out very slowly, so that the system remains essentially in thermodynamic equilibrium at all stages. The value of the element work is equal to the area of the strip shaded on the p-V diagram; the total work done during the process is obtained by integration of

$$W = \int \mathrm{d}W = \int_{V_i}^{V_f} p\,\mathrm{d}V \tag{11-3}$$

The value of the work equals the total area under the process curve connected from i to f as shown in Fig. 11-3. It is obvious that there are infinite ways to take the system from its initial state to its final state, therefore the shape of the process curve could be various, so that, the value of the work depends not only on the initial and the final states but also on the processes. Eq. (11-3) suggests that the work is positive when the gas expands ($\mathrm{d}V > 0$) and is negative when the gas is compressed ($\mathrm{d}V < 0$). Fig. 11-3 suggests that we can make the work done as small as we want (by following a path like ibf) or as large as we want (by following a path like iaf).

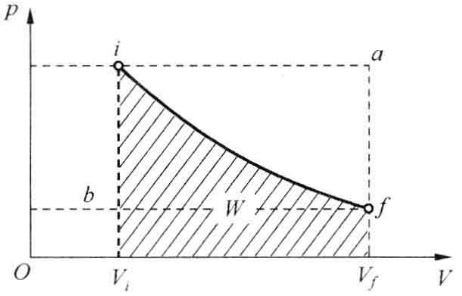


Fig. 11-3 The value of work equals the area under the process curve

Substituting Eq. (11-3) into Eq. (11-2), the first law of thermodynamics can be written as

$$Q = (E_f - E_i) + \int_{V_i}^{V_f} p\,\mathrm{d}V \tag{11-4}$$

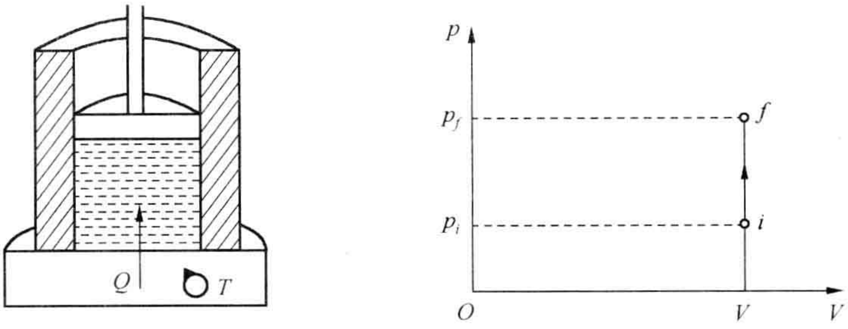
Because $(E_f - E_i)$ depends only on state i and state f and the work on the process, so the heat depends on the process. To sum up: A system can be taken from a given initial state to a given final state by an infinite number of processes. In general, the work W and the heat Q will have different values for each of these processes. We say that heat and work are path-dependent quantities.

11.3 Application of the First Law to Isochoric, Isobaric, and Isothermal Processes

We are ready to apply the first law of thermodynamics to different thermodynamic processes. Here we look at three special cases first, in each of which a certain restriction is imposed on the system. We then see what consequences follow when we use the first law to them.

11.3.1 Constant volume processes (isochoric process)

When a substance undergoes a process in which the volume remains unchanged, the process is called constant volume process or isochoric process. In Fig. 11-4(a), a gas confined to a cylinder with a movable piston is taken as our system again. The wall of the cylinder is insulating but the bottom of the cylinder rests on a heat reservoir (a hot plate, say, an electromagnetic oven) whose temperature T can be controlled at will by turning a knob. The piston is controlled mechanically at a certain position so that the volume of the gas remains constant. If heat is added to the system, the temperature increases from T_i to T_f, and therefore the pressure rises from p_i to p_f as shown in Fig. 11-4(b). Conversely, if heat is removed from the system, the temperature must drop down. The isochoric curve is a straight line parallel with p-axis on the p-V diagram.



(a) (b)

Fig. 11-4 (a) An isochoric process is carried out; (b) The isochoric curve on p-V diagram

Because $dV=0$ in isochoric process, the system can do no work, i. e. $dW = pdV = 0$. From Eq. (11-2b), for any differential change, $(dQ)_V = dE$. For a finite process like that in Fig. (11-4b), using Eq. (11-3) and Eq. (11-2a), the first law of thermodynamics, and Eq. (10-18), we have

$$W_V = 0 \tag{11-5}$$

$$Q_V = E_f - E_i = \frac{M}{\mu}\frac{i}{2}R(T_f - T_i) \tag{11-6}$$

The subscript V means the volume being constant. Eq. (11-6) tells us that all the input heat has served to increase the internal energy of the gas during an isochoric process. If the heat is withdrawn from the system in an isochoric process, the internal energy will decrease.

In practice, the very sudden increase of temperature and pressure accompanying the explosion of gasoline vapor and air in a gasoline engine may be treated mathematically as though it were an isochoric addition of heat.

11.3.2 Constant pressure processes (Isobaric process)

A process taking place at constant pressure is called constant pressure process or isobaric process. In Fig. 11-5(a), the system is same as before, except that the pressure of the confined gas is balanced by loading lead shots on the top of the piston. We can cause the volume to increase (from V_i to V_f) by slowly turning up the temperature control knob, raising temperature of the gas from T_i to some higher value T_f. During this process, work is done by the expanding gas to lift the weight on the piston, and heat is added to the system from the heat reservoir. An isobaric process is represented by a straight line parallel to V-axis on the p-V diagram as shown in Fig. 11-5(b), and the work can be interpreted graphically as the area under the curve between the limits V_i and V_f.

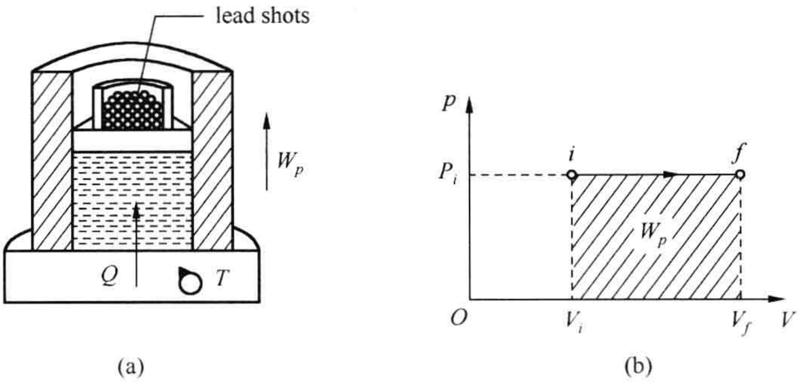


Fig. 11-5 (a) An isobaric process is carried out; (b) The isobaric expansion curve on p-V diagram

Because $dp = 0$ in any differential change in this process, the element work is

$$dW = p dV = \frac{M}{\mu} R dT$$

From the first low of thermodynamics, the heat is then

$$(dQ)_p = dE + p dV = dE + \frac{M}{\mu} R dT \tag{11-7}$$

For a finite isobaric process like that in Fig. 11-5(b), we have the total work

$$W_p = \int_{V_i}^{V_f} p dV = p(V_f - V_i) = \frac{M}{\mu} R(T_f - T_i) \tag{11-8}$$

and the heat

$$Q_p = E_f - E_i + p(V_f - V_i) \tag{11-9}$$

Here

$$E_f - E_i = \frac{M}{\mu} \frac{i}{2} R(T_f - T_i) \tag{11-10}$$

so that

$$Q_p = \frac{M}{\mu} \frac{i+2}{2} R(T_f - T_i) \tag{11-11}$$

The subscript p means the pressure being unchanged. Eq. (11-9) indicates that during an isobaric process, one part of the input heat serves to do work, another part serves to increase the internal energy of the gas. If $V_f > V_i$, i. e., the gas expands so that $W_p > 0$, the gas does work on its environment, and draws heat in from the reservoir. Conversely, if $V_f < V_i$, i. e., the gas is compressed so that $W_p < 0$, the work is done on the system by the environment, and the heat is withdrawn from the gas into the surrounding.

Note that comparing Eq. (11-6) with Eq. (11-9) leads a conclusion that no matter whether the process is an isochoric one or an isobaric, the change in internal energy will be equal, provided that the change in temperature, i. e., $T_f - T_i$, is the same.

In practice of engineering, when water enters the boiler of a steam engine and is heated to its boiling point, vaporized, and then the system is superheated, all these processes taken place can be treated as isobaric. Such processes play an important role in mechanical engineering and also in chemistry.

11.3.3 Constant temperature process (isothermal process)

A process taking place at constant temperature is said to be isothermal or constant temperature process. In order for the temperature to remain constant, by removing weight gradually from the piston as in Fig. 11-6(a), we can permit the changes in pressure and volume being carried out very slowly, so that the stage approximates thermal equilibrium very closely at every state of the process. The gas remains at constant temperature while doing so, absorbing heat Q from the reservoir and doing work W by lifting the weight. The isothermal process is represented by one of the hyperbola on the p-V diagram, called as isotherm shown in Fig. 11-6(b). In isothermal process, none of the quantities p or V is constant.

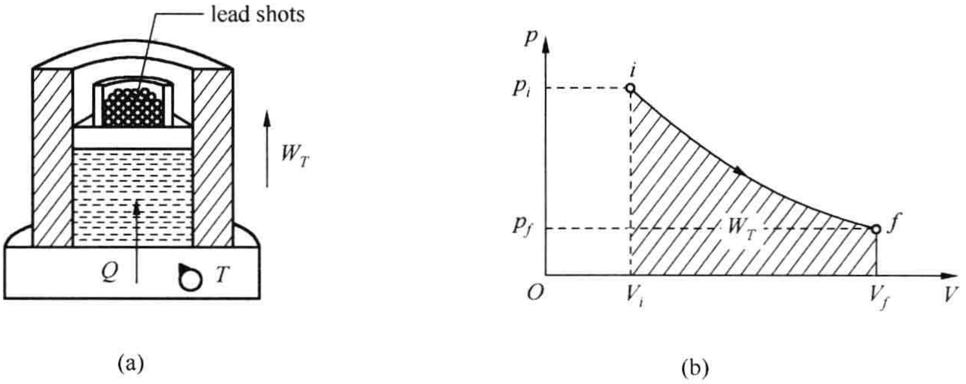


Fig. 11-6 (a) An isothermal process is carried out; (b) The isothermal expansion curve on p-V diagram

Because $\mathrm{d}T = 0$ in every differential change, so $\mathrm{d}E = \frac{M}{\mu} R\,\mathrm{d}T = 0$. For a finite process like that in Fig. 11-6(b), the internal energy of the gas remains unchanged, i. e.

$$E_f - E_i = 0 \tag{11-12}$$

From the first low of thermodynamics, we have the heat

$$Q_T = W_T = \int_{V_i}^{V_f} p\ \mathrm{d}V \tag{11-13}$$

The subscript T means the temperature being a constant. Eq. (11-13) indicates that all of the input heat serves to do work performed by the gas on its surrounding. If the process is carried out reversely, then $V_f < V_i$, the work will be negative and the heat will be withdrawn from the gas.

From the ideal gas law $pV = \frac{M}{\mu} RT$, we have $p = \frac{M}{\mu} RT \frac{1}{V}$, substituting it into Eq. (11-13) leads to

$$Q_T = W_T = \frac{M}{\mu} RT \int_{V_i}^{V_f} \frac{\mathrm{d}V}{V} = \frac{M}{\mu} RT \ln \frac{V_f}{V_i} \tag{11-14}$$

Because $p_i V_i = p_f V_f$ in isothermal process, Eq. (11-13) can be rewritten as

$$Q_T = W_T = \frac{M}{\mu} RT \ln \frac{p_i}{p_f} \tag{11-15}$$

Using ideal gas law again, we have another useful equation

$$Q_T = W_T = p_i V_i \ln \frac{V_f}{V_i} \tag{11-16}$$

Example 11-1 A monatomic ideal gas is contained in a cylinder closed with a movable piston. The initial pressure is 8 atm. and the initial volume is 0. 5 L. The gas is heated first at constant pressure until the volume becomes to 1. 5 L, then expands isothermally until the volume is doubled, finally cooled down at constant volume until the pressure drops to 1 atm.

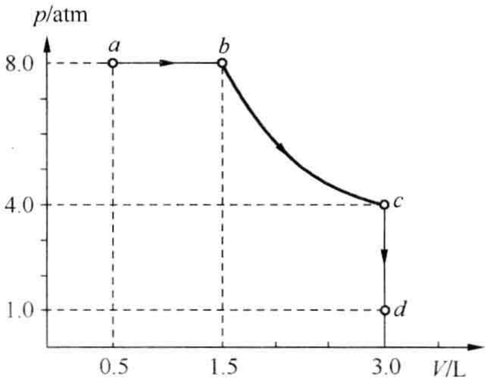


Fig. 11-7 For Example 11-1

(1) How much work is done by the gas in the three processes?

(2) What is the change in internal energy?

(3) How much heat is supplied to the system?

Solution First of all, we show each of the three processes on the p-V diagram as in Fig. 11-7, and use the arrows to describe the direction of each process.

(1) The total work done by the gas equals the sum of the work done in each stage. Using Eq. (11-8), Eq. (11-16) and Eq. (11-5), we have

$$
\begin{aligned}
W &= W_p + W_T + W_V = p_a(V_b - V_a) + p_b V_b \ln\frac{V_c}{V_b} + 0 \\
&= 8\times1.013\times10^5\times(1.5-0.5)\times10^{-3} + 8\times1.013\times10^5\times1.5\times10^{-3}\times\ln(3/1.5) \\
&= 810.4 + 842.6 = 1653(\mathrm{J})
\end{aligned}
$$

(2) The change in internal energy from the initial state a to the final state d is

$$
\begin{aligned}
E_d - E_a &= \frac{M}{\mu}\frac{i}{2}R(T_d - T_a) = \frac{i}{2}(p_d V_d - p_a V_a) \\
&= (3/2)\times(1.013\times10^5\times3\times10^{-3} - 8\times1.013\times10^5\times0.5\times10^{-3}) \\
&= 1.5\times1.013\times10^2\times(3-4) = -151.9(\mathrm{J})
\end{aligned}
$$

The negative sign indicates that the overall change in internal energy is decrement although it may be increment ($\Delta E>0$) or decrement ($\Delta E<0$) in each individual process.

(3) The total heat added to the system can be obtained in two ways:

(a) Using Eq. (11-11), Eq. (11-13) and Eq. (11-6), sum up the input heat in each of the three processes

$$
\begin{aligned}
Q = Q_p + Q_T + Q_V &= \frac{M}{\mu}\frac{i+2}{2}R\,(T_b - T_a) + W_T + \frac{M}{\mu}\frac{i}{2}R\,(T_d - T_c) \\
&= \frac{i+2}{2}p_a(V_b - V_a) + W_T + \frac{i}{2}V_c(p_d - p_c)
\end{aligned}
$$

Here $p_c = p_a\dfrac{V_b}{V_c} = 4\mathrm{L}$, and W_T can be found in step(1).

So that

$$
\begin{aligned}
Q &= (5/2)\times8\times1.013\times10^5\times10^{-3} + 842.6 + (3/2)\times3\times10^{-3}\times(1-4)\times1.013\times10^5 \\
&= 2026 + 842.6 - 1367.6 = 1501(\mathrm{J})
\end{aligned}
$$

in which -1367.6 for Q_V indicates that the heat is withdrawn from the gas into the surrounding during the isochoric cooling process.

(b) According to the first law of thermodynamics, Eq. (11-2a), we simply have the total heat

$$Q = E_d - E_a + W = -151.9 + 1653 \approx 1501(\mathrm{J})$$

The result as what we expected.

11.4 The Heat Capacities of an Ideal Gas

11.4.1 Heat capacity and specific heat capacity

Experiments indicate that the amount of heat added to an object is proportional to the mass of the object and the change in temperature that results. Thus,

$$Q = Mc(T_f - T_i)$$

in which Mc is called the heat capacity of the object, M is the mass of the object, and $(T_f - T_i)$ is the change in temperature. The value of the proportionality constant c is the same for different objects made of the same material, and is called the specific hear capacity or heat capacity per unit mass. It refers not to the object but to the material of which the object is made. The value of c equals the amount of heat needed to cause unit rise of temperature for unit mass of the material. If the mass M is just one mole, i. e., $M/\mu = 1$ mol, thus the heat capacity μc is defined as molar heat capacity of the material, labeled by C; that is $C = \mu c$, which is the amount of heat added to cause unit rise of temperature for one mole of a material. The SI unit for molar heat capacity is J/(mol · K).

An important point is that in assigning a molar heat capacity to any substance, it is necessary to know not only how much heat was added but the conditions under which that transfer took place also. The temperature of a substance may be changed under a variety of conditions. The volume may be kept constant, or the pressure may be kept constant (usually atmospheric for solids and liquids during the heat transfer), or both may be allowed to vary in some definite manner. The amount of heat necessary to cause unit rise of temperature is different in each case. In other words, a substance has many different molar heat capacities.

For gases, as we shall see, adding heat under constant volume condition and under constant pressure condition leads to quite different values for the molar heat capacities, there is a simple and important relation between these two molar heat capacities of an ideal gas.

11.4.2 The molar heat capacity at constant volume

Fig. 11-8 shows two isotherms of an ideal gas at temperature T and $T+\mathrm{d}T$. Since the internal energy of an ideal gas depends only on the temperature, it is constant if the temperature is constant and the isotherms are also curves of constant internal energy. It follows that the change in internal energy, $\mathrm{d}E$, is the same in all processes in which the gas is taken from any point on one isotherm to any point on the other. Thus $\mathrm{d}E$ is the same for all the processes *ab*, *ac* or *de* etc. as shown in Fig. 11-8.

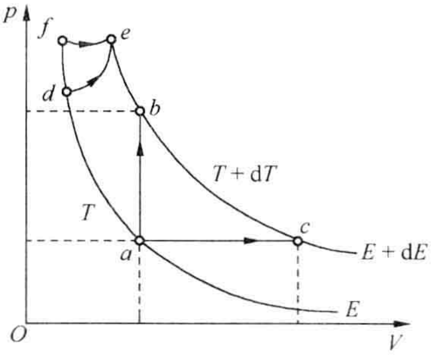


Fig. 11-8 Different processes may correspond to the same change in internal energy dE

Consider the first process *ab* at constant volume. To carry out such a process for a gas at temperature T and closed in a cylinder with a piston, we can turn the control knob of the reservoir (Fig. 11-4) to a slightly higher temperature $T+\mathrm{d}T$. There will be a flow of heat $(\mathrm{d}Q)_V$ into the gas. The molar heat capacity at constant volume, C_V, is defined as

$$C_V = \frac{(\mathrm{d}Q)_V}{\mathrm{d}T} \tag{11-17}$$

which equals the amount of heat added to cause unit rise of temperature for one mole of an ideal gas under constant volume. The pressure of the gas increases during this process, but no work is done, since the volume is constant. Hence, from the first law of thermodynamics, Eq. (11-2b), we have $(\mathrm{d}Q)_V=\mathrm{d}E$, and substituting it into Eq. (11-17) leads to

$$C_V = \frac{(\mathrm{d}Q)_V}{\mathrm{d}T} = \frac{\mathrm{d}E}{\mathrm{d}T} \tag{11-18}$$

With the internal energy of one mole in hand, that is Eq. (10-17)

$$E_0 = \frac{i}{2}RT$$

Substitute it into Eq. (11-18), we have

$$C_V = \frac{\mathrm{d}E}{\mathrm{d}T} = \frac{i}{2}R \tag{11-19}$$

in which i is the number of degrees of freedom for the ideal gas and R is the universal constant of gases. Eq. (11-19) suggests that the molar heat capacity of an ideal gas depends only on the number of degrees of freedom for that gas, not on the temperature.

For monatomic gases, $i=3$, $C_V = \frac{3}{2}R = 12.5\ \mathrm{J/(mol \cdot K)}$

For diatomic gases, $i=5$, $C_V = \frac{5}{2}R = 20.8\ \mathrm{J/(mol \cdot K)}$

For polyatomic gases, $i=6$, $C_V = 3R = 24.9\ \mathrm{J/(mol \cdot K)}$

The amount of heat needed for a gas of mass M with molar mass μ to increase the temperature from T_i to T_f at constant volume is therefore

$$Q_V = \frac{M}{\mu}C_V(T_f - T_i) \tag{11-20}$$

It is equivalent with Eq. (11-6)

$$Q_V = E_f - E_i = \frac{M}{\mu}\frac{i}{2}R(T_f - T_i)$$

Because $E_f - E_i$ is the same for all different processes for a certain gas between isotherm of T_i and isotherm of T_f, so Eq. (11-6) or Eq. (11-20) gives the change of internal energy in all those processes even if they are not at constant volume as shown in Fig. 11-8.

11.4.3 The molar heat capacity at constant pressure

Now consider the process *ac* in Fig. 11-8, at constant pressure. To carry out such a process the gas might be enclosed in a cylinder with a piston, acted on by a external force. We can turn the temperature control knob of the reservoir to rise the temperature slightly from T to $T+\mathrm{d}T$. As the heat of $(\mathrm{d}Q)_p$ flows into the gas, it expands at constant pressure and does work (Fig. 11-5), then the molar heat capacity at constant pressure, C_p, is defined as

$$C_p = \frac{(\mathrm{d}Q)_p}{\mathrm{d}T} \tag{11-21}$$

which equals to the amount of heat needed to cause unit rise of temperature for one mole of an ideal gas under the constant pressure. The work done is $\mathrm{d}W = p\mathrm{d}V$. From the first law of thermodynamics

$$\mathrm{d}Q = \mathrm{d}E + \mathrm{d}W = \mathrm{d}E + p\mathrm{d}V$$

Substituting it into Eq. (11-21) leads to

$$C_p = \frac{\mathrm{d}E}{\mathrm{d}T} + p\frac{\mathrm{d}V}{\mathrm{d}T} \tag{11-22}$$

For one mole of ideal gas, $\mathrm{d}E = C_V\mathrm{d}T$, $p\mathrm{d}V = R\mathrm{d}T$ (from the equation of state), so that

$$C_p = C_V + R \tag{11-23}$$

which means that the molar heat capacity of an ideal gas at constant pressure is greater than that at constant volume, and the difference is the universal gas constant. This is why R is expressed in the same units as C_p and C_V usually, J/(mol • K) in SI unit. Since the change in internal energy $\mathrm{d}E$ is the same corresponding to the same change in temperature $\mathrm{d}T$, for an ideal gas no matter how is the process, however, it is the difference that work is done in the isobaric process (the area under the isobaric curve ac in Fig. 11-8) but not in the isochoric one, thus, we can draw the physics meaning of R that R is the amount of work done by one mole gas in an isobaric processes during the unit rise in temperature.

Substituting Eq. (11-19) into Eq. (11-23), we have

$$C_p = \frac{i}{2}R + R = \frac{i+2}{2}R \tag{11-24}$$

For monatomic gases, $i=3$, $C_p = \frac{5}{2}R = 20.8\ \mathrm{J/(mol \cdot K)}$

For diatomic gases, $i=5$, $C_p = \frac{7}{2}R = 29.1\ \mathrm{J/(mol \cdot K)}$

For polyatomic gases, $i=6$, $C_p = 4R = 33.2\ \mathrm{J/(mol \cdot K)}$

The amount of heat needed for a gas of mass M with molar mass μ to increase the temperature from T_i to T_f at constant pressure is therefore

$$Q_p = \frac{M}{\mu}C_p(T_f - T_i) \tag{11-25}$$

It is equivalent with Eq. (11-9) and Eq. (11-11).

The ratio of molar heat capacities, denoted by γ, is defined as

$$\gamma = \frac{C_p}{C_V} \tag{11-26}$$

For an ideal gas $C_p = \frac{i+2}{2}R$, $C_V = \frac{i}{2}R$, so that

$$\gamma = \frac{i+2}{i} \tag{11-27}$$

For monatomic gases, $i=3$, $\gamma = \frac{5}{3} = 1.67$

For diatomic gases, $i=5$, $\gamma=\frac{7}{5}=1.40$

For polyatomic gases, $i=6$, $\gamma=\frac{8}{6}=1.33$

Although Eq. (11-23) was derived for an ideal gas, it is very nearly true for real gas at moderate pressure and temperature. Measured values of C_p and C_V are given in Table 11-1 for some real gases at moderate pressure, and the difference (C_p-C_V) is seen to agree quite precisely with the value of R for most of the gases.

Table 11-1 Molar heat capacities of gases at moderate pressure

[the unit of C_p and C_V is J/(mol · K)]

Type of gas	Gas	C_p	C_V	C_p-C_V	$\gamma=C_p/C_V$
Monatomic	He	20.9	12.5	8.4	1.67
	Ar	20.9	12.5	8.4	1.67
Diatomic	H_2	28.8	20.4	8.4	1.41
	N_2	28.6	20.4	8.2	1.41
	CO	29.3	21.2	8.1	1.40
	O_2	28.9	21.0	7.9	1.40
Polyatomic	H_2O	36.2	27.8	8.4	1.31
	CH_4	35.6	27.2	8.4	1.30
	CO_2	36.6	28.1	8.5	1.30
	SO_2	40.3	31.4	8.9	1.29

It is seen that the values of C_p, C_V and γ for monatomic and diatomic gases are in reasonable agreement with theoretic values in classical thermodynamics, while the measured values for polyatomic gases have fluctuation compared with those of theoretic ones. It means that the heat capacity theory in classical thermodynamics is an approximation due to our oversimplification in kinetic theory. It is only quantum theory that can solve the problem of heat capacity precisely.

11.5 Application of the First Law of Thermodynamics to Adiabatic Processes

Any process in which there is no flow of heat into or out of a system is called adiabatic. In the other words, no heat transfer occurs between the system and its environment when an adiabatic process is carried out. Fig. 11-9 shows an insulated cylinder containing an ideal gas and resting on an insulating stand. By removing weight from the piston gradually, we can allow the gas to expand adiabatically. However there is a second way of ensuring that heat exchange does not take place; for instance, sound waves are propagated through air and other gases as a series of compressions and refraction that take place so rapidly that there is no time for heat to

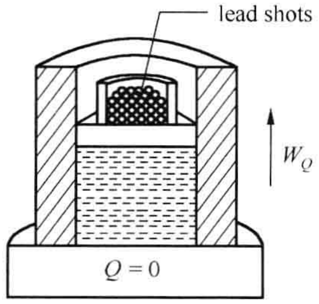


Fig. 11-9 An adiabatic expansion is carried out by removing the lead shots gradually

flow from one part of the medium to another, back and forth in synchronism with the rapidly oscillating wave, thus volume change of air can be regarded as an adiabatic process. The compressions and expansions of steam in the cylinder of a steam engine or of the hot gases in the cylinder of a gasoline and diesel engine are also approximately adiabatic.

Let an ideal gas undergo an infinitesimal adiabatic process, the $\mathrm{d}Q=0$, and from the first law of thermodynamics, we have

$$\mathrm{d}E+p\mathrm{d}V=0 \tag{11-28}$$

For a finite process, we have $Q=0$ and

$$W_Q=-(E_f-E_i) \tag{11-29}$$

in which the subscript Q means no heat exchange. Eq. (11-29) indicates that if work is done by the system (i. e. $W>0$) there must be a decrease in the internal energy of the gas, or we can say that it is at the price of decrease in the internal energy that the output work is gained. Conversely, if work is done on the system (i. e. $W<0$) there must be increase in the internal energy of the gas. From Eq. (11-19) and Eq. (11-20), we have

$$E_f-E_i=\frac{M}{\mu}C_V(T_f-T_i) \tag{11-30}$$

So that

$$W_Q=-(E_f-E_i)=-\frac{M}{\mu}C_V(T_f-T_i) \tag{11-31}$$

It is obvious that a decrease in internal energy means a decrease in temperature, which can be perceived quite readily when air is allowed to rush out of the valve of an automobile tire; this expanding air feels quite cool. While the increase in temperature due to adiabatic compression of air is quite familiar with us too when operating a manual air pump for bicycle tire, the compression of air in the barrel of the pump produces a noticeable warming of the pump.

As the volume of the gas increases in an adiabatic expansion, both the pressure and temperature drop; there are some relationships among the pressure, volume, and temperature. Let an ideal gas undergo an infinitesimal adiabatic process. Then $\mathrm{d}Q=0$, $\mathrm{d}W=p\mathrm{d}V$ and $\mathrm{d}E=\frac{M}{\mu}C_V\mathrm{d}T$. From Eq. (11-28), we have $\frac{M}{\mu}C_V\mathrm{d}T=-p\mathrm{d}V$.

By differentiating the equation $\frac{M}{\mu}RT=pV$, we find

$$\frac{M}{\mu}R\mathrm{d}T=p\mathrm{d}V+V\mathrm{d}p$$

Eliminating $\mathrm{d}T$ between these equations and making use of the fact that $C_p-C_V=R$, we obtain the relation

$$\frac{\mathrm{d}p}{p}+\frac{C_V+R}{C_V}\frac{\mathrm{d}V}{V}=0$$

or

$$\frac{\mathrm{d}p}{p}+\gamma\frac{\mathrm{d}V}{V}=0$$

To obtain the relation between p and V in a finite adiabatic change we may integrate the above equation. This gives

$$\ln p + \gamma \ln V = \ln \text{ (a constant)}$$

Using the properties of logarithms, we can rewrite this as

$$pV^{\gamma} = \text{a constant} \tag{11-32}$$

According to the ideal gas law, we can obtain the alternate forms of Eq. (11-32) as

$$V^{\gamma-1}\, T = \text{a constant} \tag{11-33}$$

and

$$p^{\gamma-1}\, T^{-\gamma} = \text{a constant} \tag{11-34}$$

in which $\gamma = C_p/C_V$, and the three constants are different from one another.

An adiabatic expansion or compression of an ideal gas can be represented graphically by a plot of Eq. (11-32), as in Fig. 11-10, in which a number of isothermal curves are shown for comparison. Note that the adiabatic curves, at any point, have a somewhat steeper slope than the isothermal curve passing through the same point. That is, as we follow along an adiabat from left to right (expansion process) the curve continually cuts across isotherms of lower and lower temperatures, in agreement with the fact that the temperature gradually decreases in an adiabatic expansion.

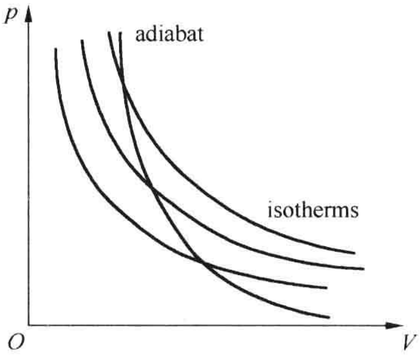


Fig. 11-10 An adiabatic curve and a number of isotherms of an ideal gas. The adiabatic is steeper than isotherm

Example 11-2 2 mol of Helium (assume an ideal gas) expanded adiabatically from an initial state ($T_i = 300$ K) until the volume is doubled.

(1) Calculate the work done by the gas.

(2) What would be the answer if the gas had expanded isothermally from the same initial state to the same final volume?

Solution (1) From Eq. (11-31), $W_Q = -\dfrac{M}{\mu}C_V(T_f - T_i)$. To find the final temperature T_f, we use Eq. (11-33), that is

$$V_f^{\gamma-1}T_f = V_i^{\gamma-1}T_i$$

$$T_f = \left(\frac{V_i}{V_f}\right)^{\gamma-1} T_i$$

For monatomic ideal gas, $i = 3$, $\gamma = 1.67$

$$T_f = (0.5)^{0.67} \times 300 = 188.6(\text{K})$$

So that

$$W_Q = -2 \times (3/2) \times 8.31 \times (188.6 - 300) = 2777(\text{J})$$

(2) If the gas expanded isothermally, from Eq. (11-14)

$$W_T = \frac{M}{\mu}RT\ln\frac{V_f}{V_i} = 2 \times 8.31 \times 300 \times \ln 2 = 3456(\text{J})$$

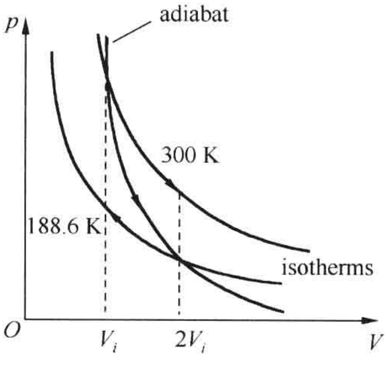


Fig. 11-11 For Example 11-2

The results indicate that $W_T > W_Q$. Because of the temperature drop, the adiabat is steeper than the isotherm passing through the same point, so the area under the isotherm is larger than that under the adiabat, as shown in Fig. 11-11. Conversely, if compressed from the same initial state to the same volume, the work done on the system adiabatically would be larger than that isothermally due to the increase in temperature (i. e. in internal energy), as shown in the same figure.

11.6 Cyclical Processes and Thermal efficiency

11.6.1 Cyclical processes

In industry, in order to operate a machine, or to propel a vehicle or projectile, mechanical energy is required, and one of the problems of the mechanical engineer is to withdraw heat from a high-temperature source and convert it as large a fraction as possible to mechanical energy. This transformation always requires the services of a heat engine, such as a steam engine, gasoline engine, diesel engine, or jet engine, etc.. We consider for simplicity an engine in which the so-called "working substance" is carried through a cyclical process, that is, a sequence of processes in which it eventually returns to its original state. All engines absorb heat from a source at a high temperature, perform some mechanical work, and reject heat into another source at a lower temperature. When a system is carried through a cyclical process, no intrinsic property of the system—including its pressure, volume, temperature and internal energy—can possibly change. And from the first law of thermodynamics, for any complete cycles, we have

$$\Delta E=0$$

$$Q = W \tag{11-35}$$

which means that the net heat flowing into the system in a cyclical process equals the net work done by the system. A cyclical process can be represented by a closed curve on the p-V diagram as shown in Fig. 11-12, and the work is represented by the area enclosed by the curve. For example, the closed curve in Fig. 11-12 shows an arbitrary cyclical process, the area under the upper curve from a to b represents the work done by the system (positive work) in the expansion from a to b, while the area under the lower curve from b to a represents the work done on the system (negative work) in the compression from b to a. Since the positive work exceeds the negative work, so the area bounded by the clockwise curve is the net positive work done by the system, and this process is called positive cycle. Conversely, if process were traversed counterclockwise the net work done by the system would be negative, and the process is called negative or reverse cycle.

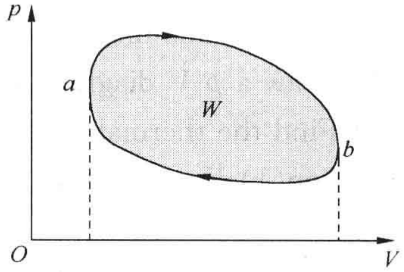


Fig. 11-12 The area enclosed by the closed curve representing a cyclical process equals the net work

11.6.2 Thermal efficiency

For a certain engine, the reservoir at high temperature is called hot reservoir; the reservoir at lower temperature is called cold reservoir (the words "hot" and "cold" are, of course, relative). For a positive cycle, we will use Q_h to represent the positive heat entering the working substance, while use Q_c to represent the absolute value of the negative heat rejected by the working substance. So, the net heat absorbed during a complete cycle is

$$Q = Q_h - Q_c$$

The net work is then

$$W = Q = Q_h - Q_c \tag{11-36}$$

or

$$Q_h = W + Q_c$$

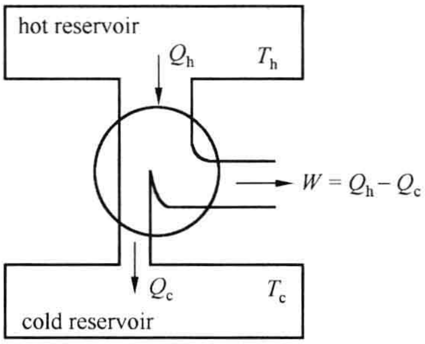


Fig. 11-13 Schematic flow diagram of a heat engine

The energy transformations in a heat engine (positive cycle) are conveniently interpreted schematically by the heat flow diagram of Fig. 11-13, in which the engine itself is represented by the circle.

The thermal efficiency of a cycle (symbol η) is defined as the ratio of the useful work to the heat supplied, that is, "what you get" divided by "what you pay for"

$$\eta = \frac{W}{Q_h} \tag{11-37a}$$

or

$$\eta = \frac{Q_h - Q_c}{Q_h} = 1 - \frac{Q_c}{Q_h} \tag{11-37b}$$

It is obvious that because $Q_c \neq 0$, thermal efficiency $\eta < 1$. Due to the friction losses, the useful work delivered by an engine is less than the work W ($= Q_h - Q_c$), and the overall efficiency is less than the thermal efficiency.

Example 11-3 One mole of a diatomic ideal gas is taken from an initial state of $p_a = 2$ atm, $V_a = 20$ L to $V_b = 2V_a$ at constant pressure, then cooled down to the original temperature with the volume unchanged, eventually compressed to the original state isothermally.

(1) Draw a p-V diagram for this cycle;

(2) Find the thermal efficiency of this cycle.

Solution(1) Fig. 11-14 is the p-V diagram for this cycle.

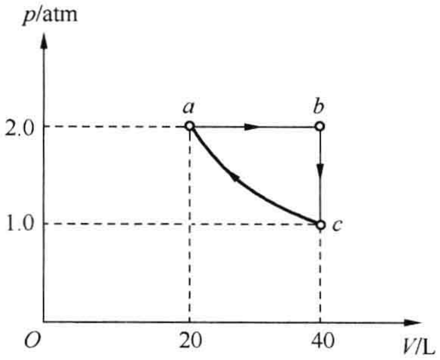


Fig. 11-14 For Example 11-3

(2) According to Eq. (11-8), the work done during the isobaric process is

$$W_p = p_a(V_b - V_a)$$
$$= 2\text{ atm} \times 20\text{ L} \times 101.3\text{ J/(atm} \cdot \text{L)} = 4052\text{ J}$$

In the isochoric process, no work is done, that is $W_V = 0$. From Eq. (11-6), the work done during the isothermal compression is

$$W_T = p_a V_a \ln \frac{V_a}{V_c} = -2\text{ atm} \times 20\text{ L} \times 101.3\text{ J/(atm} \cdot \text{L)} \times \ln 2 = -2809\text{ J}$$

The net work done by the system is then

$$W = W_p + W_V + W_T = 4052 + 0 - 2809 = 1243\text{ (J)}$$

From Eq. (11-25) the heat absorbed in is

$$Q_h = Q_p = \frac{M}{\mu}C_p(T_b - T_a) = \frac{i+2}{2}p_a(V_b - V_a)$$
$$= \frac{5+2}{2}\times 2\ \text{atm}\times 20\ \text{L}\times 101.3\ \text{J/(atm}\cdot\text{L)} = 14180\ \text{J}$$

The thermal efficiency of the cycle is therefore

$$\eta = \frac{W}{Q_h} = \frac{1243}{14180}\times 100\% = 8.76\% \approx 8.8\%$$

Another way to find the result is applying Eq. (11-37b) $\eta = \frac{Q_h - Q_c}{Q_h} = 1 - \frac{Q_c}{Q_h}$. If so, besides calculating Q_p, Q_V and Q_T, you may need to find an unknown parameter, say

$$p_c = p_a \cdot V_a / V_c$$

11.6.3 The gasoline engine and Otto cycle

The common gasoline engine in automobiles is of four-stroke-cycle type, and Fig. 11-15 is a schematic diagram of the cylinder and the piston of it. Starting with the piston at the top of its stroke, an explosive mixture of air and gasoline vapor is drawn into the cylinder, the intake valve being open and the exhaust valve closed; this is the intake stroke. At the end of this stroke, the intake valve closes and the piston rises, performing an approximately adiabatic compression of the air-gasoline mixture, which is the compression stroke. At the top of this stroke a spark ignites the mixture of air and gasoline vapor, and combustion takes place very rapidly so that the pressure and temperature increase at nearly constant volume. Then the piston is forced down, the burned gases expanding approximately adiabatically; this is the working stroke. At the end of the working stroke the exhaust valve opens, the pressure drops rapidly to atmospheric, and the rising piston on the exhaust stroke forces out most of the remaining gas, then the exhaust valve closes, while the intake valve opens, and the cycle is repeated.

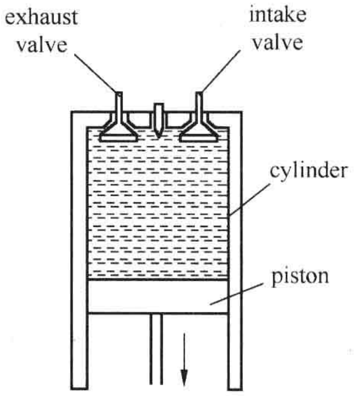


Fig. 11-15 A Schematic diagram of the cylinder and piston of a gasoline engine

The gasoline engine cycle is approximated by the Otto cycle illustrated in Fig. 11-16. The four processes include that

(1) $a\rightarrow b$, compression stroke in which the air is compressed adiabatically from state a (V_1, T_1) to state b (V_2, T_2);

(2) $b\rightarrow c$, explosion stroke, heated at constant volume from state b to state c (V_2, T_3);

(3) $c\rightarrow d$, working stroke, allowed to expand adiabatically from state c to state d (V_1, T_4);

(4) $d\rightarrow a$, exhaust stroke, cooled at constant volume from state d to the original state a, after which the cycle is repeated.

V_1 and V_2 in Fig. 11-16 are respectively the maximum and the minimum volumes of the mixture air in the cylinder. The ratio V_1/V_2 is called the compression ratio. Assuming the mixture of air and gasoline vapor to behave like an ideal gas, the heat input (stroke 2) and the heat rejected (stroke 4) can be represented by

$$Q_h = \frac{M}{\mu}C_V(T_3 - T_2)$$

and $Q_c = \frac{M}{\mu}C_V(T_4 - T_1)$ respectively, and the thermal efficiency of a Otto cycle can be proved as

$$\eta_o = 1 - \frac{Q_c}{Q_h} = 1 - \frac{1}{(V_1/V_2)^{\gamma-1}}$$

where $\gamma = C_p/C_V$. For $V_1/V_2 = 7$ and $\gamma = 1.4$, the thermal efficiency is about 55%, but the practical efficiency is only about 25%. It seems that the higher the compression ratio, the larger the efficiency. However, increasing the compression ratio also increases the temperature at the end of the adiabatic compression of the gas-air mixture. If it is too high, the mixture explodes spontaneously and prematurely, instead of burning evenly after ignition by the spark plug. The mechanical strength and wear of engine parts also pose problems, so the maximum practical value of V_1/V_2 for ordinary gasoline engines is about 7.

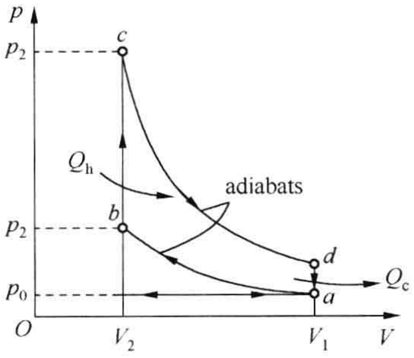


Fig. 11-16 The p-V diagram for an Otto cycle

11.7 Carnot Cycle and Reverse Cycle

11.7.1 Carnot cycle

The efficiencies of various cycles differ from one another, and none of them has an efficiency of 100%. An important question is what is the maximum attainable efficiency, given a supply of heat at a higher temperature and a reservoir at a lower temperature for cooling the exhaust?

An idealized cycle which can be shown to have maximum efficiency under these conditions was invented by the French engineer Saki Carnot in 1824 and thus is called a Carnot cycle. The Carnot cycle, shown on a p-V diagram in Fig. 11-17, consists of two isothermal and two adiabatic processes. Thus all the heat input is supplied at a single high temperature and all the heat output is rejected at a single lower temperature. The following four steps make up the cycle:

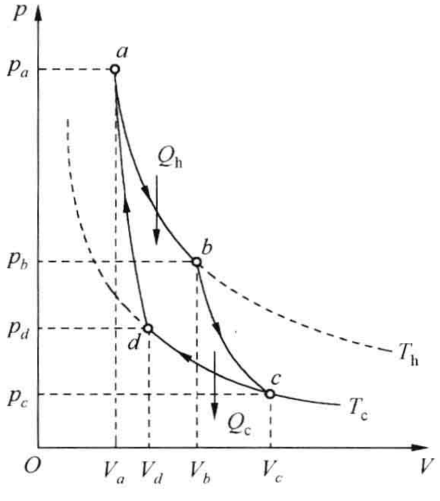


Fig. 11-17 A p-V diagram for the Carnot cycle

(1) The container (say, a cylinder) contacts with the high temperature reservoir (T_h), and the system is an ideal gas, in a state represented by point a in Fig. 11-17. The gas is allowed to expand from the initial state (V_a, p_a) to a new state (V_b, p_b). Because this process is isothermal, the internal energy of the system does not change and all added heat appears as work.

(2) The gas is allowed continuously to expand adiabatically from (V_b, p_b) to (V_c, p_c). During this expansion,

the temperature of the gas drops to T_c, because the energy to do work must come from the internal energy of the system.

(3) The cylinder contacts with the lower temperature reservoir T_c and the gas is compressed isothermally at T_c from (V_c, p_c) to (V_d, p_d). During this process, the work done on the system is converted into heat which is rejected into the lower temperature reservoir.

(4) The gas is continuously compressed adiabatically from volume V_d to the initial volume V_a and all other parameters return to their original values too. Because no heat enters or leaves the system, the work done on the system transfers into the internal energy of the gas, and its temperature rises to T_h.

To find out the thermal efficiency of a Carnot cycle, we start from Eq. (11-37b). Using Eq. (11-14), we have

$$Q_h = \frac{M}{\mu} R T_h \ln \frac{V_b}{V_a} \tag{11-38}$$

$$|Q_c| = \frac{M}{\mu} R T_c \ln \frac{V_c}{V_d} \tag{11-39}$$

Using Eq. (11-33), yields

$$T_h V_b^{\gamma-1} = T_c V_c^{\gamma-1}, \qquad T_h V_a^{\gamma-1} = T_c V_d^{\gamma-1}$$

Dividing the former equation by the later, one results in

$$\frac{V_b}{V_a} = \frac{V_c}{V_d}$$

Substituting it into Eq. (11-38) and Eq. (11-39) respectively and dividing the later by the former, we get

$$\frac{Q_c}{Q_h} = \frac{T_c}{T_h}$$

Substituting it into the definition of thermal efficiency, Eq. (11-37b), we obtain the thermal efficiency of a Carnot cycle

$$\eta_C = 1 - \frac{T_c}{T_h} \tag{11-40}$$

Eq. (11-40) points the way to the conditions in which a real engine, such as a steam engine, must fulfill to approach as closely as possible the maximum attainable efficiency. These conditions are that the intake temperature T_h must be made as high as possible and the exhaust temperature T_c as low as possible. But the exhaust temperature cannot be lower than the lowest temperature available for cooling the exhaust, which is usually the temperature of the air or perhaps of the river water available at the plant. The only recourse then is to raise the temperature of the heat of the steam engine. Since the vapor pressures of all liquids increase rapidly with increasing temperature, a limit is set by the mechanical strength of the boiler.

Example 11-4 The boiler of a steam engine with cylinder and piston produces steam at a temperature of 500℃. The engine exhausts its waste heat into the atmosphere where the temperature is 20℃. The actual efficiency of this steam engine is 15%. Compare this with the efficiency of a Carnot engine operating between the same temperatures.

Solution According to Eq. (11-40), the efficiency of a Carnot engine with $T_h = (500+$

273) K and $T_c=(20+273)$ K is

$$\eta_C = 1-\frac{T_c}{T_h} = 1-\frac{293}{773} = 62\%$$

Thus, the ideal Carnot engine is more efficient than the actual steam engine by 47%. Most modern steam engines employ a turbine wheel instead of the cylinder and piston; large engines of this kind achieve efficiencies of up to 40%.

11.7.2 Reverse cycle and refrigerator

On the contrary to the cycle represented in Fig. 11-12, a reverse cycle can be represented by a counterclockwise curve as shown in Fig. 11-18. A refrigerator may be considered to be a heat engine operated in reverse. In contrast to a heat engine, refrigerator takes in heat from a cold reservoir, the compressor supplies mechanical work input, and heat is rejected to a hot reservoir. With reference to the ordinary home refrigerator, the food and ice cubes constitute the cold reservoir, work is done by the electric motor, and the hot reservoir is the air in the surrounding. Fig. 11-19(a) is a schematic diagram of a refrigerator.

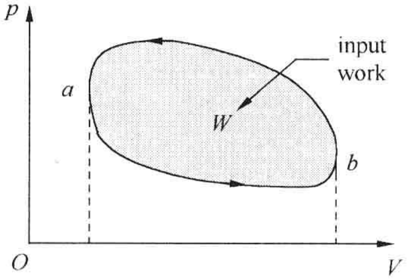


Fig. 11-18 Schematic diagram of a reverse cycle

The flow diagram of a refrigerator is given in Fig. 11-19(b). In a complete reverse cycle, heat Q_c (positive value) enters the refrigerator at low temperature T_c, work W' (negative value) is done on the refrigerator and heat Q'_h (negative value) leaves at a higher temperature T_h. Letting $W=|W'|$, $Q_h=|Q'_h|$, from the first law, we have

$$Q_h = W + Q_c$$

and the heat rejected to the hot reservoir is the sum of the heat taken from the cold reservoir and the heat equivalent of the work done by the motor.

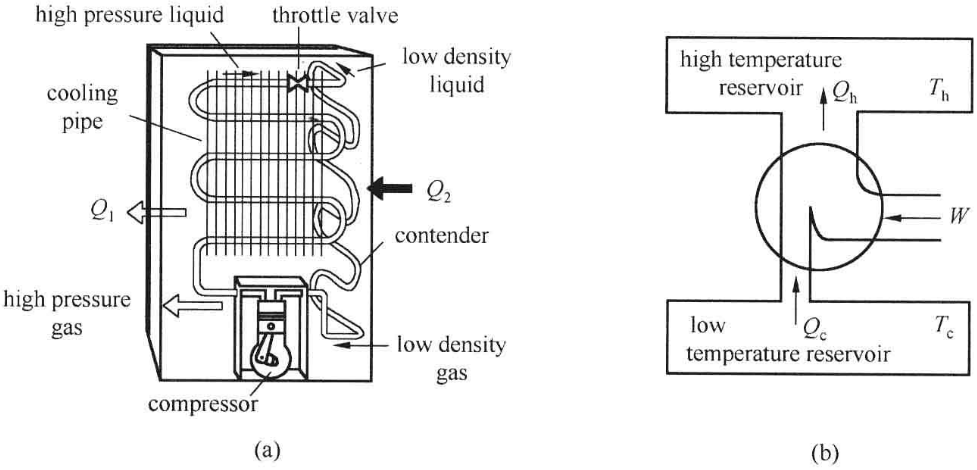


Fig. 11-19 (a) Schematic diagram of a refrigerator; (b) Schematic flow diagram of refrigerator

From an economic point of view, the best refrigeration cycle is one that removes the greatest amount of heat Q_c from the refrigerator, with the least expenditure of mechanical work W. The coefficient of performance (rather than the efficiency) of a reverse cycle is therefore defined as the ratio Q_c/W (Symbol K). Since $W = Q_h - Q_c$, so

$$K = \frac{Q_c}{W} = \frac{Q_c}{Q_h - Q_c} \tag{11-41}$$

A Carnot reverse cycle is a Carnot cycle operated in reverse, which can be interpreted by the curve in Fig. 11-20, traveled counterclockwise. Comparing Fig. 11-20 with Fig. 11-17, it is obvious that the coefficient of performance of a Carnot reverse cycle is

$$K_C = \frac{T_c}{T_h - T_c} \tag{11-42}$$

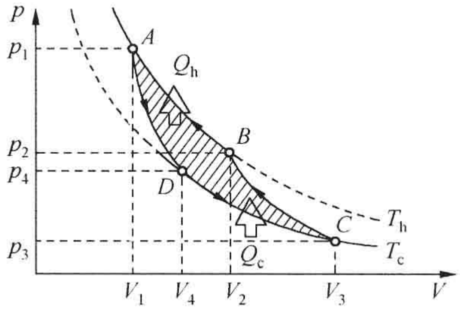


Fig. 11-20 A p-V diagram for a Carnot reverse cycle

Suppose a homeowner uses a Carnot engine operating in reverse as a refrigerator in which the temperature remains at 270 K, and the temperature of the air outside the refrigerator is 300 K, thus the coefficient of performance of the reverse Carnot cycle is

$$K_C = \frac{270}{300 - 270} = 9$$

But the practical coefficient of performance is less than this value.

An air conditioner employs a similar refrigerator cycle. In summer, its cold end is indoor and its warm end outdoor. Conversely, in winter, its cold end is out door and its warm end indoor; it is as "heat pump" to maintain the warm inside the room.

11.8 The Second Law of Thermodynamics

11.8.1 Some things consistent with first law but never happen

You may note one thing in previous discussion, although their thermal efficiencies differ from one another, none of the heat engines or positive cycles have an efficiency of 100%. That is, a heat engine fails to convert the heat absorbed completely into work and instead produces some waste heat. There is nothing in the first law of thermodynamics that precludes this possibility. The first law requires only that the energy out-put of an engine, in the form of mechanical work, should equal the difference between the energies absorbed and rejected in the form of heat. An engine with an efficiency of 100% would therefore be perfectly consistent with the first law. We know now that there is another principle, independent of the first law and not derivable from it, which determines the maximum fraction of the heat absorbed by an engine that can be transferred into mechanical work.

Moreover, there are some things that never happen but they would not be violation of the first law. Here are three examples:

(1) Coffee, resting quietly in your cup, spontaneously cools down and starts to swirl around.

(2) One end of a spoon resting on a table spontaneously gets hot while the other end cools down.

(3) The molecules of the air in the room all move to one corner and stay there.

However, these events occur quite naturally and spontaneously in the other direction. Coffee, swirling in the cup, will eventually stop swirling, its rotational energy changing into internal energy and thus heating it a little. Temperature differences set up between two ends of a spoon tend to equalize. The air molecules rush from their special corner and fill the room uniformly. The world is full of such events that happen in one way but never in the other. We are so used to this that we take them for granted if they happen in the "right" direction but we would be astonished beyond all beliefs if they happened in the other way around. The directions in which natural events happen are governed by the second law of thermodynamics. This law can be expressed in several equivalent forms, and two of them, introduced in the followings, involve simple statements about heat and work.

11.8.2 The second law of thermodynamics

(1) Kelvin statement of the second law — **It is not possible to change heat completely into work, with no other change taking place.**

You can change work into heat completely and rather easily, by rubbing your hands together briskly. For another familiar example, the kinetic energy of a moving car is changed completely into heat in the breaking system when you brake the car to a halt; the reverse process — changing heat into work — is possible as we studied before, but never completely with no other change taking place.

Let us back to Fig. 11-6, during an isothermal expansion process, the heat added from a single reservoir into working substance does transfer completely into useful work on its surrounding but with changes both in volume and pressure of the system taking place. You do eliminate every change when you take the system to undergo a cycle process — but you never be able to avoid the fact of using more than one reservoir and losing some heat. That is, the efficiency of the engine is always less than 100%. This leads to another way to put the Kelvin statement of the second law:

There are no perfect engines.

The engine with $\eta=100\%$, is also called as the second kind of perpetual motion machine. So the Kelvin statement means that it is impossible to make a perpetual motion machine, and none of the attempts and efforts in the history gets success.

(2) Clausius statement of the second law — **It is not possible for heat to flow from one body to another body at a higher temperature with no other change taking place.**

Heat flows naturally from a hot place to a cold place, as from the sun to the earth, but there is never any "natural" net heat flow in the other direction. A refrigerator does cause heat to move from a cold place to a warm place but certainly with some work done on the system by an external agent as shown schematically in Fig. 11-19. That is, the coefficient of performance of a refrigerator, $K=Q_c/W$, is never be infinite. This leads to another way to put the Clausius statement of the second law:

There are no perfect refrigerators.

This means that a refrigerator cannot be operated without some work input. The second law of thermodynamics is concluded from the experiments and observations and is verified by the agreement of the consequences deduced from it with the facts.

(3) Identity of the two statements of the second law. Although these two statements of the second law seem quite different, we shall show that they are exactly equivalent, in the sense that a violation of either implies a violation of the other. That is, if you succeed in building a perfect engine, you can also build a perfect refrigerator and conversely.

Suppose first that the Clausius statement is false and there exists a perfect refrigerator that can make heat flow Q from a cold reservoir to a hot one without any work input. If it operated jointly with an ordinary heat engine, it can be used to remove the waste heat $Q_c = Q$ from the cold reservoir and return it to the hot one. The net effect would be the complete conversion of heat $Q_h - Q$ from the hot reservoir into work, as shown in Fig. 11-21, in contradiction to the Kelvin statement. Conversely, suppose the Kelvin Statement is false and a perfect heat engine exists and converts heat Q_h from a hot reservoir completely into work. Then the work output of it can be used to operate an ordinary refrigerator and the net effect would be the transfer of heat Q_c from a cold reservoir to the hot one with no other change, as shown in Fig. 11-22, in contradiction to the Clausius statement.

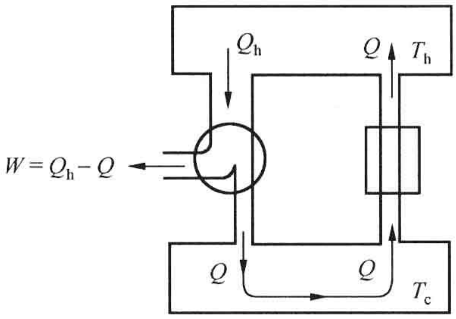


Fig. 11-21 If there were a perfect refrigerator, there would be a perfect heat engine

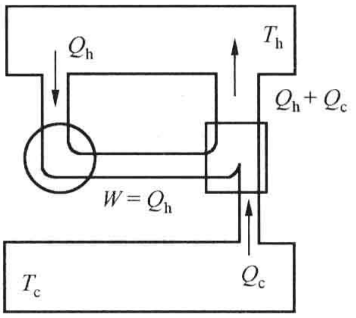


Fig. 11-22 If there were a perfect heat engine, there would be a perfect refrigerator

11.9 Reversible and Irreversible Process, Carnot Theorem

11.9.1 Irreversible processes

Suppose that a process occurs in which the system proceeds from the initial state to the final state; there does not exist any way to restore the system back to the initial state with no other change taking place, this process is then defined as an irreversible process. Conversely, if there is one way to restore the system back to the initial state with no other change taking place, the process is reversible.

The natural process spontaneously happens in one way not in the other, that is, the directionality of the natural process implies that all natural spontaneous processes are irreversible. The following processes are irreversible.

(1) The process in which work transfers into heat is irreversible: the fact that work can be dissipated completely into heat, whereas heat can not be converted entirely into work, expresses an essential one-sidedness of nature. So that, the Kelvin statement of the second law of thermodynamics simply means that the process in which **work transfers into heat is irreversible.**

(2) The process in which heat flows from a hotter to a colder body is irreversible: the fact that heat always flows spontaneously from a body to another at a lower temperature but never flows spontaneously from a colder to a hotter body indicates another essential one side of nature. Therefore, the Clausius statement of the second law of thermodynamics simply means that **the process of heat flow is irreversible.**

(3) Free expansion is irreversible:

Suppose that a gas is confined by an insulating barrier to half of an insulated container, the other half being evacuated as shown in Fig. 11-23(a). Free expansion process consists of removing the barrier and waiting until equilibrium is established, with the gas now filling both halves of the container (Fig. 11-23(b)). No heat is transferred because of the insulation. No work is done because the expanding gas rushes into an evacuated space, its motion unopposed by any counteracting pressure. Thus, $Q=0$, $W=0$ and the first law requires that $\Delta E=0$, so $\Delta T=0$.

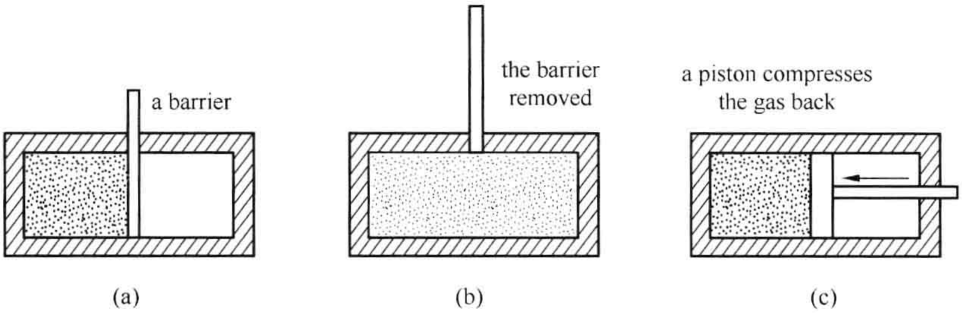


Fig. 11-23 (a) Initial state of the free expansion; (b) Final state of the free expansion; (c) If we want the gas back to the initial state, we must do work on it

If we use a piston to compress the gas adiabatically back to the left part of the container (Fig. 11-23(c)), we must do some work on the system, which is transferred into the internal energy, and the temperature of the gas is therefore increased. If we want both of the gas and the surrounding reverse to the initial state, we must take some heat out of the gas and transfer it into work. According to the Kelvin statement of the second law, it is impossible to convert the heat completely into work with no other change taking place. It is therefore that **the free expansion process is irreversible.**

These are countless examples of irreversible processes that take place naturally in only one direction. Our conclusion is that all natural spontaneous processes are irreversible.

11.9.2 Reversible processes

If it is possible to operate a process in reverse to the initial state with no other change taking place, this process is defined as reversible. A process carried out without friction and so slowly that the process can be reversed at any stage by making an infinitesimal change in the environment of the system, can be regarded as a reversible process.

The isothermal and adiabatic processes of a Carnot cycle may be imagined to process in either direction. To approach reversibility with this cycle, the motion of the piston must be sufficiently slow, so that the gas can be always in an equilibrium state. If the piston were to have a sudden, fast motion, a pressure wave would travel through the gas, and the motion of this pressure wave could not be reversed by giving the piston a sudden motion in the opposite direction—this would merely create a second pressure wave. Furthermore, the temperature of the gas must coincide with the temperature of the heat reservoir during contact. If the gas were to have, say, a lower temperature than that of the heat reservoir with which it is in contact, heat would rush from the reservoir into the gas and this flow of heat could not be reversed by any manipulation of the piston. For reversibility, the gas must remain in mechanical equilibrium with the piston and in thermal equilibrium with the heat reservoir. And finally, the process must be attended with no friction, otherwise there would be a dissipative effect, that is, some amount of work would be converted into heat.

Note that a reversible process represents the same sort of idealization in thermodynamics as a point mass, a weightless inextensible cord, or a frictionless and massless pulley represents in mechanics. It may be approximated but not be perfectly achieved.

11.9.3 Carnot theorem

From the second law of thermodynamics, it has been proved that

(1) No real engine operating between two given temperatures can have a greater efficiency than that of a Carnot engine operating between the same two temperatures.

And this leads to the following:

(2) All Carnot engine operating between the same two temperatures have the same efficiency, irrespective of the nature of the working substance.

These two statements are known as Carnot theorem. We can put them together mathematically in an equation:

$$\eta = 1 - \frac{Q_c}{Q_h} \leqslant 1 - \frac{T_c}{T_h}$$

where " $<$" refers to the irreversible engine, and "$=$" refers to the reversible engine. Carnot theorem points the ways to increase the thermo-efficiency, beside that we have discussed on the thermo efficiency of Carnot cycle, and improving the cycle as close the reversible as possible.

11.10 Statistical Meaning of the Second Law of Thermodynamics

In the preceding sections, the second law of thermodynamic was discussed from a macroscopic point of view, but what is the meaning of the second law of thermodynamics from the statistical point of view? Now, let us discuss this problem in detail.

11.10.1 Thermodynamic probability and the direction of process proceeding spontaneously

Consider again the free expansion process of an ideal gas. Assume, for simplicity, that there are four molecules *a*, *b*, *c* and *d* in the left half of the vessel before expansion (refer to Fig. 11-23(a)). After the barrier is removed, each molecule can be either in the left half or the right half of the container with an equal probability. All the possible distributions of the four molecules in two halves are shown in Table 11-2.

Table 11-2 Distribution of molecules ($N=4$) and thermodynamic probability

Distribution of molecules (microstates)		The corresponding Macrostate (configuration)		Number of microstates corresponding to the macrostate	Probability of occurrence of the macrostate
left	right	left	right		
abcd	—	4	0	1	1/16
abc	d	3	1	4	4/16
acd	b				
abd	c				
bcd	a				
ab	cd	2	2	6	6/16
ac	bd				
ad	bc				
bc	ad				
bd	ac				
cd	ab				
a	bcd	1	3	4	4/16
b	cda				
c	dab				
d	abc				
—	abcd	0	4	1	1/16

In Table 11-2, each possible distribution of the molecules between the two halves of the vessel gives one microstate of the system. But for macroscopic description, since the molecules are identical with each other, the number of molecules in the two halves determines a macro-state (or a configuration) of the system. For example, the state with (3, 1) split, and that with (2, 2) split are two different macro-state. So, there are five macro-states in our case. The number of various microstates corresponding to a given macro-state is called thermodynamic probability (labeled Ω) of that macro-state. From Table 11-2, we can see that the macro-state (2, 2) corresponds to 6 microstates, so its thermodynamic probability $\Omega=6$; while macro-state (4, 0) corresponds to 1 microstate with a thermodynamic probability $\Omega=1$. According to the rules of permutation, if there are N molecules in the container, then the macro-state or configuration with a $(N-n, n)$ split corresponds to $\frac{N!}{(N-n)!n!}$ microstates, i. e. its thermodynamic probability also called multiplicity of that configuration.

Statistical mechanics assumes that the probability for each microstate is equal, thus the probability of the occurring of a macro-states depends on number of corresponding microstates, or its thermodynamic probability Ω. The larger the Ω, the more the probability of the macro-state occurs. In our case, the sum of microstates $\sum\Omega=16$, so that, the probability of the occurring of macro-state (2, 2), that is the equilibrium state, is 6/16, the maximum value, while of macro-state (4, 0), that is reverse state, is 1/16, the minimum value.

From Table 11-2, the following conclusion can be gotten:

The free expansion process is an irreversible process from macroscopic point of view; the process that all molecules return spontaneously to the initial state is the inverse process of the free expansion, so it is impossible according to the second law of thermodynamic. In the statistical physics, if its thermodynamic probability is the smallest, the probability that it may occur is the smallest too. So, the free expansion process of gas proceeds spontaneously from the state of smaller thermodynamic probability to the state of larger thermodynamic probability i. e. $\Delta\Omega>0$.

Therefore, whether a process can proceed spontaneously, the increase of the thermodynamic probability is a criterion.

Statistical physics reveals the nature of irreversible processes occurring in an isolated system. Let us assume that first a gas of $N=10^{25}$ molecules in the left half of a vessel separated by a barrier from the right empty half. If we remove the barrier, the gas spontaneously spreads out over the entire vessel, and remains in equilibrium state-uniform distribution. This process will be irreversible because the ratio of thermodynamic probability of macro-states (10^{25}, 0) to that of macro-state ($10^{25}/2$, $10^{25}/2$) is actually zero. Hence the gas cannot concentrate in the left half of the vessel spontaneously.

For other thermal processes, such as work converting into heat and heat transfer, we can make similar analysis and conclude that for an isolated system, the process always proceeds spontaneously from the state of smaller thermodynamic probability to that of larger one. This is the statistic meaning of the second law of thermodynamics. Thus, the process of the spreading of the gas over the entire vessel is irreversible, or, the process from the state of smaller thermodynamic probability to the state of larger thermodynamic probability is irreversible.

Putting the conclusion above in the other way, we can say that whether or not $\Delta\Omega>0$ is a criterion of a process spontaneously proceeding.

11.10.2 Boltzmann relation between thermodynamic probability and entropy (a state function)

Consider the fact that there are usually a vast number of molecules in a system, say, $N=10^{15}$, or 10^{20}, even 10^{25}, so that, the thermodynamic probability of a macro-state is usually a very larger number too, which is not convenient to be used. Can we use another quantity to describe a macro-state instead of using thermodynamic probability? The answer is "yes", and we can predict the direction of a process proceeding according to the change in this quantity, which is called entropy—another function of state, labeled S. Let us leave the theoretical discussion on entropy for the time being and concentrate the major problem on the relation between entropy and thermodynamic probability, which was established first by Ludwig Boltzmann, the constant k after his name we have used, the relation is

$$S=k\ln\Omega \tag{11-43}$$

and

$$S_f - S_i = k\ln\frac{\Omega_f}{\Omega_i} \tag{11-44}$$

where S_i and Ω_i correspond to the initial macro-state of a system while S_f and Ω_f to the final macro-state for a given process. It is obvious that if

$$\Omega_f > \Omega_i \quad \text{or} \quad \frac{\Omega_f}{\Omega_i} > 1$$

then

$$S_f - S_i > 0 \quad \text{or} \quad \Delta S > 0 \tag{11-45}$$

Since for an isolated system, the process always proceeds spontaneously in the direction of increasing thermodynamic probability, Eq. (11-45) tells us that **the natural process always proceeds spontaneously in the direction of increasing entropy, that is known as the principle of the increase of entropy.**

11.10.3 A measurement of the disorder of a system

Now, the most probable question that you may ask is: from the kinetic theory point of view, what is the thermodynamic probability all about? To answer this question, let us back to the example of free expansion again. By looking closely at Table 11-2, we know that thermodynamic probability Ω, the number of microstates corresponding to a given macro-state—the ways in which positions in space can be assigned to the individual molecule of the gas symbolizes the disorder level of the molecular random motion in the gas; the larger the Ω, the higher the disorder level. So, we can say that thermodynamic probability is a measurement of the disorder of the random motion of the gas.

Another example is that if we raise the temperature of a gas by adding heat to it, according to Maxwell distribution of speed, there are more ways in which velocities can be assigned to the molecules of the gas, or, the molecular motion becomes more vigorous. On the other hand, the thermodynamic probability Ω in this case, can be proved to be increased too. Examples of this sort could be added without limit. Suffice it to say that disorder can be defined quantitatively by thermodynamic probability Ω.

Furthermore, the formal relation between thermodynamic probability Ω and entropy (as a state function) S, $S = k \ln \Omega$, implicates that entropy is also a measurement of the disorder of the molecular random motion in a system. Therefore, the signification of the principle of increase entropy, Eq. (11-45), is that natural process always proceeds spontaneously from the state of lower disorder to the state of higher disorder, that is, proceeds in the direction of increase in disorder.

11.11 Entropy

11.11.1 Entropy — a new function of state

Each of the two laws of thermodynamics is associated with a specific function of state (that is a thermodynamic variable). For the first law, the function is internal energy E. For the second law, the function is entropy S which we introduced in the last section without proof. It is our plan now to define entropy as a state function and, later to express the second law of thermodynamics in term of that variable.

Let us start by considering the Carnot cycle of Fig 11-17, in which it is proved that

$$\frac{Q_c}{Q_h}=\frac{T_c}{T_h} \quad \text{or} \quad \frac{Q_h}{T_h}=\frac{Q_c}{T_c}$$

So that

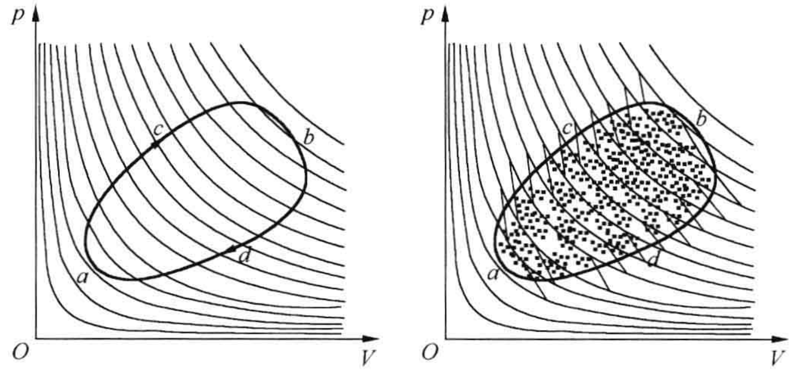
$$\frac{Q_h}{T_h}-\frac{Q_c}{T_c}=0$$

where $Q_c=$ the absolute value of the heat rejected, and thus is positive. If Q_c is defined as the heat added from the reservoir at lower temperature, we can rewrite the above equation as

$$\frac{Q_h}{T_h}+\frac{Q_c}{T_c}=0 \tag{11-46}$$

in which Q_c is negative. Eq. (11-46) tells us that the algebraic sum of the quantity Q/T, taken around the closed cycle of Fig. 11-17 (Carnot cycle) is zero.

We now wish to generalize Eq. (11-46), writing it in a form that applies not only to a Carnot cycle but to any reversible cycle. Fig. 11-24(a) shows such a generalized cycle, superimposed on a family of isotherms. We can approximate this arbitrary cycle as closely as we wish by connecting the isotherms with short, suitably-chosen adiabatic lines, as in Fig11-24(b). In this way, we form an assembly of long, thin Carnot cycles. Convince yourself that traversing the individual Carnot cycles in Fig. 11-24(b) in sequence is exactly equivalent in terms of heat transferred and work done to traversing the jagged series of isotherms and adiabats that approximate the actual cycle. This is so because adjacent Carnot cycles have a common isotherm and the two traversals, in opposite directions, and cancel each other in the region of overlap as far as heat transferred and work done are concerned.



(a) (b)

Fig. 11-24 (a) An arbitrary reversible cycle, plotted on a p-V diagram; (b) Superimposing this cycle with a family of isotherms, connecting the isotherms with short, suitably-chosen adiabatic lines, we form an assembly of long, thin Carnot cycles

We extend Eq. (11-46) by writing for the isotherm-adiabatic sequence of lines in Fig. 11-24(b)

$$\sum \frac{Q_i}{T_i}=0$$

In the limit of infinitesimal temperature differences between the isotherms of Fig. 11-24(b), this becomes

$$\oint \frac{dQ}{T} = 0 \quad \text{(reversible cycle)} \tag{11-47}$$

The cycle can be considered as a combination of two processes *acb* and *bda*, thus

$$\oint_{acbda} \frac{dQ}{T} = \int_{acb} \frac{dQ}{T} + \int_{bda} \frac{dQ}{T} = 0$$

So that

$$\int_{acb} \frac{dQ}{T} = \int_{adb} \frac{dQ}{T} = \int_a^b \frac{dQ}{T} \quad \text{(reversible process)} \tag{11-48}$$

Because the cycle is arbitrarily chosen, Eq. (11-48) indicates that the value of the integral is independent of path, but depends only on the initial state and the final state. Comparing this with the work done by conservative force and the definition of potential energy in chapter 3, we have that $\int_a^b \frac{dQ}{T}$ must represent a difference in some state function between state a and state b, and call this new function or thermodynamic variable the entropy, symbol S, of the system. Thus, if S_i and S_f represent the entropy of initial and final states of a system respectively, according to Eq. (11-48), we have

$$S_f - S_i = \int_i^f \frac{dQ}{T} \quad \text{(reversible process)} \tag{11-49}$$

where the integral is carried out over any reversible path connecting these two states. The SI unit for entropy is the joule per kelvin or J/K.

Note that Eq. (11-49) defines the difference in entropy between two states, rather than the actual entropy of a state. As for potential energy, however, our concern is always with such difference. An arbitrary constant can be added to the absolute entropy of any state without altering any of our conclusions.

Example 11-5 Calculate the entropy of an ideal gas

According to the first law and Eq. (11-49), for a differential reversible process

$$dS = \frac{dQ}{T} = \frac{dE + p dV}{T}$$

in which

$$dE = \frac{M}{\mu} C_V dT \quad \text{and} \quad p dV = \frac{M}{\mu} \frac{RT}{V} dV$$

So that $dS = \frac{M}{\mu}\left(C_V \frac{dT}{T} + \frac{R dV}{V}\right)$. Integrating the two sides, we obtain the entropy change

$$S_f - S_i = \frac{M}{\mu}\left(C_V \ln \frac{T_f}{T_i} + R \ln \frac{V_f}{V_i}\right) \tag{11-50a}$$

Dropping the subscript f and merge all integral constants into one constant, we can write the absolute entropy as

$$S = \frac{M}{\mu}(C_V \ln T + R \ln V) + S_0 \tag{11-50b}$$

Eq. (11-50b) tells us that the entropy of an ideal gas at any equilibrium state is determined by parameters T and V, and is a single valued function of the state.

11. 11. 2 Entropy statement of the second law

For any irreversible process, if its initial state and its final state are equilibrium states, thus, from Carnot theorem and Eq. (11-49), it can be proved that

$$S_f - S_i > \int_i^f \frac{dQ}{T} \quad \text{(irreversible process)} \tag{11-51}$$

Combining Eq. (11-49) and Eq. (11-51), we have

$$S_f - S_i \geqslant \int_i^f \frac{dQ}{T} \tag{11-52a}$$

Note that entropy is a single valued function of state of a system, so for given state a and state b, the change in entropy $S_f - S_i$ has a definite value. Eq. (11-52a) indicates that, the value of the integral on the right side of the equation equals $S_f - S_i$ for a reversible process, but is smaller than $S_f - S_i$ for a irreversible process.

For a differential process, using Eq. (11-52a), we have

$$dS \geqslant \frac{dQ}{T} \tag{11-52b}$$

Suppose the process is an adiabatic, $dQ = 0$, thus

$$dS \geqslant 0 \tag{11-53}$$

which means that **the possible changes for an isolated system are those in which the entropy of the system either increases or remains the same. Changes in which the entropy decreases will not happen. Eq. (11-53) is called as the principle of the entropy increasing.**

The entropy remains the same only for reversible process. No process in nature is truly reversible so that, in essentially all cases, we look for an entropy increase when we see a process occur spontaneously. It is therefore that entropy is a criterion to judge whether or not an irreversible process can spontaneously occur. If the system is not isolated, we must regard the system + environment as a large isolated system, then calculate the total change in entropy so as to judge if the process happens.

We are now ready to formulate the second law of thermodynamics in terms of entropy. The Second law — entropy statement: **In any thermodynamic process that proceeds from one equilibrium state to another, the entropy of the system + environment either remains unchanged or increases.**

There is no way in which you can make the entropy of the system + environment decrease. It is true that the entropy of the system can be made to decrease but that decrease must always be accompanied by an automatic equal or greater increase in the entropy of the environment.

The example we mentioned before, the spontaneous movement of the air in the room to one corner, is what we may call a "free compression", the opposite of a free expansion. We saw in the preceding section that a free expansion is always accompanied by an increase of entropy, while a "free compression" would result in an entropy decrease and would thus be a violation of the entropy statement of the second law; therefore, it never happens.

11.11.3 Calculate the entropy change of a system

Eq. (11-49) tells us how to calculate the entropy change of a system that undergoes a reversible process. But there are no such processes in the real world. Every actual thermodynamic process is — to a greater or lesser extent — irreversible.

How can we calculate the entropy change between the initial and final states in such cases? We take advantage of the fact that the difference in entropy between two equilibrium states does not depends on how the system passes from one state to another; to find the entropy change for an irreversible process between two equilibrium states, imagine a reversible process connecting those same states and calculate the entropy change for that process.

Example 11-6 A lump of ice whose mass is 1 kg melts to water, the temperature remaining at 0℃ (and the pressure remaining at 1 atm) throughout the process. What is the change in entropy for the ice? The heat of fusion of ice is $L=3.35\times10^5$ J/kg.

Solution The process during which ice melts to water is irreversible. Suppose that we put the ice in contact with a heat reservoir whose temperature exceeds 0℃ by only a differential amount (i. e. if we lower the reservoir temperature to a differential amount below 0℃, the melted ice will start to freeze.), thus the process can be regarded as a reversible one, and we can use Eq. (11-49) to find the entropy change:

$$S_w - S_i = \int \frac{dQ}{T} = \frac{1}{T}\int dQ = \frac{Q}{T}$$

And $Q=mL=1\ \text{kg}\times3.35\times10^5\ \text{J/kg}=3.35\times10^5$ J. Thus

$$S_w - S_i = \frac{3.35\times10^5}{273.15} = 1.23\times10^3\,(\text{J/K})$$

Example 11-7 The entropy change for free expansion

Suppose that an ideal gas (of mass M and mole mass μ) undergoes a free expansion from volume V_i to V_f in an isolated container. Calculate the change in entropy for the gas.

Solution (1) We know the free expansion being a irreversible process during which $Q=0$, $W=0$ and $T_f=T_i$. If we try to use Eq. (11-49) to find the entropy change, the fact that $Q=0$ might lead us to predict that $\Delta S=0$, which we shall see is incorrect. Following our procedure above, we must find a reversible process that has the same initial and final states as that of our free expansion. Because $T_f=T_i$ for the free expansion, a convenient reversible process is an isothermal expansion from V_i to V_f. Applying Eq. (11-49) to the isothermal expansion yields

$$S_f - S_i = \int_{V_i}^{V_f} \frac{dQ}{T} = \frac{1}{T}\int_{V_i}^{V_f} dQ = \frac{Q_T}{T}$$

where $Q_T = \frac{M}{\mu}RT\ln\frac{V_f}{V_i}$ according to Eq. (11-14). Substituting it into Q_T/T leads to

$$S_f - S_i = \frac{M}{\mu}R\ln\frac{V_f}{V_i} \quad \text{(free expansion)} \tag{11-54}$$

Although we calculated this entropy change specifically for a reversible isothermal expansion between two states, it holds for any process connecting those same states, including a free expansion.

Because $V_f > V_i$, Eq. (11-54) tells us that the entropy of the system increases during a free expansion. Note that this process takes place in an isolated rigid container, thus the entropy of the environment does not change, but the entropy of the system+ environment increases during a free expansion.

(2) Using Eq. (11-50a), the entropy formula of ideal gas, and noting that $T_f = T_i$, we have the entropy change:

$$S_f - S_i = \frac{M}{\mu}\left(C_V \ln \frac{T_f}{T_i} + R \ln \frac{V_f}{V_i}\right) = \frac{M}{\mu} R \ln \frac{V_f}{V_i}$$

This result is just as same as that obtained by using Eq. (11-49). Obviously, as long as the initial and the final states are given, the entropy change is uniquely determined, which is independent of the processes. The discussions above gave further evidence that entropy is a state function.

Questions

11-1 Analyze whether the following states are correct or not:

(1) The higher the temperature of a body, the more its heat;

(2) The higher the temperature of a body, the greater its internal energy.

11-2 Does a gas do any work when it expands adiabatically? If so, what is the source of the energy needed to do this work?

11-3 A quantity of gas occupies an initial volume V_0 at pressure p_0 and temperature T_0. It expands to a volume V:

(1) at constant temperature and (2) at constant pressure. In which case does the gas do more work? Explain why?

11-4 Explain why we use equation $E_2 - E_1 = \frac{M}{\mu} C_V (T_2 - T_1)$ to calculate the change in internal energy for all different processes including the isobaric, the isochoric, the adiabatic and any other processes.

11-5 The value of molar heat capacities $C = \mathrm{d}Q/\mathrm{d}T$ of a gas could be any amount, why? Describe the process in which the corresponding heat capacities are (1) zero, (2) infinite, (3) positive, (4) negative respectively.

11-6 During the process represented by a straight line in Fig. 11-25, if $E_A = E_B$, whether or not $\mathrm{d}E = 0$ holds for all differential stages?

11-7 Draw some schemes to illustrate the processes of (1) an isobaric, (2) an isochoric, (3) an isothermal for a ideal gas on a p-T diagram, an E-T diagram, a V-T diagram and an E-V diagram respectively.

11-8 (A) Figure 11-26(a) shows that a gas undergoes thermal processes, AC is an isotherm. Decide the signs (positive or negative) of ΔE, ΔT, W and Q for the process (1) along ABC, (2) along $CB'A$. (B) Answer the same questions for the processes (1) along ABC, (2) along $AB'C$ shown in Fig. 11-26(b) where AC is an adiabat.

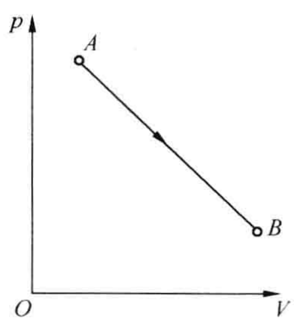


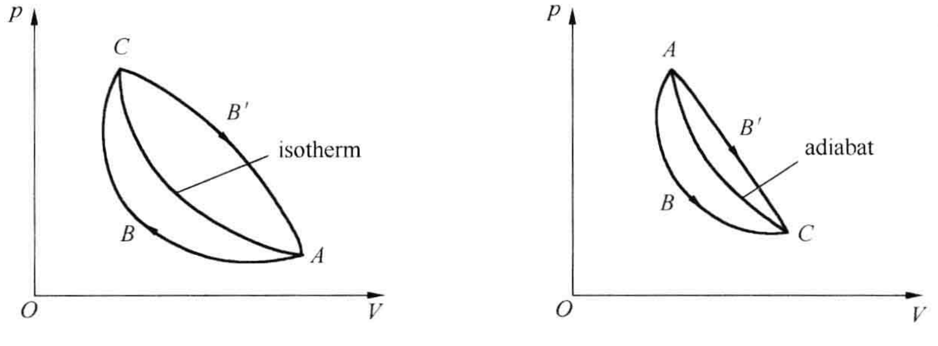
Fig. 11-25 For question 11-6

11-9 If there are Carnot engines 1 and 2 operating between the same two temperatures. Do the following equations hold definitely? (1) $\eta_{c1} = \eta_{c2}$, (2) $W_1 = W_2$ (net work), (3) $Q_{h1} = Q_{h2}$. If there is any case in which one of the equation is true, illustrate it by a figure.

11-10 Some students understand the second law of thermodynamics in such ways:

(1) work can transfer into heat at 100% while heat can not transfer into work at 100%;

(2) heat can flow from a hot body to a cold body while can not flow from a cold one to a hot one. Point out where are the misunderstandings.



(a) (b)

Fig. 11-26 For question 11-8

11-11 (1) Can one adiabatic process and two isothermal processes form a cycle? Explain.

(2) Can one isothermal process and two adiabatic processes form a cycle? Give the reason.

11-12 Two samples of a gas initially at the same temperature and pressure are compressed from volume V to a volume $V/2$, one isothermally, the other adiabatically. In which sample the final pressure greater? Does the entropy of the gas change in either process?

