# UGPHS-07

# BLOCK-1

# INTRODUCING LIGHT

Uttar Pradesh
Rajarshi Tandon Open University

# UGPHS-07
# Optics

Block

# 1

## INTRODUCING LIGHT

# COURSE INTRODUCTION

We are able to adore the wonders of nature and its creations by perceiving light through one of our sense organs. In a way, light sustains life on our planet. Though we see objects it illuminates, we cannot see light! The study of interaction of light with matter constitutes what we call optics. It is one of the most fascinating courses taught to undergraduate sciencc students. Optical studies have contributed significantly to human understanding of the laws of nature. While studying this course you will realise that there is explosive growth of this subject due to the realisation of some well known physical principles for technical applications. This is why optics occupies a prominent place in pure and applied sciences.

The subject of optics emerged as a result of the fundamental work done by scientists of eminence such as Galileo, Abbe, Newton, Huygens, Young, Fresnel, Fraunhofer, Grimaldi, Arago and Bartholims. Maxwell provided a sound mathematical basis to classical optics. Hertz qualified his work successfully. In India, Sir J.C. Bose and Sir C.V. Raman made significant contributions.

This course begins with nature of light. It is intended to establish the **transverse electromagnetic** nature of light. The phenomena of Interference and Diffraction which reveal the wave behaviour of light are discussed in two subsequent blocks. The mathematical treatment has been kept as simple as possible. For instance, without introducing Fresnel integrals, we have tried to provide insight into what is most essential in the theory of diffraction by using Fresnel zones. Similarly, we have refrained from entering into technical details of methods of observation or instrumental appliances. The development of lasers, fibre optics, holography and the progress made in optical communication, optical storage and optical computing with applications in space, defence and medicine have led to an explosive growth of optics in recent years. You will get a glimpse of some of these topics in the last block.

One last word about how to study the course material.

**Study Guide**

This is a four credit course and you have to put in 120 hours of work. Of these, you should spend about 90 hours to study course materials and solve SAQs and TQs. Some of the SAQs are quite revealing and answering these will bring joy. Answers to all questions are given at the end of each unit. But you are advised to do them yourself. We hope that you will enjoy the subject.

We wish you success.

**Acknowledgements**

We thank Shri Sunder Singh and Gopal Krishna Arora for secretarial assistance.

# BLOCK 1 INTRODUCING LIGHT

This block is intended to introduce light. From your previous physics courses, you may be familiar with some of the topics included here. But we have done so to make the block self- contained. In Unit 1 we have shown that light is a transverse electromagnetic wave. The wave equations for **E** and **B** are derived from Maxwell's field equations. In Unit 2 we have discussed reflection and refraction of e.m. waves. You will also learn that all laws of geometrical optics are inherent in Fermat's principle.

Perception of light by humans is discussed in Unit 3. You will learn that human vision involves a mix of physical and physiological processes. The role of eye as an image forming device is discussed in detail. Theories of colour vision are also given in brief. Unit 4 discusses three polarisation states of light. You will learn that light can be polarised by reflection, refraction and selective absorption. Light propagation in anisotropic crystals and phenomenon of birefringence are discussed in detail.

These units are not of equal length. We are suggesting tentative time budget for study time for each unit:

| | |
|---|---|
| Unit 1 | 4 hours |
| Unit 2 | 5 hours |
| Unit 3 | 4 hours |
| Unit 4 | 7 hours |

However, the actual study time will depend on your flair for basic mathematics. If you had opted for PHE-04 and PHE-07 courses, you will find this block rather easy to comprehend.

# UNIT 1 NATURE OF LIGHT

## Structure

## 1.1 INTRODUCTION

You all know that light is responsible for our intimate contact with the universe through one of our sense organs. We are able to admire the wonders of the world and appreciate the beauty of nature only when there is light. The reds of the sun or the ruby, the greens of the grass or emerald and the blues of the sky or sapphire involve light. In a way light plays a vital role in sustaining life on earth. Even so, we are strangely unaware of its presence. We see not light but objects, (shapes, colours, textures and motion) as constructed by the brain from information received by it.

Have you ever thought : What is light ? How light behaves when it reaches our eyes ? And so on. These questions proved very difficult even for the genius of the class of Newton and Einstein. In fact, search for answers to these gave birth to a new branch of physics: **Optics**, which is extremely relevant to the modern world. It occupies a prominent place in various branches of science, engineering and technology. Optical studies have contributed to our understanding of the laws of nature. With the development of lasers, fibre optics, holography, optical communication and computation, optics has emerged as a fertile area of practical applications. It is therefore important for you to understand the language and vocabulory of optics very thoroughly.

In this unit you will learn some important facts and developments which were made to unfold the nature of light. However, before you do so you should revise second block of PHE-02 course and fourth block of PHE-07 course. In Sec. 1.2 you will learn about corpuscular (particle) model of light. In Sec. 1.4 we have discussed the wave model of light, with particular reference to electromagnetic waves. You may now be tempted to ask: Does light behave like a particle or a Wave ? You will learn that it is like neither!

### Objectives

After going through this unit you should be able to

- name phenomena distinguishing corpuscular and wave models of light
- derive an expression for the velocity of electromagnetic waves
- specify the frequency ranges of different portions of electromagnetic spectrum, and
- explain the importance of Poynting Vector.

## 1.2 THE CORPUSCULAR MODEL

You must have read in your school physics course that corpuscular model is due to Newton. Contrary to this popular belief, the credit should be given to Descartes, although the earliest speculations about light are attributed to Pythagoras.

The corpuscular model is perhaps the simplest of the models of light. According to it, **light consists of minute invisible stream of particles called corpuscles.** A luminous body sends corpuscles out in all directions. These particles travel without being affected by earth's gravitation. Newton emphasized that corpuscles of different sizes stimulate sensation of different colours at the ratina of our eye.

In your physics courses at school you must have learnt about evidences in favour of this model. Can you recall them? The two most important experimental evidences are:

(i) Light travels in straight lines. This rectilinear propagation of light is responsible for formation of sharp (perfectly dark) shadows. If we illuminate a barrier in front of a white screen, the region of screen behind the barrier is completely dark and the region outside the barrier is completely lit. This suggests that light does not go around corners. Or does it?

(ii) Light can propagate through vacuum, i.e., light does not require any material medium, as does sound, for propagation.

The speed of propagation of light has been measured by a variety of means. The earliest measurement made by Roemer in 1676 made use of observations of the motion of the moons of Jupiter and apparent variations in the periods of their orbits resulting from the finite speed of propagation of light from Jupiter to earth. The first completely terrestrial measurement of the speed of light was made by Fizeau in 1849.

We can also predict the correct form of the laws of reflection and refraction using the corpuscular model. However, a serious flaw in this theory is encountered in respect of the speed of light. Corpuscular model predicts that light travels faster in a denser medium. This, as you now recognise, contradicts the experimental findings of Fizeau. Do you expect the speed of light to depend on the nature of the source or the medium in which light propagate? Obviously, it is a property of the medium. This means that the speed of light has a definite value for each medium. The other serious flaw in the corpuscular model came in the form of experimental observations like interference (re-distribution of energy in the form of dark and bright or coloured fringes), diffraction (bending around sharp edges) and polarization.

You may now like to answer an SAQ.

**SAQ 1**

*Spend 5 min*

Grimaldi observed that the shadow of a very small circular obstacle placed in the path of light is smaller than its actual size. Discuss how it contradicts corpuscular model.

In the experiment described in SAQ 1, Grimaldi also observed coloured fringes around the shadow. This, as we now know, is a necessary consequence of the wavelike character of light. It is interesting to observe that even though Newton had some wavelike conception of light, he continued to emphasize the particle nature. You will learn about the wave model of light in the following section.

## 1.3 THE WAVE MODEL

The earliest systematic theory of light was put forward by a contemporary of Newton, Christian Huygens. You have learnt about it in PHE-02. Using the wave model, Huygens was able to explain the laws of reflection and refraction. However, the authority and eminence of Newton was so great that no one reposed faith in Huygens' proposition. In fact, wave model was revived and shaped by Young through his interference experiments.

Young showed that the wavelength of visible light lies in the range 4000 Å to 7000 Å (Typical values of wavelength for sound range from 15 cm for a high- pitched whistle to 3 m for a deep male voice.) This explains why the wave character of light goes unnoticed (on a human scale). Interference fringes can be seen only when the spacing between two light sources is of the order of the wavelength of light. That is also why diffraction effects are small and light is said to approximately travel in straight lines. (A ray is defined as the path of energy propagation in the limit of $\lambda \to 0$). A satisfactory explanation of diffraction of light was given by Fresnel on the basis of the wave model. An important part in establishing wave model was played by polarisation- a subtle property of light. It established that light is a transverse wave; the oscillations are perpendicular to the path of propagation. But what is it that oscillates? The answer was provided by Maxwell who provided real physical significance and sound pedestal to the wave theory. Maxwell identified light with electromagnetic waves. A light wave is associated with changing electric and magnetic fields. You will learn these details now.

## 1.4 LIGHT AS AN ELECTROMAGNETIC WAVE

From the PHE-07 course on Electric and Magnetic Phenomena you will recall that a varying electric field gives rise to a time and space varying magnetic field and vice-versa. This interplay of coupled electric and magnetic fields results in the propagation of three-dimensional electromagnetic waves.To show this, we first recall Maxwell's field equations:

$$\nabla . \mathbf{D} = \rho \tag{1.1a}$$

$$\nabla . \mathbf{B} = 0 \tag{1.1b}$$

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 \tag{1.1c}$$

and

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \tag{1.1d}$$

where $\rho$ and **J** denote the free charge density and the conduction current density, respectively. **E, D, B,** and **H** respectively represent the electric field, electric displacement, magnetic induction and the magnetic field. These are connected through the following constitutive relations:

$$\mathbf{D} = \varepsilon \, \mathbf{E} \tag{1.2a}$$

$$\mathbf{B} = \mu \, \mathbf{H} \tag{1.2b}$$

and

$$\mathbf{J} = \sigma \, \mathbf{E} \tag{1.2c}$$

where $\varepsilon$, $\mu$ and $\sigma$ respectively denote the (dielectric) permittivity, magnetic permeability and the electrical conductivity of the medium.

For simplicity, we consider the field equations in vacuum so that $\rho = 0$ and $\mathbf{J} = O$. Then, if we use connecting relations [Eqs. (1.2a-c)], Eq. (1.1a-d) reduce to

$$\nabla . \mathbf{E} = 0 \tag{1.3a}$$

$$\nabla . \mathbf{H} = 0 \tag{1.3b}$$

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \tag{1.3c}$$

and

$$\nabla \times \mathbf{H} = \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \tag{1.3d}$$

where $\mu_0$ and $\varepsilon_0$ are the magnetic permeability and permittivity of free space.

Taking the curl of Eq. (1.3c), we get

$$\nabla \times \nabla \times \mathbf{E} = -\mu_0 \nabla \times (\partial \mathbf{H}/\partial t)$$

$$= -\mu_0 \frac{\partial}{\partial t}(\nabla \times \mathbf{H}) \tag{1.4}$$

since $\frac{\partial}{\partial t}$ is independent of $\nabla\times$ operation.

To simplify the left hand side of this equation, we use the vector identity

$$\nabla \times \nabla \times \mathbf{E} = \nabla(\nabla . \mathbf{E}) - \nabla^2 \mathbf{E}$$

Since $\nabla . \mathbf{E} = 0$ in view of Eq. (1.3a), we find that Eq. (1.4) reduces to

$$-\nabla^2 \mathbf{E} = -\mu_0 \frac{\partial}{\partial t}(\nabla \times \mathbf{H})$$

The 3-D wave equation has the form

$$\nabla^2 \psi = \frac{1}{v^2} \frac{\partial^2 \psi}{\partial t^2}$$

where $\psi$ is a physical quantity which propagates wavelike with speed $v$.