11-13 If in Fig. 11-23(a) there is O_2 in the left half while the air in the right half with same pressure.

(1) Removing the barrier, does the entropy of the system increase? Unchanged? Or Decrease?

(2) If both sides originally are filled by O_2, what is the answer now?

Problems

11-1 1 g of water(1 cm^3)becomes 1671 cm^3 of steam when boiled at a pressure of 1 atm. The heat of vaporization at this pressure is 2256 J/g. Find the work done by the vapor and the increase in internal energy.

11-2 A diatomic ideal gas absorbs 500cal of heat at a constant pressure. Find the work done by the gas during this process.

11-3 Oxygen gas of 10g, initially under a pressure of 3×10^5 Pa at a temperature of 283 K, is heated along the isobar until it occupies a volume of 10 liters. Find

(1) the amount of heat added to the gas;

(2) the change in the internal energy of the gas;

(3) the work performed by the gas.

11-4 When a system is taken from state i to state f along the path iaf in Fig. 11-27, it is found that $Q=50$ cal and $W=20$ cal. Along the path ibf, $Q=36$ cal.

(1) What is W along the path ibf ?

(2) If $W=-13$ cal for the curved return path fi, what is Q for this path?

(3) Taking $E_i=10$ cal, what is E_f?

(4) If $E_b=22$ cal, what are the values of Q for processes ib and bf?

11-5 A certain amount of oxygen occupies a volume of $V_1=3$ L under a pressure of $p_1=8.2\times10^5$ Pa in Fig. 11-28. If the parameters of the second state of the gas are $V_2=4.5$ L and $p=6\times10^5$ Pa, find the amount of heat received by the gas, the work performed by the gas and the change in internal energy of the gas in the following cases: the process is along the path of (1) ACB, (2) ADB.

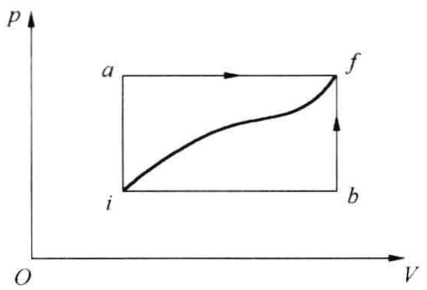


Fig. 11-27 For problem 11-4

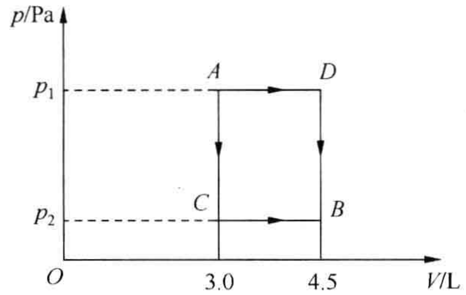


Fig. 11-28 For problem 11-5

11-6 During an isothermal expansion process, the pressure changes from $p_1=5$ atm to $p_2=4$ atm. If $V_1=2\ \mathrm{m}^3$, determine the work done.

11-7 2.8 g nitrogen gas at $t_1=27^\circ\mathrm{C}$ and $p_1=1$ atm, is heated first isobarically until the volume doubled, then continuously heated isochorically until the pressure doubled, and at last, by an isothermal expansion, its pressure drops down to 1 atm.

(1) Draw a p-V diagram for the whole process.

(2) Find the heat added into the gas in the three processes, and calculate the work done by the gas and the change in internal energy in the processes respectively.

11-8 Prove that the work done by a gas during an adiabatic expansion from initial state (p_1, V_1, T_1) to (p_2, V_2, T_2) can be written as

$$W=\frac{1}{\gamma-1}(p_1V_1-p_2V_2)=\frac{M}{\mu}\frac{RT_1}{\gamma-1}\left[1-\left(\frac{V_1}{V_2}\right)^{\gamma-1}\right]$$

11-9 A diatomic gas occupies a volume of $V_1=0.5$ L under a pressure of $p_1=0.5$ atm. The gas is compressed adiabatically to a certain volume V_2 and a pressure p_2, and is then cooled down to the initial temperature at a constant volume of V_2, its pressure becomes equal to $p_0=1$ atm.

(1) Plot the p-V diagram of this process.

(2) Determine the volume V_2 and the pressure p_2.

11-10 Ten grams of oxygen in standard condition are compressed to a volume of $1.4\times10^{-3}\ \mathrm{m}^3$. Find the pressure and temperature of the gas after compression if (1) the gas is compressed isothermally, (2) the gas is compressed adiabatically. Find the work of compression in each case.

11-11 The initial state of a diatomic ideal gas is $p_1=1$ atm, $V_1=1$ L, and it is then allowed to expand until $V_2=2$ L. Suppose that the expansion is (a) isothermal, (b) adiabatic.

(1) draw the process curves on a p-V plane;

(2) calculate the heat added to the gas, the work done by the gas and the change in internal energy in each case respectively.

11-12 An ideal gas is caused to go through the cycle shown in Fig. 11-29. Process bc is an adiabatic expansion. Prove that the thermal efficiency of this cycle can be expressed as

$$\eta=1-\gamma\frac{\left(\frac{V_1}{V_2}\right)-1}{\left(\frac{p_1}{p_2}\right)-1}$$

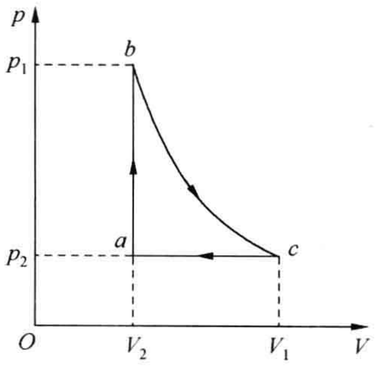


Fig. 11-29 For problem 11-12

11-13 One mole of an ideal monatomic gas initially at a volume of 10 L and a temperature of 300 K, is heated at constant volume to a temperature of 600 K, then allowed to expand isothermally to its initial pressure, and finally compressed isobarically to its original state.

(1) Draw the cycle curve on p-V plane;

(2) Compute the heat input to the system during one cycle;

(3) What is the net work done by the gas during one cycle?

(4) What is the efficiency of this cycle?

11-14 One mole of an ideal diatomic gas at the initial state of $T_1 = 300$ K and $V_1 = 4$ L, is allowed first to expand isobarically until $V_2 = 6.3$ L, then expands adiabatically to its original temperature, and finally is compressed isothermally to the original state.

(1) Draw the p-V diagram;

(2) Calculate the thermal efficiency.

11-15 An ideal heat engine undergoes a Carnot cycle and receives 600 cal of heat from a hot source each cycle. The temperature of the hot source is 400 K, and that of the cold sink 300 K. Find the efficiency and the work performed by the engine per cycle, and the amount of heat rejected to the cold sink per cycle.

11-16 An ideal heat engine undergoes a Carnot cycle and performs work equal to 7.35×10^4 J during one cycle. The temperature of the hot reservoir is 373 K and that of the cold reservoir 273 K. Find:

(1) the thermal efficiency;

(2) the amount of heat received by the engine per cycle from the hot reservoir;

(3) the amount of heat rejected to the cold sink per cycle.

11-17 To make ice, a freezer extracts 42 kcal of heat at -12 ℃. The freezer has a coefficient of performance of 5.7. The room temperature is 26℃.

(1) How much heat was delivered to the room?

(2) How much work is required to run the freezer?

11-18 A Carnot refrigerator transfers 1 kg of water at 273 K into ice. The room temperature is 300 K.

(1) How much work is required to run the refrigerator?

(2) How much heat is rejected into the surrounding? The fusion heat of ice is $L = 3.35 \times 10^5$ J/kg.

11-19 A Carnot refrigerator extracts 100J of heat from a cold reservoir and delivers it into a warm reservoir at 27℃. How much work is required to run the cycle if the temperature of the cold reservoir is at (1) 7℃, (2) -173℃, and (3) -223℃? What conclusion can you get from the results?

11-20 The Rankine cycle of an ideal steam engine is shown in Fig. 11-30.

(1) When the process of the steam entering from the boiler into the cylinder begins, the pressure increases from p_0 to p_1 at constant volume V_0 (line AB);

(2) When the piston moves from left to right (line BC), the steam expands isobarically while steam entering continues;

(3) When the piston continues to move to the right, the supply of steam is shut off and the steam expands adiabatically (line CD);

(4) When the piston is in its extreme right position the steam is rejected from the cylinder, and the pressure drops to p_0 at a constant volume V_2 (line DE);

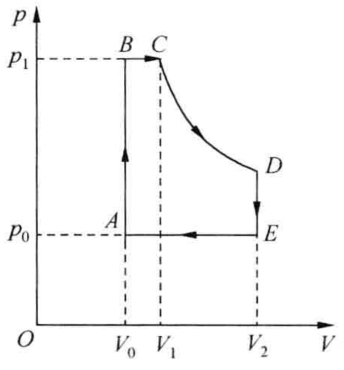


Fig. 11-30 For problem 11-20

(5) During the reverse stroke (line EA), the piston expels the remaining steam at a constant pressure p_0 and the volume drops from V_2 to V_0. Determine the work done by the engine per cycle, supposing that $V_0 = 0.5$ L, $V_1 = 1.5$ L, $V_2 = 3.0$ L, $p_0 = 1$ atm, $p_1 = 12$ atm and $\gamma = 1.33$ (using the result of problem 11-8 to calculate W_Q).

11-21 The Otto cycle of a gasoline four-stroke engine is shown in Fig. 11-16 (section 11-6), prove that the thermal efficiency of Otto cycle can be written as

$$\eta = 1 - \frac{1}{(V_1/V_2)^{\gamma-1}}$$

in which V_1/V_2 is called compression ratio.

11-22 The cycle of a four—stroke Diesel engine is shown in Fig. 11-31:

(1) the air is shucked into the cylinder (line AB) at $p_0 = 1$ atm;

(2) the air is compressed adiabatically from volume V_1 to V_2 (line BC);

(3) at the end of the compression stroke, fuel is injected into the cylinder and ignited in the hot air, then burns and expands first isobarically to V_1' (line CD) then adiabatically to V_1 (line DE);

(4) at the end of the expansion, the exhaust valve opens, the pressure drops to p_0 (line EB) isochorically;

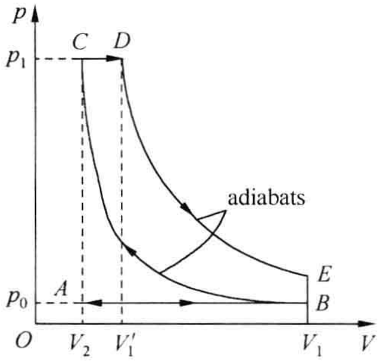


Fig. 11-31 For problem 11-22

(5) when the piston moves to the left the exhausted mixture is expelled (line BA). Prove that, thermal efficiency of the Diesel cycle is

$$\eta = 1 - \frac{\left(\frac{V_1'}{V_2}\right)^{\gamma} - 1}{\gamma \left(\frac{V_1}{V_2}\right)^{\gamma-1} \left(\frac{V_1'}{V_2} - 1\right)}$$

11-23 Show that

(1) an isotherm and an adiabatic can not intersect twice;

(2) two adiabats can not intersect with each other.

11-24 If there are 6 molecules in the left half of the box in Fig. 11-23(a), then the barrier is removed. How many possible macro-states are there? Calculate the thermodynamic possibility that is the number of microstates corresponding to a macro-state, for all possible macro-states. Refer to Table 11-2.

11-25 One mole of water is evaporated at temperature 100℃ and pressure 1atm, absorbing 4.06×10^4 J of heat. Find the change in entropy.

11-26 Find the change in entropy during the transition of 8 g of oxygen from a volume of 10 liters at a temperature of 80℃ to a volume of 40 liters at a temperature of 300℃.

11-27 Find the change in entropy during an isothermal expansion of 6 g of hydrogen from 1×10^5 to 0.5×10^5 N/m^2.

Part Four

Oscillation and Waves

Chapter12

Simple Harmonic Motion

We are surrounded by the motions which repeat themselves. In this chapter, we will study a special kind of these motions-the mechanical vibration. There are many familiar examples of oscillations: small boats bob up and down, clock pendulums swing back and forth, and the strings and reeds of musical instruments vibrate, the buildings in earthquakes vibrate. Other, less familiar examples are acoustic vibration, circuit oscillation, plasma vibration, molecular vibration, etc.. The most recognizable characteristic of oscillatory motion is that the motion is periodic, that is, it repeats itself. Thus it is important to study the vibrations in detail.

When a body is vibrating, the force acted on it is not constant, but varies during the motion. This force is called as the restoring force which causes the body to undergo a back and forth motion past the equilibrium position. The ideal vibration is periodic motion and the oscillations in the real world are usually damped, which dies out gradually. A common and very important kind of oscillatory motion is simple harmonic motion.

The vibrations are important not only because it is common, but because it is a basic constituent of wave motion. It will be seen that light and indeed the whole electromagnetic spectrum from long radio wave to short gamma rays, is an oscillating motion of electric and magnetic influence. So a study of vibration will lay the foundation for further work in many different fields of physics.

12. 1 Simple Harmonic Motion

12. 1. 1 Simple harmonic motion

The three systems shown in Fig. 12-1:

(a) A simple pendulum which consists of a mass m swinging at the end of an inextendable massless string of length L;

(b) A flat disc supported by a rigid wire through its center and oscillating through small angles in the plane of its circumference;

(c) A mass attached to a spring of force constant k fixed to a wall by one end, sliding back and forth in the x direction on a frictionless plane;

are simple harmonic oscillators if the frictions are ignored. When slightly disturbed from their equilibrium or rest position, they will oscillate with simple harmonic motion (abbreviated to

SHM). This is the most fundamental vibration of a single particle or one-dimensional system. All of these systems have their equilibrium position (represented by dashed line in Fig. 12-1), at which the external resultant force acted on the particle or one-dimensional system is zero. A small displacement x (or angular displacement) from its equilibrium position sets up as restoring force or torque that is proportional to x and acting in a direction towards the equilibrium position. The direction of the restoring force is such that the object is being either pulled or pushed toward the equilibrium position.

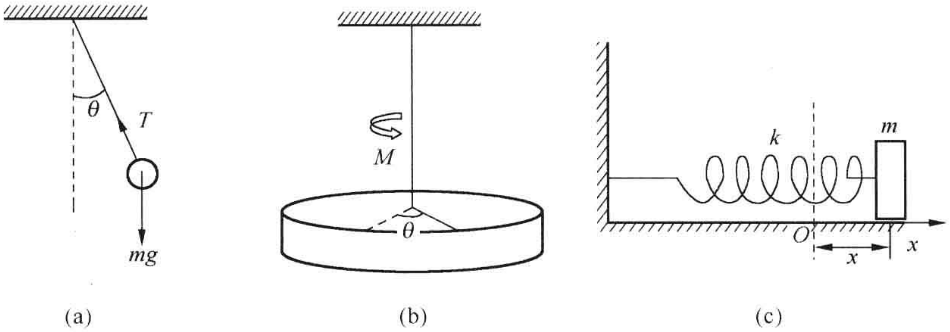


Fig. 12-1 The simple harmonic systems

The systems listed above are harmonic and have the same equation of motion. Hence, we take the system (c) as an example to study simple harmonic motion. In Fig. 12-1(c), a body of mass m is attached to a spring of force constant k (also called spring constant or stiffness) that obeys Hooke's law. Taking the equilibrium position as the origin point of x-axis, such a spring exerts on the body a restoring force that is proportional to x — the displacement of the body from the equilibrium position or the amount by which the spring is stretched (positive x) or compressed (negative x), but opposite in sign. The force exerting on the body is

$$F = -kx \tag{12-1}$$

where k is called the spring constant, the minus sign represents the restoring force in the opposite direction of the displacement.

Using Newton's second law, we have

$$F = -kx = ma = m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}$$

So

$$a = -\frac{k}{m}x = -\omega^2 x \tag{12-2}$$

where

$$\omega = \sqrt{\frac{k}{m}} \tag{12-3}$$

Eq. (12-2) can be rewritten as

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \omega^2 x = 0 \tag{12-4}$$

which has the complete or general solution

$$x = A\cos(\omega t + \varphi) \tag{12-5}$$

This equation is called as the equation of simple harmonic vibration. The constants A and φ

are determined by the values of x_0 and $v_0 = \mathrm{d}x/\mathrm{d}t$ at a specified time (usually, take the values at $t=0$).

We now can take Eq. (12-1) or Eq. (12-2) and Eq. (12-5) as the definition of the simple harmonic motion, which states:

> Simple harmonic motion is the motion executed by a particle of mass m subject to a force that is proportional to the displacement of the particle but opposite in sign;

or

> Simple harmonic motion is the motion executed by a particle of mass m which acceleration is proportional to the displacement of the particle but opposite in sign;

or

> Simple harmonic motion is the motion of a particle with the position function of Eq. (12-5).

It can be proved that the above three definitions of simple harmonic motion are equivalent.

Example 12-1 Prove that the simple pendulum as shown in Fig. 12-1(a), when the θ is small ($\theta < 5°$), is a simple harmonic oscillator.

Prove From Fig. 12-1(a), we obtain

$$F_t = -mg\sin\theta$$

The minus sign means F_t in the opposite direction of θ. Because $\theta < 5°$, $\sin\theta \approx \theta$ (rad), so that F_t is given by

$$F_t \approx -mg\theta = -\left(\frac{mg}{L}\right) \times L\theta = -kx$$

where

$$k = \frac{mg}{L} \text{ and } x = L\theta$$

Here x can be understood as the horizontal displacement of the particle. Therefore, a simple pendulum is formally like the simple harmonic oscillator of Fig. 12-1(c), and the effective spring constant k of the gravitational spring is mg/L. That is when θ is small, the motion of the simple pendulum is SHM. The motion of a simple pendulum is SHM only if the amplitude of the motion is small. Comparing with Eq. (12-3), the angular frequency and the period of the motion are

$$\omega = \sqrt{\frac{g}{L}}, \quad T = 2\pi\sqrt{\frac{L}{g}}$$

which imply the longer the length of a simple pendulum, the greater the period.

Example 12-2 The torsion pendulum as shown in Fig. 12-1(b). Prove when the disk is twisted through a small angle θ, the rotation of disk is simple harmonic motion (or a angular simple harmonic motion).

Prove If $\theta < 5°$, the restoring torque M is directly proportional to the angular displacement θ but opposite in direction, that is

$$M = -k\theta$$

where k is a constant which depends on the material from which the thin rod is made. Using the Newton's second law for rotation, the angular acceleration of the disk is given by

$$\beta = \frac{M}{I} = -\frac{k}{I}\theta = -\omega^2\theta$$

Here $\omega = \sqrt{k/I}$, I is the rotational inertia of the disk. The above equation shows that the angular acceleration is proportional to the angular displacement but opposite in sign. By the above method of analogies, we have

$$\theta = \theta_0 \cos(\omega t + \varphi)$$

Therefore, when θ is small, the motion of a torsion pendulum is SHM.

Example 12-3 The physical pendulum. As shown in Fig. 12-2, a real pendulum (usually called a physical pendulum such as grandfather clock), is a rigid body and can have a complicated distribution of mass. When the rigid body is suspended from some point O other than its center of mass C and displaced from its equilibrium position, it will oscillate. Prove that if the physical pendulum is displaced slightly from its equilibrium position, its motion is SHM.

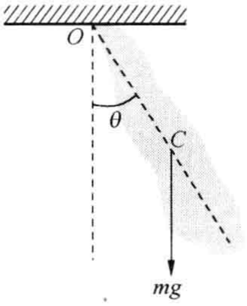


Fig. 12-2 For Example 12-3

Prove The gravitation force mg acts at its center of mass C at a distance L from the pivot point O. The torque about the point O has a magnitude

$$M = -mgL\sin\theta$$

and tends to decrease θ. Hence, we have

$$-mgL\sin\theta = I\frac{d^2\theta}{dt^2}$$

where I is the rotational inertia of the pendulum about the point O, depending on the shape of the physical pendulum. If the angular displacements are small enough so that the approximation $\sin\theta \approx \theta$ holds, we have

$$\frac{d^2\theta}{dt^2} \approx -\frac{mgL}{I}\theta \rightarrow \frac{d^2\theta}{dt^2} + \omega^2\theta = 0$$

where $\omega = \sqrt{mgL/I}$. Therefore, the angular acceleration is proportional to the displacement and opposite in sign. According to the definition of SHM, the motion of the physical pendulum is SHM. The period is

$$T = \frac{2\pi}{\omega} = 2\pi\sqrt{\frac{I}{mgL}}$$

12.1.2 Velocity and acceleration in SHM

From Eq. (12-5), the values of the velocity and acceleration in SHM are given by

$$v = \frac{dx}{dt} = -A\omega\sin(\omega t + \varphi) \tag{12-6}$$

and

$$a = \frac{dv}{dt} = -A\omega^2\cos(\omega t + \varphi) = -\omega^2 x \qquad (12\text{-}7)$$

The maximal value of velocity $A\omega$ is called the velocity amplitude and the acceleration amplitude is $A\omega^2$.

Fig. 12-3 shows the curves of displacement, velocity and acceleration with time for an example of simple harmonic motion $x = 2\cos 2t$ (SI). From Fig. 12-3, it is not difficult to see that when the magnitude of displacement is greatest, which occurs at the limits of the motion, the velocity is least, being zero and on the other hand, when the displacement is least, being zero at the equilibrium position of the motion, the velocity has its greatest magnitude. When the displacement has its greatest positive value, the acceleration has its greatest negative value and conversely, and the displacement is zero (that is $F=0$), the acceleration is also zero.

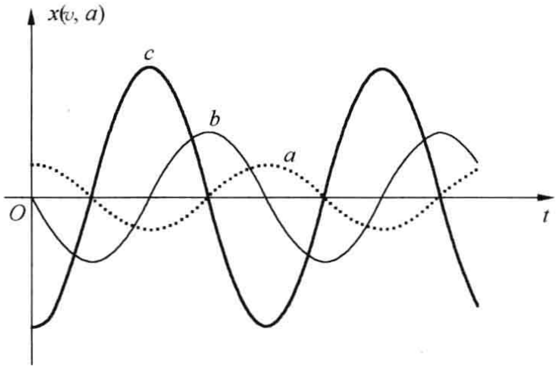


Fig. 12-3 a: Displacement; b: velocity; c: acceleration for SHM

12.1.3 Vector representation of the simple harmonic motion

In Eq. (12-5), $\omega t + \varphi$ is called as phase angle. It can be seen that the phase angle $\omega t + \varphi$ increases uniformly with time as the vibration takes place. The displacement at any moment t is proportional to the cosine of this angle. In Fig. 12-4, a vector A of length A rotates anticlockwise about a fixed axis through the origin with an angular velocity ω. The vector should make an angle φ with the x-axis at time $t = 0$. Therefore, the vector is located at the angle $\omega t + \varphi$ at time t. The point P is the projection of the end of the vector on the x-axis. It is obvious that the projection position x is given by

$$x = A\cos(\omega t + \varphi)$$

So, if a particle moves in uniform circular motion of radius A, its projection point on the x-axis will move in simple harmonic motion. Hence, the simple harmonic motion may be represented in magnitude and phase by a constant length vector (A) rotating anticlockwise with a constant angular velocity ω.

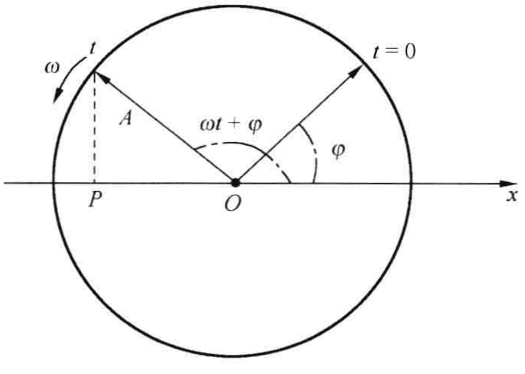


Fig. 12-4 A rotating vector **A**

According to the role to construct the rotating vector, a simple harmonic motion corresponds to a unique rotating vector. To know its corresponding rotating vector is to imply that the simple harmonic motion is determined completely. In the following, it will be seen that the rotating vector is very useful to determine the phase of a SHM and the phase difference of two SHM as well as to find the superposition of two or more simple harmonic motions by the addition of vectors.

Example 12-4 A particle undergoes a SHM with $A = 2.0$ cm and $\omega = \pi$. It passes the position of $x = \frac{\sqrt{2}}{2}A$ and toward the origin at $t = 0$. Draw the rotating vector to determine the phase constant at $t = 0$ and write the equation of this SHM.

Solution The rotating vector is shown in Fig. 12-5. When its projection is at $x = \frac{\sqrt{2}}{2}A$, the rotating vector could be at either the first quadrant or the fourth quadrant. Considering the vector rotates anticlockwise and the direction of the particle toward the origin when $t = 0$, we can choose easily the phase constant at $t = 0$ as $\varphi = \pi/4$. Therefore, the equation of this SHM is

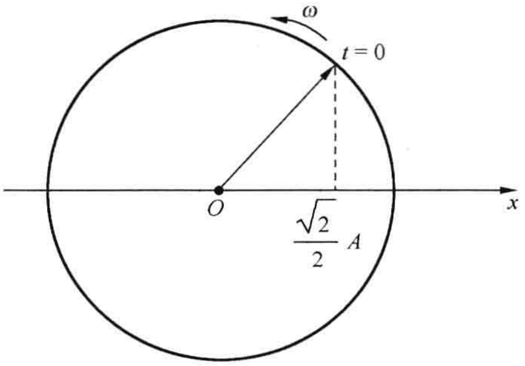


Fig. 12-5 For Example 12-4

$$x = 0.02\cos\left(\pi t + \frac{\pi}{4}\right) \text{ cm}$$

12.2 Period Frequency Amplitude and Phase Angle of SHM

There are three constants (A, ω, φ) in $x = A\cos(\omega t + \varphi)$. If A, ω and φ are known, the motion of the oscillator is determined completely. In this section, we will discuss these constants and their meanings.

12.2.1 The period and frequency

The function $A\cos(\omega t + \varphi)$ is periodic, and this fact implies that the particle of simple harmonic oscillator moves regularly and repeats with a characteristic time called as the period T. That is the time required for one complete oscillation of the particle or the system. According to $\cos(\theta + 2\pi) = \cos\theta$, if we increase the time by T in $x = A\cos(\omega t + \varphi)$, we must get the same displacement and the same velocity. Hence

$$T = \frac{2\pi}{\omega} \tag{12-8}$$

The quantity ω is called the angular frequency of the motion and has a relation with the period T

$$\omega = \frac{2\pi}{T} \tag{12-9}$$

The frequency of the motion is defined as the number of complete oscillations finished by the particle or the system each second. By definition, we have

$$\nu = \frac{1}{T} = \frac{\omega}{2\pi} \tag{12-10}$$

which unit is the hertz in SI (abbr. Hz)

For the spring oscillator as shown in Fig. 12-1(c), the above three quantities are given by

$$\omega = \sqrt{\frac{k}{m}}, \quad T = 2\pi\sqrt{\frac{m}{k}}, \quad \nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{12-11}$$

They are dependent only on the k and m that are the characteristic quantities of the system, that is why we usually call ω as intrinsic angular frequency and so as T and ν. For the two simple harmonic oscillators having the same k and m, their vibrations will be same in ω, T and ν but different in A and φ.

For simple pendulum, ω, T and ν are given by

$$\omega = \sqrt{\frac{g}{L}}, \quad T = 2\pi\sqrt{\frac{L}{g}}, \quad \nu = \frac{1}{2\pi}\sqrt{\frac{g}{L}} \tag{12-12}$$

It can be seen that the element of inertia seems to be missing and the period is independent of the mass of the particle. This is because the restoring force is proportional to the mass and the acceleration $a = F/m$ is therefore independent of the mass (that is equivalence of inertial and gravitational masses). The period T is decided by length of the string. The longer the length of the pendulum is, the greater its period. If the pendulum length is given, the period of this pendulum is constant independent of the mass of the particle. Note also that the frequency and period are independent of the amplitude of oscillation, a general feature of simple harmonic motion. However, when the amplitude of vibration of the simple pendulum is not small, its motion is periodic, but it is not simple harmonic. In particular, the period does have a slight dependence on the amplitude.

We can use a simple pendulum to make measurements of the free-fall acceleration$g$. From Eq. (12-12), we have

$$g = \frac{4\pi^2 L}{T^2} \tag{12-13}$$

for the acceleration of gravity in terms of the measured period T and length L of the pendulum. In Newton's day, the pendulum was used to obtain values for g at different latitude over the surface of the earth. For a simple pendulum. T can be determined with accuracy, but it is not possible to obtain an equally accurate value of L. This in turn limits the accuracy of the values of g obtained by using a simple pendulum. In order to gain the very precise values of g at a particular location, other method must be used.

12.2.2 The amplitude and the phase

The quantity A in $x = A\cos(\omega t + \varphi)$ is called the amplitude which represents the maximal value of the displacement x. Thus the displacement x varies between the limits $\pm A$. In the next section, it will be seen that the greater the amplitude A, the greater the total energy.

The time-varying quantity $\omega t + \varphi$ in $x = A\cos(\omega t + \varphi)$ is called the phase of the motion and the constant φ is called the phase angle (or phase constant). According to Eq. (12-5), Eq. (12-6) and Eq. (12-7), the phase $\omega t + \varphi$ determines displacement, velocity and acceleration (also kinetic energy and potential energy) of the particle at the time t.

For a given simple harmonic oscillator, it can be proved that amplitude A and phase constant are determined by the initial conditions x_0 and the velocity v_0 at $t=0$. Letting $t=0$ in Eq. (12-5) and Eq. (12-6), we have

$$x_0 = A\cos\varphi \tag{12-14}$$

$$v_0 = -A\omega\sin\varphi \tag{12-15}$$

from which it can be obtained that

$$A = \sqrt{x_0^2 + \frac{v_0^2}{\omega^2}} \tag{12-16}$$

$$\varphi = \arctan\left(-\frac{v_0}{x_0\omega}\right) \tag{12-17}$$

From Eq. (12-17), we can obtain two solutions φ_1 and φ_2 in general. How do we choose the correct one between them? According to Eq. (12-14) and Eq. (12-15), we have

$$\cos\varphi = \frac{x_0}{A} \tag{12-18}$$

$$\sin\varphi = -\frac{v_0}{A\omega} \tag{12-19}$$

The one between φ_1 and φ_2, satisfied by Eq. (12-18) and Eq. (12-19), is the correct phase constant.

Example 12-5 A cubic board floats in the water. When it is at rest, the immersing height of board is a. Slowly Pull it down and lead to an immersing height of b as shown in Fig. 12-6.

(1) Try to prove that if the friction of water is ignored, its motion is simple harmonic vibration after it is released.

(2) Find period, frequency, amplitude and initial phase.

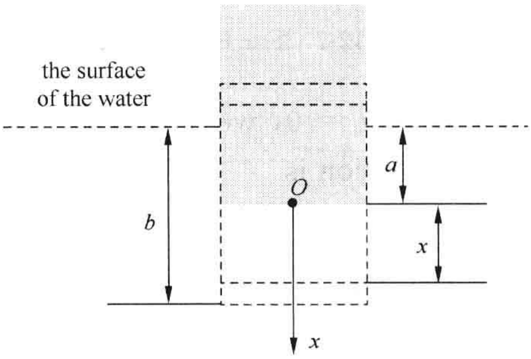


Fig. 12-6 For Example 12-5

Solution (1) When the board is at rest (equilibrium position), the gravity and the buoyancy acted on it have same value but opposite in direction:

$$Sa\rho = mg$$

where S is the cross-section area of the board, m is the mass of board and ρ the density of water. Therefore

$$m = \frac{Sa\rho}{g}$$

As shown in Fig. 12-6, set the positive direction of the coordinate axis x downward and choose the position of the down surface of the board at rest as the origin of axis. When the immersing height is $a+x$, the resultant force is given by

$$f_x = mg - S(a+x)\rho$$

Using $Sa\rho = mg$, we have

$$f_x = -S\rho x$$

so that

$$a = -\frac{S\rho}{m}x = -\omega^2 x$$

which implies that the resultant force acted on the board is proportional to the displacement of the board but opposite in sign. Hence, its motion is simple harmonic vibration.

(2) The period and frequency are expressed as

$$T = 2\pi\sqrt{\frac{m}{S\rho}} = 2\pi\sqrt{\frac{a}{g}}, \quad \nu = \frac{1}{2\pi}\sqrt{\frac{g}{a}}$$

The initial displacement is $x_0 = b - a$ and the initial velocity $v_0 = 0$. According to Eq. (12-16) and Eq. (12-17), we have $A = b - a$ and $\varphi = 0$.

Example 12-6 A particle moves with simple harmonic motion of period 6s, which curve of motion is shown in Fig. 12-7. Find

(1) the equation of motion; (2) the velocity of the particle at $t = 0$.

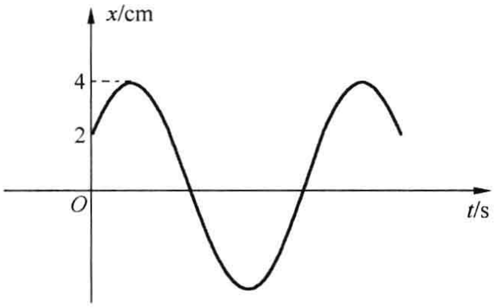


Fig. 12-7 For Example 12-6

Solution (1) The angular frequency is $\omega = \pi/3$ and $A = 4$ cm. From Fig. 12-7, we have $x_0 = 2$ cm and $v_0 > 0$ when $t = 0$. According to Eq. (12-18) and Eq. (12-19), we obtain

$$\cos\varphi = \frac{1}{2}$$

which has two solutions $\varphi = \pi/3$ and $-\pi/3$. Because $v_0 > 0$ when $t = 0$, we have $\sin\varphi < 0$. Hence, the correct initial phase is $-\pi/3$, and the equation of motion is

$$x = 0.04\cos\left(\frac{\pi t}{3} - \frac{\pi}{3}\right) \text{ m}$$

(2) The velocity of the particle at $t = 0$ s is given by

$$v_0 = -A\omega\sin\left(-\frac{\pi}{3}\right) = 0.036 \text{ m/s}$$

12.3 Kinetic and Potential Energies in SHM

In Fig. 12-1(c), if the particle and the spring are looked as a system, no external forces act on it, and then the total energy remains constant. That is the energy is conservative

$$E = E_k + E_p = \frac{1}{2}mv^2 + \frac{1}{2}kx^2 = \text{constant}$$

Using Eq. (12-6), Eq. (12-5) and Eq. (12-3), we have

$$E_k = \frac{1}{2}mv^2 = \frac{1}{2}m\omega^2A^2\sin^2(\omega t + \varphi) = \frac{1}{2}kA^2\sin^2(\omega t + \varphi) \tag{12-20}$$

$$E_p = \frac{1}{2}kx^2 = \frac{1}{2}kA^2\cos^2(\omega t + \varphi) = \frac{1}{2}m\omega^2A^2\cos^2(\omega t + \varphi) \tag{12-21}$$

According to the relation

$$\sin^2\alpha + \cos^2\alpha = 1$$

the total energy is given by

$$E = E_k + E_p = \frac{1}{2}kA^2 = \frac{1}{2}m\omega^2A^2 \tag{12-22}$$

which is indeed a constant, independent of the time.

Fig. 12-8 shows the kinetic energy and potential energy of the system as a function of time for the case $x = A\cos\omega t$. Fig. 12-9 shows these two energies as a function of the displacement x of the particle. According to Eq. (12-22), the total energy of a harmonic oscillator is determined completely by the maximal displacement and the spring constant. When the displacement has greatest value (positive or negative), the potential energy that is stored in the spring has greatest value and this means that the kinetic energy, which is stored in the particle totally converts into the potential energy. When the particle moves from the limit points ($x = \pm A$) to the equilibrium position, the potential energy reduces, and the kinetic energy increases and has greatest value at the equilibrium position. So from an energy viewpoint, the simple harmonic vibration is a continuing transformation of kinetic and potential energy, one into the other, while the total energy doesn't change.

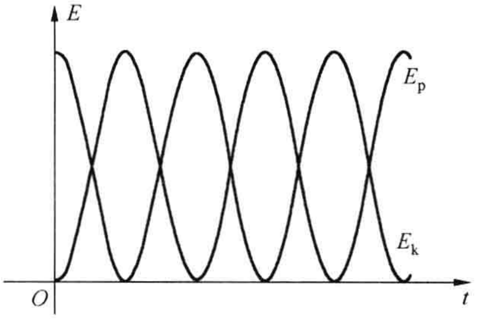


Fig. 12-8 E_k and E_p as a function of time t

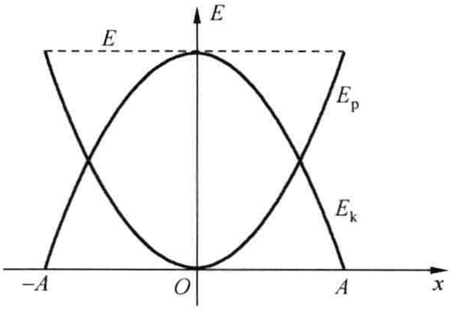


Fig. 12-9 E_k and E_p as a function of displacement x

Now we can understand why an oscillating system must contain an element of springness and an element of inertia. It needs the former to store its potential energy and the latter to store its kinetic energy. Without either element, the interchange of energy between these two forms would not be possible, and vibration can not take place.

Example 12-7 Find the positions at which the potential energy of the oscillating system is equal to its kinetic energy.

Solution From Eq. (12-22), considering $E_k = E_p$, we have

$$E = \frac{1}{2}mv^2 + \frac{1}{2}kx^2 = 2 \times \frac{1}{2}kx^2 = \frac{1}{2}kA^2$$

Hence when

$$x = \pm\frac{\sqrt{2}}{2}A$$

the potential energy is equal to the kinetic energy for a harmonic oscillator.

12.4 Damped Vibration, Forced Vibration, and Resonance

12.4.1 Damped vibration

There is friction in any real physical processes. So far, we have neglected the effect of the friction in oscillating motion. Obviously, because of the friction, any kinds of oscillations will gradually die out and finally, the particle will be at rest. Due to the frictional forces, the mechanical energy of oscillatory motion decreases with time and the motion is said to be damped. We now consider this kind of "decaying", which is called damped harmonic motion.

The friction in a vibrating system comes from its supports, the surface and the surrounding circumstances. If you try to swing a pendulum under water, because of the friction of the water, you will not have much success. This is called the overdamped harmonic motion. In air, you will do better but the oscillation will gradually die out, which is called as the underdamped harmonic motion. In a vacuum, the pendulum will swing for a long time but eventually it will also come to rest, by the action of the friction at its supports.

In the following, we will limit out consideration of the friction acted on by the surrounding medium. Experiment shows that for small velocity, the viscous frictional force exerted on a body is proportional to the magnitude of the velocity of the oscillating body but opposite to its direction. Such this force can be represented by the equation

$$f = -\gamma v = -\gamma \frac{\mathrm{d}x}{\mathrm{d}t} \tag{12-23}$$

where γ is an empirically determined constant. The total force acting on the body of simple harmonic oscillator is

$$F = -kx - \gamma \frac{\mathrm{d}x}{\mathrm{d}t} \tag{12-24}$$

Applying Newton's second law, we have

$$m \frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + \gamma \frac{\mathrm{d}x}{\mathrm{d}t} + kx = 0 \tag{12-25}$$

The solutions of this equation describe the damped harmonic motion. There are three different types of damped harmonic motion. Here, we define $\omega_0 = \sqrt{k/m}$ as the intrinsic frequency of the system and $\beta = \gamma/2m$ as the damping coefficient.

For the small values of γ (that is $\beta < \omega_0$), the frictional force on a harmonic oscillator will be small. Eq. (12-25) has the solution

$$x = A\exp(-\beta t)\cos(\omega t + \varphi) \tag{12-26}$$

$$\omega = \sqrt{\omega_0^2 - \beta^2}$$

The system should oscillate with a steady decreasing amplitude that is $A\exp(-\beta t)$. This motion illustrated in Fig. 12-10 for $\varphi = 0$ and $A=1$ is called the underdamped harmonic motion. Although the motion does not repeat itself as an undamped vibration does, it is useful to introduce the period T

$$T = \frac{2\pi}{\omega}$$

to describe as " a cycle " the events (such as passing the equilibrium position) that take place while the phase angle $\omega t + \varphi$ increases by 2π. The total mechanical energy of damped vibration is not constant but decreases with time because of the friction, appearing as thermal energy in the surrounding medium. Finally the vibration will die out and the system or the particle comes to rest.

For large values of γ, the frictional force will nearly balance the restoring force of spring as the body returns to its equilibrium position. As Eq. (12-25) indicates, this implies that the acceleration will be small throughout the motion. If displaced, we expect the body to creep back to the equilibrium position. In other words, no oscillation will occur. This motion, as shown in Fig. 12-11, is called the overdamped harmonic motion.

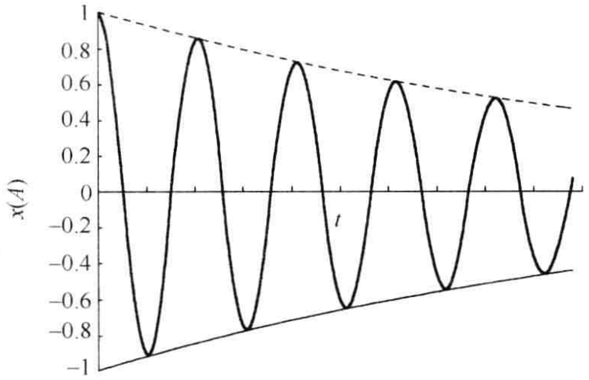


Fig. 12-10 The underdamped harmonic motion

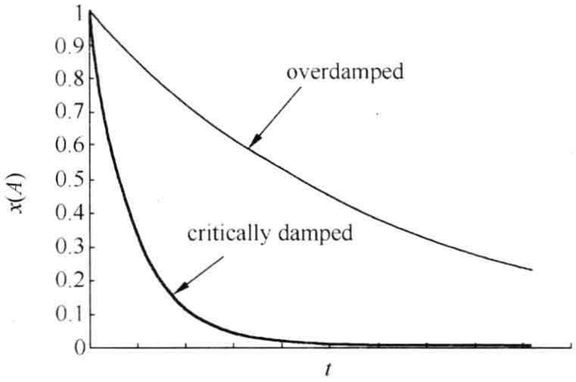


Fig. 12-11 The critically damped and overdamped motions with $A=1$

If γ is too large, the underdamped harmonic motion becomes impossible. That is the body cannot come back to the equilibrium position. Similarly, if γ is small, the overdamped harmonic motion becomes impossible. There is a critical value that separates these two types of motion. This critical value is

$$\gamma_0 = 2\sqrt{km} \tag{12-27}$$

With this value of γ, the system executes critically damped motion as shown in Fig. 12-11. x decays faster when the damped vibration is critical than when it is overdamped. This is a point of great importance in the design of any equipment containing parts that are capable vibrating.

12.4.2 Forced vibration and resonance

We now qualitatively study the behavior of a damped harmonic oscillator that is also subjected to a third force, which varies sinusoidally with time. We call such a system a forced oscillator. The external force is expressed by

$$F_{\mathrm{ex}} = F_0 \cos\omega_{\mathrm{d}} t \tag{12-28}$$

which is called driving force. There are now two angular frequencies with which to deal:

(1) the "intrinsic" angular frequency of the free oscillator of system, which we have called the natural angular frequency ω;

(2) the angular frequency of the external driving force, which we call ω_{d} simply.

The motion of a forced harmonic oscillator is governed by the equation

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \gamma\frac{\mathrm{d}x}{\mathrm{d}t} + kx = F_0\cos\omega_{\mathrm{d}}t \tag{12-29}$$

This differential equation has a general solution consisting of two parts. The transient part of the solution involves the underdamped, critically damped, or overdamped motions, depending on the value of γ. The other part of the general solution, called the steady-state solution, is our concern here. Any transient motions initially present will die out exponentially with time, but the steady-state motion will persist. After a sufficiently long time, the motion will be steady state. We assume here that the system executes steady-state motion without the presence of transients. Therefore, the displacement of a forced oscillator is given by

$$x = A(\omega_d, \omega, \beta, F_0)\cos(\omega_d t + \varphi) \tag{12-30}$$

where A is determined by ω_d, ω, β and F_0, that is

$$A = \frac{F_0}{m\sqrt{(\omega^2 - \omega_d^2)^2 + 4\beta^2 \omega_d}} \tag{12-31}$$

Much of our interest in the forced oscillator relates to the "response" of the system. By response we mean the dependence of A on ω_d.

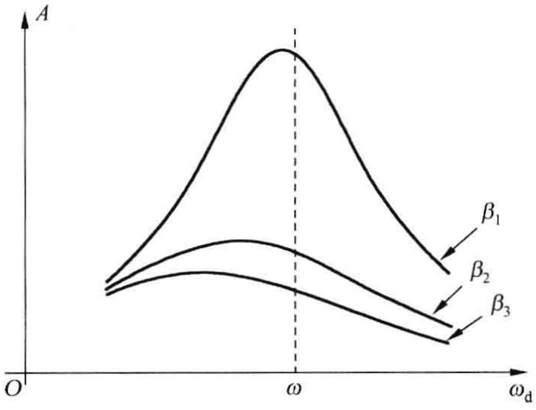


Fig. 12-12 The resonances: $\beta_1 < \beta_2 < \beta_3$

Figure 12-12 shows the amplitude A of a forced oscillator as the angular frequency ω_d of the driving force is varied. It is easy to prove that when the angular frequency of the driving force is equal to ω_r

$$\omega_r = \sqrt{\omega^2 - 2\beta^2} \tag{12-32}$$

A has a greatest value

$$A = \frac{F_0}{2m\beta\sqrt{\omega^2 - \beta^2}} \tag{12-33}$$

This phenomenon is called resonance. Fig. 12-12 shows the resonance phenomenon for three different levels of damping, that is, for three different values of the damping coefficient. The smaller the damping, the taller and the narrower the resonance peak.

The resonance is important in our daily lives. There are many familiar examples. When soldiers march across a small bridge, they often break step because it is possible that the frequency of their step might be near one of the resonance frequencies of the bridge. When we design machinery, we must consider the resonance encountered in normal use and avoid the resonance. Sometimes, we expect that the resonance become more pronounced such as in the laser.

12.5 Superposition of Two SHM with Same Direction of Vibration and Same Frequency

In this section, we will handle two or more harmonic vibrations with the same frequency and the same direction of vibration. In the real world, there are many phenomena in which a particle takes part in two or more simple harmonic motions. In the following chapters, we shall meet physical situations that involve the superposition of two or more simple harmonic vibrations on the same system. For simplicity, we only discuss the superposition of two simple harmonic motions with the same frequency in the same direction of vibration.

What happens when two harmonic driving forces

$$F_1 = F_{10}\cos(\omega t + \alpha_1)$$

$$F_2 = F_{20}\cos(\omega t + \alpha_2)$$

act simultaneously on the same simple harmonic oscillator? If F_2 were zero, we could immediately write down the steady-state solution

$$x_1 = A_1\cos(\omega t + \varphi_1)$$

Similarly if F_1 were zero, we could have

$$x_2 = A_2\cos(\omega t + \varphi_2)$$

When F_1 and F_2 exist simultaneously, it is a simple matter to verify that the displacement of the system is written as

$$x = x_1 + x_2 = A_1\cos(\omega t + \varphi_1) + A_2\cos(\omega t + \varphi_2) \tag{12-34}$$

which is the steady-state solution of the system when both of the amplitudes of the driving forces are finite. That is, if one driving force alone produces the displacement x_1 and a second driving force alone produces the displacement x_2, then forces acting together will produce the displacement $x_1 + x_2$. Clearly, the argument can be extended to any number of driving forces.

The particle takes part in the two simple harmonic motions x_1 and x_2. They have the same frequency and vibrate in the same direction. Now we deal with the superposition of the two simple harmonic motions. Because x_1 and x_2 are simple harmonic motion with the same frequency, it can be verified that the $x = x_1 + x_2$ is also harmonic. That is

$$\begin{aligned} x_2 &= A_1\cos(\omega t + \varphi_1) + A_2\cos(\omega t + \varphi_2) \\ &= A\cos(\omega t + \varphi) \end{aligned} \tag{12-35}$$

In Fig. 12-13, x_1 is represented by the rotating vector $\mathbf{A}_1$, and x_2 by the rotating vector $\mathbf{A}_2$. If x is represented by the rotating vector $\mathbf{A}$, it can be found that

$$\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2 \tag{12-36}$$

The projected position x of $\mathbf{A}$ on the x-axis is the sum of the projected position x_1 of $\mathbf{A}_1$ and x_2 of $\mathbf{A}_2$ on the x-axis. From the addition of vectors, the amplitude A and the initial phase are given by

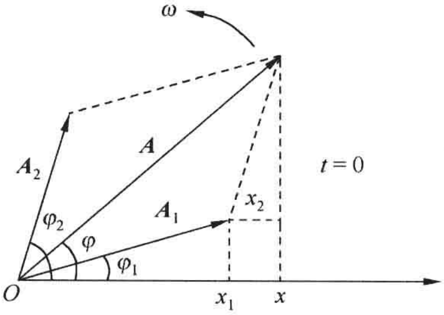


Fig. 12-13 The addition of vectors

$$A = \sqrt{A_1^2 + A_2^2 + 2A_1A_2\cos(\varphi_2 - \varphi_1)} \tag{12-37}$$

$$\varphi = \arctan\frac{A_1\sin\varphi_1 + A_2\sin\varphi_2}{A_1\cos\varphi_1 + A_2\cos\varphi_2} \tag{12-38}$$

From Eq. (12-37), we can see the influence of the phase difference $(\varphi_2 - \varphi_1)$ on the amplitude of the resultant vibration. Depending on the value of $\cos(\varphi_2 - \varphi_1)$ that can be any values from -1 to $+1$, A can have one between $|A_1 - A_2|$ and $|A_1 + A_2|$. If A comes out smaller than the larger A_1 and A_2, we say that the superposition is destructive. If A comes out larger than both, we say that the superposition is constructive.

If $\varphi_2 - \varphi_1 = 2k\pi$, $k = 0, \pm 1, \pm 2, \cdots$, A has the maximal amplitude

$$A = A_1 + A_2 \tag{12-39}$$

which is called the full constructive superposition. If $\varphi_2 - \varphi_1 = (2k+1)\pi$, $k = 0, \pm 1, \pm 2, \cdots$, A takes minimal value

$$A = |A_1 - A_2| \tag{12-40}$$

which is called the full destructive superposition. If $A_1 = A_2 = a$, we have

$$A = a\sqrt{2 + 2\cos(\varphi_2 - \varphi_1)} \tag{12-41}$$

Example 12-8 The two harmonic vibrations are $x_1 = 0.03\cos(5t + \pi/2)$ (SI) and $x_2 = 0.04\cos 5t$ (SI). Find the resultant vibration.

Solution According to Eq. (12-37) and Eq. (12-38), A and φ of the resultant vibration are given by

$$A = \sqrt{0.03^2 + 0.04^2 + 2 \times 0.03 \times 0.04 \times \cos\frac{\pi}{2}} = 0.05 \text{ (SI)}$$

$$\varphi = \arctan\frac{0.03}{0.04} = 0.64 \text{ rad}$$

Hence

$$x = x_1 + x_2 = 0.05\cos(5t + 0.64) \text{ (SI)}$$

12.6 Superposition of Two SHM with Same Frequency and Directions of Vibration Perpendicular to Each Other

Now we are going to consider the case in which a particle takes apart in two harmonic vibrations with the same frequency in the directions of vibration perpendicular to each other. Assume these two harmonic vibrations lie in the x-axis and the y-axis respectively and their vibrating equations are

$$x = A_1\cos(\omega t + \varphi_1)$$
$$y = A_2\cos(\omega t + \varphi_2)$$

The position vector $\boldsymbol{r}$ of the particle at the time t that takes part in these two harmonic vibrations is given by

$$\boldsymbol{r} = x\boldsymbol{i} + y\boldsymbol{j} = A_1\cos(\omega t + \varphi_1)\boldsymbol{i} + A_2\cos(\omega t + \varphi_2)\boldsymbol{j}$$

Eliminating time t from above equation, we obtain the equation of the trajectory of particle in the x-y plane

$$\frac{x^2}{A_1^2} + \frac{y^2}{A_2^2} - \frac{2xy}{A_1A_2}\cos(\varphi_2 - \varphi_1) = \sin^2(\varphi_2 - \varphi_1) \tag{12-42}$$

It is easy to see that the trajectory depends completely on the phase difference $\Delta\varphi = \varphi_2 - \varphi_1$. In general, the trajectory of particle is an ellipse in the x-y plane. In the following, several special cases will be discussed.

(1) $\Delta\varphi = \varphi_2 - \varphi_1 = 0$, which means two vibrations have the same initial phase. For this case, Eq. (12-42) becomes

$$\frac{x^2}{A_1^2} + \frac{y^2}{A_2^2} - \frac{2xy}{A_1A_2} = 0$$

which can be rewritten as

$$\left(\frac{x}{A_1} - \frac{y}{A_2}\right)^2 = 0$$

Hence, the trajectory of particle satisfies the below equation

$$y = \frac{A_2}{A_1}x \tag{12-43}$$

This implies that the trajectory of particle is a line passing through the origin with the slope $k = \tan\theta = A_2/A_1$ as shown in Fig. 12-14. Obviously, the particle vibrates along this line and its displacement s from the equilibrium position is given by

$$s = \pm\sqrt{x^2 + y^2} = \sqrt{A_1^2 + A_2^2}\cos(\omega t + \varphi) \tag{12-44}$$

The amplitude is $\sqrt{A_1^2 + A_2^2}$. Here, we take $\varphi_1 = \varphi_2 = \varphi$.

(2) $\Delta\varphi = \varphi_2 - \varphi_1 = \pi$ which means that two vibrations have the opposite phase. For this case, we have

$$\frac{x^2}{A_1^2} + \frac{y^2}{A_2^2} + \frac{2xy}{A_1 A_2} = 0$$

or

$$\left(\frac{x}{A_1} + \frac{y}{A_2}\right)^2 = 0$$

from which the trajectory equation of particle is obtained

$$y = -\frac{A_2}{A_1}x \tag{12-45}$$

Eq. (12-45) shows that the trajectory of particle is a line passing through the origin with the slope $k = \tan\theta = -A_2/A_1$ as shown in Fig. 12-14. The motion of particle is also the harmonic vibration along the line $y = -\frac{A_2}{A_1}x$ with the displacement s from the equilibrium position given by Eq. (12-44).

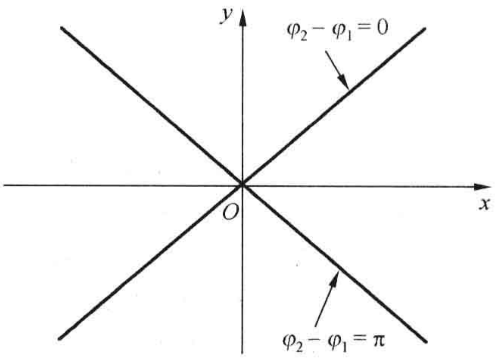
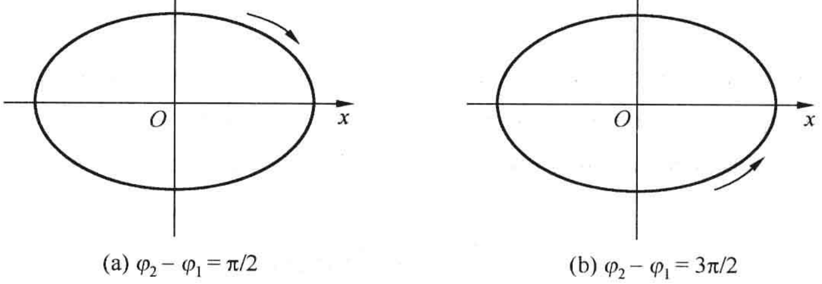


Fig. 12-14 The trajectory of particle

(3) $\Delta\varphi = \varphi_2 - \varphi_1 = \pi/2$. Eq. (12-42) becomes as

$$\frac{x^2}{A_1^2} + \frac{y^2}{A_2^2} = 1 \tag{12-46}$$

For this case, Eq. (12-46) implies that the particle moves along an ellipse which two axes are located on the x-axis and y-axis. It is easily proved that the motion of particle is clockwise as shown in Fig. 12-15(a). Especially, if $A_1 = A_2$, the trajectory of particle becomes a circumference.



(a) $\varphi_2 - \varphi_1 = \pi/2$

(b) $\varphi_2 - \varphi_1 = 3\pi/2$

Fig. 12-15 The trajectory of particle

(4) $\Delta\varphi = \varphi_2 - \varphi_1 = 3\pi/2$. The trajectory equation of particle is as same as Eq. (12-46).

For this case, the particle moves anticlockwise along an ellipse and the trajectory is presented in Fig. 12-15(b).

If the phase difference $\Delta\varphi = \varphi_2 - \varphi_1$, not being the above special values, is an arbitrary constant, the trajectory of particles is generally an ellipse with the axes that are not located in the x-axis and y-axis.

Questions

12-1 Give some examples of the motions that can be treated approximately as the simple harmonic motions. Try to explain why.

12-2 What does a simple oscillator consists of? What role does each part of a simple oscillator play in the simple harmonic motions?

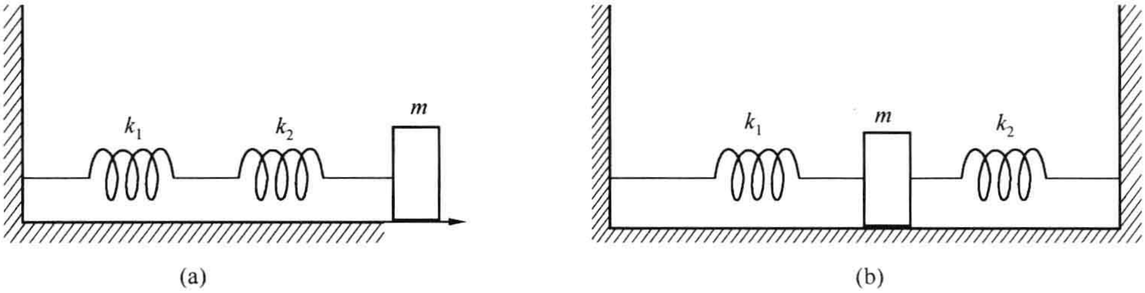


Fig. 12-16 For question 12-6

12-3 For simple pendulum, "who" provides the restoring force? Try to discuss the kinetic energy and potential energy of the simple pendulum.

12-4 What does the period of SHM depends on? What do its amplitude and initial phase depend on?

12-5 A spring, having a force constant k, is cut in half. What is the force constant of the divided half?

12-6 Systems (a) and (b) as shown in Fig. 12-16 can be simplified as simple oscillators. Try to find the equivalent force constant k for each one.

12-7 A spring has a force constant k, and a mass m is suspended from it. The spring is cut in half and the same mass is suspended from one of the halves. How are the frequencies of oscillation, before and after the spring is cut, related?

12-8 How is each of the following properties of a simple harmonic oscillator affected by doubling the amplitude: period, force constant, total mechanical energy, maximum velocity, and maximum acceleration?

12-9 Could you ever find or construct a true simple pendulum? Explain your answer.

12-10 What would happen to the motion of an oscillating system if the sign of the restoring force $-kx$ were changed?

12-11 If you have a pendulum clock, what do you have to do everyday to make it run continuously? Why?

12-12 In forced damped harmonic motion, is the damping ever useful?

12-13 Give some examples of common phenomena in which resonance plays an important role.

Problems

12-1 A particle undergoes simple harmonic motion of period 5s with the amplitude $A = 3.00 \times 10^{-2}$ m and initial phase $\varphi = \pi/4$. Write its equation of motion and draw the x-t, v-t and a-t curves.

12-2 A body of mass 0.100 kg moves with the simple harmonic motion of frequency 5 Hz. When $t = 0$, its displacement and velocity are 10.0 cm and 3.14 m/s respectively. Find

(1) amplitude and initial phase;

(2) maximum velocity, maximum acceleration and maximum restoring force;

(3) equation of motion.

12-3 A block of 4.0 kg extends a spring 16 cm from its unstretched position.

(1) What is the spring constant?

(2) The block is removed and a body of 0.50 kg is hung from the same spring. What is the period of this harmonic oscillator?

12-4 A body oscillates with simple harmonic motion according to the equation

$$x = 0.50\cos(\pi t + \pi/3) \text{ (SI)}$$

Find the frequency and period of the motion. What are the displacement, the velocity, the acceleration and the phase at the time $t = 2.0$ s.

12-5 In Fig. 12-17, several initial states of the simple pendulum with the angular frequency ω and the amplitude A are shown:

(a) θ has a maximum value θ_0 and $d\theta/dt = 0$;

(b) $\theta = 0$ and $d\theta/dt$ takes a negative maximum value;

(c) θ reaches a negative maximum value $-\theta_0$ and $d\theta/dt = 0$;

(d) $\theta = 0$ and $d\theta/dt$ is maximum.

Try to use the correspondent rotating vector to find the initial phase and the vibrating equation for each one.

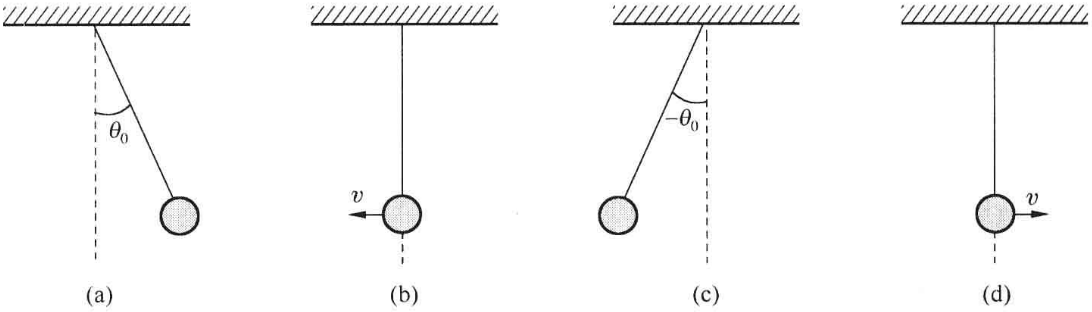


Fig. 12-17 For problem 12-5

12-6 A body of mass 50 g is attached to the bottom of a vertical spring and set vibrating. If its maximum speed is 15 cm/s and the period is 0.50 s, find

(1) the spring constant k of the spring, (2) the amplitude of the motion and (3) the frequency of oscillation.

12-7 A particle executes simple harmonic motion about the point $x = 0$. At $t = 0$, it has displacement $x_0 = 3.7$ cm and zero velocity. The frequency of the motion is 0.25 Hz. Determine (1) the period, (2) the equation of the motion and the maximum acceleration.

12-8 A particle moves in simple harmonic motion of the amplitude 0.120 m with the period 2 s. It has an initial displacement of 0.060 m and moves toward the positive direction of axis at $t = 0$. Find (1) the equation of motion, (2) the displacement, velocity and acceleration at $t = T/4$, and (3) the time needed to pass first through the equilibrium position from $t = 0$.

12-9 Two blocks ($m = 0.400$ kg and $M = 0.600$ kg) and a spring ($k = 25.0$ N/m) are arranged on a horizontal, frictionless surface as shown in Fig. 12-18. The coefficient of static friction between the two blocks is 0.50. What is the maximum possible amplitude of the simple harmonic motion if no slippage is to occur between the blocks?

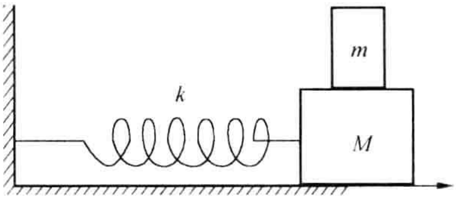


Fig. 12-18 For problem 12-9

12-10 Two bodies A and B have the same mass of 1. 0 kg and are attached to a vertical spring hung from the ceiling. Two bodies at rest cause the spring to stretch 9.8×10^{-2} m. If the body B dorp suddenly from this system as shown in Fig. 12-19, the body A start to vibrate ($t=0$). Find (1) the vibrating period of body B, (2) the amplitude and initial phase, and (3) its equation of motion.

12-11 A simple harmonic oscillator is composed of a block of mass 2. 50 kg and a spring of force constant 100 N/m. When $t=1.00$ s, the position and velocity of the block are $x=0.200$ m and $v=3.00$ m/s respectively. Find (1) the amplitude of the oscillation and (2) the position and velocity at $t=0$ s, and (3) the equation of motion.

12-12 A body of mass $m=3\times10^3$ kg is hung by a crane with a cable and descends at a speed $v=3$ m/s as shown in Fig. 12-20. The force constant of the cable is 27×10^5 N/m and its mass is negligible. The upper end of the cable is stopped suddenly by some reason. Try to prove that the motion of body is harmonic vibration after the cable is braked, and find the vibrating equation of the body.

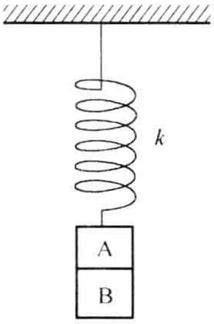


Fig. 12-19 For problem 12-10

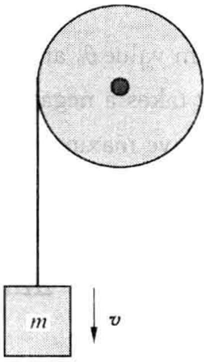


Fig. 12-20 For problem 12-12

12-13 A block of mass 4. 0 kg is suspended from a spring with a force constant $k=550$ N/m. A 50 g bullet is fired into the block from below with a speed of 200 m/s and comes to rest in the block. Find the period and the equation of the resulting simple harmonic motion.

12-14 A uniform spring whose unstressed length is L has a force constant k. The spring is cut into two pieces of unstressed lengths L_1 and L_2, with $L_1=nL_2$ (n is integer). What are the corresponding force constants k_1 and k_2 in terms of n and k? Does your result seem reasonable for $n=1$?

12-15 An oscillating block ($m=0.1$ kg)-spring system has an amplitude of 0. 01 m and a maximum acceleration of 0. 04 $\mathrm{m/s^2}$. Find (1) the period and (2) the total energy. At what position is the kinetic energy equal to the potential energy?

12-16 A block of mass M, at rest on a horizontal frictionless table, is attached to a rigid supported by a spring of force constant k. A bullet of mass m and velocity v strikes the block as shown in Fig. 12-21. The bullet remains embedded in the block. Determine (1) the vibrating period and (2) the total energy and amplitude of oscillation.

12-17 Show that the average values, over a complete cycle, of both the kinetic energy and the potential energy of a simple harmonic oscillator equal half the maximum value.

12-18 When the displacement is one-half the amplitude A, what fraction of the total energy is kinetic energy and what fraction is potential energy in the simple harmonic motion?

12-19 A body takes part in two simple harmonic motions given by

$$x_1=0.05\cos\left(10t+\frac{3\pi}{4}\right)$$

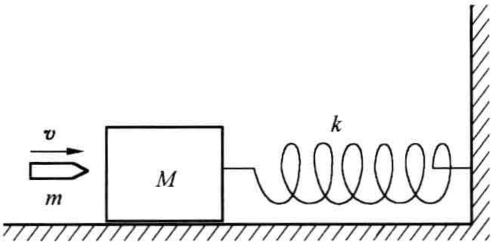


Fig. 12-21 For problem 12-16

$$x_2 = 0.06\cos\left(10t + \frac{\pi}{4}\right)$$

with same frequency in same vibrating direction. Find the equation of resultant oscillation.

12-20 A particle takes part in three simple harmonic motions which are

$$x_1 = A\cos\omega t$$

$$x_2 = A\cos\left(\omega t + \frac{\pi}{3}\right)$$

$$x_3 = A\cos\left(\omega t + \frac{2\pi}{3}\right)$$

Use the addition of their rotating vectors to find the equation of resultant oscillation.

Chapter13

Mechanical Waves

A wave is any disturbance from an equilibrium state, which propagates with time from one region of space to another. Wave phenomena are found in all branches of physics. The concept of waves, as a matter of fact, is one of the most important unifying in all of physics.

A flag is waving in the breeze; a stone is dropped into a pool of water, creating a disturbance which spreads out in concentric circles eventually reaching all parts of the pool; a sound is carried by means of traveling waves from source to the listeners . . . , they are the most familiar everyday experience that is classified among mechanical wave.

Another sort of waves, such as visible light, sunlight, infrared and ultraviolet waves, x-rays, microwaves and the waves that activate our radios and television sets etc. which exist in the atmosphere surrounding us, belong to electro-magnetic waves, which we will deal with in the next two chapters.

In this chapter, we focus on mechanical wave which is the propagation of a physical disturbance in an elastic medium as it is displaced from an equilibrium state. In other words, mechanical wave existence requires a mechanical vibration source and an elastic medium. It is important to recognize that no physical medium is necessary for the transmission of electromagnetic waves. However many of the basic concepts and principles presented in this chapter are also applicable to electromagnetic waves and all kinds of other waves.

13. 1 The Generation and Propagation of a Mechanical Wave

The most typical waves in nature are transverse waves and longitudinal waves. To understand that mechanical wave is the propagation of a mechanical oscillation in an elastic medium, let' s analyze how the transverse waves and longitudinal waves generate and propagate.

13. 1. 1 Transverse waves

In a transverse wave the vibration of the individual particles of the medium is perpendicular to the direction of wave propagation. Of all possible mechanical waves, a wave transmitted along a stretched string is perhaps the simplest. We assume an "ideal" string in which there are no friction-like forces to cause the wave to die out as it travels along. We assume further that our string is so long that we do not have to be concerned about any "echoes" that might rebound from its far end. If you give one end of such a string a single up and down jerk, as in Fig. 13-1, an impulse is passed along the string. If you move your hand up and down in continuous simple harmonic motion, giving the left end of the string a periodical displacement in the direction perpendicular to the string, an extended transverse traveling wave is generated as in Fig. 13-2.

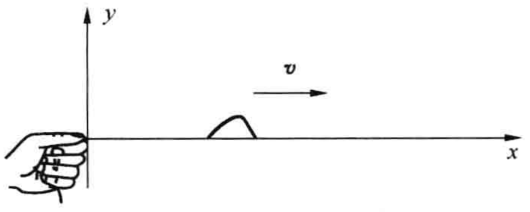


Fig. 13-1 Sending a pulse down a long string

How does the displacement propagate along the string? Because the string consists of a large number of elements, each connected to its neighbors by elastic material. If the left end is given a small displacement, which gives rise to an elastic force in the material adjacent to it; then the next element is displaced, and then the next, and so on. On the other hand, from the Newton's third law, the next element exerts an elastic reactionary force to the left one, driving it to restore to its equilibrium position. In this way, all elements vibrate one by one in simple harmonic motion.

Fig. 13-2 shows the shape of the wave at successive intervals of one-forth of a period.

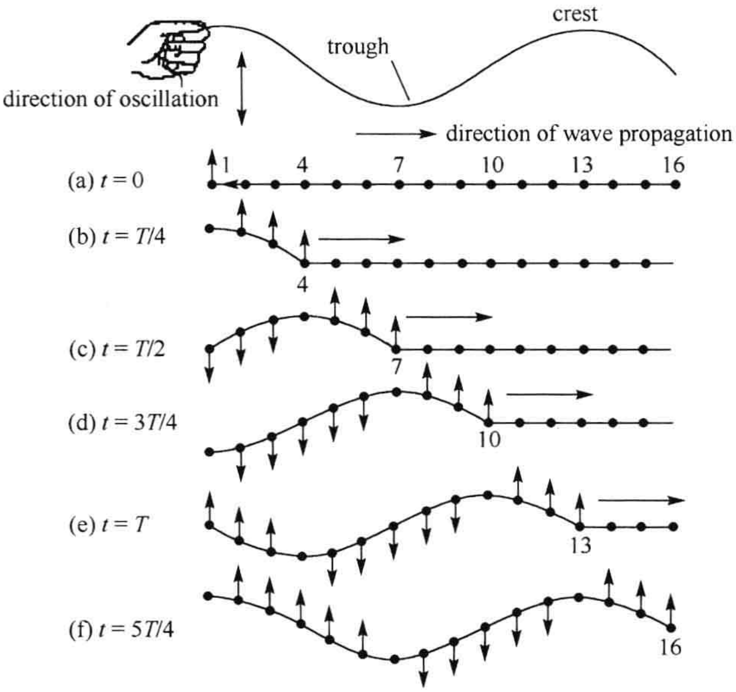


Fig. 13-2 A transverse wave generates and propagates along a stretched string, shown at successive intervals of one-forth of a period

(1) In Fig. 13-2(a), $t=0$, point 1 that your hand holds, starts to move up, its phase is $3\pi/2$, say, all others remain at rest.

(2) In Fig. 13-2(b), $t=T/4$, point 1 reaches the uppermost position, point 4 begins to move up, repeating the phase of $3\pi/2$, which means that the oscillation has propagated to point 4.

(3) In Fig. 13-2(c), $t=T/2$, point 1 passes its equilibrium position, point 7 begins to move up, repeating the phase of $3\pi/2$.

(4) In Fig. 13-2(d), $t=3T/4$, point 1 reaches the lowest position, point 10 repeats the phase of $3\pi/2$.

(5) In Fig. 13-2(e), $t=T$, your hand finishes one period vibration so that point 1 returns to its equilibrium position, going to the second vibration period, point 13 repeats the phase of $3\pi/2$, which means that the vibration has propagated to point 13, the displacements of all elements form a completely sinusoidal waveform.

(6) In Fig. 13-2(f), $t=5T/4$, the wave has traveled to point 16, the waveform repeats, and so on.

In above studying, we keep our eyes on the wave form as it moves to the right. Alternatively, we can watch the motion of a specific element, point 4, for example, as it oscillates up and down as the wave passes through it. The displacements of these elements are in the y direction, at right angle to the direction of travel of the wave, which is the positive x direction. We call such a wave motion as a transverse one.

The feature of the transverse waves is that it consists of many crests (the positions of positive maximum displacement) and troughs (the positions of negative maximum displacement) located alternatively.

13.1.2 Longitudinal wave

In a longitudinal wave the vibration of the individual particles of the medium is parallel to the direction of wave propagation. A longitudinal wave may occur in a long coiled spring, as illustrated in Fig. 13-3. In this case, we can make the same assumptions as that in the case of the string. If your hand gives the left end of the spring a push to the right slightly, the coils near the left end are pinched closely together to form a condensation as shown in Fig. 13-3(a). When the distorting force is removed, a condensation pulse is propagated throughout the spring. If you pull the coils of the spring apart at the left, a rarefaction would be formed as shown in Fig. 13-3(b); upon removal of the disturbing force, a longitudinal rarefaction pulse would be propagated along the spring. If the left end is displaced back and forth slightly in the horizontal direction, the resulting condensations and rarefactions are passed along the spring producing a periodic longitudinal wave in the horizontal direction, as shown in Fig. 13-4.

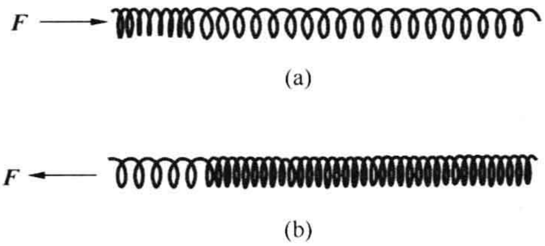


Fig. 13-3 (a) A condensation pulse formed; (b) A rarefaction pulse formed

Using the same analysis method as that of the propagation of vibration in a stretched string and treating the spring as series elements, we draw the shape of the "condensation-rarefaction wave" at successive intervals of one-forth of the first period shown in Fig. 13-4. Note that the interactive forces between the neighbor elements are in the horizontal direction so that the displacements of elements are parallel to the direction of propagation of the wave. We call such a wave motion longitudinal. Sound waves are the most common examples of longitudinal wave. The feature of the longitudinal waves is that it is formed by many condensations and rarefactions located alternatively. In the case of a spring, the condensations are the positions of maximum compression, while the rarefactions are the positions of maximum extension, instead of which they are maximum pressure and minimum pressure in the medium in the case of sound waves.

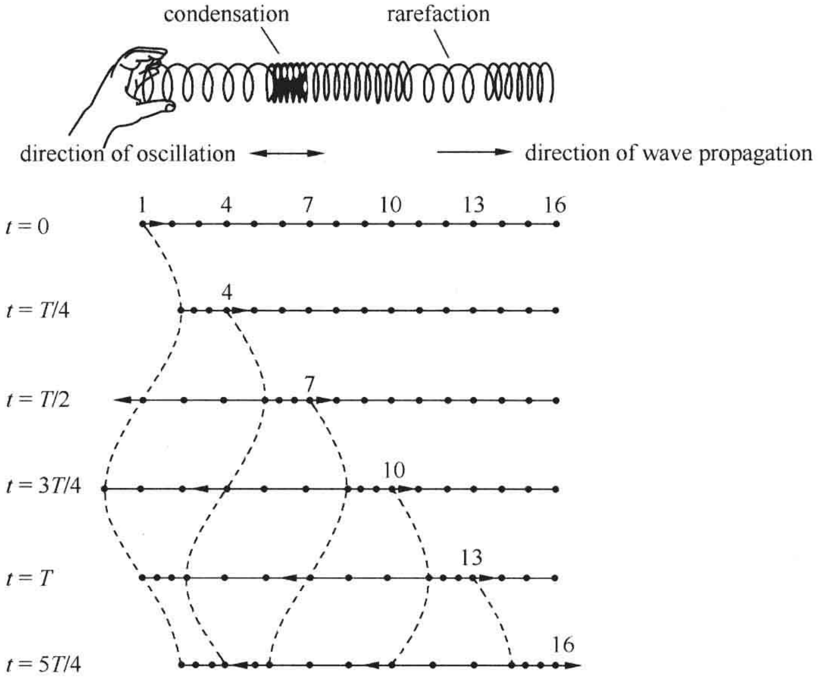


Fig. 13-4 A longitudinal wave generates and propagates in a spring shown at successive intervals of one-forth of a period

13.1.3 Wave speed, wavelength, and period

From the discussion above, we define the wave speed v as the distance that the phase of vibration propagates in a unit time, which is also called phase speed.

The wavelength λ of a periodic wave is the distance between any two adjacent elements which are in phase. In Fig. 13-2, the wavelength is the distance between point 1 and point 13, between point 4 and point 16 and so on. You can practice determining the λ in Fig. 13-4. It is obvious that the wavelength is also the distance between any two adjacent crests or troughs in a transverse wave, between any two adjacent condensation or rarefactions in a longitudinal wave. Each time the left end of the string in Fig. 13-2 makes a complete oscillation, the wave will move through a distance of one wavelength. The time required to advance one wavelength is defined as period T of wave which equals to the period of the vibrating source. Hence the wave speed v can be related to the wavelength λ and period T by the equation

$$v = \frac{\lambda}{T} \tag{13-1}$$

The frequency ν of a wave is the number of waves that pass a particular point in a unit time. It is the same as the frequency of the vibrating source and all elements and is therefore equal to the reciprocal of the period ($\nu = 1/T$). The wave speed is more often expressed in terms of its frequency rather than its period. Thus Eq. (13-1) can be rewritten as

$$v = \lambda\nu \tag{13-2}$$

The important physical relationship represented in Eq. (13-1) and Eq. (13-2) holds for all periodic waves. Note that the vibration frequency of a given source is independent from the medium, while the wave speed depends upon the elastic and inertia of the medium, which means that the wavelength varies with different medium.

Example 13-1 At room temperature, the speed of sound wave is $v = 340$ m/s in the air. $v = 1450$ m/s in the water. Find the wavelength of the sound waves of 200 Hz and 2000 Hz in the air and in the water, respectively.

Solution From Eq. (13-1), we have the wavelength of 200 Hz's sound wave:

In the air $\lambda_1 = v_a/\nu_1 = 340/200 = 1.7$ (m)

In the water $\lambda_1 = v_w/\nu_1 = 1\,450/200 = 7.25$ (m)

While, the wavelength of the 2000 Hz's sound wave:

In the air $\lambda_2 = v_a/\nu_2 = 340/2000 = 0.17$ (m)

In the water $\lambda_2 = v_w/\nu_2 = 1\,450/2000 = 0.725$ (m)

The results indicate that for a given frequency of wave, the higher the speed in a medium, the longer the wavelength in it; for a given medium, the higher the frequency, the shorter the wavelength.

13.1.4 Wave fronts and rays

It is convenient to represent a traveling wave of any sort by means of wave surfaces, or in-phase surface, defined as the locus of all points at which the phase of vibration is the same. Thus, in the case of sound waves spreading out in all directions from a point source, any spherical surface concentric with the source is a possible wave surface over which the phase of pressure vibration is the same. It is customary to draw only a few wave surfaces, usually those which pass through the maxima of the wave disturbance, separated by the wavelength. The wave having spherical wave surface is called spherical wave; the wave having plane wave surface is called plane wave, as shown in Fig. 13-5(a) and (b). At a given time, the surface formed from the locus of all points to which the wave disturbance just reached, is defined as the wave front of that time. Obviously, wave front is indeed the front wave surface of the given time.

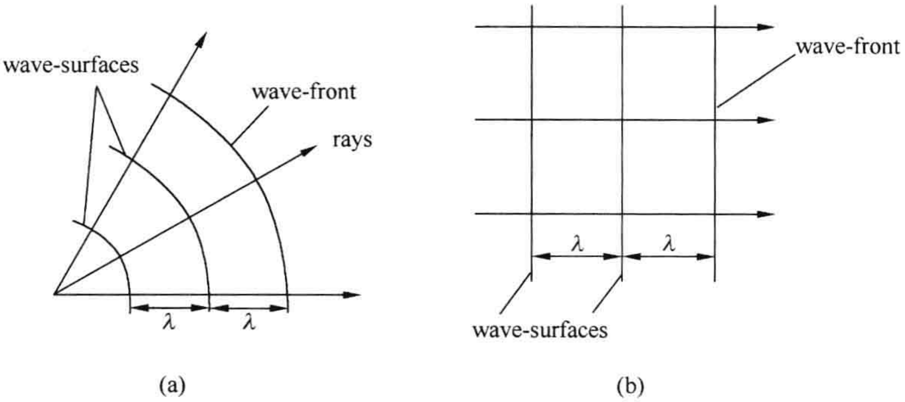


Fig. 13-5 (a) Wave surfaces, wave front and rays in a spherical wave;
(b) Wave surfaces, wave front and rays in a plane wave

We use rays to represent waves quite often in the later chapters. A ray is an imaginary line drawn in the direction in which the wave is traveling. Thus, in Fig. 13-5(a), the rays are the radii of the spherical wave surface while in Fig. 13-6(b) they are straight lines perpendicular to the wave surface. In fact, in every case in which the waves are traveling in a homogeneous isotropic medium, the rays are straight lines at right angle to the wave surfaces.

13.2 Wave Speed and Elasticity of the Medium

If a wave travels through a medium, for example, the stretched string described in the last section, it must settle the elements of that medium into oscillation as it passes. For that to happen, the medium must possess both inertia and elasticity. At the same time, as the wave passes, the medium deforms in some way. The wave speed purely depends on the elasticity and inertia of the medium.

13.2.1 Elasticity of a medium, stress, and strain

When the shape or volume of a body or a section of medium changes, we say that it is deformed. Three kinds of deformations are shown in Fig. 13-6(a), (b), and (c). Stress is defined as F/A, namely, force per area, or force divided by the area over which it acts. Associated with each type of stress there is a corresponding type of strain.

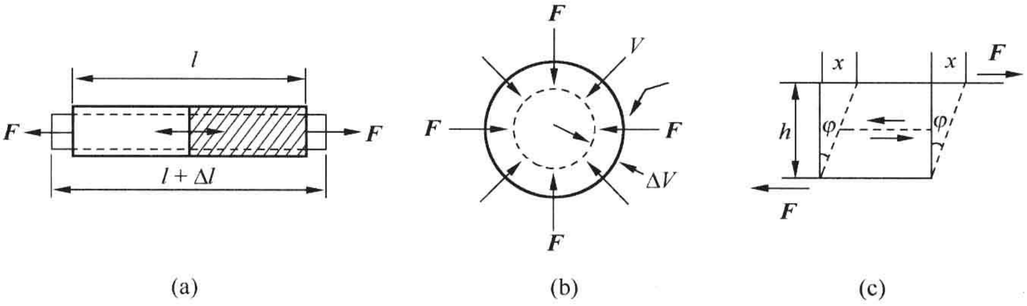


Fig. 13-6 (a) Tensile strain defined as $\Delta l/l$; (b) Volume strain defined as $\Delta V/V$; (c) Shearing strain defined as x/h

(1) Tensile strain occurs when equal and opposite pulls (or compressions) exert on a body, like the bar in Fig. 13-6(a), it is defined as the ratio of the change in length to the original length:

$$\text{Tensile strain} = \Delta l/l$$

The corresponding stress is perpendicular to the area.

(2) Volume strain occurs when a hydrostatic pressure exerts on a body, like the portion of a fluid in Fig. 13-6(b), it is defined as the ratio of the change in volume to the original volume:

$$\text{Volume strain} = \Delta V/V$$

The corresponding stress is also normal to the area.

(3) Shearing strain occurs when a shearing force exerts on the faces of a body, like the block in Fig. 13-6(c), it is defined as the ratio of the displacement x to the transverse dimension h:

$$\text{Shearing strain} = x/h = \varphi$$

The corresponding stress is tangential to the area. Shearing stress is not produced in liquids or gases.

13.2.2 Elastic Modules

The stress required to produce a given strain depends on the elastic property of the material under stress. Experiments show that stress is proportional to strain. The ratio of stress to strain, or the stress per unit strain, is called an elastic modulus of the material (or the medium). The larger the elastic modulus, the greater the stress needed for a given strain.

(1) Young's modulus or stretch modulus for the longitudinal stress (tensile or compressive), as shown in Fig. 13-6(a), is represented by Y:

$$Y = \frac{F/A}{\Delta l/l} \tag{13-3}$$

If the proportional limit is not exceeded, the radio is constant for given material.