On substituting the value of $\nabla \times \mathbf{H}$ from Eq. (1.3d), we get

$$\nabla^2 \mathbf{E} = \mu_0 \varepsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{1.5}$$

You can similarly show that

$$\nabla^2 \mathbf{H} = \mu_0 \varepsilon_0 \frac{\partial^2 \mathbf{H}}{\partial t^2} \tag{1.6}$$

*Spend 5 min*

SAQ 2

Prove Eq. (1.6)

Do you recognise Eqs.(1.5) and (1.6)? These are identical in form to 3-D wave equation derived in Unit 6 of the Oscillations and Waves course (PHE-02). This means that each component of E and H satisfies a wavelike equation. The speed of propagation of an electromagnetic wave in free space is given by

$$v = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} \tag{1.7}$$

This remarkably simple result shows that the speed of an electromagnetic wave depends only on $\mu_0$ and $\varepsilon_0$. This suggests that all e.m. waves should, irrespective of frequency or amplitude, share this speed while propagating in free space. We can easily calculate the magnitude of $v$ by noting that for free space

$$\varepsilon_0 = 8.8542 \times 10^{-12}\, \mathrm{C^2 N^{-1} m^{-2}}$$

and

$$\mu_0 = 4\pi \times 10^{-7}\, \mathrm{N s^2 C^{-2}}.$$

Thus

$$v = \frac{1}{[(8.8542 \times 10^{-12}\, \mathrm{C^2 N^{-1} m^{-2}}) \times (4 \times 10^{-7}\, \mathrm{Ns^2 C^{-2}})]^{1/2}}$$

$$= 2.99794 \times 10^8\, \mathrm{ms^{-1}}$$

This is precisely the speed of light! It is worthwhile to mention here that using the then best known value of $\varepsilon_0$, Maxwell found that electromagnetic waves should travel at a speed of $3.1074 \times 10^8$ ms$^{-1}$. This, to his amusement, was very close to the speed of light measured by Fizeau ($3.14858 \times 10^8$ ms$^{-1}$). Based on these numbers, Maxwell proposed the electromagnetic theory of light. In his own words

> "This velocity is so nearly that of light, that it seems we have strong reason tc believe that light itself is an electromagnetic disturbance in the form of waves propagated through the electromagnetic field according to electromagnetic laws."

We cannot help but wonder at such pure gold having come out of his researches on electric and magnetic phenomena. It was a rare moment of unveiled exuberance – a classic example of the unification of knowledge towards which science is ever striving. With this one calculation, Maxwell brought the entire science of optics under the umbrella of electromagnetism. Its significance is profound because it identifies light with structures consisting of electric and magnetic fields travelling freely through free space.

The direct experimental evidence for electromagnetic waves came through a series of brilliant experiments by Hertz . He found that he could detect the effect of electromagnetic induction at considerable distances from his apparatus. His apparatus is shown in Fig. 1.1. By measuring the wavelength and frequency of electromagnetic waves, Hertz calculated their speed. He found it to be precisely equal to the speed of light. He also demonstrated properties like reflection. refraction, interference, etc and demonstrated conclusively that light is an electromagnetic wave.

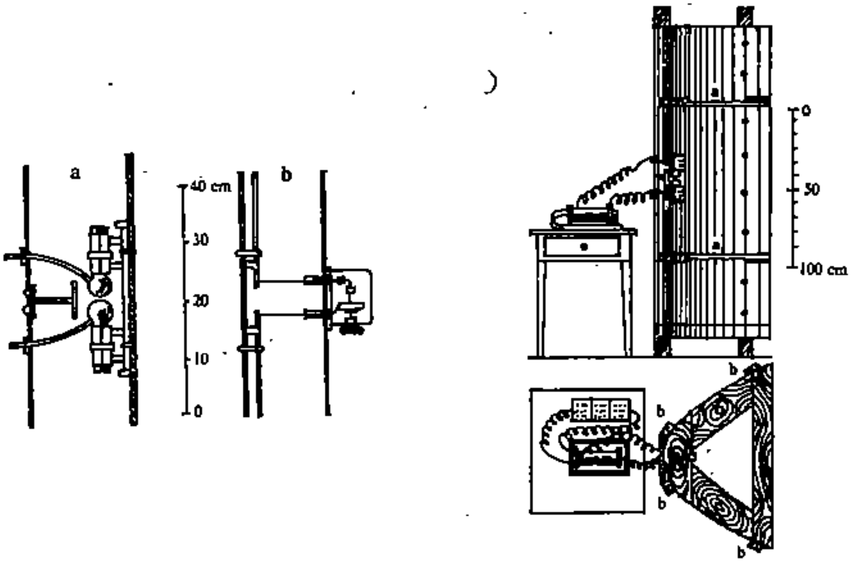


**Fig.1.1: Hertz's apparatus for the generation and detection of electromagnetic waves**

You now know that electromagnetic waves are generated by time varying electric and magnetic fields. So these are described by the amplitudes and phases of these fields. The simplest electromagnetic wave is the plane wave. You may recall that in a plane wave the phases of all points on a plane normal to the direction of propagation are same. And for a plane electromagnetic wave propagating along the $+z-$ direction, the phase is $(kz - \omega t)$, where $k$ is the wave number and $\omega$ is the angular frequency of electromagnetic plane wave. And the scalar electric and magnetic fields can be expressed as

$$E = E_0 \exp\left[i(kz - \omega t)\right]$$

$$H = H_0 \exp\left[i(kz - \omega t)\right]$$

where $E_0$ and $H_0$ are amplitudes of $E$ and $H$.

For a wave propagating along the + $z$-direction, the field vectors **E** and **H** are independent of $x$ and $y$. Then Eqs. (1.3a) and (1.3b) reduce to

$$\frac{\partial E_z}{\partial z} = 0 \qquad (1.8a)$$

and

$$\frac{\partial H_z}{\partial z} = 0 \qquad (1.8b)$$

By the same argument you will find that the time variation of $E_z$ and $H_z$ can be expressed as

$$\frac{\partial E_z}{\partial t} = 0 \qquad (1.9a)$$

$$\frac{\partial H_z}{\partial t} = 0 \qquad (1.9b)$$

To arrive at Eqs. (1.9 a,b), we write the z-components of Eqs. (1.3c) and(1.3d) as

$$\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = -\mu_0 \frac{\partial H_z}{\partial t}$$

and

$$\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = -\varepsilon_0 \frac{\partial E_z}{\partial t}$$

Since **E** and **H** are independent of $x$ and $y$, the LHS will be identically equal to zero.

What do these equations convey? Physically, these imply that the components of **E** and **H** along the direction of propagation of an electromagnetic wave (+$z$-direction in this case) does not depend upon time and the space coordinate $z$. So we must have

$$E_z = 0 = H_z \qquad (1.10)$$

You should convince yourself why any other constant value of $E_z$ and $H_y$ would not represent a wave. We can now draw the following conclusions:

1. Plane electromagnetic waves have no longitudinal component. That is, they are transverse. This implies that if electric field is along the $x$- axis, the magnetic field will be along the $y$-axis so that we may write

$$\mathbf{E} = \hat{\mathbf{x}} E_0 e^{i(kz - \omega t)}$$

and

$$\mathbf{H} = \hat{\mathbf{y}} H_0 e^{i(kz - \omega t)} \qquad (1.11)$$

You may now ask: Are $E_0$ *and* $H_0$ connected? If so, what is the relation between them? To discover answer to this question you have to solve TQ2:

$$H_0 = \frac{k}{\mu_0 \omega} E_0$$

2. Since $\frac{k}{\mu_o \omega}$ is a real number, the electric and magnetic vectors should be in phase. Thus if **E** becomes zero (maximum) at some instant, **H** must also necessarily be zero (maximum) and so on. This also shows that neither electric nor magnetic wave can exist without the other. An electric field varying in

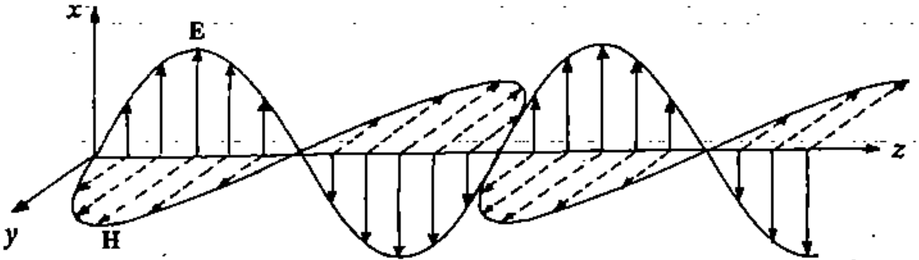


**Fig.1.2: The electric and magnetic fields associated with a plane electromagnetic wave**

time sets up a space-time varying magnetic field, which, in turn, produces an electric field varying in space and time, and so on. You cannot separate them. This mutually supporting role results in the generation of electromagnetic waves. The pictorial representation of fields of a plane electromagnetic waves (propagating along the + $z$- direction) is shown in Fig. 1.2. You will note that electric and magnetic fields are oriented at right angles to one another and to the direction of wave motion. Moreover, the variation in the spacing of the field lines and their reversal from one region of densely spaced lines to another reflect the spatial sinusoidal dependence of the wave fields.

### 1.4.1 Energy Transfer: The Poynting Vector

From Unit 6 of PHE-02 course you will recall that a general characteristic of wave motion is: **Wave carries energy, not matter.** Is it true even for electromagnetic waves? To know the answer, you should again consider the two field vectors (E and H) and calculate the divergence of their cross product. You can express it as

$$\nabla . (\mathbf{E} \times \mathbf{H}) = \mathbf{H} . (\nabla \times \mathbf{E}) - \mathbf{E} . (\nabla \times \mathbf{H}) \qquad (1.12)$$

Recall the identity $\nabla.(\mathbf{A} \times \mathbf{B}) = \mathbf{B} . (\nabla \times \mathbf{A}) - \mathbf{A} . (\nabla \times \mathbf{B})$ from unit 2 of PHE-04 course on Mathematical Methods in Physics-I.

If you now substitute for the cross products on the right-hand side from Maxwell's third and fourth equations respectively for free space, you will get

$$\nabla . (\mathbf{E} \times \mathbf{H}) = - \mathbf{H} . \mu_0 \frac{\partial \mathbf{H}}{\partial t} - \mathbf{E} . \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}$$

The time derivatives on the right-hand side can be written as

$$\mathbf{H} . \mu_0 \frac{\partial \mathbf{H}}{\partial t} = \mathbf{H} . \mu_0 \frac{\partial \mathbf{H}}{\partial t} = \frac{1}{2} \mu_0 \frac{\partial}{\partial t} (\mathbf{H} . \mathbf{H})$$

and

$$\varepsilon_0 \mathbf{E} . \frac{\partial \mathbf{E}}{\partial t} = \varepsilon_0 \mathbf{E} . \frac{\partial \mathbf{E}}{\partial t} = \frac{\varepsilon_0}{2} \frac{\partial}{\partial t} (\mathbf{E} . \mathbf{E})$$

so that

$$\nabla . (\mathbf{E} \times \mathbf{H}) = - \frac{\partial}{\partial t} \frac{1}{2} (\varepsilon_0 \mathbf{E} . \mathbf{E} + \mu_0 \mathbf{H} . \mathbf{H}) \qquad (1.13)$$

Gauss' divergence theorem relates the surface integral of a vector function to the volume integral of the divergence of this same function:

$$\int_S \mathbf{D} . d\mathbf{A} = \int_V \nabla . \mathbf{D}\, dV$$

The surface integral is taken over the closed surface, $S$ bounding the volume, $V$.

Do you recognise Eq. (1.13)? If so, can you identify it with some known equation in physics? This equation resembles the equation of continuity in hydrostatics. To discover the physical significance of Eq. (1.13), you should integrate it over volume $V$ bound by the surface $S$ and use Gauss' theorem. This yields

$$\int_V \nabla . (\mathbf{E} \times \mathbf{H})\, dV = - \frac{\partial}{\partial t} \int_V \frac{1}{2} (\varepsilon_0 \mathbf{E} . \mathbf{E} + \mu_0 \mathbf{H} . \mathbf{H})\, dV$$

or

$$\int_S (\mathbf{E} \times \mathbf{H}) . d\mathbf{A} = - \frac{\partial}{\partial t} \int_V \frac{1}{2} (\varepsilon_0 \mathbf{E} . \mathbf{E} + \mu_0 \mathbf{H} . \mathbf{H})\, dV$$

The integrand on the right hand side refers to the time rate of flow of electromagnetic energy in free space. You will note that both **E** and **H** contribute to it equally. The vector

$$\mathbf{S} = \mathbf{E} \times \mathbf{H} \qquad (1.14)$$

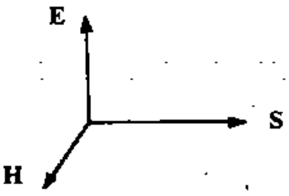


Fig. 1.3: The Poynting Vector

is called the **Poynting Vector**. It is obvious that S, E and H are mutually orthogonal. Physically it implies that S points in the direction of propagation of the wave since electromagnetic waves are transverse. This is illustrated in Fig. 1.3.