(2) Bulk modulus for volume strain, as shown in Fig. 13-6(b), is represented by B:

$$B = \frac{F/A}{\Delta V/V} \tag{13-4}$$

(3) Shearing modulus for shearing strain, as shown in Fig. 13-6(c), is represented by N:

$$N = \frac{F/A}{\varphi} \tag{13-5}$$

13.2.3 Wave speed in the elastic medium

Wave speed depends only on the density and elasticity of the medium in which wave propagates.

1. Speed of propagation for transverse waves

If a transverse wave travels in a medium, the deformation of the medium is shearing strain that corresponds to a shearing stress. Because liquid and gas can not be able to produce shearing stress, transverse waves propagate only in solids not in liquids or gases. The speed of propagation for a transverse wave is given by

$$v = \sqrt{N/\rho} \tag{13-6}$$

in which N is shearing modulus, ρ is the mass per unit volume of the medium. The speed of a wave traveling along a stretched string is given by

$$v = \sqrt{T/\mu} \tag{13-7}$$

where T is the tension in the string, μ is the mass per unit length of the string.

2. Speed of propagation for longitudinal waves

If a longitudinal wave travels in a medium which subjects to tension or compression, it corresponds to a normal stress. Because solids, liquids and gases all are able to produce a normal stress, longitudinal waves can propagate in them. The speed of propagation for a longitudinal wave is

$$v = \sqrt{B/\rho} \tag{13-8}$$

Here B is bulk modulus, ρ is the mass per unit volume of the medium. If longitudinal wave travels along a bar, B could be substituted by Young's modulus:

$$v = \sqrt{Y/\rho} \tag{13-9}$$

From the discussion above, we can make the following conclusion: the speed of wave propagation depends on the modulus and mass density of the medium, but the period of the wave is simply the period of the oscillation of wave source and has nothing to do with medium. Therefore wavelength is related to medium, according to Eq. (13-1) $v = \lambda/T$. For example, when a sound wave of given frequency is set up in air and in water, the corresponding wavelength will be different.

13. 3 Wave Function of a Plans Simple Harmonic Wave

An important special case of a traveling wave is the simple harmonic wave or harmonic wave in which all elements of medium oscillate in the way of SHM. Any wave of arbitrary shape can be regarded as a superposition of several harmonic waves. Thus, if we understand the motion of harmonic waves, we can understand the motion of any kind of waves as well. Hereafter we will concentrate on harmonic waves, particularly on the plane harmonic wave in which the wave front is a plane.

13. 3. 1 How to describe a Harmonic wave

To describe a harmonic wave, we need to find the displacement of any point in the wave line at any instant. Suppose that a plane harmonic transverse wave propagates along the positive x direction, as shown in Fig. 13-7. Assume that there is no friction force in the medium. To determine the displacement y of point P at time t, let us compare the motion of point P (its equilibrium position is x at x axis) with the motion of the origin ($x=0$). We find that particle P moves in the same manner as the particle at $x=0$, but after a lapse of time that is proportional to the distance between them. In other words, they oscillate with the same amplitude and same frequency, but different phase which depends on the time required for the wave to travel from $x=0$ to point P. Say, the displacement of the origin at the time t is given by

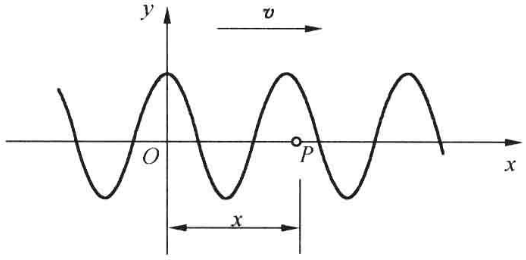


Fig. 13-7 A wave traveling in positive x direction

$$y = A\cos \omega t \tag{13-10}$$

The time required for the propagation of vibration from $x=0$ to point P is given by $\frac{x}{v}$ where v is the wave speed. It is clear that the motion of point P at time t is the same as the motion of point $x=0$ at the earlier time $\left(t-\frac{x}{v}\right)$. Thus the displacement of point P at time t is obtained simply by replacing t in Eq. (13-10) by $\left(t-\frac{x}{v}\right)$, that is

$$y(x,t) = A\cos\omega\left(t - \frac{x}{v}\right) \tag{13-11}$$

From $\omega = 2\pi/T$ and $\lambda = vT$, Eq. (13-11) can be rewritten as

$$y(x,t) = A\cos 2\pi\left(\frac{t}{T} - \frac{x}{\lambda}\right)$$

or

$$y(x,t) = A\cos 2\pi\left(\nu t - \frac{x}{\lambda}\right) \tag{13-12}$$

Any one of these various formulas can be used in determination of the displacement of any point in the wave line. Therefore, we call Eq. (13-11) or Eq. (13-12) as wave function of a plane harmonic wave, which holds for both transverse waves and longitudinal waves.

13.3.2 The physical meanings of wave function

At any given coordinate x, Eq. (13-11) or Eq. (13-12) gives the displacement y of the particle at that coordinate, as a function of time t. For example, at the position $x = \lambda/4$ we have

$$y = A\cos 2\pi\left(\frac{t}{T} - \frac{1}{4}\right) = A\cos\left(\omega t - \frac{\pi}{2}\right)$$

This is the oscillation function of that point, which is plotted in Fig. 13-8.

At any given time t, Eq. (13-11) or Eq. (13-12) gives the displacement y of any particle from its equilibrium position, as a function of the coordinate x of the particle. For the case of a transverse wave, the equation represents the wave pattern or waveform at that instant, as if we had taken a photograph of it. For example, at time $t = T/2$, we have

$$y = A\cos\left(\pi - \frac{2\pi x}{\lambda}\right) = A\cos\left(\frac{2\pi x}{\lambda} - \pi\right)$$

This is the waveform curve at that instant, which is plotted in Fig. 13-9.

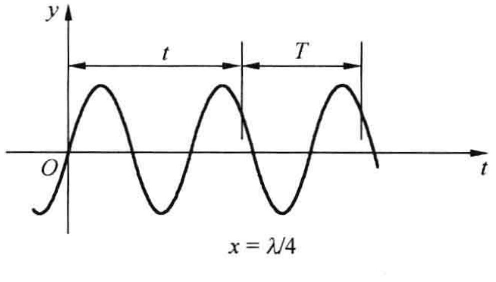


Fig. 13-8 The oscillation curve at $x=\lambda/4$

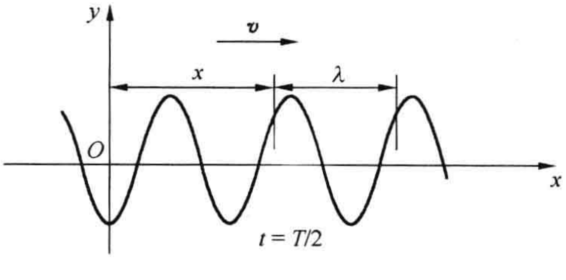


Fig. 13-9 The wave pattern at $t=T/2$

Fig. 13-10 shows the waveforms at time t (solid curve) and the later time $t+\Delta t$ (dashed curve). Suppose A and B are two points that equilibrium positions are x and $x+\Delta x$ respectively, the displacement of point B at $t+\Delta t$ is the same as that of point A at time t. From Eq. (13-11),

$$\omega\left(t - \frac{x}{v}\right) = \omega\left(t + \Delta t - \frac{x+\Delta x}{v}\right)$$

we obtain

$$\Delta x = v\Delta t$$

where v is the wave speed, which suggests that the wave travels a distance $x = v\Delta t$ during the time interval Δt.

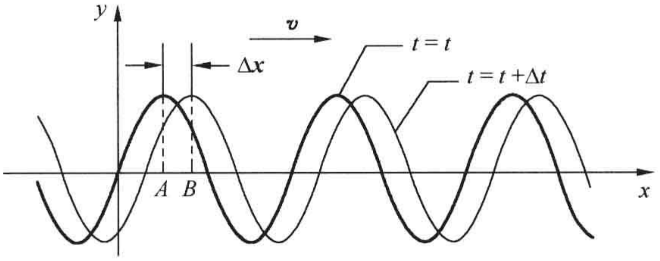


Fig. 13-10 The waves forms at t and $t+\Delta t$

13.3.3 Wave function for a wave traveling in the negative *x* direction

The formulas above may be used to describe a wave traveling in the negative x-direction by making a simple modification. In this case, the motion of point x at time t is the same as the motion of point $x=0$ at $(t+x/v)$, thus the displacement of point P at time t is obtained simply by replacing t in Eq. (13-11) by $(t+x/v)$

$$y = A\cos\omega\left(t+\frac{x}{v}\right) \tag{13-13}$$

which is the wave function for a wave traveling in the negative x-direction.

It is necessary to distinguish v of the wave and the vibration speed of a particle of the medium; the latter one is given by taking the partial derivative with respect to t in Eq. (13-11)

$$u = \frac{\partial y}{\partial t} = -\omega A \sin\omega\left(t-\frac{x}{v}\right) \tag{13-14}$$

It represents the oscillation speed of particle located at x. To find the acceleration of the point, you can take the partial derivative with respect to time in Eq. (13-14).

Example 13-2 A plane harmonic wave is moving in the direction of decreasing x with the wave speed $v=5$ m/s. The oscillation curve of point O is given in Fig. 13-11(a).

(1) Find the wave function.

(2) Draw the oscillation curve of point $x=25$ m.

(3) Draw the wave pattern at $t=3$ s and find the wavelength.

(4) Give the coordination of the wave crests at that time.

Solution (1) From Fig. 13-11(a), $A=2\times10^{-2}$ m, $T=4$ s, $\omega = 2\pi/T = \frac{\pi}{2}\ \text{s}^{-1}$, and at $t=0$ the origin point is upward passing its equilibrium position corresponding to the initial phase $\varphi = -\frac{\pi}{2}$. The oscillation function of point O is then

$$y_0 = 2\times10^{-2}\cos\left(\frac{\pi}{2}t-\frac{\pi}{2}\right)\text{m} \qquad ①$$

Note the wave propagates in $-x$ direction so that the wave function is given by

$$y(x,t) = 2\times10^{-2}\cos\left[\frac{\pi}{2}\left(t+\frac{x}{5}\right)-\frac{\pi}{2}\right]\ \text{m} \qquad ②$$

(2) Letting $x=25$ m, we have the oscillation function at that point:

$$y = 2\times10^{-2}\cos\left[\frac{\pi}{2}\ (t+5)-\frac{\pi}{2}\right]= 2\times10^{-2}\cos\left(\frac{\pi}{2}\ t\right)\text{m} \qquad ③$$

The corresponding curve is shown in Fig. 13-11(b).

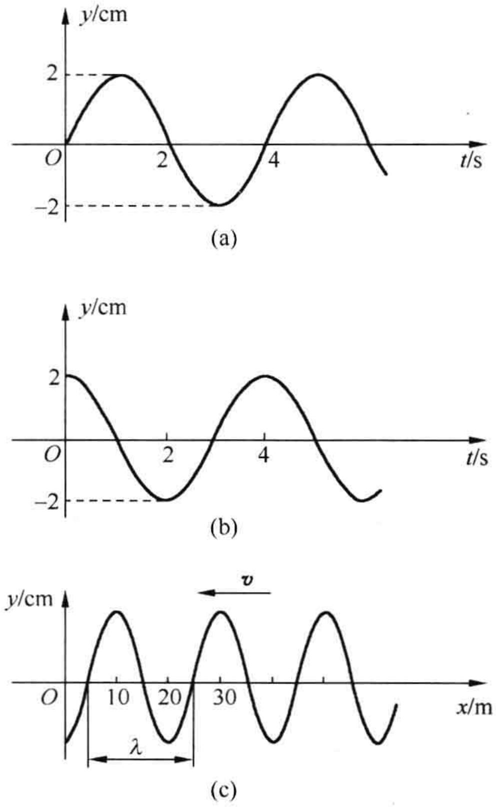


Fig. 13-11 For Example 13-2

(3) Letting t=3s in Eq. ②, we get the wave pattern function at that moment

$$y = 2\times10^{-2}\cos\left[\frac{1}{2}\pi\left(3+\frac{x}{5}\right)-\frac{\pi}{2}\right]$$

$$= 2\times10^{-2}\cos\left(\frac{\pi\, x}{10}+\pi\right)\ \text{m} \qquad ④$$

The curve is shown in Fig. 13-11(c) in which the wavelength is found as $\lambda = 20$ m.

(4) Letting $y = 2\times10^{-2}$ m in Eq. ④, we have

$$\cos\left(\frac{\pi\, x}{10}+\pi\right)= 1$$

so that

$$\frac{\pi\, x}{10}+\pi = 2k\pi,\quad k = 0,\pm1,\pm2,\cdots$$

The coordination of wave crests meet

$$x = 10\ (2k-1)\ \text{m},\quad k = 0,\pm1,\pm2,\cdots$$

let $k = -1, 0, 1, 2, \cdots$ successively, then $x = -30, -10, 10, 30, \cdots$ respectively, see Fig. 13-11(c).

Example 13-3 The waveform of a plane transverse wave at t=0 is shown in Fig. 13-12. Suppose the wave is traveling along the direction of $+x$ with the speed of $v = 0.08$ m/s.

(1) Determine the vibration direction of elements A and B at t=0.

(2) Find the vibration function of point O.

(3) Write the wave function.

Solution (1) The waveform shifts to the right after t as shown in dashed line. By comparing the new positions of points A and B at $t = \Delta t$ with their position at t=0, we see that A moves downward and B moves upward at t=0.

(2) Fig. 13-12 suggests that A=0.2 m, λ=0.4 m, so $T=\lambda/v$=5 s; the oscillation function of point O is

$$y_0 = A\cos\left(2\pi\frac{t}{T}+\varphi\right)= 0.2\cos\left(\frac{2}{5}\pi\, t+\varphi\right)\ \text{m} \qquad ①$$

The phase constant φ is determined in the following: the vibration speed is

$$u = \frac{\mathrm{d}y}{\mathrm{d}t} = -0.08\ \pi\ \sin\left(\frac{2}{5}\pi\, t+\varphi\right) \qquad ②$$

Letting t=0, from Fig. 13-12 $y_0 = 0$, so $\cos\varphi = 0$, that is

$$\varphi = \pi/2 \quad \text{or} \quad 3\pi/2$$

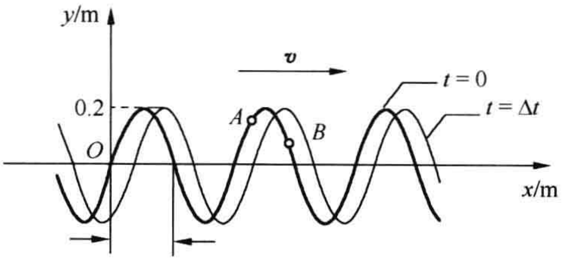


Fig. 13-12 For Example 13-3

Because point O moves downward at $t=0$, so that

$$u < 0, \quad \sin\varphi > 0$$

so take

$$\varphi = \pi/2 \qquad ③$$

the oscillation equation of that point is then

$$y = 0.08\cos\left(\frac{2}{5}\pi t + \frac{\pi}{2}\right)\ \text{m} \qquad ④$$

(3) From Eq. (13-11), the wave function

$$y = 0.2\cos\left[\frac{2}{5}\pi\left(t - \frac{x}{0.08}\right) + \frac{\pi}{2}\right]\text{m} \qquad ⑤$$

obtained.

Example 13-4 A plane harmonic transverse wave is propagating along the direction of increasing x, the amplitude $A = 0.10$ m, angular frequency $\omega = 7\pi \cdot \text{s}^{-1}$. At the instance of $t = 1.0$ s, point A of $x = 0.10$ m is passing through its equilibrium position downward to the direction of $-y$ and the point B of $x = 0.20$ m is passing through $y = 0.05$ m upward to the direction of $+y$. Suppose the wavelength $\lambda > 0.1$ m, write the wavefunction.

Solution According to Eq. (13-11), the wavefunction can be written as

$$y = 0.1\cos\left(7\pi t - \frac{2\pi}{\lambda}x + \varphi\right)\ \text{m}$$

We find λ and φ in the following: When $t = 1.0$ s we have

$$y = 0.1\cos\left(7\pi - \frac{2\pi}{\lambda}x + \varphi\right)\ \text{m}$$

we know that point A of $x = 0.1$ m is passing its equilibrium position

$$y_A = 0.1\cos\left[7\pi - 2\pi(0.1/\lambda) + \varphi\right] = 0$$

And it is moving in $-y$ direction, so that

$$7\pi - 2\pi(0.1/\lambda) + \varphi = \frac{1}{2}\pi \qquad ①$$

At the same time point B of $x = 0.2$ m

$$y_B = 0.1\cos\left[7\pi - 2\pi(0.2/\lambda) + \varphi\right] = 0.05\ \text{m}$$

generally

$$7\pi - 2\pi(0.2/\lambda) + \varphi = 2k\pi \pm \frac{\pi}{3} \qquad ②$$

consider that B is moving in $+y$ direction so take $-\frac{\pi}{3}$ in Eq. ②, and $x_B - x_A = 0.1\ \text{m} < \lambda$, which means that the phase difference between A and B is smaller than 2π, and k in Eq. ② must also satisfy $\lambda > 0$, so take $k = 0$ that

$$7\pi - 2\pi(0.2/\lambda) + \varphi = -\frac{1}{3}\pi \qquad ③$$

Combine Eq. ① and Eq. ③ , we obtain $\lambda = 0.24$ m, $\varphi = -17\pi/3$.

The wavefunction is given by

$$y = 0.1\cos\left(7\pi t - \frac{2\pi x}{0.24} - \frac{17}{3}\pi\right)\ \text{m} = 0.1\cos\left(7\pi t - \frac{\pi x}{0.12} + \frac{1}{3}\pi\right)\ \text{m}$$

13.4 Energy, Energy Flow, and Wave Intensity

13.4.1 Energy and energy density

A transverse wave on a string has kinetic energy because the particles are in motion, and it has potential energy because the segments of string deform due to the elasticity of the string.

Suppose that a wave travels along x axis described by Eq. (13-11). Consider a small interval $\mathrm{d}x$ along the string, μ is the mass per unit length. The mass of the segment is $\mathrm{d}m = \mu\,\mathrm{d}x$ and the velocity of it is $\partial y/\partial t$; hence the kinetic energy associated with this piece of string is

$$\mathrm{d}E_{\mathrm{k}} = \frac{1}{2}\mu\,\mathrm{d}x\left(\frac{\partial y}{\partial t}\right)^2$$

Replacing $\frac{\partial y}{\partial t}$ by Eq. (13-14), we obtain

$$\mathrm{d}E_{\mathrm{k}} = \frac{1}{2}\mu\,\mathrm{d}x\,A^2\omega^2\sin^2\omega\left(t - \frac{x}{v}\right) \tag{13-15}$$

To find the potential energy, note that the piece of string is slightly elongated when the wave passes; see the lift side in Fig. 13-13, the wave stretches the string from its original length $\mathrm{d}x$ to a length $\sqrt{(\mathrm{d}x)^2 + (\mathrm{d}y)^2}$, and hence the change of length is $\delta l = \sqrt{(\mathrm{d}x)^2 + (\mathrm{d}y)^2} - \mathrm{d}x$ or

$$\delta l = \mathrm{d}x\left(\sqrt{1 + \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2} - 1\right)$$

Assuming $\mathrm{d}y/\mathrm{d}x$ is very small (so the string only makes a small angle with its original, horizontal direction), we can take advantage of the following approximation $\sqrt{1 + \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2} \approx 1 + \frac{1}{2}\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2$, so that the change of length becomes $\delta l = \frac{1}{2}\mathrm{d}x\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2$. The potential energy associated with the interval $\mathrm{d}x$ is simply the work that must be done against the tension F to stretch the string by the amount δl, that is

$$\mathrm{d}E_{\mathrm{p}} = F\,\delta l = \frac{1}{2}F\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2\mathrm{d}x$$

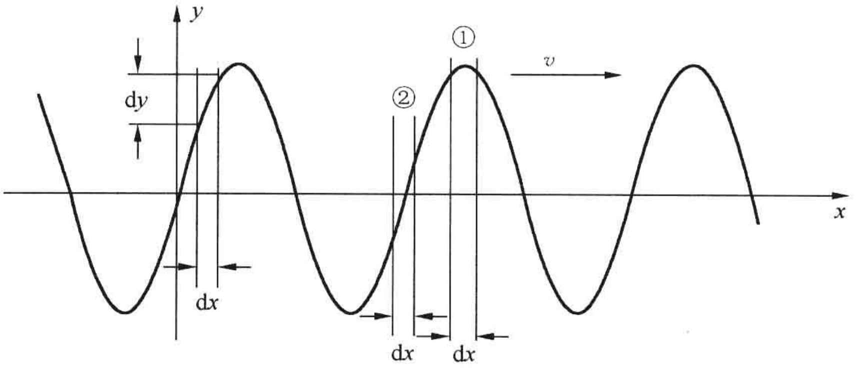


Fig. 13-13 In element ①, the stored kinetic energy and potential energy are zero respectively; In element ②, the stored energy are maximum

Replacing dy/dx by $\partial y/\partial x$, we obtain the potential energy

$$dE_p = \frac{1}{2}F\left(\frac{\partial y}{\partial x}\right)^2 dx$$

Taking the partial derivative with respect to x in Eq. (13-11)

$$\left(\frac{\partial y}{\partial x}\right)^2 = \frac{A^2\omega^2}{v^2}\sin^2\omega\left(t-\frac{x}{v}\right)$$

Using Eq. (13-7), we have $F=\mu v^2$, thus

$$dE_p = \frac{1}{2}\mu\, dx\, A^2\omega^2\sin^2\omega\left(t-\frac{x}{v}\right) \tag{13-16}$$

The total energy associated with the interval dx is the sum of Eq. (13-15) and Eq. (13-16)

$$dE = dE_k + dE_p = \mu\, dx\, A^2\omega^2\sin^2\omega\left(t-\frac{x}{v}\right) \tag{13-17}$$

To find the physical meaning of Eq. (13-15)、Eq. (13-16) and Eq. (13-17), see the elements ① and ② in Fig. 13-13. When the string element is in the position of $y=A$ (wave crest, see ①), its oscillation velocity — and thus its kinetic energy — is zero. On the other hand, its length has its normal undisturbed value dx, so that its stored potential energy is also zero. However, when the element is rushing through its $y=0$ position, its vibration velocity — and thus its kinetic energy — is a maximum. On the other hand, it is steeply tilted and thus stretched to its maximum extend(see ②); its stored potential energy then has the maximum value. From above discussion, we can draw the following conclusion: when a harmonic wave travels along a stretched string, both of its kinetic energy and potential energy are changing periodically; they are always in-phase, and have maximum value at $y=0$ and minimum value at $y=\pm A$. The total energy stored in the string is not conserved, but transported by the wave.

Energy density is the energy stored in a unit length defined by

$$w = dE/dx = \mu A^2\omega^2\sin^2\omega\left(t-\frac{x}{v}\right) \tag{13-18}$$

The average energy density is defined as the average value of energy density over one period that is

$$\overline{w} = \mu A^2\omega^2\frac{1}{T}\int_0^T \sin^2\omega\left(t-\frac{x}{v}\right)dt = \frac{1}{2}\mu A^2\omega^2 \tag{13-19}$$

The factors in this equation that the average energy density is proportional to the mass density of the medium and also to the square of amplitude and square of frequency of the wave is a general result. It is true for mechanical waves of all types.

For the case of longitudinal wave shown in Fig. 13-4, the small interval of medium is an element volume dV instead dx; the mass density is ρ, that is mass per unit volume instead μ in Eq. (13-17) and Eq. (13-18). Therefore the average energy density becomes

$$\overline{w} = \frac{1}{2}\rho A^2\omega^2 \tag{13-20}$$

13.4.2 Energy flow and wave intensity

As discussed above, a traveling wave passing through a medium is associated with energy transported through the medium. The amount of energy that moves past a given area in one

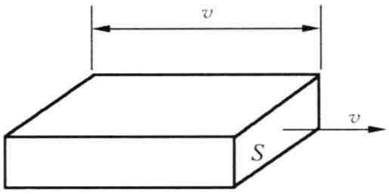


Fig. 13-14 The energy transported per second from an area S equals the energy stored in the volume of vS

unit of time is defined as energy flow through that area. Suppose area S is perpendicular to the direction of wave propagating as shown in Fig. 13-14, the average energy stored in volume of vS, is $vS\overline{w}$ called as the average energy flow or the average power transmitted by the wave, symbolized as $\overline{P}$, thus

$$\overline{P} = vS\overline{w} = \frac{1}{2}\rho A^2\omega^2 vS \qquad (13\text{-}21)$$

The rate of $\overline{P}/S$ is the average energy transported per second through one unit of area perpendicular to the direction of v defined as average energy flow density, or intensity of the wave symbolized as I. That is

$$I = \frac{\overline{P}}{S} = \frac{1}{2}\rho A^2\omega^2 v \qquad (13\text{-}22)$$

which is along the direction of v. Eq. (13-22) is also called as the power of a mechanical wave.

13. 4. 3 The amplitudes of plane waves and spherical waves

In the wave function of the plane harmonic wave

$$y = A\cos\left(\omega t - \frac{x}{v}\right)$$

We have assumed that the amplitudes are a constant for all points in the medium. Let's see what it means in view of energy. In Fig. 13-15, suppose there is a cuboids (rectangular) with the end surfaces S_1 and S_2 perpendicular to the direction of wave speed, $\overline{P}_1$ are $\overline{P}_2$ the average energy flux through S_1 and S_2 respectively, thus

$$\overline{P}_1 = \overline{w}_1 vS_1 = \frac{1}{2}\rho\omega^2 A_1^2 vS_1, \quad \overline{P}_2 = \overline{w}_2 v S_2 = \frac{1}{2}\rho\omega^2 A_2^2 vS_2$$

where A_1 and A_2 are the amplitudes in S_1 and S_2. $A_1 = A_2$ then, we have $\overline{P}_1 = \overline{P}_2$, that means if amplitude remains unchanged, the average energy flow through S_1 and S_2 has the same amount. It is the same amount of energy flow as it moves in the cuboids from S_1 that moves out of the cuboids from S_2, which indicates that the medium does not absorb energy at all.

However the situation is different in the case of a spherical wave, the amplitude is decreased although the medium does not absorb energy. In Fig. 13-16, suppose S_1 and S_2 are two spheres that depart from source O by r_1 and r_2 respectively, that is $S_1 = 4\pi r_1^2$ and $S_2 = 4\pi r_2^2$. If there is no energy absorbed in the medium, then the average energy flows through S_1 and S_2 have the same amount, thus

$$\frac{1}{2}\rho\omega^2 A_1^2\, v\, 4\pi r_1^2 = \frac{1}{2}\rho\omega^2 A_2^2\, v 4\pi r_2^2$$

where A_1 and A_2 are the amplitude in S_1 and S_2 respectively. Hence

$$\frac{A_1}{A_2} = \frac{r_2}{r_1}$$

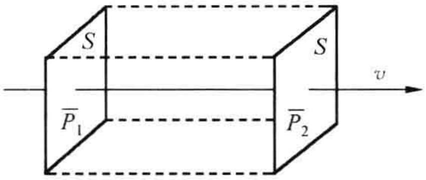


Fig. 13-15 Energy flow in the case of a plane wave

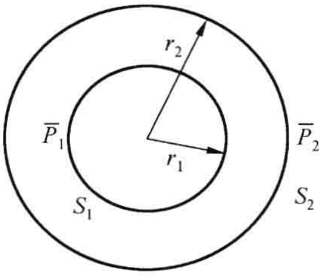


Fig. 13-16 Energy flow in the case of a spherical wave

is obtained, which means that the amplitude is inversely proportional to the distance from the source. Therefore the wave function of spherical harmonic wave should be

$$y = \frac{A}{r}\cos\omega\left(t - \frac{r}{v}\right)$$

where r is the distance from the source; A is the amplitude in the points of one unit distance from the source.

13.5 Huygens' Principle

We have known that mechanical wave existence requires a mechanical vibration source and an elastic medium and mechanical wave is the propagation of a mechanical oscillation in an elastic medium. A mechanical wave originates in the disturbance of some portion of an elastic medium, in which the particles displace from their equilibrium positions, hence causing them to oscillate about their equilibrium positions. In order to explain the manner in which a wave is propagated through a medium, it was postulated by Huygens that each point on the wave front of a disturbance may be considered as a centre (vibration source) from which secondary waves are emitted throughout the medium. Huygens' principle is: **All points on a wave front serve as point sources of spherical secondary wavelets; after a time t, the new position of the wave front will be the envelope of these secondary wavelets.**

Huygens' principle is a geometrical method for finding the new wave front at later instant from known shape of a portion of a wave front at some instant. What shape will be at some later instant? Huygens' principle allows us to find the new wave front at any instant; it is therefore employed with a great advantage in discussing various problems concerning the propagation of waves. Fig. 13-17 illustrates Huygens' principle.

Suppose that the original wave front AA' of a spherical wave is traveling as shown in Fig. 13-17(a). We wish to find the shape of the new wave front after the time interval Δt. Let v represent the speed of wave propagating in the medium. Construct a number of circle

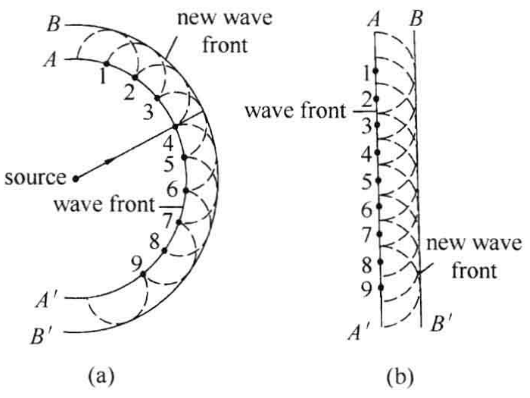


Fig. 13-17 Huygens' principle is applied to (a) aspherical wave-front and (b) a plane wave-front

(traces of spherical wavelets) of radius $r = v\Delta t$, with centers along AA'. Then the trace of envelope of these wavelets is the new wave front which is the surface BB'. The speed v of wave propagation in the medium has been assumed to be the same at all points and in all directions. The new wave front of a plane wave will be found in the same way as shown in Fig. 13-17(b).

The diffraction of a wave can be explained by Huygens' principle. The bending of wave into the shadow region of obstacles and related phenomena are called diffraction.

All types of waves exhibit diffraction, but it is most prominent when the wavelength is larger compared with the size of the opening or of the obstacle. Diffraction is less apparent when λ is much smaller than obstacles or openings. Sound waves, with λ of the order of a meter, diffract readily around typical obstacles, while light, with λ of a few hundred nanometers, does not. It is because sound diffracts prominently while light does not, so we can talk with a person through an open doorway even though we can not see him. In fact, the extent to which a wave bends around the edges of an opening is determined by the ratio λ/d, where d is the width of the opening. Fig. 13-18 illustrates the effect of this ratio on the diffraction of water waves. In Fig. 13-18(a), plane waves are incident on a partition with a broad opening. The wave-fronts move through the opening with little bending or diffraction into the shadow region of obstacles since the ratio λ/d is small as the width d is large. In Fig. 13-18(b), only a narrow passage remains, and the wave pattern closely resembles that from a point source corresponding to the prediction of Huygens' principle. The ratio λ/d becomes larger because the width d becomes smaller and the degree of diffraction becomes more pronounced with the wave-fronts penetrating more into the shadow region of obstacles.

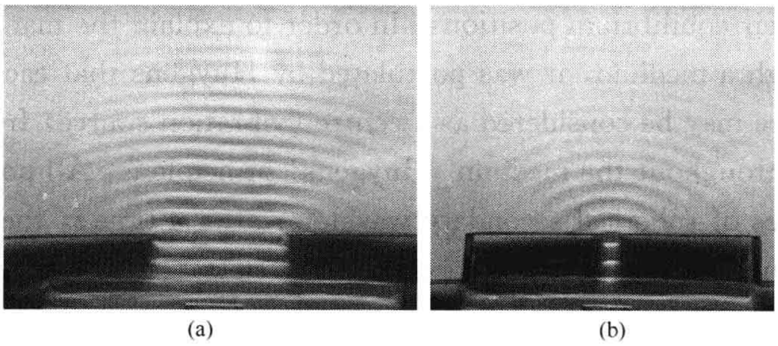

(a) (b)

Fig. 13-18 The bending of wave-front as predicted by Huygens' principle and the effect of λ/d on the diffraction of water waves

13.6 Principle of Superposition of Waves, and Interference of Waves

13.6.1 Principle of superposition of waves

As we have known, the propagation of sound from any source is not affected by other sound waves spreading in any manner through the same areas of the medium. And the same is observed for light waves. Their propagation from any object to our eye and everything we can see owing to these waves never depend on the multitude of other light waves that intersect in all possible directions of the path of the light from the object being observed.

What occur in the regions of space in which the waves are superposed? The experiment shows that for many kinds of waves two or more waves can travel in the same space independently of one another. If the particle is in path of two waves simultaneously, it participates in the vibrations of both waves, that is, its displacement at a given time is simply the resultant of the displacements that individual wave would alone give it. This process of vector addition of the displacements of a particle is called superposition of waves.

13. 6. 2 Interference of waves

In general, the superposition of two or more waves whose frequencies are different from each other is very complex. Here we will discuss a special case, that is, the superposition of two waves of equal frequency but with a constant phase difference φ between them.

Two wave sources with equal frequency, the same direction of vibration and constant phase difference are called coherent wave sources. The waves from coherent wave sources are called coherent waves.

The superposition of coherent waves forms interference pattern in the space within which alternating maximum and minimum of vibration occur. The phenomenon of superposition of waves that produces this pattern is called interference of waves. Let us discuss the condition for the formation of maxima and minima in the phenomenon of wave interference. Suppose that the equations of two coherent sources are

$$y_1 = A_1\cos(\omega t + \varphi_1)$$
$$y_2 = A_2\cos(\omega t + \varphi_2)$$

If the waves are meeting at any point P (Fig. 13-19) in the space, the particle at point P in the space participates in the vibrations of both waves, that is, its displacement at a given time t is the resultant of the displacements that each wave alone would give it. We assume that the distance from point P to the first source S_1 is r_1 and r_2 to the second source S_2. Then the equations of the vibrations from two coherent waves are

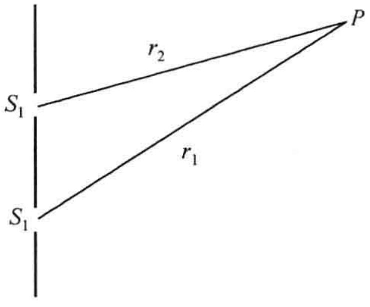


Fig. 13-19 Interference

$$y_1 = A_1\cos\left(\omega t + \varphi_1 - \frac{2\pi r_1}{\lambda}\right) \tag{13-23a}$$

$$y_2 = A_2\cos\left(\omega t + \varphi_2 - \frac{2\pi r_2}{\lambda}\right) \tag{13-23b}$$

Obviously, the equation of resultant motion at point P is given by

$$y = y_1 + y_2 = A\cos(\omega t + \varphi)$$

where A is resultant amplitude and φ is initial phase of resultant motion. They can be determined by combinations of simple harmonic motions (see chapter 12). Thus, we have

$$A = \sqrt{A_1^2 + A_2^2 + 2A_1A_2\cos\left(\varphi_2 - \varphi_1 - 2\pi\frac{r_2 - r_1}{\lambda}\right)} \tag{13-24a}$$

and

$$\tan\varphi = \frac{A_1 \sin\left(\varphi_1 - \frac{2\pi r_1}{\lambda}\right) + A_2 \sin\left(\varphi_2 - \frac{2\pi r_2}{\lambda}\right)}{A_1 \cos\left(\varphi_1 - \frac{2\pi r_1}{\lambda}\right) + A_2 \cos\left(\varphi_2 - \frac{2\pi r_2}{\lambda}\right)} \qquad (13\text{-}24b)$$

The quantity $\left(\varphi_2 - \varphi_1 - 2\pi \frac{r_2 - r_1}{\lambda}\right)$ is a constant, the phase difference of two waves at point P. And the quantity $(r_2 - r_1)$ is called the path difference which is a difference between the distances from any point P to sources S_1 and S_2 as shown in Fig. 13-19. If the phase difference $\Delta\varphi$ of two waves at point P is given by

$$\Delta\varphi = \varphi_2 - \varphi_1 - 2\pi \frac{r_2 - r_1}{\lambda} = \pm 2k\pi, \qquad k = 0, 1, 2, \cdots$$

the resultant amplitude A at point P has the maximum value $A = A_1 + A_2$. If the phase difference $\Delta\varphi$ of two waves at point P is given by

$$\Delta\varphi = \varphi_2 - \varphi_1 - 2\pi \frac{r_2 - r_1}{\lambda} = \pm (2k + 1)\pi, \qquad k = 0, 1, 2, \cdots$$

the resultant amplitude at point P has minimum value $|A_1 - A_2|$.

If we assume $\varphi_1 = \varphi_2$, then above conditions may be reduced to

$$\delta = r_2 - r_1 = \pm 2k \frac{\lambda}{2}, \quad k = 0, 1, 2, \cdots \text{ (maxima)} \qquad (13\text{-}25a)$$

$$\delta = r_2 - r_1 = \pm (2k + 1) \frac{\lambda}{2}, \quad k = 0, 1, 2, \cdots \text{ (minima)} \qquad (13\text{-}25b)$$

where δ is the path difference from two coherent sources S_1 and S_2 to any point P. Eq. (13-25a) and Eq. (13-25b) show that when the path difference from two coherent sources to some point P is equal to an integral number of wavelengths or to an even number of half wavelengths, the maximum of vibration occurs at that point. When the path difference is equal to an odd number of half-wavelengths, the minimum of vibration occurs at that point. Fig. 13-20 is a photograph of an interference pattern produced in a ripple tank by two identical circular waves emanating from two points. The alternating distribution of the maxima (bright in the photo) and minima (dark in the photo) is clear.

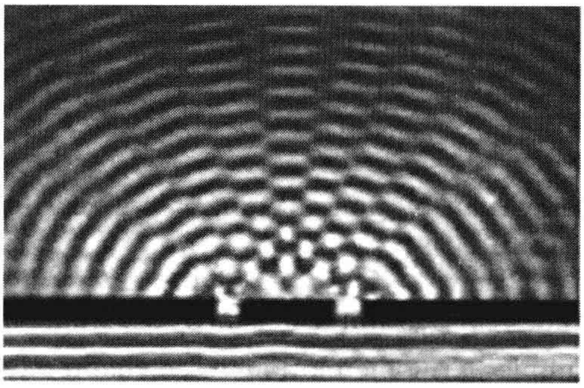


Fig. 13-20 A photograph of the interference pattern of water waves from two point sources

13.7 Standing Waves

Sometimes waves do not seem to move. We have known that if a wave train encounters a boundary, the reflected part of the wave train interferes with the incident part of the wave train. This interference can create a stationary wave pattern called a standing wave. The standing waves are special interference phenomena and it is important in the design of buildings, bridges and musical instruments.

Fig. 13-21 shows an experiment of a standing wave produced on a long string. Point A, on the tuning fork, is the wave source and B is a fixed obstacle. The incident wave, traveling to the right, is indicated by a solid line and the reflected wave, traveling to the left, is indicated by a dotted line. Note that, if the distance of AB is suitable, there are certain points along the string would not be moved up and down by the wave motion, e. g. they remain at rest and are called nodes. Between the nodal points, the particles of the string move up and down with simple harmonic motion. The points of maximum amplitude occur midway between the nodes and are called antinodes or loops. Generally, when two waves with the same vibration direction, the same frequency and the same amplitude are traveling in opposite directions along a straight line, the resultant wave of the two waves is called a standing wave. In this situation, a stationary pattern may be formed so that to the eye the waves appear to be standing still rather than progressing.

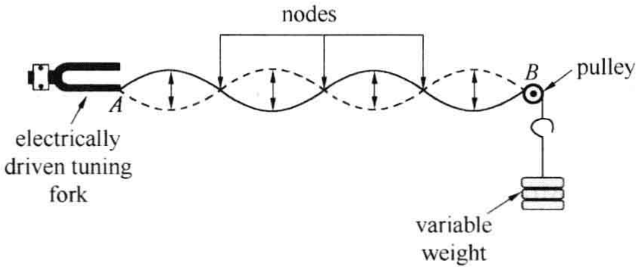


Fig. 13-21 The experiment of a standing wave

How are standing waves formed? Considering the superposition of two harmonic waves that have the same amplitude A, frequency ν and wavelength λ, but travel along x axis in opposite directions:

$$y_1 = A\cos 2\pi\left(\nu t - \frac{x}{\lambda}\right) \quad \text{(incident wave)}$$

$$y_2 = A\cos 2\pi\left(\nu t + \frac{x}{\lambda}\right) \quad \text{(reflected wave)}$$

According to the principle of superposition, we find that the resultant wave y is

$$y = y_1 + y_2 = A\cos 2\pi\left(\nu t - \frac{x}{\lambda}\right) + A\cos 2\pi\left(\nu t + \frac{x}{\lambda}\right)$$

By using the trigonometric identity, the resultant wave can be expressed more simply as

$$y = 2A\cos 2\pi\frac{x}{\lambda}\cos 2\pi\nu t \tag{13-26}$$

This is the wave function for a standing wave. We note that it is a simple harmonic motion equation in which amplitude equal to $2A\cos 2\pi\frac{x}{\lambda}$.

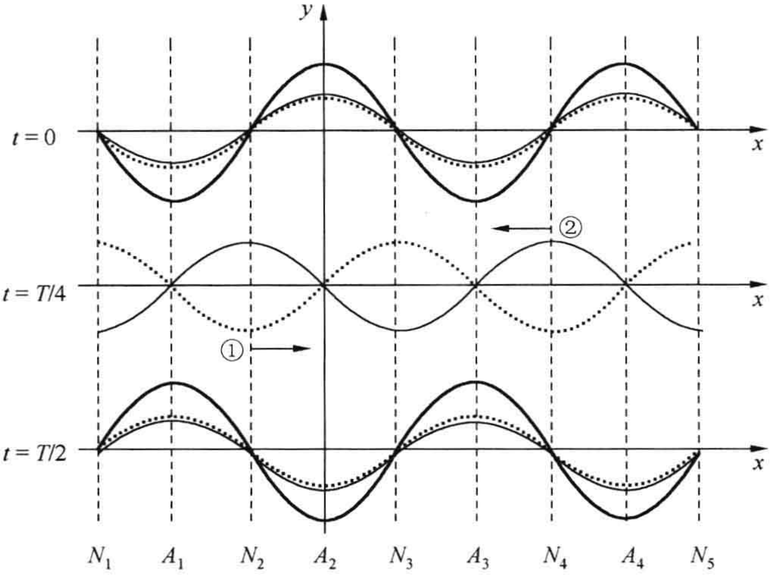


Fig. 13-22 Production of a standing wave

To understand the formation of a standing wave, we'd better consider Fig. 13-22. The dotted lines ①and ②represent incident wave and reflected wave respectively, and the solid line represents resultant wave. At time $t=0$, the incident wave, traveling to the right, is coincide with the reflected wave traveling to the left, and the resulting displacement is maximum. At time $t=T/4$, e. g. one fourth of a period later, the resulting displacement is zero. Therefore, the shape of the resultant wave is a straight line that coincides with the x-axis. At time $t=T/2$, e. g. after another quarter of a period, the resulting displacement is maximum again but in the opposite direction.

From Eq. (13-26), the amplitude has its maximum value of $2A$ at positions where $\left|\cos 2\pi \frac{x}{\lambda}\right| = 1$, or $2\pi \frac{x}{\lambda} = \pm n\pi$. We have known that these positions of maximum amplitude are called antinodes. Hence, the anti-nodal positions are

$$x = \pm \frac{1}{2} n\lambda, \qquad n = 0,1,2,\cdots \tag{13-27}$$

These positions are labeled $A_1, A_2, A_3, \cdots$ in Fig. 13-22.

Similarly, the amplitude has its minimum value of zero at positions where $\left|\cos 2\pi \frac{x}{\lambda}\right| = 0$ or $2\pi \frac{x}{\lambda} = \pm (2n+1)\frac{\pi}{2}$. We have known that elements located at these positions do not move and these positions of minimum amplitude are called nodes. Hence, the nodal positions are

$$x = \pm (2n+1)\frac{\lambda}{4}, \quad n = 0,1,2,\cdots \tag{13-28}$$

The nodal positions are labeled $N_1, N_2, N_3, \cdots$ in Fig. 13-22.

From Eq. (13-27) and Eq. (13-28), the distance between two adjacent antinodes or nodes is half-wavelength in a standing wave. We can obtain the wavelength of a standing wave by measuring above distance.

Let us consider the phases of different points in the resultant wave. From Eq. (13-26) and

Fig. (13-22), we can see that the phase of all particles between two adjacent nodes is the same. The particles in adjacent segments of length $\lambda/2$ have a phase difference of π radian. Dislike a traveling wave, the waveform of a standing wave does not advance. And no translation of energy occurs in a standing wave, but there is energy associated with the wave.

In Fig. 13-21, note that there is a node at point B, a fixed end. The reason is that the reflected wave is always 180°(or π) out of phase with the incident wave at a fixed boundary. We called this phenomenon as half-wave loss. This idea will be applied to light wave associated with the reflection from material of larger refractive index to that of smaller refractive index.

13.8 The Doppler Effect

In above discussions, we suppose that the wave source and observer are all at rest with respect to the medium in which wave is traveling. In this situation, the wave frequency is just the same as the vibrating frequency of the source. However, if the source (or the observer, or both) is in motion in the medium, the frequency received by the observer may differ from the vibrating frequency of the source. The most common example is that the pitch (frequency) of the whistle of a locomotive heard by a person (observer) is higher when the source is approaching to the hearer (observer) than when it is at rest with the hearer. An alteration in the observed frequency of a sound due to motion of either the source or the observer is known as the Doppler effect. It is named for the Austrian physicist Christian Johann Doppler (1803—1853), who did experiments with both moving sources and moving observers.

Of course, the frequency of the source remains unchanged as does the velocity of the wave. The essence of Doppler effect is that the time interval of two events measured in the source system is different from the time interval of these events received by the observer measured in his own system. Let us now discuss it in detail.

We now consider the application of the Doppler effect to sound waves, treating only the special case in which the source and observer move along the line joining them.

(1) Observer is moving and source is at rest. In Fig. 13-23, an observer O (represented by an ear) is moving toward a stationary source at speed v_0. The wave fronts are drawn to be one wavelength apart. Clearly, waves pass the observer at a faster rate when the observer is moving toward the source than they would if it were standing still, which accounts for the higher frequency that the person hears.

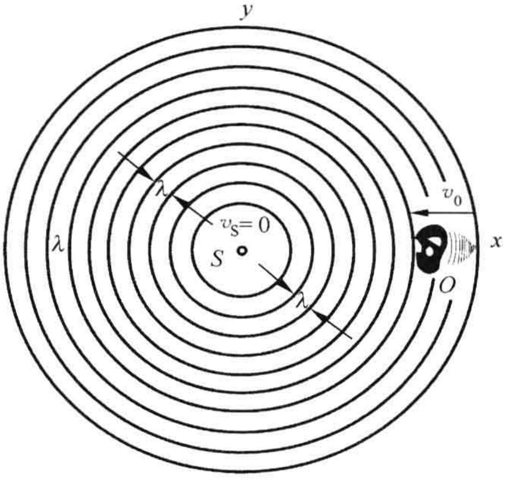


Fig. 13-23 The Doppler effect due to motion of the observer (ear). The source is at rest

Let us consider the observer at time t. If he were at rest, vt/λ waves would pass by at this time. Because of his motion, an additional $v_0 t/\lambda$ must be added to this. The frequency at which the observer receives waves—which is its perceived frequency—is then

$$\nu' = \frac{vt/\lambda + v_0 t/\lambda}{t} = \frac{v + v_0}{\lambda} = \frac{v + v_0}{v/\nu} = \nu \frac{v + v_0}{v}$$

For the general case of a resting source and a moving observer, we have

$$\nu' = \nu \frac{v \pm v_0}{v} \quad \text{(source at rest; observer moving)} \tag{13-29}$$

Here the plus sign indicates the observer moving toward the source and the minus sign indicates the observer moving away from the source.

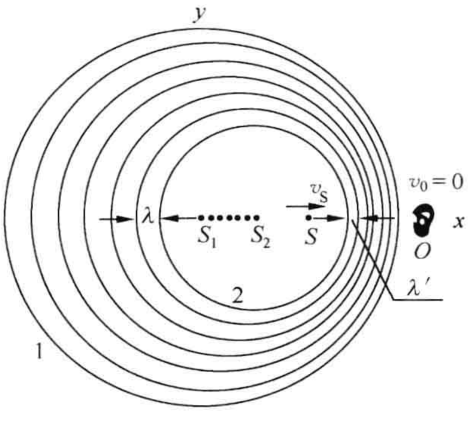


Fig. 13-24 The Doppler effect due to motion of the source. The observer is at rest

(2) Source is moving and the observer O is at rest. In Fig. 13-24, the source S is moving toward observer at speed v_S. Wave front 1 was emitted when the source was at S_1, wave front 2 was emitted when it was at S_2, and so on. At the moment depicted, the source was at S. The observer perceives a higher frequency because the moving source chases its own wave fronts, emits a reduced wavelength in its direction of motion. The effect here in Fig. 13-24 is $\lambda' < \lambda$.

Let $T(T = 1/\nu)$ be the period of oscillation of the source of sound. During this interval, the source moves toward the observer by a distance $v_S T$ or v_S/ν and the wavelength is shortened by that amount. Thus, the wavelength of the sound arriving at the observer is not v/ν but is the difference of $v/\nu - v_S/\nu$. The frequency heard by the observer is

$$\nu' = \frac{v}{\lambda'} = \frac{v}{v/\nu - v_S/\nu} = \nu \frac{v}{v - v_S}$$

For the general case of a moving source and a resting observer, we have

$$\nu' = \nu \frac{v}{v \mp v_S} \quad \text{(source moving; observer at rest)} \tag{13-30}$$

(3) Both source and observer are moving. According to above discussion, we can produce the general Doppler effect equation in which both the source and the observer are moving with respect to the medium. Replacing the ν (the frequency of the source) in Eq. (13-30) by the ν' (the frequency associated with motion of the moving observer) in Eq. (13-29) leads to

$$\nu' = \nu \frac{v \pm v_0}{v \mp v_S} \quad \text{(both of observer and source moving)} \tag{13-31}$$

Putting $v_S = 0$ in Eq. (13-31) reduces it to Eq. (13-29) and putting $v_0 = 0$ reduces it to Eq. (13-30), as we expect. There are four possible combinations of signs in Eq. (13-31).

Questions

13-1 What are the difference and connection, if any, between the following pair of concepts:

(1) mechanical wave and oscillation (or vibration);

(2) waveform(or wave pattern) and oscillation curve;

(3) wave speed and oscillation speed;

(4) the traveling direction of wave and the direction of vibration;

(5) the energy of wave and the energy of oscillation?

13-2 What is the difference between transverse and longitudinal waves? If you have a long, thin steel rod and a hammer, now must you hit the end of the rod to generate a longitudinal wave? a transverse wave?

13-3 Describe which quantity among the wavelength, period, frequency and wave speed will change if a wave travels from one medium into another?

13-4 The waveform of a transverse wave is shown in Fig. 13-25. Point out

(1) the moving directions of points a, b, c, and d;

(2) the phases of points O, a, and b;

(3) inverse the direction of wave speed, then answer question (1) and (2).

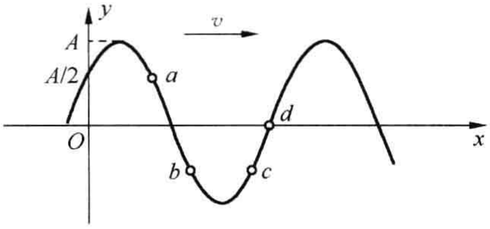


Fig. 13-25 For question 13-4

13-5 A harmonic wave is traveling along a string. Where are the maximum of kinetic energy in this wave? Where the maximum of potential energy? Where the maximum of total energy?

13-6 A plane harmonic wave from air falls on the interface between the air and the water as shown in Fig. 13-26. AB is the wave-front at t, the time taken for the wave traveling from B to C is τ, the speed of wave in the air is u_1, while in water is u_2, and $u_1 > u_2$. Draw the wave-front in water at the instant $(t+\tau)$ according to Huygens' principle.

13-7 Two coherent waves of λ are meeting at point P as shown in Fig. 13-27. The distance from source S_2 to P is $r_2 = 3\lambda$, and that from S_1 to P is $r_1 = 10\lambda/3$. What is the phase relation satisfied by the two sources to form a maximum resultant amplitude at P?

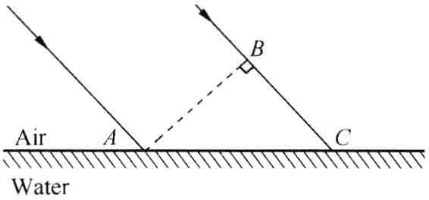


Fig. 13-26 For question 13-6

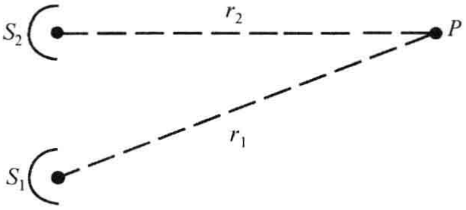


Fig. 13-27 For question 13-7

13-8 In your experimental physics course, using a "resonance tube", you measure the speed of a sound wave in air generated from a source arranged above the water filling in a tube. There will be a node formed at the interface between the air and the water in the tube when the sound wave reflects from the water surface to the air. Explain.

13-9 In tuning a guitar or violin, by what means do you change the frequency of a string?

13-10 What is the difference between a traveling wave in a string and a standing wave formed in a string?

13-11 There is a SHW $y_1 = 2.0 \times 10^{-2} \cos\left[2\pi\left(\frac{t}{0.02} - \frac{x}{20}\right) + \frac{\pi}{3}\right]$ (SI) formed on a long string. In order to form a standing wave on this string with a node at $x=0$, there must be another SHW of y_2 on the string. What should be the wave function of y_2 ?

13-12 In an earthquake of 19 September 1985, in the south area of North America, the areas with high damage alternated with areas of low damage. Also, buildings between 5 and 15 stories high sustained the most damage. Discuss these effects in terms of standing waves and resonance.

13-13 Have you ever observed the phenomenon that you feel the tune of the steam whistle from the train rising up when the train is approaching the platform where you are standing? If so, explain.

Problems

13-1 The speed of sound wave in air is 340 m/s, in water 1500 m/s, and in steel 5300 m/s. Find the wavelengths in the air, in the water and in the steel respectively for

(1) the sound wave with frequency of 600 Hz;

(2) the ultrasonic wave with frequency of 2×10^5 Hz.

13-2 A source of a SHW vibrates in y direction with period of 0.02s, suppose that the wave is traveling along positive x axis with speed of 100 m/s, and the source is passing the equilibrium position toward the positive y axis when $t=0$. Find

(1) the wave function;

(2) the vibration function for the point of $x=5$;

(3) the phase difference between the points of $x=16$ m and $x=17$ m.

13-3 A plane harmonic wave travels in negative direction of x axis, having a wavelength of 1 m, the period of 0.5 s, and the amplitude of 0.1 m. The point of $x=0$ is passing the equilibrium position toward the negative y direction when $t=0$. Find the wave function.

13-4 A traveling wave on a string is described by $y=0.2\sin\pi(0.5+x-2.5t)$ m. Find

(1) the wavelength, period, the wave speed;

(2) the oscillation equation at $x=1$ m;

(3) the displacement and velocity of $x=1$ m for $t=0.4$ s, when did the origin have the same phase as that of $x=1$ m having now; after $\Delta t=0.4$ s from this moment, where will this phase reach?

13-5 The function of a simple harmonic wave is given by $y=0.2\cos(10t+6x)$ m. Find

(1) the waveform equation for $t=6$ s,

(2) the positions of crests and troughs for $t=6$ s,

(3) what is the wave speed? How long will be the time taken for those crests to reach the origin?

13-6 A plane SHW is propagating along positive x direction with the wavelength λ, and the oscillation equation for $x_0=\lambda/4$ is known as $y_{x_0}=A\cos\omega t$ m.

(1) Write the wave function;

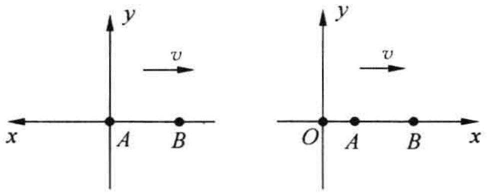
(2) Draw the waveform for $t=T$ and $t=5/4\,T$ respectively.

13-7 Fig. 13-28 shows that a wave is traveling from left to right direction with wave speed of 20 m/s. The oscillation function for point A is $y=3\cos(4\pi t-\pi)$ (SI), point B is 9 m apart from A on the right.

(1) Take the positive x direction to the left, write the wave function with the origin point at A, and find the oscillation function of point B.

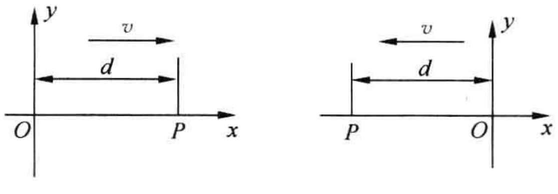
(2) Take the positive x direction to the right with origin at point O of 5 m apart from A on the left, find the answers of problem (1).

13-8 A plane cosinusoidal wave is propagating in the space, the vibration function of point P is $y=A\cos(\omega t+\varphi)$. Find the wave functions for the two different cases as shown in Fig. 13-29, in which d is the distance between point P and the origin and v is the wave speed.



(a) (b)

Fig. 13-28 For problem 13-7



(a) (b)

Fig. 13-29 For problem 13-8

13-9 A plane harmonic wave travels along positive x direction with wavelength of 4 m, period of 4s. The oscillation curve of point $x=0$ is shown in Fig. 13-30 where $y_0 = 1\times 10^{-2}$ m for $t=0$. Find

(1) Oscillation function for the point of $x=0$;

(2) Wave function;

(3) Draw the wave pattern at $t=1$s.

13-10 Solve Example 13-4 again by means of reference circle diagram and Eq. (13-11).

13-11 Fig. 13-31 shows the wave pattern of a cosinusoidal wave for $t=0$ as the point P moves downward. Suppose that the frequency of the source is 250 Hz. Find

(1) the wave function;

(2) the oscillation and velocity equations for the point of 100 m apart from the origin.

13-12 A transverse wave generated on a long string with mass density of 1.5×10^{-2} kg/m, stretched by a horizontal force of 6 N. Suppose that the wave propagates in negative x direction, the wave pattern at $t=0$ as shown in Fig. 13-32.

(1) Show the moving direction for points a, b, c, and d on the curve.

(2) Find the wave function.

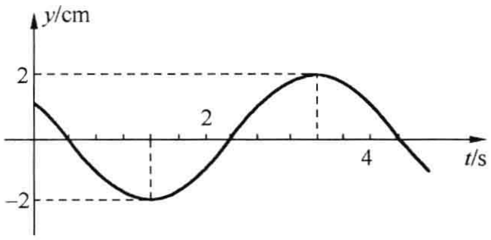


Fig. 13-30 For problem 13-9

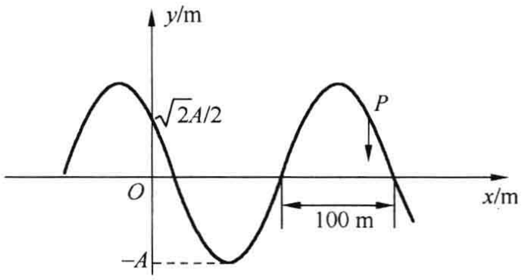


Fig. 13-31 For problem 13-11

13-13 In Fig. 13-33, the curve I is the wave pattern of a traveling wave at $t=0$; after $\Delta t=0.5$s, the pattern shifts a distance to the right in x axis, becoming the dashed curve II. Suppose the period $T \geqslant 1$ s. Find

(1) the wave function;

(2) the position function for point a.

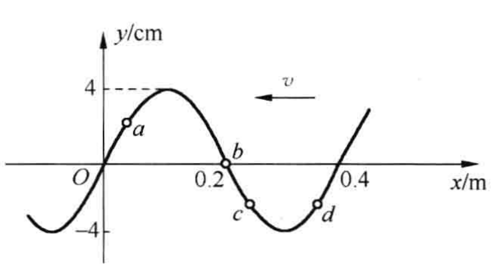


Fig. 13-32 For problem 13-12

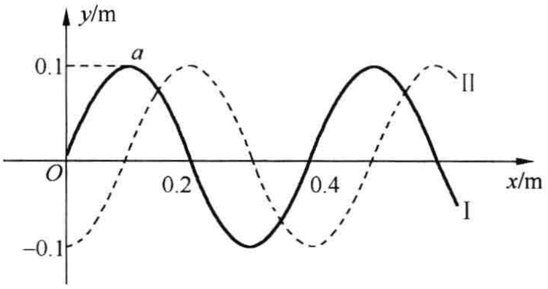


Fig. 13-33 For problem 13-13

13-14 A SHW travels in a medium, having the wave speed of 10^3 m/s, amplitude of 1.0×10^{-4} m, frequency of 10^3 Hz. Suppose that the mass density of the medium is 800 kg/m^3. Find

(1) the average energy flow density of wave;

(2) the energy transported through the area of 4×10^{-4} m^2 in one minute.

13-15 There is a spherical electromagnetic wave traveling in space. In order to keep the amplitude of the source unchanged, a power of 4W is supported. Find the intensity of the wave at the distance of 5 m and 10 m from the source. Supposing the electromagnetic wave speed is 3×10^8 m/s, calculate the corresponding average energy densities.

13-16 A wave travels in the air along a cylindrical tube with diameter of 0.14 m, the average intensity of the wave is 9×10^{-3} W/m^2, frequency is 300 Hz, the wave speed is 300 m/s. Find

(1) the average energy density and the maximum energy density of the wave;

(2) the energy contained between two adjacent wave surfaces apart by a wavelength.

13-17 Two simple harmonic wave sources A and B apart from each other by 20 m, vibrate in the same direction with the same frequency of 100 Hz and same amplitude, and the wave speed is 200 m/s. As a crest forms at A while a trough forms at B. Find the positions between A and B, whose amplitude are zero resulted from the interference.

13-18 In Figure 13-34, point O is a source vibrating in the direction perpendicular to the page, the wavelength is λ, $AO=h$, $Ox/\!/AB$, suppose the wave is reflected by AB plane without the phase shift. Find the x coordinates of the points with maximum resultant-amplitude on the x axis.

13-19 Two sources of plane harmonic wave, S_1 and S_2, vibrate in the same direction, same phase, with frequency of 30 Hz, the wave speed is 0.5 m/s. Find

(1) the phase difference at point P where the two waves meet, supposing $S_1P=3$ m, $\angle PS_1S_2=30°$, $S_1S_2=0.07$ m in Fig. 13-35;

(2) the initial phase of point P, if the two coherent waves have the same amplitude and zero initial phases.

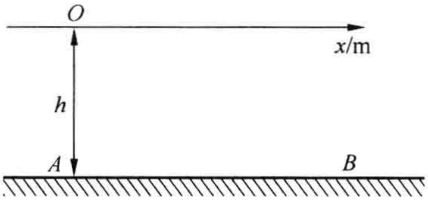


Fig. 13-34 For problem 13-18

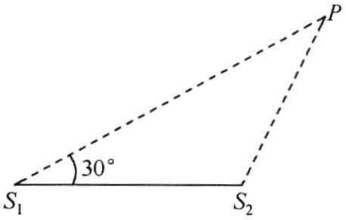


Fig. 13-35 For problem 13-19

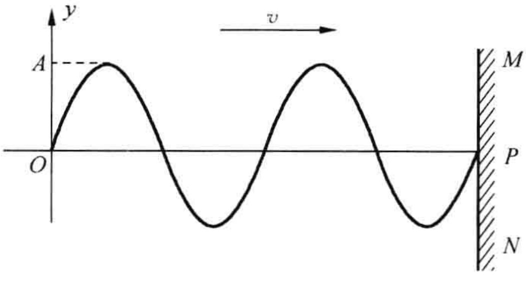


Fig. 13-36 For problem 13-21

13-20 S_1 and S_2 are two coherent sources, apart by $17/4\lambda$, with the same amplitude A. The vibration of S_1 is earlier by $\pi/2$ in phase than that of S_2. Find the amplitude distribution on the line connecting S_1 and S_2:

(1) outside S_1;

(2) outside S_2;

(3) between S_1 and S_2.

13-21 A simple harmonic wave travels to the right direction and the wave forms at time t as shown in Fig. 13-36 where the incident wave is reflected at point P with phase shift of π. Draw the reflecting waveform at time t.

13-22 The incident wave function along a string is $y_1=A\cos\left(\omega t-\frac{2\pi x}{\lambda}\right)$ (SI). Suppose it reflects freely from point B apart from origin by L on the right. Find the reflecting wave function y_2, if the amplitude remains unchanged.

13-23 An SHW propagates with the speed of 1 m/s, its function is $y_1=0.05\cos\pi\left(\frac{x}{v}+t\right)$ m, reflected from $x=0$, the reflecting wave function is $y_2=0.05\cos\pi\left(\frac{x}{v}-t\right)$ m. Find

(1) the resultant wave function;

(2) the positions of the nodes and antinodes;

(3) the amplitude for $x=1.2$ m.

13-24 The function of a standing wave on a string is known as $y = 0.08\sin 2\pi x\cos 50\pi t$ where x, y are in meter, t in second. Find the wavelength, frequency, speed and amplitude of the two waves propagating in opposite direction, forming the standing wave.

13-25 Suppose that you are at rest on the platform of a train station, a train approaches toward you at 33 m/s sounding its steam whistle which has frequency of 1000 Hz.

(1) What frequency do you hear?

(2) What is the answer if the train is going away from you at the same speed (Suppose the sound speed is 340 m/s in the air)?

13-26 In problem 13-25.

(1) If you drive a car on a road parallel to the railway toward the approaching train with the same speed as the train , what frequency do you hear?

(2) If you and the train are driving away from each other, what is the answer? The speed of the train and the sound speed are the same as in problem 13-25.

13-27 A man on a railroad car blows a trumpet sounding at 440 Hz. The car is moving toward a wall at 20 m/s. Calculate

(1) the frequency of the sound as received at the wall;

(2) the frequency of the reflected sound arriving back at the source.

Suppose that the sound speed in the air is 340 m/s.

Chapter14

Electromagnetic Oscillations and Waves

James Clerk Maxwell first formally postulated electromagnetic waves in 1865. He demonstrated that electric and magnetic fields travel through space in the form of waves, and at the constant speed of light. In 1887, Heinrich Hertz produced electromagnetic wave with oscillating circuits. Electromagnetic radiation is classified according to the frequency of its wave. The electromagnetic spectrum, in order of increasing frequency and decreasing wavelength, consists of radio waves, microwaves, infrared radiation, visible light, ultraviolet radiation, X-rays and gamma rays. Electromagnetic waves in the form of visible light enable us to view the world around us, and radio-frequency waves carry our favorite television and radio program. In this chapter, to understand the origin of the electromagnetic waves sent out by an oscillating current, we discuss the electromagnetic oscillation of *LC* circuit in detail and the generation of electromagnetic waves. Further, we summarize the properties of electromagnetic waves.

14.1 *LC* Oscillations

14.1.1 *LC* Circuit

An *LC* circuit, also called a resonant circuit, consists of an inductor L and a capacitor C, but with no resistance as shown in Fig. 14-1. When the switch K is closed in the position A, the capacitor is charged to a potential difference and with a charge of Q, and the electric field exists between two plates of capacitor so that the electric energy is stored in the capacitor. Then the switch K is flipped to the position B, and the capacitor immediately starts to discharge through the inductor. Therefore a current is established and is varying with time due to the self-induced emf in the inductor so that the magnetic field appears around the inductor and the magnetic energy is stored in the inductor. It will be proved that the charge in the capacitor and current in *LC* circuit vary sinusoidally. The resulting oscillations of the electric field and magnetic field in *LC* circuit are said to be electromagnetic oscillations. Such a circuit is said to oscillate. Steps (a) through (f) of Fig. 14-2 show succeeding stages of the oscillations in a period T after the switch K is flipped to the position B in the *LC* circuit as shown in Fig. 14-2.

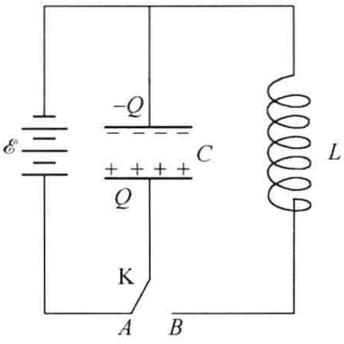
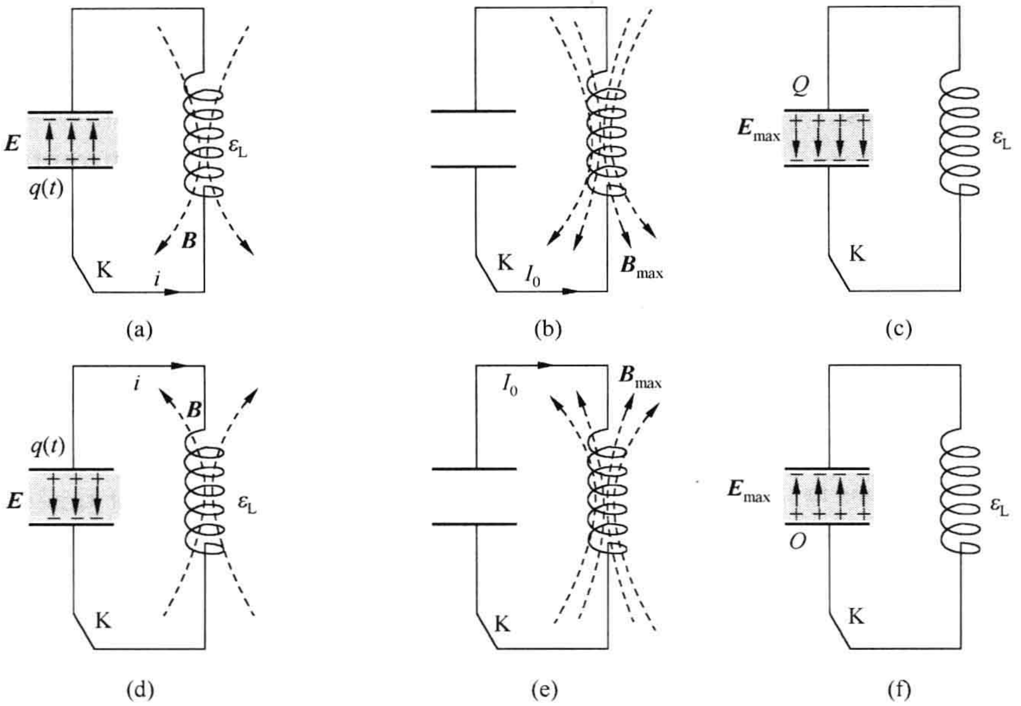


Fig. 14-1 An *LC* circuit



(a) (b) (c)

(d) (e) (f)

Fig. 14-2 Oscillations in *LC* circuit during a period *T*

Let us discuss the process above from the energy point of view. In Fig. 14-2(a), the capacitor with its maximum charge Q and maximum electric energy given by Eq. (7-24)

$$W_e = \frac{Q^2}{2C}$$

starts to discharge through the inductor, positive charge carriers moving counterclockwise (really, the electron are moving clockwise) and a current i is formed so that a magnetic field appears in the space around inductor. As the charge q and the electric energy in the capacitor

$$W_e(t) = \frac{q^2}{2C} \tag{14-1}$$

decrease, the current i in the circuit and the magnetic energy (see Eq. (9-29))

$$W_m(t) = \frac{1}{2}Li^2 \tag{14-2}$$

increase until all the charge on the plates of the capacitor has gone and the magnetic energy reaches its maximum

$$W_m = \frac{1}{2}Li_{\max}^2 = \frac{1}{2}LI_0^2$$

where I_0 is the current maximum $i_{\max}$ as shown in Fig. 14-2(b). This implies that the electric energy has been fully transferred into the magnetic energy, that is the energy has been transferred from the electric field to the magnetic field.

The current in Fig. 14-2(b) must continue because the inductor does not allow it to change suddenly to zero, and will begin to charge the capacitor with a voltage of opposite polarity. However the current in the counterclockwise direction decreases, and eventually has decreased to zero (momentarily) as shown in Fig. 14-2(c). Here, the capacitor has been charged to its original charge with opposite polarity and the energy has been transferred completely back to the capacitor. Then the cycle will begin again, with the current flowing in the clockwise direction through the inductor, until the circuit eventually return to its intial situation as shown from Fig. 14-2(d), (e) to (f). The process then repeats periodically.

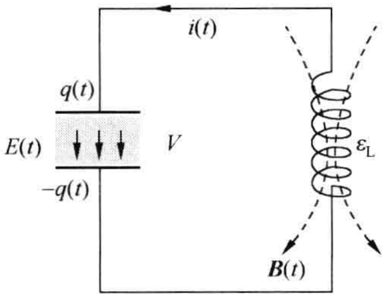


Fig. 14-3 Oscillations in LC circuit at time t

In a word, oscillation occurs in the current in the circuit and in the charge on the capacitor. If the resistance is neglected, no energy is dissipated, and the oscillations continue indefinitely. The charge flows back and forth between the plates of the capacitor through the inductor, and the electric energy in the capacitor and the magnetic energy in the inductor vary with time. Fig. 14-3 shows the situation of LC circuit at time t and the conservation of energy means that the total energy at time t is

$$\frac{q^2}{2C}+\frac{1}{2}Li^2=\frac{Q^2}{2C}=\frac{1}{2}LI_0^2 \tag{14-3}$$

Note that the maximum current and the maximum charge do not occur at the same time.

14.1.2 *LC* oscillating equation

At time t, the potential difference V between two plates of capacitor is equal to self-induced emf ε_L in Fig. 14-3. That is

$$V=\frac{q}{C}=-L\frac{\mathrm{d}i}{\mathrm{d}t} \quad \text{or} \quad L\frac{\mathrm{d}i}{\mathrm{d}t}+\frac{q}{C}=0$$

Considering $i=\mathrm{d}q/\mathrm{d}t$ and $\mathrm{d}i/\mathrm{d}t=\mathrm{d}^2q/\mathrm{d}t^2$, we have

$$\frac{\mathrm{d}^2q}{\mathrm{d}t^2}+\frac{q}{LC}=0 \tag{14-4}$$

This is the differential equation that describes the oscillation of an (resistanceless) LC circuit. Eq. (14-4) is mathematically identical in the equation of the block-spring oscillator. Because q corresponds to x, the general solution of Eq. (14-4) can be written as

$$q=Q\cos(\omega t+\varphi) \tag{14-5}$$

where Q is the amplitude of the charge variations and ω is the angular frequency of the electromagnetic oscillations

$$\omega=\frac{1}{\sqrt{LC}} \tag{14-6}$$

Therefore, the frequency of oscillations is given by

$$\nu=\frac{1}{2\pi\sqrt{LC}} \tag{14-7}$$

The phase constant in Eq. (14-5) is determined by the initial conditions of the charge q_0 and the current i_0 at $t=0$. The current of LC oscillator is

$$i=\frac{\mathrm{d}q}{\mathrm{d}t}=-\omega Q\sin(\omega t+\varphi)=I_0\sin(\omega t+\varphi') \tag{14-8}$$

The stored electric and magnetic energies in the LC circuit, using Eq. (14-1) and Eq. (14-2), are

$$W_e(t)=\frac{Q^2}{2C}\cos^2(\omega t+\varphi) \tag{14-9}$$

$$W_m(t)=\frac{Q^2}{2C}\sin^2(\omega t+\varphi) \tag{14-10}$$

At any instant, the total energy is

$$W_e(t)+W_m(t)=\frac{Q^2}{2C}\cos^2(\omega t+\varphi)+\frac{Q^2}{2C}\sin^2(\omega t+\varphi)=\frac{Q^2}{2C}=\frac{1}{2}LI_0{}^2$$

which is the conservation of energy indicated in Eq. (14-3).

One can continue the discussion on the LC circuit in analogy of the simple harmonic oscillator. In an actual LC circuit, the oscillations will not continue infinitely because there is always some resistance present that will drain energy from the electric and magnetic field and dissipate it as thermal energy as shown in Fig. 14-4 (a). The RLC circuit is called to be damped. For the case of small R, RLC circuit is underdamped and oscillate at angular frequency $1/\sqrt{LC}$ with an exponentially decaying amplitude $Qe^{-Rt/2L}$. Therefore, the oscillations, once started, will die away eventually, and all the energy originally stored in the capacitor is dissipated as the thermal energy. The angular frequency $\omega=1/\sqrt{LC}$ is called the circuit's natural angular frequency. When the external alternating emf of

$$\varepsilon_{ex}=\varepsilon_0\cos\omega_d t \tag{14-11}$$

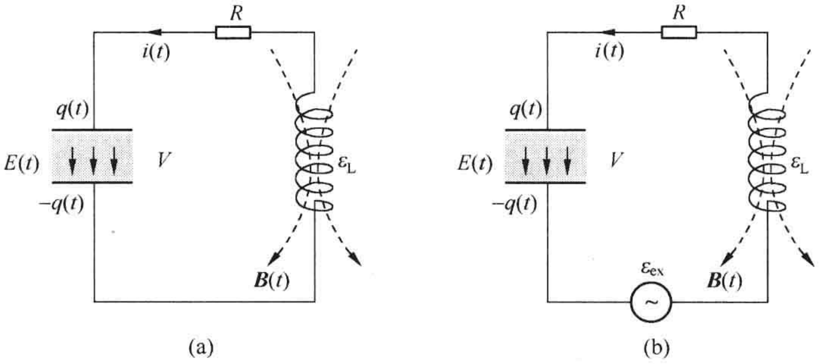


Fig. 14-4 (a) Damped oscillations and (b) forced oscillations in RLC circuit

is connected to an RLC circuit and provide energy to the circuit, the oscillations will continue and are said to be driven ocsillations or forced oscillations as shown in Fig. 14-4(b). It can be proved that whatever the natural angular frequency ω may be, the forced oscillations in circuit always occur at the driving angular frequency ω_d, and when ω_d approaches to the natural angular frequency ω

$$\omega_d=\frac{1}{\sqrt{LC}} \tag{14-12}$$

the resonance will occur after the alternating emf has been applied for some time. That is the amplitude of the current i in the circuit is maximum.

14. 2 Electromagnetic Waves

14. 2. 1 Production of electromagnetic waves by an electric dipole

The LC oscillator has been discussed in the last section, in which we assumed that energy is conserved but the thermal loses in the circuit. In fact, other energy losses have existed due to the time-varying electric field in the capacitor and inductor, that is the electromagnetic radiation we will discuss in this section. Can the above LC circuit produce the electromagnetic wave? And if so, can the electromagnetic wave be sent out? The electric and magnetic fields in the LC circuit have a low frequency and are localized in the capacitor and the induction coil so that the varying fields with time are hardly possible to be sent out. To generate the electromagnetic wave, one can alternate the discussed LC circuit as shown in Fig. 14-5 to be a line eventually, equivalent to an electric dipole, usually called an antenna. Obviously, the electromagnet field produced by an antenna is dispersed in an open space. Because L and C of an antenna are quite small, the frequency of oscillation is quite high. Therefore, it can produce electromagnetic wave and is called a transmitter.

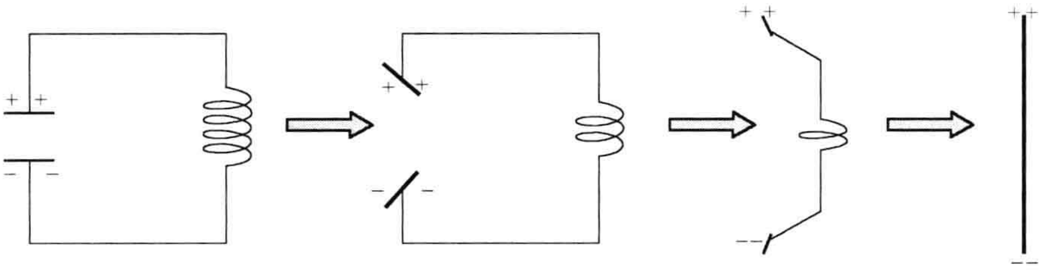


Fig. 14-5 The LC circuit is alternated gradually to be an antenna

Fig. 14-6 shows, in broad outline, a generator of electromagnetic wave. At its heart is a LC oscillator, which establishes an angular frequency $\omega = 1/\sqrt{LC}$. Charges and currents in this circuit vary sinusoidally at this frequency, as depicted in Eq. (14-5) and Eq. (14-8). An external source supples the energy needed to compensate both for thermal losses in the circuit and for energy carried away by the radiated electromagnetic wave. The LC oscillator is coupled by a transformer and a transmission line to feed the antenna (an air gap between two metal spheres). The current varying sinusoidally with time in the LC oscillator forces the charge in the antenna to oscillate (discharge). That is an electric dipole moment which varies sinusoidally with time. So, the electromagnetic wave is produced by oscillating elecrtic dipole in an antenna.

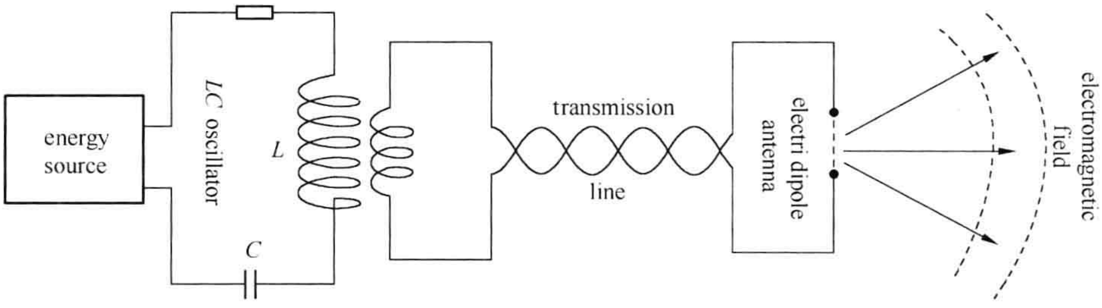


Fig. 14-6 Schematic diagram of apparatus for generating an electromagnetc wave

The electromagnet radiation from an oscillating electric diploe is complicated in the region near the antenna. The radiation is not a plane wave, but it travels out in all directions from the source. At points far from the antenna, the electromagnetic wave can be considered to be a plane wave. Let the dipole moment vary with time being $\boldsymbol{p} = \boldsymbol{p}_0 \cos\omega t$ where $\boldsymbol{p}_0$ is the amplitude of the dipole moment variation, then a varied electric and magnetic field with time is generated around the dipole by the accelerating charges in the dipole antenna. For the radiation fields far from the dipole as shown in Fig. 14-7, it can be proved that the electric and magnetic fields are given by

$$\begin{cases} E(r,t) = \dfrac{\omega^2 p_0 \sin\theta}{4\pi\varepsilon v^2 r}\cos\omega\left(t-\dfrac{r}{v}\right) \\ H(r,t) = \dfrac{\omega^2 p_0 \sin\theta}{4\pi v r}\cos\omega\left(t-\dfrac{r}{v}\right) \end{cases} \tag{14-13}$$

where r is the distance of the field point from the center of the dipole, and θ is the angle between the vector $\boldsymbol{r}$ and dipole moment $\boldsymbol{p}$, and

$$v = \frac{1}{\sqrt{\varepsilon\mu}} \tag{14-14}$$

which is the speed of electromagnetic wave.

By Eq. (14-13) we know that intensity of the traveling wave in any direction is proportional to $\sin^2\theta$, being zero in the direction of the axis of the dipole ($\theta=0°$or $180°$) and a maximum in the equatorial plane of the dipole ($\theta=90°$). Moreover, the amplitude of the electromagnetic wave at points far from the antenna are proportional inversely to the distance from the antenna.

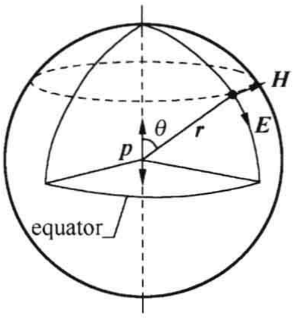


Fig. 14-7 Diagram of the coordinate system

14. 2. 2 The properties of the electromagnetic wave

We have already discussed some properties of electromagnetic wave produced by an antenna. Maxwell's equations predict the existence of electromagnetic waves that can propagate through free space and travel with a speed equal to the value of speed of light. The simplest such wave is a plane wave, in which $\boldsymbol{E}$ and $\boldsymbol{B}$ are uniform over any plane perpendicular to the propagation direction. Now we take a plane wave as an example and summarize what we know about electromagnetic wave thus far.

A plane electromagnetic wave is a wave traveling from a very distance source and can be expressed as

$$\begin{cases} \boldsymbol{E}(x,t) = \boldsymbol{E}_0\cos\omega\left(t-\dfrac{x}{v}\right) \\ \boldsymbol{H}(x,t) = \boldsymbol{H}_0\cos\omega\left(t-\dfrac{x}{v}\right) \end{cases} \tag{14-15}$$

Here, the electromagnetic wave propagates in a direction of the x axis, and the oscillations of the electric and magnetic fields take place in planes perpendicular to the x axis. The picture of the electromagnetic wave at a given instant of time is shown in Fig. 14-8, in which the electric field $\boldsymbol{E}$ is in the y direction and the magnetic field $\boldsymbol{H}$ is in the z direction.

The basic properties of the electromagnetic wave traveling through free space are summarized briefly as follow.

(1) The electric and magnetic fields are perpendicular to the direction of propagation of the wave and to each other. That is, the electromagnetic waves are transverse waves. The direction of the cross product $\boldsymbol{E}\times\boldsymbol{H}$ is the direction of propagation of electromagnetic waves.

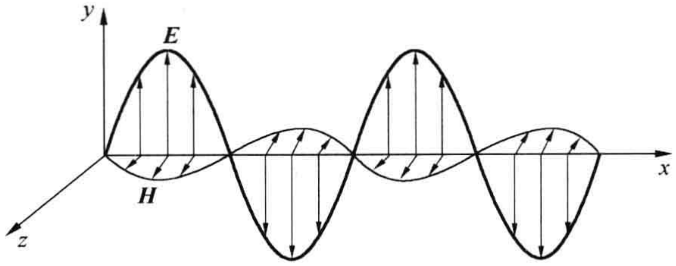


Fig. 14-8 An plane electromagnetic wave traveling along the x axis

(2) The electric and magnetic fields are in phase. This implies that at any point, the two fields reach their maximum values and zero at the same instant.