You may now like to know the time-average of energy carried by electromagnetic waves (light) per unit area. If you substitute for E and H in Eq. (1.14) and average over time, you will obtain

$$\langle S \rangle = \hat{z}\frac{k}{2\mu_0\omega}E_0^2 \tag{1.15}$$

Before you proceed, you should convince yourself about the validity of this result. To ensure this we wish you to solve SAQ 3.

*Spend 5 min*

SAQ 3

Prove Eq.(1.15).

### 1.4.2 The Electromagnetic Spectrum

Soon after Hertz demonstrated the existence of electromagnetic waves in 1888, intense interest and activity got generated. In 1895, J.C. Bose, working at Calcutta, produced electromagnetic waves of wavelengths in the range 25 mm to 5 m. (In 1901, Marconi succeded in transmitting electromagnetic waves across the Atlantic Ocean. This created public sensation. In fact, this pioneering work marked the beginning of the era of communication using electromagnetic waves.) X-rays, discovered in 1898 by Roentgen, were shown in 1906 to be e.m. waves of wavelength much smaller than the wavelength of light waves. Our knowledge of e.m. waves of various wavelengths has grown continuously since then. The e.m. spectrum, as we know it today, is shown in Fig. 1.4.

The range of wavelengths (and their applications in modern technologies) is very wide. However, the boundaries of various regions are not sharply defined. The visible light is confined to a very limited portion of the spectrum from about 4000 Å to 7000 Å. As you know, different wavelengths correspond to different colours. The red is at the long wavelength-end of visible region and the violet at the short wavelength -end. For centuries our only information about the universe beyond earth has come from visible light. All electromagnetic waves from 1 m to $10^6$ m are referred to as radiowaves. These are used in transmission of radio and television signals. The ordinary AM radio corresponds to waves with $\lambda = 100$m, whereas FM radio corresponds to 1m. The microwaves are used for radar and satellite communications ( $\lambda \sim 0.5\text{m} - 10^{-3}$ m ).

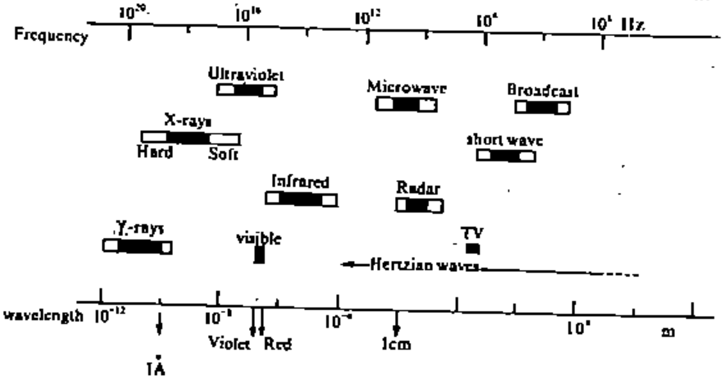


Fig.1.4: The electromagnetic spectrum

Between two radio waves and visible light lies the infrared region. Beyond the visible region we encounter the ultraviolet rays, X-rays and gamma rays. You must convince yourself that all phenomena from radio waves to gamma rays are essentially the same; they are all electromagnetic waves which differ only in wavelength (or frequency). You may now be tempted to enquire: Why do we attribute different nomenclature to different portions of the electromagnetic spectrum? The distinction is a mere convenience while identifying their practical applications.

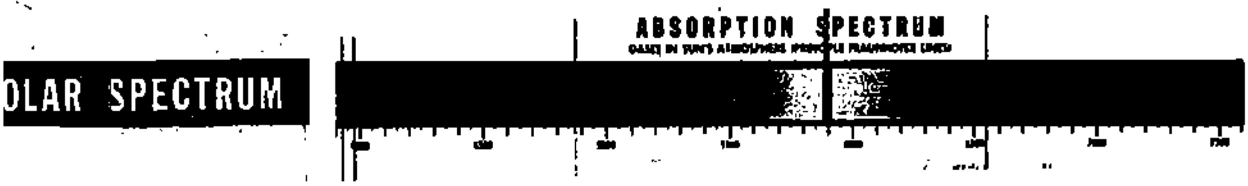


Fig. 1.5: The solar spectrum received on the earth

In our solar system, the sun is the major source of e.m. waves. If you closely examine the solar spectrum received on the earth, you will observe broad continuous spectrum crossed by Fraunhofer dark absorption lines (Fig. 1.5).

Let us now sum up what you have learnt in this unit.

## 1.5 SUMMARY

- Light is an electromagnetic wave.
- The electric and magnetic fields constituting an electromagnetic wave satisfy the equations

$$\nabla^2 \mathbf{E} = \mu_0 \, \varepsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2}$$

and

$$\nabla^2 \mathbf{H} = \mu_0 \, \varepsilon_0 \frac{\partial^2 \mathbf{H}}{\partial t^2}$$

- For a plane electromagnetic wave propagating along the + $z$- direction, the electric and magnetic fields can be expressed as

$$\mathbf{E} = \hat{\mathbf{x}} E_0 \exp \left[ i (kz - \omega t) \right]$$

and

$$\mathbf{H} = \hat{\mathbf{y}} H_0 \exp \left[ i (kz - \omega t) \right]$$

- The electromagnetic waves are transverse.
- The pointing vector **S** = **E** x **H** defines the direction of propagation of an electromagnetic wave.
- The visible light is confined to a very limited portion (4000 Å - 7000 Å) of electromagnetic spectrum.

## 1.6 TERMINAL QUESTIONS

1. Derive the wave equation for the propagation of electromagnetic waves in a conducting medium.

2. Starting from Eqs. (1.3c) and (1.3d) show that

$$H_0 = \frac{k}{\mu_0\,\omega} E_0$$

3. The energy radiated by the sun per second is approximately 4.0 x $10^{26}$ $Js^{-1}$. Assuming the sun to be a sphere of radius 7 x $10^8$m, calculate the value of Poynting vector at its surface. How much of it is incident on the earth? The average distance between the sun and earth is 1.5 x $10^{11}$m.

## 1.7 SOLUTIONS AND ANSWERS

**SAQs**

1. According to the corpuscular model, light travels in straight lines. As a result, the size of the shadow should be equal to the size of the object. Grimaldi's observation - the size of the shadow is smaller than the size of the obstacle - indicates that light bends around edges, contradicting corpuscular model.

2. Taking the curl of Eq. (1.3d), we get

$$\nabla \times \nabla \times \mathbf{H} = \varepsilon_0 \nabla \times \left(\frac{\partial \mathbf{E}}{\partial t}\right)$$

$$= \varepsilon_0 \frac{\partial}{\partial t}(\nabla \times \mathbf{E})$$

Using the vector identity

$$\text{curl curl } \mathbf{H} = \text{grad div } \mathbf{H} - \nabla^2 \mathbf{H}$$

we have

$$\nabla(\nabla . \mathbf{H}) - \nabla^2 \mathbf{H} = \mu_0 \varepsilon_0 \frac{\partial}{\partial t}\left(-\frac{\partial \mathbf{H}}{\partial t}\right)$$

Since $\nabla . \mathbf{H} = 0$, we get

$$\nabla^2 \mathbf{H} = \mu_0 \varepsilon_0 \frac{\partial^2 \mathbf{H}}{\partial t^2}$$

3. From Eq. (1.14), we have for Poynting vector

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}$$

Taking only the real part of Eq. (1.11), the electric and magnetic field vectors can be represented as

$$\mathbf{E} = \hat{\mathbf{x}} E_0 \cos(kz - \omega t)$$

$$\mathbf{H} = \hat{\mathbf{y}} H_0 \cos(kz - \omega t)$$

$$= \hat{\mathbf{y}} \frac{k}{\mu_0\,\omega} E_0 \cos(kz - \omega t) \qquad \left(\because H_0 = \frac{k}{\mu_0\,\omega} E_0\right)$$

So

$$\mathbf{E} \times \mathbf{H} = (\hat{\mathbf{x}} \times \hat{\mathbf{y}})\, \frac{k}{\mu_0\,\omega} E_0 \cos^2(kz - \omega t)$$

or

$$\mathbf{S} = \hat{z}\frac{k}{\mu_0\,\omega}E_0^2\cos^2(kz - \omega t)$$

This gives the amount of energy crossing a unit area perpendicular to $z$-axis per unit time. Typical frequency for an optical beam is of the order of $10^{15}\ \mathrm{s}^{-1}$ and the cosine term will fluctuate rapidly. Therefore, any measuring device placed in the path would record only an average value. The time average of the cosine term, as you know, is 1/2. Hence

$$\langle \mathbf{S} \rangle = \hat{z}\frac{k}{2\,\mu_0}\frac{E_0^2}{\omega}$$

## TQs

1. While deriving the wave-equation for electromagnetic waves in free space, we assumed that the electric cuurrent density is zero:

$$\mathbf{J} = \sigma\mathbf{E} = 0$$

This is because the conductivity (σ) of the free space was taken to be zero However in case of conducting medium, σ is non-zero. Hence

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

and

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

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

where symbols have their usual meaning.

With the help of above relations, Maxwell's relations in a conducting medium can be written as

$$\nabla . \mathbf{E} = 0 \tag{1a}$$

$$\nabla . \mathbf{B} = 0 \tag{1b}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{1c}$$

and

$$\nabla \times \mathbf{B} = \mu\left(\sigma\mathbf{E} + \varepsilon\frac{\partial \mathbf{E}}{\partial t}\right) \tag{1d}$$

Taking curl of equation (1c), we get

$$\nabla \times \nabla \times \mathbf{E} = \nabla \times \left(-\frac{\partial \mathbf{B}}{\partial t}\right)$$

Using the identity $\nabla \times \nabla \times \mathbf{A} = \text{grad div}\,\mathbf{A} - \nabla^2\mathbf{A}$, we have

$$\text{grad div}\,\mathbf{E} - \nabla^2\mathbf{E} = -\frac{\partial}{\partial t}(\nabla \times \mathbf{B})$$

Using (1a) and (1d) in this expression, we get

$$-\nabla^2\mathbf{E} = -\frac{\partial}{\partial t}\left[\mu\left\{\sigma\mathbf{E} + \varepsilon\frac{\partial \mathbf{E}}{\partial t}\right\}\right]$$

which implies that

$$- \nabla^2 \mathbf{E} = - \mu \sigma \frac{\partial \mathbf{E}}{\partial t} - \mu \varepsilon \frac{\partial^2 \mathbf{E}}{\partial t^2}$$

or

$$\nabla^2 \mathbf{E} - \mu \sigma \frac{\partial \mathbf{E}}{\partial t} - \mu \varepsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0 \qquad (2)$$

This is the wave equation for the propagation of the electromagnetic waves in a conducting medium.