(3)The values of the electric and the magnetic fields in an electromagnetic wave at any time are related by

$$\frac{E}{B}=\frac{1}{\sqrt{\varepsilon\mu}}=v \tag{14-16}$$

or

$$\sqrt{\varepsilon}E=\sqrt{\mu}H \tag{14-17}$$

which are validated to be correct by Eq. 14-13 for the electromagnetic wave produced by an antenna.

(4) The speed of electromagnetic wave is given by

$$v=\frac{1}{\sqrt{\varepsilon\mu}}$$

which is determined by the permittivity and permeability of the medium and independent of the reference frames and observers. For free space (vacuum), the speed, usually denoted by c, is equal to

$$c=\frac{1}{\sqrt{\varepsilon_0\mu_0}}=299792458\ \mathrm{m/s}\approx 3\times 10^8\ \mathrm{m/s}$$

which is precisely the same as the speed of light in vacuum. One is led to believe that light is an electromagnetic wave.

The emergence of the speed of light from purely electromagnetic consideration is the crowning achievement of Maxwell's electromagnetic theory. Maxwell made this prediction before radio wave was known and before it was realized that light was electromagnetic in nature. His prediction led to the concept of the electromagnetic spectrum and to the discovery of radio wave by Heinrich Hertz in 1890. It made it possible to discuss wave optics as a branch of electriomagnetic waves.

In the following chapter, we will realize that the electromagnetic waves such as light exhibit wave properties: interference, diffraction, reflection, refraction, and polarization.

14. 3 Energy in Electromagnetic Waves

One of the important characteristics of electromagnetic waves is that they carry energy, and as they travel through space, they can transfer energy to objects placed in their paths. The energy density w in a region of space where $\boldsymbol{E}$ and $\boldsymbol{H}$ field are present is given by

$$w = \frac{1}{2}\varepsilon E^2 + \frac{1}{2}\mu H^2 \tag{14-18}$$

The wave energy transfer is conveniently characterized by considering the energy transferred per unit time, per unit cross-sectional area perpendicular to the direction of wave propagation. That is the energy flux density of an electromagnetic wave, denoted by **S**.

To see how the energy flow is related to the field, we consider a plane wave traveling along the x-axis as shown in Fig. 14-9. In a time dt after this, the wave front moves a distance v dt to the right. Considering a cuboid with an area dA perpendicular to the x axis and a length of vdt, we note that the energy in the cuboid have passed through the area dA in the interval dt. The total energy dW in the cuboid is

$$\mathrm{d}W = wv\,\mathrm{d}t\mathrm{d}A \tag{14-19}$$

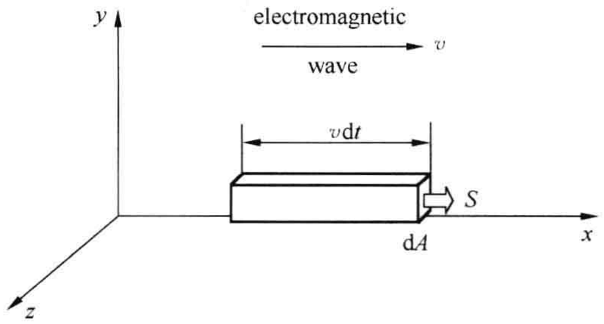


Fig. 14-9 The energy flow density S

Hence the energy flow density S is equal to

$$S = \frac{\mathrm{d}W}{\mathrm{d}t\mathrm{d}A} = wv \tag{14-20}$$

Using Eq. (14-14), Eq. (14-17) and Eq. (14-18), we obtain

$$S = EH \tag{14-21}$$

To describe both the magnitude and direction of the energy flow rate, we define a vector **S**

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H} \tag{14-22}$$

S is called the Poynting vector and its SI units are W/m^2. The magnitude S of the Poynting vector is the rate at which energy is transported by a wave across a unit area at any instant and the direction of **S** at any point gives the direction of energy transport at that point because **E** and **H** are perpendicular to each other and the direction of cross product **E** × **H** is the direction of propagation of an electromagnetic wave.

In practice, we are more interested in the average value of the time-varying quantity S, which is the intensity I of the wave. From Eq. (14-15) and Eq. (14-21), we can then write

$$I = \bar{S} = E_0 H_0 \overline{\cos^2\left(\omega t - \frac{x}{v}\right)} = \frac{1}{2} E_0 H_0 \tag{14-23}$$

By using Eq. (14-16), Eq. (14-23) can be written as

$$I = \frac{1}{2\mu} E_0 B_0 = \frac{1}{2\mu v} E_0^2 = \frac{v}{2\mu} B_0^2 \tag{14-24}$$

A detailed analysis would show that the energy carried by an electromagnetic wave is shared equally by the electric and magnetic fields.

Substituting Eq. (14-13) into Eq. (14-23), the intensity radiated by electric dipole is given by

$$S = \frac{\omega^4 p_0^2 \sin^2\theta}{32\pi^2 \varepsilon v^3 r^2} \tag{14-25}$$

We can calculate the radiation power of an antenna by integrating on a spherical surface of radius r, that is

$$\bar{P} = \int_0^{2\pi}\int_0^{\pi} \bar{S} r^2 \sin^2\theta \mathrm{d}\theta \mathrm{d}\varphi = \frac{\omega^4 p_0^2}{12\pi\varepsilon v^3} \tag{14-26}$$

Electromagnetic waves carry linear momentum as well as energy. Hence it follows that pressure is exerted on a surface when an electromagnetic wave strikes it and the momentum is absorbed by the surface. Although radiation pressures are very small, they have been measured with a device.

Example 14-1 The sun delivers about 10^3 W/m^2 of energy to the Earth's surface via electromagnetic radiation. Assume that the average magnitude of the Poynting vector for solar radiation at the surface of the Earth is $S_{av} = 1000$ W/m^2. Calculate the total power that is incident normal to a flat plate collector of dimensions 1.00 m×2.00 m of solar water heater.

Solution The average total power is

$$P_{av} = A \times S_{av} = 1.00 \times 2.00 \text{ m}^2 \times 1000 \text{ W/m}^2 = 2000 \text{ W}$$

If this energy were transformed into electrical energy which is stored, it would provide more than enough energy for the average home. However, solar energy is not easily harnessed.

14.4 The Electromagnetic Wave Spectrum

In Maxwells day, visible light and the adjoining infrared and ultraviolet radiation were the only electromagnetic radiation known. Spurred on by Maxwell's predictions, however, Heinrich Hertz discovered what we now call radio wave and verified that they move through the laboratory at the same speed as visible light. It is for this achievement that a Hertz is commemorated by using his name as the SI unit of frequency. It is now known that other forms of electromagnetic waves exist that are distinguished by their frequencies and wavelengths. Everyone is surrounded by the electromagnetic waves from the different sources: the Sun, stars in sky, the galaxies, TV stations, electric bulb, ..., all things around us.

Because all the electromagnetic waves have a same speed c through free space, their frequency and wavelength are related by

$$c = \nu\lambda \tag{14-27}$$

In Fig. 14-10, the various types of electromagnetic waves are listed in terms of frequency as well as wavelength, which shows the electromagnetic spectrum. The names given to the types of waves are simply for convenience in describing the region of the spectrum in which they lie. 1 nm is equal to 10^{-9} m and the wavelengths of visible light range from 400 nm to 760 nm. Note that all forms of radiation are generated by accelerating charges. There are no gaps in the electromagnetic spectrum. We stress that all electromagnetic waves, no matter where they lie in the spectrum, travel through free space with the same speed c.

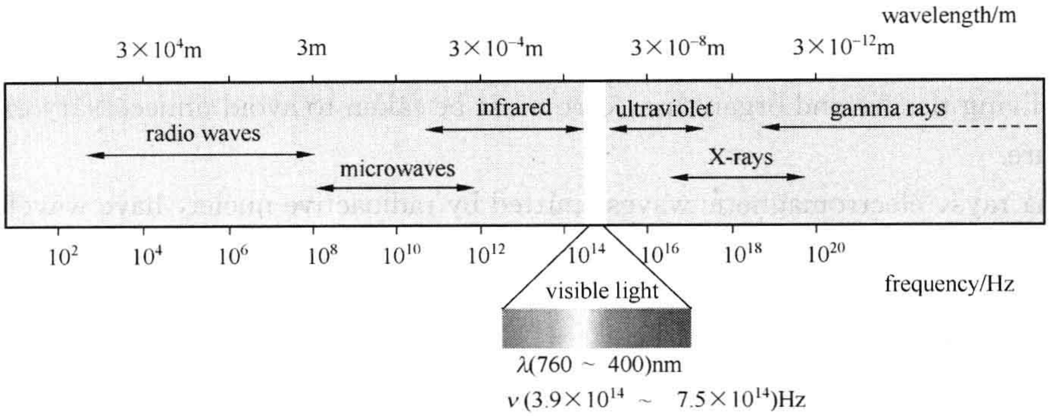


Fig. 14-10 The electromagnetic spectrum

Fig. 14-10 shows an entire spectrum of electromagnetic waves that we have now know. Radio waves are the result of charges accelerating through conducing wires, which are used in mobile radio communication, broadcasting, radar and other navigation systems, satellite communication, computer networks and innumerable other applications. For example, radio and TV bring the world into our home in sight and sound.

Microwaves are electromagnetic waves with wavelengths ranging between about 1 mm and 30 cm, which are produced by electronic devices and well suited for the radar systems in aircraft navigations. Microwaves are also used in communication, navigation, radio astronomy, spectroscopy, and heating in microwave oven.

Infrared waves, radiated by hot objects and molecules, have wavelength ranging from 1 mm to the longest wavelength of visible light, including most of the thermal radiation emitted by objects near room temperature. Infrared radiation is popularly known as “heat” or sometimes known as “heat radiation”. They are readily absorbed by most materials. The infrared energy absorbed by a substance causes it to get warmer. Infrared radiation has many practical and scientific applications, including physical therapy, infrared photography (such as night vision), thermography, the study of the vibrations of atoms and so on. Sunlight at zenith provides an irradiance of just over 1 kilowatt per square meter at sea level. Of this energy, 527 watts is infrared radiation, 445 watts is visible light, and 32 watts is ultraviolet radiation.

Visible light, the most familiar form of electromagnetic waves, may be defined as the part of the spectrum that is detected by the human eye. The wavelengths of visible are classified as colors ranging from violet ($\lambda \sim 400$ nm) to red ($\lambda \sim 760$ nm). The eye’s sensitivity is a function of wavelength and is greatest at a wavelength of about 560 nm.

Ultraviolet(UV) light is electromagnetic radiation with a wavelength shorter than that of visible light, but longer than X-rays, in the range 0. 6 nm to 400 nm. The Sun is an important source of ultraviolet light, most of which is absorbed by atoms in the upper atmosphere, or stratosphere. This is fortunate, because UV light in large quantities has harmful effects on humans. Although ultraviolet radiation is invisible to the human eye, most people are aware of the effects of UV through sunburn. UV light is widely used in everyday life.

X-rays are electromagnetic waves with wavelengths from about 10 nm down to 10^{-4} nm.

The most common source of X-rays is the acceleration of high-energy electrons bombarding a metal target. X-rays are widely used in medicine. Because X-rays easily penetrate and damage or destroy living tissues and organisms, care must be taken to avoid unnecessary exposure and overexposure.

Gamma rays, electromagnetic waves emitted by radioactive nuclei, have wavelengths ranging from about 0.1 nm to less than 10^{-5} nm. They are highly penetrating and cause serious damage when absorbed by living tissues.

In a word, we are submerging in electromagnetic waves from light bulbs, from TV and radio stations, from the heated engine blocks of automobiles, from X-ray machines, from lightning flashes, from γ-radiation, from the radioactive materials in the earth. Beyond this, radiation reaches us from stars and other objects in our galaxy and from other galaxies.

Questions

14-1 If all charges on the capacitor in a *LC* circuit have discharged, is the current equal zero in the *LC* circuit?

14-2 Compare the cases varying with time between the current in the inductor and the charges on the capacitor for the resistanceless *LC* circuit.

14-3 Why do the energy losses exit in a resistanceless *LC* circuit?

14-4 As the radiation field is farther from the dipole, what property have the electric and magnetic fields?

14-5 Why can we say that light is an electromagnetic wave?

14-6 If we have known the electric field in the vacuum for an electromagnetic wave, can we calculate the magnetic field in the same region?

Problems

14-1 A 1.5 μF capacitor is charged to 57 V. The charging battery is then disconnected and a 12 mH coil is connected across the capacitor, so that *LC* oscillations occur. What is the maximum current in the coil? (Assume that the circuit contains no resistance.)

14-2 A 1.5 mH inductor in an *LC* circuit stores a maximum energy of 17 μJ. What is the peak current I_m?

14-3 In an oscillating *LC* circuit, how much time is required for the energy of electric field varied from zero to equal that of the magnetic field? Assume that $L=12$ mH and $C=1.7$ μF.

14-4 An *LC* oscillating circuit consists of a 1.2 mH inductor and a 0.03 μF capacitor. Find the oscillating frequency. (Assume that the circuit contains no resistance.)

14-5 A plane electromagnetic wave has a maximum electric field of 3.2×10^{-4} V/m. Find the maximum magnetic field.

14-6 The electric field associated with a plane electromagnetic wave is given by $E_x = 0$, E_y, $E_z = 2.0\cos[\pi\times10^{15}(t-x/c)]$, with $c=3\times10^8$ m/s and all quantities in SI unit. The wave is propagating in the $+x$ direction. Write expressions for components of the magnetic field of the wave.

14-7 An electromagnetic wave is traveling in the negative y direction. At a particular position and time, the electric field is along the positive z axis and has a magnetic of 100 V/m. What are the direction and magnitude of the magnetic field at that position and at that time?

14-8 what is the average intensity of a plane traveling electromagnetic wave if B_0, the maximum value of its magnetic field component, is 1.0×10^{-4} T?

14-9 Assume that 10% of the power input of a 100W lamp is radiated uniformly as light of wavelength 500 nm. At a distance of 2.00 m from the source, the electric and magnetic intensities vary sinusoidally according to the equations $E=E_0\sin(\omega t+\varphi)$ and $H=H_0\sin(\omega t+\varphi)$. Calculate E and H.

14-10 Sunlight strikes the earth outside its atmosphere with an intensity of 1.40 kW/m^2. Calculate E_0 and B_0 for sunlight, assuming the electromagnetic field is plane harmonic wave.

14-11 You walk 150 m directly toward a street lamp and find that the intensity increases to 1.5 times the intensity at your original position. How far from the lamp were you first standing(the lamp radiates uniformly in all directions)?

14-12 An oscillating dipole radiates electromagnetic wave. Assume the amplitude of dipole moment $p_0=2.26\times10^{-4}$ cm, and frequency $f=800$ Hz. Find

(1) the wavelength of the electromagnetic wave;

(2) the electric and magnetic field intensity on the dipole equatorial of 200 km distance from the dipole.

Chapter15

Optics

Optics is the branch of physics which deals with the behavior and properties of light, including its interactions with matter and the construction of instruments. In general, the term "light" in physics refers to electromagnetic radiation of any wavelength, whether visible or not. In this chapter, "light" is visible light that is visible to the human eye and electromagnetic radiation with a wavelength in a range from about 380 nm to about 740 nm, with a frequency range of about 405 THz to 790 THz.

One of our principal contacts with the world around us is through light. We personally depend on light to convey visual information and have many instruments which use light to operate such as microscope, telescope and spectrometer. Most of what we know about the Universe comes from information that has been carried to us by light. Our knowledge on the structure of atoms comes largely from observing the radiations they emit. Therefore, light is very important to us, and nothing is visible to humans when light is totally absent.

The history of elucidating the nature of light can be traced back to the beginning of human being. In ancient China, India and Greek, "light" is one of the fundamental "subtle" elements such as Water, Wood, Metal, Fire, and Earth which compose of everything (that should not be confused with the ordinary meaning of these terms). In about 300 BC, Euclid studied the properties of light. Euclid postulated that light travelled in straight lines and he described the laws of reflection and studied them mathematically. René Descartes (1596—1650) held that light was a mechanical property of the luminous body and assumed that light behaved like a wave.

Isaac Newton stated in 1675 that light was composed of corpuscles (particles of matter) which were emitted in all directions from a source. Newton's theory could be used to predict the reflection of light, but could only explain refraction by incorrectly assuming that light accelerated upon entering a denser medium because the gravitational pull was greater. Although there were some outstanding questions, Newton's view was accepted by many scientists and popularized by his followers for over 100 years.

Christiaan Huygens in 1678 proposed that light was emitted in all directions as a series of waves in a medium. As waves are not affected by gravity, it was assumed that they slowed down upon entering a denser medium. The wave theory predicted that light waves could interfere with each other like sound waves, and that light could be polarized, if it were a transverse wave. Early in the 19th century, Young showed that light behaved as waves by means of his famous double-slit diffraction experiment. Young's experiments supported the theory that light consists of waves. A subsequent work by Fresnel soon convinced scientists that light must be a wave, who independently worked out his own wave theory of light. In 1862, James Clerk Maxwell came up with some brilliant equations, today known as Maxwell's equations and concluded that light was a form of electromagnetic radiation. His theory was later experimentally proven correct by Heinrich Hertz. The weakness of the wave theory was that light waves, like sound waves, would need a medium for transmission. Hence, "ether" (the medium) was invented. However, Michelson-Morley experiments demonstrated that ether does not exist.

Max Planck in 1900 suggested the idea that black bodies emit light (and other electromagnetic radiation) only as discrete bundles or packets of energy (quanta) and accurately predicted the results of "black-body radiation" observed. Albert Einstein in 1905 developed the particle theory of black-body radiation to explain the photoelectric effect. He suggested that the quantization used by Planck reflects a basic aspect of the reality and light exhibits as a particle. Later, the particle of light was given the name photon with an energy E and a momentum $\boldsymbol{P}$. His ideas were experimentally confirmed by American scientist Robert Millikan a year later. However, his ideas were finally accepted throughout the world only a decade later when the American physicist Compton made theoretical predictions for the scattering of photons from the electron-Compton effect.

Louis de Broglie in 1924 stated that everything has both a particle nature and a wave nature as light. In a word, the modern theory explains the nature of light as wave-particle duality, that is light sometimes exhibits a particle nature and sometimes a wave nature. The quantum theory of light and electromagnetic radiation continued to evolve through the 1920s and 1930s, and culminated with the development during the 1940s of the theory of quantum electrodynamics (QED).

In this chapter, we briefly introduce the geometrical optics and its applications, and use the knowledge of wave motion in chapters 13 and 14 to treat interference, diffraction, and polarization of light.

15.1 Laws of Reflection and Refraction

15.1.1 The ray approximation in geometric optics

Geometrical optics describes light propagation in terms of "rays". A beam of light from its source, on common experience, travels in a straight-line path at a finite speed in a homogeneous medium until it counters a boundary between two different materials. When a beam of light strikes a boundary, it is reflected from that boundary or passes into the material on the other side of the boundary or partially does both. These phenomena are reflection and refraction. The ray approximation is used to represent beams of light. That is a ray of light is an imaginary line drawn along the direction of travel of the light beam. The "ray" is used to approximately model how light will propagate. Light rays are defined to propagate in a rectilinear path as far as they travel in a homogeneous medium. Rays bend (and may split in two) at the interface between two dissimilar media. A light ray is a line or curve that is perpendicular to the light's wavefronts.

Light is the fastest thing observed in universe. The speed of light in a vacuum is exactly 299,792,458 m/s, and approximately 3×10^8 m/s, labeled as c, a physical constant. When light enters a material, it slows down. For example, light travels about 30% slower in water than it does in a vacuum, while in diamonds, it travels at about half the speed it does in a vacuum. The index of refraction of medium is introduced to measure the speed of light in that medium, defined as

$$n=\frac{c}{v} \tag{15-1}$$

where v is the speed of light in the material. Eq. (15-1) implies that the index of vacuum is taken as unity "1", and the index of refraction of air differs little from unity and in most situations it is considered to be unity without significant difference. As an example, the index of refraction of water is 1. 333, meaning that light travels 1. 333 times faster in a vacuum than it does in water. Index of refraction of medium 2 relative to medium 1 is equal to the ratio of the speed of light in the first medium v_1 to the speed of light in the second medium v_2

$$n_{21}=\frac{n_2}{n_1}=\frac{v_1}{v_2} \tag{15-2}$$

15. 1. 2 Laws of reflection and refraction

When light goes from one transparent medium such as water or glass to another with different optical properties, usually there is a reflected bean as well as a refracted one as shown in Fig. 15-1, in which the ray 1 represents the incident light, the ray 2 reflected light and the ray 3 refracted light. The incident beam of light in the medium of refractive index n_1 encounters the interface (assuming to be a plane) between two media. Part of the light is reflected by the interface to form the ray 2 (reflected light), traveling as if the original beam had bounced from the interface. The rest of the light passes through the interface and into the medium of refractive index n_2, forming the ray 3(refracted light). The travel of light through an interface between two media is called refraction. Unless an incident beam of light is perpendicular to the interface, the directions of reflected light and refracted light have some changes with respect to the direction of incident light. For this reason, the beam is said to be bent by refraction.

The dashed line is the normal, a line perpendicular to the surface at the point where the ray hits. The plane containing the incident ray and the normal is called the plane of incidence, which is in the plane of the page in Fig. 15-1. The angle i between the normal and incident ray is called the angle of incidence, the angle i' between the normal and reflected ray is the angle of reflection and the angle r of between the normal and refracted ray is the angle of refraction. These angles satisfy law of reflection and law of refraction as follow.

Law of reflection: ① The angle of reflection is equal to the angle of incidence. ② The incident ray, the reflected ray, and the normal lie in the plane of incidence.

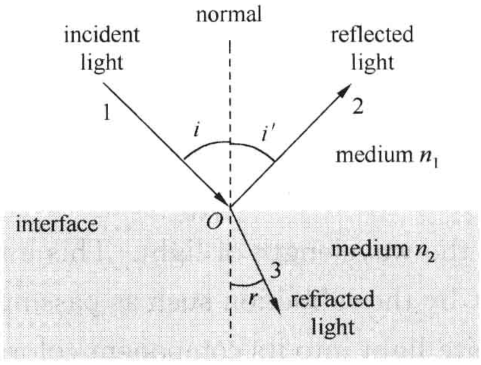


Fig. 15-1 Schematic diagram of reflection and refraction of light

Law of refraction: ① The incident ray, the refracted ray, and the normal lie in the plane of incidence. ② The angle of refraction is related to the angle of incidence by

$$n_1 \sin i = n_2 \sin r \tag{15-3}$$

Law of refraction is called Snell's law. Eq. (15-3) can be rewritten as

$$\sin r = \frac{n_1}{n_2}\sin i = \frac{\sin i}{n_{21}} \tag{15-4}$$

Eq. (15-4) implies that for a given angle i of incidence, the value of r depends on the relative refraction index n_{21}. For example, if $n_{21} > 1$ (that is $n_2 > n_1$), r is less than i, and this means that refraction bends the light beam away from the undeflected direction and toward the normal.

If n_2 is less than n_1, r is greater than i. In this case, refraction bends the light beam away from the undeflected direction and away from the normal. For the case, if the angle of incidence is greater than the critical angle i_c

$$i_c = \arcsin\frac{n_2}{n_1} \tag{15-5}$$

for which the angle of refraction is 90°, there is no refracted light and all the light is reflected. This effect is called total reflection. Total reflection has found many applications such as in optical fiber communications and in medical technology.

There are many application examples of the refraction and the total reflection in everyday life. For example, one can see a fish moving in water near to lakeside and find that the apparent depth of the fish in water is less than the actual depth as shown in Fig. 15-2. Moreover, one cannot find the fish far from lakeside. The law of refraction and total reflection can explain the daily experience.

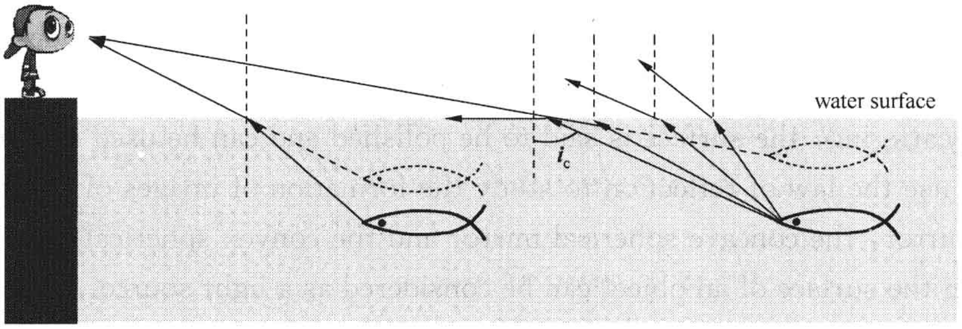


Fig. 15-2 Example of total reflection

The laws of reflection and refraction can be used to predict the deflection of light rays and the formations of images for mirrors and lenses.

Example 15-1 Find the critical angle of total reflection for a water-air boundary if the index of reflection of water is 1.33.

Solution From Eq. (15-5), we find the critical angle to be

$$i_c = \arcsin\frac{n_2}{n_1} = \arcsin\frac{1}{1.33} = \arcsin 0.752 \approx 48.8^\circ$$

15.1.3 Chromatic dispersion

The careful measurements have shown that the index of refraction in anything but vacuum depends on the wavelength of light. The dependence of the index of refraction on wavelength indicates that when light is refracted by a surface of material, the angle of refraction depends on the wavelength of light. This implies that a very narrow beam of white light will be spread out by the refraction such as passing through a prism as shown in Fig. 15-3. This separation of white light into its component colors is called chromatic dispersion. The index of refraction for a material usually decreases with increasing wavelength so that violet light refracts more than red light when passing from air into a material.

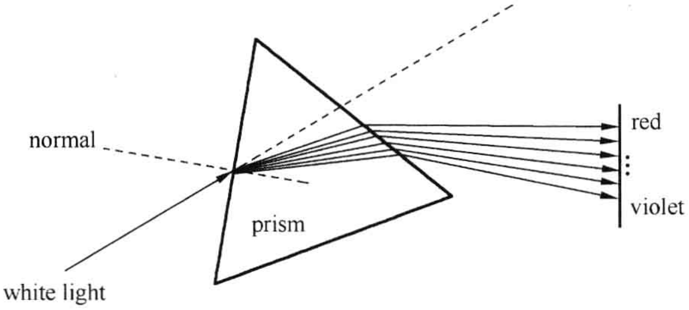


Fig. 15-3 Schematic diagram of chromatic dispersion

The dispersion of light is demonstrated most vividly in nature through the formation of a rainbow, often seen by an observer positioned between the Sun and a rain shower.

15.2 The Plane Mirror and Spherical Mirrors

A mirror is a surface that can reflect a beam of light in one direction instead of either scattering it widely in many directions or absorbing it. When light falls on a rough, opaque surface, the incident light is scattered in all directions. If the surface is so smooth that there is little random scattering, the surface is said to be polished and can be used as a mirror. In the following, we use the law of reflection to study the formation of images of an object reflected by the plane mirror, the concave spherical mirror and the convex spherical mirror.

A point on the surface of an object can be considered as a light source. Generally a source of light can be considered a point source if the resolution of the imaging instrument is too low to resolve its size, or if the object is at a very great distance.

15.2.1 Plane mirror

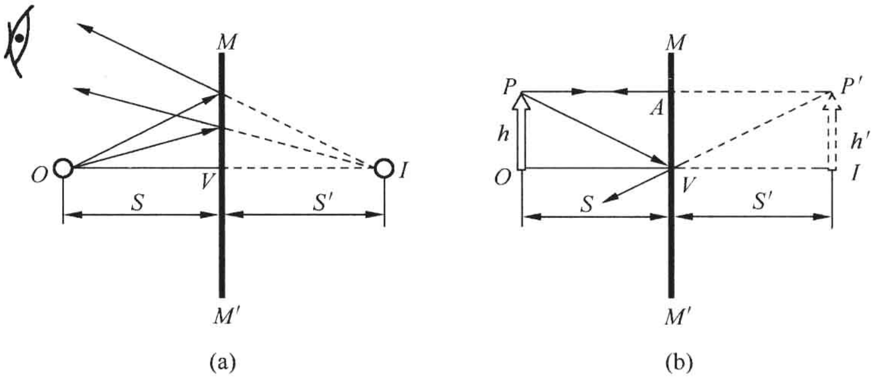
A plane mirror is a mirror with a planar reflective surface. For light rays striking a plane mirror, the angle of reflection equals the angle of incidence. This implies that images of object are upright and the same distance behind the mirror as in front of the mirror. The image size is the same as the object size. That is the magnification of a flat mirror is equal to one. Images are classified as real or virtual. A real image is one in which light actually passes through the image point and a virtual image is one in which the light does not pass through the image point but appears to come (diverge) from that point.

As shown in Fig. 15-4(a), when a point object O is placed in frond of a plane mirror MM' and sends light in all directions, some of rays are reflected. To an observer in front of the mirror all the reflected rays appear to come from the point I behind the mirror. The observer therefore sees a bright spot which appears to be behind the mirror and which we call the image of the point O. The image behind the mirror is a virtual image because the light rays do not actually come from I. The distance S is the object distance and the distance S' is the image distance.

The image of the object OP with height of h can be found by graphical ray tracing as shown in Fig. 15-4(b). Two rays start from point P and then reflect from mirror. One follows a horizontal path PA to the mirror, and reflects back on itself. The second ray follows the oblique path PV and reflects. An observer to the left of the mirror would trace the two reflected rays back to the point P' from which they appear to have originated, that is the image of point P. Because triangles PAV and $P'AV$ are identical, the image is far behind the mirror as the object in front, and the object height h equals the image height h'. Lateral magnification M is defined as follows

$$M = \frac{\text{image height}}{\text{object height}} = \frac{h'}{h} \tag{15-6}$$

For a plane mirror, $M = 1$ because $h = h'$.



(a) (b)

Fig. 15-4 Formation of image by reflection from a plane mirror

However, the law of reflection also implies that the image is a laterally-inverted "mirror image" of the object. If a person is reflected in a plane mirror, the image of his right hand appears to be the left hand of the image.

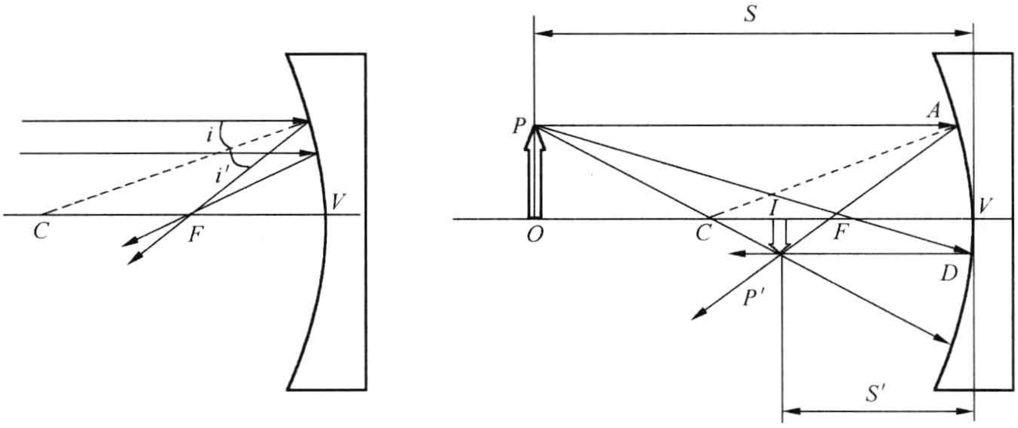
15.2.2 Spherical mirrors

A curved mirror is a mirror with a curved reflective surface, which may be either convex or concave. Most curved mirrors have surfaces that are shaped like part of a sphere, but other shapes are sometimes used in optical devices.

A concave spherical mirror is part of a spherical shell with its inner surface polished as shown in Fig. 15-5(a). The mirror has radius of curvature R, and its center of curvature is at point C. Point V (vertex) is the center of the spherical segment, and a line drawn from C to V is called the principal axis of the mirror. When the rays parallel to the principal axis reach a concave mirror, those near the central axis are reflected through a common point F which is called the focal point (or focus) of the mirror (a real focal point) and its distance from the center of the mirror is the focal length of the mirror. That is

$$f = \frac{R}{2} \tag{15-7}$$

If an incident ray of light passes through the focus F, it is reflected back in the direction of principal axis.



(a) (b)

Fig. 15-5 (a) The concave mirror and (b) formation of image

As shown in Fig. 15-5(b), an object OP is placed before a concave mirror, we can locate an image by graphical ray tracing. The ray PA parallel to the principal axis is reflected back through F, and the ray PF strikes the mirror at point D and is reflected back parallel to the principal axis. The ray PC comes to the spherical surface along the radius and is reflected directly back on itself. The IP' is the image of object OP and it is a real image since the rays of light come from the IP'. In this case, the image is inverted. Let S and S' denote the distances of object and image respectively. They are related by a simple formula which can be derived easily

$$\frac{1}{S} + \frac{1}{S'} = \frac{1}{f} \tag{15-8}$$

This equation applies to any concave, convex, or plane. For a convex or plane mirror, only a virtual image can be formed, regardless of the object's location on the central axis. For a concave mirror, different image types can be formed and depend on the distance between the object and the mirror. Note, according to Eq. (15-8), that an object and it's image are conjugate.

For a convex mirror, we find that the parallel rays are no longer reflected through a common point. Instead, they diverge. As shown in Fig. 15-6, the convex mirror is a diverging mirror because rays initially parallel diverge after reflection. A ray of light parallel to the principal axis is reflected away from the principal axis, and the reflected ray is extended backward to cross the principal axis as point F, the principal focus for this convex mirror. The principal focus is virtual. If we place an object OP, an erect virtual image IP' is formed and reduce in size. The image on a convex mirror is always virtual, diminished, and upright. Eq. (15-8) can be used for both concave and convex mirrors if we are careful to use the proper signs for various quantities shown in Table 15-1.

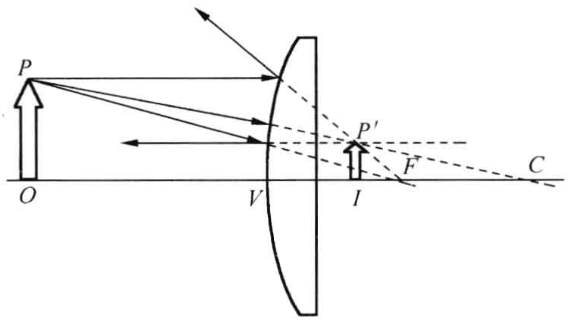


Fig. 15-6 The convex mirror and formation of image

Table 15-1 Sign conventions for mirrors

Quantity	Positive when	Negative when
Object location S	Object is in front of Mirror	Object is behind mirror
Image location S'	Image is behind mirror	Image is in front of Mirror
Image height h'	Image is upright	Image is inverted
Radius R(or focal length f)	Mirror is concave	Mirror is convex

The lateral magnification can be proved to be

$$M=\frac{\text{Height of image}}{\text{Height of object}}=\frac{h'}{h}=-\frac{S'}{S} \tag{15-9}$$

By convention, the lateral magnification includes a plus sign when the image orientation is that of the object and a minus sign when the orientation is opposite that of the object. The image-forming characteristics of curved mirrors obviously determine their uses.

Example 15-2 An object of height 4.00 cm is placed 25.00 cm from a convex mirror with a focal length of 10.00 cm. Find

(1) the position of the formed image;

(2) the magnification of the mirror.

Solution (1) Because the mirror is convex, its focal length is negative. According to Eq. (15-8), we have

$$\frac{1}{25.0}+\frac{1}{S'}=\frac{1}{-10.0} \rightarrow S'=-7.14\ (\text{cm})$$

The negative value of S' indicates that the image is virtual and behind the mirror.

(2) The magnification of mirror is

$$M=-\frac{-7.14}{25.00}\approx 0.286$$

Hence, the image is upright.

15.3 Images Formed by Refraction and Thin Lenses

15.3.1 Images formed by refraction

We now turn from images formed by reflections through spherical surfaces of transparent materials, such as glass. Consider two transparent media with indices of refraction n_1 and n_2, where the boundary between the two media is a spherical surface of radius R. At the point of refraction of each ray, the normal to the refracting surface is a radial line through the center of curvature. Because of refraction, the ray bends toward the normal if it is entering a medium of greater index of refraction, and away from the normal if it is entering a medium of lesser index of refraction.

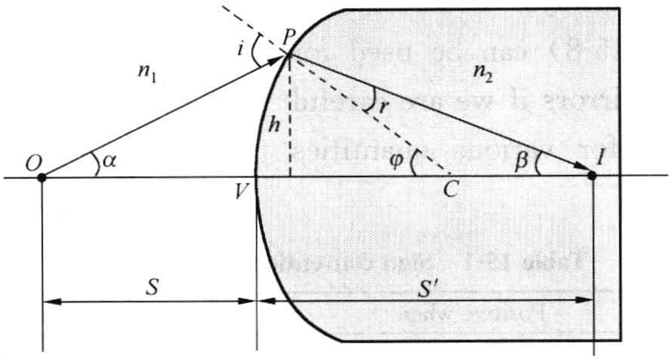


Fig. 15-7 An Image formed by refraction through a spherical surface

Consider a ray OP from the point source O in medium 1 which is incident upon a spherical surface (convex) of medium 2 with center of curvature at C as shown in Fig. 15-7. A line drawn from C to the center V of the segment of spherical surface is called the principal axis of the system. Ray OP is refracted at surface and an image is formed. According to law of refraction, we have

$$n_1 \sin i = n_2 \sin r \tag{15-10}$$

It is easy to prove that

$$i = \alpha + \varphi$$
$$\varphi = r + \beta$$

For the paraxial rays, i. e., rays which make small angles with the axis, their sine, their tangent and themselves are essentially equal. Hence we have

$$\begin{cases} \sin i \approx i = \alpha + \varphi \approx \dfrac{h}{S} + \dfrac{h}{R} \\ \sin r \approx r = \varphi - \beta \approx \dfrac{h}{R} - \dfrac{h}{S'} \end{cases} \tag{15-11}$$

Then, by Eq. (15-10), we have

$$\frac{n_1}{S} + \frac{n_2}{S'} = \frac{n_2 - n_1}{R} \tag{15-12}$$

Note that real images are formed on the side of surface opposite the side from which the light comes. This is in contrast with mirrors. Eq. (15-12) applies to a variety of circumstances: concave convex and plane surfaces(R approaches infinity). The sign convention for spherical refracting surface is summarized in Table 15-2.

Table 15-2 Sign conventions for refracting surface

Quantity	Positive when	Negative when
Object location S	Object is in front of surface	Object is in back of surface
Image location S'	Image is in back of surface	Image is in front of surface
Image height h'	Image is upright	Image is inverted
Radius R	Center of curvature is in back of surface	Center of curvature is in front of surface

Furthermore, the lateral magnification of a refracting surface is

$$M=\frac{\text{Height of image}}{\text{Height of object}}=\frac{h'}{h}=-\frac{n_1 S'}{n_2 S} \tag{15-13}$$

15.3.2 Thin lenses

Our knowledge of how light rays are bent when they pass from one material to another can now be applied to lenses, which are the basic elements of most optical instruments. A lens is a transparent object with two refracting surfaces whose central axes coincide. The common central axis is the principal axis of the lens. Simple lenses are bounded by faces which are small sections of sphere. When a lens is surrounded by air, light refracts from the air into the lens, crosses through the lens, and then refracts back into the air. Each refraction can change the direction of travel of the light. A lens that causes light rays initially parallel to the central axis to converge is called a converging lens. If, instead, it causes such rays to diverge, the lens is a diverging lens. Fig. 15-8 shows six possible types of spherical lenses with their name: double-convex, plano-convex, plano-concave, concavo-convex, convexo-concave, and double-concave.

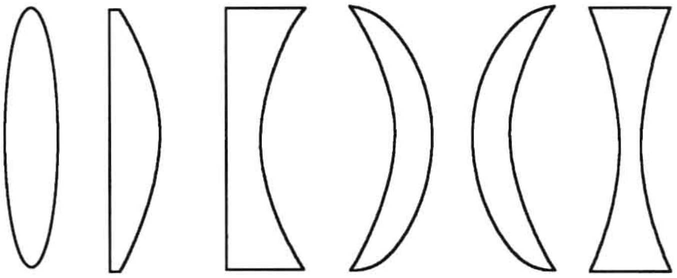


Fig. 15-8 Six possible types of spherical lenses

A thin lens is a lens in which the thickest part is thin compared to the object distance S, the image distance S' and the radii of curvature R_1 and R_2 of the two surfaces of lens. The thin lens approximation ignores optical effects due to the thickness of lenses and simplifies ray tracing calculations. Rays which approach a converging lens parallel to the principal axis are deviated so that they pass through a common point, the principal focus, on the principal axis. A lens has two principal foci, one for light incident from the left, which we denote by F, and the second F' for light incident from the right. For thin lenses the principal foci are equidistant from the optical center of the lens as shown in Fig. 15-9. The focal length f of a thin lens with index of refraction n surrounded by air is given by the Lensmaker's equation

$$\frac{1}{f}=(n-1)\left(\frac{1}{R_1}-\frac{1}{R_2}\right) \tag{15-14}$$

where R_1 and R_2 are the radii of curvature of the two surfaces. Here, R_1 is taken to be positive if the first surface is convex, and negative if the surface is concave. The signs are reversed for the back surface of the lens: R_2 is positive if the surface is concave, and negative if it is convex. The line FF' is called as principal optical axis, which passes through the center of curvature of each surface. The front and rear (or back) focal planes are defined as the planes which

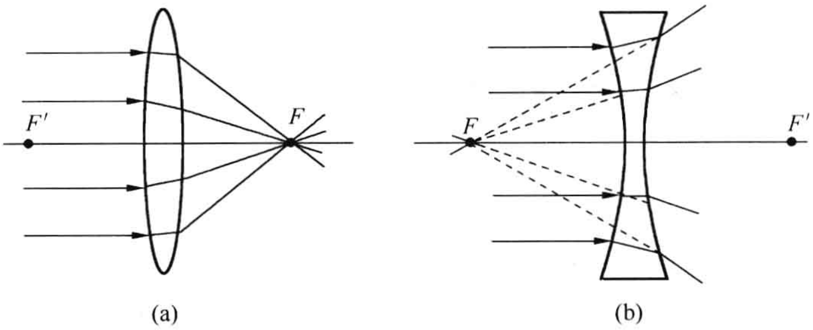


Fig. 15-9 The foci of thin lenses: (a) double-convex; (b) double-concave

are perpendicular to the principal optical axis and pass through the front and rear focal points. The line PP' through the center of thin lens in Fig. 15-10 is defined as the secondary optical axis.

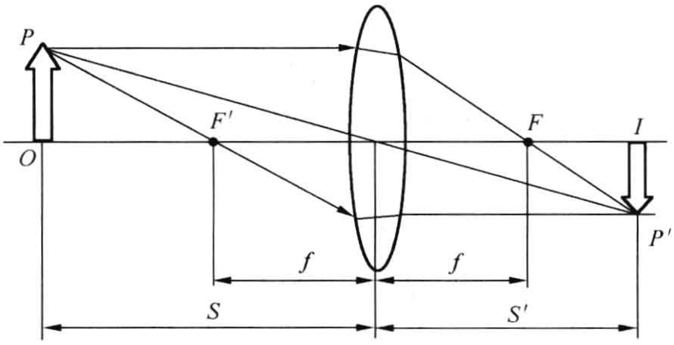


Fig. 15-10 Formation of a real image by a converging lens

We shall also consider only light rays that make small angles with the central axis. That is the paraxial approximation. Take a converging lens as an example. As shown in Fig. 15-10, an object OP outside the focus F is a distance S from the converging lens and an image IP' is formed at a distance S' from the lens, which in this case is real and inverted. The image can be located by tracing of three standard rays:

(1) A ray from P parallel to the principal axis is deviated so that it passes through the principal focus of a converging lens (or diverge as though it came from the principal focus of a diverging lens).

(2) The ray from P which approaches the lens along the line through the principal focus F' is deviated so that it leaves the lens parallel to the principal axis.

(3) A ray through the optical center of lens passes through the lens undeviated.

For a thin lens, in the paraxial approximation, the object S and image S' distances are related by the equation

$$\frac{1}{S}+\frac{1}{S'}=\frac{1}{f} \tag{15-15}$$

which is called as the thin-lens equation, and can be used with both converging and diverging lenses. The sign convention of object distance, image distance and focal distance used in this text is listed in Table 15-3.

Table 15-3 Sign Convention for Thin Lenses

Quantity	Positive when	Negative when
Object location S	Object is in front of lens	Object is in back of lens
Image location S'	Image is in back of lens	Image is in front of lens
Image height h'	Image is upright	Image is inverted
Focal length f	Converging lens	Diverging lens

The lateral magnification is

$$M=\frac{\text{Height of image}}{\text{Height of object}}=\frac{h'}{h}=-\frac{S'}{S} \tag{15-16}$$

which is the same as Eq. (15-10) for magnification by a mirror.

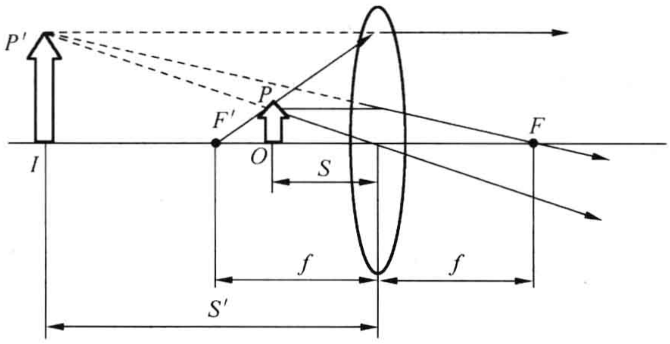


Fig. 15-11 Formation of a virtual image by a converging lens

When an object is inside the principal focus of a converging lens, the image formed is erect, virtual, and enlarged in Fig. 15-11. In this case, the image distance is negative. When an object is placed in front of a system of lenses such as two lenses whose principal axes coincide, the positive of the final image can be calculated by repeated use of the lens equation. The object for the second lens is the image formed by the first. If there is a third lens, the image formed by the second lens acts as its object. The final image formed by a very complicated optical system can be located by successive application of the lens (or mirror) equation. Great care must be taken to assign the proper sign to each distance.

Example 15-3 An object is placed 30.0 cm from a thin diverging lens of focal length 10.0 cm. Find the image distance and the lateral magnification.

Solution From Eq. (15-15), we have

$$\frac{1}{30.0}+\frac{1}{S'}=\frac{1}{-10.0}\rightarrow S'=-7.50\ (\text{cm})$$

The lateral magnification is

$$M=-\frac{S'}{S}=-\frac{-7.50}{30.0}=0.250$$

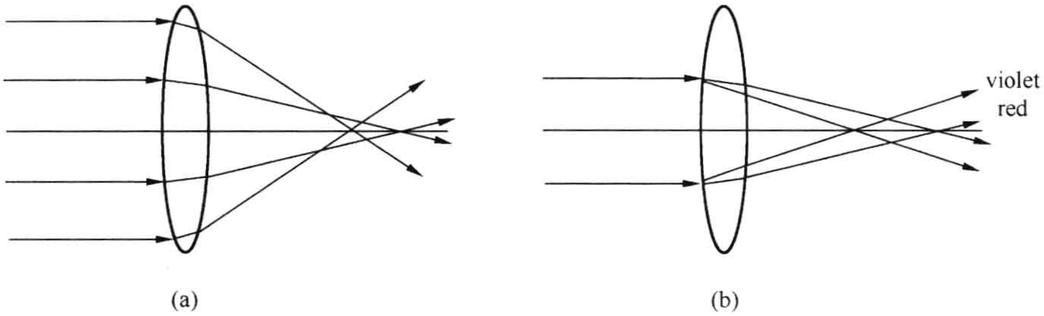
This result implies that the image is virtual, upright and smaller than the object.

15.3.3 Mirror and lens aberrations

Mirrors and lenses are widely used to form images in many optical instruments. If we had an ideal lens, every ray of light from any given point on the object would cross at exactly the same point on the image. Further, the image would be similar to the object in every respect.

Any circle on the object would be a perfect circle on the image. If the object were all in one plane, the image would be all in one plane. Actually no such ideal lens exists. If most of the rays from a point on the object cross at the desired image point but some of them cross off on one side, the image becomes blurred. With some lenses, the image of a rectangular object is distorted so that it is barrel-shaped and in some case it may look like a pincushion. The imperfect quality of the images is largely the result of defects in shape and form, and is also the result of aberrations: spherical aberration, chromatic aberration.

Spherical aberration results from the fact that when rays of light parallel to the principal axis of a spherical lens (or mirror) pass through zones near its edges, they cross the axis nearer the lens than those rays which pass through nearer to the center so that the image is blurred as shown in Fig. 15-12(a). Spherical aberration can be reduced by some methods such as using an adjustable aperture and choosing the proper curvature of the lens.



(a) (b)

Fig. 15-12 Schematic Diagrams of (a) spherical aberration and (b) chromatic aberration

Chromatic aberration is due to the dependence of the index of refraction on wavelength. When rays of white light parallel to the principal axis pass through an ordinary lens, they are refracted in such a way that different colors are brought to focus at different distances from the lens as shown in Fig15-12(b). As result the image is not sharp and is likely to have colored rings or markings. The chromatic aberration can be corrected by combing a converging lens of crown glass with a diverging lens of flint glass since their dispersions are opposite. For better correction of chromatic aberration, more lenses can be used.

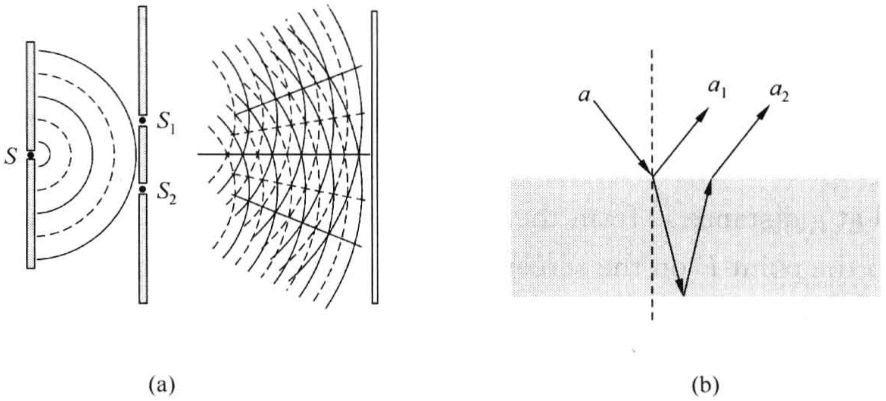
15.4 The Coherence of Light

We will see that two light waves emitted by two coherent sources can produce interference pattern. The existence of interference phenomena is perhaps the most convincing evidence that light is a wave since interference cannot be explained other than with waves. In order to understand the interference of light, you must understand firstly the coherence of light.

When discussing the superposition of mechanical waves in chapter 13, we saw that two waves emitted by two coherent sources could produce interference pattern in crossing space. However, no interference pattern would be observed when two light waves emitted by two completely independent light sources. It means that two completely independent light sources are not coherent sources.

For common sources of visible light, the fundamental light emission processes owe to individual atoms and these atoms do not act together in a coherent way. The action of light emission by a single atom in a typical case takes about 10^{-9} s and the emitted light is properly described as a wavetrain that length is less than one meter. No constant phase relationship exists between different wavetrains, even if the wavetrains maintain identical frequencies. Lasers are coherent sources of light since its output is a single-frequency radiation of in-phase parallel waves, whereas incandescent light bulbs and fluorescent lamps are incoherent sources.

How can we obtain coherent light from common sources of light? If light waves emitted by an identical point of a monochromatic source can be divided into two beams by using some arrangements, the two beams are coherent light. For example, we can use the arrangements as shown in Fig. 15-13 to obtain coherent light. In Fig. 15-13(a), S_1 and S_2 are two wavelet sources on an identical wavefront emitted by a point S of a monochromatic source. This ensures that, by diffraction, the same families of wavetrains fall on slits S_1 and S_2. When the phase of the light emitted from S changes, this change is transmitted simultaneously to S_1 and S_2, the diffraction beams emerging from S_1 and S_2 are thus coherent. This way to obtain coherent light is called the way of division of wavefront. Fig. 15-13(b) shows another easy way to obtain coherent light. Because the beams a_1 and a_2 are two reflected light waves that come from the same wavetrain a, they are coherent light. This way to obtain coherent light is called the way of division of amplitude.



(a) (b)

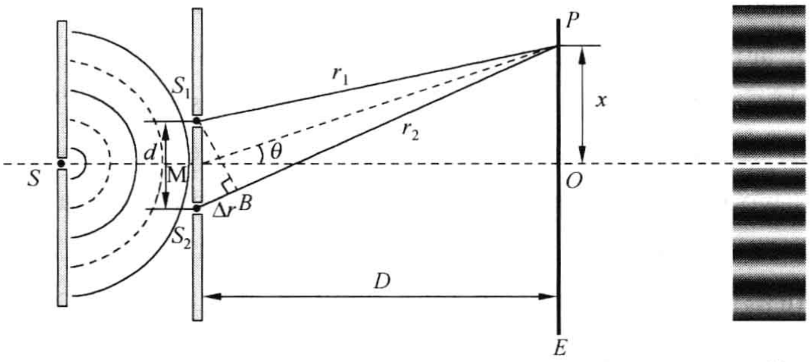
Fig. 15-13 Easy ways to obtain coherent light: (a) the way of division of wave front and (b) the way of division of amplitude

Interference is not limited to mechanical waves but is a characteristic of all wave phenomena. The existence of interference phenomena of light is perhaps our most convincing evidence that light is a wave.

15.5 Two Beams Interference by Division of Wavefront

15.5.1 Young's double-slit experiment

The historical experiment that established the wave theory of light is Young's experiment. In 1801, English physicist Thomas Young performed a famous experiment that demonstrated the wave nature of light by showing two overlapping light waves interfered with each other. The experiment also showed how this phenomenon supports the wave theory of Huygens that we introduced before. And Young's experiment was especially convincing because he was able to deduce the wavelength of incident light from his measurement, the first determination of this important quantity.



(a) (b)

Fig. 15-14 (a) Young's double-slit experiment and (b) the interference pattern

Two narrow slits S_1 and S_2 are equidistant from the source S, as shown in Fig. 15-14(a). As a beam of monochromatic light from S reaches S_1 and S_2, each slit serves as a new light source producing new wavefront in phase with each other. The sources are said to be coherent. These waves travel out from S_1 and S_2 producing bright fringes on the viewing screen E where constructive interference occurs and dark fringes where destructive interference occurs. Figure 15-14(b) shows the interference pattern.

Let us now consider the theoretical conditions necessary for the production of bright and dark fringes. In Fig. 15-14(a), the separation of the two slits is represented by d, and the screen is located at a distance D from the slits. Two beams travel the distances r_1 and r_2 from slits S_1 and S_2 to the point P on the screen. The point M is the midpoint between the two slits and line MP makes an angle θ with axis MO. The line S_1B is drawn so that the length of line BP equals to r_1. As the distance D is much longer than the slit separation d, the path difference Δr coming from S_1 and S_2 is approximately given by

$$\Delta r = r_2 - r_1 = d\sin\theta$$

If angle θ is small enough, then

$$\sin\theta \approx \tan\theta = \frac{d}{D}$$

Therefore, we have

$$\Delta r \approx \frac{xd}{D}$$

Constructive interference will occur at P when

$$\Delta r = \pm k\lambda, \quad k = 0, 1, 2, \cdots$$

where λ is the wavelength of incident light. That is bright fringes will occur at

$$x = \pm k\frac{D}{d}\lambda, \quad k = 0, 1, 2, \cdots \tag{15-17}$$

Destructive interference will occur at P when

$$\Delta r = \pm (2k-1)\frac{\lambda}{2}, \quad k = 1,2,3,\cdots$$

That is, dark fringes will occur at

$$x = \pm (2k-1)\frac{D}{d}\frac{\lambda}{2}, \quad k=1,2,3,\cdots \tag{15-18}$$

The central bright fringe, $k=0$, is called the zero-order bright fringe. The bright (or dark) fringes, $k=1$, are the first-order bright (or dark) fringes which are symmetrically located about O and so forth. From Eq. (15-17) and Eq. (15-18), the central distance between two adjacent bright (or dark) fringes is

$$\Delta x = \frac{D}{d}\lambda \tag{15-19}$$

which is also called the fringe spacing.

The result of the interference is to produce a series of alternating bright and dark fringes on the screen, which are equally spaced. We call them as interference fringes. Since the spacing Δx is proportional to the wavelength λ, the separation for red light is wider than that for violet light. When the incident light contains more than one wavelength, the separate interference patterns with different fringe spacing will be superimposed on the screen. For example, when white light is used in Young's double-slit experiment, except for the central fringe, which is white, the bright fringes are a rainbow of colors.

Example 15-4 In Young's experiment, the slit separation is 0.04 mm and the slit-screen separation is 2 m. The third-order bright fringe is formed at 8.3 cm from the central fringe.

(1) What is the wavelength of light used?

(2) Where will the second-order dark fringes appear?

(3) What is the spacing of two adjacent bright fringes?

Solution (1) For the third-order bright fringe, $k=3$ in Eq. (15-17), thus $\frac{xd}{D} = 3\lambda$, so that

$$\lambda = \frac{xd}{3D} = \frac{8.3\times10^{-2}\times4\times10^{-5}}{3\times2} = 5.53\times10^{-7}\,(\mathrm{m})$$

(2) The displacement of the second-order dark fringe is found by setting $k=2$ in Eq. (15-18) That is

$$x = \pm 3\frac{D}{d}\frac{\lambda}{2} = \pm\frac{3\times2\times5.53\times10^{-7}}{2\times4\times10^{-5}} = \pm 4.15\times10^{-2}\,(\mathrm{m})$$

Where the sign "±" shows the second-order dark fringes are symmetrically located about central fringe.

(3) From Eq. (15-19), the spacing of two adjacent bright fringes is

$$\Delta x = \frac{D}{d}\lambda = \frac{2}{4\times10^{-5}}\times5.53\times10^{-7} \approx 2.76\times10^{-2}\,(\mathrm{m})$$

Example 15-5 A beam of parallel white light ($\lambda = 400 \sim 760$ nm) is used in Young's experiment with the slits separated by a distance $d=0.250$ mm. The screen is located at a distance of $D=0.500$ m from the slits.

(1) Where will the second-order bright fringes appear for red light ($\lambda_r = 760$ nm)? Where will the forth-order bright fringes appear for violet light ($\lambda_v = 400$ nm)?

(2) What is the spacing of two adjacent dark fringes for red light and for violet light, respectively? And give the answers again when d is changed into $d'=0.500$ mm.

Solution (1) According to Eq. (15-17), for the second-order bright fringe of red light, $k=2$, $\lambda_r=760$ nm, we have

$$x_{2r}=\pm 2\frac{D}{d}\lambda_r=\pm 2\times\frac{0.500}{0.250\times 10^{-3}}\times 760\times 10^{-9}=\pm 3.04\times 10^{-3}\,(\mathrm{m})=\pm 3.04\ (\mathrm{mm})$$

For the forth-order bright fringe of violet light, $k=4$, $\lambda_v=400$ nm, we have

$$x_{4v}=\pm 4\frac{D}{d}\lambda_v=\pm 4\times\frac{0.500}{0.250\times 10^{-3}}\times 400\times 10^{-9}=\pm 3.20\times 10^{-3}\,(\mathrm{m})=\pm 3.20\ (\mathrm{mm})$$

(2) From Eq. (15-19), the spacing of two adjacent dark fringes for red light is

$$\Delta x_r=\frac{D}{d}\lambda_r=\frac{0.500}{0.250\times 10^{-3}}\times 760\times 10^{-9}=1.52\times 10^{-3}\,(\mathrm{m})=1.52\ (\mathrm{mm})$$

and the spacing for violet light is

$$\Delta x_v=\frac{D}{d}\lambda_v=\frac{0.500}{0.250\times 10^{-3}}\times 400\times 10^{-9}=8.00\times 10^{-4}\,(\mathrm{m})=0.800\ (\mathrm{mm})$$

When d is changed into $d'=0.500$ mm, we have

$$\Delta x_{r'}=\frac{D}{d'}\lambda_r=\frac{0.500}{0.500\times 10^{-3}}\times 760\times 10^{-9}=7.60\times 10^{-4}\,(\mathrm{m})=0.760\ (\mathrm{mm})$$

$$\Delta x_{v'}=\frac{D}{d'}\lambda_v=\frac{0.500}{0.500\times 10^{-3}}\times 400\times 10^{-9}=4.00\times 10^{-4}\,(\mathrm{m})=0.400\ (\mathrm{mm})$$

We can see that the forth-order bright fringes of violet light will coincide with the second-order bright fringes of red light. The fringe separation for red light is wider than that of for violet light when d keeps a constant and the wavelength of incident light keeps unchanged. The wider the d is, the narrower the separation is. When we use white light as incident light in Young's experiment, except for the central bright fringe, all fringes on the screen overlap one another and form rainbow color fringes.

15.5.2 Lloyd's mirror

In chapter 13, we have discussed that if mechanical wave incident from a lower density medium to a higher density medium, the reflected waves from the interface between them will undergo a phase shift π or a wave-path shift $\lambda/2$. This phenomenon also occurs for light waves' reflection, which can be demonstrated by an arrangement called Lloyd's mirror, as shown in Fig. 15-15.

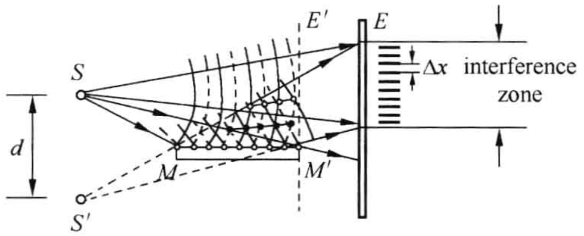


Fig. 15-15 Lloyd's mirror

A narrow slit S in an opaque screen is irradiated by one beam monochromatic light of wavelength λ. A plane mirror MM' is located a distance $d/2$ below S. The actual source S and its virtual image S' formed by the reflected light waves from the surface of plane mirror are a pair of coherent sources. The fringes formed by interference between the coherent light waves from S and S' can be viewed in the interference zone on screen E. When we move E to E' which is in contact with the edge of the mirror (Fig. 15-16), the fringe nearest the edge is dark. This is, however, the fringe corresponding to zero path difference. We have to recognize that the zeroth fringe is dark indicates that the waves reflected from the surface of mirror have undergone a phase shift of π. The experiment of Lloyd's mirror demonstrates that a one-half a wavelength (or a π phase shift) occurs when light reflects from a medium having an index of refraction higher than that of the medium in which it is originally traveling.

Example 15-6 Fig. 15-16 shows an experiment of Lloyd's mirror. The perpendicular distance from linear source S_1 to the surface of mirror is 1 mm. The screen is located at a distance $D=1.5$ m from the source. The total length of mirror is $MM'=D/2$ and the distance from the right edge of mirror M' to the screen is $M'O=D/4$.

(1) Calculate OA and OB which are the distances from the top and bottom edges of the interference zone to the center of the screen.

(2) What is the spacing of two adjacent bright fringes when the incident wavelength is $\lambda=600$ nm? And how many bright fringes are there on the screen?

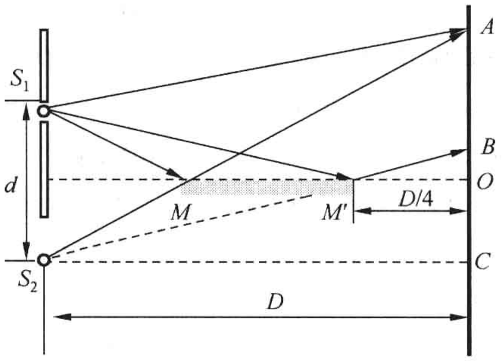


Fig. 15-16 For Example 15-5

Solution (1) Suppose point C is the projection of virtual light source S_2 as shown in Fig. 15-16. From $\triangle MOA \backsim \triangle S_2CA$ and $\triangle M'OB \backsim \triangle S_2CB$, we have

$$\frac{OA}{OA+d/2}=\frac{D/2+D/4}{D}=\frac{3}{4}, \quad \frac{OB}{OB+d/2}=\frac{D/4}{D}=\frac{1}{4}$$

Substituting $d=2$ mm into above two equations, we have

$$OA=3.0 \text{ mm}, \quad OB=0.33 \text{ mm}$$

(2) Comparing Fig. 15-16 with Fig. 15-14, we can find that Lloyd's mirror experiment is similar to Young's experiment. Thus, the spacing of two adjacent bright fringes is also

$$\Delta x=\frac{D}{d}\lambda$$

Substituting the values of D, d and λ into above equation, we have

$$\Delta x=\frac{1.5\times 6\times 10^{-7}}{2\times 10^{-3}}=4.5\times 10^{-4}\,(\text{m})=0.45\ (\text{mm})$$

Consider that the half-wavelength shift of reflection, the wave-path difference is

$$\delta=\frac{d}{D}x+\frac{\lambda}{2}$$

for bright fringes

$$\frac{d}{D}x+\frac{\lambda}{2}=k\lambda, \quad k=1,2,3,\cdots$$

that is

$$x=\left(k-\frac{1}{2}\right)\frac{D}{d}\lambda$$

we have

for $k=1$, $x_1 = 0.225$ mm $< OB$, it is not in the interference zone.

for $k=2$, $x_2 = x_1 + \Delta x = 0.675$ mm $> OB$, bright fringes can be observed on the screen.

Substituting $OA=3$ mm into the condition of bright fringes, we have $k=7.17$, it tells us that the seventh-order bright fringe is the upper most one on the screen. Therefore, there are six bright fringes, corresponding to $k=2,3,4,5,6,7$, on the viewing screen.

Light is a kind of waves, for instance, demonstrated by experiments of Young's double-slit and Lloyd's mirror. The wavelength of incident light can be calculated from the interference pattern. Lloyd's mirror experiment indicates further that light traveling from a higher index medium undergoes (gain or lose) one-half a wavelength when it is reflected at an interface. This might be remembered as "higher means half".

However, so far we have considered interference of light only as light waves traveling in the same medium. As light waves traveling in two or more media, how do we decide the conditions of constructive or destructive interference? We will give an experimental fact to show that an optical path difference is used for the conditions of interference during the light beams propagating in different media.

15.6 Optical Path and Optical Path Difference, a Property of Thin Lens

In Young's experiment, a shift in the fringe pattern can be caused by the insertion of a thin transparent material (e. g. a thin glass plate) into the path of one of the interfering beams. We know that the position of fringes is decided upon the phase difference of coherent light. From Eq. (13-24a), we can write the phase difference of two coherent light waves as

$$\Delta\varphi = 2\pi\left(\frac{r_2}{\lambda_2} - \frac{r_1}{\lambda_1}\right)$$

where λ_1 and r_1 are wavelength and path (or geometric path) in the medium with a refractive index n_1, λ_2 and r_2 are wavelength and path in the medium of refractive index n_2.

Since the wavelength λ_n in a given medium in terms of the wavelength λ in a vacuum can be written as

$$\lambda_n = \frac{\lambda}{n}$$

Thus

$$\Delta\varphi = 2\pi\left(\frac{n_2 r_2 - n_1 r_1}{\lambda}\right)$$

If one or both paths are through media with different indices of refraction, we must take into account the fact that the phase difference depends on $(n_2 r_2 - n_1 r_1)$ instead of $(r_2 - r_1)$. The product of the refractive index and the geometrical path within the medium is called optical path. What is the physical meaning of optical path? Consider two coherent light wave trains L_1 and L_2 propagating through different media as shown in Fig. 15-17. Assume that L_1 propagates in vacuum, and L_2 propagates through a medium with refractive index n. The time required by L_2 to travel a distance l in the medium is $t = l/v$, where v is the speed of light propagating in the medium. During this time, L_1 will have traveled a distance $S = ct = lc/v$, that is $S = nl$. It means that the light wave travels a distance l in medium with refractive index n is equivalent to a distance nl if it travels in vacuum during the same time. Thus, as Fig. 15-17 makes clear, the optical path is the distance in vacuum containing the same number of waves as the actual geometry path in the medium. Let δ be the optical path difference, the phase difference of two coherent light waves is $\Delta\varphi = 2\pi\frac{\delta}{\lambda}$. Therefore, when discussing the interference of light in different media we should convert geometry path into optical path for calculating the phase difference between the coherent light waves reaching given point. Now the condition of interference can be rewrite as

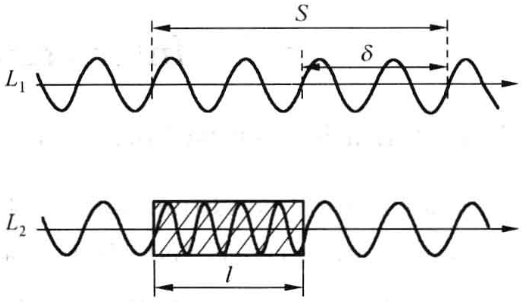


Fig. 15-17 Two coherent light waves: L_1 in vacuum and L_2 propagating through different media

$$\Delta\varphi = 2\pi\frac{\delta}{\lambda} = \begin{cases} \pm 2k\pi, & k = 0,1,2,\cdots, \quad \text{bright fringes} \\ \pm(2k+1)\pi, & k = 0,1,2,\cdots, \quad \text{dark fringes} \end{cases} \tag{15-20a}$$

or

$$\delta = \begin{cases} \pm 2k\frac{\lambda}{2} & k = 0,1,2,\cdots, \quad \text{bright fringes} \\ \pm(2k+1)\frac{\lambda}{2}, & k = 0,1,2,\cdots, \quad \text{dark fringes} \end{cases} \tag{15-20b}$$

Interference pattern is often obtained experimentally by using a converging lens. A convex lens converges all rays parallel to the principal optical axis to its focus. Moreover, the rays parallel to a secondary optical axis are also converged to point P at which the secondary optical axis intersects with the rear focal plane as shown in Fig. 15-18. Now obviously different rays in the beam of light parallel to the secondary optical axis will travel different distances in reaching the point P. Does a lens yield any additional phase shift? If rays diverging from a point source form a real image after passing through a converging lens, the rays travel over paths of different geometric length but all paths contain the same number of waves, otherwise the rays would not be in phase when they recombine at the image point. It is, in fact, the function of the lens — taking advantage of the fact that the wavelength in glass is less than that in air — to cause the rays to follow paths that contain the same number of waves. Look up optical path difference. The rays on the outer edges of the lens travel a greater distance through the air but a shorter distance through the lens. The greater geometric length in air is just compensated by the shorter path through the lens. Simply, a lens does not create any additional optical path difference or phase shift.

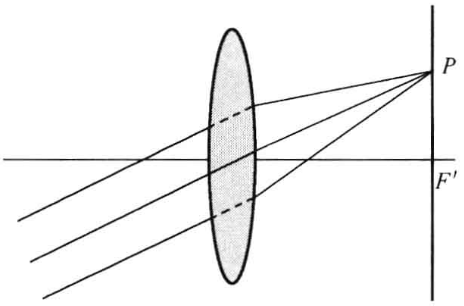


Fig. 15-18 A property of thin lens

15.7 Interference by Division of Amplitude

15.7.1 Equal-inclination interference

Under natural conditions, thin films such as a soap bubble or an oil slick on water have a multicolored appearance that often changes while you are watching them. This is an interference phenomenon caused by interference of light waves reflected from the front and back surfaces of a thin transparent film. In this case, the amplitude of incident light wave is divided into two parts because of reflecting of the film. This interference, one of interferences by division of amplitude, is often called thin film interference. Fig. 15-19 shows how light reflected from the top and bottom surfaces of a film of uniform thickness can interfere. One ray of monochromatic light is incident on a transparent thin film at a small angle i. Some light is reflected at B and some is refracted toward C. The refractive index n of the transparent film is larger than that of air. The refracted light is then partially reflected at C, emerging from the film along DE'. The two reflected waves of rays a_1 and a_2 are coherent because they both originated from the same point A on the monochromatic source. When rays a_1 and a_2 are intercepted by the observer's eye, interference pattern is observed.

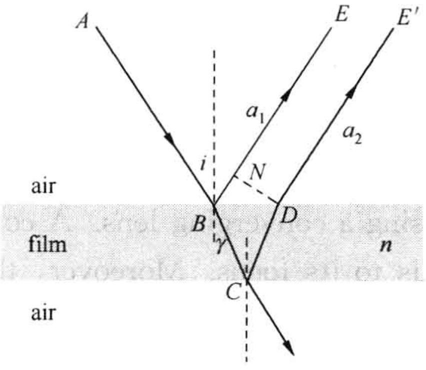


Fig. 15-19 Thin film interference

Let e be thickness of the film, n be refractive index of the film material, and λ_n be wavelength in film ($\lambda_n = \lambda/n$). Because the refractive index of air is very near to 1, for the common use, it is safety to use the wavelength in vacuum λ to represent the wavelength in air. From Fig. 15-19, the optical path difference of two coherent rays a_1 and a_2 is

$$\delta = n(BC + CD) - BN + \lambda/2 \tag{15-21}$$

where the term $\lambda/2$ is introduced because of half-wave loss of the reflection on front of the film. Since $DN \perp BN$ and refractive angle is γ, thus

$$BC = CD = e/\cos\gamma$$

$$BN = BD\sin i = 2e\tan\gamma\sin i = \frac{2e\sin\gamma\sin i}{\cos\gamma}$$

From refraction law, we have

$$\sin i = n\sin\gamma$$

Substituting above relations into Eq. (15-21), we obtain

$$\delta = \frac{2ne}{\cos\gamma}(1 - \sin^2\gamma) + \frac{\lambda}{2} = 2ne\cos\gamma + \frac{\lambda}{2}$$

That is

$$\delta = 2e\sqrt{n^2 - \sin^2 i} + \frac{\lambda}{2} \tag{15-22}$$

Constructive or destructive interferences occur here when

$$\delta = \begin{cases} 2k\dfrac{\lambda}{2}, & k = 1,2,3,\cdots, \quad \text{bright fringes} \\ (2k+1)\dfrac{\lambda}{2}, & k = 1,2,3,\cdots, \quad \text{dark fringes} \end{cases} \tag{15-23}$$

Eq. (15-22) and Eq. (15-23) tell us that for a given incident wavelength λ, the optical path difference varies only owing to the variation of the incident angle i due to e and n keep unchanged for a plane-parallel film. The fringe pattern that depends only on the incident angle is called equal-inclination interference.

We now discuss the applications of interference of plane-parallel films.

1. Antireflective coating

The light energy lost much when light is reflected from many interfaces in a complex optical instrument. For example, the reflecting lose of light energy is about 50% of the incident light as the optical instrument is composed of six lenses. Thus, if optical devices consist of a large number of surfaces, the losses of light energy at the interfaces can be considerable. In order to reduce these losses, lens surfaces, like a camera lens, are often coated with a plane-parallel thin film of transparent structure whose refractive index being less than that of the lens. This transparent thin film is called antireflective coating.

Example 15-7 In Fig. 15-20, a glass optical component of refractive index $n_1 = 1.50$ is coated on one side with transparent material MgF_2 to reduce reflection from optical component. The refractive index of MgF_2 is $n_2 = 1.38$. What are the least thickness and optical-thickness of MgF_2 coating which produce a minimum reflection with incident light of wavelength $\lambda = 550$nm? Assume that the incident light is approximately perpendicular to the component surface.

Solution The product of refractive index and thickness of the material is called optical-thickness. Now, we seek destructive interference between light rays 1 and 2 in Fig. 15-20. At the upper and lower surfaces of MgF_2 coating, a phase shift of π exists respectively. Thus, the optical path difference between rays 1 and 2 is

$$\delta = 2n_2 e$$

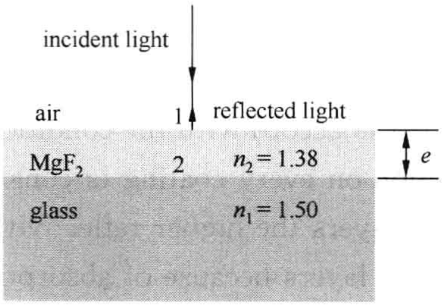


Fig. 15-20 For Example 15-7

For destructive interference, we have

$$\delta = (2k+1)\frac{\lambda}{2}, \quad k = 0,1,2,\cdots$$

Therefore, the thickness of the coating is

$$e = (2k+1)\frac{\lambda}{4n_2}$$

and the optical-thickness of the coating is

$$n_2 e = (2k+1)\frac{\lambda}{4}$$

Substituting $k=0$, $\lambda = 5.50\times10^{-7}$ m and $n_2 = 1.38$ into above two equations, we obtain

$$e = \frac{5.50\times10^{-7}}{4\times1.38} = 9.96\times10^{-8}\,(\text{m}) \approx 100\ (\text{nm})$$

$$n_2 e = \frac{5.50\times10^{-7}}{4} = 1.375\times10^{-7}\,(\text{m}) \approx 138\ (\text{nm})$$

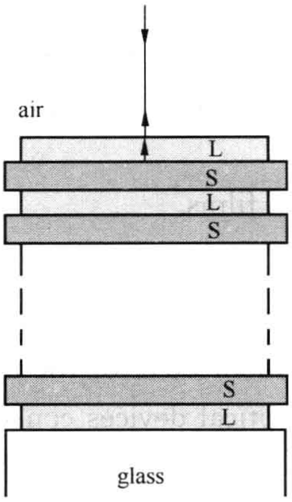


Fig. 15-21 High-reflective coating

2. High-reflective coating

Another important application of film interference is converse of the procedure just discussed. The optical surface is coated by suitable transparent material when high reflection is desired. The coating film, which is usually multicoating, is called high-reflective coating. The reflecting mirror of resonant cavity of a He-Ne laser device, for example, is coated multicoating (15~17 layers) to reach 99% reflectivity for the laser of $\lambda_0 = 632.8$ nm. In order to increase the reflectivity of light, lens surface is often coated with large refractive index films (L) and small refractive index films (S) alternatively, as shown in Fig. 15-21. ZnS, refractive index $n=2.35$, is usually used as a large refractive index material, and MgF_2, $n=1.38$, is usually used as a small refractive index material. When λ_0 is the light wavelength whose high reflection is desired, the optical thickness of each coating is $\lambda_0/4$. Since the light is reflected by each coating, a π phase shift occurs only on one surface. For normal incidence, the optical path difference of reflecting rays on each coating is

$$\delta = 2ne + \frac{\lambda_0}{2}$$

if $ne = \lambda_0/4$, then

$$\delta = 2 \times \frac{\lambda_0}{4} + \frac{\lambda_0}{2} = \lambda_0$$

It is accord with the condition of constructive interference. Because the interference of reflection on every coating is constructive interference, high reflectivity can be obtained. The more layers the higher reflectivity. However, the layers of the films should be no more than 15~17 layers because of absorption.

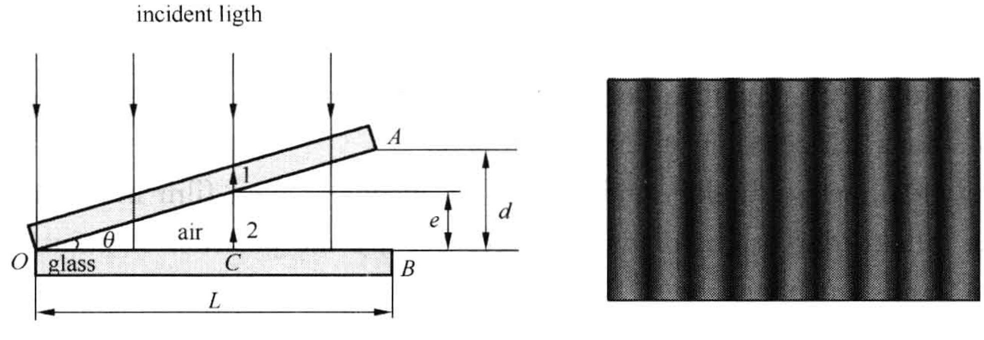
15.7.2 Equal-thickness interference

Thin film interference depends on film thickness, wavelength of light and refractive indices. Equal-thickness interference is observed when the incident angle i keeps a constant on the surface of film that varies in thickness. In this case, the conditions of interference are the same for points having the same values of thickness e. Two arrangements to obtain equal-thickness interference will be discussed.

1. Wedge-shaped film

Let us discuss the interference pattern produced by a film of varying thickness called a wedge-shaped film which consists of two non-parallel plane surfaces.

Example 15-8 In Fig. 15-22(a), two flat glass plates are separated at the right end by a thin spacer with a thickness d, and contact with each other at the left end. One beam parallel light incidents normally on the upper surface of the wedge-shaped arrangement.



(a) (b)

Fig. 15-22 Experiment of wedged-shaped film:

(a) the scheme; (b) interference fringes

(1) Give the conditions of interference and locate the interference maximum with respect to the left end.

(2) What is the distance between two adjacent bright (or dark) fringes?

(3) Is there a bright or dark fringe at the contact edge, i. e. the left end?

Solution (1) So far as that the common light source (except for laser) is concerned, the thickness of the glass plates is very large compared with the incident wavelength. The observable interference fringes are due entirely to the wedge-shaped air gap. The ray 2 which reflected from the upper surface of the second glass undergoes a half-wave loss at reflecting point C. The optical path difference between coherent waves of rays 1 and 2 is

$$\delta = 2n_{\text{air}}e + \frac{\lambda}{2} = 2e + \frac{\lambda}{2}$$

where e is the thickness of wedge-shaped air film at point C. Then, the conditions of interference are

$$\begin{cases} \text{bright fringes} \quad \delta = 2e + \dfrac{\lambda}{2} = 2k\dfrac{\lambda}{2}, & k = 1,2,3,\cdots \\ \text{dark fringes} \quad \delta = 2e + \dfrac{\lambda}{2} = (2k+1)\dfrac{\lambda}{2}, & k = 0,1,2,\cdots \end{cases} \tag{15-24}$$

From geometry of Fig. 15-22, we have

$$e = \frac{d}{L}x$$

where $x = OC$, d is the distance between A and B. Therefore, interference maxima occur at position for which

$$2\frac{d}{L}x + \frac{\lambda}{2} = 2k\frac{\lambda}{2}$$

or

$$x = \frac{L}{2d}\left(k - \frac{1}{2}\right)\lambda, \quad k = 1,2,3,\cdots$$

(2) From Eq. (15-24), the distance l between adjacent bright or dark fringes is

$$l = \frac{e_{k+1} - e_k}{\sin\theta} = \frac{1}{\sin\theta}\frac{\lambda}{2}$$

where θ is the wedge angle. Since θ is small, we have

$$l = \frac{L\lambda}{2d}$$

(3) The thickness of contact edge is $e=0$. From Eq. (15-24), we have $\delta=\frac{\lambda}{2}$ which is the one-half a wavelength loss due to reflection from the upper surface of the second glass. Therefore, there is a dark fringe where the plates touch.

From this example we understand that for a wedge-shaped film in the normal incident condition, the same order bright or dark fringes occupy the position of points for which the film thickness keeps unchangeable. That means a series of parallel equally spaced straight fringes can be seen on the upper surface of the film, as shown in Fig. 15-22(b). These fringes are called equal-thickness fringes. Because the larger the wedge angle θ is, the denser the interference fringes are, we can only observe interference fringes when θ is very small. The interference of wedge-shaped film may be applied both check on smoothness of an optical flat plane or to measure a small angle or a small thickness.

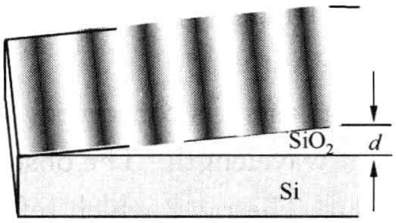


Fig. 15-23 For Example 15-9

Example 15-9 A piece of SiO_2 film is formed by oxygenation over a perfectly flat piece of Si. To determine the thickness of the SiO_2 film precisely, it is usually corroded to a wedge shape as shown in Fig. 15-23. The refractive indexes of SiO_2 and Si are $n_{SiO_2}=1.50$ and $n_{si}=3.42$, respectively. One light beam of wavelength $\lambda=589.3$ nm incident normally to the film. 7 bright fringes are seen over the length of film. Calculate the thickness d of the SiO_2 film.

Solution On the SiO_2 film where the thickness is e, the optical path difference between two reflected light rays which are reflected from the upper and lower surfaces of the film is

$$\delta = 2n_{SiO_2}e$$

When

$$\delta = 2n_{SiO_2}e = k\lambda, \quad k=0,1,2,\cdots$$

the bright fringes occur. The thickness of the SiO_2 film is

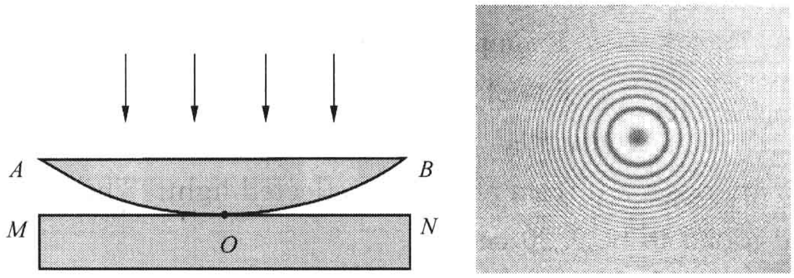
$$e_k = \frac{k\lambda}{2n_{SiO_2}}, \quad k=0,1,2,\cdots$$

There is a bright fringe at the contact edge where $e_k=0$ or $k=0$. Thus, the thickness d should correspond to the bright fringe of $k=6$. Therefore, from above equation, we obtain

$$d = e_6 = \frac{6\lambda}{2n_{SiO_2}} = \frac{6\times 589.3\times 10^{-9}}{2\times 1.5} = 1.18\times 10^{-6}(\mathrm{m})$$

2. Newton's rings

One classical example of equal-thickness fringes is Newton's rings. A plane-convex lens with radius of curvature R rests on an accurately plane glass plate. An air film of gradually increasing thickness outward is formed between the lens and the glass plate, as shown in Fig. 15-24(a). The thickness of the film at the point of contact O is zero. If it is illuminated from above by monochromatic light with wavelength λ then the light reflected from the surface AOB interferes with the light reflected from the surface MON. A series of alternate bright and dark concentric rings appear around the point of contact. These circular interference fringes, called Newton's rings, are shown in Fig. 15-24(b). It is associated with the variable thickness of the air film between the lens and the plate. We will find out the relations among r, λ and R. Here r represents one of radius of Newton's rings.



(a) (b)

Fig. 15-24 Newton's Rings: (a) Arrangement of Newton's rings; (b) Interference pattern

In Fig. 15-25, at point P, the thickness of film is e. The optical path difference between the two reflected coherent waves is

$$\delta = 2n_{\text{air}}e + \frac{\lambda}{2} = 2e + \frac{\lambda}{2}$$

where $\lambda/2$ corresponds to the one half-wave loss due to the reflection of light on the bottom surface of the air film. The conditions that deduce circular bright or dark fringes are

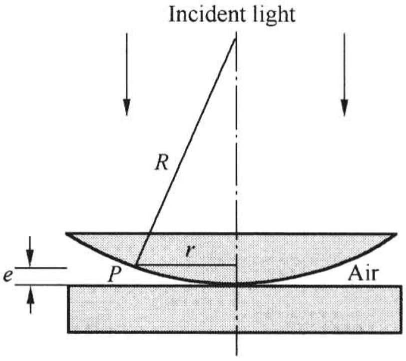


Fig. 15-25 Illustration for Newton's rings experiment

$$\begin{cases} \text{bright fringes} \quad 2e + \dfrac{\lambda}{2} = k\lambda, & k = 1,2,3,\cdots \\ \text{dark fringes} \quad 2e + \dfrac{\lambda}{2} = (2k+1)\dfrac{\lambda}{2}, & k = 0,1,2,\cdots \end{cases} \tag{15-25}$$

There is a dark spot at the point of contact ($e=0$), as shown in Fig. 15-24(b). From geometry of Fig. 15-25, we have

$$r^2 = R^2 - (R-e)^2 = e(2R-e)$$

Assuming that $e \ll 2R$, thus

$$e = \frac{r^2}{2R} \tag{15-26}$$

Substituting Eq. (15-26) into Eq. (15-25), we obtain

$$\begin{cases} \text{bright fringes} \quad r_k = \sqrt{(2k-1)R\dfrac{\lambda}{2}}, & k = 1,2,3,\cdots \\ \text{dark fringes} \quad r_k = \sqrt{kR\lambda}, & k = 0,1,2,\cdots \end{cases} \tag{15-27}$$

Let the radius of the k_1th dark ring be r_1, the radius of the k_2th dark ring be r_2 and then, from Eq. (15-27), we have

$$r_1^2 = k_1 R\lambda$$
$$r_2^2 = k_2 R\lambda$$

Therefore

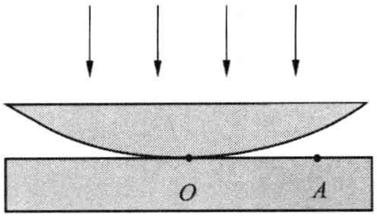


Fig. 15-26 For Example 15-10

$$R = \frac{r_2^2 - r_1^2}{(k_2 - k_1)\lambda}$$

We can calculate the radius of curvature R or the wavelength λ by using above equation.

Example 15-10 Fig. 15-26 shows a plane-convex lens with a radius of curvature $R=4.0$ m lying on a flat glass plate and illuminated from above by parallel beams with wavelength λ. The Newton's rings are observed from above by reflected light. The radius of the fifth bright ring is measured and found to be 0. 30 cm.

(1) Find the wavelength λ.

(2) In Fig. 15-26, if $OA = 1.0$ cm, how many bright rings would be seen within the range of OA?

Solution (1) The radii of bright rings are

$$r_k = \sqrt{(2k-1)R\frac{\lambda}{2}}, \quad k = 1,2,3,\cdots$$

Thus

$$\lambda = \frac{2r_k^2}{(2k-1)R}$$

With $k=5$, $r_k = r_5 = 0.30\ \text{cm} = 3.0\times10^{-3}$ m, $R=4.0$ m, we obtain

$$\lambda = 5.0\times10^{-7}\ \text{m} \quad \text{or} \quad \lambda = 500\ \text{nm}$$

(2) From $r_k = \sqrt{(2k-1)R\frac{\lambda}{2}}$, we have $k = \frac{r_k^2}{R\lambda} + \frac{1}{2}$; Let $r_k = OA = 1.0\ \text{cm} = 1.0\times10^{-2}$ m, and entering known values yields

$$k = \frac{(1.0\times10^{-2})^2}{4.0\times5.0\times10^{-7}} + \frac{1}{2} = 50.5$$

Therefore, there are 50 bright rings which would be seen within the range of OA.

15. 8 Michelson Interferometer

An interferometer is an apparatus that can be used to measure lengths or changes in length with great accuracy by utilizing interference between two light waves. One particularly historical interferometer is Michelson interferometer that developed by Albert A. Michelson (1852—1931). A schematic diagram of Michelson interferometer is shown in Fig. 15-27. S is an extended source. G_1 and G_2 are two beam splitters with the same material and thickness, i. e., a beam of light that strikes on G_1 and G_2 gets partially reflected and partially transmitted. M_1 and M_2 are two good quality plane mirrors with very high reflectivity. One of the mirrors (usually M_1) is fixed and the other (usually M_2) is capable of moving away or towards the glass plate G_2 along an accurately machined track by means of a screw. In the normal adjustment of the interferometer, the mirrors M_1 and M_2 are perpendicular to each other and G_1 or G_2 makes a 45° angle with M_2.

A beam of light emitted by the monochromatic light source S gets partially reflected and partially transmitted by G_1. The reflected portion labeled (1) travels a distance l_1 to mirror M_1. It is reflected by M_1 and partially transmitted by G_1 and arrives finally at the eye E of the observer. The transmitted portion of the original beam labeled (2) travels a distance l_2 to mirror M_2, which reflects the beam (2) back toward G_1. It is partially reflected and arrives finally at the eye E of the observer. Note that beam (1) passes through G_1 three times in reaching E, while beam (2) passes through it only once. The plate G_2, called the compensating plate, is introduced into the path of beam (2) to equalize the optical path lengths of two beams (1) and (2). As being divided from the same point on the source, the beam (1) and (2) are coherent and their interference pattern depends only on the difference lengths l_1 and l_2. Considering the effect of compensator plate G_2, the mirror M_2 may be replaced with its virtual image M_2'. Beams (1) and (2) can be considered as emitted from coherent sources M_1 and M_2' with optical path $2l_1$ and $2l_2$, respectively. If the mirror M_1 and M_2 are exactly perpendicular to each other, the interference fringes are a series of circular equal-inclination fringes. If the mirror M_1 and M_2 are nearly but not exactly perpendicular to each other, they construct a wedge-shaped film. The observer then can see a series of parallel equal-thickness fringes. If the optical path difference $2|l_2 - l_1|$ is an integral number of wavelength of incident beam, beams (1) and (2) will constructively interfere, and they will destructively interfere if $2|l_2 - l_1|$ is an odd half-integral number of the wavelength.

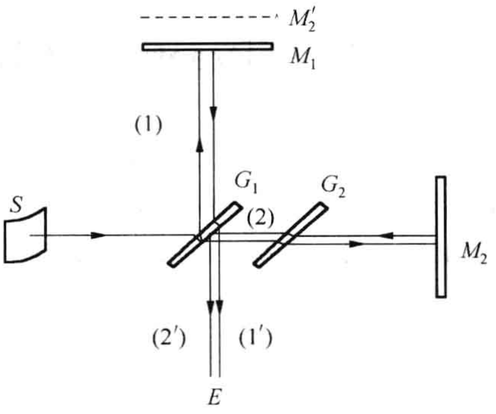


Fig. 15-27 A schematic diagram of Michelson interferometer

If the mirror M_2 moves backward or forward along the direction perpendicular to its surface, the length l_2 is changed, and the pattern of interference fringes is shifted. This shift can be seen by using a telescope with a cross-hair eyepiece. When M_2 moves a distance of $\lambda/2$, the shift of optical path difference is λ, and the observer can see a total fringe passing the cross-hair, thus, the distance d that M_2 moves is determined by counting the number of fringes N that pass the cross-hair as the movement of M_2 is made, that is

$$d = N\frac{\lambda}{2} \tag{15-28}$$

A careful observer using a well-constructed instrument can detect a shift as small as 1/40 of one fringe spacing. If $\lambda = 500$ nm, it means that we can measure distance as small as 7 nm using Michelson interferometer. By such techniques, the lengths of objects can be expressed in terms of the wavelength of light. In Michelson's day, the standard of length—the meter, was chosen by international agreement to be the distance between two fine scratches on a certain metal bar preserved at Sevres, near Paris. Michelson was able to show, using his interferometer, that the standard meter was equivalent to 1553163. 5 wavelengths of a certain monochromatic red light emitted from a light source containing cadmium. Michelson, for this careful measurement, received the 1907 Nobel Prize in physics, American's first Nobel Prize winner in science.

15. 9 Diffraction of Light and Huygens-Fresnel's Principle

Interference of light gives the evidence that light has wave-like properties. According to the wave theory, waves should bend around obstacles. The bending of sound waves around obstacles, the diffraction of sound waves, is constantly observed in our everyday life. Since light has a wave-like nature, it is expected that light wave exhibits diffraction. We will discuss the diffraction phenomena of light in the following.

15. 9. 1 Two kinds of diffraction: Fresnel diffraction and Fraunhofer diffraction

Fig. 15-28 shows a plane light wave strikes on a long narrow slit of width d. According to geometrical optics, the region AB on screen E directly opposite the slit would be illuminated and the remaining portion, the geometrical shadow, to be absolutely dark. However, if the slit width a is small enough to make the ratio λ/a sufficiently large, the light intensity in the region AB is not uniform and the incident light certainly "flares out" into the geometric shadow of the slit, as shown in Fig. 15-29. The narrower the slit is, the more flaring out the light is. This flaring out of a light wave when it passes through a narrow opening is known as diffraction of light and intensity distribution on viewing screen is called diffraction pattern. It is possible to see diffraction fringes if one peers between two fingers at a distance light source. With the fingers held close to the eye and brought together to form a slit opening, dark fringes will be seen before the light is shut out.

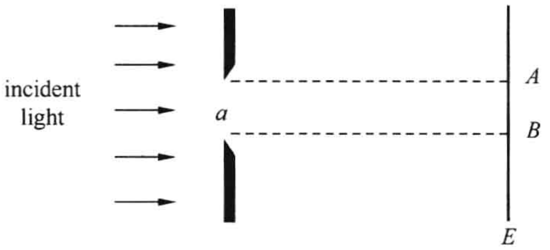


Fig. 15-28 A sharp shadow will be cast in the region AB of the screen E according to geometrical optics

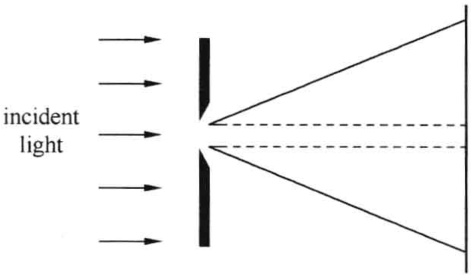


Fig. 15-29 The incident light flares out in the geometrical shadow

The diffraction phenomena are usually divided into two categories: Fresnel diffraction and Fraunhofer diffraction.

(1) Fresnel diffraction refers to the cases in which the source of light or the screen or both are at finite distances from the diffracting obstacle, as shown in Fig. 15-30(a). Experimentally Fresnel diffraction is easily achieved, but mathematical analysis proves to be very complex for it concerns non-plane wavefronts.

(2) Fraunhofer diffraction refers to the cases in which the source of light and the screen are at infinite distances from the obstacle, as shown in Fig. 15-30(b). Experimentally, this is often achieved by using an ordinary source and two lenses as Fig. 15-30(c) illustrates. Mathematically, the analysis on Fraunhofer diffraction is easy to achieve because only parallel rays

are concerned. Taking advantage of the highly paralleled laser beam, if choosing laser as the source of light, Fraunhofer diffraction may be achieved readily without lenses. Actually, lasers have been employed as sources to demonstrate the wave properties of light for years. We will only focus on Fraunhofer diffraction in following content because it is easy to handle both in experiment and in mathematics.

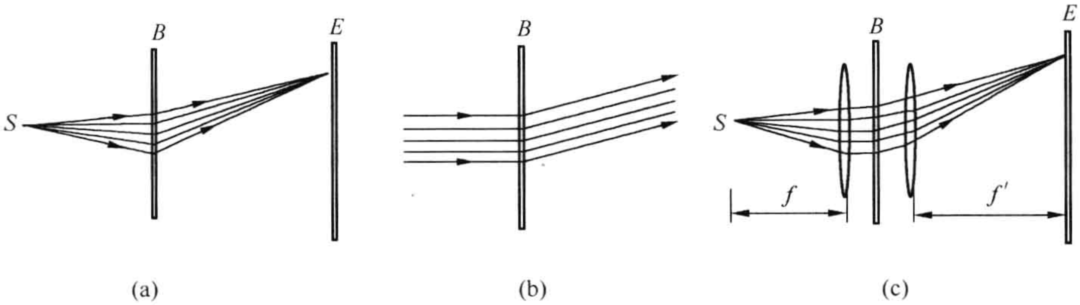


Fig. 15-30 (a) Fresnel diffraction; (b) and (c) Fraunhofer diffraction

15.9.2 Huygens-Fresnel's principle

The penetration of light waves into the geometrical shadow can be explained with the aid of Huygens' principle which has been introduced in 13.5. This principle, however, gives no information on the intensity distribution of light on viewing screen. French physicist Augustin Fresnel (1788—1827) supplemented Huygens' principle with the concept of interference of secondary wavelets. Taking into account the amplitudes and phases of the secondary wavelets makes it possible to find the amplitude of the resultant wave for any point in space, i. e. the intensity distribution of light on viewing screen. Huygens' principle developed in this way was known as the Huygens-Fresnel principle.

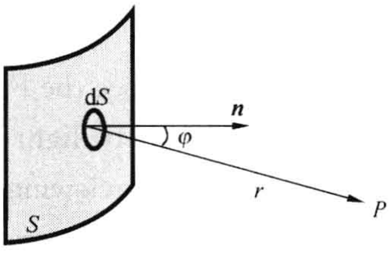


Fig. 15-31 The explanation of diffraction

In Fig. 15-31, according to the Huygens-Fresnel principle, every element of wavefront S emits a secondary spherical wavelet whose amplitude is proportional to the element area $\mathrm{d}s$ and is inversely proportional to the distance r. Consequently, the oscillation at point P where spherical wavelets from each section $\mathrm{d}s$ converge together can be written as

$$\mathrm{d}y = \frac{k(\varphi)\mathrm{d}s}{r}\cos\left(\omega t - \frac{2\pi r}{\lambda}\right) \tag{15-29}$$

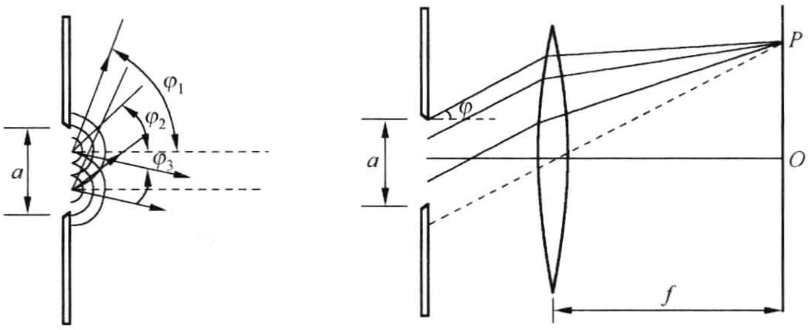
where r is the distance from surface element $\mathrm{d}s$ to point P, $k(\varphi)$ is a function of angle φ between r and the normal of $\mathrm{d}s$, ω is the circular frequency of light and λ is the wavelength of light. Therefore, the resultant oscillation at point P is given by

$$y = \int_S \frac{k(\varphi)}{r}\cos\left(\omega t - \frac{2\pi r}{\lambda}\right)\mathrm{d}s \tag{15-30}$$

Eq. (15-30) is an analytical expression of Huygens-Fresnel principle.

15.9.3 Parallel beams interference—a simplified discussion

Huygens-Fresnel principle can be applied to simplify the determination of intensity distribution of light on viewing screen. When a beam of plane wave is incident normally on a slit of width a, the area of the slit opening is a part of wavefront whose every point could be regarded as a source of secondary wavelet which emits a spherical wavelet with radial rays in all directions. Fig. 15-32 (a) shows a plane wavefront approaching the slit and singles out two sources of wavelets and three arbitrary parallel beams with different directions represented by diffraction angles of φ_1, φ_2 and φ_3 respectively. Actually, the rays emerging from slit are made up of a random array of parallel beams in all directions corresponding to diffraction angles φ changing from 0 to 180°. We now focus on one parallel beam in arbitrary direction represented by the diffraction angle φ as shown in Fig. 15-32(b). According to geometrical optics, this parallel beam will convergent to point P on the focal plane where the screen is located. All rays in this parallel beam are coherent with each other, and their superposition determines the intensity of light at point P. We will study the diffraction of light from a single slit with above idea.



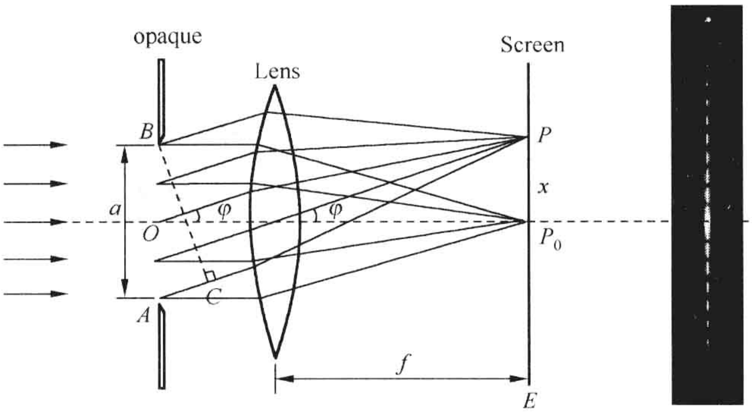
(a) (b)

Fig. 15-32 (a) Wavelet rays in all different directions; (b) a parallel beam in arbitrary direction represented by diffraction angle

15.10 Diffraction from a Single Slit

We now introduce the Fraunhofer diffraction produced by a single slit. Suppose a beam of parallel monochromatic light is incident from the left of an opaque plate with a narrow horizontal slit of width a, a viewing screen is placed at the focal plane of the lens on the right of the slit, as Fig. 15-33(a) illustrates. The beam spreads out perpendicularly after passing through the slit and Fig. 15-33(b) shows the diffraction pattern formed on the screen. It consists of a broad central maximum and many narrower and less intense maxima on either side.

Let us now discuss how light is superimposed at point P by using Huygens-Fresnel principle. One set of parallel rays of wavelets emerging from the slit at an angle φ will be focused at point P on the screen, as shown in Fig. 15-33(a). The angle φ between the emerging direction and the normal direction of the slit is usually known as diffraction angle. Make a plane BC be perpendicular to the diffraction rays of angle φ. The optical paths of the rays from the wavefront BA to the plane BC are totally different. Note that optical difference between the upmost ray and the bottommost ray is $AC = \delta = a\sin\varphi$ and it is the maximum optical path difference in the rays. We will see that it is the maximum optical path difference $a\sin\varphi$ determines the intensity at point P on the screen.



(a) (b)

Fig. 15-33 Fraunhofer diffraction from a single slit
(a) Experiment arrangement; (b) diffraction pattern

15.10.1 Central bright fringe

When rays travel along the optical axis, that is $\varphi = 0$, these rays are all in phase at the plane of the slit, and they remain in phase at the center point P_0 on the screen E because a lens does not lead to any additional optical path difference. As a result, constructive interference causes a bright central fringe, also called central maximum, at point P_0 on the screen.

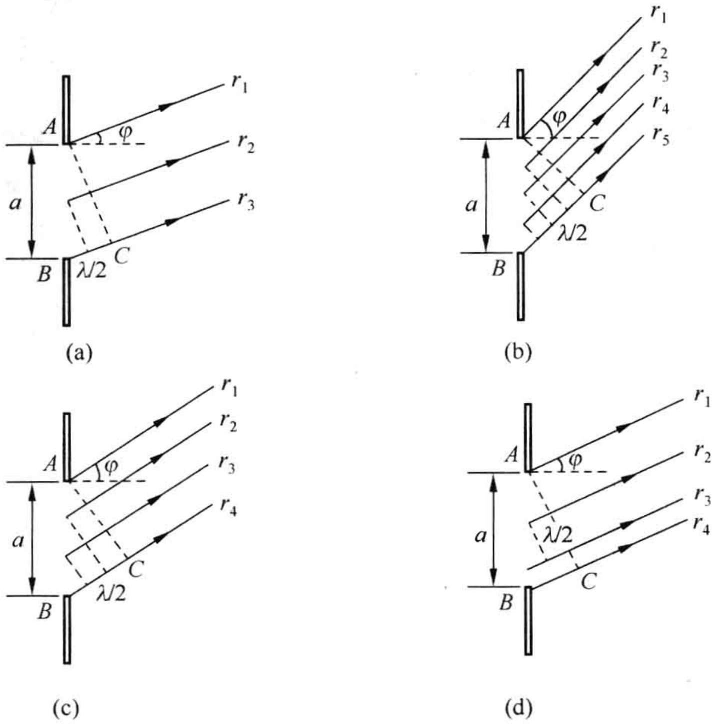
15.10.2 Location of the dark fringes

When rays travel at an angle φ, each ray undergoes a different path to a common location P on viewing screen E, and they can arrive on or out of phase. To find the dark fringes, we will introduce a clever and effective method that involves paring up all the rays coming through the slit and then finding what conditions cause the wavelets of the rays in each pair to destructive interference. The wavefront BA on the slit opening is mentally divided into many zones of equal widths a/n and a bunch of parallel planes, which represented by dash lines and are always perpendicular to the emitting rays of diffraction angle φ in Fig. 15-34, being $\lambda/2$ apart between any adjacent two of them. In Fig. 15-34(a), the wavefront BA is mentally divided into two zones of equal widths $a/2$ and three rays labeled r_1, r_2 and r_3 are drawn. Ray r_1 originates at the top of the slit, r_2 at the center and r_3 at the bottom. $\Delta\delta$ is the optical path difference between r_2 and r_1. If the diffraction angle φ is given so that $\Delta\delta$ equals to $\lambda/2$ (in this case, we usually say that wavefront BA is divided into two one-half a wavelength zones), ray r_1 and r_2 will be out of phase and interfere destructively. A ray from slightly above the central ray r_2 and one from slightly above the bottom ray r_3 will also cancel one another. That is, each ray from the upper one-half a wavelength zone will be cancelled by a corresponding one from the lower zone because they always keep an optical path difference $\lambda/2$, and a dark fringe will occur at convergent point P on the viewing screen. Therefore, the first minimum of the diffraction pattern occurs, when

$$\frac{a}{2}\sin\varphi = \frac{\lambda}{2}$$

or

$$a\sin\varphi = \lambda$$



(a) (b) (c) (d)

Fig. 15-34 For "location of minima" (a) $a\sin\varphi=2\cdot\frac{\lambda}{2}$; (b) $a\sin\varphi=4\cdot\frac{\lambda}{2}$; (c) $a\sin\varphi=3\cdot\frac{\lambda}{2}$; (d) $2\cdot\frac{\lambda}{2}<a\sin\varphi<3\cdot\frac{\lambda}{2}$

If φ is given so that $BC=\delta=a\sin\varphi=2\lambda$ and $\Delta\delta$, the optical path difference between r_{k+1} and r_k $(k=1,2,3,4)$, equals to $\frac{\lambda}{2}$, the wavefront BA is divided into four "one-half a wavelength zones", as shown in Fig. 15-34(b). Rays emit from the upper two "one-half a wavelength zones" will interfere destructively with each other at convergent point P, and rays emit from the lower two "one-half a wavelength zones" do likewise. Therefore, the second minimum of the diffraction pattern occurs at point P, when

$$\frac{a}{4}\sin\varphi=\frac{\lambda}{2}$$

or

$$a\sin\varphi=2\lambda$$

By extension of above discussion to the situation that the wavefront BA on the slit opening is divided into even number of one-half a wavelength zones, the general equation for the minima in the diffraction pattern on viewing screen is given by

$$a\sin\varphi=\pm 2k\frac{\lambda}{2},\quad k=1,2,\cdots \tag{15-31a}$$

or

$$a\sin\varphi=\pm k\lambda,\quad k=1,2,\cdots \tag{15-31b}$$

where the sign "±" shows that the same order dark fringes on the screen are symmetrically located about the central maximum. It tells us that in the case of diffraction from a single-slit, dark fringes are produced where the maximum optical path difference (i. e. the optical path difference between the top and the bottom rays) equals to odd number times the wavelength of incidence.

15.10.3 Location of the bright fringes (secondary maxima)

When φ is given so that the wavefront BA on the slit opening is divided into three one-half a wavelength zones, as shown in Fig. 15-34(c), those rays emitting from the top zone and the middle zone will interfere destructively. Only do the rays emitting from the bottom zone remain intensity at point P on the viewing screen. It means that point P will turn out bright fringe with less intensity comparing to the central maximum. This bright fringe which is known as the first secondary maximum can be located with the equation

$$a\sin\varphi = 3 \cdot \frac{\lambda}{2}$$

In terms of above strategy, when the wavefront BA on the slit opening is divided into odd number of one-half a wawelength zones, only the rays emitting from the bottommost zone remains intensity at point P on viewing screen. Therefore, all other bright fringes, i.e. secondary maxima, above and below the central maximum can be located with the following general equation

$$a\sin\varphi=\pm(2k+1)\frac{\lambda}{2}, \quad k=1,2,\cdots \tag{15-32}$$

The fringe pattern is symmetrical about the central maximum as Fig. 15-33(b) shows.

If the diffraction angle φ is given so that $2k\frac{\lambda}{2} < a\sin\varphi < (2k+1)\frac{\lambda}{2}$, as shown in Fig. 15-34(d) for $k=1$, it is detected that the intensity at P will be weaker than that of the closed secondary maximum.

The variation of intensity with distance from the centre of the central maximum has been graphically depicted in Fig. 15-35. It shows that the central maximum is wider and much more intense than those of the sides. The intensity of maxima falls off rapidly as the points go away from the centre O because the greater the order k is, the more the zones are and the smaller the area of uncanceled zones is.

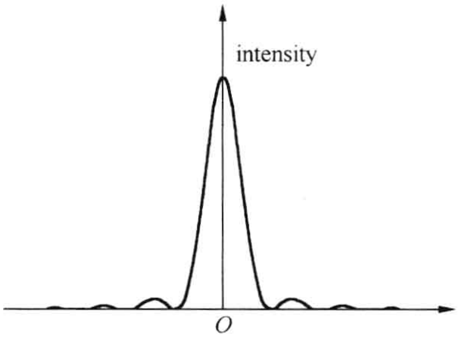


Fig. 15-35 Distribution of intensity of the diffraction pattern from a single slit

The width of the central bright fringe, represented by l_0, is defined as the distance between the positions of the two first minima. The width l of the other bright fringe is the distance between the positions of any other adjacent pair of minima, say, between the kth and the $(k+1)$th minima. Fig. 15-35 shows that l_0 is approximately the double of l.

Example 15-11 In an experiment of Fraunhofer diffraction from a single slit, the width of the slit is $a = 0.1$ mm and the focal length of the lens is 0.5 m. Monochromatic light of wavelength 500nm is incident on the slit. Find

(1) the width of central bright fringe;

(2) the width of the first secondary bright fringe.

Solution (1) The width l_0 of central bright fringe equals the spacing of the two first minima as $\sin\varphi_1 = \lambda$ and $\sin\varphi_2 = -\lambda$. Let x_1 be distance of the first minimum and the centre point P_0. From Fig. 15-33(a), we have

$$l_0 = 2x_1 = 2f\tan\varphi_1$$

For lower order diffraction fringes, φ is very small, thus $\tan\varphi_1 = \sin\varphi_1$.

Hence

$$l_0 = 2f\sin\varphi_1 = \frac{2f\lambda}{a} \tag{15-33}$$

Where

$$a = 0.1\ \mathrm{mm} = 0.1 \times 10^{-3}\ \mathrm{m},\quad f = 0.5\ \mathrm{m}$$
$$\lambda = 500\ \mathrm{nm} = 5.00 \times 10^{-7}\ \mathrm{m}$$

Substituting known values into Eq. (15-33) gives

$$l_0 = 0.005\ \mathrm{m} = 5\ \mathrm{mm}$$

(2) The width l of the first secondary bright fringe equals the spacing of the first minimum $a\sin\varphi_1 = \lambda$ and the second minimum $a\sin\varphi_2 = 2\lambda$.

From Fig. 15-33 (a), the distance of the second minimum and the centre point P_0 is $f\tan\varphi_2 \approx f\sin\varphi_2$, thus

$$l = f\sin\varphi_2 - f\sin\varphi_1 = f\left(\frac{2\lambda}{a} - \frac{\lambda}{a}\right) = \frac{f\lambda}{a} \tag{15-34}$$

Substituting the values of f and λ into above equation yields

$$l = 0.0025\ \mathrm{m} = 2.5\ \mathrm{mm}$$

Eq. (15-33) and Eq. (15-34) imply that the width of other bright fringes equals approximately $f\lambda/a$ when φ is very small and the central bright fringe is approximately double as wide as other bright fringes. And it is obvious that as wavelength λ is unchanged, the smaller the width a is, the wider the fringes become and as the width a is determined, the longer the wavelength λ is, the wider the fringes become.

Example 15-12 A slit of width $a = 2.5\ \mu\mathrm{m}$ is illuminated by white light.

(1) What is the diffraction angle that the first minimum for red light $\lambda = 650$ nm falls at?

(2) What is the wavelength λ' of the light whose first diffraction maximum (not counting the central maximum) falls at φ_1, thus coinciding with the first minimum for red light?

Solution (1) From equation $a\sin\varphi = k\lambda$, where $k = 1$ for the first minimum, we have

$$\sin\varphi_1 = \frac{k\lambda}{a} = \frac{\lambda}{a}$$

That is

$$\varphi_1 = \arcsin\frac{\lambda}{a}$$

Substituting known values, $\lambda = 650$ nm, $a = 2.5\ \mu\mathrm{m}$, we obtain

$$\varphi_1 = 15^\circ$$

(2) From equation $\varphi' = (2k+1)\frac{\lambda}{2}$, where $k = 1$ and $\varphi' = 15^\circ = \varphi_1$ for the first maximum, thus

$$a\sin\varphi_1 = \frac{3}{2}\lambda'$$

Problem (1) gives

$$a\sin\varphi_1 = \lambda$$

Therefore

$$\lambda' = \frac{2}{3}\lambda = \frac{2}{3} \times 650 = 433\ (\mathrm{nm})$$

This is the wavelength of violet. The first diffraction maximum of violet light with wavelength 433 nm will always coincide with the first minimum of red light with wavelength 650 nm, no matter what the slit width is. If the slit is relatively narrow, the overlapping of the angle φ will be relatively large, and conversely. Therefore, each fringe is a rainbow of color except for the central maximum when white light falls on a slit.

15.11 Diffraction Grating

A logical extension of Young's double-slit interference experiment is to increase the number of slits from two to a large number of N. An optical component consisting of a large number of parallel, closely spaced slits- for example, as many as 10^3/mm is not uncommon is called a diffraction grating. A diffraction grating can be used to determine the wavelength of light with high precision. Diffraction gratings are usually made by ruling equally spaced parallel grooves on a polished glass plate using a diamond-tipped cutting tool. The grooves are effectively opaque and they scatter the light and the space between the grooves behaves as a slit. Thus, the action of a diffraction grating can be described in terms of a regular array of parallel slits. The width of a slit is a and the width of a groove is b. $d = (a+b)$ is called grating constant as illustrated in Fig. 15-36. Usually d has orders of $10^{-6} \sim 10^{-5}$ m, meaning that there are 1000～10000 grooves per centimeter. Gratings are widely used to measure wavelengths and to study the structure and intensity of spectral lines. Intensity patterns of bright and dark fringes can be seen on the viewing screen when monochromatic light passes through a single or double slit. The fringe patterns also result when light falls on a grating. Fig. 15-37 shows a grating with total number of slits N ($N=5$). At any point on the screen, the available light intensity from each slit, considered separately, is given by the diffraction pattern of that slit. The diffraction patterns for each separate slit coincide with each other because parallel rays in Fraunhofer diffraction are focused on the same point of the lens' focus plane. The diffraction rays will interference with each other since they are coherent. The diffraction patterns for N equals 1, 2, 3, 4, 5 and 20 are shown respectively in Fig. 15-38. Two important changes occur when the number of slits increases:

(1) The bright fringes become narrower and brighter;

(2) Faint secondary maxima appear between fringes.

It indicates that the diffraction pattern of a grating is total result of interference of the lights from slits and diffraction of a single slit. The more slits, the more principal maxima and the weaker the secondary maxima become. As N increase, perhaps to 10^4 for a useful grating, the bright fringes become very sharp and bright indeed while the secondary maxima become so reduced in intensity as to be negligible in their effects. We will ignore the secondary maxima in what follows and discuss the locations of the bright fringes of a diffraction grating. The optical path difference between rays from adjacent slits shown in Fig. 15-38 is

$$\delta = (a+b)\sin\varphi$$

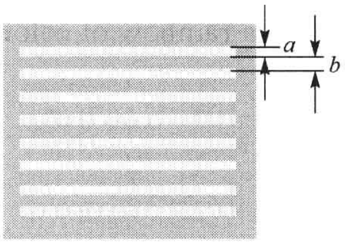


Fig. 15-36 Grating constant $d=a+b$

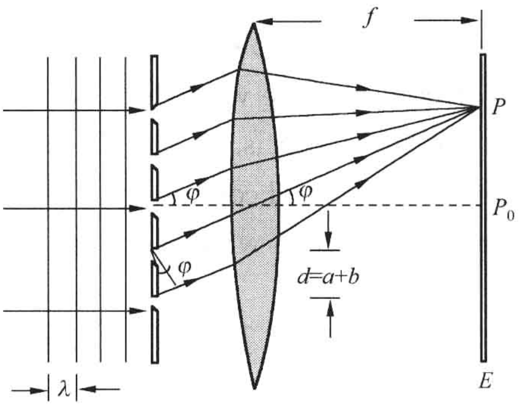


Fig. 15-37 Arrangement for a grating of $N=5$

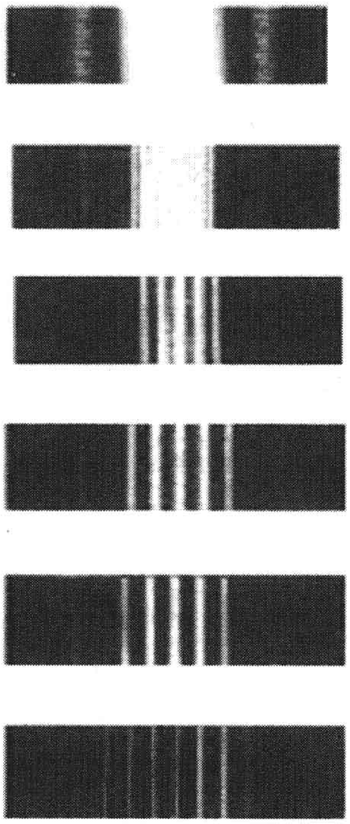


Fig. 15-38 Diffraction pattern for $N=1,2,3,4,5,20$(from up to bottom)

where φ is diffraction angle. Constructive interference creates bright fringes. These sharp bright fringes which are sometimes called principal maxima will occur when

$$(a+b)\sin\varphi = \pm k\lambda, \quad k = 0,1,2,3,\cdots \tag{15-35}$$

Where λ is the wavelength of incident light, k is the order of the fringe (principal maximum) and $k=0$ corresponding to the central fringe (or central principal maximum). Eq. (15-35) is called grating equation which gives the condition necessary to obtain constructive interference for a diffraction grating in the case of normal incidence.

Missing orders occur for a diffraction grating when an interference maximum coincides with a diffraction minimum of a single slit. For example, if both the conditions $(a+b)\sin\varphi = \pm k\lambda$ and $a\sin\varphi = \pm k'\lambda$ are satisfied for a given φ, the kth principal maximum of the grating diffraction is coincident with the k'th minimum of the single-slit diffraction. As a result, there is no the kth principal maximum on the viewing screen, that is such bright fringe disappears.

Eq. (15-35) can be used to study the dependence of the diffraction angle φ on the wavelength λ. When the grating constant $(a+b)$ keeps a constant, the diffraction angle φ of the kth bright fringe is determined only by the wavelength λ. The longer the wavelength λ is, the larger the angleφ is. When we use white light as incident light, except for the central fringe keeps white, all the higher-orders maxima disperse white light into their rainbow of colors on the screen. The whole of all the same order fringe of the incident light emitted by a polychromatic source is called grating spectrum, the fringe on the grating spectrum is called spectrum line. The grating spectrum formed by one material is unique, that is, a kind of material has a characteristic spectrum. We can determine components of material by its grating spectrum.

Example 15-13 A beam of white light ($\lambda = 400 \sim 760$ nm) falls normally on a grating, as shown in Fig. 15-39. The grating constant is $(a+b) = 2.4 \times 10^{-4}$ cm, the focal length is $f = 0.25$ m. Find the distances from the 3rd order spectrum line of violet light ($\lambda_1 = 400$ nm) and the 2nd order spectrum line of red light ($\lambda_2 = 760$ nm) to the centre P_0 on the screen, respectively.

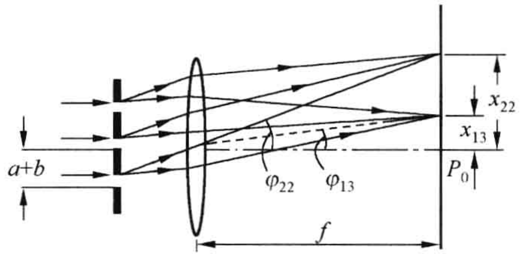


Fig. 15-39 for Example 15-13

Solution Assuming the diffraction angles of the 3rd order spectrum line of the violet light and the 2nd order spectrum line of the red light are φ_{13} and φ_{22}, respectively. The distance from them to the centre P_0 are x_{13} and x_{22}, respectively. From the grating Eq. (15-35), we have

$$\varphi_{13} = \arcsin\left(\frac{3\lambda_1}{a+b}\right) = \arcsin\left(\frac{3 \times 400 \times 10^{-9}}{2.4 \times 10^{-6}}\right) = \arcsin(0.50) = 30^\circ$$

$$\varphi_{22} = \arcsin\left(\frac{2\lambda_2}{a+b}\right) = \arcsin\left(\frac{2 \times 760 \times 10^{-9}}{2.4 \times 10^{-6}}\right) = \arcsin(0.633) = 39^\circ 18'$$

From the geometry shown in Fig. 15-39, we have

$$x_{13} = f \cdot \tan\varphi_{13} = 0.25 \times \tan 30^\circ = 0.14 \text{ (m)}$$

$$x_{22} = f \cdot \tan\varphi_{22} = 0.25 \times \tan 39^\circ 18' = 0.20 \text{ (m)}$$

We can see that x_{13} is smaller than x_{22}. It indicates that the 3rd order spectrum line of the violet light is more near the central maximum line than the 2nd order spectrum line of the red light and the two spectrums overlap with each other. Please note that when $(a+b)$ is very small (e. g. 10^{-6}m), the relationship $\tan\varphi \approx \sin\varphi \approx \varphi$ that is used in discussing of single-slit diffraction may not suitable for discussing of diffraction grating.

Example 15-14 Light of wavelength $\lambda = 643.8$nm is incident obliquely on a grating at an angle $\theta = 30^\circ$ as shown in Fig. 15-40. The diffraction grating has 5000 lines per centimeter.

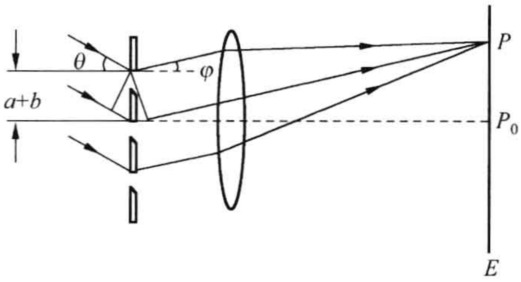


Fig. 15-40 For Example 15-14

(1) Find the diffraction angles for bright fringes corresponding to $k=0$, $k=+1$ and $k=-1$, respectively.

(2) Find the highest order of maximum that can be seen. Compare it with the one in the case of normal incidence.

Solution (1) As light is oblique incidence, the optical path difference between rays from adjacent slits in Fig. 15-40 is

$$\delta = (a+b)(\sin\theta + \sin\varphi)$$

The principal maxima will occur when

$$(a+b)(\sin\theta+\sin\varphi)=\pm k\lambda,\quad k=0,1,2,3,\cdots \tag{15-36}$$

Eq. (15-36) is called grating equation for oblique incidence. In this problem, the grating constant

$$(a+b)=\frac{1}{N}=\frac{1\times10^{-2}}{5000}\ \mathrm{m}=2.0\times10^{-6}\ \mathrm{m}$$

From Eq. (15-36), the diffraction angle for fringe of $k=0$ is

$$\varphi_0=\arcsin(-\sin 30^\circ)=-30^\circ$$

It indicates the 0th order bright fringe located below the centre of the screen.

The diffraction angles corresponding to fringes of $k=+1$ and $k=-1$ are respectively

$$\varphi_{+1}=\arcsin\left(\frac{\lambda}{d}-\sin\theta\right)=\arcsin\left(\frac{634.8\times10^{-9}}{2.0\times10^{-6}}-0.5\right)=-10^\circ16'$$

and

$$\varphi_{-1}=\arcsin\left(-\frac{\lambda}{d}-\sin\theta\right)=\arcsin\left(-\frac{634.8\times10^{-9}}{2.0\times10^{-6}}-0.5\right)=-55^\circ17'$$

These tell us that the distribution of bright fringes is not symmetrical.

(2) The highest order $k_{\max}$ of maximum corresponds with the maximum of $\sin\varphi$, that is $\varphi=90^\circ$. From Eq. (15-36), we have

$$k_{\max}=\frac{a+b}{\lambda}(\sin 30^\circ+\sin 90^\circ)=\frac{2.0\times10^{-6}}{643.8\times10^{-9}}\times(0.5+1)=4.7$$

This tells us that the highest order of maximum which can be seen is the fourth-order fringe when oblique incidence occurs.

Assuming the highest order of the spectral line is $k'_{\max}$ in the case of normal incidence, according to the grating equation for normal incidence

$$(a+b)\sin\varphi=\pm k\lambda,\quad k=0,1,2,\cdots$$

we have

$$k'_{\max}=\frac{a+b}{\lambda}\sin 90^\circ=\frac{2.0\times10^{-6}}{643.8\times10^{-9}}\times1=3.1$$

which means that the highest order of maximum which can be seen is the third-order fringe when normal incidence occurs.

Therefore, we can observe higher order of bright fringes in the case of oblique incidence. The highest order of spectral line is only decided by the incident angle θ when the wavelength λ and the grating constant $(a+b)$ are unchanged.

15.12 Resolving Power of Optical Instruments

Diffraction comes from the wave nature of light and is determined by the finite aperture of the optical elements. An important property of any optical instrument, such as a camera or a telescope, is its resolving power. Resolving power is the ability to distinguish between two closely spaced objects. We have known that light passing a small opening or a boundary of a circular lens is bended so that the images are fuzzy. Diffraction fringes near the edges of images usually make it difficult to determine the exact shape of the source. It limits the resolving power of an optical instrument because the diffraction pattern has certain intensity distribution. Fig. 15-41 shows a circular diffraction pattern formed by a small circular aperture. Note that the large central maximum (bright spot), called Airy disk, is surrounded by alternating bright and dark rings. This diffraction is of extreme importance in an optical instrument because it sets the ultimate limit on the possible magnification.

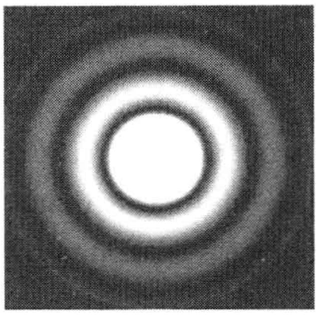
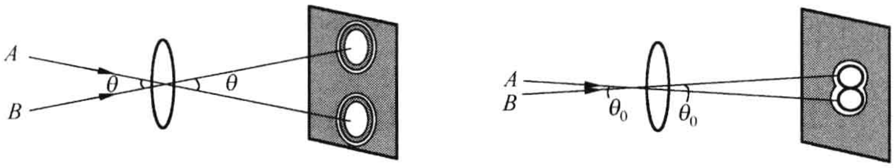


Fig. 15-41 Diffraction pattern of a circular aperture

Fig. 15-42 shows that light from two sources A and B pass through a small circular opening in an opaque barrier. In Fig. 15-42(a), the images of the sources A and B are distinguished as separate images. In this situation, the sources are said to be resolved. If they are brought closer together as shown in Fig. 15-42(b), however, their images overlap resulting in a confused image. When the sources are so close together (or the opening is so small) that the separate images can no longer be distinguished, the sources are said to be unresolved. The resolving power of an optical instrument is a measurement of its ability to produce well-defined separate images.



(a) (b)

Fig. 15-42 (a) The images of the sources A and B are easily distinguished; (b) As the sources are brought closer together, the images overlap, resulting in a confused image.

It can be proved that the first minimum of the diffraction pattern of a circular aperture with a diameter d (Fig. 15-43) is given by

$$\theta_1 = 1.22\frac{\lambda}{d} \tag{15-37}$$

where θ_1 is called the half-angle of Airy disk and has units of radians, λ is the wavelength of incidence. In Fig. 15-43, since $\theta_1 \approx \frac{\rho_1}{f}$, ρ_1 is the radius of Airy disk and f is the focal length of lens L, we have

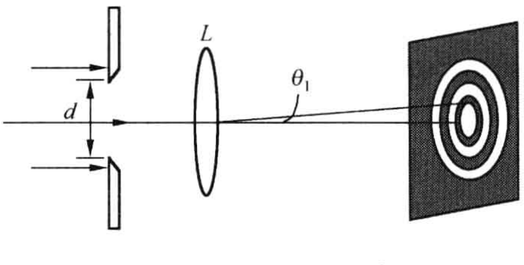


Fig. 15-43 Ariy disk

$$\rho_1 = 1.22\frac{\lambda}{d}f \tag{15-38}$$

From Eq. (15-37) and Eq. (15-38), we can see that the size of the central maximum is directly proportional to λ/d. That is, the central maximum is more spread out for longer wavelength and smaller aperture. No matter how perfectly a lens is constructed, the image of a point source of light will not be focused at a point. What is the condition of two images just resolved? The accepted criterion for resolution is the Rayleigh criterion, first proposed by Lord Rayleigh (1842—1919): two images are just resolved when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other. It tells us that two images are just resolved when the center of central maximum of one pattern coincides with the first dark fringe of the other. According to Rayleigh criterion and Eq. (15-37), two point objects separated by an angle θ_0 are just resolved when

$$\theta_0 = \theta_1 = 1.22\frac{\lambda}{d} \tag{15-39}$$

where θ_0 is called the minimum angle in radians, $1/\theta_0$ is called the resolving power of an optical instrument. It depends on the wavelength of the light and the diameter of the instrument's aperture. Fig. 15-44 shows two point objects when they are clearly resolved, just resolved and unresolved.

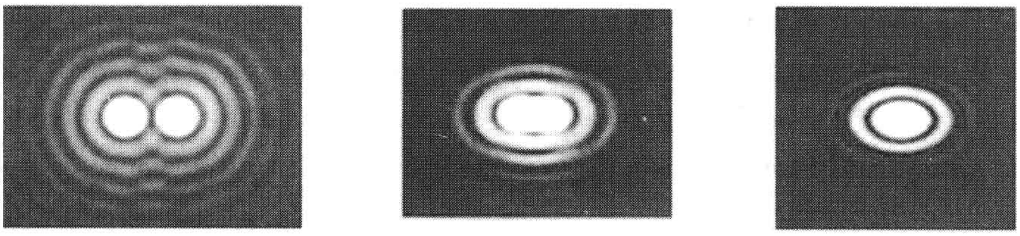


Fig. 15-44 Two point objects when they are clearly resolved, just resolved and unresolved

15.13 X-Ray Diffraction by Crystals

During the autumn of 1895, when studying the conduction of electricity, W. K. Roentgen discovered a mysterious radiation which was able to penetrate thin layers of material. He called this radiation as X-ray, with the "x" standing for "unknown". But it was soon established that X-ray is a type of electromagnetic radiation with wavelength in the range of 0.01 to 10 nanometers. Fig. 15-45 shows an X-ray tube — a device causes the emission of X-ray.

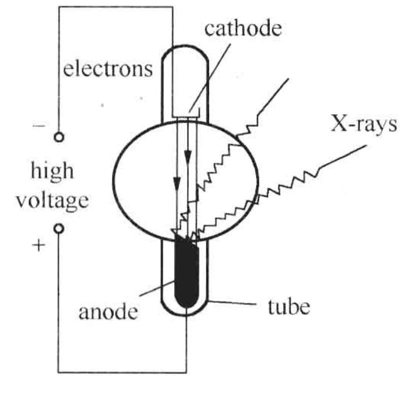


Fig. 15-45 X-ray tube

Diffraction is a natural property of waves. The diffraction phenomena of X-ray ought to be observed. But it is very difficult to observe X-ray diffraction because the wavelength of X-ray is too small for ordinary diffraction gratings. Not all diffraction gratings are commercially made. Crystalline solids supply natural gratings. The regular array of atoms in a crystal forms a natural 3-dimensional diffraction grating with a spacing that falls in the wavelength region of X-ray. The suggestion that crystal could be used to diffract X-ray was first made by German physicist M. Von Laue. In 1912, Laue carried out an experiment to prove the wave nature of X-ray. The experiment arrangement of X-ray diffraction is shown in Fig. 15-46. When a collimated beam of X-rays with a continuous distribution of wavelength strikes a single crystal, it is strongly scattered only in certain sharply defined directions. One certain pattern of spots occurs when the diffracted rays are fallen on a photographic film. These spots, called Laue Spots as shown in Fig. 15-47, are related in complicated ways to the internal structure of the crystal. The Laue Spots is the diffraction pattern of X-ray through a single crystal and it verified that X-ray is a kind of wave.

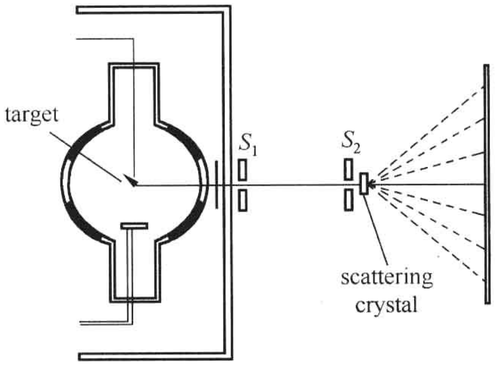


Fig. 15-46 Experiment arrangement for X-ray diffraction

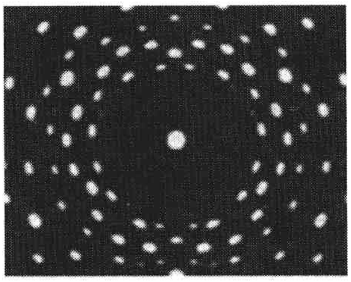


Fig. 15-47 Laue spots

After soon the experiment suggested by M. Von Laue, W. L. Bragg and his father W. H. Bragg suggested another method to study X-ray diffraction. The father and his son shared the 1915 Nobel Prize for their use of X-ray to study the structures of crystals. Fig. 15-48 shows a two dimensional representation of a three dimensional crystal and the rows of dots portray planes of atoms. The X-rays of a single wavelength are in phase before being scattered from the atoms in plane P_1 and in plane P_2. For constructive interference of the X-rays scattered from each plane of atoms, the glancing angle is equal to the emergence angle. Note that both angles are measured relative to the planes. The path difference is AOB in Fig. 15-48 since rays scattered from the atoms in plane P_2 travels a greater distance than those scattered from the atoms in plane P_1. Let the distance between adjacent planes be d. As $AO=OB=d\sin\theta$, we have

$$2d\sin\theta = k\lambda, \quad k = 1,2,3,\cdots \tag{15-40}$$

where k is the order number of an intensity maximum. The angle of glance and emergence in above equation is also called a Bragg angle. The waves scattered from the atoms in plane P_1 will arrive at the detector in phase with the waves scattered from the atoms in plane P_2. An interference maximum appears in the direction of emergence angle θ. Similarly, constructive interference will occur for rays scattered from the atoms of each of many planes that are parallel to planes P_1 and P_2. Eq. (15-40) was first developed by W. L. Bragg and is called Bragg's law. Each of the many families of planes has its own characteristic interplanar spacing d. Bragg's law gives the angles which locate the maxima produced by the constructive interference of X-ray.

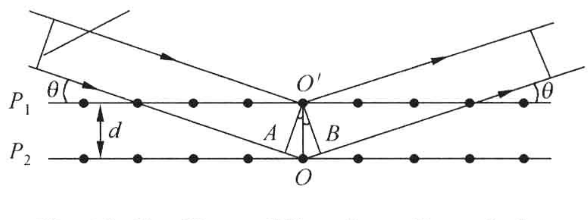


Fig. 15-48 X-ray diffraction—Bragg's low

X-ray diffraction is a powerful tool for studying both X-ray spectra and the arrangement of atoms in crystals. X-ray diffraction has also been applied to understand the structure of biological molecules such as proteins and nucleic acids. It is X-ray diffraction that reveals the molecular structure of deoxyribonucleic acid (DNA) is a double helix.

15.14 Polarization of Light and Malus's Law

The Interference and diffraction of light give the best evidence that light has wave-like characteristics. However, the phenomena of interference and diffraction can occur with any sort of waves, such as sound waves or waves on a stretched string. As noted in Chapter 13, the sound waves and waves on a stretched string are longitudinal and transverse wave disturbance, respectively. More evidence is needed to determine whether light waves are longitudinal or transverse. As we will see below, the polarization is a characteristic of transverse wave. The light can be polarized gives the best evidence that light is a transverse wave.

Light is one type of electromagnetic wave, as noted in Chap. 14. The electromagnetic waves are transverse waves consisting of varying electric and magnetic fields that oscillate perpendicular to the direction of propagation. And the directions for the oscillations of the electric and magnetic fields are always perpendicular to each other, as shown in Fig. 14-7. Polarization is an attribute that the oscillations of a wave have a definite direction relative to the direction of wave's propagation. Waves having such a direction are side to be polarized. For an electromagnetic wave, the polarizing direction is defined as the direction parallel to the oscillating direction of electric field that most detectors (including human's eyes) of electromagnetic radiation are sensitive.

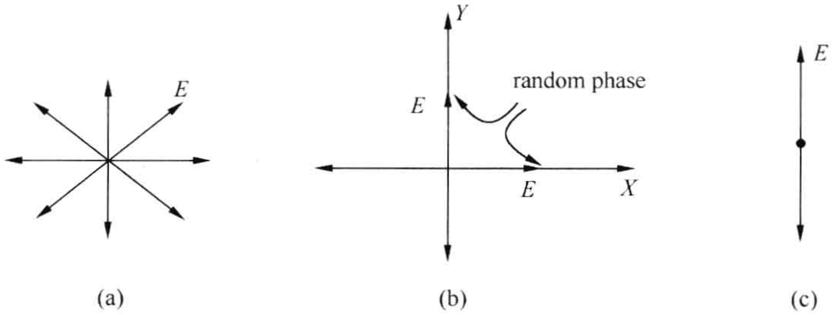


Fig. 15-49 (a) Random polarization; (b) unpolarized light is showed diagrammatically; (c) one polarized light (the beam moving out of the plane of the page)

The light emitted by ordinary sources or the sun light are randomly polarized as they are composed of many waves with all possible directions of polarization. Such light is called nature light or unpolarized light. Nature light is unpolarized because of its radiation mechanism. A large number of atoms and molecules that emit light oscillate independently so that the orientation of the electric fields produced by these emitters may not be correlated. Diagrammatically, we can represent unpolarized light by double arrow electric field vector **E** oriented randomly, as shown in Fig. 15-49(a). It is customary to resolve these vectors into two perpendicular components, but the phase difference between them is random, as shown in Fig. 15-49(b). Logically, the polarized light can be represented by a double arrow electric field vector ***E***, as shown in Fig. 15-49(c).

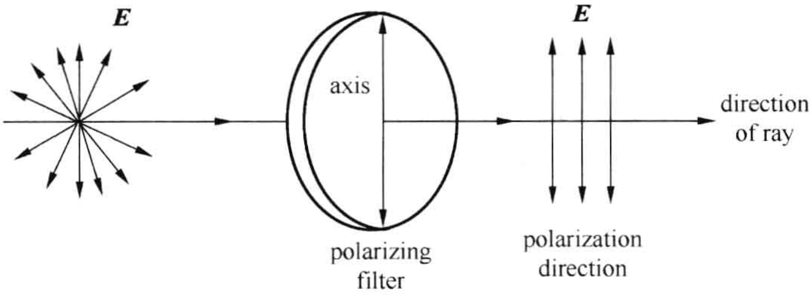


Fig. 15-50 Polarizing direction

A wave can be polarized by use of a polarizing filter which acts as a polarizing slit for light. A polarizing filter allows just only polarization in one to pass through. The axis of a polarizing filter (also called polarizing direction) is the direction along which the filter passes the vector ***E*** of light, as shown in Fig. 15-50. In fact, the polarizing direction of a light wave is defined to be the direction of its electric field.

Let us place two polarizing filters in the way of light emitted from a bulb, as shown in Fig. 15-51. The polarizing filters P_1 and P_2 are usually called a polarizer and an analyzer, respectively. The polarizing directions of P_1 and P_2 are represented by the parallel lines in filters. If we rotate P_2 about the direction of propagation of light, the intensity of light transmitted from P_2 varies with angle θ, as shown in Fig. 15-51(a). Two positions where the polarizing directions P_1 and P_2 are at right angle exist so that the transmitted light intensity is almost zero. This fact reveals the light emerging from P_1 is polarized. The light can be polarized gives the good evidence that light is a transverse wave. Furthermore, if we remove P_2 and rotate P_1, the transmitted light intensity keeps unvarying, which shows that light emitted from ordinary light sources such as bulb or the sun is unpolarized.

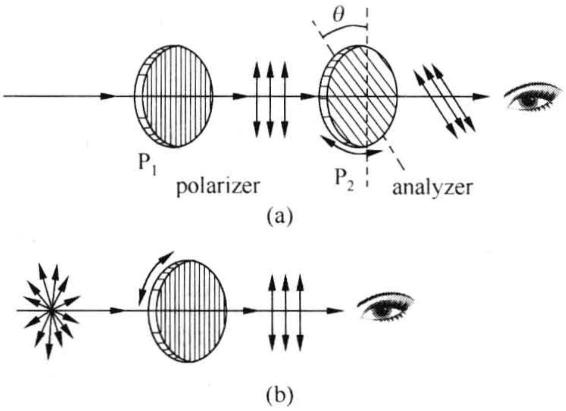


Fig. 15-51 (a) The light transmitted from P_1 is polarized; (b) the light emitted from a bulb is unpolarized

As shown in Fig. 15-51, the intensity of emerging light from P_2 varies with angle θ. What law does it follow? Etienne Louis Malus gave the answer in 1809.

If the amplitude of the polarized light falling on P_2 is E_m, then, the amplitude of the light that emerges is $E_m\cos\theta$, where θ is the angle between the polarizing directions of P_1 and P_2. Since the intensity of light is proportional to its amplitude squared, so we have

$$\frac{I}{I_0}=\frac{E_m^2\cos^2\theta}{E_m^2}=\cos^2\theta \tag{15-41}$$

in which I_0 is the intensity of polarized light that incidents on P_2, I is the intensity of the light that emerges from P_2.

Eq. (15-41) can be rearranged as:

$$I = I_0 \cos^2\theta \tag{15-42}$$

which is called Malus's Law. In Fig. 15-51(a), the maximum occurs at $\theta = 0°$ or $180°$. When P_2 or P_1 is rotated so that θ in Eq. (15-42) has value of 90° or 270°, the intensity of light that transmits through P_2 is a minimum, say, no light is passed.

Example 15-15 Unpolarized light falls on two polarizing filters placed one on top of the other. What must be the angle between the polarizing directions of two filters if the intensity of the transmitted light is one fourth of intensity of the incident beam? Assume that each polarizing filter is ideal, that is, it reduces the intensity of unpolarized light by exactly 50%.

Solution Suppose the angle between the polarizing directions of two filters is θ, I_0 is the intensity of incident unpolarized beam. The intensity of the light transmitted from the first filter is $I_1 = I_0/2$ as the filter is ideal. According to Malus's law, the intensity of the light transmitted from the second filter is

$$I_2 = I_1 \cos^2\theta = \frac{1}{2} I_0 \cos^2\theta$$

So we have

$$\frac{1}{4} I_0 = \frac{1}{2} I_0 \cos^2\theta, \quad \theta = \pm\frac{\pi}{4} \text{ or } \pm\frac{3\pi}{4}$$

15.15 Polarization by Reflection

If you rotate a polarizing filter in front of one-eye, you can reduce or eliminate the glare from sunlight reflected from water or from any glossy surface. Such reflected light is fully or partially polarized by the process of reflection from the surface.

Fig. 15-52 shows an unpolarized beam falling on a glass surface. The electric field vector for each wave train in the beam can be resolved into two components—a perpendicular component (perpendicular to the plane of paper), represented by dots in Fig. 15-52, and a parallel component (lying in the plane of paper), represented by straight short lines. For unpolarized incident light, these two components are of equal amplitude. It can be proved experimentally that at a particular angle of incidence, called the Brewster's angle θ_B, the reflection for the parallel component is zero, which means that the reflected light is completely polarized, as shown in Fig. 15-52. The refracted beams, however, are not completely polarized, but partially polarized.

At the Brewster's angle (sometimes be called polarizing angle), it can be shown experimentally that the refracted beams are at right angles with the reflected beams, or equivalently

$$\theta_B + \theta_\tau = 90° \tag{15-43}$$

From the Snell's law we have

$$n_1 \sin\theta_B = n_2 \sin\theta_\tau \tag{15-44}$$

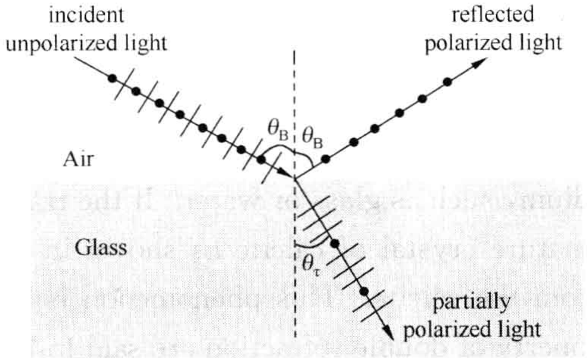


Fig. 15-52 At the Brewster angle, the reflected lights is completely polarized but of low intensity

Combining these equations leads to

$$n_1 \sin\theta_B = n_2 \sin(90^\circ - \theta_B) = n_2 \cos\theta_B$$

or

$$\tan\theta_B = \frac{n_2}{n_1} = n_{21} \tag{15-45}$$

where n_1 is the index of refraction of the medium in which the reflected light travels and n_2 is the index of refraction of the medium from which the light is refracted. Eq. (15-45) is known as Brewster's Law after Sir. David Brewster deduced it empirically in 1812.

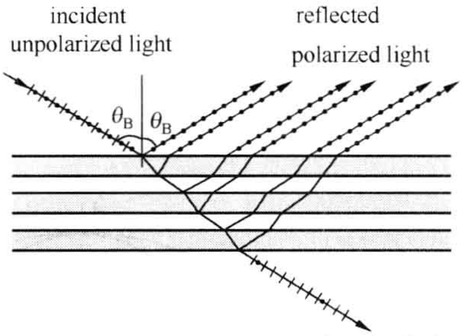


Fig. 15-53 Stacking glass used to incrcease the intensity of reflected polarized light. The transmitted beam is almost polarized

At the Brewster's angle, the reflected beam is of low intensity. Of the components perpendicular to the plane of incidence, about 15% are reflected if the reflecting surface is glass. The refracted beam, which is not completely polarized, is bright. The intensity of the reflected polarized beam can be increased by combining reflections as a result of stacking several plates, as shown in Fig. 15-53. The combined refracted beam becomes less intense but more completely polarized.

Example 15-16 One wish to use a plate of glass ($n=1.57$) as a polarizer. (1) What is the Brewster's angle? That is, at what angle of incidence will the reflected beam be fully polarized? (2) What angle of refraction corresponds to this angle of incidence?

Solution From Brewster's law, we have

$$\theta_B = \arctan n = \arctan 1.57 = 57.5^\circ$$

since

$$\theta_B + \theta_r = 90^\circ$$

therefore

$$\theta_r = 90^\circ - \theta_B = 32.5^\circ$$

15.16 Double Refraction

Corresponding to one beam of incident light, only one beam of light emerges through a homogeneous isotropic medium, such as glass or water. If the transparent medium is an anisotropic medium, such as a nature crystal of calcite as shown in Fig. 15-54, not one but two beams of light emerging from the calcite. This phenomenon is known as double refraction. The crystals having the property of double refraction are said to be double refraction crystals. Double refraction, also called birefringence, an optical property in which one beam of unpolarized light entering an anisotropic medium is split into two rays, each traveling in a different direction. One ray that follows Snell's law is called the ordinary ray (o-ray or o-beam), the other ray does not obey Snell's law is called the extraordinary ray (e-ray or e-beam). In double refraction, the ordinary ray and the extraordinary ray are polarized vibrating at right angles to each other.

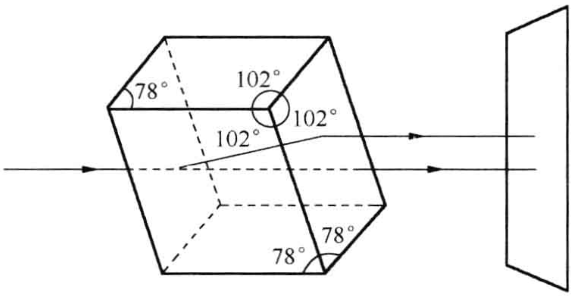
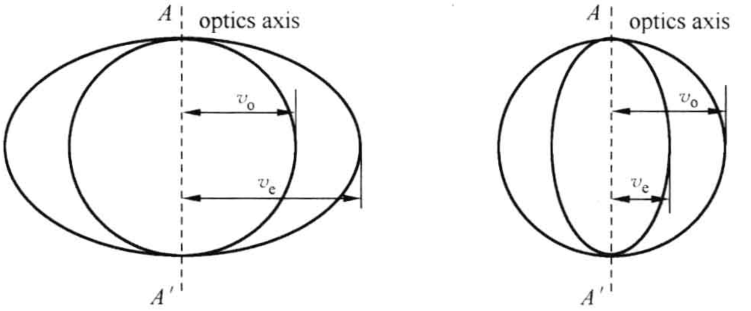


Fig. 15-54 One incident beam be split into two beams



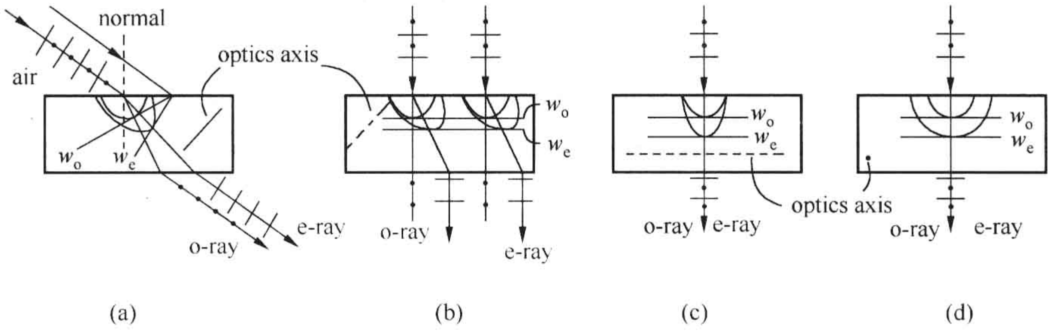
(a) Negative uniaxial crystal (b) Positive uniaxial crystal

Fig. 15-55 The wave fronts of o-ray and e-ray from a point source o in a uniaxial crystal

Using monochromatic sodium light of 589.3 nm and measuring the angles of incident and refraction, o-ray yields a constant index of refraction of 1.655. However, the e-ray shows an index of refraction that varies from 1.486 to 1.655 depending on the angle of incidence. This difference suggests that the e-ray propagates through the crystal at different speeds in different directions.

The double refraction phenomena can be interpreted By Huygens' principle. According to the Huygens' principle, since the speed of o-ray is the same in all directions, the secondary wavelets wave front of o-ray from a point source within a doubly refracting crystal is a spherical surface. The e-ray, however, has different speed in different direction, the wave front of e-ray's secondary wavelets is not a spherical surface, but an elliptical surface, as shown in Fig. 15-55. The two wave fronts of o-ray and e-ray are tangent to each other in direction (AA' in Fig. 15-55), which is called the optic axis of crystal. Thus the optic axis is such a direction in which the speed of o-ray and e-ray are equal. Fig. 15-55(a) represents a crystal in which the speed of o-ray is always smaller than that of e-ray in all directions is called negative uniaxial crystal, such as Calcite and Tourmaline. Fig. 15-55(b) represents for the positive uniaxial crystal, such as quartz and ice, in which the speed of the o-ray is always faster that of e-ray in all directions. In some crystals there are two different directions in which the speeds are equal. These crystals are called biaxial crystals. Almost all of the doubly refracting crystals used in optical instruments (chiefly quartz and calcite) are uniaxial, however, we will concentrate on uniaxial crystals.

Fig. 15-56(a), (b), (c), (d) are diagrams showing the double refraction of calcite. The w_o and w_e in the Fig. 15-56 is the wave front of o-ray and e-ray, respectively. The plane formed by normal and optic axis of crystal is called the principal section of crystal. In Fig. 15-56(a), (b), (c), the principal section is the plane of page, while the principal section of Fig. 15-56(d) is perpendicular to the page. When the incidence plane (the plane formed by incident light and normal) is parallel to the principal section, the vibration planes of o-ray and e-ray are perpendicular and parallel to the principal section, respectively. That means their planes of polarization to be perpendicular to each other. If the incidence plane is not parallel to the principal section, the vibration planes of o-ray and e-ray are not perpendicular to each other, but the angle between them nearly equal to 90°. Even light incidents normally on the crystal as shown in Fig. 15-56(b), the light will also be split into two beams. When the optic axis of crystal is parallel to the crystal's surface and the light incidents normally, the o-ray and e-ray pass through crystal in original direction of incident, not separated, but with different speeds, as shown in Fig. 15-56(c) and (d).



(a) (b) (c) (d)

Fig. 15-56 Diagrams showing the double refraction of negative uniaxial crystal: Incident at arbitrary angle in (a); normally incident with different optic axis directions in (b),(c) and (d)

The speed difference between o-ray and e-ray has a maximum in the direction at right angles to the optic axis; it is customary to define the index of extraordinary ray in the direction at right angles to the optic axis, i. e. the ratio of the speed of light in a vacuum to the speed of e-ray in the direction at right angles to the optic axis. Some values of n_o and n_e, the indices for ordinary ray and extraordinary ray in the direction at right angles to the optic axis are listed in Table 15-4.

Table 15-4 Indices of refraction of doubly refracting crystals
(for light wavelength 589.3 nm)

Material	n_o	n_e	
Calcite	1.655	1.486	$n_o > n_e$ negative uniaxial crystal
Tourmaline	1.640	1.620	
Quartz	1.544	1.553	$n_o < n_e$ positive uniaxial crystal
Ice	1.306	1.307	

Usually, the angle between o-ray and e-ray is not very large even with the highly doubly refracting crystal calcite, so that some extra technique is needed to separate them to obtain polarized light from nature light. A prism originally developed by Glan and Thompson and then modified by Ammann and Massey (1968) is such an example. The modified Glan-Thompson prism consists of a glass prism with index of refraction 1.655 and a calcite prism with index of refraction 1.486 that are cemented together by their long faces. The optical cement has the same index of refraction with glass. As shown in Fig. 15-57, where the thickness of cement is exaggerated, the natural unpolarized light incident from the left is resolved into two perpendicular components (represented by dots and double arrows). Both beams travel the same path with equal speed in the glass. The parallel component is extraordinary ray (e-ray) in calcite with an index of refraction 1.486. It will totally internal reflected when traveling from cement ($n = 1.655$) toward calcite. The vertical component is ordinary ray (o-ray) in calcite with an index of refraction 1.655 and proceeds without deflection from glass to cement and to calcite. It emerges into the air on the right side since its polarization plane is perpendicular to the principal section, e.g. the plane of the page. The Glan-Thompson prism can be used as a polarizer or an analyzer. When the incident light is polarized light with its vibration plane parallel to the principal section, no transmitted light can be viewed by observer.

Certain doubly refracting crystals exhibit dichroism, that is, one of the polarized components is absorbed much more strongly than the other. Hence, if the crystal is cut of a proper thickness, one of the components is practically extinguished by absorption, while the other is transmitted in appreciable amount, as indicated in Fig. 15-58. Tourmaline is one example of such a dichroic crystal. Dichroism is the basic operating principle of the commercial Polaroid sheet.

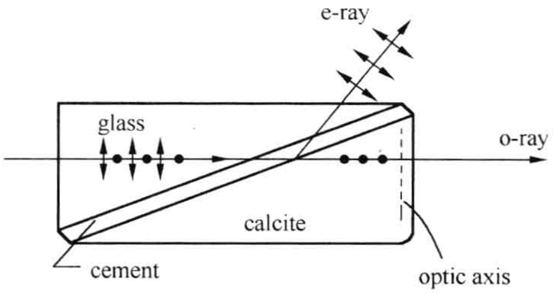


Fig. 15-57 Glan-Thompson polarizing prism, as modified by Ammann and Massey

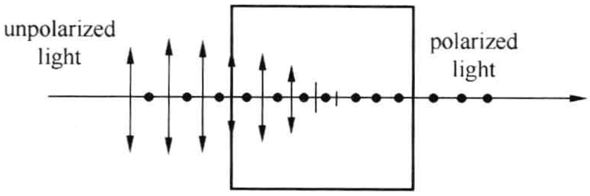


Fig. 15-58 Plane-polarized light transmitted by a dichroism crystal

15.17 Optical Stress Analysis

The two beams of light emerge from a double refracting crystal can not interfere with each other even though they could meet again. This is because their vibration directions nearly perpendicular to each other. As shown in Fig. 15-59, when a doubly refracting crystal is inserted between two crossed Polaroid sheets, the emerging beams from the analyzer are found to interfere. In Fig. 15-59, the o-vibration and e-vibration refer to ordinary ray and extraordinary ray, respectively. The polarized light falling on the crystal can be resolved into two components, o-vibration and e-vibration. o-vibration and e-vibration are perpendicular and parallel to the principal section, respectively. Only the components of o-ray passes through the analyzer and e-ray still lies in its polarizing direction, so the components of o-ray and e-ray emerge analyzer and lie in the same direction of X-axis in the figure. They will interfere with each other. When the two beams have an optic path difference of an odd multiple of one-half a wavelength, they interfere destructively; while the optic path difference is of even multiple of one-half a wavelength, they interfere constructively. The interference fringes will be viewed through analyzer. If the light applied is of many colors, the colorful interference fringes will appear.

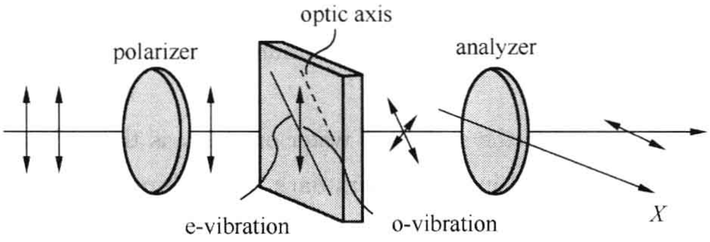


Fig. 15-59 A diagram showing the interference of polarized o-ray and e-ray

Some substances, such as glass, celluloid and various plastics, with no natural doubly refracting, may become of double refracting ones when subjected to mechanical stress. From a study of the interference pattern due to the specimen between crossed polarized filters, much information regarding the stresses can be obtained.

Improperly annealed glass, for example, may be internally stressed to an extent, which might cause it later to develop cracks. It is evidently important that optical glass should be free from such a condition before it is subjected to expensive grinding and polishing. Hence such glass is always examined by use of two crossed polarized filters before grinding operations begin.

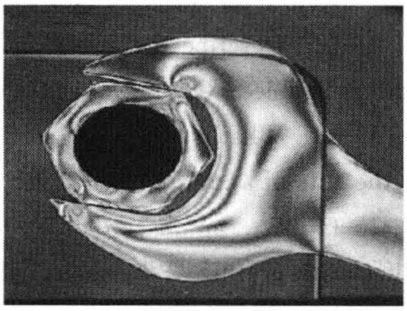


Fig. 15-60 A photo-elastic stress pattern

The double refraction produced by stress is the basis of the technique of photo-elasticity. The stresses in opaque engineering materials, such as grinders, boilerplates, gear teeth etc., can be analyzed by constructing a transparent model of the object, usually of a plastic. The transparent model of the object will be placed between a polarizer and an analyzer which are in the crossed position and the interference pattern of the model will be obtained. The stress distribution inside the object can be calculated by analyzing the interference pattern. Very complicated stress distributions, such as those around a hole or a gear tooth, which would be practically impossible to analyze mathematically, may thus be studied by optical methods. Usually, the denser the fringes are, the higher the stress is. Fig. 15-60 shows a model under light and heavy stress.

Questions

15-1 What type of experimental evidence indicates that light is a wave?

15-2 Light has wave characteristics in various media as well as in a vacuum. The wavelength of light is smaller in any medium than it is in a vacuum. Does the color change when light going from one medium to another? Why?

15-3 Why are interference effects not more commonly observed? How do you get coherent light rays? Explain your methods.

15-4 Why does sound bend around the corner of a building while light does not?

15-5 Suppose the Young's double-slit experiment is performed with a glass plate with index of refraction n covering the entrance to one of the slits. Describe the effect of this plate on the resulting pattern on the screen.

15-6 Suppose we perform the double-slit experiment under water. How would the pattern be affected?

15-7 Imagine observing the double-slit pattern for light of a given wavelength and gradually reducing the slit spacing d. What happens to the pattern? Is there a minimum spacing for observing a pattern? If so, what is this spacing?

15-8 When light rays in air encounter the surface of water or glass, is there phase reversed upon reflection?

15-9 A soap film on a wire loop held in air appears dark at its thinnest portion when viewed by reflected light. On the other hand, a thin oil film floating on water appears bright at its thinnest portion when similarly viewed from the air above. Explain this phenomenon.

15-10 Describe how the Newton's ring pattern would change if the lens radius was doubled; if the incident wavelength were doubled; if both the lens radius and wavelength were doubled.

15-11 If the Newton's ring arrangement was composed by three transparent materials (Fig. 15-61) whose index of refraction are not the same. What is the shape of Newton's ring? Is the point P bright or dark?

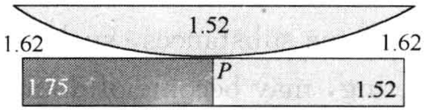


Fig. 15-61 For problem 15-11

15-12 Describe the effect on the Fraunhofer diffraction pattern produced by a single slit if (1) the slit width is doubled; (2) the wavelength is doubled; (3) both the slit width and wavelength are doubled.

15-13 What is the purpose of a diffraction grating? Why does a grating have a large number of slits? Why are the slits of a grating spaced very close together?

15-14 We can observe diffraction pattern when a monochromatic beam incident normally to a single-slit or a diffraction grating. Why the equations for bright fringes are different in above two arrangements?

15-15 Bragg's law, $2d\sin\theta=k\lambda$, is closely related to the grating equation $d\sin\theta=k\lambda$. Explain the origin of the factor 2 in Bragg's law.

15-16 Unpolarized light falls on two polarizing filters so oriented that no light is transmitted. If the third polarizing filter is placed between them, can light be transmitted?

15-17 Devise a way to identify the polarizing direction of a filter of Polaroid without the help of another polarizing filter.

15-18 Can polarization by reflection occur if the light is incident on the interface from the side with the larger index of refraction (from glass to air, for example)?

15-19 3D movies were very popular in the early 1950's. Viewers wear polarizing glasses and a polarizing filter was placed in front of each of the two projectors needed. Explain how the system worked.

15-20 A photographer usually fixes a polarizing filter in front of the lens of his camera when he takes a picture of clouds or objects reflecting light. Explain how the polarizing filter works and how the amount of light exposed should be changed.

15-21 If ice is doubly refraction, why can't we see two images of objects viewed through an ice cube?

Problems

15-1 Diamond has an index of refraction 2. 42. Find its critical angle of total internal reflection in air.

15-2 Assume that a certain concave spherical mirror has a focal length of 10. 0 cm. Determine the images for object distances of (1) 25. 0 cm, (2) 10. 0 cm, and (3) 5. 00 cm.

15-3 A converging thin lens of focal length of 10. 0 cm is used to form images of objects placed at (1) 30. 0 cm, (2) 10. 0 cm and (3) 5. 00 cm from the lens. Find the image distance in each case.

15-4 In Fig. 15-62, light with a wavelength of $\lambda = 600$ nm falls on a transparent material. The refractive index and thickness of the material is $n = 1.23$ and $d = 1.00$ cm, respectively. The incident angle $\theta = 30°$ and $SA = BC = 5.00$ cm.

(1) What is the angle θ_1?

(2) What are the frequency, speed and wavelength of light traveling in the transparent material?

(3) What is the geometric path from S to C? And what is the optical path from S to C?

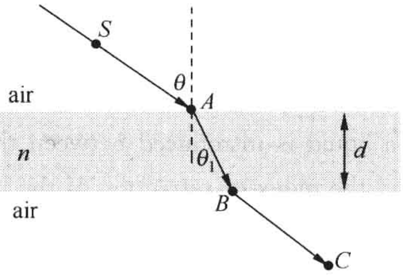


Fig. 15-62 For problem 15-4

15-5 In Young's interference experiment, the separation between the slits is 5. 00 mm and a screen is located 1. 00 m away from the slits. Two interference patterns can be seen on the screen, one due to light with wavelength 480 nm and the other due to light with wavelength 600 nm. What is the separation on the screen between the third-order interference fringes of the two different patterns?

15-6 In a double-slit arrangement, the slits are separated by a distance equal to 100 times the wavelength of the light passing through the slits.

(1) What is the angular separation between the central maximum and an adjacent maximum?

(2) What is the linear distance between these maxima if the screen is at a distance of 50. 0 cm from the slits?

15-7 A Lloyd's mirror arrangement is shown in Fig. 15-63, the screen is in contact with right edge of the mirror. The wavelength of incident beam is $\lambda = 720$ nm. Calculate the distance from right edge of the mirror to the first-order bright fringe on the screen.

15-8 A transparent flake of mica ($n = 1.58$) covers one slit of a double-slit arrangement. When the double slits are illuminated by monochromatic light of wavelength $\lambda = 550$ nm, the center of the screen appears the seventh-order bright fringe rather than the zero-order bright fringe. What is the thickness of the mica?

15-9 In Fig. 15-64, light of $\lambda = 500$ nm is incident normally on an air wedge-shaped film formed by two pieces of flat glass ($n = 1.50$). The 4th dark fringe is observed at position B. $AB = L = 1.56$ cm.

(1) Find $\theta = ?$

(2) If the wavelength of incident light is changed to $\lambda' = 600$ nm, does bright or dark fringe appear at position B?

(3) In the situation of (2), how many bright fringes and dark fringes are seen within region AB?

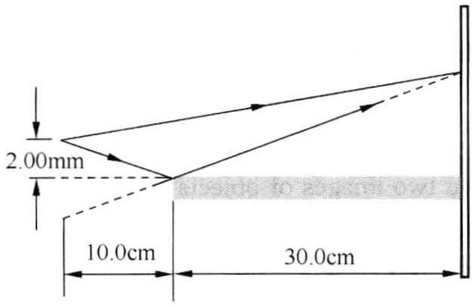


Fig. 15-63 For problem 15-7

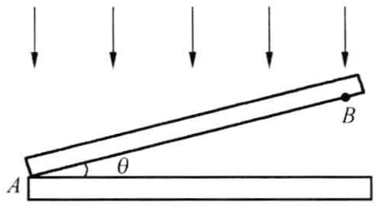
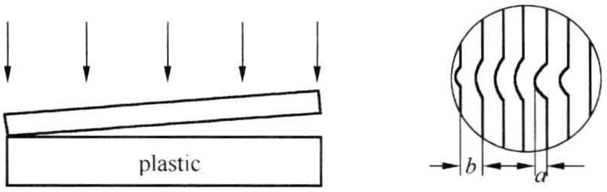


Fig. 15-64 For problem 15-9

15-10 An air wedge-shaped film is formed by a perfectly flat glass plate placed over a piece of black plastic as shown in Fig. 15-65(a). If the upper surface of the plastic has a little of roughness and the interference pattern was observed in the case of normal incidence as shown in Fig. 15-65(b). Is the veining on the upper surface of the plastic concave or convex? Prove that the depth H of the veining is $H = \frac{a}{b} \cdot \frac{\lambda}{2}$.



(a) (b)

Fig. 15-65 For problem 15-10

15-11 A thin film of oil floats in the air. The thickness and the refractive index of the oil film are $t = 500$ nm and $n = 1.46$, respectively. White light shines perpendicularly on the oil film. Find the interference maxima of the reflected beams for wavelength from 300 nm to 700 nm.

15-12 The diameter of the nth dark ring in a Newton's rings apparatus changes from 1.40 cm to 1.27 cm as a liquid is introduced between the lens and the plate. Find the index of refraction of the liquid.