2. If we write Eqs. (1.3c) and (1.3d) in component form, we get

$$\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = - \mu_0 \frac{\partial H_x}{\partial t} \qquad (a)$$

$$\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} = - \mu_0 \frac{\partial H_y}{\partial t} \qquad (b)$$

$$\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = - \mu_0 \frac{\partial H_z}{\partial t} \qquad (c)$$

and

$$\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} = \varepsilon_0 \frac{\partial E_x}{\partial t} \qquad (d)$$

$$\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} = \varepsilon_0 \frac{\partial E_y}{\partial t} \qquad (e)$$

$$\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = \varepsilon_0 \frac{\partial E_z}{\partial t} \qquad (f)$$

Now let us take $x$-axis along the electric field vector **E**. Then

$$E_y = 0 \qquad (g)$$

From Eq. (1.10) it follows for electromagnetic wave travelling along the $x$-axis that

$$E_z = 0 = H_z \qquad (h)$$

On using these results in Eq. (a) to (f), we get

$$\frac{\partial H_x}{\partial t} = 0 \qquad (i) \qquad \frac{\partial H_y}{\partial z} = - \varepsilon_0 \frac{\partial E_x}{\partial t} \qquad (l)$$

$$\frac{\partial E_x}{\partial z} = - \mu_0 \frac{\partial H_y}{\partial t} \qquad (j) \qquad \frac{\partial H_x}{\partial z} = 0 \qquad (m)$$

$$\frac{\partial E_x}{\partial y} = \mu_0 \frac{\partial H_z}{\partial t} \qquad (k) \qquad \frac{\partial H_y}{\partial x} = \frac{\partial H_x}{\partial y} \qquad (n)$$

The plane wave representation of the electric and magnetic field vectors is given as

$$E = E_0 \exp \left[ i (kz - \omega t) \right]$$

$$H = H_0 \exp \left[ i (kz - \omega t) \right]$$

Substituting these in Eq. (j), we get

$$i\,k E_0 = i\,\mu_0\,\omega\, H_0 \Rightarrow H_0 = \frac{k}{\mu_0\,\omega} E_0$$

3. We know that Poynting vector denotes the rate at which energy is radiated per unit area. So we can write the total average energy radiated from the surface of sun per unit time as

   or

$$E = \langle S \rangle \times 4\pi R^2$$

$$\langle S \rangle = \frac{E}{4\pi R^2} = \frac{4.0 \times 10^{26}\,\mathrm{Js}^{-1}}{4 \times 3.1416 \times (7 \times 10^8\,\mathrm{m})^2}$$

$$= 6.5 \times 10^7\,\mathrm{J\,m^{-2}\,s^{-1}}$$

To calculate the energy incident on the earth, we should know the average Poynting vector $\langle S_E \rangle$ at the surface of earth. To do so, we denote the distance between the surface of earth and centre of the sun as $R_E = R_{ES} + R$ and note that

$$\langle S_E \rangle 4\pi R_E = \langle S \rangle \times 4\pi R^2$$

so that

$$\langle S_E \rangle = \left(\frac{R}{R_E}\right)^2 \langle S \rangle$$

$$= \left(\frac{7 \times 10^8\,\mathrm{m}}{1.5 \times 10^{11}\,\mathrm{m}}\right)^2 \times \left(6.5 \times 10^7\,\mathrm{J\,m^{-2}\,s^{-1}}\right)$$

$$= 1.42 \times 10^3\,\mathrm{Jm^{-2}\,s^{-1}}.$$

# UNIT 2 REFLECTION AND REFRACTION OF LIGHT

**Structure**

## 2.1 INTRODUCTION

In the previous unit you have learnt that light is an electromagnetic wave. It is made up of mutually supporting electric and magnetic fields, which vary continuously in space and time. An interesting question related to e.m. waves is: What happens to these fields when such a wave is incident on the boundary separating two optically different media? From Unit 7 of PHE-02 Course you may recall that when a wave passes from air to water or air to glass, we get a reflected wave and a refracted wave. Reflection of light from a silvered surface, a looking mirror say, is the most common optical effect. Reflection of e.m. waves governs the working of a radar. Reflection of radiowaves by the ionosphere makes signal transmission possible and is so crucial in the area of communication.

In your earlier school years you have learnt that refraction explains the working of lenses and is responsible for seeing; our contact with surroundings. Even the grand spectacle of sun-set or a rainbow can be explained in terms of refraction of light. Refraction of e. m. waves forms the basis of one of the greatest technological applications in signal transmission. In fact, electro-optics has seen tremendous growth via optical fibres for a variety of applications.

In Unit 7 of PHE-02 course on Oscillations and Waves, you learnt to explain reflection and refraction of waves on the basis of Huygens' wave model. Now the question arises: Can we extend this analysis to electromagnetic waves, which include visible light, radiowaves, microwaves and X-rays? In Sec. 2.2 you will learn to derive the equations for reflected and transmitted fields (**E** and **B**) when an e.m. wave is incident normally as well as obliquely on the boundary of two media.

You are aware that many physical systems behave according to optimisation principle. In PHE-06 course you have learnt that when several fluids at different temperatures are mixed, the heat exchange takes place so that the total entropy of the system is maximum. A ball rolling on an undulating surface comes to rest at the lowest point. The profoundness of such situations and scientific laws governing them led Fermat to speculate: Does light also obey some optimization principle? And he concluded: **Ray of light chooses a path of extremum between two points.** This is known as Fermat's principle. Implicit in it are the assumptions

(i) Light travels at a finite speed, and

(ii) The speed of light is lower in a denser medium.

In Sec. 2.4 you will learn about Fermat's principle. We have shown that all laws of geometrical optics are contained in it.

## Objectives

After studying this unit you should be able to

- explain reflection and refraction of e.m. waves incident normally and obliquely on the interface separating two optically different media
- apply Fermat's principle to explain the reflection and refraction of light, and
- solve problems based on reflection and refraction of e.m. waves.

## 2.2 ELECTROMAGNETIC WAVES AT THE INTERFACE SEPARATING TWO MEDIA

Consider a plane electromagnetic wave that is incident on a boundary between two linear media. That is, **D** and **H** are proportional to **E** and **B**, respectively, and the constants of proportionality are independent of position and direction. You can visualise it as light passing from air (medium 1) to glass (medium 2). Let us assume that there are no free charges or currents in the materials.

Fig. 2.1 shows a plane boundary between two media having different permittivity and permeability: $\varepsilon_1$, $\mu_1$ for medium 1 and $\varepsilon_2$, $\mu_2$ for medium 2. A uniform plane wave travelling to the right in medium 1 is incident on the interface **normal to the boundary**. As in the case of waves on a string, we expect a reflected wave propagating back into the medium and a transmitted (or refracted) wave travelling in the second medium. We wish (i) to derive expressions for the fields associated with reflected and refracted waves in terms of the field associated with the incident wave and (ii) know the fraction of the incident energy that is reflected and transmitted. To do so we need to know the **boundary conditions** satisfied by these waves at the interface separating the two media. We obtain these conditions by

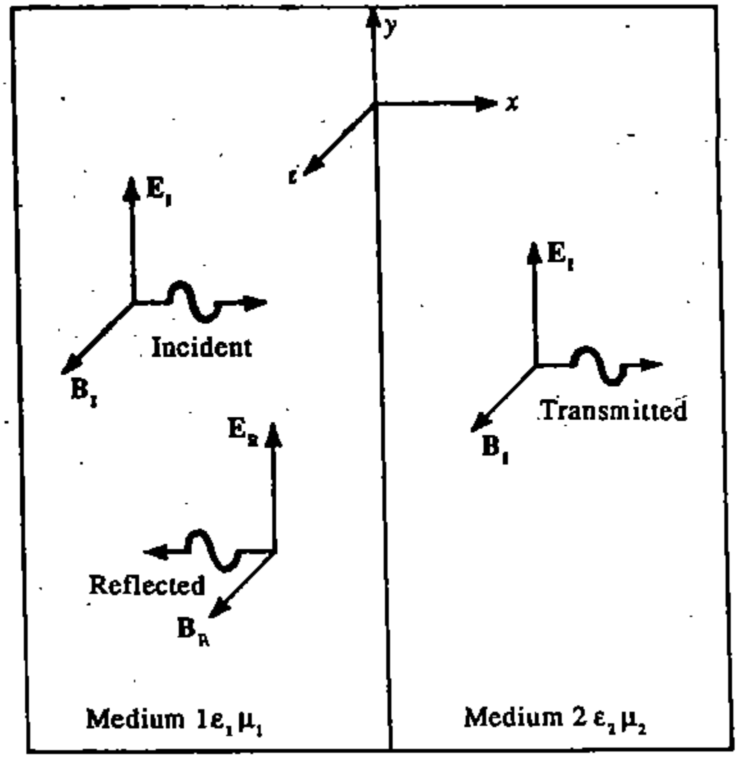


**Fig.2.1: A uniform plane wave is incident normally on a plane boundary. The reflected and refracted (transmitted) waves are also shown. The angle of incidence is α and angle of refraction is β.**

stipulating that Maxwell's equations must be satisfied at the boundary between these media. We first state the appropriate conditions. Their proof is given in the appendix to this Unit.

**Boundary Conditions**

You learnt to derive the boundary conditions from Maxwell's equations for a medium free of charges and currents in Unit 15 of the PHE-07 course on electric and magnetic phenomena. For your convenience, we rewrite appropriate integral form of these equations:

$$\varepsilon \int_S \mathbf{E} \cdot d\mathbf{S} = 0 \tag{2.1a}$$

$$\int_S \mathbf{B} \cdot d\mathbf{S} = 0 \tag{2.1b}$$

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = \frac{d}{dt}\int_s \mathbf{B} \cdot d\mathbf{S} \tag{2.1c}$$

and

$$\frac{1}{\mu} \oint_C \mathbf{B} \cdot d\mathbf{l} = \varepsilon \frac{d}{dt}\int_S \mathbf{E} \cdot d\mathbf{S} \tag{2.1d}$$

where S is a surface bound by the closed loop *C*.

The electric field can oscillate either parallel or normal to the plane of incidence. The magnetic field **B** will then be normal or parallel to the plane of incidence. We will denote these with subscripts ∥ (parallel) and ⊥ (normal). The boundary conditions for normal and parallel components of electric and magnetic fields take the form (Appendix A).

$$\varepsilon_1 E_{1\perp} - \varepsilon_2 E_{2\perp} = 0 \tag{2.2a}$$

$$B_{1\perp} - B_{2\perp} = 0 \tag{2.2b}$$

$$E_{1\parallel} - E_{2\parallel} = 0 \tag{2.2c}$$

and

$$\frac{1}{\mu_1} B_{1\parallel} - \frac{1}{\mu_2} B_{2\parallel} = 0 \tag{2.2d}$$

We shall now use the boundary conditions expressed by Eqs. (2.2a- d) to study reflection and refraction (transmission) at normal as well as oblique incidence.

## 2. 2. 1 Normal Incidence

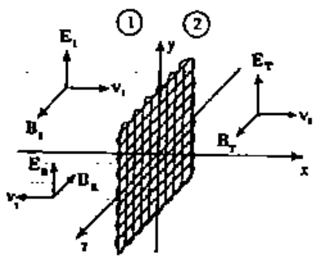


Fig.2.2: A sinusoidal plane e.m. wave incident normally at the boundary of two optically transparent media

Refer to Fig. 2.2. The *yz*-plane ($x = 0$) forms the interface of two optically transparent (non-absorbing) media (refractive indices $n_1$ and $n_2$). A sinusoidal plane wave of frequency ω travelling in *x*-direction is incident from the left. From Unit 7 of the Oscillations and Waves course you will recall that progressive waves are partially reflected and partially refracted at the boundary separating two physically different media. However, the energy of the reflected or transmitted e.m. waves depends upon their refractive indices.

The appropriate magnetic fields to be associated with electric fields are obtained from the equation

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

Let us suppose that the electric field is along the *y*-direction. Then the electric and magnetic fields associated with the incident wave are given by

$$\mathbf{E_I}(x,t) = E_{0I}\hat{\mathbf{j}} \exp\left[i(k_I x - \omega t)\right] \qquad (2.3a)$$

and

$$\mathbf{B_I}(x,t) = \frac{E_{0I}}{v_1}\hat{\mathbf{k}} \exp\left[i(k_I x - \omega t)\right] \qquad (2.3b)$$

The reflected wave propagates back into the first medium and can be represented by the following fields:

$$\mathbf{E_R}(x,t) = E_{oR}\hat{\mathbf{j}} \exp\left[-i(k_I x + \omega t)\right] \qquad (2.4a)$$

and

$$\mathbf{B_R}(x,t) = -\frac{E_{oR}}{v_1}\hat{\mathbf{k}} \exp\left[-i(k_I x + \omega t)\right] \qquad (2.4b)$$

The minus sign in the exponents in Eqs. (2.4a,b) indicates that propagation of the wave is in the -*x* direction. But the negative sign with the amplitude in Eq. (2.4b) arises because of transverse nature of e.m. waves and that the electric and magnetic field vectors should obey the relation

$$\mathbf{B_R} = \frac{1}{v_1}(\hat{\mathbf{k}}_\mathbf{I} \times \mathbf{E_R})$$

where $\hat{\mathbf{k}}_\mathbf{I}$ is unit vector along the direction of incidence.

If you visualise Eqs. (2.3) and (2.4) diagramatically, you will note that the electric vectors have been kept fixed in the same direction but the magnetic field vectors have been oriented. The orientation of the magnetic field vector ensures that the flow of energy is always along the direction of propagation of the wave (Poynting theorem).