15-13 A Newton's rings apparatus is used to determine the radius of curvature of a lens. When the apparatus is illuminated from above by light with wavelength $\lambda = 546$ nm, the radii of the nth and $(n+20)$th bright rings are measured and found to be 1.62 mm and 3.68 mm, respectively. Find the radius of curvature of the lower surface of the lens.

15-14 A vacuum chamber and an argon chamber with pressure 1.01×10^5 Pa are placed respectively in two arms of a Michelson interferometer. The lengths of two chambers with glass windows are 0.200 m. Light of $\lambda = 546$ nm is used. The argon is slowly evacuated from the chamber by a vacuum pump. While the argon is being removed, 205 fringes are observed to pass through the field of vision. Find the refractive index of argon at pressure 1.01×10^5 Pa.

15-15 White light passes through a slit and produces a diffraction pattern on the viewing screen.

(1) It is found that the first minimum for red light ($\lambda = 650$ nm) falls at $\varphi = 30°$. What is the width of the slit?

(2) What is the wavelength λ' of the light whose the 2nd maximum falls at $\varphi = 30°$, thus coinciding with the first minimum of red light ($\lambda = 650$ nm)?

15-16 A plane light wave with wavelength $\lambda = 700$ nm illuminates a narrow slit, and the diffracted light is focused on a screen by a lens with focal length $f = 0.700$ m. The width of the central maximum on the screen is $l_0 = 2.00$ mm.

(1) What is the width a of the slit?

(2) Assuming that the wavelength of incident light is changed to λ' from λ, then the width of the central maximum becomes $l'_0 = 1.50$ mm. Determine the wavelength λ'.

15-17 Visible light that contains all wavelengths between 400 nm and 760 nm is incident normally on a diffraction grating. The grating has 2408 lines per centimeter and produces a principal maximum at $\varphi = 30.0°$. What is (are) the wavelength(s) of the incident light that could have produced this maximum?

15-18 Light of wavelength 500 nm is incident obliquely on a diffraction grating. The position of the central maximum is occupied by the second maximum as incident angle changes from 0° to 30°.

(1) Find the grating constant.

(2) What is the largest order of the spectral line?

15-19 A parallel beam of wavelength $\lambda=500$ nm is incident obliquely on a diffraction grating whose grating constant is $d=2.10\ \mu\text{m}$ and the width of one slit is $a=0.700\ \mu\text{m}$. The angle of incidence is $\theta=30.0°$. How many bright fringes (diffraction principal maxima) can be observed?

15-20 Diffraction by pupil of eye limits the eye's acuity.

(1) What is the angle between two just resolving points of light for a 3.00 mm diameter pupil, assuming an average wavelength of 550 nm?

(2) What is the greatest possible distance a car can be from you if you can resolve its two headlights which are 1.30 m apart?

15-21 A parallel beam of X-rays is diffracted by a rock salt crystal. The first-order strong reflection is obtained when the incident angle is $83°10'$. The distance between reflection planes in the crystal is 28.1 nm. What is the wavelength of the X-rays?

15-22 What angle is needed between the direction of polarized light and the axis of a polarizing filter to reduce original intensity by 90%?

15-23 Unpolarized light falls on two polarizing filters placed one on the top of the other. What must be the angle between the polarizing directions of the filters if the intensity of the transmitted light is (1) one-third the maximum intensity of the transmitted beam or (2) one-third the intensity of the incident beam?

15-24 An unpolarized beam of light is incident on a group of four polarizing filters which are lined up so that the characteristic direction of each is rotated by 30° clockwise with respect to the preceding filter. What fraction of the incident intensity is transmitted?

15-25 A beam of light is a mixture of polarized light and randomly polarized light. When it is sent through a polaroid filter, it is found that the transmitted intensity can be varied by a factor of five depending on the orientation of the polaroid. Find the relative intensities of these two components of the incident beam.

15-26 (1) The refractive index of diamond is 2.42. Light reflected at 62.5° from a gemstone in a ring is completely polarized. Can this gemstone be a diamond? (2) At what angle would the light be completely polarized if this gemstone were in water with refractive index 1.33?

15-27 When a beam of parallel unpolarized light is incident on the surface of a plane glass at an angle 58°, the reflected light is completely polarized. Find

(1) the refraction angle of refraction light;

(2) the refractive index of the glass.

15-28 A beam of light strikes the upper surface of a flat of glass plate at polarizing angle (Brewster's angle). Prove that the reflected light from the lower surface of the glass is also completely polarized.

Part Five

The Modern Physics

Here "modern physics" implies the quantitative description and understanding of phenomena which could not be analyzed in terms of the "classical physics" we have been studying. The kinematics developed by Galileo and the mechanics developed by Newton, which form the basis of what we call classical physics, had many triumphs. Particularly noteworthy are the understanding of the motion of the planets and the use of kinetic theory of explaining certain observed properties of gases. By expressing the laws of electricity in terms of differential equations and introducing some new ideas, Maxwell showed that light is electromagnetic in origin and thus brought unity to optics and electricity. During the later part of the ninetieth century, and in the early years of twentieth century, a great number of new phenomena had been found in the world. If we try to use the classical physics to explain these phenomena, it will be proved that the classical physics is inadequate.

Among the topics which are typically regarded as foundations of modern physics are three major experimental discoveries and three revolutionary theoretical developments which challenged the validity of some seemingly well-established ideas of classical physics. The three experimental milestones were:

(1) The discovery of X-rays by Rontgen in 1895. Once this new type of radiation has been announced, hundreds of physicists began to work with X-rays. The origin, properties and uses of X-rays were explored and X-rays became a valuable tool in areas as diverse as medicine, crystallography and engineering.

(2) The discovery of radioactivity in 1896 by Becquerel. Work with natural radioactive sources led Rutherford to propose the nuclear model of the atom in 1911. By the early 1930s accelerators were available for producing nuclear reactions, and nuclear physics had become one of the most active branches of physics; during the 1940s nuclear energy emerged as an important new source.

(3) The discovery of the electron as one constituent of the atom. In 1897 J. J. Thomson measured e/m by deflecting a beam of electrons by electric and magnetic fields. By 1909 Millikan had developed a method of measuring e capable of good accuracy.

The three theoretical milestones were:

(1) The idea that electromagnetic radiation has particle characteristics as well as wave properties. In 1900 Planck proposed that harmonic oscillators radiate energy only in quanta of energy $h\nu$. In 1905 Einstein extended Planck's idea to explain the photoelectric effect. In 1913 Bohr combined the nuclear model, the electron and the particle nature of radiation to provide a model of the hydrogen atom which correctly predicted the spectrum of atomic hydrogen.

(2) The theory of relativity proposed by Einstein in 1905, which revolutionized our thinking about space and time and which introduced the mass-energy relationship.

(3) The idea that electron and other classical "particles" have wave properties. This hypothesis was advanced by De Broglie in 1924, and the wave equation satisfied by the electron was proposed by Schrödinger in 1926. Within a few years wave mechanics (quantum mechanics) was used to solve many previously intractable problems and to give some understanding of structures of atoms and of solids as well as an explanation of the periodic table of the elements.

In the short span of less than 40 year radically new ideas were established in physics.

In the following chapters, we will study only several main topics of the modern physics. The chapter 16 will discuss the relativity proposed by Einstein. In the chapter 17, we shall study the quantization of light. The following chapter is devoted to the quantum theory of atoms. The band theory of solids is introduced briefly in the last chapter.

Chapter 16

The Special Theory of Relativity

Maxwell's electromagnetic theory, made a great success in classical physics, led to new question about the propagation of light. Through what medium are the electromagnetic waves propagated? There was no obvious medium to which one could assign electromagnetic waves, since they were known to pass through a vacuum. However, most of physicists believed that some medium was necessary for energy to be transmitted through a distance, and this medium was given the name "***ether***". Does such as ***ether*** exist? If it exists, how can one detect it and determine its characteristics?

If *ether* exists, the speed of light measured in a reference frame at rest with respect to this ***ether***, according to the classical mechanics, should be different from the speed determined in a reference moving with respect to the ***ether***. Thus the ***ether*** should provide an absolute reference frame to which the motions of all other frames and of all bodies in the universe could be referred. Then, how fast does the earth move relative to the ***ether***?

Maxwell's theory told us that the speed of light in vacuum, given by

$$c=\frac{1}{\sqrt{\mu_0\varepsilon_0}} \tag{16-1}$$

is constant, independent of the reference frame. According to Galilean transformation of velocities, the speed of light in a reference frame is not equal to the speed in others. This means that Maxwell's theory is proper in a particular reference frame: ***ether***. If so, we must find the ***ether***. If we cannot detect the ***ether***, what is the next step?

In 1905, Einstein proposed a theory called the Special relativity. He abandoned the idea of the ***ether*** and insisted that there was not any absolute reference frame and all the inertial frames are equivalent with respect to all the physical laws. One principal focus of relativity has to do with measurements of events (things that happen): where and when they happen, and by how much any two events are separated in space and time. In addition, relativity has to do with transforming such measurements and others between reference frames that move relative to each other. This chapter introduces the Lorentz transformation, some of the consequences of the relativity and the famous mass-energy formula.

16.1 The Michelson-Morley Experiment

In previous chapter, we have described the Michelson's interferometer. This section is devoted to an application of Michelson's interferometer in which the effects of ***ether*** (if it exists) on the earth is measured. Historically, Einstein may have known only vaguely of the Michelson-Morley result. He was concerned instead with problems in the electrodynamics of moving bodies—the appearance of the fields in a light wave to an observer moving at speed c. However, the result of this experiment is helpful for one to abandon the idea of ***ether*** and to verify the relativity.

As shown in Fig. 16-1, a beam of light emitted by S is split into two parts by G. The light (1) goes from G to mirror M_1 and then to G and recombines with the light (2) which goes from G to mirror M_2 and then to G at T. Assume that the earth is moving with velocity $\boldsymbol{u}$ relative to the ***ether*** in the direction GM_2. That is the apparatus is moving in the ether.

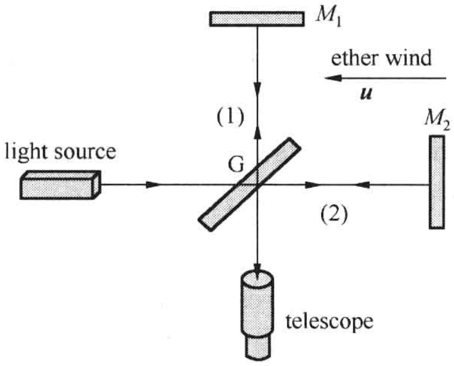


Fig. 16-1 Schematics of Michelson -Morley Experiment

Given $GM_1 = GM_2 = l$, we first consider the transit time of a light beam (1) moving perpendicular to the velocity of the apparatus relative to the ***ether***. The speed of light relative to the ***ether*** is c. According to Galilean transformation of velocity, the speed V of light (1) relative to the earth is given by

$$V = \sqrt{c^2 - u^2} \tag{16-2}$$

The time required for the light (1) to go from G to M_1 and then return to G is equal to

$$t_1 = \frac{2l}{\sqrt{c^2 - u^2}} = \frac{2l}{c}\left(1 - \frac{u^2}{c^2}\right)^{-\frac{1}{2}} \approx \frac{2l}{c}\left(1 + \frac{u^2}{2c^2}\right) \tag{16-3}$$

In the same way, we obtain the time required for the light (2) parallel to $\boldsymbol{u}$ to travel from G to M_2 and then return to G

$$t_2 = \frac{l}{c-u} + \frac{l}{c+u} = \frac{2cl}{c^2 - u^2} = \frac{2l}{c}\left(1 - \frac{u^2}{c^2}\right)^{-1} \approx \frac{2l}{c}\left(1 + \frac{u^2}{c^2}\right) \tag{16-4}$$

The difference Δt is

$$\Delta t = t_2 - t_1 = \frac{lu^2}{c^3} \tag{16-5}$$

Finally the difference of optical path lengths of two beams of light is

$$\delta = c\Delta t = \frac{lu^2}{c^2} \tag{16-6}$$

When the interferometer is rotated, interchanging the parallel and perpendicular paths, the total change of the difference of optical path lengths equals to 2δ. If the earth were moving through the ***ether***, the interference fringes should have shifted through about ΔN

$$\Delta N = \frac{2\delta}{\lambda} = \frac{2lu^2}{\lambda c^2} \tag{16-7}$$

fringes as the interferometer was rotated. Taken $c = 3\times 10^8$ m/s, $l = 11$ m (effective length in Michelson-Morley experiment), $\lambda = 5.9 \times 10^{-7}$ and $u = 3 \times 10^4$ m/s (the velocity with which earth rotates about the Sun), ΔN has a value 0.4 which can be observed by Michelson interferometer.

The instrument used by Michelson and Morley was capable of detecting a shift in the fringe pattern as small as 0. 01 fringes. However, no shift was detected in the experiments. Michelson and Morley performed this experiment repeatedly over many months. No effect was observed which could be interpreted as showing a motion of earth through the ***ether*** , in spite of the fact that they made measurements throughout the year, when the earth was moving in many different directions relative to the solar system. The experiment has been repeated many times by different scientists under a wide variety of conditions and no fringe shift has ever been detected. There had been various proposed explanations to interpret the results of the experiments and to save the ether frame concept and the Galilean addition law for the velocity of light. Lorentz and Fitzgerad suggested that the arm of the interferometer moving in the direction of the earth's motion shrinks just enough to lead to the null result. All proposals resulting from these efforts have been shown to be wrong. Therefore, this experiment shows that "*ether* "does not exist and there is not any motion of earth relative to the ***ether*** . As we shall see, the results of this experiment are supported by Einstein's theory of relativity.

16. 2 The Postulates of Special Relativity

Einstein was keenly aware of the successes of classical mechanics, which predicted with great precision the motions of planets, aircraft, and baseballs. He proposed his special theory of relativity in 1905, not out of any attempt to explain the outcome of the Michelson-Morley experiment but based a thought experiment that he had devised. As a 16-year-old student, Einstein had learned Maxwell's theory of electromagnetism and had thought about a paradox: if you were to move at the speed of light parallel to light beam traveling in empty space, you would observe "*static* " electric and magnetic field patterns. However, Einstein knew that such static electric and magnetic field patterns in empty space violated Maxwell's theory.

Einstein was faced with two choices to resolve this paradox: either Maxwell's theory was wrong or also the classical kinematics that permits an observer to travel along with a light beam was wrong. With the intuition that was perhaps his greatest attribute, Einstein put his faith in Maxwell's theory and sought an alternative to the kinematics of Galileo and Newton.

Einstein's special theory rests on two postulates:

> The principle of the relativity: the laws of physics are the same in all inertial reference frames.
>
> The principle of the constancy of the speed of light: the speed of light c is the same for every inertial reference frame and is independent of any motion of the source or observer.

Galileo assumed that the laws of mechanics were the same in all inertial reference frames. Einstein extended that idea to include all the laws of physics, including especially electromagnetism and optics. The first postulate declares that the laws of physics are absolute, universal, and the same for all inertial observers. Laws that hold for one inertial observer cannot be violated for any inertial observer. There exists no preferred frame in the universe; there is no *ether* relative to which everything can be measured. One inertial reference frame is equivalent to another. The first postulate does not say that the measured values of all physical quantities are the same for all inertial observers; most are not the same. It is the laws of physics, which relates these measurements to each other, are same.

The second postulate means that the speed of light in a vacuum has the same value, regardless of the velocity of the observer or the velocity of the source emitting the light. This postulate is much more difficult to accept, because it violates our "*common sense* ", which is firmly grounded in the Galilean kinematics that we have learned from everyday experience. The first postulate implies that Maxwell's theory is the same in all the inertial reference frames and then the speed of light in the empty space (as shown in Eq. (16-1)) is the same in all the inertial reference frames. However, the second postulate is not a special case of the first postulate. The second postulate also implies there is in nature an ultimate speed c (as we shall see in the section 16-5). The ultimate speed c sets a limit to which any material particles such as an electron can be accelerated.

According to Newton's mechanics, the kinetic energy of a particle can be increased toward very large values so that the speed of the particle exceeds the speed c. The above two postulates imply that it is impossible to accelerate a particle to a speed greater than c, no matter how much kinetic energy we give it.

According to the special theory, the speed of light is same in all the inertial reference frames. Therefore, the times required by the two beams of light for the Michelson-Morley experiment are equal and Δt is zero, which implies there are no shifts in the experiments. The results of Michelson-Morley experiments also confirm the postulates of special relativity.

16. 3 The Lorentz Transformation

16. 3. 1 The space-time coordinates of an event

An event is just something that happens, and anyone in any reference frame may measure it and assign space-time coordinates (x, y, z, t) to it. For example, the turning on or off of a tiny light bulb are two events; the motion of a body in the space can be looked as many successive events. The events are represented by the symbols P_1, P_2 or A, B with the space-time coordinates (x, y, z, t) in physics.

A given event may be recorded by any number of observers, each in their own inertial reference frame. In general, all such observers will assign different space-time coordinates to the same event. Note that an event does not, in any sense, "*belong* " to a particular inertial reference frame. An event is just something that happens and anyone may look at it and assign space-time coordinates to it. Therefore, we require a set of relationships called transformation equations that relate observations of a single event by two different observers.

The transformation equations have three ingredients:

(1) an observer S at rest in one inertial frame (S-frame);

(2) another observer S' at rest in a different inertial frame (S'-frame);

(3) a single event that is observed by both S and S'.

When the event P occurs, the time-space coordinates are recorded by S and S', which are $P(x,y,z,t)$ and $P(x',y',z',t')$ respectively. Knowing the relative velocity u of S and S', we wish to find the relation between (x,y,z,t) and (x',y',z',t'). We simplify this problem somewhat, without losing generality, by always choosing the x and x' axes to be along the direction of u, as shown in Fig. 16-2.

In chapter two, we can calculate (x',y',z',t') from (x,y,z,t) by Galilean transformation Eq. (2-29)

$$\begin{cases} x' = x - ut \\ y' = y \\ z' = z \\ t' = t \end{cases} \tag{16-8}$$

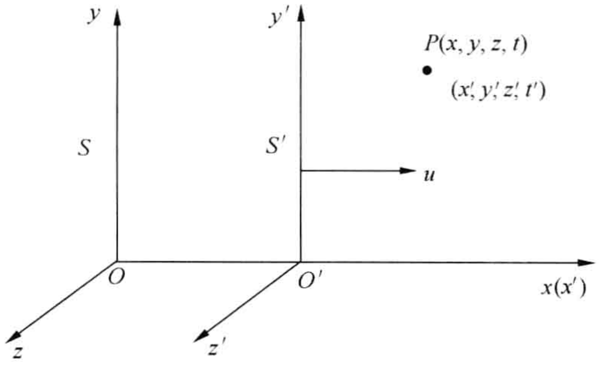


Fig. 16-2 S' moves with speed u relative to S along the common $x(x')$ direction

Here, when $t = t' = 0$, let the origins of two inertial frames coincide. If a body has a velocity v in the frame S with components (v_x, v_y, v_z), the corresponding components in the frame S' are given by Eq. (2-30)

$$\begin{cases} v'_x = v_x - u \\ v'_y = v_y \\ v'_z = v_z \end{cases} \tag{16-9}$$

Galilean transformations predict that if a beam of light has a velocity c in the x direction of the rest frame, its velocity in the moving frame is $c \pm u$. This is not consistent with the Einstein's postulates. Hence we need a new set of relationships which reduce to Galilean transformation for low velocities but permit us to satisfy Einstein's postulates for very high velocities.

16. 3. 2 The Lorenz transformation

It was Einstein's good fortune that such a transformation had already been derived by Lortenz. It is (see appendix 1)

$$\begin{cases} x' = \dfrac{x - ut}{\sqrt{1 - \dfrac{u^2}{c^2}}} \\ y' = y \\ z' = z \\ t' = \dfrac{t - \dfrac{u}{c^2}x}{\sqrt{1 - \dfrac{u^2}{c^2}}} \end{cases} \tag{16-10}$$

which is called Lortenz transformation. Eq. (16-10) can be derived directly from Einstein's postulates as Einstein did, if we invoke certain reasonable assumptions about the symmetry and the homogeneity of space and time. Note Eq. (16-10) is written with the assumption that $t = t' = 0$ when the origins of S and S' coincide.

The inverse transformation of Eq. (16-10) can be obtained by switching primed and unprimed coordinates in Eq. (16-10) and substituting $-u$ for u:

$$\begin{cases} x = \dfrac{x' + ut'}{\sqrt{1 - \dfrac{u^2}{c^2}}} \\ y = y' \\ z = z' \\ t = \dfrac{t' + \dfrac{u}{c^2}x'}{\sqrt{1 - \dfrac{u^2}{c^2}}} \end{cases} \tag{16-11}$$

Introducing the Lortenz factor γ and the speed parameter β:

$$\beta = \frac{u}{c} \tag{16-12}$$

$$\gamma = \frac{1}{\sqrt{1 - \dfrac{u^2}{c^2}}} = \frac{1}{\sqrt{1 - \beta^2}} \tag{16-13}$$

We can rewrite Eq. (16-10) and Eq. (16-11) as

$$\begin{cases} x' = \gamma(x - ut) \\ y' = y \\ z' = z \\ t' = \gamma\left(t - \dfrac{\beta}{c}x\right) \end{cases} \tag{16-14}$$

and

$$\begin{cases} x = \gamma(x' + ut') \\ y = y' \\ z = z' \\ t = \gamma\left(t' + \dfrac{\beta}{c}x'\right) \end{cases} \tag{16-15}$$

It can be easily proved that when $u \ll c$ or $\beta \to 0$, the Lorentz transformation reduces to Galilean transformation.

The relationship between β and γ is shown in Fig. 16-3, which implies that only at the high speeds must we take relativistic effects into account. When $u \ll c$ ($\beta \to 0$), the relativistic effects are very small and may be negligible.

Two events P_1 and P_2 occur respectively at the time-space coordinates (x_1, y_1, z_1, t_1) and (x_2, y_2, z_2, t_2) according to the inertial frame S. Corresponding their time-space coordinates observed by S' in the moving inertial frame S' are (x'_1, y'_1, z'_1, t'_1) and (x'_2, y'_2, z'_2, t'_2). From Eq. (16-10), we have

$$\begin{cases}\Delta x' = \gamma(\Delta x - u\Delta t) \\ \Delta y' = \Delta y \\ \Delta z' = \Delta z \\ \Delta t' = \gamma(\Delta t - \frac{\beta}{c}\Delta x)\end{cases} \tag{16-16}$$

and from Eq. (16-11), the inverse Eq. (16-16) are obtained

$$\begin{cases}\Delta x = \gamma(\Delta x' + u\Delta t') \\ \Delta y = \Delta y' \\ \Delta z = \Delta z' \\ \Delta t = \gamma(\Delta t' + \frac{\beta}{c}\Delta x')\end{cases} \tag{16-17}$$

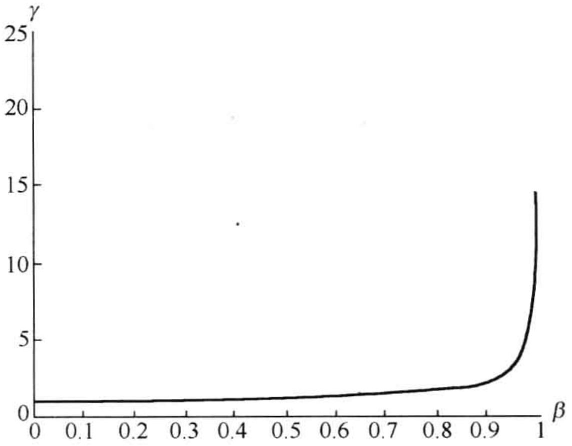


Fig. 16-3 The Lorentz factor γ as a function of β

where $(\Delta x, \Delta y, \Delta z, \Delta t)$ and $(\Delta x', \Delta y', \Delta z', \Delta t')$ are the time-space intervals between events measured by both S and S', respectively.

Example 16-1 In inertial frame S, a red light and a blue light are separated by a distance $\Delta x = 2.45$ km, with the red light at the larger value of x. The blue light flashes, and 5.35 μs later the red light flashes. Frame S' is moving in the direction of increasing x with a speed of $u = 0.855c$. What is the distance between the two flashes and the time between them as measured in S' ?

Solution The Lorentz parameter is

$$\gamma = \frac{1}{\sqrt{1-\frac{u^2}{c^2}}} = \frac{1}{\sqrt{1-(0.855)^2}} = 1.93$$

We are given the intervals in S as $\Delta x = 2.45$ km and $\Delta t = 5.35 \times 10^{-6}$ s. From Eq. (16-16), we have the interval transformations

$$\Delta x' = \gamma(\Delta x - u\Delta t) \approx 2.08 \text{ km}$$

and

$$\Delta t' = \gamma\left(\Delta t - \frac{u}{c^2}\Delta x\right) \approx -3.15\ \mu\text{s}$$

In S', the red flash is also located at the more distance coordinates, but the distance is 2.08 km rather than 2.45 km. Also, in S' the red flash comes before the blue flash (in contrast to what is observed in S); the time between flashes is 3.15 μs according to S'.

16.4 Some Consequences of the Lorentz Transformation

According to Galilean transformation, the time-space intervals between two events are absolute with respect to all the inertial reference frames (observers). If two events take place at the same time in the inertial reference frame, they also occur at the same time in all other inertial reference frames. That is, the simultaneity is absolute in classical physics. However, the results of Example 16-1 show that the time-space interval between two events is related to the motion of the observer. In this section, we will discuss in detail the time-space intervals and the simultaneity in the special theory of relativity.

16.4.1 The relativity of simultaneity

As shown in Fig. 16-4, let us imagine a high-velocity train moving along the direction x-axis of S (earth) with the velocity u. There is a blue lamp in the point midway of the train. The beams of light, emitted by the lamp, will arrive at the ends A and B of train at the same time, according to S' (the train). The arrival of light at the end A is looked as event P_A and the arrival at the end B as event P_B. The time-space coordinates of two events P_A and P_B, measured by S', are (x'_A, t') and (x'_B, t') respectively. The time interval between two events in the moving inertial frame (the train) is zero. That is $\Delta t' = 0$ and the two events occur at the same time measured by the observer in S'. Now consider the situation from the point of view of S (the observer at rest on the earth). Clearly the light signal reaches A before the light signal reaches B, because the speed of light is constant. Hence, two events P_A and P_B does not take place simultaneously with respect to S.

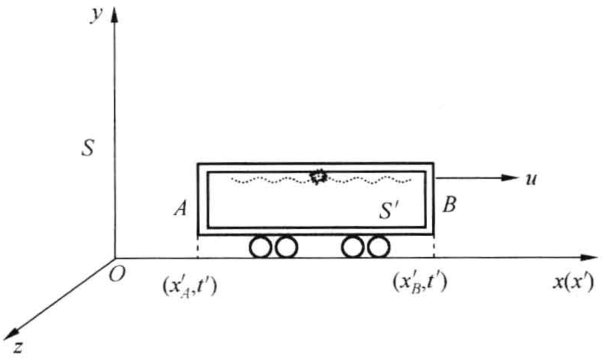


Fig. 16-4 A high-velocity train moves along the x-direction of S. According to S' (train), the light, emitted from the point midway of train, arrive at the ends A and B of the train at the same time

From Eq. (16-17), the time interval between two events measured by S is given by

$$\Delta t = \gamma \frac{u}{c^2}(x'_B - x'_A) = \gamma \frac{u}{c^2} L \tag{16-18}$$

where L is the length of the train. Because $\Delta x' = x'_B - x'_A = L \neq 0, \Delta t \neq 0$. We therefore reach the following conclusion:

> If two observers are in relative motion, in general they do not agree on whether two events at different location are simultaneous. If one observer finds the two events to be simultaneous, the other does not.

Note that this occurs only when the two events occur at different locations according to S'. If the two events take place at the same location and are simultaneous according to S', they are simultaneous in all other inertial frames as well.

We cannot say that one observer is right and other wrong. The situation is completely symmetrical and there is no reason to choose one observer over the others because there is no preferred inertial reference frame and any inertial reference frame can be used to describe events and do physics. Although the two observers reach different conclusions, both are correct in their own reference frames. Hence we have:

> Simultaneity is not an absolute concept but a relative one, depending on the state of motion of the observer.

Of course, if the relative speed of the observers is very much less than the speed of light, the measured departures from simultaneity become so small that they are not noticeable. Such is the case for all our experience of daily living; this is why the relativity of simultaneity is unfamiliar.

16.4.2 The relativity of time (The time dilation effect)

In Fig. 16-5, clock is at rest in the frame of S', who moves at speed u relative to S. S' measures the time interval $\Delta t' = t'_B - t'_A$ in which the hand of the clock moves between two marks, passing the first mark at time t'_A (looked as event P_A) and the second at time t'_B (looked as event P_B).

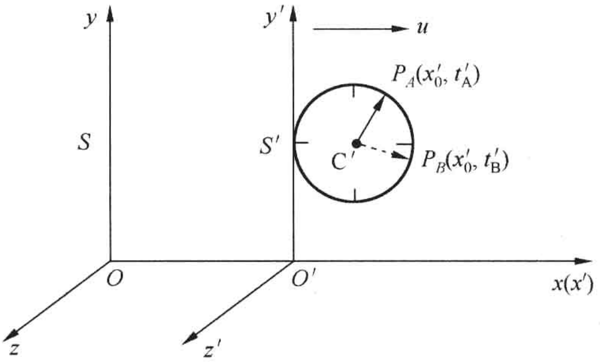


Fig. 16-5 Clock C′ is at rest in reference S'. With respect to reference S, clock C′ is in motion at velocity **u**

The hand of clock C′ passing the two marks is two events which occur at the same location x'_0 according to S' (because clock C′ is at rest in the frame S'). However, S observes the hand of clock C′ to pass the first mark (event P_A) at the location x_1 at time t_1 and to pass the second mark (event P_B) at the location x_2 at time t_2. The time interval Δt between two events P_A and P_B measured by S is $t_2 - t_1$. We can find the relation between the time intervals Δt and $\Delta t'$ directly from Eq. (16-17)

$$\Delta t = \frac{\Delta t'}{\sqrt{1-\frac{u^2}{c^2}}} \tag{16-19}$$

Because S' is at rest relative to clock C′, the time interval $\Delta t'$ measured by S' is a proper time interval, which we represent as Δt_0. Hence, we have

$$\Delta t = \frac{\Delta t_0}{\sqrt{1-\frac{u^2}{c^2}}} \tag{16-20}$$

Obviously Δt is longer than Δt_0. Thus, when an observer (S) is moving relative to a clock (or the clock relative to the observer), the time interval as seen by this observer is longer than that seen by another observer (S') at rest relative to the clock. This effect is called time dilation. Δt_0 is called the proper time, the time between two events as measured by an observer who sees two events occur at the same position. That is, the proper time interval between events is the smallest interval between them that any observer can measure; all observers in motion relative to the clock C′ measure longer intervals. That is the time interval Δt between two events measured by an observer moving with respect to a clock is longer than the time interval Δt_0 between the same two events measured by an observer at rest with respect to the clock.

As shown in Fig. 16-5, Δt_0 is the time interval in which the hand of the clock moves between two marks, and since $\Delta t > \Delta t_0$, observer S concludes that the clock C ′ is running slow. So, time dilation is often summarized by the phrase, "*moving clocks run slow* ". The time dilation effect is completely symmetric. If a clock C at rest in the inertial reference frame S is observed by S', then S' concludes the clock C is running slow. Each observer believes that the other's clock (moving relative to him) is running slower than the ones at rest in the reference frame of the observer. Note that the factor in the denominator differs appreciably from 1 only at speeds that approach the speed of light. That is, when the speed u approaches the speed of light, the time dilation is pronounced.

This time dilation effect is very real and has nothing to do with any mechanical change that takes place in a clock because of its motion. It is simply the nature of time. The time dilation effect is verified by the experiments of high energy physics. An interesting example of time dilation involves the observation of muons, unstable elementary particles that have a charge equal to that of the electron and a mass 207 times that of the electron.

Example 16-2 Muons are unstable, and when at rest in the laboratory, decay with an average lifetime of 2.20 μs. This average lifetime, measured for resting muons with a single resting laboratory clock, is thus a proper time interval. Find the average lifetime of muons which is accelerated to a speed of 0.996c.

Solution The accelerated muons can serve as tiny moving clock, to which the average lifetime of the muons is 2.20 μs. From Eq. (16-20), the average lifetime by the observer at rest in the laboratory is given by

$$\Delta t = \frac{\Delta t_0}{\sqrt{1-\frac{u^2}{c^2}}} = \frac{2.20}{\sqrt{1-(0.996)^2}} = 26.7(\mu s)$$

This value is in excellent agreement with the experimental data within the experimental error.

16.4.3 The relativity of length

Suppose that a rod lies along the $O'x'$, moving with a velocity u relative to S, at rest in the moving reference frame S', as shown in Fig. 16-6. The ends of a measuring rod are determined to be at x'_1 and x'_2 according to S', relative to whom the rod is rest. To determine the length of the rod, S must make a simultaneous measurement of the coordinates x_1 and x_2 of its endpoints. According to S', the ends of the rod are at coordinates x'_2 and x'_1, such that $\Delta x' = x'_2 - x'_1 = L_0$ is called as the rest length of the rod (also known as the proper length).

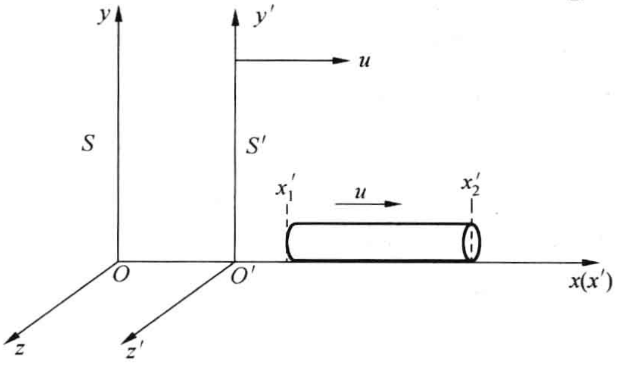


Fig. 16-6 A rod moves with speed u relative to S along the common $x(x')$ direction

In order to measure the length of the rod, observer S must make a simultaneous determination of the coordinates x_1 and x_2 of the ends of the rod. The interval $\Delta x = x_2 - x_1$ gives the length L of the rod according to S. From Eq. (16-16), we have

$$\Delta x' = \gamma(\Delta x - u\Delta t)$$

For $\Delta t = 0$, we obtain

$$L = \Delta x = \frac{\Delta x'}{\gamma} = L_0\sqrt{1-\frac{u^2}{c^2}} \qquad (16\text{-}21)$$

Thus, a rod moving relative to an observer appears to be contracted in the direction of motion by the factor $\sqrt{1-u^2/c^2}$. This effect is known as length contraction.

From Eq. (16-21), we conclude that the length L_0 measured by an observer who is at rest relative to the rod is the largest and all other inertial observers will measure a shorter length. That is the length of a rod is not absolute. Note that the contraction of rod happens only for dimensions along the direction of motion and length measurements transverse to the direction of motion are unaffected.

Under ordinary circumstances $u \ll c$, the effects of length contraction are far too small to be observed. For example, a rocket of length 100 m launched from Earth with the high speed sufficient to escape the Earth's gravity ($u = 11.2$ km/s) would be measured to contract, according to an observer on the earth, by an amount roughly equivalent to only 2 atomic diameters!

Like time dilation, length contraction is an effect that holds for all observers in relative motion. If a rod is at rest along Ox in the frame S, moving with a speed u relative to the frame S', the length measured by the S' is also contracted. The questions, "Does the rod really shrink ?" and "Do the atoms in the rod really get pushed closer together ?" are not proper questions within the framework of relativity. The length of a rod is what you measure it to be and motion affects measurements.

Example 16-3 A rod lies parallel to the x-axis of reference frame S, moving along this axis at a speed of $0.632c$. Its rest length is 1.70 m. What will be its measured length in frame S?

Solution The rod is moving along the x-axis of reference frame S relative to the frame S at a speed of $0.632c$. According to the S, its length is contracted. From the Eq. (16-21), we have

$$L = L_0\sqrt{1-\frac{u^2}{c^2}} = 1.70 \times \sqrt{1-0.632^2} = 1.32\ (\mathrm{m})$$

which is the value of the length of rod measured by S.

Example16-4 The length of a meterstick moving parallel to its length with a speed of u relative to an observer is measured to be 0.866 m by this observer. Find the relative speed u.

Solution The rest length of a meterstick is 1.000 m. By Eq. (16-21), we have

$$0.866 = 1.000 \times \sqrt{1-\frac{u^2}{c^2}}$$

From above formula, we obtain $u = 0.5c$.

16.5 The Lorentz Transformation of Velocities

In this section we use the equation of the Lorentz transformation to relate the velocity $\boldsymbol{v}$ of a particle measured by an observer in the frame S to the velocity $\boldsymbol{v}'$ of the same particle measured by an observer in the frame S', who is in turn moving with velocity u relative to S along the common $x(x')$ axis. Here, it is important to keep in mind the meanings of these three velocities.

Seeing Fig. 16-7, suppose observer S finds the particle to move from time-space point (x_1,

y_1, z_1, t_1) to (x_2, y_2, z_2, t_2). Observer S', on the other hand records the observations of the initial and final coordinates of the same particle as (x'_1, y'_1, z'_1, t'_1) and (x'_2, y'_2, z'_2, t'_2).

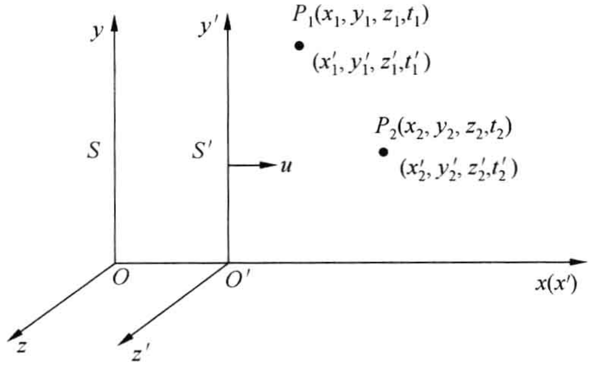


Fig. 16-7 A particle is moving in the space. The velocity of the particle is measured by both S and S'

v_x, the x component of the velocity of the particle measured by S, and v'_x, the x' component of the velocity of the same particle measured by S', are given by

$$v_x = \operatorname*{limit}_{\Delta t \to 0} \frac{\Delta x}{\Delta t} \tag{16-22a}$$

$$v'_x = \operatorname*{limit}_{\Delta t' \to 0} \frac{\Delta x'}{\Delta t'} \tag{16-22b}$$

where $\Delta x = x_2 - x_1$, $\Delta t = t_2 - t_1$ and $\Delta x' = x'_2 - x'_1$, $\Delta t' = t'_2 - t'_1$. Using Eq. (16-16), we have

$$v'_x = \operatorname*{limit}_{\Delta t \to 0} \frac{\gamma(\Delta x - u\Delta t)}{\gamma\left(\Delta t - \frac{u\Delta x}{c^2}\right)} = \frac{v_x - u}{1 - \frac{uv_x}{c^2}} \tag{16-23a}$$

In similar method, we obtain the transformation equation for the y and z components of the velocity

$$v'_y = \frac{v_y}{1 - \frac{uv_x}{c^2}} \frac{1}{\gamma} \tag{16-23b}$$

$$v'_z = \frac{v_z}{1 - \frac{uv_x}{c^2}} \frac{1}{\gamma} \tag{16-23c}$$

Note that $v'_y \neq v_y$, even though $\Delta y' = \Delta y$, because $\Delta t' \neq \Delta t$. Similar consideration holds for v'_z.

Switching primed and unprimed velocities in Eq. (16-23a), Eq. (16-23b) and Eq. (16-23c), and substituting $-u$ for u, we have

$$v_x = \frac{v'_x + u}{1 + \frac{ux}{c^2}} \tag{16-24a}$$

$$v_y = \frac{v'_y}{1 + \frac{uv'_x}{c^2}} \frac{1}{\gamma} \tag{16-24b}$$

$$v_z = \frac{v'_z}{1 + \frac{uv'_x}{c^2}} \frac{1}{\gamma} \tag{16-24c}$$

which is the inverse Lorentz transformation of velocities.

Eq. (16-23a)、Eq. (16-23b)、Eq. (16-23c) and Eq. (16-24a)、Eq. (16-24b)、Eq. (16-24c) give the Lorentz transformations of velocities. They are analogous to the equations of the Lorentz transformations of coordinates. They relate observations in one reference frame to observations in another. When $u \ll c$ (or equivalently, $c \to \infty$), it can be proved that Eq. (16-23a)、Eq. (16-23b)、Eq. (16-23c) reduce to

$$\begin{cases} v'_x = v_x - u \\ v'_y = v_y \\ v'_z = v_z \end{cases} \tag{16-25}$$

which is indeed the Galilean results.

Suppose that the common event being observed by both S and S' is the passage of a light beam along the x direction. Observer S measures $v_x = c$ and $v_y = v_z = 0$. What velocity does observer S' measure? Using Eq. (16-23a)、Eq. (16-23b)、Eq. (16-23c), we find the velocity components measured by S' to be

$$v'_x = \frac{v_x - u}{1 - \frac{uv_x}{c^2}} = \frac{c - u}{1 - \frac{uc}{c^2}} = c, \quad v'_y = 0, \quad v'_z = 0$$

Hence a speed of c measured by one observer must also be measured to be c by any other observer. This is the result of Einstein's second postulate. Thus the speed of light is indeed the same for all observers. The same conclusion holds for any direction of travel of light beam.

If $u > c$, the Lorentz factor $\gamma = 1/\sqrt{1 - \frac{u^2}{c^2}}$ will be complex. However, our world is real so that the speed of c is the ultimate speed limit. That is, no real body can move with a speed which exceeds c.

Example 16-5 A proton moving eastward with a speed of $0.60c$ in a nuclear-physics laboratory passes an electron moving westward with a speed of $0.90c$. Find the speed of the electron relative to a frame of reference riding with the proton.

Solution We call eastward the x positive direction. Hence

$$u = 0.60c, \quad v_x = -0.90c$$

By Eq. (16-23a), we obtain

$$v'_x = \frac{-0.90c - 0.60c}{1 - 0.90 \times 0.60} = \frac{-1.50c}{1.54} = -0.97c$$

which is the speed of the electron relative to the proton.

16.6 The Relativistic Dynamic Theory

So far we have investigated the effect of Einstein's two postulates on the kinematical variable time, displacement and velocity as viewed from two different inertial frames. In this section, we will discuss the dynamical variable momentum and mass and energy of a particle.

16.6.1 Relativistic mass and relativistic momentum

If a constant force $\boldsymbol{F}$ acts on a body of mass m indefinitely, it is clear that Newton's second law $\boldsymbol{F} = m\boldsymbol{a}$ must break down as the speed of the body approaches the speed of light. In the precious chapters, we have learned that the momentum conservation derived from Newton's second law is a universal theorem. According to Einstein's first postulates, if the momentum of a system is conservative in an inertial frame, it also is conservative in all others. Using the classical formula $\boldsymbol{p} = m\boldsymbol{v}$, it can be proved that the momentum may be not conserved in all the inertial frame. Therefore, we need to modify the Newton's law of mechanics.

Einstein showed by application of the law of conservation of momentum that if m_0 is the mass of a body when it is at rest relative to an inertial frame, its mass m, moving with a velocity $\boldsymbol{v}$ relative to this frame, is given by(see appendix 2)

$$m = \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}} \tag{16-26}$$

and its momentum is defined as

$$\boldsymbol{p} = \frac{m_0\,\boldsymbol{v}}{\sqrt{1-\frac{v^2}{c^2}}} \tag{16-27}$$

Here, m_0 is called the rest mass, and m the moving mass.

Using the definition (16-27), the Newton's second law has the same form as the classical case. That is

$$\boldsymbol{F} = \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{m_0\,\boldsymbol{v}}{\sqrt{1-\frac{v^2}{c^2}}}\right) \tag{16-28}$$

Eq. (16-26) and Eq. (16-28) are the relativistic extension of the classical Newton's first and second law. They are satisfied with Lorentz transformation. They have the same form in all the inertial reference frames. In the special relativistic theory, the Newton's third law contains the same meanings as the classical case.

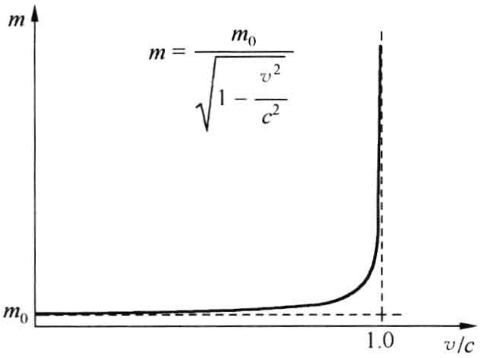


Fig. 16-8 The moving mass m as a function of speed v

According to Eq. (16-26), when a constant force acts on a body and the velocity of the body increases with the time, its mass (that is its inertial) will increases with the increase of the velocity. If the velocity of the body approaches the speed of light, its mass will approaches ∞, as shown in Fig. 16-8. Hence the speed of a body can not be in excess of the speed of light, which is the up limit of velocities.

16.6.2 Relativistic energy

From the Eq. (16-28), the relativistic expression of the kinetic energy of a particle can be derived by using essentially the same procedure we used to derive the classical expression, starting with the particle form of the work-energy theorem. Using the Newton's second law, we have

$$\begin{aligned}\mathrm{d}E_k &= \boldsymbol{F}\cdot\mathrm{d}\boldsymbol{s} = \boldsymbol{v}\cdot\mathrm{d}\boldsymbol{p} = \mathrm{d}(\boldsymbol{v}\cdot\boldsymbol{p}) - \boldsymbol{p}\cdot\mathrm{d}\boldsymbol{v}\\ &= \mathrm{d}(\boldsymbol{v}\cdot\boldsymbol{p}) - \frac{m_0\,\boldsymbol{v}\cdot\mathrm{d}\boldsymbol{v}}{\sqrt{1-\frac{v^2}{c^2}}}\\ &= \mathrm{d}(\boldsymbol{v}\cdot\boldsymbol{p}) - \frac{m_0 v\mathrm{d}v}{\sqrt{1-\frac{v^2}{c^2}}}\end{aligned}$$

Assuming the initial velocity is zero, the kinetic energy E_k at the time t is equal to

$$\begin{aligned}E_{\mathrm{k}} &= (\boldsymbol{v}\cdot\boldsymbol{p})\Big|_0^v - \int_0^v \frac{m_0 v}{\sqrt{1-\frac{v^2}{c^2}}}\mathrm{d}v \\ &= mc^2 - m_0c^2\end{aligned} \tag{16-29}$$

Eq. (16-29) looks very different from the classical result $\frac{1}{2}m_0v^2$. You can also see from the first term of Eq. (16-29) that $E_{\mathrm{k}}\to\infty$ as $v\to c$. Thus we can increase the kinetic energy of a particle without limit, and its speed will not exceed c.

We can also express Eq. (16-29) as

$$E_{\mathrm{k}} = E - E_0 \tag{16-30}$$

where the total relativistic energy $\boldsymbol{E}$ is defined as

$$E = \frac{m_0c^2}{\sqrt{1-\frac{v^2}{c^2}}} = mc^2 \tag{16-31}$$

and E_0 is defined as

$$E_0 = m_0c^2 \tag{16-32}$$

which is known as the rest energy. The rest energy is the total relativistic energy of a particle measured in a reference frame in which the particle is at rest. The total relativistic energy is given by Eq. (16-39) as

$$E = E_{\mathrm{k}} + E_0 = mc^2 \tag{16-33}$$

That is the total relativistic energy is equal to the sum of the kinetic energy and the rest energy. This is Einstein's famous mass-energy equation.

In interaction of particles at relativistic speeds, we can replace our previous principle of conservation of energy with one based on the total relativistic energy: in an isolated system of particles, the total energy relativistic energy remains constant. That is

$$\sum_i E_i = \sum_i m_ic^2 = \text{constant} \tag{16-34}$$

where m_i is the moving mass of the particle i. Eq. (16-34) can be expressed as

$$\sum_i m_i = \text{constant} \tag{16-35}$$

which implies that the total mass of the system is conservative. The system is isolated—no external work is done by its environment.

If we expand the square root in Eq. (16-29) assuming that v^2/c^2 is small compared with 1, we obtain

$$\begin{aligned}E_{\mathrm{k}} &= m_0c^2\left(1+\frac{1}{2}\frac{v^2}{c^2}+\frac{3}{8}\frac{v^4}{c^4}+\cdots-1\right) \\ &= \frac{1}{2}m_0v^2\left(1+\frac{3}{4}\frac{v^2}{c^2}+\cdots\right)\end{aligned} \tag{16-36}$$

In the limiting case, where v/c is small compared with 1, we obtain the classical formula

$$E_{\mathrm{k}} = \frac{1}{2}m_0v^2 \tag{16-37}$$

for the kinetic energy of a body. As v approaches c, we must turn to the relativistic relation for the kinetic energy as given in Eq. (16-29).

From Eq. (16-34) and Eq. (16-35), the conservation of total energy is equivalent to the conservation of relativistic mass. Einstein's well known $E = mc^2$ relation asserts that the rest energy is freely convertible into other forms. All reaction, whether chemical or nuclear, in which energy is released or absorbed involve a corresponding change in the rest energy of the reactants. The understanding of the mass converting into the energy is incorrect. The more total relativistic energy of a body, the greater its inertial. The increase of the mass of a body does not imply the increase of the numbers of atoms contained in the body in the relativistic theory.

The revolutionary formula $E = mc^2$ has proved to be a cornerstone of modern physics, permitting a new understanding of nuclear physics and leading to the development of nuclear fission as a practical energy source.

16.6.3 The relationship between the total energy and the momentum

In the classical mechanics, the momentum $\boldsymbol{p}$ of a particle is

$$\boldsymbol{p} = m_0 \boldsymbol{v}$$

and its kinetic energy

$$E_k = \frac{1}{2} m_0 v^2$$

If we eliminate v between these two expressions, we find a directive relation between the momentum and the kinetic energy

$$p^2 = 2E_k m_0 \quad \text{(classical)} \tag{16-38}$$

We can find a similar connection in relativity by eliminating v between the relativistic definition of momentum Eq. (16-27) and the relativistic definition of kinetic energy Eq. (1-29). Doing so leads, after some algebra, to

$$(pc)^2 = E_k^2 + 2E_k m_0 c^2 \quad \text{(relativistic)} \tag{16-39}$$

With the aid of Eq. (16-33), we can transform Eq. (16-39) into a relation between the momentum $\boldsymbol{p}$ and the total energy E of a particle

$$E^2 = (pc)^2 + (m_0 c^2)^2 \tag{16-40}$$

or

$$E = \sqrt{p^2 c^2 + m_0^2 c^4} \tag{16-41}$$

Eq. (16-41) is universally applicable. When the rest mass m_0 of a particle (photon) is equal to zero, we have from Eq. (16-41)

$$E = pc \tag{16-42}$$

which is very useful in the following chapter.

The relationships between kinetic energy and velocity and between kinetic energy and momentum can be tested in the relativistic regime by accelerating particles to high speeds or by using high-speed particles (namely, electrons) emitted in certain radioactive decay processes. The experimental data are in perfect agreement with the relativistic expression and in disagreement with the classical expression. Similar results are obtained indirectly today at every large accelerator facility in the world. Particles are accelerated to speeds very close to c, and the design parameters of the accelerators must be based on relativistic dynamics. Thus every modern accelerator is in effect a laboratory for testing special relativity. Needless to say, the success of these accelerators is a dramatic confirmation of special relativity.

Example 16-6 The total energy of an electron is $2m_0c^2$. Find the velocity of the electron.

Solution According to Eq. (16-31), we have

$$E = \frac{m_0 c^2}{\sqrt{1-\frac{v^2}{c^2}}} = 2m_0 c^2$$

Hence

$$\frac{1}{\sqrt{1-\frac{v^2}{c^2}}} = 2$$

from which we obtain

$$v = \frac{\sqrt{3}}{2}c$$

Example16-7 Find the work which must be done on an electron to increase its speed from (1) 0. 60c to 0. 80c; (2) 0. 80c to 0. 90c, where c is the speed of light.

Solution The work is equal to the increase of the total energy of the electron. When the speed increase from 0. 60c to 0. 80c, the work done on the electron is

$$W_1 = \frac{m_0 c^2}{\sqrt{1-0.80^2}} - \frac{m_0 c^2}{\sqrt{1-0.60^2}} = \frac{5}{12}m_0 c^2 = 0.42 m_0 c^2$$

where $m_0c^2 = 0.511$ MeV. The work done on the electron with an increase of speed from 0. 80c to 0. 90c is

$$W_2 = \frac{m_0 c^2}{\sqrt{1-0.90^2}} - \frac{m_0 c^2}{\sqrt{1-0.80^2}} = 1.0 m_0 c^2$$

Example16-8 What must be the momentum of a particle with rest mass m_0 in order that its total energy is three times its rest energy?

Solution According to the given condition, we have

$$E = 3m_0 c^2$$

Using Eq. (16-40), we obtain

$$p = \frac{\sqrt{E^2 - m_0^2 c^4}}{c} = \sqrt{8} m_0 c$$

Appendix 1

As shown in Fig. 16-2, considering the symmetry and the homogeneity of space and time, the needed transformation is written as

$$\begin{cases} x' = Ax + Bt \\ y' = y \\ z' = z \\ t' = Cx + Dt \end{cases} \tag{A1-1}$$

where A, B, C and D are, independent of the (x, y, z, t), constants. Noting that the coordinate of the origin O' is zero relative to the S' and $x = ut$ relative to the S, we have

$$B = -Au \tag{A1-2}$$

Considering the origin o in the same way, we obtain

$$D = A \tag{A1-3}$$

Assume that $x' = x$ at $t' = t = 0$ at which the beam of light is emitted in the direction $Ox(O'x')$. The position of the wavefront is $x = ct$ measured by S and $x' = ct'$ measured by S'. This fact implies

$$C = -\frac{Au}{c^2} \tag{A1-4}$$

Using Eq. (A1-2), Eq. (A1-3) and Eq. (A1-4), the transformation (A1-1) reduces to

$$\begin{cases} x' = A(x - ut) \\ y' = y \\ z' = z \\ t' = A\left(t - \frac{ux}{c^2}\right) \end{cases} \tag{A1-5}$$

The description of the wavefront of light emitted at time $t = t' = 0$ in the direction Oy is given by

$$\begin{cases} x = 0 \\ y = ct \end{cases} \quad (\text{to } S) \tag{A1-6}$$

or

$$x'^2 + y'^2 = c^2 t'^2 \quad (\text{to } S') \tag{A1-7}$$

Inserting Eq. (A1-5) into Eq. (A1-7) and using Eq. (A1-6) yields

$$A = \frac{1}{\sqrt{1 - \frac{u^2}{c^2}}} \tag{A1-8}$$

Hence we obtain the Lorentz transformation (16-10).

Appendix 2

Here, we consider a sample example in which a particle decays into two same particles. As shown in Fig. 16-9, the particle C, located at the origin O', is at rest relative to the reference S' and broken into two same particles A, B (having the same mass). According to the conservation of the momentum, if the particles A moves in the direction $-O'x'$ with the velocity $-v$, the particle B will move in the direction $O'x'$ with the velocity v.

Suppose the reference S moves along the $O'x'$ with the velocity $-v$. To S, the particle A is at rest with the mass m_0 (the rest mass), and the particle B have the mass m (the moving mass) and moves at the velocity

$$v_B = \frac{2v}{1 + \frac{v^2}{c^2}} \tag{A2-1}$$

Assuming the mass of the particle C is M, the conservation of the momentum measured by S implies

$$Mu = mv_B \tag{A2-2}$$

It is reasonable that we take

$$M = m_0 + m \tag{A2-3}$$

That is, the mass of system is conservative. From Eq. (A2-1), Eq. (A2-2) and Eq. (A2-3), we obtain

$$m = m_0 \frac{1+\frac{v^2}{c^2}}{1-\frac{v^2}{c^2}} \tag{A2-4}$$

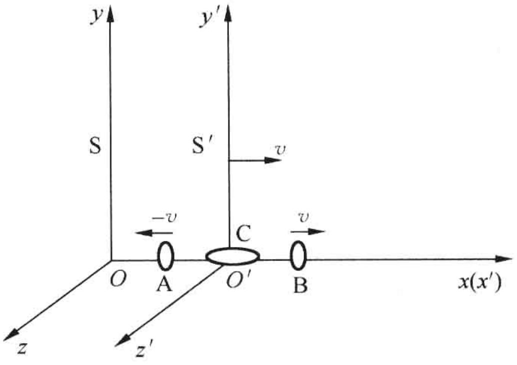


Fig. 16-9

Using Eq. (A2-1), v is equal to

$$v = \frac{c^2}{v_B}\left(1 - \sqrt{1 - v_B^2/c^2}\right) \tag{A2-5}$$

Substituting Eq. (A2-5) into Eq. (A2-4), we have

$$m = \frac{m_0}{\sqrt{1 - \frac{v_B^2}{c^2}}} \tag{A2-6}$$

The equation tells us that the particles A and B, having the same mass relative to the observer in the reference S', have the different mass measured by the observer in the reference S. If the particle A and B are at rest relative to S, their masses would be m_0. Hence, the moving mass of the particle, which has the rest mass of m_0 and moves at the velocity v, is given by

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{A2-7}$$

That is Eq. (16-26).

Questions

16-1 Why don't we observe relativistic effects in our everyday lives?

16-2 A quasar (quasi-stellar object) travels away from the Earth at half the speed of light. What is the speed, with respect to the Earth, of the light we detect coming from it?

16-3 A beam from a laser falls at right angle on a plane mirror and reflects from it. What is the speed of the reflected beam if the mirror is (1) fixed in the laboratory and (2) moving directly toward the laser with speed V?

16-4 Although in relativity (where motion is relative and not absolute) we find that "moving clocks run slow", this effect had nothing to do with the motion altering the way a clock works. With what does it have to do?

16-5 We have seen that if several observers watch two events, labeled A and B, one of them may say that event A occurred first but another may claim that it was event B that did so. What would you say to a friend who asked you which event really did occur first?

16-6 Two observers, one at rest in S and one at rest in S', each carries a meterstick oriented parallel to their relative motion. Each observer finds upon measurement that the other observer's meterstick is the shorter of two sticks. Does this seem like a paradox to you?

16-7 If zero-mass particles have a speed c in one reference frame, can they be found at rest in any other frame? Can such particles have any speed other than c?

16-8 A particle with zero mass (a Neutrino, possibly) can transport momentum. But, by Eq. (16-27)

$$p=\frac{m_0 v}{\sqrt{1-\frac{v^2}{c^2}}}$$

the momentum is directly proportional to the mass and therefore should be zero if the mass is zero. Explain.

16-9 Is the mass of a stable, composite particle (a gold nucleus, for example) greater than, equal to or less than the sum of the masses of its constituents? Explain.

16-10 "The relation $E=mc^2$ is essential to the operation of a power plant based on nuclear fission but has only a negligible relevance for a fossil-fuel plant." Is this a true statement? Explain why or why not.

Problems

16-1 Observer S assigns to an event the coordinates $x=100$ km, $t=200$ μs. Find the coordinates of this event in frame S', which moves in the direction of increasing x with speed 0. 950c. Assume $x=x'=0$ at $t=t'=0$.

16-2 Inertial frame S' moves at a speed of 0. 60c with respect to frame S in the direction of increasing x. In frame S, event 1 occurs at the origin at $t=0$ and event 2 occurs on the x axis at $x=3.0$ kmand at $t=4.0$ μs. What times of occurrence does observer S' record for these same events? Explain the reversal of the time order.

16-3 An experimenter arranges to trigger two flashbulbs simultaneously, a blue flash located at the origin of his reference frame and a red flash at $x=34.4$ km. A second observer, moving at a speed of 0. 247c in the direction of increasing of x, also views the flashes.

(1) What time interval between them does she find?

(2) Which flash does she say occurs first?

16-4 An observer S sees a flash of red light 1200 m from his position and a flash of blue light 720 m closer to him and on the same straight line. He measures the time interval between the occurrences of the flashes to be 5. 00 μs, the red flash occurring first.

(1) What is the relative velocity u (magnitude and direction) of a second observer S' who would record these flashes as occurring at same place?

(2) From the point of view of S', which flash occurs first?

(3) What time interval between them would S' measures?

16-5 The mean lifetime of muons stopped in a lead block in laboratory is measured to be 2. 20 μs. The mean lifetime of high-speed muons in a burst of cosmic rays observed the earth is measured to be 16 μs. Find the speed of these cosmic ray muons.

16-6 An unstable high-energy particle enters a detector and leaves a track 1. 05 mm long before it decays. Its speed relative to the detector was 0. 992c. What is its proper time? That is, how long would it have lasted before decay had it been at rest with respect to the detector?

16-7 The length of a spaceship is measured to be exactly half its rest length.

(1) What is the speed of the spaceship relative to the observer's frame?

(2) By what factor do the spaceship's clocks run slow, compared to clocks in the observer's frame?

16-8 A space traveler takes off from Earth and moves at speed 0. 988c toward the star Vega, which is 26. 0 ly distant. How much time will have elapsed by Earth clock (1) when the traveler reaches Vega and (2) when Earth observers receive word from him that he has arrived? (3) how much older will the observers calculate the traveler to be when he reaches Vega than he was when he started the trip?

16-9 (1) Can a person, in principle, travel from Earth to the galactic center (which is about 23 000 ly distant) in a normal lifetime? Explain, using either time dilation or length-contraction arguments.

(2) What constant speed would be needed to make the trip in 30 y (proper time).

16-10 An alpha particle moving east with a speed of $0.6c$ is passed by an electron moving west with a speed of $0.9c$. Find the speed of the electron relative to the alpha particle.

16-11 Find the kinetic energy, total energy, mass and momentum of an electron moving at a speed of $0.900c$.

16-12 Through what potential difference must protons be accelerated to achieve a speed of $0.800c$? What is the momentum of such a proton? The total energy?

16-13 Calculate the speed of a particle (1) whose kinetic energy is equal to twice its rest energy and (2) whose total energy is equal to twice its kinetic energy.

16-14 Find the momentum of a particle of mass m_0 in order that its total energy is three times its rest energy.

16-15 How much work must be done to increase the speed of an electron (1) $0.18c$ to $0.19c$ and (2) $0.98c$ to $0.99c$? Note that the speed increase ($0.01c$) is the same in each case.

Chapter 17

The Quantization of Light

The birthday of quantum physics is considered to be the date when Max Planck read his paper, "on the theory of the energy distribution law of the normal spectrum" on Dec. 14, 1900. This paper, which first attracted little attention, was the start of a revolution in physics. In this and following chapters, we shall examine the major milestones and study the basic theory of quantum mechanics.

Like relativity, quantum physics represents a generalization of classical physics that include the classical law as special cases. Just as relativity extends the range of application of physical laws to the region of high velocities, so quantum physics extends that range to the region of small dimensions. Just as a universal constant of fundamental significance, the velocity of light c, characterizes relativity, so a universal constant of fundamental significance, now called Planck's constant h, characterizes quantum physics. It was trying to explain the observed properties of thermal radiation that Planck introduced this constant in his 1900 paper. Let us now begin to examine thermal radiation ourselves.

17.1 Thermal Radiation and Planck's Theory of Radiation

17.1.1 Thermal radiation

It is a matter of common experience that a hot body radiates electromagnetic energy in the form of heat. In fact, at any temperature, a body emits radiation of all wavelengths, but the distribution in wavelength, the spectral distribution, depends on temperature. At low temperature, most of the energy is in the form of low frequency infra-red radiation, but as the temperature increases more energy is radiated at higher frequencies, up to about 500℃ radiation of visible light is observed. At still higher temperature, such as that of an incandescent lamp filament the spectral distribution has shifted sufficiently to the higher frequencies for the body to be white hot. Not only the spectral distribution changes with temperature, but the total energy radiated also changes, increasing as a body becomes hotter.

All bodies emit radiation to their surroundings and the radiation may fall on a body. When radiation falls on a body, some is reflected and some is absorbed. For example, dark bodies absorb most of the radiation falling on them, while light colored bodies reflect most. Hence, the three processes take place simultaneously on a body as shown in Fig. 17-1. If a body is at first hotter than its surroundings, it will cool off because its rate emitting energy exceeds its rate of absorbing energy. When thermal equilibrium is reached, the rates of emission and absorption are equal and the body must emit and absorb the same amount of radiant energy per unit time. The radiation emitted or absorbed under this circumstance is called thermal radiation.

A blackbody is defined as a body that absorbs all radiations falling upon it, which is an ideal system. In the following, we will study only the blackbody radiation. Generally speaking, the detailed form of the spectrum of the thermal radiation emitted by a hot body depends somewhat upon the composition of the body. However, experiment shows that the blackbody radiation has thermal spectra of a universal character. Independent of the details of their composition, it is found that all black bodies at the same temperature emit thermal radiation with the same spectrum. The universal properties of the radiation emitted by blackbodies make them of particular theoretical interest and physicists sought to explain the specific features of their spectrum. As shown in Fig. 17-2, a cavity, which is connected to the outside by a small hole, may be looked as a blackbody. The radiation incident upon the hole from the outside enters the cavity and is reflected back and forth by the walls of the cavity, eventually being absorbed on these walls. If the area of the hole is very small compared to the area of the inner surface of the cavity, a negligible amount of the incident radiation will be reflected back through the hole. Essentially all the radiation incident upon the hole is absorbed; therefore, the hole must have the properties of the surface of a blackbody and may be looked as a blackbody. The idea blackbody does not exist in the word. Most black bodies used in laboratory experiments are constructed along these lines and the emitted radiation from the hole is measured by the receiver.

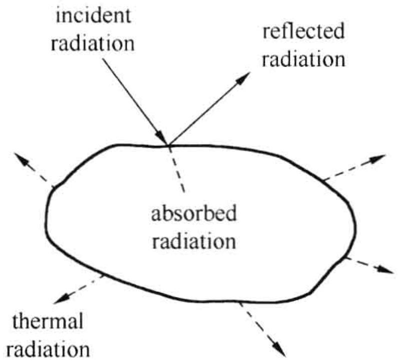


Fig. 17-1 While the body emits radiation, it reflects some of the radiations falling on it and absorbs others

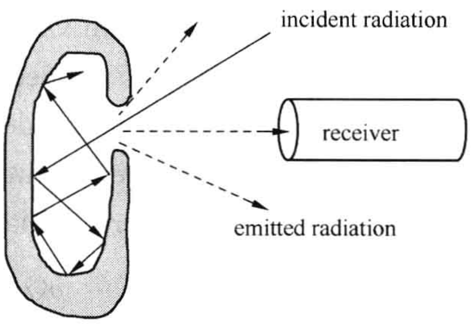


Fig. 17-2 A cavity in a body connected by a small hole to the outside can be looked as a blackbody. The hole emits like a black body

17. 1. 2 The spectral radiancy of blackbody radiation

The emissivity M of a surface is defined as the energy of the radiation emitted from unit area of body per unit time. The spectral radiancy, represented by $M(\lambda, T)$, is defined so that $M(\lambda, T)\mathrm{d}\lambda$ is equal to the energy emitted per unit time in radiation of frequency in the interval λ to $\lambda+\mathrm{d}\lambda$ from a unit area of the surface at absolute temperature T. The spectral distribution of blackbody radiation is specified by the quantity $M(\lambda, T)$. The dependence of $M(\lambda, T)$ on λ and T is shown in Fig. 17-3. Note that the wavelength at which the maximum radiancy occurs (dashed line) decreases linearly with increasing temperature, and that the total power per square meter of the radiator (area under curve) increases very rapidly with temperature.

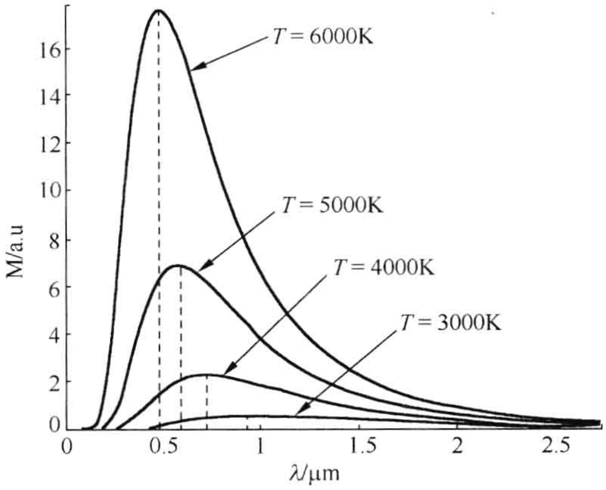


Fig. 17-3 The spectral radiancy of a blackbody as a function of the wavelength of radiation

The integral of spectral radiancy $M(\lambda, T)$ over all λ is the total energy emitted per unit time per unit area from blackbody at temperature T. It is called the radiancy $E(T)$. That is

$$E(T) = \int_0^{\infty} M(\lambda, T)\,\mathrm{d}\lambda \qquad (17\text{-}1)$$

$E(T)$, equal to the area under the curve in Fig. 17-3, increases rapidly with increasing temperature. In fact, this result is called Stefan's law, and it was first stated in 1879 in the form of an empirical equation

$$E(T) = \sigma T^4 \qquad (17\text{-}2)$$

where

$$\sigma = 5.67 \times 10^{-8}\,\mathrm{W/(m^2 \cdot K^4)}$$

is called the Stefan-Boltzmann constant. Fig. 17-3 also shows us that the spectrum shifts toward lower wavelength as T increases. This result is called Wien's displacement law

$$\nu_{\max} \propto T \qquad (17\text{-}3)$$

or

$$\lambda_{\max} T = b \qquad (17\text{-}4)$$

The constant b has the value of $2.898 \times 10^{-3}\,\mathrm{m \cdot K}$. $\nu_{\max}$ (or $\lambda_{\max}$) is the frequency (wavelength) at which $M(\lambda, T)$ has its maximum value for a particular T.

It is more useful, however, to specify the spectrum of radiation inside the cavity in terms of an energy density $\rho(\lambda, T)$, which is defined as the energy contained in a unit volume of cavity at temperature T in the frequency interval λ to $\lambda + \mathrm{d}\lambda$. It is evident that $\rho(\lambda, T)$ is related by a geometrical factor to the spectral radiancy $M(\lambda, T)$. It can be shown that

$$\rho(\lambda, T) = \frac{4}{c} M(\lambda, T) \qquad (17\text{-}5)$$

Hence, the radiation inside a cavity whose walls are at temperature T has the same character as the radiation emitted by the surface of a blackbody at temperature T.

17.1.3 Planck's theory of cavity radiation

Shortly after the turn of the present century, using the statistical thermodynamics and the ordinary laws of mechanics and electromagnetic theory, Rayleigh and also Jeans, made a calculation of the energy density of cavity (or blackbody) radiation that points up a serious conflict between classical physics and experiment results. Their result is

$$\rho(\lambda, T) = \frac{8\pi k T}{\lambda^4} \qquad (17\text{-}6)$$

This is the Rayleigh—Jeans formula for blackbody radiation. Eq. (17-6) has shown that in the limit of long wavelengths, the classical spectrum approaches the experimental results, but as the wavelength becomes short, the theoretical prediction goes to infinity. Experiment shows that the energy density always remains finite, as it obviously must, and in fact, that the energy density goes to zero at very long wavelengths. The grossly unrealistic behavior of the prediction of classical theory at short wavelengths is known in physics as the "ultraviolet catastrophe". This term is suggestive of the importance of the failure of the theory.

In December, 1900, in trying to resolve the discrepancy between theory and experiment, M. Planck presented a new form of the blackbody radiation spectral distribution, based on a revolutionary hypothesis. He postulated that there are many oscillators in the blackbody and the energy of an oscillator of a given frequency ν cannot take arbitrary values between zero and infinity, but can only take on the discrete values $n\varepsilon_0$, where n is a positive integer or zero, and ε_0 is a finite amount, or quantum, of energy, which may depend on the frequency. In order to satisfy Wien's displacement law, ε_0 must be taken to be proportional to the frequency ν

$$\varepsilon_0 = h\nu \tag{17-7}$$

where h is a fundamental constant, called Planck's constant, having a value of

$$h = 6.62618 \times 10^{-34}\,\mathrm{J \cdot s}$$

The exchange of energy between oscillator and radiation must be equal to $nh\nu$. Applying the above postulate to the problem of blackbody radiation, Planck obtained

$$\rho(\lambda, T) = \frac{8\pi hc}{\lambda^5} \frac{1}{e^{\frac{hc}{kT\lambda}} - 1} \tag{17-8}$$

or

$$M(\lambda, T) = \frac{2\pi hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{kT\lambda}} - 1} \tag{17-9}$$

where k is the Boltzmann constant. This is Planck's blackbody radiation formula. Using the values of h and k, Planck obtained very good agreement with the experimental data for $M(\lambda, T)$ over the entire range of wavelength.

Stefan's law is obtained by integrating Planck's formula over the entire spectrum of frequencies. The radiancy is found to be proportional to the fourth power of the temperature with the theoretical value of Stefan's constant

$$\sigma = \frac{2\pi^5 k^4}{15c^2 h^3} \tag{17-10}$$

Wien's displacement law is obtained by setting $\mathrm{d}M(\lambda, T)/\mathrm{d}\lambda = 0$

$$\lambda_{\max} T = \frac{hc}{4.965k} = b \tag{17-11}$$

Using the experimental value of c, σ and b, the values of h and k can be calculated from Eq. (17-10) and Eq. (17-11). Indeed, this was done by Planck, his values agreeing very well with those obtained subsequently by other methods.

The idea of quantization of energy, in which the energy of a system can only take certain discrete values, was totally at variance with classical physics, and Planck's theory was not accepted readily. However, it was not long before the quantum concept was used to explain other phenomena.

17.2 The Photoelectric Effect

In the course of experiments investigating the properties of electromagnetic waves, H. Hertz discovered in 1887 that ultraviolet light falling on metallic electrodes facilitates the passage of a spark. Further work by W. Hallwacks, M. Stoletov, P. Lenard and others showed that charged particles are ejected from metal surface irradiated by high frequency electromagnetic waves. This phenomenon is called the photoelectric effect. In 1900, Lenard measured the charge to mass ratio of the charged particles in experiments similar to those of J. J. Thomson, and in this way he was able to identify the particles as electrons.

17.2.1 The results of the experiments

Fig. 17-4 shows an apparatus used to study the photoelectric effect. A glass envelope encloses the apparatus in an evacuated space. Monochromatic light, incident through a quartz window, falls on the metal plate K and liberates electrons, called photoelectrons. The electrons can be detected as a current if they are attracted to the metal plate A by means of a potential difference V applied between K and A. The sensitive ammeter G serves to measure this photoelectric current.

The curve of Fig. 17-5 is a plot of the photoelectric current, in an apparatus like that of Fig. 17-4, as a function of the potential difference V. If V is made large enough, the photoelectric current reaches a certain limiting (saturation) value at which all photoelectrons ejected from K are collected by the plate A. If V is reversed in sign, the photoelectric current does not immediately drop to zero, which suggests that the electrons are emitted from K with kinetic energy. Some will reach up A in spite of the fact the electric field opposes their motion. However, if this reversed potential difference is made large enough, a value V_0 called the stopping potential is reached at which the photoelectric current does drop to zero. This potential difference V_0, multiplied by electron charge, measures the kinetic energy $E_{\max}$ of the fastest ejected photoelectron. That is

$$E_{\max} = \frac{1}{2}mv_{\max}^2 = eV_0 \tag{17-12}$$

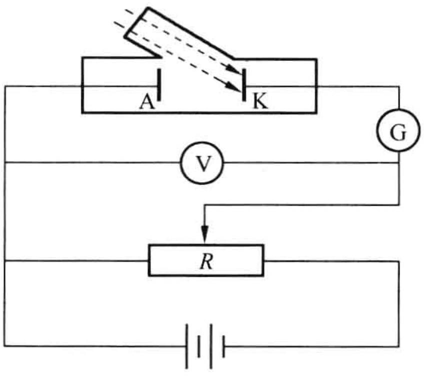


Fig. 17-4 Schematic diagram of an apparatus used to study the photoelectric effect

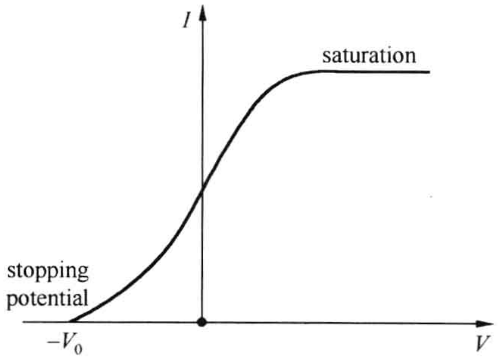


Fig. 17-5 Graphs of current I as a function of potential difference V

The quantity $E_{\max}$ turns out experimentally to be independent of the intensity of the light.

The most important features of the experimental data are the following :

(1) There is a minimum, or cutoff frequency ν_0 of the radiation, below which no emission of electron takes place and no photoelectric effect occurs, no matter what the intensity of the incident radiation, or for how long it falls on the surface;

(2) Electrons emerge with a range of velocities from zero up to a maximum $v_{\max}$ and the maximum kinetic energy, $mv_{\max}^2/2$, is found to depend linearly on the frequency of the radiation and to be independent of its intensity;

(3) For incident radiation of a given frequency, the number of electrons emitted per unit time is proportional to the intensity of the radiation;

(4) Electron emission takes place immediately as the light shines on the surface with no detectable time delay.

The above features cannot be explained in terms of the classical wave theory of light. According to the classical physics, it would be natural to suppose that the maximum kinetic energy of the emitted electrons would increase with the energy density (or intensity) of the incident radiation, independently of the frequency. Another important aspect of classical theory is that the incident energy is spread uniformly over the illuminated surface. To eject an electron from an atom, this energy would have to be concentrated over an area of atomic dimensions, and to achieve such a concentration would require a certain time delay. According to the wave theory the photoelectric effect should occur for any frequency of the light, provided only that the light is intense enough to give the energy needed to eject the photoelectrons.

17.2.2 Einstein's quantum theory of the photoelectric effect

In 1905, Einstein offered an explanation for these seemingly strange aspects of the photoelectric effect, based on an extension of Planck's idea of the quantization of blackbody radiation. In Planck's theory, the oscillators representing the source of the electromagnetic field could only vibrate with energies given by $E = nh\nu$. Planck believed that electromagnetic energy, once radiated, spreads through space as if water waves spread through water. In contrast, Einstein supposed that the electromagnetic field itself was quantized and that light consists of corpuscles, called light quanta or photons and that each photon travels with the velocity of light c and carries a quantum of energy of magnitude

$$E = h\nu = \frac{hc}{\lambda} \tag{17-13}$$

The photons are sufficiently localized, so that the whole quantum of energy can be absorbed by a single atom at one time.

When a photon falls on a metallic surface, its entire energy $h\nu$ is used to eject an electron from an atom. Because of the interaction of the ejected electron with other atoms it requires a certain minimum energy to escape from the surface. The minimum energy required to escape depends on the metal and is called the work function W. It follows that the maximum kinetic energy of a photoelectron is given by

$$\frac{1}{2}mv_{\max}^2 = h\nu - W \tag{17-14}$$

This result is called Einstein's equation. The cutoff frequency ν_0 is determined by the work function since in this case, $v_{\max} = 0$, from which

$$\nu_0 = \frac{W}{h} \tag{17-15}$$

If the frequency is reduced below ν_0, the individual photon, no matter how many of them there are (that is, no matter how intense the illumination), will not have enough energy individually to eject photoelectron.

The number of electrons emerging from the metal surface per unit time is proportional to the number of photons striking the surface per unit time, but the intensity of the radiation is also proportional to the number of photons falling on a certain area per unit time, since each photon carries a fixed energy $h\nu$. It follows that the number of electrons emitted (or the photoelectric current) is proportional to the intensity of the radiation.

According to Eq. (17-14), the maximum kinetic energy of ejected electron depends only on the frequency of the radiation. Doubling the light intensity merely doubles the number of the photons and thus doubles the photoelectric current; it does not change the energy $h\nu$ of the individual photons and the kinetic energy of the ejected electrons.

If there is any illumination at all incident on the plate K, then there will be at least one photon that hit it, and this photon will be immediately absorbed by some atom, if the frequency of the photon is greater than ν_0, leading to the immediate emission of a photoelectron.

Combining Eq. (17-12) with Eq. (17-14), we see that stopping potential satisfies

$$eV_0 = h\nu - W \tag{17-16}$$

Thus Einstein's theory predicts a linear relationship between the stopping potential V_0 and the frequency ν, in complete agreement with experimental results. Eq. (17-16) can be rewritten as

$$V_0 = \frac{h}{e}\nu - \frac{W}{e} \tag{17-17}$$

The slope in Eq. (17-17) is h/e. Hence, h/e can be determined from the experimental data. Therefore all the experimental observations are explained by Einstein's theory.

Notice that the photons are absorbed in the photoelectric process. This requires the electrons to be bound to atoms, or solids, for a truly free electron cannot absorb a photon and conserve both total relativistic energy and momentum in the process. We must have a bound electron, and therefore, the bonding forces serve to transmit momentum to the atom or solid. Due to the large mass of an atom, or solid comparing to the electron, our photoelectric energy Eq. (17-14) remains valid.

Although the photoelectric effect provides compelling evidence for a corpuscular theory of light, it must not be forgotten that the existence of diffraction and interference phenomena demonstrates that light also exhibits a wave behavior. In the next section, we will study another important experiment which further confirms the particle—like properties of radiation.

Example 17-1 The potassium foil, its work function being 2.21 eV, is irradiated by a beam of light of wavelength $\lambda = 2500 \times 10^{-10}$ m with the intensity of 2 W/m^2. Find

(1) the maximum kinetic energy of the photoelectrons;

(2) the maximum number of the photoelectrons per unit time per unit area from the surface of the potassium foil.

Solution (1) According to the Eq. (17-14), we have

$$E_{\max} = \frac{1}{2}mv_{\max}^2 = h\frac{c}{\lambda} - W = 2.76(\text{eV})$$

which is the maximum kinetic energy of the photoelectrons.

(2) The energy of a photon is equal to

$$\varepsilon = h\nu = 4.97 \text{ eV} = 7.95 \times 10^{-19} \text{ J}$$

Assuming a photon can knock out an electron, the maximum number of photoelectrons is as same as the number of photons falling on the surface of the potassium foil. Therefore, the maximum number of photoelectrons per unit time per unit area from the surface of the potassium foil is given by

$$I = Nh\nu \rightarrow N = \frac{2}{7.95 \times 10^{-19}} = 2.52 \times 10^{18} (\text{m}^{-2} \cdot \text{s}^{-1})$$

17.3 The Compton Effect

The corpuscular nature of electromagnetic radiation received dramatic confirmation in 1923 from the experiments of A. H. Compton. These experiments proved to be a great "convincer" of the reality of photons because they introduced the photon momentum, as well as the photon energy. Further, it showed that the photon model applies not only to visible and ultraviolet light—the domain of the photoelectric effect, but also to X-rays.

Compton allowed a beam of X-rays of sharply defined wavelength λ to fall on a graphite target, as shown in Fig. 17-6. The X-rays were scattered through the scatterer (target). For various angles of scattering, he measured the intensity of the scattered X-rays as a function of their wavelengths. Fig. 17-7 shows his experimental results. We see that although part of the scattered radiation had the same wavelength λ as the incident radiation, there was also a second component of wavelength λ', where $\lambda' > \lambda$. The shift in wavelength between the incident and scattered radiation, the Compton shift $\Delta\lambda = \lambda' - \lambda$ was found to vary with the angle of scattering and to be proportional to $\sin^2\varphi/2$ where φ is the angle between the incident and scattered beams

$$\Delta\lambda = 2\lambda_c \sin^2\frac{\varphi}{2} \tag{17-18}$$

where $\lambda_c = 2.43 \times 10^{-12}$ m. This phenomenon is called the Compton effect. Further investigation showed $\Delta\lambda$ to be independent of both λ and the material used as the scatterer.

The Compton effect could not be explained by the classical physics. In the classical model the oscillating electric field (X-rays: classical electromagnetic wave) vector in the incident wave of frequency ν acts on the free electron (the outer electron of the atoms) in the scattering target and sets them oscillating at that same frequency. These oscillating electrons, like charges surging back and forth in a small radio transmitting antenna, radiate electromagnetic

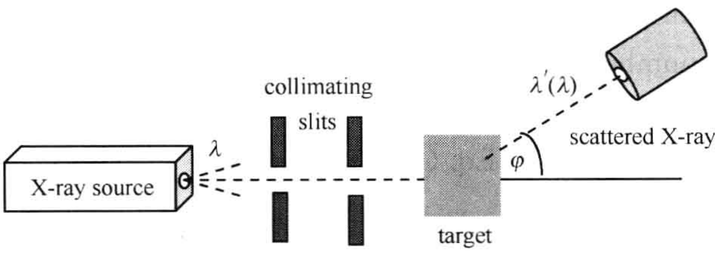


Fig. 17-6 Schematic diagram of Compton apparatus

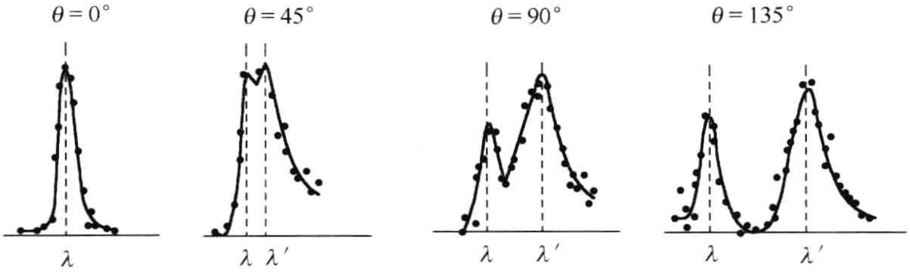


Fig. 17-7 Compton effect of the scattering of X-ray by graphite

waves that again have this same frequency ν. Hence, in the classical picture the scattered wave should have the same frequency ν and the same wave length λ as the incident wave.

Compton interpreted his experimental results by postulating that the incident X-ray beam was not a wave of frequency ν but a collection of photons, each of energy

$$E = h\nu_0 \tag{17-19}$$

and momentum

$$p = \frac{E}{c} = \frac{h\nu}{c} \tag{17-20}$$

and that these photons collided with the electrons in the scattering target as in a collision between billiard balls.

The electron in the scattering can be divided into two parts:

(1) The outer bound electrons in the atoms of the target, looked as free since their binding energies of a few electron volts are small compared with the energy of an X-ray photon ($\sim 10^4$ eV). The collisions between the X-ray photons and these electrons are equivalent to the collisions between the X-ray photons and free electrons.

(2) The inner bound electrons in the atoms of the target, which can not be treated as free since they are tightly bound to the atoms of the target. The collisions between the X-ray photons and this kind of electrons are equivalent to the collisions between the X-ray photons and the whole atoms of target.

Therefore, there are two kinds of collisions in the Compton experiments.

Let us consider the scattering of an X-ray photon by a free electron, which can be taken to be at rest initially. Since the energies involved in the collisions may be large, we need to use relativistic kinematics. As shown in Fig. 17-8, conservation of momentum yields

$$\boldsymbol{p} = \boldsymbol{p}' + \boldsymbol{p}_e \tag{17-21}$$

and conservation of energy is expressed as

$$h\nu + m_0c^2 = h\nu' + mc^2 \qquad (17\text{-}22)$$

where m_0 is the rest mass of electron and m is the moving mass of recoiling electron.

From Eq. (17-21), we obtain

$$p_e^2 = p^2 + p'^2 - 2pp'\cos\varphi \qquad (17\text{-}23)$$

Conservation of energy then gives

$$h\nu + m_0c^2 = h\nu' + (m_0^2c^4 + p_e^2c^2)^{1/2} \qquad (17\text{-}24)$$

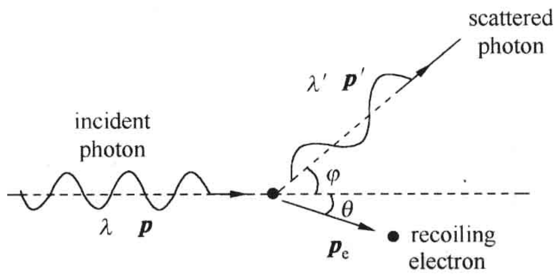


Fig. 17-8 Diagram of Compton scattering of a photon by an electron

and defining the kinetic energy of the electron after the collision as E_k, we have

$$\begin{aligned} E_k &= (m_0^2c^4 + p_e^2c^2)^{1/2} - m_0c^2 \\ &= h\nu - h\nu' = c(p - p') \end{aligned} \qquad (17\text{-}25)$$

From Eq. (17-25), we have that

$$p_e^2 = (p - p')^2 + 2m_0c(p - p') \qquad (17\text{-}26)$$

and combining Eq. (17-26) and Eq. (17-23), we find that

$$m_0c(p - p') = 2pp'\sin^2\varphi/2 \qquad (17\text{-}27)$$

Since $\lambda = h/p$ and $\lambda' = h/p'$, this can be written in form

$$\Delta\lambda = \lambda' - \lambda = 2\lambda_c\sin^2\varphi/2 \qquad (17\text{-}28)$$

where the constant is given by

$$\lambda_c = \frac{h}{m_0c} \qquad (17\text{-}29)$$

and called the Compton wavelength of the electron. Eq. (17-28) is known as the Compton equation. The calculated value of λ_c is 2.43×10^{-12} m, and this agrees very well with the experimental data. The Compton Eq. (17-28) predicts the experimentally observed Compton shifts.

Considering collisions between the X-photons and the whole atom (or the tightly bound electrons in the atom of target) in the same way, we have

$$\Delta\lambda' = \lambda' - \lambda = 2\lambda'_c\sin^2\varphi/2 \qquad (17\text{-}30)$$

where

$$\lambda'_c = \frac{h}{Mc} \qquad (17\text{-}31)$$

in which M is the mass of the atom. Since $M \gg m_0$, $\lambda'_c \approx 0$, and $\lambda' \approx \lambda_0$. Therefore the Compton shift for collisions with tightly bound electrons is immeasurably small, and the scattered photon is essentially unmodified in wavelength. To summarize, some photons are scattered from "free" electrons, which are modified in wavelength by the collision, and other photons are scattered from the "whole" atom without the change in wavelength.

For the reason, there is no Compton shift for light in the visible region, because the energy of photon in this case is not large compared with the binding energy of even the loosely bound electrons. In contrast, for very energetic γ-rays only the shifted line is observed, since the photon energies are larger compared with the binding energies of even the tightly bound electrons.

In conclusion, the experimental results are explained by the postulate that the radiation consists of the photons and this fact shows that we must accept the particle aspect of radiation.

Example 17-2 Find the energy, momentum and mass of a photon which has a wavelength of (1) 700 nm, (2) 0.500×10^{-10} mand (3) 1.20×10^{-12} m.

Solution The energy, momentum and mass of photon are given by the following equations

$$E = h\nu = \frac{hc}{\lambda}, \quad p = \frac{E}{c} = \frac{h}{\lambda}, \quad m = \frac{E}{c^2} = \frac{h}{\lambda c}$$

(1) When $\lambda = 700$ nm, E, p, and m are equal to

$$E = 1.78 \text{ eV}, \quad p = 9.46 \times 10^{-28} \text{ kg} \cdot \text{m/s}, \quad m = 3.15 \times 10^{-36} \text{ kg}$$

(2) When $\lambda = 0.500 \times 10^{-10}$ m, E, p and m are equal to

$$E = 2.48 \times 10^{4} \text{ eV}, \quad p = 1.25 \times 10^{-23} \text{kg} \cdot \text{m/s}, \quad m = 4.17 \times 10^{-32} \text{ kg}$$

(3) When $\lambda = 1.20 \times 10^{-12}$ m, E, p and m are equal to

$$E = 1.03 \times 10^{6} \text{ eV}, \quad p = 5.52 \times 10^{-22} \text{kg} \cdot \text{m/s}, \quad m = 1.84 \times 10^{-30} \text{ kg}$$

Example 17-3 The X-rays of wavelength 2.2×10^{-11} m are scattered from a carbon target and the scattered radiation is viewed at 85° to the incident beam.

(1) What is the Compton shift?

(2) What percentage of its initial energy does an incident X-ray photon lose?

Solution (1) From Eq. (17-28), we obtain

$$\Delta\lambda = \frac{2h}{m_0 c} \sin^2 \frac{\varphi}{2} = 2.2 \times 10^{-12} \text{ m}$$

(2) The fractional energy loss r is

$$r = \frac{h\nu - h\nu'}{h\nu} = \frac{\lambda' - \lambda}{\lambda'} = \frac{\Delta\lambda}{\lambda + \Delta\lambda} = 9.1\%$$

Questions

17-1 Does a blackbody always appear black? Explain the term blackbody.

17-2 Pockets formed by coals in a coal fire seem brighter than the coals themselves. Is the temperature in such pockets appreciably higher than the surface temperature of an exposed glowing coal?

17-3 In the photoelectric experiments, the current (number of electrons emitted per unit time) is proportional to the intensity of light. Can this result alone be used to distinguish between the classical and quantum theories?

17-4 The existence of a cutoff frequency in the photoelectric effect is often regarded as the most potent objection to a wave theory. Explain.

17-5 Does a photon of energy E has mass? If so, evaluate it.

17-6 Why, in Compton scattering, would you expect $\Delta\lambda$ to be independent of the materials of which the scatterer is composed?

17-7 Do you observe a Compton effect with visible light? Why?

Problems

17-1 At what wavelength does a cavity at 6000K radiate most per unit wavelength?

17-2 (1) Assuming the surface temperature of Sun to be 5700K, use Stefan's law, to determine the rest mass lost per second to radiation by the Sun. Take the Sun's diameter to be 1.4×10^{9} m.

(2) What fraction of the Sun's rest mass is lost each year from electromagnetic radiation? Take the Sun's mass to be 2.0×10^{30} kg.

17-3 At what wavelength does the human body emits its maximum temperature radiation? List assumption you make in arriving at an answer.

17-4 (1) The energy needed to remove an electron from metallic sodium is 2.28 eV. Does sodium show a photoelectric effect for red light, with $\lambda=680$ nm?

(2) What is the cutoff wavelength for photoelectric emission from sodium and to what color does this wavelength correspond?

17-5 Light of wavelength 200 nm falls on an aluminum surface. In aluminum 4.2 eV are required to remove an electron. What is the kinetic energy of (1) the fastest and (2) the lowest emitted photoelectron? (3) What is the stopping potential? (4) What is the cutoff wavelength for aluminum?

17-6 The stopping potential for photoelectrons emitted from a surface illuminated by light of wavelength 491 nm is 0.71 V. When the incident wavelength is changed to a new value, the stopping potential is found to be 1.43 V.

(1) What is this new wavelength?

(2) What is the work function for the surface?

17-7 (1) Show that a free electron cannot absorb a photon and conserve both energy and momentum in the process. Hence, the photoelectric process requires a bound electron.

(2) In Compton effect, the electron can be free. Explain.

17-8 Solar radiation falls on the earth at a rate 1.94 cal/ cm^2 min on a surface normal to the incoming rays. Assuming an average wavelength of 5500 Å, how many photons per cm^2 min is this?

17-9 The wavelength of incident light is 0.05 Å in the Compton experiment. Find the wavelength of the photons which are scattered at $\varphi = 30^\circ, 90^\circ, 180^\circ$.

17-10 Photons of wavelength 2.4 Å are incident on a target containing free electron.

(1) Find the wavelength of a photon that is scattered at 30° from the incident direction.

(2) Do the same for a scattering angle of 120°?

17-11 What fractional increase in wavelength leads to a 75% loss of photon energy in a Compton collision with a free electron?

17-12 Find the maximum wavelength shift for a Compton collision between a photon and a free proton.

Chapter 18

Quantum Theory of the Atoms

By the early years of last century, the atomic nature of matter had been well established. It was known that atoms contained electrons and that an electron was much lighter than even the lightest atom (hydrogen). It had also been shown that electrons could be removed from atoms of a given species, producing positively charged ions, but that only a finite number of electrons could be obtained from atom. This number is characteristic of the atoms of each element and is called the atomic number Z. In the normal state an atom is electrically neutral so that it must contain positive charge of an amount Ze, where $-e$ is the charge on the electron. The question now arises as to how the mass and positive charge are distributed with the atom and how the electrons move in the atom. In this chapter, we will briefly discuss these questions.

18. 1 Rutherford's Experiment and the Nuclear Atom

18. 1. 1 Thomson's model

Many models of the atom were proposed in history, among which Thomson's model was of important. J. J. Thomson gave a tentative description, or model, of an atom according to which the negatively charged electrons were located within a continuous distribution of positive charge. The positive charge distribution was assumed to be spherical in shape with a radius of the known order of magnitude of the radius of an atom, 10^{-10} m. Owing to their mutual repulsions, the electrons would be uniformly distributed through the sphere of positive charge.

Fig. 18-1 illustrates this "plum pudding" model of the atom. In an atom in its lowest possible energy state, electron would be fixed at their equilibrium positions. In excited atom (that is, atoms in a material at high temperature), the electrons would vibrate about their equilibrium positions. Since classical electromagnetic theory predicts that an accelerated charged body, such as a vibrating electron, emits electromagnetic radiation, it was possible to understand qualitatively the emission of such radiation by excited atoms on the basis of Thomson's model. Quantitative agreement with experimentally observed spectra was lacking however.

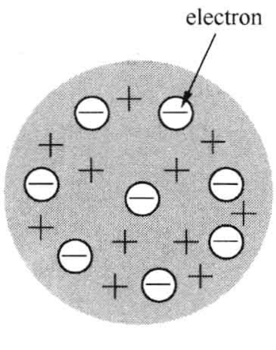


Fig. 18-1 The "plum pudding" model of atom

Conclusive proof of the inadequacy of Thomson's model was obtained by Ernest Rutherford in 1911, a former student of Thomson's, from the analysis of experiments on the scattering of α particles by atoms.

18.1.2 Rutherford's experiment

Fig. 18-2 shows a typical arrangement that Rutherford and his colleagues used to study the scattering of α particles on passing through thin foils of various substances. The α particles are the doubly ionized helium. The radioactive source emits α particles that are collimated into a narrow parallel beam by a pair of diaphragms. The parallel beam is incident upon a foil of some substance, usually a metal. The foil is so thin that the particles pass completely through with only a small decrease in speed. In traversing the foil, however, each α particle experiences many small deflections due to the Coulomb force acting between its charge and the positive and negative charge of the atoms of the foil. Since the deflection of an α particle in passing through a single atom depend on the details of its trajectory through the atom, the net deflection in passing through the entire foil will be different for different α particles in the beam. As a result, the beam emerges from the foil not as a parallel beam but as a divergent beam.

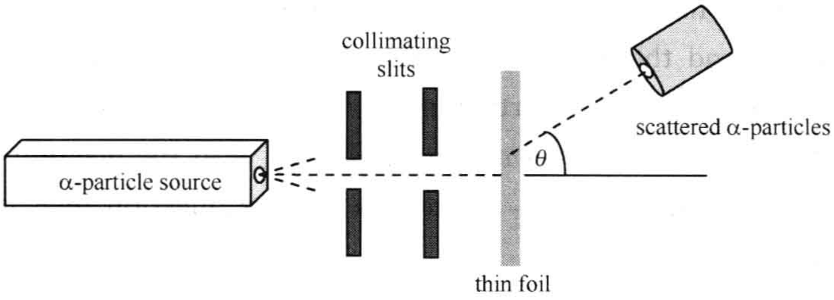


Fig. 18-2 Arrangement of an α-particle scattering experiment. The region traversed by the particle is evacuated

The results of Rutherford's experiment showed that most of the α particles are deflected through very small angles ($<10°$), but some are deflected through large angles, about 1 in 8000 being deflected through angles greater than 90°, up to 180°. This fact is very important for Rutherford to propose the nuclear atom. In explaining the experimental data, we can neglect the influence of the electrons with the atom and assume that the scattering is due to the positive charge alone. This is because the α-particles are about 8000 times as heavy as an electron. It can be shown that if the positive charge is spread uniformly throughout the atom, as in the Thomson's model of the atom, there would be no appreciable scattering at large angles, and even at 1° the number of scattered α particles would be negligible. To scientists accustomed to thinking in terms of Thomson's model, it came as a great surprise that some α particles were deflected through very large angles.

18.1.3 Rutherford's model—the nuclear atom

In order to explain the experimental data of the α particles scattering, Rutherford proposed a model in 1911. In his model, all the positive charge of the atom, and consequently essentially all its mass, are assumed to be concentrated in a small region in the center of the atom while the electrons were distributed over the remaining volume. The small core is called the nucleus. Surrounding this core are the electrons, held to the nucleus by electrostatic attraction. The radius of even the largest nucleus is only about 10^{-14} m, while the radius of a typical atom is nearer 10^{-10} m. Thus all the positive charge and almost all the mass of the atom are concentrated in about 10^{-12} the volume of the atom.

Rutherford made a detailed calculation of the angular distribution to be expected for the scattering of α particles from atoms of the type proposed in his model. Because the dimensions of the nucleus are very small, an α particle passing very near it can be scattered by a strong coulomb repulsion through a large angle in the traversal of a single atom. When Rutherford's theoretical values were compared with the experimental ones, they were found to be in good agreement.

18.1.4 Limitation of the Rutherford's model

In the Rutherford's model, the electrons move in the Coulomb field of the nucleus in orbits, like a planetary system. According to the laws of classical physics, a particle moving on a curved trajectory is accelerating and an accelerating charged particle radiates electromagnetic waves and loses energy. This implies that in a time of order 10^{-10} s all the energy of the atom would be radiated away and the electrons would collapse into the nucleus so that our Earth would become the size of an apple. Further more, the continuous spectrum of the radiation that would be emitted in this process is not in agreement with the discrete spectrum which is known to be emitted by atoms. Hence, there is a serious question in Rutherford's model. That is the stability of such an atom.

However, the experiments had shown that the atom consists of the nucleus and the electrons. The next step is to find the new laws applied to electrons of an atom and to phenomena on the atomic scale.

18.2 Atomic Spectra and the Bohr Model of Hydrogen

18.2.1 Atomic spectra

The experiments show that each chemical element emits its own unique spectrum when its electrons are excited in electric discharges or in flames, called emission lines, and that atoms can only absorb light at certain discrete frequencies when radiation falls on the material made up of these atoms, called absorption lines. G. R. Kirchoff was the first to show that the emission frequencies coincide with the absorption frequencies for a given atom. The fact that each element has its own characteristic line spectrum is of the greatest practical importance, because it makes spectroscopy a very useful addition to the usual techniques of chemical analysis. Chiefly for this reason much effort was devoted to the accurate measurement of atomic spectra, and in fact, much effort was needed because the spectra consist of many hundreds of lines and in general are very complicated.

However, the spectrum of hydrogen is relatively simple. This is perhaps not surprising since hydrogen, which contains just one electron, is itself the simplest atom. Most of the universe consists of isolated hydrogen atoms so that the hydrogen spectrum is of considerable practical interest. Fig. 18-3 shows that The Balmer series of the hydrogen atom spectrum that falls approximately within the wavelength range of visible light. We see that the spacing, in wavelengths, between adjacent lines of the spectrum continuously decreases with decreasing wavelength of the lines, so that the series of lines converges to the so-called series limit at 3645.6 Å(Å$= 10^{-10}$ m).

The most important discovery in the reach for regularities in the line spectra of atoms was made in 1885 by J. Balmer, who showed that the frequencies of a series of lines in the visible part of the spectrum of atomic hydrogen were among those given by the empirical formula

$$\widetilde{\nu}_{ab} = \frac{1}{\lambda} = R\left(\frac{1}{n_a^2} - \frac{1}{n_b^2}\right), \quad n_a = 2, \; n_b = 3, 4, 5, \cdots \tag{18-1}$$

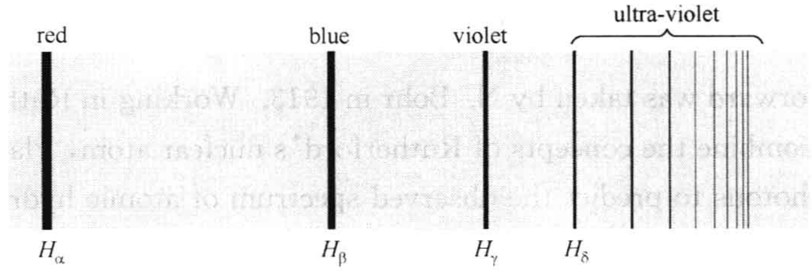


Fig. 18-3 Schematic Diagram of the visible part of the hydrogen spectrum

where $\widetilde{\nu}_{ab}$ is the wave number of either an emission or absorption line, n_a and n_b are positive integers with $n_b > n_a$, and R is a constant, known as Rydberg's constant for hydrogen, equal to

$$R = 10967757.6 \pm 1.2\text{m}^{-1} \tag{18-2}$$

It was subsequently discovered that Balmer's formula (18-2) is not only applicable to visible region, but in fact describes the complete spectrum of atomic hydrogen. For instance, we now know of the existence of five series of line in the hydrogen spectrum, as shown in Table 18-1. The Balmer was denoted by $H_\alpha, H_\beta, H_\gamma, \cdots$, respectively for $n = 3, 4, 5, \cdots$. The Balmer H_α line ($n=3$) was first discovered by J. Von Fraunhofer in the solar spectrum and was called by him the C line.

The set of quantities R/n^2, are called terms. J. R. Rydberg showed that for other atoms, particularly the alkalis, the wave number of lines could be represented approximately as difference of terms T_n, where

Table 18-1 The Hydrogen Series

Names	Wavelength Ranges	Formula
Lyman	Ultraviolet	$\widetilde{\nu} = R\left(\frac{1}{1^2} - \frac{1}{n^2}\right), \quad n = 2,3,\cdots$
Balmer	Near ultraviolet and visible	$\widetilde{\nu} = R\left(\frac{1}{2^2} - \frac{1}{n^2}\right), \quad n = 3,4,\cdots$
Paschen	Infrared	$\widetilde{\nu} = R\left(\frac{1}{3^2} - \frac{1}{n^2}\right), \quad n = 4,5,\cdots$
Brackeff	Infrared	$\widetilde{\nu} = R\left(\frac{1}{4^2} - \frac{1}{n^2}\right), \quad n = 5,6,\cdots$
Pfund	Infrared	$\widetilde{\nu} = R\left(\frac{1}{5^2} - \frac{1}{n^2}\right), \quad n = 6,7,\cdots$

$$T_n = \frac{R}{(n+\alpha)^2}, \quad n = 1, 2, 3, \cdots \tag{18-3}$$

and where α was a constant for each particular series. Subsequently it was discovered by W. Ritz that the wave number of all lines can be represented by the difference terms, even if the terms cannot be represented by a simple formula such as Eq. (18-3). That is

$$\widetilde{\nu}_{n_1 n_2} = T_{n_1} - T_{n_2} \tag{18-4}$$

and

$$\begin{aligned}\widetilde{\nu}_{n_1 n_3} &= (T_{n_1} - T_{n_2}) + (T_{n_2} - T_{n_3}) \\ &= \widetilde{\nu}_{n_1 n_2} + \widetilde{\nu}_{n_2 n_3}, \quad n_3 > n_2 > n_1\end{aligned} \tag{18-5}$$

known as the Ritz combination principle. However, not all combinations of lines are observed because certain selection rules operate which will be discussed in the following section.

18.2.2 Bohr's model of the hydrogen atom

A major step forward was taken by N. Bohr in 1913. Working in Rutherford's Laboratory, he was able to combine the concepts of Rutherford's nuclear atom, Planck's energy quanta and Einstein's photons to predict the observed spectrum of atomic hydrogen.

Bohr postulated:

(1) An electron in an atom moves in a circular orbit about the nucleus under the influence of the Coulomb attraction between the electron and the nucleus, obeying the laws of classical mechanics.

(2) Instead of the infinity of orbits which would be possible in classical mechanics, it is only possible for an electron to move in an orbit for which its orbital angular momentum L is an integral multiple of $\hbar$, Planck's constant h divided by 2π

$$L = n\hbar \tag{18-6}$$

(3) Despite the fact that it is constantly accelerating, an electron moving in such an allowed orbit does not radiate electromagnetic energy. Thus, its total energy E remains constant. Electromagnetic radiation is emitted if an electron, initially moving in an orbit of total energy E_i, discontinuously changes its motion so that it moves in an orbit of total energy E_f. The frequency of the emitted radiation ν is equal to the quantity $(E_i - E_f)$ divided by Planck's constant h. That is

$$\nu = \frac{E_i - E_f}{h} \tag{18-7}$$

The first postulate bases Bohr model on the existence of the atomic nucleus. The second postulate introduces quantization to the world of atoms. The third postulate removes the problem of the stability of an electron moving in a circular orbit, due to the emission of the electromagnetic radiation of the electron required by the classical theory. By simply postulating that this particular feature of the classical theory is not valid for the case of an atomic electron, Eq. (18-7) is really just Einstein's postulate that

$$h\nu = E_i - E_f \tag{18-8}$$

from the conservation of the energy.

18.2.3 The results of Bohr's model

We shall make the approximation that the nucleus is infinitely heavy compared with the bound electron and is therefore at rest. The electron will be taken to be moving in a circular orbit of radius r, in which case the Coulomb attractive force acting on the electron, due to its electrostatic interaction with the nucleus of charge Ze, can be equated with the electron m times the centripetal acceleration

$$\frac{Ze^2}{4\pi\varepsilon_0 r^2} = \frac{mv^2}{r} \tag{18-9}$$

where v is the velocity of the electron. A second equation is obtained from Bohr's second postulate that the orbital angular momentum is quantized

$$L = mvr = n\hbar, \quad n = 1, 2, 3, \cdots \tag{18-10}$$

From Eq. (18-9) and Eq. (18-10), we obtain the possible values of v and r

$$v_n = \frac{Ze^2}{4\pi\varepsilon_0 n\hbar} \tag{18-11}$$

$$r_n = \frac{4\pi\varepsilon_0 \hbar^2 n^2}{Ze^2 m} \tag{18-12}$$

Eq. (18-11) and Eq. (18-12) show that the radius and the velocity of an electron cannot take the arbitrary values. Note that these radii are proportional to the square of the integer n. If we evaluate the radius of the smallest orbit ($n=1$) for hydrogen atom ($Z=1$) by inserting the known values of h, m and e, we obtain $r_1 = 5.3\times10^{-11}\,\text{m} = 0.53\,\text{Å}$, which is called the Bohr's radius. Evaluating the orbital velocity of an electron in the smallest orbit of a hydrogen atom from Eq. (18-11), we find $v_1 \approx 2.2\times10^6\,\text{m/s}$. It is apparent from Eq. (18-11) that this is the largest velocity possible for a hydrogen atom electron. The fact that this velocity is less than 1% of the velocity of light is the justification for using classical mechanics instead of relativistic mechanics in the Bohr model. Eq. (18-11) and Eq. (18-12) can be rewritten as

$$\begin{cases} v_n = \dfrac{v_1}{n} \\ r_n = n^2 r_1 \end{cases} \tag{18-13}$$

for $Z=1$.

The kinetic energy of the electron E_k, is then found to

$$E_k = \frac{1}{2}mv^2 = \frac{m}{2\hbar^2}\left(\frac{Ze^2}{4\pi\varepsilon_0}\right)^2 \frac{1}{n^2} \tag{18-14}$$

and the potential energy E_p is, correspondingly

$$E_p = -\frac{Ze^2}{4\pi\varepsilon_0 r} = -\frac{m}{\hbar^2}\left(\frac{Ze^2}{4\pi\varepsilon_0}\right)^2 \frac{1}{n^2} \tag{18-15}$$

from which the total energy E_n of the system is equal to

$$E_n = E_k + E_p = -\frac{m}{2\hbar^2}\left(\frac{Ze^2}{4\pi\varepsilon_0}\right)^2 \frac{1}{n^2} \tag{18-16}$$

Taking $Z=1$ for hydrogen atom, we have the energy of hydrogen atom

$$E_n = -\frac{m}{2\hbar^2}\left(\frac{e^2}{4\pi\varepsilon_0}\right)^2\frac{1}{n^2} \tag{18-17}$$

We see that the quantization of the orbital angular momentum of the electron leads to a quantization of its total energy. Eq. (18-17) can be rewritten as

$$E_n = -\frac{I_p}{n^2}, \quad n = 1,2,3,\cdots \tag{18-18}$$

where

$$I_p = \frac{m}{2\hbar^2}\left(\frac{e^2}{4\pi\varepsilon_0}\right)^2 = 13.6\ \text{eV} \tag{18-19}$$

which is the ionization energy of the hydrogen atom for $n=1$. The information contained in Eq. (18-18) is presented as an energy-level diagram in Fig. 18-4. The positive integer n is called the quantum number. The energy of each level is shown on the left in electron volts and the quantum number of the level is shown on the right. The diagram is so constructed that the distance from any level to the level of zero energy is proportional to the energy of that level. The negative sign in Eq. (18-18) tells us that the hydrogen atom states whose energies are given by this equation are bound states. As n increases, the total energy of the quantum state becomes less negative, with E_n approaching zero as n approaches infinity.

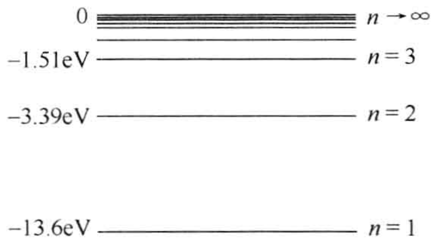


Fig. 18-4 The energy-level diagram

The level with the lowest energy ($n=1$) is known as the ground state. It corresponds, according to Eq. (18-12), to the smallest orbit. The level with $n=2$ is the first excited state, and the level with $n=3$ is the second excited state and so on. The highest energy $E = 0$ results when $n \to \infty$. That is when the electron and proton are separated by an infinite distance. Since the state of lowest total energy is, of course, the most stable state for the electron, we see that the normal state of the electron in a one-electron atom is the ground state for which $n=1$. The ionization energy of the hydrogen atom, that is, the least amount of energy needed to remove an electron from the $n=1$ orbit to infinity, is equal to $I_p = 13.6$ eV. In an electric discharge, or in some other process, the atom receives energy due to collision, etc; this means that the electron must make a transition to a state of higher energy, or excited state, in which $n>1$.

Obeying the common tendency of all physical system, the atom will emit its excess energy and return to the ground state. This is accomplished by a series of transitions in which the electron drops to excited states of successively lower energy, finally reaching the ground state. When the hydrogen atom is excited into a state n_i and then drops to a state n_f, the frequency of light emitted in this transition is given by

$$\nu_{fi} = \frac{m}{4\pi\hbar^3}\left(\frac{e^2}{4\pi\varepsilon_0}\right)^2\left(\frac{1}{{n_f}^2} - \frac{1}{{n_i}^2}\right) \tag{18-20}$$

in agreement with Eq. (18-2) provided R is taken to be

$$R_\infty = \frac{m}{4\pi\hbar^3 c}\left(\frac{e^2}{4\pi\varepsilon_0}\right)^2 \tag{18-21}$$

Here we have written R_∞ to recall that we are using the infinite nuclear mass approximation. Considering the finite mass M of nucleus, Eq. (18-21) becomes as

$$R_M = \frac{M}{m+M}R_\infty \tag{18-22}$$

If we evaluate R_H from Eq. (18-22), using the currently accepted values of the quantities m, M, e, ε_0 and $\hbar$, we find

$$R_H = 10\ 968\ 100\ \mathrm{m}^{-1}$$

Comparing this with the experimental values of R_H, we see that the Bohr's model agrees with the spectroscopic data to within three parts in 100 000! The last recommended value of R_H can be found in Appendix 1.

According to Bohr's model, each of the five known series of the hydrogen spectrum arises from a subset of transitions in which the electron goes to a certain final quantum state n_f. For the Lyman series $n_f=1$; for the Balmer series $n_f = 2$; for the Paschen series $n_f = 3$; for the Brackett series $n_f = 4$; and for the Pfund series $n_f = 5$. The first three of these series are conveniently illustrated in terms of the energy-level diagram of Fig. 18-5. The wavelengths of the lines of all these series are fitted very accurately by Eq. (18-20) by using the appropriate value of n_f. This was a great triumph for Bohr's model. The success of the model was particularly impressive because the Lyman, Brackett, and Pfund series had not been discovered at the time the model was developed by Bohr. The existence of these series was predicted, and the series were soon found experimentally by the persons after whom they are named.

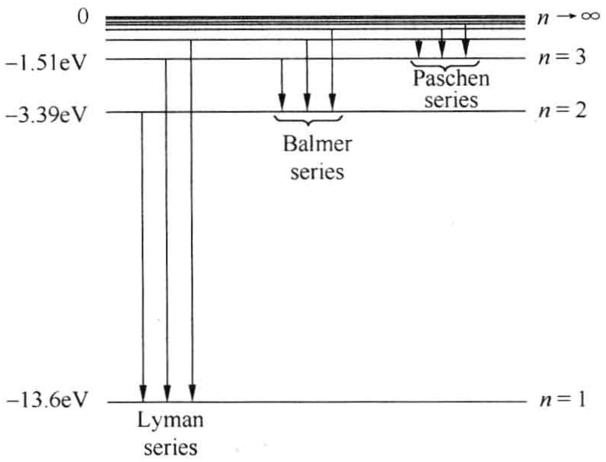


Fig. 18-5 The energy-level diagram and the corresponding spectral lines for hydrogen

The model worked equally well when applied to the case of one-electron atoms with $Z=2, 3, \cdots$, for example, singly ionized helium atoms He^+. The properties of the absorption spectrum of one-electron atom are also easy to understand in terms of the Bohr's model. Since the atomic electron must have a total energy exactly equal to the energy of one of the allowed energy states, the atom can only absorb discrete amounts of energy from the incident electromagnetic radiation. The process of absorbing electromagnetic radiation is then just the inverse of the normal emission process, and the lines of the absorption spectrum will have exactly the same wavelength as the lines of the emission spectrum.

The Bohr's model predicts that the total energy of an atomic electron is quantized. This fact was confirmed by a simple experiment performed by J. Franck and G. Hertz.

18.2.4 The limitation of Bohr's model

The Bohr's theory can completely explain the regulation of the hydrogen and the hydrogen-like ions and shows that the total energies of the atoms are quantized. Its success was certainly very striking. Bohr quite properly pushed his theory for all it was worth, taking as his goal the development of a theoretical basic for the periodic table of the elements. Unfortunately, his model only worked for hydrogen and hydrogen-like ions such as He^+, but his ideas were so revolutionary that they broke up the classical view of electrons' behavior and paved the way for fresh new ideas in what would become quantum physics and quantum mechanics.

According to Bohr's model, an electron bound to a proton to form a hydrogen atom can exist only in certain orbits characterized by a total energy given by Eq. (18-16). The hypothesis that only circular orbits are allowed, is inexplicable and arbitrary. The concept of orbit is a classical concept and the electron in orbit is obeying the laws of classical mechanics but not satisfying the classical electromagnetic theory. This is incomprehensive. The Bohr's theory cannot calculate the rate of transitions between the different energy levels when radiation is emitted or absorbed, and had not the ability to handle unbound systems.

Another problem with the Bohr's model was its ad hoc nature. The angular momentum quantization was not the consequence of a more general theory or the result of some fundamental insight. Hence we need more general theory—quantum mechanics.

18. 3 De Broglie's Postulate and Matter Waves

18. 3. 1 De Broglie's postulate

M. De Broglie was a French experimental physicist who, from the outset, had supported-Compton's view of the particle nature of radiates. His experiments and discussions impressed his brother Louis so much with the philosophic problems of physics at the time that Louis changed his career from history to physics. In his doctoral thesis, presented in 1924 to the faculty of Science at the University of Paris, Louis De Broglie proposed the existence of matter waves. The thoroughness and originality of his thesis were recognized at once but, because of the apparent lack of experimental evidence, De Broglie's ideas were not considered to have any physical reality. It was Albert Einstein who recognized their importance and validity and in turn called them to the attention of other physicists. Five years later De Broglie won the Nobel Prize in physics, his ideas having been dramatically confirmed by experiment.

In 1924, L. De Broglie made a great speculative hypothesis that just as radiation has particle-like properties, electron and other material particles possess wave-like properties. For free material particles, L. De Broglie assumed that the associated wave also has a frequency ν and a wavelength λ, related to the energy E and the momentum p of the particle by

$$\nu = \frac{E}{h} \tag{18-23}$$

and

$$\lambda = \frac{h}{p} \tag{18-24}$$

In particular, for a particle of mass m moving at a non-relativistic speed v, one has $p = mv$ so that

$$\lambda = \frac{h}{mv} \tag{18-25}$$

A wavelength calculated from Eq. (18-24) or Eq. (18-25) is called a De Broglie wavelength. Look again at Eq. (18-24) and note the central role played by the Planck constant h in connecting the wave and the particle aspects of both light and matter.

Example 18-1 What is the De Broglie wavelength of a baseball moving at a speed $v = 10$ m/s?

Solution Assume the mass of a baseball is 1.0 kg. From Eq. (18-25), we have

$$\lambda = \frac{h}{mv} = \frac{6.6 \times 10^{-34}}{1.0 \times 10} = 6.6 \times 10^{-35}\,(\text{m}) = 6.6 \times 10^{-25}\,(\text{Å})$$

which is too small to be measured by the experiments.

Example 18-2 What is the De Broglie wavelength of an electron whose kinetic energy is 100 eV?

Solution According to Eq. (18-25), the De Broglie wavelength of the electron is

$$\lambda = \frac{h}{p} = \frac{h}{\sqrt{2mE_\text{k}}}$$

$$= \frac{6.6 \times 10^{-34}}{\sqrt{2 \times 9.1 \times 10^{-31} \times 100 \times 1.6 \times 10^{-19}}} = 1.2 \times 10^{-10}\,(\text{m})$$

which is in the range of X-rays.

The above examples imply that for macroscopic bodies De Broglie wavelength is so tiny that we have no practical means of measuring it, but for electrons and other atomic particles, wavelength of measurable magnitude are available. Measurements on such particles confirm the De Broglie Hypothesis.

18.3.2 Testing De Broglie hypothesis

When waves are scattered or pass through slits, which have dimensions comparable to their wavelength, interference and diffraction effects are observed. Now, as seen from Eq. (18-25), the De Broglie wavelengths associated with electrons of energy 1 eV, 10 eV and 100 eV, are, respectively, 12 Å, 3.9 Å and 1.2 Å. Thus, in macroscopic situations, the De Broglie electron wavelengths are exceedingly small compared with the dimensions of any obstacles or slits in the apparatus and no interference or diffraction effects can be observed. However, the spacing of atoms in a crystal lattice is of the order of a few angstroms (Å) and the wave nature of matter might be tested in the same way that the wave nature of X-rays was first tested, namely by allowing a beam of electrons of appropriate energy to fall on a crystalline solid.

1. Davisson-Germer experiment

Fig. 18-6 shows schematically the apparatus of Davisson and Germer experiment. Electrons from a heated filament F are accelerated through a potential difference V and emerge from the "electron gun" G with kinetic energy. This electron beam falls at normal incidence on a single crystal of nickel at C. The detector D is set at a particular angle θ and readings of the intensity of the scattered beam is taken at various values of the accelerating potential V.

Fig. 18-7, for example, shows that a strong scattered electron beam is detected at $\theta = 50°$

for $V = 54$ V. The existence of this peak at 50° can only be explained by constructive interference of the waves scattered by the regular crystal lattice. If the atoms in crystal are spaced at a distance d, then the Bragg condition for constructive interference at a scattering angle is

$$n\lambda = 2d\sin\varphi \tag{18-26}$$

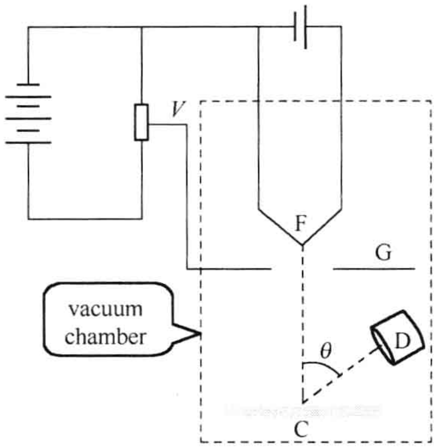


Fig. 18-6 Schematic diagram of Davisson-Germer experiment

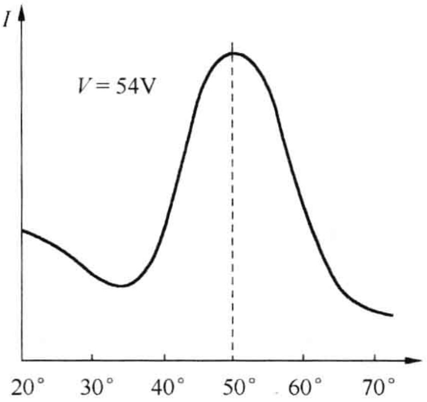


Fig. 18-7 The result of Davisson-Germer

where

$$\lambda = \frac{h}{\sqrt{2mE_k}} \tag{18-27}$$

and $\varphi = 90° - \theta/2$. For the condition of Fig. 18-8 the effective interplanar spacing d can be shown by X-ray scattering from the same crystal to be 0.91 Å. Since $\theta = 50°$, it follows that $\varphi = 65°$. The wavelength calculated from Eq. (18-26), assuming $n=1$, is $\lambda = 2 \times 0.91 \times \sin 65° = 1.65$ Å. The De Broglie wavelength for 54 eV electrons, calculated from the Eq. (18-27), is

$$\lambda = \frac{h}{\sqrt{2mE_k}} = \frac{6.6 \times 10^{-34}}{\sqrt{2 \times 9.1 \times 10^{-31} \times 54 \times 1.6 \times 10^{-19}}} = 1.65(\text{Å})$$

which agrees with the value of 1.65 Å with the experimental error. This impressive agreement gives quantitative confirmation of De Broglie's relation between λ and p.

2. Thomson Experiment

In 1927, G. P. Thomson showed the diffraction of electron beams passing through thin films and independently confirmed the De Broglie's relation (18-24) in detail. Fig. 18-9(a) shows schematically the apparatus of Thomson experiment, in which the incident electron beams pass through the thin metal film M made up of a large number of tiny, randomly oriented crystallites, and if a photographic plate P is placed at right angle to the incident beam, the central beam spot will be surround by diffraction rings. Thomson used higher-energy electrons, which are much more penetrating so that many hundred atomic planes contribute to the diffracted wave. The resulting diffraction rings have a

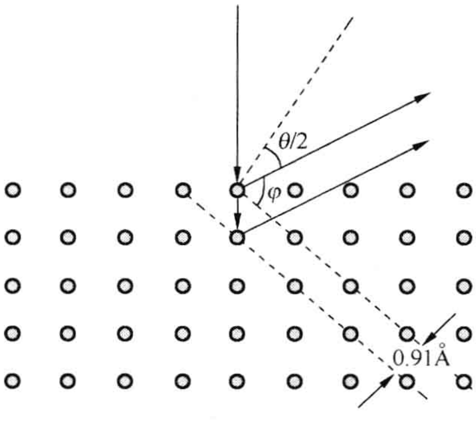
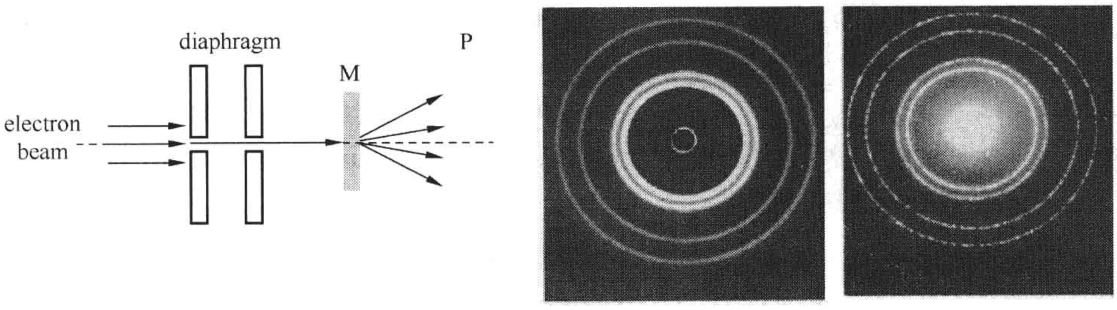


Fig. 18-8 The diffraction of the electrons by the crystal

sharp structure. For comparison, in Fig. 18-9(b) and (c) we show an X-ray diffraction pattern and an electron diffraction pattern from polycrystalline substances. The result of Thomson experiment showed that the electrons exhibit wave property as the X-ray.



(a) (b) (c)

Fig. 18-9 (a) Schematic diagram of Thomson experiment; (b) diffraction pattern of X-ray (c) diffraction pattern of electron beam

It is of interest that J. J. Thomson, who in 1897 discovered the electron (which he characterized as a particle with a definite charge-to-mass ratio) and was awarded the Nobel prize in 1906, was the father of G. P. Thomson, who in 1927 experimentally discovered electron diffraction and was awarded the Nobel prize (with Davisson) in 1937.

18.3.3 The matter waves

It has been proved that the microscopical particles such as electrons, atoms and neutrons have the properties of wave. The waves associated with the matter particles is called the matter waves, which are not simple waves, compared with the classical waves. The wave and particle models for the matter are complementary, and if a measurement proves the wave character of radiation or matter, then it is impossible to prove the particle character in the same measurement, and conversely. Which model we use is determined by the nature of the measurement. Furthermore, our understanding of radiation, or of matter, is incomplete unless we take into account measurements that reveal the wave aspects and also those that reveal the particle aspects. Hence, radiation and matter are not simply waves and not simply particles. The character of the matter is called the wave-particle duality. A more general and, to the classical mind, a more complicated model is needed to describe their behavior. The link between wave model and particle model is provided by a probability interpretation of the wave-particle duality. This will be discussed in the following sections.

Today the wave nature of matter is taken for granted and diffraction studies by beams of electrons or neutrons are used routinely to study the atomic structures of solids and liquids. The equipments to analyze the microstructure of materials using electron beams are electron microscopes which mainly include two kinds, scanning electron microscope and transmission electron microscope, briefly called as SEM and TEM. TEM is used to study the microstructure at much higher magnification. For example, Fig. 18-10 shows the TEM images of the microstructure from a Co-base alloy, taken at magnification of 30 000. Matter waves are a valuable supplement to X-rays in studying the atomic structures of solids. Electrons, for example, are less penetrating than X-rays and so are particularly useful for studying surface features.

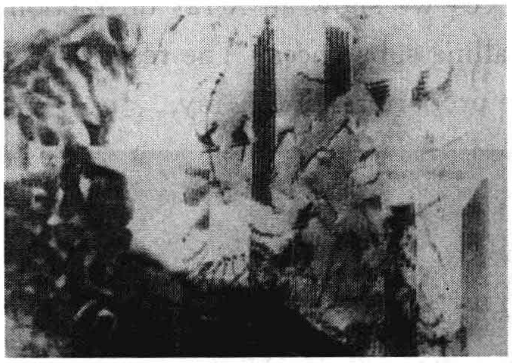


Fig. 18-10 TEM image of the microstructure in a Co base alloy (×30000)

18.4 The Uncertainty Principle

In classical mechanics the equations of motion of a system with given forces can be solved to give us the position and momentum of a particle at all values of the time. All we need to know are the precise position and momentum of the particle at some value of the time $t=0$ (the initial condition) and the future motion is determined exactly. This mechanics has been used with great success in the macroscopic world, for example in astronomy, to predict the subsequent motions of objects in terms of their initial motions. Hence, it was naturally assumed by classical physicists that in the realm of microscopic systems the position and momentum of an object, such as an electron, could be determined precisely by observations in a similar way. Heisenberg and Bohr questioned this assumption.

Can we determine by actual experiment at the same instant both the position and momentum of matter or radiation? The answer given by quantum theory is: not more accurately than is allowed by the Heisenberg uncertainty principle.

Here we make a measurement on the y coordinate of an electron, which is a member of a broad parallel electron beam moving in the x direction, by introducing the beam into a slit of narrow width Δy, as shown in Fig. 18-11. If the electrons with the momentum of p were classical particle, they would strike in a range of Δy near point O in the y direction at time t. If so, this would imply that at time $t-D/v_x=t-mD/p$ their y coordinates were 0 with an error of Δy. That is the position of the electron can be measured by observing the striking point y_P of electron on the screen S. However, the experiments have shown that the striking points are distributed on the screen S in the same way as in the single-slit diffraction experiment of light. If the y component of momentum of the electron, which has only the x component of momentum before passing the silt, is zero when it passes the slit, the electron will move along the x direction. The fact that the electrons have departures from the x direction after passing the slit means they have nonzero y component of momentum at the position of slit.

In propagating through the silt, the waves of electrons will be diffracted by the slit in the view of wave. The results of experiments show that if the width Δy decreases (we try to determine the y coordinate of the electron), the diffraction shall be more pronounced. The angle θ to the first minimum of the single-slit diffraction pattern is given by

$$\sin\theta=\frac{\lambda}{\Delta y} \qquad (18\text{-}28)$$

The wavelength λ of the electron moving with the momentum p in the x direction is equal to

$$\lambda=\frac{h}{p} \qquad (18\text{-}29)$$

Therefore, Eq. (18-28) becomes as

$$\sin\theta = \frac{h}{p\Delta y} \tag{18-30}$$

Assume the distance between the slit and photographic plate is D. The y coordinate of the first minimum of the diffraction pattern is given by

$$y = D \times \tan\theta \approx D \times \sin\theta = \frac{Dh}{p\Delta y} \tag{18-31}$$

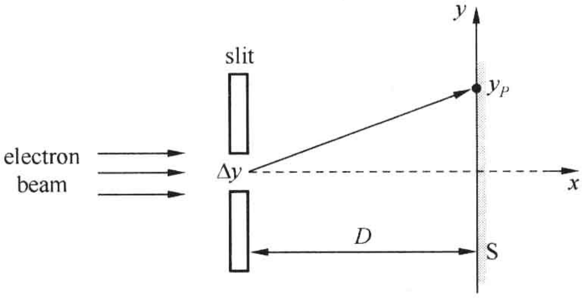


Fig. 18-11 Schematic diagram of single-slit diffraction experiment of a beam of electrons

The results of experiments implies in the view of particle that the electron, after passing through the slit where the measurement of its y position was made, has y component of momentum which can be anywhere within a range from about $-\Delta p_y$ to Δp_y. Hence, the $y_{\max}$ coordinate of the maximum position at which the electron arrives equals to

$$y_{\max} = v_{y\max} \times t = \frac{\Delta p_y}{m} \times \frac{D}{v_x} = \frac{D\Delta p_y}{p} \tag{18-32}$$

Because $y_{\max} \geqslant y$, we obtain

$$\Delta y \Delta p_y \geqslant h \tag{18-33}$$

Hence, we can not determine both the position and momentum of an electron at the same instant.

There are two parts to the Heisenberg uncertainty principle. The first has to do with the simultaneous measurement of position and momentum. It states that experiment cannot simultaneously determine the exact value of a component of momentum, p_x say, of a particle and also the exact value of its corresponding coordinate, x. Instead, our precision of measurement is inherently limited by the measurement process itself such that

$$\begin{cases} \Delta p_x \Delta x \geqslant \dfrac{\hbar}{2} \\ \Delta p_y \Delta y \geqslant \dfrac{\hbar}{2} \\ \Delta p_z \Delta z \geqslant \dfrac{\hbar}{2} \end{cases} \tag{18-34}$$

where $\hbar$ is a shorthand symbol for $h/2\pi$. That is, if you design an experiment to pin down the position of an electron as closely as possible (by making Δx smaller), you will find that you are not able to measure its momentum very well (Δp_x will get bigger). If you tinker with the experiment to improve the precision of your momentum measurement, the precision of your position measurement will deteriorate. There is nothing that you can do about it. It is important to realize that this principle has nothing to do with improvements in instrumentation leading to better simultaneous determinations of p_x and x. Rather the principle says that even with ideal instruments we can never in principle do better than $\Delta p_x \Delta x \geqslant \hbar/2$.

The second part of the uncertainty principle has to do with measurements of the energy E and the time t required for the measurements, as for example, the time interval Δt during which a photon of energy spread ΔE is emitted from an atom. In this case

$$\Delta E \cdot \Delta t \geqslant \frac{\hbar}{2} \tag{18-35}$$

Thus, if you try to measure the energy of a particle, allowing yourself a time interval Δt to do so, your energy measurement will be uncertain by an amount $\Delta E \left(\geqslant \frac{\hbar}{2\Delta t}\right)$. To improve the precision of your energy measurement, you must allow more time.

If h, or $\hbar$ in Eq. (18-34) and Eq. (18-35) were zero, there would be no basic limitation on our measurement at all, which is the classical view. Again it is the smallness of h that takes the principle out of the range of our ordinary experiences.

If we cannot determine x and p_x simultaneously, we cannot specify the initial conditions of motions exactly. Therefore, we cannot precisely determine the future behavior of a system. Instead of making deterministic predictions, we can only state the possible results of an observation, giving the relative probabilities of their occurrence.

Example 18-3 The speed of a bullet ($m=50$ g) is measured to be 300m/s, with an uncertainty of 0.01%. With what fundamental accuracy could we have located the position of each, if the position is measured simultaneously with the speed.

Solution The momentum of the bullet is

$$p = mv = 0.05 \times 300 = 15 \text{ (kg} \cdot \text{m/s)}$$

and

$$\Delta p = 0.0001 \times 15 = 1.5 \times 10^{-3} \text{(kg} \cdot \text{m/s)}$$

so that

$$\Delta x \geqslant \frac{h}{4\pi\Delta p} = 3 \times 10^{-32} \text{ m}$$

Example 18-4 In the Example 18-3, the bullet is substituted by an electron with the same speed. Find the Δx of the electron.

Solution For the electron

$$p = mv = 9.1 \times 10^{-31} \times 300 = 2.7 \times 10^{-28} \text{(kg} \cdot \text{m/s)}$$

and

$$\Delta p = 0.0001 \times 2.7 \times 10^{-28} = 2.7 \times 10^{-32} \text{(kg} \cdot \text{m/s)}$$

so that

$$\Delta x \geqslant \frac{h}{4\pi\Delta p} = 0.2 \text{ cm}$$

From the results of the above examples, we can see that for macroscopic objects such as bullets the uncertainty principle sets no practical limit to our measuring procedure, but for microscopic objects such as electrons there are practical limits.

18.5 The Wave Function, Schrödinger Equation, and Born's Interpretation of Wave Functions

18.5.1 The wave function

According to the De Broglie postulate, the microscopic particles have the wave-particle duality. This implies that the particles of microscopic systems move according to the laws of some form of wave motion, and not according to the Newtonian laws of motion obeyed by the particles of macroscopic systems. Thus a microscopic particle acts as if certain aspects of its behavior are governed by the behavior of an associated De Broglie wave, or wave function.

With the electromagnetic field we must associate a particle, the photon, whose energy E and magnitude of momentum $\boldsymbol{p}$ are related to the frequency ν and the wavelength λ of the electromagnetic radiation by

$$\begin{cases} E = h\nu \\ p = \dfrac{h}{\lambda} \end{cases}$$

The plane electromagnetic waves, or the free photons, are governed by the plane wave function. In the same way, let us consider a free particle of mass m, having a well defined momentum $\boldsymbol{p} = p_x\boldsymbol{i}$ directed along the positive x direction ($\Delta x \to \infty$) and a non-relativistic energy $E = p_x^2/2m$. We associate with this particle a plane wave

$$\Psi(x,t) = A\mathrm{e}^{\mathrm{i}(kx-\omega t)} \tag{18-36}$$

where

$$\begin{cases} E = \hbar\omega = h\nu \\ \omega = 2\pi\nu \end{cases} \tag{18-37}$$

and

$$\begin{cases} p = \dfrac{h}{\lambda} = \hbar k \\ k = \dfrac{2\pi}{\lambda} \end{cases} \tag{18-38}$$

Here k is called the wave number and A is a constant. ω is the angular frequency, which can be expressed as

$$\omega = \frac{\hbar k^2}{2m} \tag{18-39}$$

The function $\Psi(x, t)$ is known as a wave function.

This one-dimensional treatment is easily extended to three dimensions. To a free particle of mass m, having the well-defined momentum $\boldsymbol{p}$ and the energy $E = p^2/2m$, we now associate a plane wave

$$\Psi(r,t) = A\mathrm{e}^{\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)} \tag{18-40}$$

where $\boldsymbol{k}$, called the propagation vector, is related to the momentum $\boldsymbol{p}$ by

$$\boldsymbol{p} = \hbar\boldsymbol{k} \tag{18-41}$$

with

$$k = |\boldsymbol{k}| = \frac{|\boldsymbol{p}|}{\hbar} = \frac{2\pi}{\lambda} \tag{18-42}$$

The vector $\boldsymbol{r}$ is the position vector of the particle and the angular frequency ω is still given by $\omega = \dfrac{E}{\hbar} = \dfrac{\hbar k^2}{2m}$.

To a particle that is acted on by a force, the function $\Psi(r,t)$, describing the associated wave, is more complicated than the plane wave function. For this case, the linear momentum of the particle is of changing magnitude, and the wavelength of the associated wave is not well defined. Hence we shall need help to find these more complicated wave functions. The quantum mechanics, the physical theory of atomic world, will help us to do this.

18.5.2 The Schrödinger equation

To a particle of mass m, moving in a potential field, the wave function is given by

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi(\boldsymbol{r},t)=\left[-\frac{\hbar^2}{2m}\nabla^2+V\right]\Psi(\boldsymbol{r},t) \tag{18-43}$$

where

$$\nabla^2=\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2} \tag{18-44}$$

and V is the potential energy of the particle. Eq. (18-43) is called the time-dependent Schrödinger equation, which was discovered in 1926 by Erwin Schrödinger. It is this equation which tells us the form of the wave function $\Psi(\boldsymbol{r},t)$, if we tell it about the force acting on the associated particle by specifying the potential energy V corresponding to the force. In other words, the wave function is a solution to the Schrödinger equation for that potential energy.

Introducing the operator

$$H=-\frac{\hbar^2}{2m}\nabla^2+V \tag{18-45}$$

which is known as the hamiltonian operator, Eq. (18-43) may be rewritten as

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi(\boldsymbol{r},t)=H\Psi(\boldsymbol{r},t) \tag{18-46}$$

The above equation is the "Newtonian equation" of the atomic world and the major task of quantum mechanics (also called wave mechanics) is to find out the solutions of Eq. (18-46) when the potential energy V is given.

In fact, the Schrödinger Eq. (18-46) is a differential equation, and is linear in $\Psi(\boldsymbol{r},t)$. As a result, it satisfies the superposition principle, according to which a linear superposition of possible wave functions is also a possible wave function. That is, if $\Psi_1(\boldsymbol{r},t)$ and $\Psi_2(\boldsymbol{r},t)$ are two different solutions of the equation, then any arbitrary linear combination of these solutions

$$\Psi(\boldsymbol{r},t)=C_1\Psi_1(\boldsymbol{r},t)+C_2\Psi_2(\boldsymbol{r},t) \tag{18-47}$$

is also a solution. The fact that the superposition principle applies is directly related to the wave nature of matter, and in particular to the existence of interference effect for the De Broglie waves.

We also note that the Schrödinger Eq. (18-46) is of first order in the time derivative $\frac{\partial}{\partial t}$, so that once the initial value of the wave function Ψ is given at some time t_0, namely $\Psi(\boldsymbol{r},t_0)$, its values at all other times can be found.

When the potential energy does not depend on the time t, the Schrödinger Eq. (18-46) admits stationary state solutions of the form

$$\Psi(\boldsymbol{r},t)=\psi_E(\boldsymbol{r})\mathrm{e}^{-\frac{\mathrm{i}Et}{\hbar}} \tag{18-48}$$

where E is a constant and $\psi_E(\boldsymbol{r})$ satisfies

$$\left[-\frac{\hbar^2}{2m}\nabla^2+V(\boldsymbol{r})\right]\psi_E(\boldsymbol{r})=E\psi_E(\boldsymbol{r}) \tag{18-49}$$

or

$$H\psi_E(\boldsymbol{r})=E\psi_E(\boldsymbol{r}) \tag{18-50}$$

which is known as the time-independent Schrödinger equation.

18.5.3 Born's interpretation of wave function

In the last section, we had discussed the equation, developed by Erwin Schrödinger in 1926, which tells us the behavior of any wave function of interest. In the following we shall study the relation, developed by Max Born, which connects the behavior of the wave functions to the behavior of the associated particles.

From the Schrödinger Eq. (18-46), we see that the presence of an "i"(that is $\mathrm{i}=\sqrt{-1}$) in the equation implies that in general case (for any potential energy function) the wave functions which are its solutions will be complex. This is in contrast to a "wave function" of classical mechanics. We should not attempt to give to wave function a physical existence in the same sense that water waves have a physical existence. The reason is that a complex quantity cannot be measured by any actual physical instrument. These comments should not be taken to imply that the wave function has no physical interest.

The basic connection between the properties of the wave function $\Psi(\boldsymbol{r},t)$ and the behavior of the associated particle is given by Max Born in terms of the probability density $\rho(\boldsymbol{r},t)$

$$\rho(\boldsymbol{r},t)=\Psi^*(\boldsymbol{r},t)\Psi(\boldsymbol{r},t) \tag{18-51}$$

where the symbol $\Psi^*(\boldsymbol{r},t)$ represents the complex conjugate of $\Psi(\boldsymbol{r},t)$. In 1926, M. Born made a fundamental postulate that if a particle is described by a wave function $\Psi(\boldsymbol{r},t)$, the probability of finding the particle within the volume element $\mathrm{d}V=\mathrm{d}x\mathrm{d}y\mathrm{d}z$ about the point $\boldsymbol{r}$ at the time t is

$$\rho(\boldsymbol{r},t)\mathrm{d}V=\Psi^*(\boldsymbol{r},t)\Psi(\boldsymbol{r},t)\mathrm{d}V=|\Psi(\boldsymbol{r},t)|^2\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{18-52}$$

Here $\rho(\boldsymbol{r},t)$, called the probability density, is the probability of finding the particle per unit volume about the point $\boldsymbol{r}$ at the time t. Since $\rho(\boldsymbol{r},t)$ is unique everywhere, one must require that the wave function $\Psi(\boldsymbol{r},t)$ should be "single-valued, continuous and finite".

Since $\rho(\boldsymbol{r},t)$ is the probability density, it must be unique everywhere. Therefore we must require that

The wave function $\Psi(r,t)$ should be continuous and single-valued and finite.

Since the probability of finding the particles somewhere must be unity, we deduce from Eq. (18-52) that the wave function $\Psi(\boldsymbol{r},t)$ should be normalized so that

$$\int\Psi^*(\boldsymbol{r},t)\Psi(\boldsymbol{r},t)\mathrm{d}V=1 \tag{18-53}$$

where the integral extends over all space. Eq. (18-53) expresses the simple fact that the probability of finding the particle anywhere in space at all is unity. However, not every wave function can be normalized like Eq. (8-53), for example the plane wave. In this case,

$|\Psi(\boldsymbol{r},t)|^2 \mathrm{d}V$ can be interpreted as the relative probability of finding the particle at time t in a volume element dV centered about $\boldsymbol{r}$, so that the ratio $|\Psi(\boldsymbol{r}_1,t)|^2/|\Psi(\boldsymbol{r}_2,t)|^2$ gives the probability of finding the particle within a volume element centered about $\boldsymbol{r}_1$, compared with that of finding it within the same volume element at $\boldsymbol{r}_2$.

From the Schrödinger Eq. (18-46), when the potential energy is real, we can obtain

$$\frac{\partial}{\partial t}\int_V \rho(\boldsymbol{r},t)\mathrm{d}V = -\int_S \boldsymbol{j}\cdot \mathrm{d}\boldsymbol{s} \tag{18-54}$$

where S is the surface bounding the volume V and $\boldsymbol{j}$ is defined as

$$\boldsymbol{j} = \frac{\hbar}{2mi}[\Psi^* \nabla \Psi - \nabla \Psi^* \Psi] \tag{18-55}$$

Eq. (18-54) may be also written as

$$\nabla \cdot \boldsymbol{j} + \frac{\partial \rho}{\partial t} = 0 \tag{18-56}$$

which implies the probability conservation, that is if the probability of finding the particle in some bounded region of space decreases as time goes on, then the probability of finding it outside of this region must increase by the same amount. Correspondingly the vector $\boldsymbol{j}$ defined by Eq. (18-55) can be interpreted as probability current density.

According to the probabilistic interpretation of wave function, the quantum theory cannot tell us that a particle in a given energy state will be found in a precise location at a certain time, but only the relative probability that the particle will be found in various location at that time. Hence, this interpretation is statistical, and so do the predictions of quantum mechanics.

18.6 The Infinite Potential Well

Of all Schrödinger equations the one for a constant potential is mathematically the simplest. For simplicity and as an example, consider a particle of m moving in a one-dimensional potential well that is zero over the interval $0 < x < a$ and infinite elsewhere as shown in Fig. 18-12. It has the feature that it will bind a particle with any finite total energy $E \geqslant 0$.

The infinite potential well, given by

$$V(x) = \begin{cases} 0, & 0 < x < a \\ \infty, & x \leqslant 0 \text{ or } x \geqslant a \end{cases} \tag{18-57}$$

does not depend on the time t. Hence, the motion of the particle is described by the wave function

$$\Psi(x,t) = \psi_E(x)\mathrm{e}^{-\frac{\mathrm{i}Et}{\hbar}} \tag{18-58}$$

where the $\psi_E(x)$ satisfies the time-independent Schrödinger equation for the potential. That is

$$\begin{gathered} H\psi_E(x) = E\psi_E(x) \\ H = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x) \end{gathered} \tag{18-59}$$

Outside the region $0 < x < a$, the potential is infinite, the particle does not move in this region because the energy of

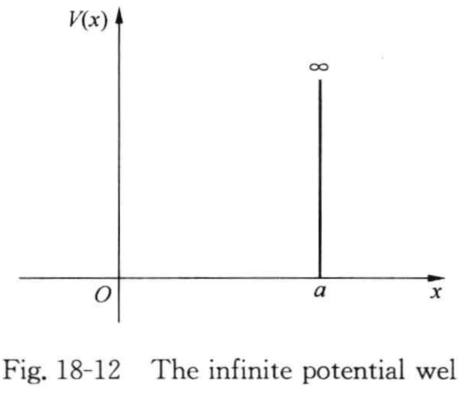


Fig. 18-12 The infinite potential well

the particle is impossible to be infinite. Therefore

$$\psi_E(x) \equiv 0, \quad x \leqslant 0 \text{ or } x \geqslant a \tag{18-60}$$

In the region $0 < x < a$, Eq. (18-59) may be expressed as

$$\frac{\partial^2}{\partial x^2}\psi_E(x) + k^2\psi_E(x) = 0 \tag{18-61}$$

where

$$k = \sqrt{2mE/\hbar^2} \tag{18-62}$$

For $E \geqslant 0, k \geqslant 0$, then the general solution to Eq. (18-61) is given by

$$\psi_E(x) = A\sin kx + B\cos kx, \quad 0 < x < a$$

where A and B are constant. Applying the continuous boundary conditions

$$\begin{cases} \psi_E(0) = 0 \\ \psi_E(a) = 0 \end{cases}$$

we obtain

$$\begin{cases} B = 0 \\ A\sin ka = 0 \end{cases} \tag{18-63}$$

Because the particle is in the well, A can not be equal to zero. Hence, we can satisfy Eq. (18-63) by choosing k so that

$$\sin ka = 0 \tag{18-64}$$

That is

$$ka = n\pi, \quad n = 1,2,3,\cdots \tag{18-65}$$

This can be expressed as

$$k_n = \frac{n\pi}{a}, \quad n = 1,2,3,\cdots \tag{18-66}$$

Knowing the allowed values of k, we can then obtain the solutions of the time-independent Schrödinger equation for the infinite square well

$$\psi_n = A\sin\frac{n\pi}{a}x, \quad n = 1,2,3,\cdots \tag{18-67}$$

where A is a constant which is determined by the normalization condition (18-53)

$$A = \sqrt{2/a}$$

so

$$\psi_n = \sqrt{\frac{2}{a}}\sin\frac{n\pi}{a}x, \quad n = 1,2,3,\cdots \tag{18-68}$$

Here n is called the quantum number.

Using Eq. (18-62) and Eq. (18-66), we find

$$E_n = \frac{\pi^2\hbar^2 n^2}{2ma^2}, \quad n = 1,2,3,\cdots \tag{18-69}$$

Thus we conclude that only certain values of the total energy E are allowed. The total energy of the particle in the box is quantized. From this example, we see that the quantization of total energy automatically appears in the quantum mechanics.

The solution corresponding to $n = 0$ is $\psi_0(x) = A\sin 0 = 0$ and it is ruled out because it does not describe a particle in a box. If $n=0$, this implies $E_0=0$, and the particle is at rest in the infinite potential well. In the classical physics, this can occur, but in the quantum mechanics, this is impossible to take place because if so, the uncertainty principle tells us that the momentum of the particle would be infinite (for $\Delta x=0, \Delta p \rightarrow \infty$). Therefore, the lowest energy is

$$E_1=\frac{\pi^2 \hbar^2}{2ma^2} \tag{18-70}$$

This is called the zero-point energy. The fact that the particle can not have zero total energy is basically a result of the uncertainty principle and the fact that the particle has a lowest energy is basically a phenomenon in the quantum mechanics. This is in sharp contrast to the idea, of classical physics, that all motion ceases when a system has its minimum energy content at the temperature of absolute zero.

The quantum number n has been used to label the different solutions of the transcendental equations, and the corresponding eigenfunctions. The first few eigenfunctions of the infinite square potential well with the corresponding energy are shown in Fig. 18-13. Note that for the number of half wavelength of each eigenfunction is equal to its quantum number n, and therefore the number of the nodes is $n+1$.

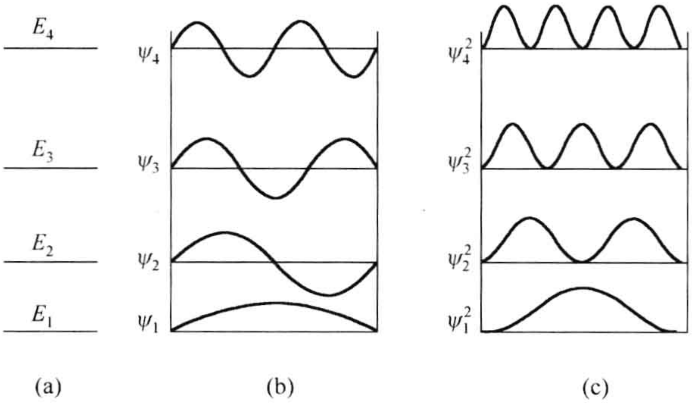


Fig. 18-13 (a) The energy diagram; (b) the wave functions and (c) the probability density for $n=1,2,3,4$

Example 18-5 Derive the infinite square potential well energy quantization law (18-69), directly from the De Broglie relation $\lambda=h/p$ by fitting an integral number of half De Broglie wavelengths $\lambda/2$ into the width a of the well.

Solution It is clear from Fig. 18-12 that for the infinite square potential well wave functions satisfy the following relation between the De Broglie wavelengths and the length of the well

$$n\frac{\lambda}{2}=a$$

This means

$$\lambda=\frac{2a}{n}, \quad n=1,2,3,\cdots$$

Hence, according to Eq. (18-24), the corresponding values of the momentum of the particle are

$$p=\frac{h}{\lambda}=\frac{nh}{2a}$$

As the potential energy of the particle is zero within the well, its total energy equals its kinetic energy. Thus

$$E_n = \frac{p^2}{2m} = \frac{h^2 n^2}{8ma^2} = \frac{\pi^2 \hbar^2 n^2}{2ma^2}$$

is in agreement with Eq. (18-69).

18.7 The Hydrogen Atom and Electron Spin

18.7.1 The hydrogen atom

The hydrogen atom is the simplest system that consists of an electron and a proton. This is also the most important case. It is of historical importance because it was the first system which Schrödinger treated with his theory of quantum mechanics. We shall see that the eigenvalues which the theory predicts for the hydrogen atom agree with those predicted by the Bohr's model and observed by experiment. This provided the first verification of the Schrödinger theory. The Schrödinger theory also predicts the eigenfunction and other properties of the hydrogen atom. Beyond these, the Schrödinger theory of the hydrogen atom is of great importance because it forms the foundation of the quantum mechanical treatment of all multi-electron atoms, as well as of molecules and nuclei.

The hydrogen atom is the simplest bound system that occurs in nature. However, it is more complicated than the system we have dealt with in the previous section because it contains two particles, and because it is three-dimensional. We denote by m the mass of the electron and assume that the mass of the nucleus is infinite so that the origin of the coordinate is located at the nucleus. Taking the potential at the infinite points to be zero, the Coulomb potential of this system is

$$V(r) = -\frac{Ze^2}{4\pi\varepsilon_0 r} \tag{18-71}$$

where r is the distance between the two particles. The time independent Schrödinger equation is

$$\left(-\frac{\hbar^2}{2m}\nabla^2 - \frac{Ze^2}{4\pi\varepsilon_0 r}\right)\psi(\boldsymbol{r}) = E\psi(\boldsymbol{r}) \tag{18-72}$$

which has the analytic solutions. The mathematical procedures that must be used in the treatment of Eq. (18-72) are complicated. In the following, we will give only the results obtained by solving Eq. (18-72).

One of the important results of the Schrödinger theory of the hydrogen atom is the prediction of the allowed values of total energy of the bound states of the atom that is

$$E_n = -\frac{m}{2\hbar^2}\left(\frac{Ze^2}{4\pi\varepsilon_0}\right)^2 \frac{1}{n^2} \tag{18-73}$$

which agrees exactly with those found in the Bohr model. The agreement of this energy spectrum with the experimental spectrum was pointed out when we analyzed the Bohr results. Therefore, Schrödinger's derivation of Eq. (18-73) provided the first convincing verification of his theory of quantum mechanics.

The complete eigenfunctions of the discrete spectrum for the hydrogen atom may be written as

$$\psi_{nlm}(r,\theta,\varphi) = R_{nl}(r)Y_{lm}(\theta,\varphi) \tag{18-74}$$

where (r, θ, φ) are the spherical coordinates of the electron, and (n, l, m) are the quantum numbers. The fact that three quantum numbers arise is a consequence of the fact that the time-independent Schrödinger equation contains three independent variables, one for each space coordinate. From Eq. (18-73) and Eq. (18-74), we see that although the eigenvalues of the hydrogen atom depend on only the quantum number n, the eigenfunctions depend on all three quantum numbers n, l, m. The quantum number n is sometimes called the principal quantum number, l is sometimes called the azimuthal quantum number, and m the magnetic quantum number.

The quantum numbers (n, l, m) satisfy

$$\begin{cases} n = 1,2,3,\cdots \\ l = 0,1,2,\cdots,n-1 \\ m = -l, -l+1,\cdots,0,1,\cdots,l-1,l \end{cases} \tag{18-75}$$

which make it apparent that for a given value of n there are generally several different possible values of l and m. Since the form of the eigenfunctions depends on all three quantum numbers, it is apparent that there will be situations in which two or more completely different eigenfunctions correspond to exactly the same eigenvalue E_n. As the eigenfunctions describe the behavior of the atom, we see that it has state with completely different behaviors that nevertheless have the same total energy. The physical word used to characterize this phenomenon is degeneracy, and eigenfunctions corresponding to the same eigenvalue are said to be degenerate. It is proved that the total degeneracy of the energy level E_n is given by

$$\sum_{l=0}^{n-1}(2l+1) = n^2 \tag{18-76}$$

Therefore, for each value of n, there are a total of n^2 degenerate eigenfunctions.

The magnitude L of the angular momentum of electron in its motion about the center of an atom is given by

$$L = \sqrt{l(l+1)}\,\hbar \tag{18-77}$$

and the z component L_Z is

$$L_Z = m\hbar \tag{18-78}$$

The angular momentum and its z component are quantized. Note that Eq. (18-77) is completely different with Eq. (18-10).

It can be proved that the selection rules of transitions (so-called electric dipole transitions, or the most common transitions) are

$$\begin{cases} \Delta l = l - l' = \pm 1 \\ \Delta m = m - m' = 0, \pm 1 \end{cases} \tag{18-79}$$

while $\Delta n = n - n'$ is arbitrary. Here the symbols n, l, m refer to the quantum numbers of the upper state and n', l', m' to those of the lower state of the transition.

The experiments show that the agreement of the total energy spectrum (18-73) with the results of experiments is not perfect and the difference can be explained by the various corrections such as the relativistic effects，the electron spin，the Lamb shift and the hyperfine structure.

18.7.2 Electron clouds

The solution of Eq. (18-72) yields a wave function $\psi_{nlm}(r, \theta, \varphi)$ that depends on the quantum numbers (n, l, m). What does $\psi_{nlm}(r, \theta, \varphi)$ tell us about the hydrogen atom? According to Born's interpretation of wave function，the probability density $|\psi_{nlm}(r, \theta, \varphi)|^2$ can be obtained consequently，which is the probability of finding the electron per unit volume about the position $\boldsymbol{r}$ at the time t. Fig. 18-14 gives the probability $P(r) = \psi_{100}^2 4\pi r^2$ of finding the electron at various radial distances from the nucleus in the state ψ_{100}. According to Fig. 18-14，one can see that $P(r)$ has a maximum value at a radius of 0.53 Å，the Bohr's radius. This means that there is a maximum probability of finding the electron in a small interval centered at this distance from the nucleus. However，as the curve indicates，there is also a probability of finding the electron in a small interval centered at any other distance from the nucleus. In other words，the electron is confined to a particular orbital distance from the nucleus，and the electron may be found at various distances from the nucleus，but the probability of finding it at a distance of the first Bohr orbit is a maximum.

The quantum mechanical result is often interpreted by viewing the electron as a cloud surrounding the nucleus. This "cloud" is called "electron cloud" which is used to describe the area around an atomic nucleus where an electron will probably be. It is also described as the "fuzzy" orbit of an atomic electron. The method of "electron cloud" visualizes the statistical explanation of the wave function of the hydrogen atom. The densest regions of the cloud represent those locations where the electron is most likely to be found. Fig. 18-15 shows two dimensional distributions of the electron clouds of two states of (2，0，0) and (2，1，1) for the electron in hydrogen atom，which are plotted by demonstration of computer random sampling simulation.

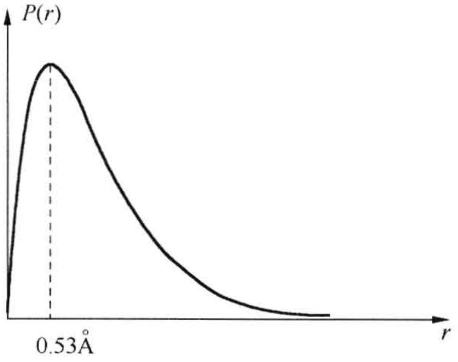
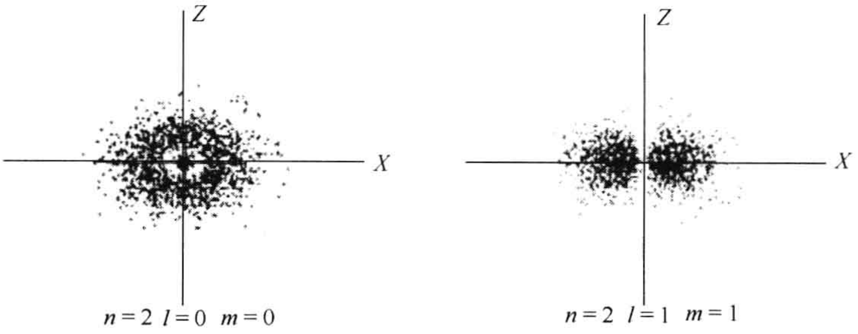


Fig. 18-14 The probability density versus distance from the nucleus for the hydrogen atom in ψ_{100}



(a) (b)

Fig. 18-15 The electron clouds of two states in hydrogen atom: (a) ψ_{200} and (b) ψ_{211}

18. 7. 3 Electron spin

In 1925 Uhlenbeck and Goudsmit assumed that an electron not only revolves around a nucleus but has a property, called spin, which may be visualized as a rotation of the electron about its own axis. This is similar to that of Earth that revolves around the Sun but has a rotation about its own axis. The discovery of this intrinsic property of the electron is of fundamental importance. In fact, it is now known that all particles can be assigned an internal angular momentum. In some cases, such as the pion (π meson) it is zero, but in others, such as the electron, the proton, the neutron, it is not zero.

We also assume that the magnitude S and the z component S_Z of the spin angular momentum are related to two quantum numbers, s and m_s, by quantization relations which are identical to those for orbital angular momentum. That is

$$\begin{cases} S = \sqrt{s(s+1)}\,\hbar \\ S_z = m_s\,\hbar \end{cases} \tag{18-80}$$

The s has the single value

$$s = \frac{1}{2}$$

and m_s has two possible values

$$m_s = \pm\frac{1}{2}$$

for the electrons. For a single electron, the total angular momentum is the vector resultant of the orbital angular momentum and the spin angular momentum.

Considering the quantum numbers (n, l, m), four quantum numbers (n, l, m, m_s) are required to specify the hydrogen states. The first three arise because three independent coordinates are required to describe its location in the three-dimensional space. The fourth arises because it is also necessary to describe the orientation in space of its spin, which can be either "up" or "down" relative to some z-axis. The relativistic theory of quantum mechanics developed in 1929 by Dirac shows that an electron must have an intrinsic $s = 1/2$ angular momentum.

18. 8 Multielectron Atoms and the Periodic Table

18. 8. 1 The multielectron atoms

A multielectron atom of atomic number Z contains a nucleus of charge $+Ze$ surrounded by Z electrons each of charge $-e$. Every electron moves under the influence of an attractive coulomb interaction exacted by the nucleus and the repulsive Coulomb interactions exacted by all the other $Z-1$ electrons, as well as certain weaker interaction involving the angular momenta. When the Schrödinger equation is applied to multielectron atoms, the resulting differential equation is awesome and only approximation solutions are available.

The quantitative information about multielectron atoms can be obtained from the approximation procedures such as the Hartree-Fock theory, but the required calculations must be carried out on large computers. However, it is found that one can describe complex atoms reasonably well if one thinks of each electron as existing in a hydrogen state called the individual state and moving in the electric field due to the nucleus and to the average contributions of the other electrons. In this very rough model, we imagine each electron with its own orbital and spin angular momentum moving in a fuzzy average hydrogen-like "orbit". When we treat complex atoms from the point of view of hydrogen-like "orbit", we find that we can specify the state of each electron in terms of four quantum numbers (n, l, m, m_s):

(1) **The principle quantum number n, on which the energy of a state depends, taking on integral values 1, 2, 3,⋯.**

(2) **The azimuthal quantum number l, which takes on integral values 0, 1, 2, ⋯, $n-2, n-1$.**

(3) **The quantum number m, which is a measure of the z component of the orbital momentum in units of $\hbar$.**

(4) **The quantum number m_s, which describes the z component of the spin angular momentum.**

In 1925, W. Pauli proposed an important rule, known as the Pauli exclusion principle:

> **In a multielectron atom there can never be more than one electron in the same quantum state.**

or

> **No two electrons in any one atom can have the same four quantum numbers.**

This principle has many applications and forms one of the foundation stones for our understanding of atomic structure.

The energy corresponding to the individual state (n, l, m, m_s) is proved to be dependent on both the principal quantum number n and the azimuthal quantum number l so that we write the energy of an electron in a multielectron atom as E_{nl}. For a fixed value of l, the energy E_{nl} is an increasing function of n and for a given value of n it is an increasing function of l. In general, the individual electron energy E_{nl} is an increasing function of the sum $n+l$. Following the usual spectroscopic notation, the level corresponding to E_{nl} is labeled by two symbols. The first one gives the value of the principle quantum number n and the second indicates the azimuthal quantum number l according to the correspondence

value of l:	0	1	2	3	4	5 ⋯
code letter:	s	p	d	f	g	h ⋯

Using this symbol, the ground state of hydrogen is a 1s state, and the first excited state is fourfold degenerate and contains a 2s state and three 2p states with $m=-1, 0, +1$.

Electrons having the same value of n and l are said to belong to the same subshell. It can be proved that there are $2(2l+1)$ states having the same values of n and l but different values of m and m_s. Such states are called equivalent. The electrons belonging to the same subshell are known as equivalent electrons. According to the Pauli exclusion principle, the maximum number of electrons in a subshell is $2(2l+1)$. The order of succession of the individual energy levels E_{nl} is nearly the same for all atoms, and is given in Table 18-2.

Table 18-2 The ordering of the individual energy levels E_{nl} (The energy increases from bottom to top.)

Quantum numbers n, l	Spectroscopic notation for subshell (nl)	Maximum number of electron allowed in the subshell $= 2(2l+1)$
6,2	6d	10
5,3	5f	14
7,0	7s	2
6,1	6p	6
5,2	5d	10
4,3	4f	14
6,0	6s	2
5,1	5p	6
4,2	4d	10
5,0	5s	2
4,1	4p	6
3,2	3d	10
4,0	4s	2
3,1	3p	6
3,0	3s	2
2,1	2p	6
2,0	2s	2
1,0	1s	2

Electrons which have the same value of the principle quantum number n are said to belong to the shell. The value of the principle quantum number n is sometimes specified by a capital letter according to the correspondence

value of n:	1	2	3	4	5	6
code letter:	K	L	M	N	O	P

The maximum number of electrons in a shell is proved to be $2n^2$, in which case has a closed (or filled) shell.

18.8.2 The periodic table of the elements

Most of the properties of the chemical elements are periodic functions of the atomic number Z that specifies the number of electrons in an atom of the element. It was first emphasized by Mendeleev in 1869 that these periodicities can be made most apparent by constructing a periodic table of the elements. A modern version of his table is presented in Table 18-3. Each element is represented in the table by its chemical symbol, and also by its atomic number. Elements with similar chemical and physical properties are in the same column. The discovery of the periodic table was a great breakthrough of chemistry. Its interpretation was an equally significant development of physics.

Table 18-3 The Periodic Table of the Elements

1 H 1.00794						Atomic number **Symbol** Standard atomic weight											2 He 4.0026
3 Li 6.941	4 Be 9.01218											5 B 10.811	6 C 12.0107	7 N 14.0067	8 O 15.9994	9 F 18.9984	10 Ne 20.1797
11 Na 22.9897	12 Mg 24.3050											13 Al 26.981	14 Si 28.085	15 P 30.974	16 S 32.065	17 Cl 35.453	18 Ar 39.948
19 K 39.0983	20 Ca 40.078	21 Sc 44.9559	22 Ti 47.867	23 V 50.941	24 Cr 51.996	25 Mn 54.938	26 Fe 55.845	27 Co 58.933	28 Ni 58.693	29 Cu 63.546	30 Zn 65.382	31 Ga 69.723	32 Ge 72.64	33 As 74.921	34 Se 78.96	35 Br 79.904	36 Kr 83.798
37 Rb 85.4678	38 Sr 87.62	39 Y 88.9058	40 Zr 91.224	41 Nb 92.906	42 Mo 95.96	43 Tc (98)	44 Ru 101.07	45 Rh 102.905	46 Pd 106.42	47 Ag 107.868	48 Cd 112.411	49 In 114.818	50 Sn 118.710	51 Sb 121.760	52 Te 127.60	53 I 126.904	54 Xe 131.293
55 Cs 132.905	56 Ba 137.327	57-71 Rare Earths	72 Hf 178.49	73 Ta 180.948	74 W 183.84	75 Re 186.207	76 Os 190.23	77 Ir 192.217	78 Pt 195.084	79 Au 196.966	80 Hg 200.59	81 Tl 204.383	82 Pb 207.2	83 Bi 208.980	84 Po (209)	85 At (210)	86 Rn (222)
87 Fr (223)	88 Ra (226.02)	89-103 Actinides	104 Rf (265.11)	105 Db (268)	106 Sg (269)	107 Bh (270)	108 Hs (269)	109 Mt (278)	110 Ds (281)	111 Rg (281)	112 Cn (285)	113 Uut (286)	114 Fl (289)	115 Uup (288)	116 Lv (293)	117 Uus (294)	118 Uuo (294)

Rare Earths (Lanthanides)	57 La 138.905	58 Ce 140.116	59 Pr 140.907	60 Nd 144.242	61 Pm (145)	62 Sm 150.36	63 Eu 151.964	64 Gd 157.25	65 Tb 158.925	66 Dy 162.500	67 Ho 164.930	68 Er 167.259	69 Tm 168.9342	70 Yb 173.04	71 Lu 174.967
Actinides	89 Ac (227)	90 Th 232.0381	91 Pa 231.0358	92 U 238.0289	93 Np (237)	94 Pu (244)	95 Am (243)	96 Cm (247)	97 Bk (247)	98 Cf (251)	99 Es (252)	100 Fm (257)	101 Md (258)	102 No (259)	103 Lr (262)

Now the characteristics of an atom depend on the behavior of its electrons. The behavior of an electron is specified by the set of four quantum numbers that specify its quantum state. However, E_{nl} depend only on the quantum number n and l, and then an atom can be characterized by specifying the n and l quantum numbers of all the electrons. This specification of the subshells occupied by the various electrons is called the configuration of the atom, which is represented by the symbol

$$(n_1 l_1)^{x_1} \ (n_2 l_2)^{x_2} \cdots$$

where n_i and l_i are the quantum numbers n and l and where x_i is the number of the electrons in the $n_i l_i$ subshell. For example, the ground state of hydrogen is $1s^1$; the ground state of helium is $1s^2$. In the ground state:

> **the electrons must fill all the subshell in such a way as to minimize the total energy of the atom and yet not exceed the capacity $2(2l+1)$ of any subshell.**

which is called as the principle of the minimum energy. The subshells will fill in order of increasing energy, as listed in Table18-2. In the following, we briefly discuss how the subshells are occupied by the electrons of an atom in the ground state.