The electric and magnetic fields of the transmitted wave, which travels to the right in medium 2, are given by

$$\mathbf{E_T}(x,t) = E_{oT}\hat{\mathbf{j}} \exp\left[i(\omega_T t + k_T x)\right] \qquad (2.5a)$$

and

$$\mathbf{B_T}(x,t) = \frac{1}{v_2}\left[\hat{\mathbf{k}}_\mathbf{T} \times \mathbf{E_T}(x,t)\right] \qquad (2.5b)$$

The phenomenon of reflection and refraction is usually analysed in two parts:

(i) To determine the relations between the field vectors of the reflected and refracted waves in terms of that of the incident wave. These relations determine the reflection and the transmission coefficients. In this derivation, we match the **E** and **B** fields in the two media at the interface with the help of appropriate boundary conditions there.

(ii) To establish relations between the angle of incidence and the angles of reflection and refraction we may emphasize that so far as the laws of reflection and refraction are concerned, explicit use of any boundary condition is not required.

## Fresnel's Amplitude Relations

To derive expressions for the amplitudes of the reflected and the refracted waves in terms of the amplitude of the incident wave, we apply boundary conditions given by Eq. (2.2a-d) at **every point** on the interface at **all times**. At $x = 0$, the combined field to the left ($\mathbf{E}_I + \mathbf{E}_R$ and $\mathbf{B}_I$ and $\mathbf{B}_R$) must join the fields to the right ($\mathbf{E}_T$ and $\mathbf{B}_T$). For normal incidence, there are no normal field components (perpendicular to the interface). But why? This is because neither **E** nor **B** field is in the $x$-direction. This means that Eqs. (2.2a,b) are trivial and only tangential components of the electric and magnetic fields should be matched at the plane $x = 0$. Thus

$$E_{0I} + E_{0R} = E_{0T} \tag{2.6a}$$

and

$$\frac{1}{\mu_1}\left( B_{0I} + B_{0R} \right) = \frac{1}{\mu_2} B_{0T}$$

or

$$\frac{1}{\mu_1}\left( \frac{E_{0I}}{\nu_1} - \frac{E_{0R}}{\nu_1} \right) = \frac{1}{\mu_2}\frac{E_{0I}}{\nu_2}$$

which, on simplification yields

$$E_{0I} - E_{0R} = \alpha E_{0T} \tag{2.6b}$$

where

$$\alpha = \frac{\mu_1 \nu_I}{\mu_2 \nu_2} = \sqrt{\frac{\mu_1 \varepsilon_1}{\mu_2 \varepsilon_2}} = \frac{\mu_1 n_2}{\mu_2 n_1} \tag{2.6c}$$

Solving Eqs. (2.6a) and (2.6b) for the reflected and transmitted electric field amplitudes in terms of the incident amplitude, you will find that

$$E_{0R} = \left( \frac{1 - \alpha}{1 + \alpha} \right) E_{0I} \tag{2.7a}$$

and

$$E_{0T} = \frac{2}{1 + \alpha} E_{0I} \tag{2.7b}$$

For most optical media, the permeabilities are close to their values in vacuum ($\mu_1 \approx \mu_2 \approx \mu_0$). In such cases $\alpha = \frac{\nu_1}{\nu_2}$ and we have

$$E_{0R} = \left( \frac{\nu_2 - \nu_1}{\nu_2 + \nu_1} \right) E_{0I}$$

and

$$E_{0T} = \frac{2\nu_2}{\nu_2 + \nu_1} E_{0I} \tag{2.8}$$

This suggests that when $\nu_2 > \nu_1$, the reflected wave will be in phase with the incident wave and for $\nu_2 < \nu_1$, the reflected and incident waves will be out of phase. This is illustrated in Fig. 2.3.

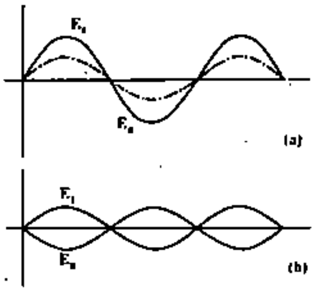


**Fig.2.3: The phase relationship between reflected wave and the incident wave**

In terms of the index of refraction $n \left( = \dfrac{c}{v} \right)$, we can rewrite Eq. (2.8) as

$$E_{0R} = \frac{n_1 - n_2}{n_1 + n_2} E_{0I}$$

and

$$E_{0T} = \frac{2n_1}{n_1 + n_2} E_{0I} \qquad (2.9)$$

When an e.m. wave passes from a rarer medium to a denser medium ($n_1 < n_2$), the ratio $\dfrac{E_{0R}}{E_{0I}}$ will be negative. Physically, it means that the reflected wave is 180° out of phase with the incident wave. You have already learnt it in case of reflection of sound waves in the course on Oscillations and Waves. When an e.m. wave is incident from a denser medium on the interface separating it from a rarer medium ($n_1 > n_2$), the ratio $\dfrac{E_{0R}}{E_{0I}}$ is positive and no such phase change occurs.

We can now easily calculate the **reflection** and the **transmission coefficients,** which respectively measure the fraction of incident energy that is reflected and transmitted. The first step in this calculation is to recall that

$$R = \frac{I_R}{I_I}$$

and

$$T = \frac{I_T}{I_I}$$

where $I_R$, $I_T$ and $I_I$ respectively denote the reflected, transmitted and incident wave intensity. Intensity is defined as the average power per unit area, $(1/2)\, v E^2$. So you can readily show that

$$R = \frac{I_R}{I_I} = \left( \frac{n_1 - n_2}{n_1 + n_2} \right)^2 \qquad (2.10a)$$

and

$$T = \frac{I_T}{I_I} = \frac{n_2}{n_1} \left( \frac{2n_1}{n_1 + n_2} \right)^2 \qquad (2.10b)$$

You can convince yourself that $R + T = 1$. For air ($n_1 = 1$) - glass ($n_2 = 1.5$) interface, the $R$ and $T$ coefficients have the values $R = 0.04$ and $T = 0.96$. There is no energy stored (or absorbed) at the interface and you can now realise why most of the light is transmitted.

We will now repeat this exercise for the case of oblique incidence.

### 2.2.2 Oblique Incidence

Refer to Fig. 2.4. A plane electromagnetic wave is incident at an angle $\theta_I$. Let the angles of reflection and refraction be $\theta_R$ and $\theta_T$. We can represent the fields associated with these three plane electromagnetic waves as

Incident Wave

$$\mathbf{E}_\mathrm{I} = \mathbf{E}_{0\mathrm{I}} \exp\left[-i(\omega_I t - \mathbf{k}_\mathrm{I} \cdot \mathbf{r})\right]$$

$$\mathbf{B}_\mathrm{I} = \frac{1}{v_1}(\hat{\mathbf{k}}_\mathrm{I} \times \mathbf{E}_\mathrm{I}) \qquad (2.11a)$$

Reflected Wave

$$\mathbf{E}_\mathrm{R} = \mathbf{E}_{0\mathrm{R}} \exp\left[-i(\omega_R t - \mathbf{k}_\mathrm{R} \cdot \mathbf{r})\right]$$

$$\mathbf{B}_\mathrm{R} = \frac{1}{v_1}(\hat{\mathbf{k}}_\mathrm{R} \times \mathbf{E}_\mathrm{R}) \qquad (2.11b)$$

Transmitted Wave

$$\mathbf{E}_\mathrm{T} = \mathbf{E}_{0\mathrm{T}} \exp\left[-i(\omega_T t - \mathbf{k}_\mathrm{T} \cdot \mathbf{r})\right]$$

$$\mathbf{B}_\mathrm{T} = \frac{1}{v_2}(\hat{\mathbf{k}}_\mathrm{T} \times \mathbf{E}_\mathrm{T}) \qquad (2.11c)$$

You may recall that the boundary conditions must hold at every point on the interface at all times. If the boundary conditions hold at a point and at sometime, they will hold at all points in space for all subsequent times only if the exponential parts in above expressions for each wave are the same, i.e.

$$\omega_I t - \mathbf{k}_\mathrm{I} \cdot \mathbf{r} = \omega_R t - \mathbf{k}_\mathrm{R} \cdot \mathbf{r} = \omega_T t - \mathbf{k}_\mathrm{T} \cdot \mathbf{r}$$

at the interface. This implies that for equality of phases at all times we must have

$$\omega_I = \omega_R = \omega_T = \omega \quad (\text{say}) \qquad (2.12a)$$

That is, the frequency of an e.m. wave does not change when it undergoes reflection and refraction: all waves have the same frequency. Since the fields must satisfy Maxwell's equations, we must have for the wave vectors

$$\frac{k_I^2}{\omega^2} = \frac{1}{c^2} = \varepsilon_1 \mu_1 \qquad (2.13a)$$

$$\frac{k_T^2}{\omega^2} = \frac{1}{c^2} = \varepsilon_2 \mu_2 \qquad (2.13b)$$

$$\frac{k_R^2}{\omega^2} = \frac{1}{c^2} = \varepsilon_1 \mu_1 \qquad (2.13c)$$

Further, let $k_{Ix}$, $k_{Iy}$ and $k_{Iz}$ represent the $x$, $y$ and $z$ components of $\mathbf{k_I}$. We can use similar notation for $\mathbf{k_T}$ and $\mathbf{k_R}$. For the continuity conditions to be satisfied at all points on the interface, we must have

$$k_{Iy} = k_{Ty} = k_{Ry} \tag{2.14a}$$

and

$$k_{Iz} = k_{Tz} = k_{Rz} \tag{2.14b}$$

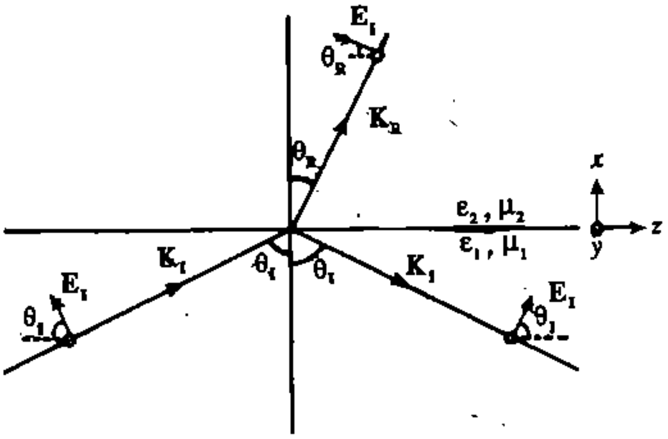


Fig.2.4: The reflection of a plane wave with its electric vector parallel to the plane of incidence

Let us choose the $y$-axis such that

$$k_{Iy} = 0$$

(i.e. we assume $\mathbf{k_I}$ to lie in the $x$-$z$ plane - see Fig. 2.4). Consequently

$$k_{Ty} = k_{Ry} = 0 \tag{2.14c}$$

This result implies that the vectors $\mathbf{k_I}$, $\mathbf{k_T}$ and $\mathbf{k_R}$ will lie in the same plane. Further, from Eq. (2.14b) we get

$$k_I \sin\theta_I = k_T \sin\theta_T = k_R \sin\theta_R \tag{2.15}$$

Since $|\mathbf{k_I}| = |\mathbf{k_R}|$ (see Eq. 2.13a and c), we must have

$$\theta_I = \theta_R \tag{2.16}$$

That is, the angle of incidence is equal to the angle of reflection, which is the law of reflection. Further,

$$\frac{\sin\theta_I}{\sin\theta_T} = \frac{k_T}{k_I} = \frac{\omega\sqrt{\varepsilon_2\mu_2}}{\omega\sqrt{\varepsilon_1\mu_1}}$$

or

$$\frac{\sin\theta_I}{\sin\theta_T} = \sqrt{\frac{\varepsilon_2\mu_2}{\varepsilon_1\mu_1}} \tag{2.17}$$

If we denote the speeds of propagation of the waves in media 1 and 2 by $v_1\left(=\frac{1}{\sqrt{\varepsilon_1\mu_1}}\right)$ and $v_2\left(=\frac{1}{\sqrt{\varepsilon_2\mu_2}}\right)$ we find that Eq. (2.17) can be rewritten as

$$\frac{\sin\theta_I}{\sin\theta_T} = \frac{v_1}{v_2} = \frac{n_2}{n_1} \tag{2.18}$$

where $n_1 = \dfrac{c}{v_1} = c\sqrt{\varepsilon_1\mu_1}$ and $n_2 = \dfrac{c}{v_2} = c\sqrt{\varepsilon_2\mu_2}$.

represent the refractive indices of media 1 and 2 respectively. Do you recognise Eq. (2.18)? It is the well known **Snell's law.**

Eqs. (2.16) and (2.18) constitute the **laws of reflection and refraction** in optics.