Hydrogen has the ground state configuration $1s^1$. The next element, helium, has the ground state configuration $1s^2$, the two electrons of which fill the K shell ($n = 1$). The hydrogen and helium constitute the first period (row) in the period table.

The third element, Lithium, has the ground state configuration $1s^2 2s^1$, and the third of three electrons cannot have $n = 1$ but goes into the lowest available state for which $n = 2$, $l = 0$, from which the second period begins. The ground state of Beryllium is $1s^2 2s^2$ and the 2p subshell becomes progressively filled beginning with Boron ($Z=5$, ground state configuration $1s^2 2s^2 2p$) up to Neon ($Z=10$, ground state configuration $1s^2 2s^2 2p^6$). Neon has the maximum number of electrons allowed in the $n = 2$ (L) shell, and the second period ends at this point.

With $Z=11$ (Sodium) the eleventh electron must go into the 3s subshell. From $Z=11$ to $Z=18$ the 3s and 3p subshells are progressively filled, the ground state configuration of Argon ($Z=18$) being [Ne] $3s^2 3p^6$ (here [Ne] $= 1s^2 2s^2 2p^6$). The third period consists of these elements.

The process of filling the $n=3$ states is temporarily interrupted at $Z=19$ (Potassium). Looking at Table 18-2, the added electrons in Potassium ($Z=19$) and Calcium ($Z=20$) go into the 4s rather than the 3d subshell, the 3d level being energetically less favorable because of the screening by the Argon core [Ar]. The filling of the 3d subshell is therefore deferred until Scandium ($Z=21$), which has the ground state configuration [Ar] $4s^2 3d^1$. It is the first element of the so-called first transition or iron group, extending from $Z=21$ to $Z=30$ (Zinc). It is worth noting that because the 4s and 3d states are very close in energy, a competition between these two states develops, and the process of filling is not so regular as for the previous subshells. The detail is shown in the Table 18-3. The last element of the first transition or iron group, Zinc ($Z=30$) has the ground state configuration [Ar] $4s^2 3d^{10}$.

Two other sets of transition elements, the second transition or Palladium group (from 39 to $Z=48$) and the third transition or Platinum group (from $Z=71$ to $Z=80$) correspond to similar situations in first transition, but with $n=4$ and $n=5$, respectively. Here again a competition develops between the nd and the $(n+1)$s levels, and irregularities in the filling of the subshell occur.

The rare earth elements or Lanthanides contains the 14 elements, beginning with Lanthanum ($Z=57$) which corresponds to the filling of the 4f subshell, the 4s subshell being already complete. We note that this filling of the 4f subshell is irregular, with a competition taking place between the 4f and 5d levels. Analogous to the rare—earth elements are the actinide, beginning with actinium ($Z=89$) in which the 5s subshell is complete and a competition occurs between the 5f and the 6s states.

The electron configurations of atoms with large values of Z are clearly difficult to explain on the basis of the simple qualitative arguments developed above. One reason, which we have already mentioned, is that various energy levels are very close in energy. Also, for large Z, relativistic effects become important. However, the period table is very well explained by the quantum mechanics.

Questions

18-1 In a collision between an α particle and an electron, what general considerations limit the momentum transfer? Does the fact that the force is Coulombic play any role in this respect?

18-2 How does the Thomson atom differ from a random distribution of protons and electrons in a spherical region?

18-3 What is the experimental evidence that all the positive charge and most of the mass of an atom are concentrated in a nucleus?

18-4 Did Bohr postulate the quantization of energy? What did he postulate?

18-5 For the Bohr hydrogen atom orbits, the potential energy is negative and greater in magnitude than the kinetic energy. What does this imply?

18-6 Can a hydrogen atom in its ground state absorb a photon with an energy of 7. 5 eV? One with an energy of 10. 2 eV? One with a wavelength equal to the short—wavelength limit of the Balmer series?

18-7 Can a hydrogen atom absorb a photon whose energy exceeds its binding energy of 13. 6 eV?

18-8 What is the maximum number of photons that can be emitted when a hydrogen atom excited to a $n = 4$ level goes to the ground state? What is the minimum number(Assume there are no nonradiative transitions)?

18-9 Why is the wave nature of matter not more apparent to us in our daily observations?

18-10 Is an electron a particle? Is it a wave? Explain?

18-11 What features of the Bohr model of hydrogen are retained in wave mechanics? What features are abandoned?

18-12 Discuss similarities and differences between a matter wave and an electromagnetic wave.

18-13 Why is it so important for the Schröodinger equation to be linear in the wave function?

18-14 The mass m of a particle appears explicitly in Schrödinger equation, but its charge q does not, even though both may affect its motion. Why?

18-15 What is the basic connection between the properties of a wave function and the behavior of the associated particle?

18-16 Why does the probability density function have to be everywhere real, non-negative, of finite and definite value?

18-17 Since the wave function describing the behavior of a particle satisfies a differential equation, its evolution in time is perfectly predictable. How does this fact fit in with the uncertainty principle?

18-18 In the $n = 3$ state, the probability density function for a particle in a box is zero at two positions between the walls of the box. How then can the particle ever more across these positions?

18-19 What is the Pauli exclusion principle? How does it help us understand the period table?

18-20 What is the consequences of the fact that the sizes of all atoms are about the same? What are the reasons for this fact?

Problems

18-1 Prove that Planck's constant has the dimensions of angular momentum.

18-2 Compute the wavelengths of the first two lines and of the series limit for the Lyman series of hydrogen. What are the energies of photons with these wavelengths?

18-3 Calculate the speed of electron in a ground state hydrogen atom.

18-4 What is the energy, momentum, and wavelength of a photon that is emitted by a hydrogen atom making a direct transition from an excited state with $n = 10$ to the ground state? Find the recoil speed of the hydrogen atom in this process.

18-5 Calculate the shortest wavelength of the Lyman series line in hydrogen; of the Paschen series; of the Pfund series. In what region of the electromagnetic spectrum does each lie?

18-6 How much energy is required to remove an electron from a hydrogen atom in a state with $n = 4$?

18-7 A hydrogen atom in the ground state absorbs a 20. 0eV photon. What is the speed of the liberated electron?

18-8 A bullet of mass 40g travels at 1000 m/s. (1) What wavelength can we associate with it? (2) Why does the wave nature of the bullet not reveal itself through diffraction effects?

18-9 An electron and a photon each have a wavelength of 2. 0 Å. What are their (1) momenta and (2) total energies? (3) Compare the kinetic energies of the electron and the photon.

18-10 A particle moving with kinetic energy equal to its rest energy has a de Broglie wavelength of 1.7898×10^{-6} Å. If the kinetic energy doubles, what is the now de Broglie wavelength?

18-11 If $\Delta\lambda/\lambda = 10^{-7}$ for a photon, what is the simultaneous value of Δx for (1) $\lambda = 5.00\times10^{-4}$ (γ-ray)? (2) $\lambda = 5.00$ Å (X-ray)? (3) $\lambda = 5000$ Å (light)?

18-12 The velocity of a positron is measured to be: $v_x = (4.00\pm0.18)\times10^5$ m/s, $v_y = (0.34\pm0.12)\times10^5$ m/s, $v_z = (1.41\pm0.10)\times10^5$ m/s. What minimum volume was the positron located at the moment the measurement was carried out?

18-13 If the wave functions $\Psi_1(x,t)$, $\Psi_2(x,t)$, $\Psi_3(x,t)$ are three solutions to the Schrödinger equation for a particular potential $V(x,t)$, show that the arbitrary linear combination $\Psi(x,t) = c_1\Psi_1(x,t) + c_2\Psi_2(x,t) + c_3\Psi_3(x,t)$ is also a solution to that equation.

18-14 Find a solution to Schrödinger equation for $V(x) = 0.0$

18-15 For a particle in a box, show that the fractional difference in the energy adjacent eigenvalues is

$$\frac{\Delta E_n}{E_n} = \frac{E_{n+1} - E_n}{E_n} = \frac{2n+1}{n^2}$$

and discuss the meaning of the obtained formula.

18-16 Apply the normalization condition to show that the value of constant A for the wave function (18-67) is $\sqrt{2/a}$.

18-17 Find the probability with which the particle in a state $n = 2$ within $(0, a/3)$ in the infinite potential well.

18-18 Find the maximum number of electrons which are contained in the 4f subshell.

Chapter19

The Band Theory of Solids

In past chapter, we have discussed many properties of a single atom in the term of quantum theory. Now we will study the electric property of solids, which consist of a great deal of atoms. Solids have many properties—mechanical, thermal, electrical, magnetic, and optical. Some aspects have been discussed in the earlier chapters. Here we shall focus on the electrical conductivity of solids.

19.1 The Crystals

19.1.1 The types of solids

The matter is in the gaseous state or the solid state or the liquid state at ordinary temperatures and pressures. In the gaseous state the average distance between molecules is large compared to the size of a molecule, so the molecules may be regard as isolated from one another apart from the collisions of molecules. However, in the solid state molecules (or atom) can no longer be regarded as isolated. Their separation is comparable to the molecular size, and the strength of the forces holding them together is of the same order of magnitude as the forces binding the atoms into a molecule, and they form a block substance which maintains both its shape and its volume. Such a mass is commonly known as a solid. Hence, the properties of a molecule in solids are altered by the presence of neighboring molecules.

According to the origin of the forces between molecules, solids can be classified as molecular, ionic, covalent, and metallic solids:

(1) Molecular solids consist of molecules that are so stable that they retain much of their individuality when brought in close proximity. The electrons in the molecule are all paired so these atoms in different molecules cannot form covalent bonds with one another. The forces between molecules are the van der forces, which are as weak as in the gaseous state. Because of the weakness of van der forces, the binding energies of molecular solids are low so that these solids have low melting point, and are highly compressible and mechanically weak. The absence of free electrons makes molecular solids very poor conductors of heat or electricity.

(2) Ionic solids, such as sodium chloride, are composed of positive and negative ions. The binding forces, which come essentially from the coulomb attraction between these ions, are relatively large. Because there are no free electrons to carry energy or charges from one part of the solid to another, such solids are poor conductors of heat or electricity. Because of

the strong electrostatic forces between the ions, ionic solids are usually hard and have high melting points. Ordinary, ionic crystals are relatively transparent to visible.

(3) Covalent solids contain atoms that are bound by shared valence electrons, as in covalent binding of molecules. Examples of crystals exhibiting covalent binding are germanium, silicon and diamond. The bonds are directional and determine the geometrical arrangement of atoms in the crystal structure. Due to the rigidity of their electronic structure, covalent solids are ordinarily hard and brittle and have high electric resistance and high melting points at low temperature. As the temperature is increased, there is a marked decrease in resistance. Practical semiconductors are based on covalent solids.

(4) Metallic solids, in which the valence electrons are not tied to any particular atom or pair of atoms but are free to wander through the crystal, can be thought of as a limiting case of covalent solids in which electrons are shared by all the ions in the crystal. A metallic solid is a regular lattice of spherically symmetrical positive ions, arranged like close—packed spheres, through which the electron moves. Metallic solids are obviously excellent conductors of electricity, or heat, the electrons of which easily absorb energy from incident radiation or lattice vibrations, and move under the influence of an applied electric field, or thermal gradient. Because radiation in the visible portion of electromagnetic spectrum is easily absorbed, such solids are opaque.

(5) Hydrogen binding solids is a special type of solids. Since the hydrogen atom has a single electron, one might expect it to form a bond with only one other atom, but hydrogen frequently forms a strong bond between two hydrogen atoms. We may think of this bond as primarily ionic in character, since it is formed only with strongly electronegative atoms. The hydrogen atom loses or transfers its electron to one of these atoms, and the proton binds two negative ions, nestling between these very much larger ions. Only hydrogen forms such bonds; no other ion has the very small size of the proton—so small that there are only two nearest neighbors. The hydrogen bond occurs in ice $(H_2O)_n$, solid hydrogen fluoride, and a wind variety of proteins and other organic compounds. Therefore, it is of great interest in biophysics and biochemistry.

The type of binding that a particular solid has is determined experimentally by studies of X-ray diffraction, dielectric properties, optical emissions, and so forth. There are some solids whose binding must be interpreted as a mixture of the principal types we have described. In addition, not all solids have the ideal structure implied by the discussion so far. Indeed, the so—called lattice imperfections, or deviations from ideal crystal structure, lead to many properties of solids that have practical consequences.

Every solid has an enormous range of properties that we can choose to examine. Is it transparent? Can you hammer it out into a flat sheet? What kinds of waves travel through it and at what speeds? Does it have interesting magnetic properties? Is it a good heat conductor? What is its crystal structure? Does it have special surface properties? ··· The list goes on. We choose here to focus on a single question: "What are the mechanisms by which a solid conducts, or does not conduct, electricity?"

19.1.2 The crystal structures of solids

To the naked eye, a solid appears as a continuous rigid body. In fact, all solids are composed of discrete basic units—atoms. These atoms are not distributed randomly, but are arranged in a highly ordered manner relative to each other. Such a group of ordered atoms is refereed to as a crystal. A solid is said to be a crystal if the atoms are arranged in such a way that their positions are exactly periodic. There are several types of crystalline structure, depending on the geometry of the atomic arrangement, and the knowledge of these is important in solid physics because these structures usually influence the physical properties of solids.

The atoms in some solids appear to be randomly arranged, i. e. , the crystalline structure is absent. These solids are called as noncrystalline, or amorphous solids which will not be discussed in this text. The most familiar example of an amorphous solid is ordinary window glass. Chemically it is a silicon oxide. Structurally it has no crystal structure at all.

In order to study the geometrical properties of the crystal, one replaces each atom by a geometrical point located at equilibrium position of that atom. This geometrical pattern is the crystal lattice, or simply the lattice, and all the atom sites have been replaced by lattice sites. Fig. 19-1 illustrates the periodicity of the lattice. The distance between any two nearest neighbors along the x direction is a, and along the y direction is b (the x and y axes are not necessarily orthogonal). A perfect crystal maintains this periodicity in both the x and y directions from $-\infty$ to ∞. The constants a and b are called as the periods along the x or y directions. It fellows from the periodicity that the atom A, B, C, etc. , are equivalent. In other words, to an observer located at any of these atomic (lattice) sites, the crystal appears exactly the same. This idea is often expressed by saying that a crystal possessed a transitional symmetry, meaning that if the crystal is translated by any vector joining two atoms (lattice sites), say $\boldsymbol{R}$ in Fig. 19-1, the crystal appears exactly the same as it did before the translation. In other words, the crystal remains invariant under any such translation. The periods along different direction may be unequal. Strictly speaking, one cannot prepare a perfect crystal because the practical solids are finite and always contains some foreign atoms, i. e. , impurities. Even with the best crystal-growing techniques, some impurities remain, which spoil the perfect crystal structure. The imperfection of crystal also results into the thermal vibrations of atoms around their equilibrium positions for any temperature $T>0$ K. If the effects of imperfection are negligible, we speak of a "perfect" crystal. Many of most interesting phenomena in solids are associated with imperfections.

Fig. 19-2 shows the periodicity of three-dimension solids, where $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ are three noncoplanar vectors joining the lattice point at the origin to its near neighbors. The parallelepiped whose sides are these vectors is a unit cell. If such a unit call is translated by all the vectors $\boldsymbol{R}_n$

$$\boldsymbol{R}_n = n_1\boldsymbol{a} + n_2\boldsymbol{b} + n_3\boldsymbol{c} \tag{19-1}$$

where n_1, n_2, n_3 are a triplet of integers 0, ±1, ±2, etc. , the whole lattice is covered. It is noted that the choice of a unit cell is not unique. A unit cell, if there is no lattice point inside the cell or on its surface, is called a primitive cell. It can be proved that all primitive cells have equal volumes. As shown in Fig. 19-2, the primitive cell has eight points at its corners, but each of these points is shared by eight adjacent cells. Hence each primitive cell has only one lattice point.

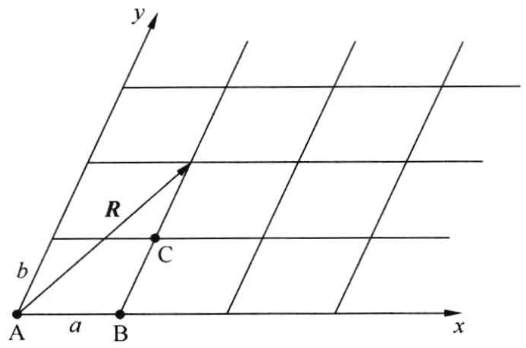


Fig. 19-1 A crystalline solid in which all the atoms are arranged periodically

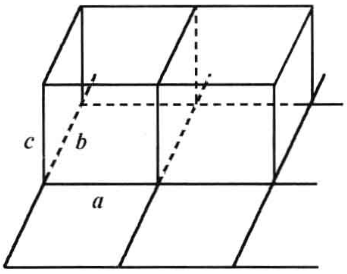


Fig. 19-2 A three-dimensional lattice

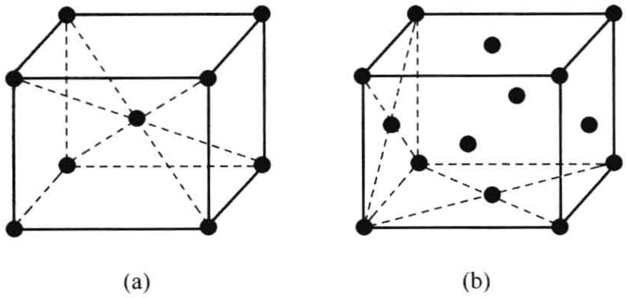

(a) (b)

Fig. 19-3 (a) A body-centered lattice and (b) a face-centered lattice

Apart from the periodicity of crystal, the crystal has its particular symmetry. Hence, it is sometimes convenient to deal with nonprimitive cells, ones which have additional points either inside the cell or on its surface. Fig. 19-3(a) shows a body-centered lattice which has one additional point at the center of the cell. All body-centered lattice has two lattice points and is not a primitive cell. Fig. 19-3(b) shows a face-centered lattice which has six additional points, one on each face. The points on the surface are shared by the two adjacent cells. So, a face-centered lattice has four lattice points. If $a = b = c$, the body-centered lattice or the face-centered lattice are called as the body-centered or face-centered cubic. The most familiar metals: Ag, Al, Au, Cu, Co, Fe, Ne, Pb and Pt, all crystallize in the face-centered cubic structure. Some of the metals which crystallize in the body-centered cubic structure are: Fe and the alkalis Li, Na, K, Rb, and Cs. The structure of the ordinary table salt NaCl is cubic, and there is an alternation of Na and Cl atoms. Its cell is a face-centered cubic one, which can be looked as a structure composed of two interpenetrating face-centered structure sublattices: one made up of Na atoms and the other of Cl atoms.

19.2 Energy Bands in Solids

19.2.1 The exchange degeneracy

Let's discuss two hydrogen atoms that are far apart and may be looked as two free atoms. Assuming two free atoms are in the ground state, the energy of this system is given by

$$E = E_0 + E_0 = 2E_0 \tag{19-2}$$

where E_0 is the energy of the ground state for hydrogen atom. In the formal language of quantum physics, these two free atoms are identical particles. As shown in Fig. 19-4, exchanging the positions of two free atoms, the system has no change because the two free atoms are indistinguishable in micro-world. This property is called as exchange symmetry. That is, measurable results obtained from accurate quantum mechanical calculations should not depend on the assignment of labels to identical particles. This leads to important effects which have no classical analogies because the indistinguishability itself is purely quantum mechanical.

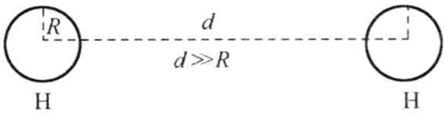


Fig. 19-4 Two isolated hydrogen atoms

According to the quantum mechanics, the wave function of the system with the energy Eq. (19-2) is not unique. Therefore, the energy level is degeneracy. It can be proved that the energy levels of this two-atom system have a twofold exchange degeneracy, which results from the indistinguishability of two atoms.

19.2.2 Energy bands in solids

A solid is a quantum system containing the vast number of particles. In principle, we can study its properties by using of the quantum theory. This is a many-body problem that deals with the complex mathematics. Here, we discuss the energy levels in a simply directive way.

If we bring the atoms of Fig. 19-4 close together, they will, speaking loosely, gradually begin to sense each other's presence. That is their wave function (the electric cloud) will begin to overlap and there will be interactions between the atoms. The system is not a group of two independent and isolated atoms, and becomes a single two-atom system (that is H_2) in which there are interactions. According to the quantum mechanics, the exchange degeneracy is removed and each energy level of the isolated atom splits into two levels for the two atom system. The splitting increases as the separation of the atoms decreases.

If we had started with three isolated atoms, we would have a threefold exchange degeneracy of the energy levels. When the atoms are brought together in a uniform linear lattice, each of the levels splits into three distinct levels. As we go to a system containing N atoms of a given species, each level of one of these atoms leads to an N-fold degeneracy level of the system when the atoms are well separated. With decreasing separation, each of these levels splits into N levels of which a set consists. The spread in energy between the lowest and highest level of a particular set depends on the separation distance. As shown in Fig. 19-5, a set of levels is called as an energy band.

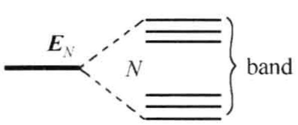


Fig. 19-5 A set of levels

At the values of R found in a solid, a few angstroms, the energy spread is of the order of a few volts. Since N is of the order of the Avogadro number (10^{23} per mole), we can see that the individual energy levels with a band are so extremely closely spaced in energy that they form a practically continuous energy band. Fig. 19-6 suggests the band structure of the levels in a hypothetical solid in which we have assumed, for simplicity, that the band do not overlap. The gaps between the bands represent ranges of energy that no electron may possess.

Note in Fig. 19-6 that the bands that lie lower in energy are narrower than these bands that lie higher. This is because these low energy bands correspond to levels in the isolated atom that are occupied by electrons that spend most of their time deep within the electron cloud of the atom. The reason is that the electrons in lower levels are electrons in inner subshells of the atoms, which are not significantly influenced by the presence of nearly atoms. These electrons are localized on particular atoms, even when the separation distance is small. The valence electrons, on the other hand, are not localized at all for small separation distance, but they become part of the whole system. The overlap of their wave functions results in a spreading of their energy levels. Therefore, the wave functions of the core electrons do not overlap as much as do the wave functions of the outer or valence electrons and for this reason the splitting of the levels, although it must occur, is not as great as it is for the levels normally occupied by the outer electrons.

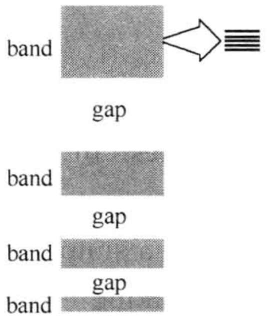


Fig. 19-6 The band-gap pattern for the energy levels of a solid

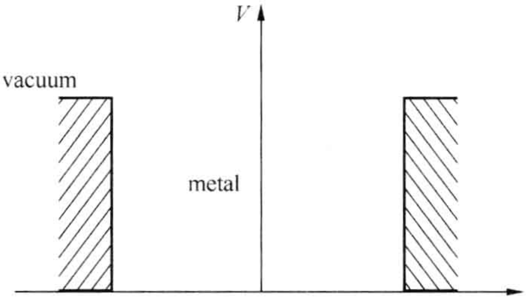


Fig. 19-7 The potential in the free-electron

Now that we have established the pattern of levels for a solid, we are ready to consider how these levels are filled with electrons and then explain the properties of solid.

19.3 Metals, Insulators, and Semiconductors

19.3.1 Metals—the quantum free-electron model and Fermi distribution

There are many free electrons (that is the electrons can move in the whole solid) in the metal. These electrons are called as the conduction electrons, and assumed to be completely free, except for a potential at the surface, as shown in Fig. 19-7, which has the effect of confining the electrons to the interior of the solid. In the absence of an applied electric field, the directions in which these electrons move are random. When an electric field is applied to a metal, the electrons modify their random motion in such a way that, on the average, they drift slowly in the direction opposite to that of the field.

For the free electrons in the metal, we treat them as an ideal gas of fermions inside the solid. At the absolute zero of temperature, as shown in Fig. 19-7, we can assume with little error that the potential energy of the conduction electrons remains constant as they move about with in the solid. If we set this constant equal to zero (this is permitted to do), the total energy E associated with any level is equal to the kinetic energy of the electron that occupied that level. The electron gas is different from the ideal gas in chapter 10, and it is the quantum free-electron gas. This approximation for metal, which greatly simplifies quantum mechanical calculations, turns out to be surprisingly good in determining many of the observed properties of metals.

The Pauli exclusion principle requires that free electrons in a metal have a broad rang of kinetic energies. Even at 0 K the most energetic electrons have a kinetic energy of a few electron volts. The energy distribution of the conduction electrons in a metal, Fermi distribution, can be proved to be given by

$$n(E)=\frac{8\pi V\,(2m^3)^{1/2}}{h^3}\,\frac{E^{1/2}}{e^{(E-E_f)/kT}+1} \tag{19-3}$$

where V is the volume of solid, k is the Boltzmann constant and m is the mass of the electron. E_f is called the Fermi energy, the energy of the highest state occupied by electrons at 0 K.

Fermi distribution is plotted in Fig. 19-8, which shows that at the absolute zero of temperature, all states with energies less than the Fermi energy are filled and all states with energies greater than that energy are vacant. When the temperature is increased, a few electrons are raised to slightly higher energies, leaving vacancies in levels just below the Fermi level. A glance at Fig. 19-8 should be enough to shatter the popular misconception that all motion ceases at the absolute zero of temperature. The conduction electrons in a metal have plenty of energy at 0 K (Note that the average momentum of the valence electrons is zero).

It can be proved that

$$E_f=\frac{0.121h^2}{m}n^{\frac{2}{3}} \tag{19-4}$$

Here n is the number of conduction electrons per unit volume. Because n of metals is different, the Fermi energy of metals is unequal.

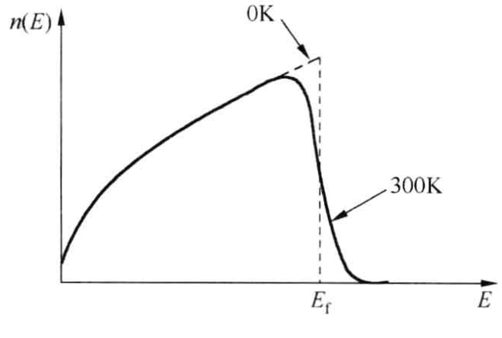


Fig. 19-8 Fermi distributions at 0K and 300K

One can use the quantum free-electron model and the above formula to study the properties of metals such as thermal energy, the heat capacity of the electrons, electrical conductivity, which can be explained.

19.3.2 Metal—the band structure

Fig. 19-9 shows the feature of a metal at the absolute zero of temperature. The highest occupied level falls somehow in the middle of a band. The energy band contains some electron but is not completely filled. The electrons that occupy this partially filled band are the valence electrons of the atoms, and being free to move throughout the solid, become the conduction electrons of the solid.

If you apply an electric field $\boldsymbol{E}$ to a metal, it exerts a force $-e\boldsymbol{E}$ on each electron. In a metal, this force, during times Δt, causes every conduction electron to acquire a velocity increment Δv in the direction of $-\boldsymbol{E}$. This will require the electron to change their energies, but there are vacant levels available so that these rearrangements can be made. Hence, the average momentum of valence electron in an electric field is not zero and in the direction of $-\boldsymbol{E}$, and the electric current will form in the metal. So, an electric field can produce a flow of these electrons, and the material is an excellent electric conductor at all temperature.

The velocities of the individual conduction electrons do not increase without limit, however, because of collisions associated with the thermal vibrations of the lattice. Thus, after a certain time τ called the relaxation time, the drift velocity of the conduction electrons settles down to a constant limiting value, which we associated with the constant current that is set up by the applied electric field.

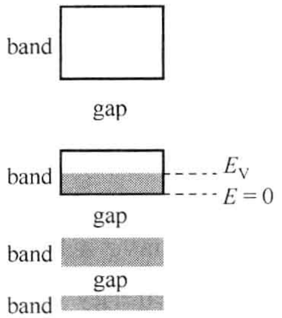


Fig. 19-9 An idealized representation of the band-gap pattern for a metal at the absolute zero of temperature

In order to remove one of the least tightly bound electrons from a metal at 0K, the electron must be supplied enough energy to take it from the Fermi level to the potential just outside the metal. We found evidence for this statement in the photoelectric effect. Similarly in order to achieve thermal emission, the temperature must be raised high enough for electron to escape via thermal excitation.

19.3.3 Insulators

Fig. 19-10 shows the feature of energy band of an insulator, in which the highest occupied level coincides with the top of a band and the gap between the highest filled band and the next permitted (but empty) band is large. For diamond, $E_g = 5.4$ eV, a value some 140 times larger than the average energy of a free particle at room temperature. Hence, modest temperatures are not sufficient to lift electrons to the conduction bands. Electrons in the filled band are not free, and the material serves as an excellent insulator.

If you apply an electric field $\boldsymbol{E}$ to an insulator, it will exert a force $-e\boldsymbol{E}$ on each electron. If the energy of an electron changes, the electron must move to a different energy level within the solid. However, the Pauli principle prevents from doing so because all other levels within the band into which the electron might move are already occupied. These electrons are in total rigid lock. There are plenty of vacant levels in the band above the filled band in Fig. 19-10, but, if an electron is to occupy one of these levels, it must somehow jump across the gap that separates the two bands. Of course, if the temperature rises sufficiently high, electrons can be excited to conduction levels and then no material is an insulator at extremely high temperatures. However, if the energy gap is reasonably large, practically no electrons are lifted to the conduction band at room temperature.

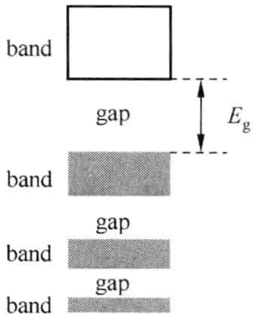


Fig. 19-10 An representation of he band-gap pattern for an insulator

19.3.4 Semiconductor

A semiconductor is like an insulator in that its uppermost filled level (at the absolute zero of temperature) lies at the top of a band. A semiconductor differs from an insulator, however, in that the gap between this filled band and the next vacant band above it is much smaller than that of an insulator, as shown in Fig. 19-11. Therefore, there is a real possibility for electron to "jump the gap" into the empty band by thermal agitation. The semiconductors are the good insulators at the absolute zero of temperature. For semiconducting material, the highest filled band is called the valence band because the electrons that occupy it are the valence electrons of the isolated atoms. The band above the valence band, which is vacant at $T=0$ K, is called the conduction band.

As the temperature is raised, thermal agitation gives some electrons enough energy to shift to the higher band. The electrons so excited serve as conduction electrons. Further, the holes left by the electrons when they move to the higher band also serve as charge carriers. Such an arrangement of energy levels is characteristic of semiconductors, which are insulators at 0 K but become poor conductors at somewhat higher temperature. Semiconductors become steadily better conductors as the temperature is increased because more electrons are transferred to the conduction band. Each electron raised to conduction band leaves a vacancy, or holes, in the valence band.

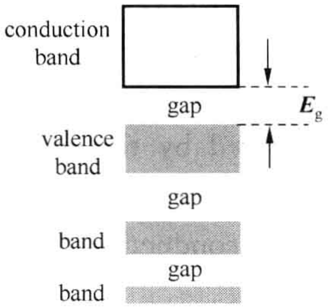


Fig. 19-11 An representation of the band-gap pattern for a semiconductor

The revolution in microelectronics that has so influenced our lives is based on semiconductors, so we would do well to learn more about them in the next section.

19.4 Doping and p-n Junction

19.4.1 Doping

Crystals of pure germanium are well known as a semiconductor material. Each atom of germanium has four valence electrons, which form four covalent bonds with neighboring atoms. This structure has the semiconductor arrangement of energy states, a completely filled energy band separated from an empty higher-lying conduction level by a small energy gap. Such a crystal is called as intrinsic semiconductor. At any temperature above 0 K a few electrons in the semiconductor are thermally excited to the conduction band. The minimum energy required to move an electron from one of the covalent bonds into the conduction band is called the gap E_g. For silicon E_g is 1.09 eV; for germanium, E_g is 0.72 eV.

The versatility of semiconductors can be marvelously improved by introducing a small number of suitable replacement atoms into the semiconductor lattice. This process is called doping. The doped semiconductor becomes an impurity semiconductor. Essentially all practical semiconducting devices today are based on doped crystals in which the electric properties are drastically changed by the impurity atoms. They are of two varieties, called n-type and p-type.

Addition of foreign phosphorus atoms to an intrinsic semiconductor produces an n-type semiconductor. As shown in Fig. 19-12, arsenic impurity atoms occupy sites in the crystal normally filled by germanium atoms. Each arsenic atom brings a total of five valence electrons, of which four are used in forming covalent bonds with the four nearest germanium neighbors. The fifth electron finds no convenient neighbor with which to form a bond and is very loosely tied to the arsenic ion, even at 0 K. Only a small amount of energy is needed to free this electron and permit it to wander about in the lattice. Room temperature is enough to supply this energy. Almost every arsenic atom introduced into the crystal provides one electron that is available for conduction. The arsenic atoms are called donors, since each atom donates an electron to the crystal. The conduction of n-type semiconductor is primarily by electrons provided by the donor atoms. The negative charge carriers in n-type semiconductor greatly outnumber the positive carriers. The former, called the majority carriers, are the electrons in the conduction band. The latter, called the minority carriers, are the holes in the valence band.

Next consider what happens when we add to a germanium crystal atom of valence 3, such as indium (or gallium, aluminum, boron). Each indium atom in Fig. 19-13 brings only three electrons to form the four covalent bonds expected from each atom in the germanium crystal, producing a site at which one electron is missing from the four bonds. It is easy for the indium ion core to "steal" or " capture" a valence electron from a nearly silicon atom, and utilizes it to complete the bonding structure, thus creating a hole in the valence band. Atoms of valence 3 elements are called the acceptor atoms because they so readily accept electrons from the valence band. They form p-type semiconductor, since in a crystal with such impurities there are sites at which an electron is missing and there are positive holes available for conduction. The positive charge carriers in p-type semiconductor greatly outnumber the negative carriers. In p-type semiconductors the majority carriers are the holes in the valence band and the minority carriers are the electrons in the conduction band.

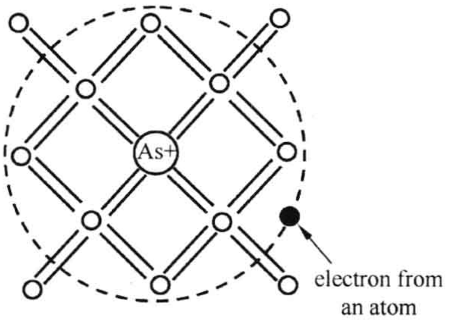


Fig. 19-12 A two-dimensional representation of a doped germanium crystal in which the arsenic atoms are added

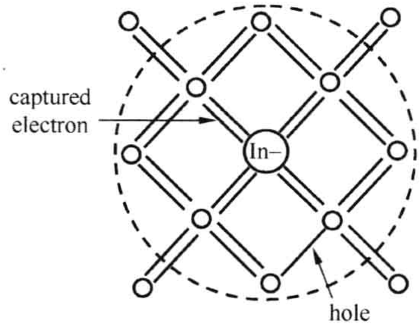


Fig. 19-13 A two-dimensional representation of a doped germanium crystal in which the indium atoms are added

19.4.2 p-n Junction

If a region of n-type material is in intimate contact with a region of p-type material, they form a p-n junction. Such junctions are of great importance, since they are the building blocks for semiconductor devices. One cannot form a p-n junction by pressing n-type material against p-type material. One way of making a p-n junction is to melt an indium pellet which has been placed on top of an n-type germanium crystal where the germanium has been saturated with molten indium, p-type material is produced. A p-n junction is formed at the interface between the indium-saturated p region and the original n material.

Fig. 19-14 represents a p-n junction at the imagined moment of its creation. Let us first discuss the motions of the majority carriers. At a p-n junction, where n-type and p-type material are in intimate contact, electrons from the donor atoms near the junction fill the nearly acceptor states just across the junction. This makes the p material negative, and the loss of electrons in the n region leaves it positively charged. In the immediate vicinity of the junction there is a scarcity of free carriers of either sign, because the electrons from the n side have filled acceptor states and holes on the p side. Thus at the junction we have a depletion layer perhaps 20 nm ($\approx$50 atom layers) thick which is a relatively poor conductor together with a double layer of bound charger, negative on the p side and positive on the n side. These fixed charges (negative and positive ions) cause a contact potential difference to build up across the junction, as Fig. 19-15 shows. This potential difference is such that it serves as a barrier to limit further diffusion of electrons. Thus, at equilibrium, a p-n junction resting on a shelf develops a contact potential difference V_0 between its ends.

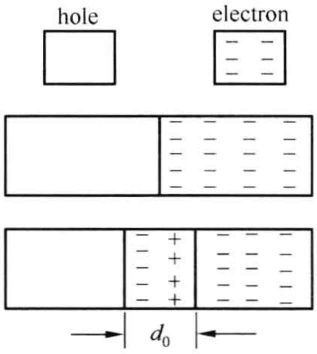


Fig. 19-14 A p-n junction at the imagined moment of its creation

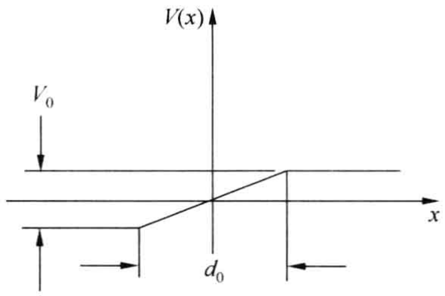


Fig. 19-15 The space charge established a contact potential difference V_0 across the junction plane

New let us turn our attention to the minority carriers. Although the majority carriers in n-type material are electrons, there are nevertheless also a few holes, the minority carriers. Likewise in p-type material, although the majority carriers are holes there are also a few conduction electrons. The potential difference V_0, being a barrier for majority, is a downhill trip for the minority carriers, be they electron or holes. Hence, the contact potential difference causes the minority carriers (electrons or holes) to drift steadily across the junction plane.

The current represented by the motions of minority carriers, called the drift current i_{drift}, is in the opposite direction to the diffusion current and just compensates it at equilibrium. So at equilibrium, there also are charge transfers across the junction but the charge transfers in both directions are equal.

When we apply a potential difference across the p-n junction, making the p-type material positive relative to the n-type material, the hill down which the electron slides is reduced. This has essentially no effect on the saturation current i_{drift} due to electrons going from the p to n side. However many more electrons from the n-type material diffuse to the p-type material, so the electrons flow from the n to p side is greatly increased by this bias. The current in what is known as the forward direction rises rapidly with increasing forward bias (or potential difference), as shown in Fig. 19-16. On the other hand, if we make the n-type material positive relative to the p-type material, the hill is increased and electron flow from n to p is substantially reduced while the current i_{drift} remains the same. Consequently, with negative on reverse bias there is a very small current essentially independent of bias. Thus a p-n junction acts as a rectifier. A p-n junction is basically a two-terminal rectifier.

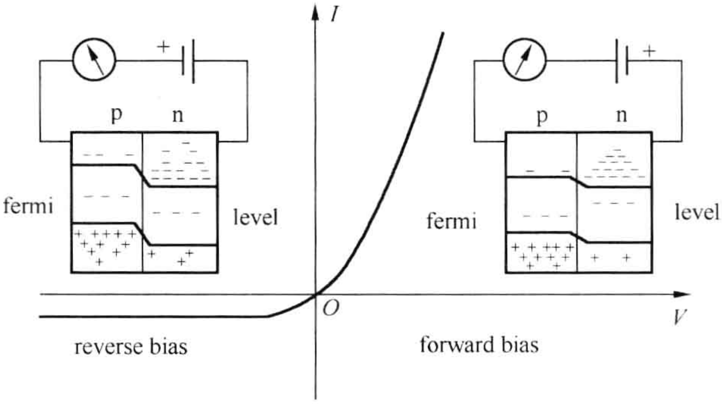


Fig. 19-16 How current varies with bias potential difference across a p-n junction

19.5 Laser

A device that produces a beam of coherent light by stimulating emission from a large population of excited radiators is called a laser, an acronym for "light amplification by stimulated emission of radiation". The first stimulated emission amplifiers operated in the microwave region and received the name masers with the m coming from microwave, but maser is now a general term for stimulated-emission devices which operate over a broad spectral range, including microwave, infrared, visible, and ultraviolet frequencies. For example, lasers are often called optical masers. Now, the laser is leading the way in a new field-sometimes called photonics-which deals with the interaction between photons and bulk matter. Due to the intense coherent nature of the radiation, the lasers have led to increasing application of lasers in fields such as radio astronomy, microwave spectroscopy, photography, biophysics, communications, drilling, cutting, spot welding, precision surveying and other applications too numerous to list.

19.5.1 The stimulated emission

In the chapter 18, we have discussed the absorption processes and spontaneous emissions of atoms. New let us study the stimulated emissions of atoms, which were introduced into physics in 1917 by Einstein. In Fig. 19-17 we illustrate schematically the three transition processes of an atom, namely the spontaneous emission, stimulated absorption and stimulated emission. In the spontaneous emission process, the atom is initially in the upper state of energy E_2 and decays to the lower state of energy E_1 by the emission of a photon of frequency $\nu = (E_2 - E_1)/h$. In the stimulated absorption process, an incident photon of frequency ν, from an electromagnetic field applied to the atom, stimulates the atom to make a transition from the lower to the high energy state, the photon being absorbed by the atom. In the stimulated emission process, an incident photon of frequency ν stimulates the atom to make a transition from the higher state energy to the lower and the atom is left in this lower state at the emergence of two photons, the incident one and the emitted one. The emitted photon is in every way identical to the stimulating photon. It has the same energy (or frequency), direction, phase, and polarization. In the stimulated emission process, one photon comes in and two photons come out. Laser light is produced in this way. The spontaneous emission process, however, is a strictly quantum effect. In all three processes, we have to deal with the interaction of radiation with the atom. It can be proved that for atoms in thermal equilibrium with the radiation, spontaneous emission is far more probable than stimulated emission if $E_2 - E_1 \ll kT$ and the stimulated emission can be ignored. Stimulated emission can become significant, however, if $E_2 - E_1 \approx kT$, and it may be dominant if $E_2 - E_1 \gg kT$.

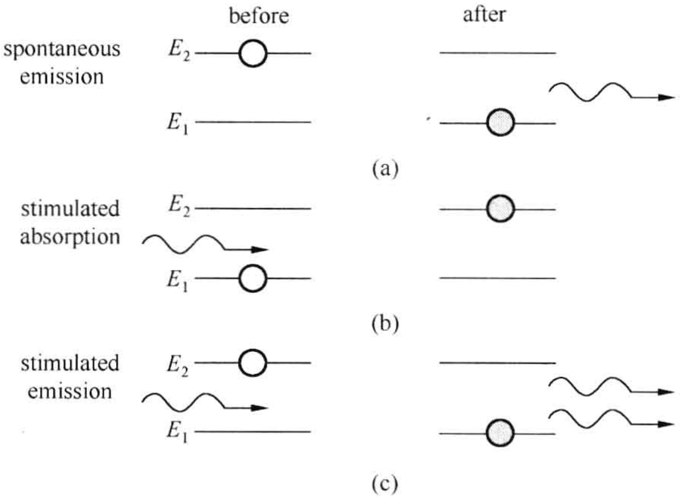


Fig. 19-17 Illustrating: (a) the spontaneous emission process; (b) the stimulated absorption process; (c) the stimulated emission process for two energy states of an atom

Let us consider a system consisting of N same atoms. If $E_2 - E_1 \ll kT$, it can be proved that

$$\frac{\text{rate of emission}}{\text{rate of absorption}} \approx \frac{n_2}{n_1}$$

where n_1 is the number of atoms in state 1 and n_2 is the number of atoms in state 2 of the system. In situations of thermal equilibrium, we have $n_2 < n_1$ and therefore, when the system is in the electromagnetic field, the net effect will be absorption. If now we have a means of inverting the normal population of states so that $n_2 > n_1$, then the emission would exceed the absorption rate. This means that the applied radiation of frequency $\nu = (E_2 - E_1)/h$ will be amplified in intensity by the interaction process, more such radiation emerging than entering. Of course, such a process will reduce the population of the upper state until equilibrium is re-established. In order to sustain the process, therefore, we must use some method to maintain the population inversion of the states.

19.5.2 Lasers

Generally speaking, a laser is a device in which a material is prepared so that the higher of two energy levels is more highly populated than the lower energy level, the material being enclosed in an appropriate resonator of sharp response. The system produces coherent radiation at those frequencies common to the resonator and the difference in energy of the levels.

To produce the laser, it is necessary ① to produce an unnaturally large number of excited atoms (or molecules) ready to emit the desired radiation and ② to stimulate the emission in such a way as to obtain a coherent beam in the desired direction. The process of providing the required number of excited atoms in a laser is known as pumping. In gas lasers pumping comes from an electric discharge through the gas, and in semiconductor lasers passing a current through a p-n junction provides the pumping action. The upper state of the states between which transitions are made must be a metastate, whose relatively long lifetime allows it to be highly populated (that is to maintain the population inversion of the states). The lower state is the ground state of infinitely long lifetime. From the uncertainty relation $\Delta E \Delta t \approx \hbar$, with Δt equal to the long lifetime of the upper state, we conclude that the energy uncertainty in the energy difference of states is small and the emitted transition frequency is sharp, giving a highly monochromatic beam.

Figure 19-18 shows the solid state laser that operates with a ruby crystal. Some Al atoms in the Al_2O_3 molecules are replaced by Cr atoms which account for the laser action. The level of energy E_1 is the ground and the level of energy E_3 is the unstable upper state (that is really a multiplet) with a short lifetime (10^{-8} sec.), the energy difference $E_3 - E_1$ corresponding to a wavelength of about 5500 Å. Level E_2 is an intermediate excited state which is metastable, its lifetime against spontaneous decay being about 3×10^{-3} s. If the chromium atoms are in thermal equilibrium, the population numbers of the states are such that $n_3 < n_2 < n_1$. By pumping in radiation of wavelength 5500 Å, however, we stimulate absorption of incoming photons by Cr atoms in the ground state, thereby raising the population number of the energy state E_3 and depleting energy state E_1 of occupants. Spontaneous emission, bringing atoms from state 3 to state 2, then enhances the occupancy of state 2, which is relatively long-lived. The result of this optical pumping is to decrease n_1 and increase n_2, so that $n_2 > n_1$ and population inversion exists. Now, when an atom does make a transition from state 2 to state 1, the emitted photon of the wavelength 6943 Å will stimulate further transitions. Stimulated emission will dominate stimulated absorption (because $n_2 > n_1$) and the output of photons of wavelength 6943 Å is much enhanced. We obtain an intensified coherent monochromatic beam.

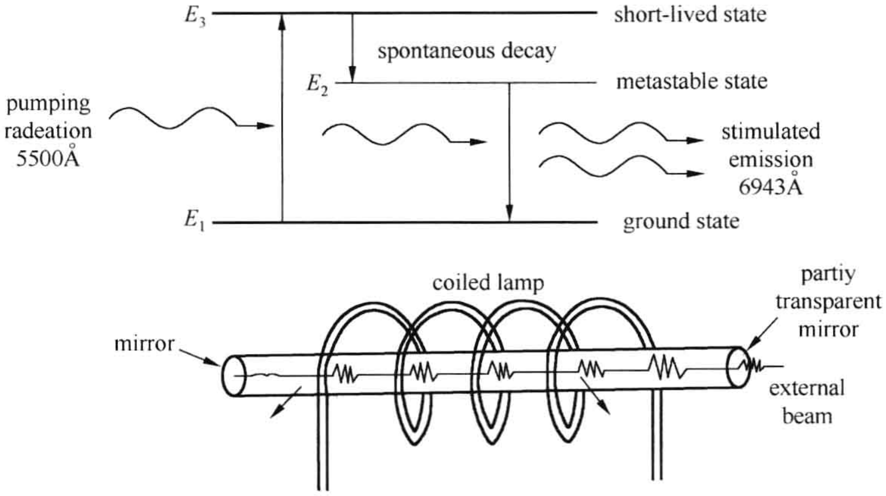


Fig. 19-18 Top: the relevant energy levels of chromium atoms in a ruby laser. Bottom: a schematic of the ruby laser

In practice, the ruby laser is a cylindrical rod with parallel, optically flat reflecting mirrors M_1 and M_2, one of which is only partly reflecting as shown in figure 19-18. The emitted photons that do not travel along the axis escape through the sides before they are able to cause much stimulated emission. But those stimulated photons that are parallel to the axis can be made to move back and forth through the rod by successive reflections from mirrors M_1 and M_2, and they are capable of stimulating emission repeatedly. It is perhaps more useful to think of the entire arrangement of mirrors M_1 and M_2 as an optical resonant cavity and the sharp resonance at one (or more) wavelength can be built. Thus the number of photons is increased rapidly in the direction of axis, those escaping from the partially reflecting mirror giving a unidirectional beam of great intensity and sharply defined wavelength.

Here, we list some of the important characteristics of laser light:

(1) laser light is highly monochromatic;

(2) laser light is highly coherent, wave trains for which may be several hundred kilometers long;

(3) laser light is highly directional;

(4) laser light can be sharply focused, energy flux densities for which of $\sim 10^{16}$ W/cm^2 are readily possible.

There is now a wide variety of lasers—gas lasers, liquid lasers, and solid state lasers—covering various regions of the electromagnetic spectrum. These devices have found many uses and opened the door to a whole new world of research and application.

Questions

19-1 It is often said that a crystal is one giant molecule. Explain. Can we regard a diatomic molecule as a small solid?

19-2 Why does metallic binding usually occur with atoms having a small number of valence electrons?

19-3 Of the five types of binding in solids discussed in the text, which one (or ones) is most likely to produce an insulator? A conductor? A semiconductor?

19-4 How is the optical transparency of a semiconductor related to the energy gap of the forbidden band ?

19-5 Does the Fermi energy for a given metal depend on the volume of the sample?

19-6 Why do the curves in Fig. 19-8 differ so little from each other?

19-7 What role does the Pauli exclusion principle play in accounting for the electrical conductivity of a metal ?

19-8 At room temperature a given applied electric field will generate a drift speed for the conduction electrons of silicon that is about 40 times as great as that for the conduction electrons of copper. Why isn't silicon a better conductor of electricity than copper?

19-9 How does the number of conduction electrons compare with the number of holes in (1) an ideal intrinsic semiconductor, (2) a p-type semiconductor, and (3) an n-type semiconductor ? Explain.

19-10 Does an ordinary semiconductor at constant temperature obey Ohm's law? Does a p-n junction? Explain.

19-11 How do you account for the fact that the resistivity of metals increases with temperature but that of semiconductors decreases?

19-12 Identify the following as p-type or n-type semiconductors: (1) Sb in Si ; (2) In in Ge ; (3) Al in Ge ; (4) P in Si.

19-13 What is the distinction between coherent and incoherent radiation?

19-14 What are some of the advantages of lasers and masers over the more familiar light sources? What possible uses and applications can you envision for lasers and masers?

19-15 Why is a population inversion between two atomic levels necessary for laser action to occur?

19-16 What is a metastable state? What role do such states play in the operation of a laser?

19-17 Why is it difficult to build an X-ray laser?

Appendix 1

The International System of Units

1. The SI Base Units

Quantity	Name	Symbol	Definition
length	meter	m	It is defined as the length of the path travelled by light in vacuum in 1/299792458 of a second(1983)
mass	kilogram	kg	The kilogram is defined as being equal to the mass of the International Prototype Kilogram(a certain platinum-iridium cylinder, 1889)
time	second	s	The second has been defined to be the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom(1967)
electric current	Ampere	A	The constant current that will produce an attractive force of 2×10^{-7} Newton per meter of length between two straight, parallel conductors of infinite length and negligible circular cross section placed one metre apart in a vacuum(1946)
thermodynamic temperature	Kelvin	K	The kelvin is defined as the fraction 1/273. 16 of the thermodynamic temperature of the triple point of water(1967)
amount of substance	mole	mol	The mole is defined as an amount of a substance that contains as many elementary entities as there are atoms in 12 grams of pure carbon-12 (1971)
luminous intensity	candela	cd	The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} Hertz and that has a radiant intensity in that direction of 1/683 Watt per steradian(1979)

2. Some SI Derived Units

Quantity	Name of unit	Symbol
area	square meter	m^2
volume	cubic meter	m^3

table continued

Quantity	Name of unit	Symbol
frequency	Hertz	Hz s^{-1}
mass density	kilogram per cubic meter	kg/m^3
speed, velocity	meter per second	m/s
angular velocity	radian per second	rad/s
acceleration	meter per second squared	m/s^2
angular acceleration	radian per second squared	rad/s^2
force	Newton	N $kg \cdot m/s^2$
pressure	Pascal	Pa N/m^2
work, energy, heat	Joule	J $N \cdot m$
power	Watt	W J/s
quantity of electricity	Coulomb	C $A \cdot s$
potential difference electromotive force	Volt	V W/A
electric field strength	Volt per meter	V/m
electric resistance	Ohm	Ω V/A
capacitance	Farad	F $A \cdot s/V$
magnetic flux	Weber	Wb $V \cdot s$
inductance	Henry	H $V \cdot s/A$
magnetic field	tesla	T Wb/m^2
magnetic field strength	ampere per meter	A/m
entropy	joule per kelvin	J/K
specific heat capacity	joule per kilogram kelvin	$J/(kg \cdot K)$
thermal conductivity	watt per meter kelvin	$J/(m \cdot K)$
radiant intensity	watt per steradian	W/sr

Appendix 2

Some Fundamental Constants of Physics*

Constant	Symbol	Values and units
speed of light in a vacuum	c	299 792 458 m/s
vacuum permeability	μ_0	$12.566\ 370\ 61\times10^{-7}$ N/A^2
vacuum permittivity	ε_0	$8.854\ 187\ 817\times10^{-12}$ F/m
gravitational constant	G	$6.67384(80)\times10^{-11}$ m^3/(kg · s^2)
Planck constant	h	$6.626\ 069\ 57(29)\times10^{-34}$ J · s
reduced Planck constant	$\hbar$	$1.054\ 571\ 726(47)\times10^{-34}$ J · s
elementary charge	e	$1.602\ 176\ 565(35)\times10^{-19}$ C
electron mass	m_e	$9.109\ 382\ 91(40)\times10^{-31}$ kg
proton mass	m_p	$1.672\ 621\ 777(74)\times10^{-27}$ kg
proton-electron mass ratio	m_p/m_e	$1836.152\ 672\ 45(75)$
neutron mass	m_n	$1.674\ 927\ 351(74)\times10^{-27}$ kg
Avogadro constant	N_A	$6.02214129(27)\times10^{23}$ mol^{-1}
universal gas constant	R	$8.314\ 4621(75)$ J/(K · mol)
Boltzmann constant	k	$1.380\ 6488(13)\times10^{-23}$ J/K
atomic mass unit	u	$1.660\ 538\ 921(73)\times10^{-27}$ kg
Rydberg constant	R	$10\ 973\ 731.568\ 539(55)$ m^{-1}

---

* The 2010 CODATA recommended values.

Words

Chapter 10

ammonia	氨（水）
argument	辩论,证论
associated	关联的,连带的
aviation	飞行,航空
barometer	气压计
bombard	炮击,轰炸
botanist	植物学家
bump	撞,碰撞
calibrate	校准
climb	攀登
compressor	压缩机
condensation	浓缩,凝固
crucial	极重要的,决定性的
cylinder	汽缸
diatomic	双原子的
diffusion	扩散
dimensionless	无量纲
discrete	分立的
dissolution	分解,溶解
disturbance	扰动
drumbeat	敲击,敲鼓
equilibrium	平衡,均衡
equipartition	均分,均等
erratic	无常的,不稳定的
excessively	过分,过度
expansion	膨胀
fluctuation	涨落,起伏
gauge	测量表,口径
grain	颗粒,微粒
hypothesis	假说,假设
intrinsically	内在地,固有地
ionization gauge	离子真空压力计
ionization	电离,游离
jaggedly	锯齿状
methane	甲烷
monatomic	单原子的
normalizing	归一化
pointlike	点状的
pollen	花粉
polyatomic	多原子的
probable speed	最概然速率
realm	领域,领土
repulsive	排斥的
riddle	筛子,探究
temperature	温度
thermometer	温度计
Tibet	西藏
vaporization	蒸发
vessel	容器,船,舰
vigorous	有活力的,强健的
withdraw	取出,收回
zigzag	曲折的,锯齿形

Chapter 11

adiabatic	绝热的
adiabat	绝热线
astonish	使吃惊
barrel	桶,圆桶
briskly	轻快地
by virtue of	依靠,由于
claimant	要求者
combustion	燃烧
compression	压缩,密集
contradiction	矛盾
criterion	判据,标准,准则
cyclical	循环的
directivity	方向性
dissipate	消耗
entropy	熵
evacuate	抽空

exhaust value	排气阀	moderate	中等的,适度的
expenditure	消耗,支出	perceive	发觉,看出
gasoline	汽油	permutation	置换,互换
ignite	点火	perpetual motion	永动
infinitesimal	无限小,微元	propel	推进
intake value	进气阀	quantum theory	量子理论
interstellar	星际的	rarefaction	稀疏
intrinsic	内在的,固有的	recourse	求助,依靠
irrespective	不考虑的,不顾	reservoir	库,存储器
irrelative	不相干的	schematic	示意
irreversible	不可逆的	seep	泄漏
isobaric	等压的	steep	陡的
isochoric	等容的	suck	吸吮,吸入
isothermal	等温的	superimpose	重叠,附加
knob	旋钮,调节器	swindle	欺骗
lead shots	铅丸	synchronism	同步性
logarithm	对数	take for granted	认为当然
merge	合并	violation	违背

Chapter 12

acoustic	声学的	plasma	等离子体
amplitude	振幅	projection	投影
circumference	圆周	quadrant	象限
critical	临界的	resonance	共振
damped vibration	阻尼振动	response	响应
ellipse	椭圆	restoring force	恢复力
equilibrium	平衡	SHM	简谐振动
execute	实行	simple pendulum	单摆
forced vibration	受迫振动	stiffness	倔强系数
frequency	频率	suspend	吊
immerse	沉浸	swing	摆动
initial phase	初相位	torsion pendulum	扭摆
mechanical vibration	机械振动	transient	瞬变的
overdamped	过阻尼	twist	使扭转
period	周期	underdamping	欠阻尼
phase angle	相位		

Chapter 13

activate	使激活	emanate	放射,发射,发散
adjacent	相邻的	energy flow	能流
alternatively	交替的	half a wavelength	半波长
condensation	密部	homogeneous	均匀的
crest	波峰	Huygens' principle	惠更斯原理
deformation	形变	incident & reflected	入射的和反射的
diffract	衍射,绕射	infrared	红外线
disturbance	扰动	interference	干涉

isotropic	各向同性的
longitudinal	纵向的，经度的
medium	介质
modulus	模量，模数
node & antinode	波节和波腹
obstacle	障碍
partial derivative	偏微分
periodical	周期的
pinch	挤压
pitch	音调，音高，高度
postulate	假说，假设
prominent	突出显著的，凸起的
propagate	传播
rarefaction	疏部
ripple	波纹，脉动
segment	段，节
spherical	球形的，球面的
standing wave	驻波
straightedge	直尺，直边，弄直
strain	应变
stress	应力
string	弦，绳
successively	逐次地，相继地
tensile	张力的，拉力的
transverse	横向的
trough	波谷
trumpet	喇叭
tuning fork	音叉
ultraviolet	紫外线
waveform	波形
wave front	波前
wavefunction	波方程，波函数
wave intensity	波的强度
wavelength	波长
wavelets	子波
whistle	号笛，哨子，振鸣声

Chapter 14

antenna	天线
astronomy	天文学
compensate	补偿
cuboid	长方体
deliver	释放
microwave	微波
oven	烤炉
radioactive	放射性的
spectroscopy	光谱学
spectrum	光谱
spur	刺激
stratosphere	［气］同温层
therapy	治疗
transformer	变压器
transmission	传输
transmitter	发射机
zenith	天顶

Chapter 15

aberration	像差
Airy disk	艾里斑
analyzer	检偏器
angstrom	［物理］埃（长度单位）
anisotropic	各向异性的
antireflecting film	增透膜
aperture	孔，孔径
biaxial	双轴的
Brewster angle	布儒斯特角
bright fringe	明条纹
cadmium	［化］镉
calcite	方解石
celluloid	赛璐珞
cement	胶接剂
chromatic dispersion	色散
component	元件
concave	凹的
conjugate	［物理学］共轭的
contour	轮廓等值线
converging	会聚的
convex	凸的
corrode	腐蚀
curvature	曲率
cylindrical	柱面的
dark fringe	暗条纹
dichroism	二色性
diffraction	衍射
diverging	发散的
dualistic	二元的
elaborate	精心制作的
elliptical surface	椭球面
embed	嵌入

equal-inclination	等倾
extraordinary	非常的
eyepiece	目镜
filament	灯丝
focus	焦点
Fresnel	[人名]菲涅耳
fuzzy	模糊的
geometrical	几何的
glare	眩光
glossy surface	光滑表面
grating	光栅
grind	研磨
groove	槽,沟
homogeneous	均匀的
image	像
incandescent lamp	白炽灯
index of refraction	折射率
interface	界面
interference	干涉
interferometer	干涉仪
isotropic	各向同性的
lateral	横(向)的
magnification	放大倍数
monochromatic	单色的
nanometer	[物理]纳米(长度单位)
normal	法线
oblique	倾斜的
opaque	不透明的
optical-thickness	光学厚度
orientation	方位
paraxial	旁轴的
pincushion	针垫
plane-polarized	平面偏振的
polarization	偏振
polarizer	起偏器
polaroid	偏振片
prism	棱镜
Rayleigh criterion	瑞利判据
resolving power	分辨本领
resonant cavity	谐振腔
sodium	[化]钠
spacing	间隔
spectral line	光谱线
spliter	分光片
stack	堆叠
stress	应力
superimpose	叠加
symmetrical	对称的
thin lens	薄透镜
tourmaline	电气石
transparent	透明的
uniaxial	单轴的
unpolarized light	非偏振光
unpolarized	非偏振的
violet	紫色
virtual	虚的
wavetrain	波列
wedge-shaped film	劈形膜(劈尖)

Chapter 16

alpha particle	α 粒子
assign	指定
complex	复数
consistent	一致的
contract	使缩短
cornerstone	基础
cosmic	宇宙的
departure	违背
dilation	膨胀
ether	以太
fission	裂变
galactic	银河的
homogeneity	均匀
ingredient	成分
intuition	直觉
lifetime	寿命
meterstick	米尺
moving mass	运动质量
muon	μ 介子
neutrino	中微子
paradox	佯谬
photon	光子
proper	固有的
quasar	类星体
radioactive	放射性的
relativity	相对性,相对论
rest mass	静止质量
shrink	收缩
simultaneity	同时性
simultaneously	同时地
time-space interval	时空间隔
transformation	变换
ultimate	极限的

Chapter 17

absorption	吸收
absorptivity	吸收率
ammeter	电流表
billiard	台球的
black body	黑体
catastrophe	灾变
Compton	[人]康普顿
conflict	冲突
corpuscle	微粒
discrepancy	不一致
discrete	分离的
emission	发射
emissivity	发射率
empirical	经验主义的
facilitate	促进
foil	金属薄片
graphite	石墨
incandescent	白热的
milestone	里程碑
photoelectric	光电的
photoelectron	光电子
Planck constant	普朗克常量
potassium	钾
quantization	量子化
radiancy	发射本领
saturation	饱和
stopping potential	遏止电势差
thermal radiation	热辐射

Chapter 18

ad hoc	特别
alkalis	碱金属元素
awesome	可畏的
azimuthal	角的
Bohr's radius	玻尔半径
cease	停
coincide	一致
collimate	使成平行
configuration	组态
crystallite	微晶
De Broglie's postulate	德布罗意假设
degeneracy	简并
deteriorate	变坏
diaphragm	隔板
dislocation	位错
eigenfunction	本征函数
eigenvalue	本征值
eutectic	共晶的
excess	过量
excited state	激发态
exclusion	不相容
fault	断层
ground state	基态
helical	螺旋形的
hydrogenlike	类氢原子
hyperfine	超精细
inadequacy	不充足
inexplicable	费解的
infinite potential well	无限深势阱
inherently	固有地
interplanar	平面间的
ionize	电离
lattice	晶格
morphology	结构
normalize	归一化
orientation	方位
outset	开端
pin	圈住
plum pudding	葡萄干布丁
polycrystalline	多晶的
proton	质子
quantum number	量子数
quantum	量子
realm	范围
Rydberg constant	里德伯常量
screw	螺旋
spin	自旋
stationary	定态
tentative	尝试的
tinker	调整
transcendental	超越的
transition	跃迁
triumph	成功
uncertainty principle	测不准关系
visualize	想象
wave-particle duality	波粒二象性

Chapter 19

acronym	首字母缩写词
adjacent	相邻的
agitation	扰动
alternation	交替
alter	改变
amorphous	非结晶的
Arsenic	砷
bias	偏压
brittle	易碎的
compressible	可压缩的
covalent	共价的
depletion	竭尽
deviation	偏离
dielectric	绝缘的
disruptive	分裂的
doping	掺杂质
focus	集中
gradient	梯度
hypothetical	假设的
imperfection	有缺陷
impurity	不纯
indistinguishable	不可区分的
Indium	铟
interpenetrate	渗透
intimate	密切的
junction	结
laser	激光
marvelously	惊奇地
metastable	亚稳态的
nestle	偎依
noncrystalline	非晶体的
overlap	重叠
parallelepiped	平行六面体
pellet	小球
protein	蛋白质
pumping	泵浦
rectifier	整流器
resonator	谐振腔
retain	保持
ruby	红宝石
scarcity	不足
shatter	破坏
silicon	硅
split	使分裂
spoil	破坏
spontaneous	自发的
stimulate	激励
sustain	维持
twofold	双重
versatility	多面性

Answers to Problems

Chapter 10

10-1 (1) $2.4\times10^{-3}\,\text{m}^3$, $4.14\,\text{kg/m}^3$ (2) 1170K, $1.0\,\text{kg/m}^3$ **10-2** 0.732cm **10-3** 76.7cmHg

10-4 $1.4\times10^5\,\text{N/m}^2$ **10-5** $1890\,\text{N/m}^2$ **10-6** $2nmv^2$ **10-8** 400m/s **10-9** $2.69\times10^{11}/\text{m}^3$

10-10 (1)551m/s (2)32×10^{-3} kg/mol, Oxygen

10-11 (1) $2.45\times10^{25}/\text{m}^3$ (2) $1.3\,\text{kg/m}^3$ (3) $5.3\times10^{25}\,\text{g}$ (4) $3.44\times10^{-9}\,\text{m}$ (5) $6.21\times10^{-21}\,\text{J}$
(6) $1.56\times10^{-25}\,\text{J}$, 0.25×10^{-4}

10-12 $6.21\times10^{-21}\,\text{J}$, 483m/s, 1934m/s **10-13** (1) $3.74\times10^3\,\text{J}$ (2) $2.49\times10^3\,\text{J}$, 20.8J

10-14 $1.60\times10^{-19}\,\text{J}$, $7.74\times10^3\,\text{K}$ **10-15** (1) 12.5J, 20.8J, 24.9J (2) 0.74J, 10.4J

10-16 (1) $5.23\times10^9\,\text{s}^{-1}$ (2) $7.03\times10^{28}\,\text{s}^{-1}$ **10-17** (1)$8.2\times10^{-8}\,\text{m}$ (2) $8.5\times10^{-4}\,\text{m}$

10-18 5.6 km

10-19 (1) average score: 60.6 (2) root mean square score: 63.3
(3)the most probable score: 65.0, the fraction of the number lines in this range: 24.2%

10-21 0.4% **10-22** 1.1 **10-23** 0.83%

10-24 (1)$a=2N/3v_0$ (2) $v_p=v_0$ (3) 33.3%, 50%, 16.7%

10-25 $\mathrm{d}N_{E_k}=\dfrac{2N}{\sqrt{\pi}}(kT)^{-3/2}\sqrt{E_k}\,\mathrm{e}^{-E_k/kT}\,\mathrm{d}E_k$, $E_{kp}=\dfrac{1}{2}kT$, $\dfrac{1}{2}mv_p^2=2E_{kp}$ **10-26** $4.5\times10^7\,\text{s}^{-1}$

10-28 $N_2/N_1=\mathrm{e}^{-394}=2.74^{-394}\to0$ it means that nearly all of the atoms are at the ground state, at room temperature.

10-29 2.3 km **10-30** 510mmHg $=0.68\times10^5\,\text{Pa}$ **10-31** 342K, 608K, 78%

Chapter 11

11-1 169J, 2087J **11-2** 597J **11-3** (1)7928J (2) 5663J (3) 2265J

11-4 (1) 6cal (2) −43cal (3) 10cal (4) 18cal, 18cal

11-5 (1) 1500 J, 900 J, 600 J (2) 1830 J, 1230 J, 600 J

11-6 $2.26\times10^5\,\text{J}$ **11-7** (2) 872 J, 249 J, 623 J; 1245 J, 0, 1245 J; 691 J, 691 J, 0

11-9 (2) 0.25 L, 1.32 atm **11-10** (1) 5 atm, 273 K, −1141 K; (2) 9.5 atm, 250 K, −1595 J

11-11 (1) 70.2 J, 70.2 J, 0 (2) 0, 61.3 J, −61.3 J **11-13** (2) 7200 J (3) 960 J (4) 13%

11-14 (2)21% **11-15** 25%, 627J, 1881J **11-16** (1) 26.8% (2) $27.42\times10^4\,\text{J}$ (3) $20.05\times10^4\,\text{J}$

11-17 (1) 49.4kcal (2) 31.0 kJ **11-18** (1) 3.3×10^4 J (2) 3.67×10^5 J

11-19 (1) 7.14 J (2) 200 J (3) 500 J **11-20** 2092 J

11-24 There are 7 possible macro-states: (6, 0), (5, 1), (4, 2), (3, 3), (2, 4), (1, 5), (0, 6). The corresponding numbers of microstates are 1, 6, 15, 20, 15, 6, 1 respectively.

11-25 108.9 J/K **11-26** 5.4 J/K **11-27** 17.3 J/K

Chapter 12

12-1 $x=3.00\times10^{-2}\cos\left(\frac{2\pi}{5}t+\frac{\pi}{4}\right)\text{m}$, $v=-3.77\times10^{-2}\sin\left(\frac{2\pi}{5}t+\frac{\pi}{4}\right)\text{m/s}$

$a=-4.74\times10^{-2}\cos\left(\frac{2\pi}{5}t+\frac{\pi}{4}\right)\text{m/s}^2$

12-2 (1) 0.140m, $-\pi/4$ (2) 4.44m/s, 140m/s², 140 N (3) $x = 0.141\cos\left[10\pi t + \left(-\frac{\pi}{4}\right)\right]$m

12-3 (1) 25×10 N, (2) 0.28 s **12-4** 2s, 0.25m, −1.4m/s, −2.5m/s², $7\pi/3$

12-5 (a) 0 or 2π (b) $\pi/2$ (c) π (d) $-\pi/2$

12-6 (1) 2.5 N/m (2) 1.2×10^{-2}m (3) 2 Hz

12-7 (1) 4 s (2) $x = 3.7\times10^{-2}\cos(0.5\pi t)$m, 2.9×10^{-2}m/s²

12-8 (1) $x=0.120\cos(\pi t-\pi/3)$m (2) 0.104m, −0.188m/s, −1.03m/s² (3) 0.83 s

12-9 0.196m **12-10** (1) 0.44 s (2) 4.9×10^{-2}m, 0 (3) $x=4.9\times10^{-2}\cos10\sqrt{2}t$ m

12-11 (1) 0.510 m (2) 0.200 m, 3.00 m/s (3) $x=0.510\cos(2\sqrt{10}t-1.20)$m

12-12 $x = 0.100\cos(30t-\pi/2)$m **12-13** $T = 0.54$ s, $x = 0.2\cos(3.7\pi t+\pi/2)$m

12-14 $k_1 = \frac{n+1}{n}k$, $k_2 = (n+1)k$ ($k_1 = k_2 = 2k$ for $n = 1$)

12-15 (1) πs (2) 2×10^{-5} J (3) $\pm7\times10^{-3}$ m

12-16 (1) $T=2\pi\sqrt{\frac{m+M}{k}}$ (2) $E=\frac{m^2v^2}{2(m+M)}$, $A=\frac{mv}{\sqrt{k(m+M)}}$

12-18 $E_k = 3E/4$, $E_p = E/4$ **12-19** $x = 0.078\cos(10t+1.48)$ (SI)

12-20 $x = 2A\cos(\omega t+\pi/3)$

Chapter 13

13-1 (1) 0.567 m, 2.5 m, 8.83 m (2) 1.7×10^{-3}m, 7.5×10^{-3}m, 2.65×10^{-2}m

13-2 (1) $y=A\cos\left(100\pi-\pi x-\frac{\pi}{2}\right)$(SI) (2) $y_5=A\cos(100\pi t-5.5\pi)$(SI) (3) π

13-3 $y = 0.1\cos\left[4\pi\left(t+\frac{x}{2}\right)+\frac{\pi}{2}\right]$ m

13-4 (1) 2 m, 0.8s, 2.5 m/s (2) $y_1=0.2\cos(2.5\pi t-\pi)$m (3) $x=2$ m

13-5 (1) $y=0.2\cos(60+6x)$m

(2) $x_c=\left(k\frac{\pi}{3}-10\right)$m, $k=0,\pm1,\pm2,\cdots$; $x_t=\left(\frac{2k+1}{6}\pi-10\right)$m, $k=0,\pm1,\pm2\cdots$

(3) $v=5/3$ m/s$=1.67$ m/s, $t_c=\left(\frac{k\pi}{5}-6\right)$s, $k=10,11,12,\cdots$

13-6 (1) $y=A\cos\left(\omega t-\frac{2\pi}{\lambda}x+\frac{\pi}{2}\right)$(SI)

(2) $y_T=y_0=A\sin\frac{2\pi}{\lambda}x$(SI), $y_{5T/4}=y_{T/4}=A\cos\left(\frac{2\pi}{\lambda}x-\pi\right)$(SI)

13-7 (1) $y = 3\cos\left(4\pi t+\frac{\pi}{5}x-\pi\right)$(SI), $y_B = 3\cos\left(4\pi t-\frac{4}{5}\pi\right)$(SI)

(2) $y = 3\cos\left(4\pi t-\frac{\pi}{5}x\right)$(SI), $y_B = 3\cos\left(4\pi t-\frac{4}{5}\pi\right)$(SI)

13-8 (1) $y = A\cos\left[\omega\left(t-\frac{x-d}{v}\right)+\phi\right]$ (2) $y = A\cos\left[\omega\left(t+\frac{x+d}{v}\right)+\phi\right]$

13-9 (1) $y_0 = 2\times10^{-2}\cos\left(\frac{\pi}{2}t+\frac{\pi}{3}\right)$ m (2) $y = 2\times10^{-2}\cos\left[2\pi\left(\frac{t}{4}-\frac{x}{4}\right)+\frac{\pi}{3}\right]$m

(3) $y_{t=1} = 2\times10^{-2}\cos\left(\frac{\pi}{2}x-\frac{5}{6}\pi\right)$m

13-10 $y = 0.1\cos\left(7\pi t-\frac{\pi x}{0.12}+\frac{\pi}{3}\right)$m

13-11 (1) $y = A\cos\left[2\pi\left(250t+\frac{x}{200}\right)+\frac{\pi}{4}\right]$m

(2) $y_{100} = A\cos\left(500\pi t+\frac{5}{4}\pi\right)$m, $u_{100} = 500\pi A\sin\left(500\pi t+\frac{1}{4}\pi\right)$m

13-12 (1) a and d up, b and c down (2) $y = 4\times10^{-2}\cos\left(100\pi t+5\pi x-\frac{\pi}{2}\right)$m

13-13 (1) $y = 0.1\cos\left[\pi(t-5x)+\frac{\pi}{2}\right]$m (2) $y_a = 0.1\cos\pi t$ m

13-14 (1) 1.58×10^{5} W/m^2; (2) 3.79×10^{3} J

13-15 (1) 1.27×10^{-2} W/m^2, 3.18×10^{-3} W/m^2 (2) 4.2×10^{-11} J/m^3, 1.06×10^{-11} J/m^3

13-16 (1) 3×10^{-5} J/m^3, 6×10^{-5} J/m^3 (2) 4.62×10^{-7} J

13-17 $x=(10-k)$m $(k=0,\pm1,\pm2,\cdots,\pm9)$ **13-18** $x=\frac{4h^2-k^2\lambda^2}{2k\lambda}, k=\pm1,\pm2,\pm3,\cdots,\pm\left[\frac{2h}{\lambda}\right]$

13-19 (1) $\Delta\varphi = 7.2\pi$ (2) $\varphi=-72^{\circ}=-0.4\pi$

13-20 (1) $A=0$ (2) $2A$

(3) $2A$ where $r=\frac{\lambda}{4}(2k+1)(k=0,1,2,\cdots,8)$, 0 where $r=\frac{\lambda}{4}(2k+2)(k=-1, 0, 1, \cdots,7)$

13-21 The reflected waveform is the same as that of incident waveform at time *t*.

13-22 $y_2 = A\cos\left(\omega t+\frac{2\pi}{\lambda}x-\frac{4\pi}{\lambda}L\right)$ (SI)

13-23 (1) $y=0.1\cos\pi x\cos\pi t$ m

(2) $x=k$ m$(k=0,1,2,\cdots)$ for crests; $x=\frac{1}{2}(2k+1)$m$(k=0,1,2,\cdots)$ for troughs

(3) $A=0.081$m for the point $x=1.2$ m

13-24 $\lambda=1$ m, $\nu=25$ Hz, $v=25$ m/s, $A=0.04$ m **13-25** (1) 1107 Hz (2) 912 Hz

13-26 (1) 1214 Hz (2) 823 Hz **13-27** (1) 467 Hz (2) 494 Hz

Chapter 14

14-1 $I_m=0.64$ A **14-2** $I_m=0.15$ A **14-3** 1.1×10^{-4} s **14-4** 2.65×10^{-4} Hz

14-5 8.5×10^{-7} A/m **14-6** $H_y=-5.7\times10^{-3}\cos[\pi 10^{15}(t-x/c)]$

14-7 0.265A/m, along negative x axis **14-8** 1.2×10^{-6} W/m^2

14-9 $E=12.2\sin[1.22\pi\times10^{15}t+\varphi]$V/m, $H=3.25\times10^{-3}\sin[1.22\pi\times10^{15}t+\varphi]$A/m

14-10 $E_0=103.0\times10$ V/m, $B_0=3.42\times10^{-6}$ T **14-11** 300 m

14-12 (1) 3.75×10^{5} m (2) $E_0=8.61$ V/m, $H_0=2.28\times10^{-2}$ A/m

Chapter 15

15-1 24.0° **15-2** 16.7cm, ∞, −10.0cm **15-3** 15.0cm, ∞, −10.0cm

15-4 (1) $\theta_1=24.0^{\circ}$ (2) 5.00×10^{14} Hz, 2.44×10^{8} m/s, 4.88×10^{-7}m (3) 11.1 cm, 11.3 cm

15-5 7.20×10^{-2} mm **15-6** (1) 0.573° (2) 5.00 mm

15-7 3.60×10^{-3} cm **15-8** 6.64×10^{-6} m

15-9 (1) 3.2×10^{-5} rad (2) the 3rd bright fringe (3) the 3rd bright fringes and the 3rd dark fringes

15-10 concave **15-11** 584nm, 417nm, 324nm **15-12** 1.22 **15-13** 1.00 m

15-14 1.003 **15-15** (1) 1.3×10^{-6} m (2) 433 nm **15-16** (1) 0.49 mm (2) 525 nm

15-17 692 nm, 519 nm, 415 nm **15-18** (1) 2.0×10^{-6} m; (2) 6 **15-19** 7 bright fringes

15-20 (1) 2.24×10^{-4} rad (2) 5.80 km **15-21** 6.69×10^{-2} nm **15-22** 71.6°

15-23 (1) 54.7° (2) 35.3° **15-24** 21% **15-25** $\frac{I_{\text{polarized}}}{I_{\text{unpolarized}}}=2$

15-26 (1) It is not diamond (2) 55.3° **15-27** (1) 32° (2) 1.6

Chapter 16

16-1 138 km, −372 μs **16-2** 0.0, 2.5 μs, event 2 earlier for S'.

16-3 (1) 29.2 μs (2) The blue first **16-4** (1) $-0.84c$ (2) The red first (3) 4.38 μs

16-5 $0.99c$ **16-6** 4.45×10^{-3} s **16-7** (1) $\sqrt{3}/2c$ (2) 2

16-8 (1) 26.0 years (2) 52.0 years (3) 26.0 years

16-9 (2) 0.999998c **16-10** 0.97c

16-11 $E=2.29m_0c^2$, $E_k=1.29m_0c^2$, $m=2.29\ m_0$, $p=2.06m_0c$

16-12 6.3×10^8 V, $p=1.33m_pc$, $E=1.67m_pc^2$

16-13 (1) 0.94c (2) 0.87c **16-14** $2\sqrt{2}m_0c$ **16-15** (1) $0.002m_0c^2$ (2) $2.064\ m_0c^2$

Chapter 17

17-1 483 nm **17-2** (1) 408×10^9 kg (2) 6.4×10^{-14} **17-3** 936.35 nm **17-4** (1) 545.4 nm

17-5 (1) 2.01 eV (2) 0 (3) 2.01 (4) 295.6 nm **17-6** (1) 1.82 eV (2) 383 nm

17-8 2.245×10^{19} **17-9** 0.053×10^{-10} m, 0.074×10^{-10} m, 0.098×10^{-10} m

17-10 (1) 2.403×10^{-10} m (2) 2.436×10^{-10} m **17-11** 3λ **17-12** 180°, $2\lambda_c$

Chapter 18

18-2 121.5 nm, 102.5 nm, 912 nm, 0.2 eV, 12.1 eV, 13.6 eV **18-3** 2.18×10^6 m/s

18-4 921 nm, 13.46 eV, 7.18×10^{-27}, 4.3 m/s **18-5** 912 nm, 828 nm, 228 nm

18-6 0.85 eV **18-7** 1.5×10^6 m/s **18-8** (1) 1.65×10^{-35} m

18-9 (1) 3.31×10^{-24} (2) 9.93×10^{-16} (3) 8.18×10^{-14}, 6/1000 **18-10** 1.096×10^{-16} m

18-11 (1) 5×10^{-7} m (2) 5 mm (3) 5 m **18-12** 1.83×10^{-22} m^3

18-14 $A\mathrm{e}^{-\mathrm{i}Et/\hbar}\mathrm{e}^{\mathrm{i}p_xx/\hbar}$ **18-15** $\sqrt{2/a}$ **18-17** 0.40 **18-18** 14

```
{
  "filename": "5aSn5a2m54mp55CG566A5piO5pWZ56iL77yI6Iux5paH54mI77yJ56ysMueJiCDkuIvlhozvvJ1BIENPTkNJU0UgQ09VUlNFIElOIFVOSVZFUlNJVFkgUEhZU0lDUyBTRUNPTkQgRURJVE9OIFZPTFVNRSAyXzEzMjY3MzE4LnppcA==",
  "filename_decoded": "\u5927\u5b66\u7269\u7406\u7b80\u660e\u6559\u7a0b\uff08\u82f1\u6587\u7248\uff09\u7b2c2\u7248\u4e0b\u518c\uff1dA CONCISE COURSE IN UNIVERSITY PHYSICS SECOND EDITON VOLUME 2_13267318.zip",
  "filesize": 71128229,
  "md5": "f52eca04449c35020377a88a9c93cf07",
  "header_md5": "c7c693d44049b175eb122849b014b5ae",
  "sha1": "fc67972cde6cb134d020045e9b47feeecbe40e82",
  "sha256": "db4f2d9892213e269cf14748d8860c44bcab4e1f13928d7d432ffdf213fb4ce6",
  "crc32": 1427503803,
  "zip_password": "",
  "uncompressed_size": 81262870,
  "pdg_dir_name": "\u2524\u2264\u2564\u00ba\u256c\u2229\u2514\u03c6\u255d\u2265\u251c\u2248\u255c\u2560\u2502\u2560\u00fa\u00bf\u2559\u00f3\u256c\u2500\u2591\u00b5\u00fa\u2310\u2561\u250c2\u2591\u00b5 \u2567\u252c\u2593\u00df\u00fa\u255cA CONCISE COURSE IN UNIVERSITY PHYSICS SECOND EDITON VOLUME 2_13267318",
  "pdg_main_pages_found": 279,
  "pdg_main_pages_max": 279,
  "total_pages": 289,
  "total_pixels": 1803158104,
  "pdf_generation_missing_pages": false
}
```