You can now derive Fresnel's amplitude relations following the procedure outlined for the case of normal incidence. For brevity, we just quote the results without going into details. (You will not be examined for the same in the term-end examination.) When **E** oscillates parallel to the plane of incidence, we have

$$\frac{E_{R\parallel}}{E_{I\parallel}} = \frac{\tan(\theta_I - \theta_T)}{\tan(\theta_I + \theta_T)} \tag{2.19a}$$

$$\frac{E_{T\parallel}}{E_{I\parallel}} = \frac{2\cos\theta_I\sin\theta_T}{\sin(\theta_I + \theta_T)\cos(\theta_I - \theta_T)} \tag{2.19b}$$

When **E** oscillates normal to the plane of incidence, we have

$$\frac{E_{T\perp}}{E_{I\perp}} = -\frac{\sin(\theta_I - \theta_T)}{\sin(\theta_I + \theta_T)} \tag{2.20a}$$

$$\frac{E_{T\perp}}{E_{I\perp}} = \frac{2\sin\theta_T\cos\theta_I}{\sin(\theta_I + \theta_T)} \tag{2.20b}$$

You can easily verify that for normal incidence these equations reduce to Eq. (2.9).

The corresponding expressions for reflections and transmission coefficients for normal and parallel oscillations of **E** when a plane wave is incident obliquely are

$$R_{\parallel} = \frac{\tan^2(\theta_I - \theta_T)}{\tan^2(\theta_I + \theta_T)} \tag{2.21a}$$

$$T_{\parallel} = \frac{\sin 2\theta_I \sin 2\theta_T}{\sin^2(\theta_I + \theta_T)\cos^2(\theta_I - \theta_T)} \tag{2.21b}$$

$$R_{\perp} = \frac{\sin^2(\theta_I - \theta_T)}{\sin^2(\theta_I + \theta_T)} \tag{2.21c}$$

and

$$T_{\perp} = \frac{\sin 2\theta_I \sin 2\theta_T}{\sin^2(\theta_I + \theta_T)} \tag{2.21d}$$

As before, you can easily show that for normal incidence these equations reduce to Eq. (2.10a, b).

## 2.3 IDEALIZATION OF WAVES AS LIGHT RAYS

So far you have learnt to explain reflection and refraction of plane electromagnetic waves at a plane interface. This signifies a relatively simple situation where the solutions of Maxwell's equations give the laws of propagation of light. It is not true in general and we invariably seek approximations to describe a phenomenon well. One such approximation makes use of smallness of wavelength of light. You know that the wavelength of light is very small ( $\sim 10^{-7}$m). It is orders of magnitude less than the dimensions of optical instruments such as telescopes and microscopes. In such cases, the passage of light is most easily shown by geometrical rays. A ray is the path of propagation of energy in the zero wavelength limit ( $\lambda \rightarrow 0$ ). The way in which rays may represent the propagation of wavefronts for some familiar situations is shown in Fig. 2.5. You will note that a plane wavefront corresponds to parallel rays and spherical wavefronts correspond to rays diverging from a point or converging to a point. You will agree that all parts of the wavefront take the same time to travel from the source.

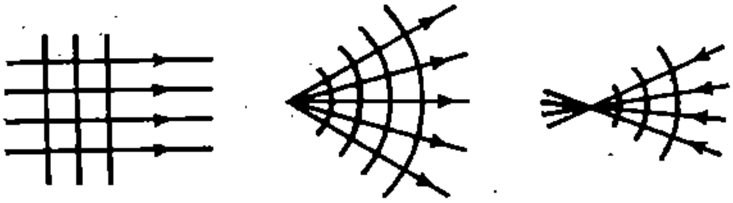


**Fig.2.5: Ray representation of a plane, diverging spherical and converging spherical wavefronts moving from left to right**

Huygens proposed that light propagates as a wavefront (a surface of constant phase) progresses in a medium perpendicular to itself with the speed of light. The zero wavelength approximation of wave optics is known as **geometrical optics.**

The laws of geometrical optics are incorporated in Fermat's principle. We will now discuss it in detail.

## 2.4 FERMAT'S PRINCIPLE

In its original form, Fermat's principle may be stated as follows:

> **Any light ray travels between two end points along a line requiring the minimum transit time.**

If $v$ is the speed of light at a given point in a medium, the time taken to cover the distance $dl$ is

$$dt = \frac{dl}{v} \qquad (2.22)$$

In your earlier years you have learnt that the refractive index of a medium is defined as the ratio of the speed of light in vacuum to its speed in the medium, i.e.

$$n = \frac{c}{v}$$

Using this relation in Eq. (2.22), we get

$$dt = \frac{1}{c}\, n\, dl$$

Hence, the time taken by light in covering the distance from point $A$ to $B$ is

$$\tau = \frac{1}{c}\int_A^B n\, dl$$

The quantity

$$L = \int_A^B n\, dl \qquad (2.23)$$

has the dimensions of length and is called the **optical distance** or **optical path length** between two given points. You must realise that optical distance is different from the physical (geometrical) distance ( $= \int_A^B dl$ ). However, in a homogeneous medium, the optical distance is equal to the product of the geometrical length and the refractive index of the medium. Thus, we can write

$$\tau = \frac{L}{c}$$

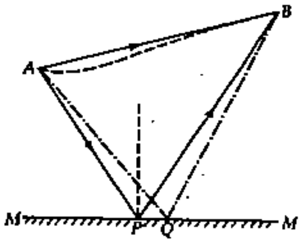


Fig. 2.6

This is Fermat's principle of least time. Let us pause for a moment and ask: Is there any exception to this law? Yes, there are cases where the optical path corresponds to maximum time or it is neither a maximum nor a minimum, i.e. stationary. To incorporate such situations, this principle is modified as follows:

**Out of many paths connecting two given points, the light ray follows that path for which the time required is an extremum. In other words, the optical path length between any two points is a maximum, minimum, or stationary.**

The essential point involved in Fermat's principle is that slight variation in the actual path causes a second-order variation in the actual path. Let us consider that light propagates from point $A$ in the medium characterised by the refractive index $n$ to the point $B$ as shown in Fig. 2.6. According to this principle,

$$\delta\int_A^B n(x, y, z)\, dl = 0 \qquad (2.24)$$

For a homogeneous medium, the rays are straight lines, since the shortest optical path between two points is along a straight line.

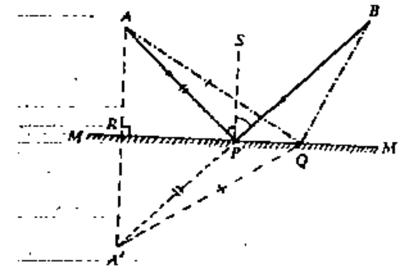


Fig. 2.7 : Reflection of rays at a plane interface

In effect, Fermat's principle prohibits the consideration of an isolated ray of light. It tells us that a path is real only when we extend our examination to the paths in **immediate neighbourhood** of the rays. To understand the meaning of this statement, let us consider the case of finding the path of a ray from a point $A$ to a point $B$ when both of them lie on the same side of a mirror $M$ (Fig. 2.7). It can be seen that the ray can go directly from $A$ to $B$ without suffering any reflection. Alternatively, it can go along the path $APB$ after suffering a single reflection from the mirror. If Fermat's principle had asked for, say, an absolute minimum, then the path $APB$ would be prohibited; but that is not the actual case. The path $APB$ is also minimum in the neighbourhood involving paths like $AQB$. The phrase "immediate neighbourhood of path" would mean those paths that lie near the path under consideration and are similar to it. For example, the path $AQB$ lies near $APB$ and is similar to it; along both paths the ray suffers one reflection at the mirror. Thus Fermat's principle requires an extremum in the immediate neighbourhood of the actual path, and in general, there may be more than one ray path connecting two points.

All the laws of geometrical optics are incorporated in Fermat's principle. We now illustrate Fermat's principle by applying it to reflection of light.

### Example 1

Using Fermat's principle, derive the laws of reflection.

### Solution

Let us first consider the case of reflection. Refer to Fig. 2.8. Light from a point $A$ is reflected at a mirror $MM$ towards a point $B$. $A$ ray $APB$ connects $A$ and $B$. $\theta_I$ and $\theta_R$ are the angles of incidence and reflection, respectively. We have denoted the vertical distances of $A$ and $B$ from the mirror $MM$ by $a$ and $b$. From the construction in Fig. 2.8 and Pythagoras' theorem, we find that the total path length $l$ of this ray from $A$ to $MM$ to $B$ is

$$l = \sqrt{a^2 + x^2} + \sqrt{b^2 + (d - x)^2} \tag{2.25}$$

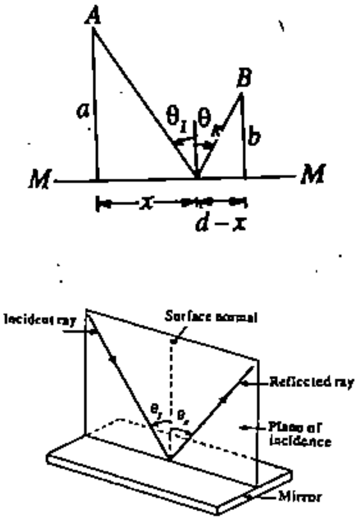


Fig. 2.8 Derivation of the laws of reflection using Fermat's principle.

where $x$ is the distance between the foot of the perpendicular from $A$ and the point $P$ at which the ray touches the mirror.

According to Fermat's principle, $P$ will have a position such that the time of travel of the light must be a minimum (a maximum or stationary). Expressed in another way, the total length $l$ of the ray must be a minimum or maximum or stationary. In other words, for Fermat's principle to hold, the derivative of $l$ with respect to $x$ must be zero, i.e. $dl/dx = 0$. Hence, on differentiating Eq. (2.25) with respect to $x$, we get

$$\frac{dl}{dx} = \frac{1}{2}(a^2 + x^2)^{-1/2}(2x) + \frac{1}{2}\left[b^2 + (d - x)^2\right]^{-1/2} \times 2(d - x)(-1) = 0 \tag{2.26}$$

which can be rewritten as

$$\frac{x}{(a^2 + x^2)^{1/2}} = \frac{d - x}{[b^2 + (d - x)^2]^{1/2}} \tag{2.27}$$

By examining Fig. (2.8) you will note that this gives

$$\sin \theta_I = \sin \theta_R$$

or

$$\theta_I = \theta_R \tag{2.28}$$

which is (part of) the law of reflection. You will also note that the incident ray, the reflected ray and the normal to $MM$ lie in the same incidence plane.

In the above example time required or the optical path length can be seen to be minimum by calculating the second deviative and finding its value at $x$ for which $dl/dx = 0$. The 2nd derivative turns out to be positive, showing it to be minimum. You can convince yourself by carrying out this simple caculation.

We now summarise what you have learnt in this unit.

## 2.5 SUMMARY

- When an e.m. wave is incident normally on the interface separating two optically different media, the reflected and transmitted electric field amplitudes are given by

  $$E_{0R} = \frac{1-\alpha}{1+\alpha} E_{0I}$$

  and

  $$E_{0T} = \frac{2}{1+\alpha} E_{0I}$$

  where $\alpha = \sqrt{\mu_1 \varepsilon_1 / \mu_2 \varepsilon_2}$ and $E_{0I}$ is amplitude of incident electric field.

- The frequency of an e.m. wave is not affected when it undergoes reflection or refraction.

- Fermat's principle states that a ray of light travels between two given points along that path for which the time required is an extremum:

  $$\delta \int_A^B n\,(x, y, z)\, dl = 0$$

## 2.6 TERMINAL QUESTIONS

1. Derive Snell's law from Fermat's principle.
2. A collimated beam is incident parallel to the axis of a concave mirror. It is reflected into a converging beam. Using Fermat's principle show that the mirror is parabolic.

## 2.7 SOLUTIONS AND ANSWERS

**TQs**

1. To prove the law of refraction from Fermat's principle, consider Fig. 2.9, which shows that the points $A$ and $B$ are in two optically different media. (If the refractive index on both sides of the boundary $SS$ were the same, the path from $A$ to $B$ would be a straight line, irrespective of the magnitude of the refractive index. But the refractive indices are not the same and the ray $APB$ is not a straight line.) Suppose that the velocities of light on the two sides of the boundary are $v_1$ and $v_2$. Since $v = l/t$, the time light takes to traverse the paths $AP$ and $PB$ is

   $$t = \frac{\sqrt{a^2 + x^2}}{v_1} + \frac{\sqrt{b^2 + (d-x)^2}}{v_2} = \frac{l_1}{v_1} + \frac{l_2}{v_2} \qquad \text{(i)}$$

   Using the relation $n = c / v$, this can be rewritten as

   $$t = \frac{n_1 l_1 + n_2 l_2}{c} = \frac{l}{c}$$

   where $l\,(= n_1 l_1 + n_2 l_2)$ is the optical path length of the ray. The geometrical path in this case is $l_1 + l_2$. If $\lambda$ is the wavelength of light in vacuum and $\lambda_n$ in a medium of refractive index $n$, then $\lambda = n\,\lambda_n$. This shows that the optical path length is equal to the length that the same number of waves would have if the medium were a vacuum.

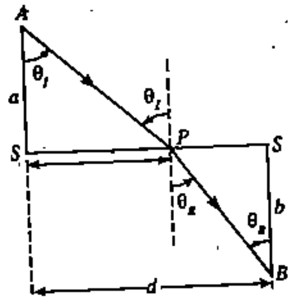


Fig.2.9: A ray from *A* passes to *B* after refraction at *P*

Fermat's principle requires that $dl / dx = 0$ for some values of $x$. The optical path length

$$l = l_1 n_1 + l_2 n_2$$

$$= n_1 \sqrt{(a^2 + x^2)} + n_2 \sqrt{b^2 + (d - x)^2} \qquad \text{(ii)}$$

so that $$\frac{dl}{dx} = n_1 \frac{x}{\left(a^2 + x^2\right)^{1/2}} - n_2 \frac{d - x}{\left\{b^2 + (d - x)^2\right\}^{1/2}} = 0$$

or $$n_2 \frac{x}{\left(a^2 + x^2\right)^{1/2}} = n_2 \frac{d - x}{\left\{b^2 + (d - x)^2\right\}^{1/2}} \qquad \text{(iii)}$$

As before, we can write it in terms of the angles of incidence and refraction as

$$n_I \sin \theta_I = n_2 \sin \theta_R \qquad \text{(iv)}$$

which is **Snell's law of refraction.** It shows that when light passes from a medium of lower refractive index (rarer medium) to a medium of higher refractive index (denser medium), it bends towards the surface normal.

2

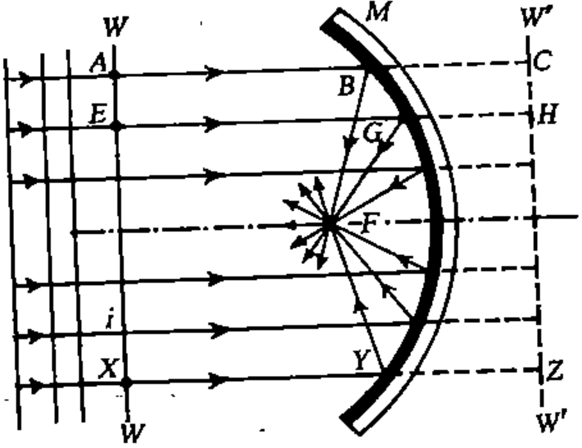


Fig. 2.10: Reflection of light incident on a concave mirror

Fig. 2.10: depicts cross sectional view of parallel rays corresponding to a plane wave *WW* incident on the mirror. The reflected rays converge on *F*. The optical path lengths of all rays reaching *F* must be the same:

$$n_i(\overline{AB} + \overline{BF}) = n_i(\overline{EG} + \overline{GF}) = \dots n_i(\overline{XY} + \overline{YF})$$

Now let the line segments $\overline{AB}, \overline{EG}, \dots, \overline{XY}$ be prolonged through the mirror to points *C, H, …, Z* such that

$$\overline{BC} = \overline{BF}, \quad \overline{GH} = \overline{GF}, \dots, \overline{YZ} = \overline{YF}$$

The two sets of equalities above imply that $\overline{AB} + \overline{BC} = \overline{EG} + \overline{GH} = \dots = \overline{XY} + \overline{YZ}$, which tells us that the distance between *WW* and *W′ W′* through *C, H, …, Z* is constant. We have thus constructed a straight line *W′ W′* such that the points of *M* are equidistant from it and point *F*. By definition, then M is parabolic (with *Focus F*).

# APPENDIX - A

## BOUNADARY CONDITIONS

Let us first consider the components of **E** and **B** fields that are normal to the boundary. We construct a thin Gaussian pill box - extending just a little bit (hair-like) on either side of the boundary of the media, as shown in Fig. A.1.

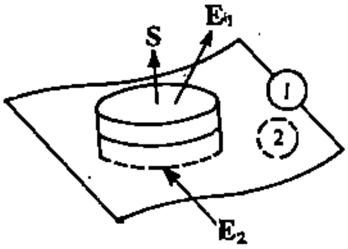


Fig.A.1: The positive direction of S and E is from medium 2 towards medium 1

Eq. (2.1a) implies that

$$\varepsilon\int \mathbf{E}.\,d\mathbf{S} = \varepsilon_1\int \mathbf{E}_1.\,d\mathbf{S} + \varepsilon_2\int \mathbf{E}_2.\,d\mathbf{S} = 0$$

or

$$\varepsilon_1\,\mathbf{E}_1.\,\mathbf{S} - \varepsilon_2\,\mathbf{E}_2.\,\mathbf{S} = 0$$

In the limit thickness of the wafer goes to zero, the edges of the wafer do not contribute. Thus, the components of the electric fields perpendicular to the interface satisfy the condition

$$\varepsilon_1 E_{1\,\perp} - \varepsilon_2 E_{2\,\perp} = 0 \tag{A.1}$$

That is, the normal component of electric displacement is continuous across the boundary.

By a similar argument for normal components of magnetic fields we obtain the following boundary condition from Eq. (2.1b):

$$B_{1\perp} - B_{2\perp} = 0 \tag{A.2}$$

It may be emphasized here that only the normal components of **D** and **B** are equal on both sides of the boundary. Their total magnitudes may not be equal and their directions need not be the same. In fact, these fields may well be reflected or refracted and may also change directions

We now consider the components of two fields parallel to the boundary and apply Eq. (2.1c) to a thin Amperian loop across the surface. This yields

$$\mathbf{E}_1.\,\mathbf{l} - \mathbf{E}_2.\,\mathbf{l} = -\frac{d}{dt}\int_S \mathbf{B}.\,d\mathbf{S} = -\frac{d\phi_B}{dt}$$

where **B** is magnetic flux. As the width of the loop goes to zero, the magnetic flux vanishes. Therefore,

$$(\mathbf{E}_1 - \mathbf{E}_2)\,.\,\mathbf{l} = 0$$

which implies that

$$E_{1\parallel} - E_{2\parallel} = 0 \tag{A.3}$$

That is, the components of **E** parallel to the interface are continuous across the boundary.

In the same way, from Eq. (2.1d) we find that the parallel components of the magnetic field are equal and continuous. Mathematically we write

$$\frac{1}{\mu_1}B_{1\parallel} = \frac{1}{\mu_2}B_{2\parallel} \tag{A.4}$$

# UNIT 3 PERCEPTION OF LIGHT

## Structure

## 3.1 INTRODUCTION

The sense of vision is one of our most prized possessions. It enables us to enjoy splendours of nature, stimulates our thinking and enriches our lives in many ways. We become aware of the infinite variety of objects around us, especially their shapes, colours, textures and motion etc only due to our ability to see them. But have you ever thought: What makes us to see? It all begins with eyes but also depends on what happens behind the eye. Every object viewed is seen with light. Eye responds to illumination. We all know that all living species - from one celled amoeba to the great bald eagle - have developed special mechanisms for responding to light. Human perception of light, i.e. vision is a more developed process. It takes place almost spontaneously without any one, other than the perceiver, knowing what is happening. Perception of light involves formation of sharp images (in the visual apparatus) and their interpretation. Vision begins in the eye, but light is sensed by the brain. In fact, what we see is the world created by our visual apparatus inside our head. So we can say that vision involves a mix of physical and physiological phenomena. You are already familiar with some of the aspects about light and visual systems from your earlier classes. Therefore, you are advised to glance through NCERT physics text book. In this unit we will develop on what you already know. In Sec. 3.2 you will get an opportunity to review internal eye structure and know how light is sensed. Sec. 3.3 is devoted to colour vision where you will learn about dimensions of colour, the trichromatic and opponent-colour theories.

The amoeba reacts only to extreme changes in light intensity such as light and darkness. The earthworms react to light through light sensitive cells present on their skin. This ability to sense only general level of light intensity is termed **photosensitivity.**

### Objectives

After studying this unit, you should be able to

- explain the functions of different parts of the eye
- list common eye defects and suggest remedial measures
- describe how list common eye defects and suggest remedial measures human eye responds to colour, and
- explain trichromatic and opponent-colour theories of colour vision.

## 3.2 HUMAN VISION

Human vision also has a rich relationship with other senses. In fact, all our five senses cooperate and augment each other.

In medicine, the study of structure, functions and diseases of the eye is called **opthalmology**.

Vision involves a mix of physical phenomena and physiological processes. We can understand how the image of an object is formed within the eye purely in terms of physical principles and processes. But from image formation to its perception by the brain, the process is physiological. In this section our emphasis will be on the physics of vision. We shall also discuss very briefly the physiology of vision. Let us begin our study of human vision with the eyes - our windows to the external world. Our eyes are very versatile. They possess a staggering degree of adaptability and precision. They are capable of extremely rapid movement. That is why we can instantaneously shift the focus from a book in hand to a distant star, adapt to bright or dim light, distinguish colours, scan space, estimate distance, size and direction of movement. You may now ask: How vision begins in the eye? What is the internal structure of the eye? How brain interprets images? The answers to such questions have fascinated man for thousands of years. Physiologists say that human eye is an image-making device. (In a way, human eye has striking similarities to a camera of automatic intensity and focal control.) To know the details of mechanism of vision, some knowledge of the visual apparatus is necessary. You will now learn about the structure of eye and how it works as an optical instrument.

Human eyes are very versatile and highly accurate. Their overall visual horizon is broad. But they are less acute than a hawk's eyes and less wide-seeing than those of a deer. Moreover, human eyes are not ideally suited for seeing underwater, nor are they very efficient at night. Even in twilight, eyes lose all colour perceptions.

### 3.2.1 Viewing Apparatus: The Eye

Our eyes, as you know, are located in the bony sockets and are cushioned in fatty connective tissue. The adult human eye measures about 1.5 cm in diameter. Now refer to Fig. 3.1. It shows a labelled diagram of human eye.

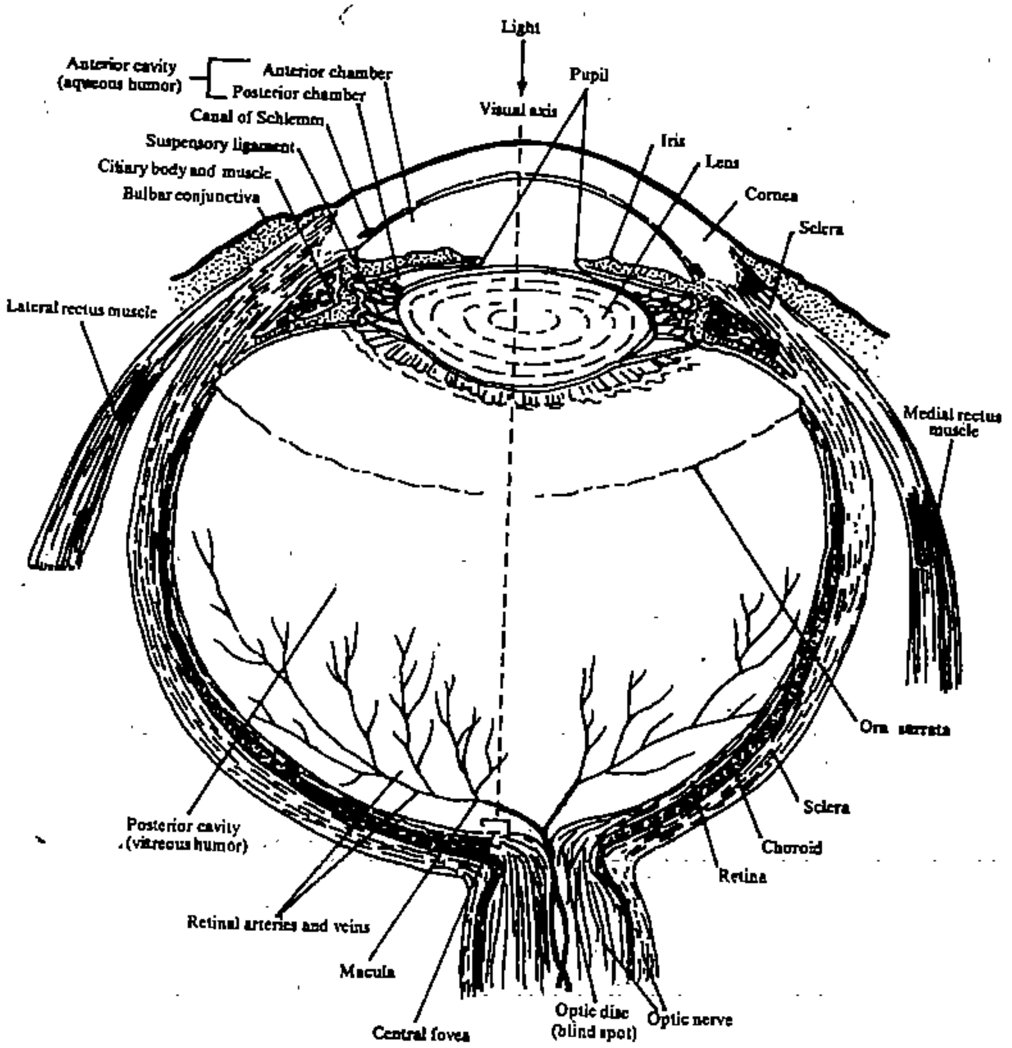


**Fig.3.1: A schematic labelled diagram of human eye**

The **sclera** or 'white' of the eyeball is an opaque, fibro-elastic capsule. It is fairly tough and gives shape to the eyeball, protects its inner parts and withstands the intraocular pressure in the eye. The muscle fibres which control eyeball movement are inserted on the sclera. The **cornea** is a tough curved front membrane that covers the **iris**, the coloured circular curtain in the eye. The cornea acts as transparent window to the eye and is the major converging element.

The cornea is followed by a chamber filled with a transparent watery liquid, the **aqueous humor**, which is produced continuously in the eye. It is mainly responsible for the maintenance of intraocular pressure. Besides this, aqueous humor is the only link between the circulatory system and the eye-lens or cornea. (Neither the lens nor the cornea has blood vessels.) The intraocular pressure maintains the shape of the eye, helps to keep the retina smoothly applied to the choroid and form clear images. Near the rear of this chamber is the iris. The iris is opaque but has a small central hole (aperture), called **pupil**. In our common observation, pupil appears more like a black solid screen. Why? This is because behind the opening is the dark interior of the eye. The size of pupil in normal eye is about 2 mm. The light enters the eye ball through this area. The iris is suspended between the cornea and the lens. The principal function of the iris is to regulate the intensity of light entering the eyeball. When the light is bright, the iris contracts and the size of the pupil decreases and vice versa.

The intraocular pressure maintains the shape of the eye, helps to keep the retina smoothly applied to the choroid and form clear images.

Thread-like suspensory ligaments hold the biconvex crystalline eye-lens, which is just behind the pupil and iris. The muscle responsible for changes in the shape of the lens for near as well as far vision is called the **ciliary muscle.** The eye-lens is an elastic structure made of protein fibres arranged like the layers of an onion. It is perfectly transparent and its focal length is about 3 cm.

The crystalline lens is followed by a dark chamber, which is filled with **vitreous humor**. It is a transparent jelly-like substance. It augments the functions of aqueous humor and helps the eye hold its shape. The rear boundary of this chamber is **retina**, where the image of the object is formed. Microscopic structure of retina is shown in Fig. 3.2(a). It consists of a nervous layer and a pigmented layer. Apart from sensing the shape, and the movement of an object, the retina also senses its colour. The retina consists of five types of **neuronal cells**: the photoreceptors, bipolar, horizontal, amacrine and ganglion neurons. A magnified view of the arrangement of neuronal cells in the retina is shown in Fig. 3.2(b).

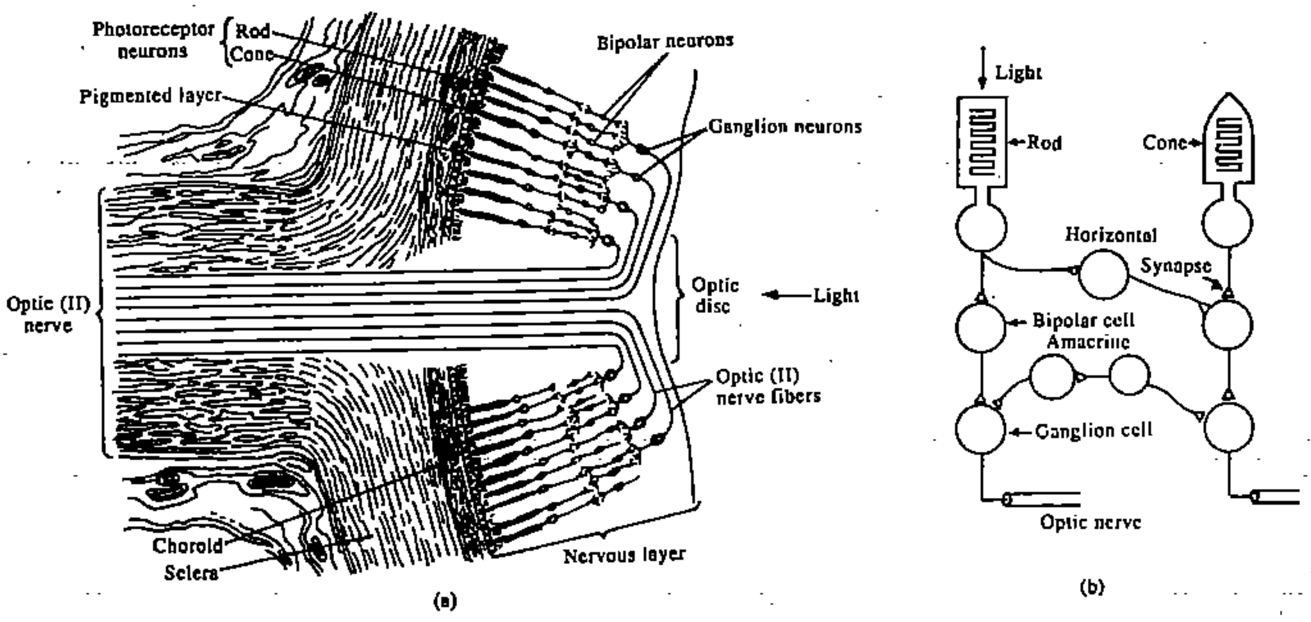


**Fig.3.2: (a) Microscopic structure of retina (b) A Magnified view of arrangement of neuronal cells in the retina**

The photoreceptor neurons are of two types: **rods and cones.** (This nomenclature is due to their geometrical shapes.) It is estimated that about 130 million rods and cones are found lining the retina. Of these, about six million are cones and about twenty times as many are rods. The light sensitive pigments of photoreceptors are formed from vitamin A.

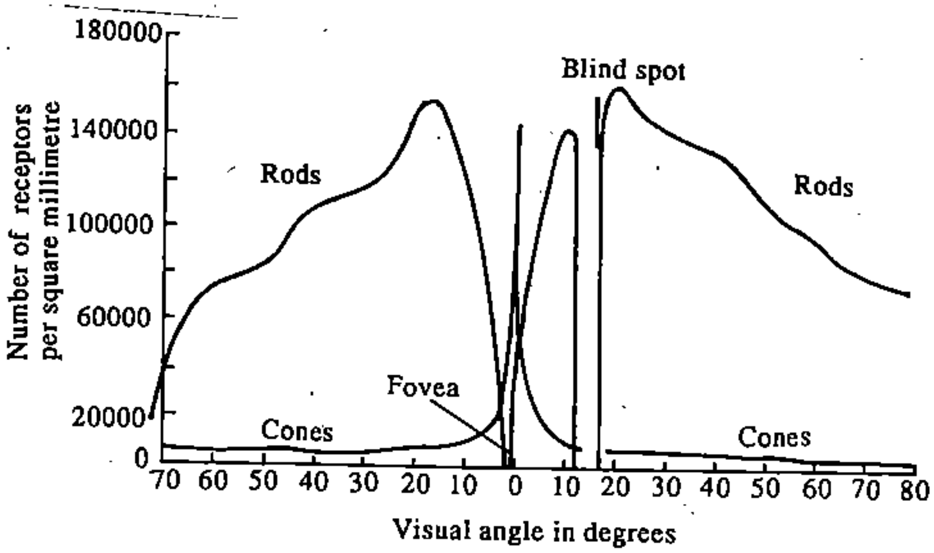


**Fig. 3.3: Distribution of rods and cones in the retina of human eye**

At the very centre of the retina is a small yellowish depression, called **fovea.** This small valley (of about 5mm diameter) contains a large number (~ 110,000 ) of cones and no rods. The distribution of rods and cones across the human retina is shown in Fig. 3.3. The horizontal axis shows the distances in degrees of visual angle from the fovea located at 0°.

Rods are highly specialized for vision in dim light. They enable us to discriminate between different shades of dark and light, see shapes and movements. That is, rods provide a high sensitivity. Cones contain light sensitive pigments which make colour vision and sharpness of vision (high visual acuity) possible. .

When light is absorbed by photoreceptor cells, the light sensitive pigments are broken up by specific wavelengths of light. The bipolar nerve cells carry nerve impulses generated by rods and cones to the ganglion cells. The axons of the ganglion cells converge on a small area of the retina. It is lateral to the fovea and is free from rods and cones. Can you say anything about its ability for vision? Since this area contains only nerve fibres, no image is formed on it. That is, it is devoid of vision. For this reason, it is called the **blind spot.** You may be tempted to ask: Is there a spot in the eye for maximum vision? Certainly yes, the fovea is the valley of the sharpest vision. This remarkable perceptive ability is provided by the cones. Muscles for moving the eye spring from the sclera. The conjunctiva - a supple protective membrane - joins the front of the eye to the inside of the eyelids.

### 3.2.2 Image formation

Before stimulating rods and/or cones, light passes through the cornea, aqueous humor, pupil, eye-lens and vitreous humor. For clear vision, the image formed on the retina should be sharp. Image formation on the retina involves refraction of light, accommodation of eye-lens, constriction of pupil, and convergence of the eyes. We will now discuss these.

**Refraction and Accommodation**

The light entering the eye through the transparent window - **cornea** - undergoes refraction four times. This is because the eye has four optically different media: cornea ($n$ = 1.38), aqueous humor ($n$ = 1.33), eye-lens ($n$ = 1.40), and vitreous humor ($n = 1.34$). Most of the refraction occurs at the air-cornea interface. Can you say why? This is because the cornea has a considerably larger refractive index than air ($n = 1.0$). Moreover, due to the curved shape, the cornea bends the light towards the retina. Additional bending is provided by the eye-lens, which is surrounded on both sides by eye- fluids (Fig. 3.1). However, the power of the lens to refract light is less than that of the cornea. So the main function of lens is to make fine adjustments in focussing. The focussing power of eye lens depends on the tension in the ciliary muscle. When the ciliary muscle is relaxed, the lens is stretched and thinned. When a visual object is 6m or more away from the eye,cornea receives almost parallel light rays. When the eye is focussing an object nearer than 6m, the ciliary muscles contract. As a result, the lens shortens, thickens and bulges and its focussing power increases. These features are illustrated in Fig. 3.4. The great value of the lens lies in its unique ability to automatically change its focal control. This ability is called **accommodation**. Since accommodation means work for the muscles attached to the eye lens, viewing an object nearer than 6m for a long time can cause eye- strain.

While a healthy cornea is transparent, disease or injury may result in blindness. But eye surgeons have now acquired competence in replacing damaged cornea with clear one from human donors. Any imperfection in the shape of the cornea may cause distortion in visual images.

The eye-lens of elderly people tends to be less flexible and loses ability to accommodate. This condition is called **prestyopia**. For extra focussing power, they use glasses (spectacles or contact lens).

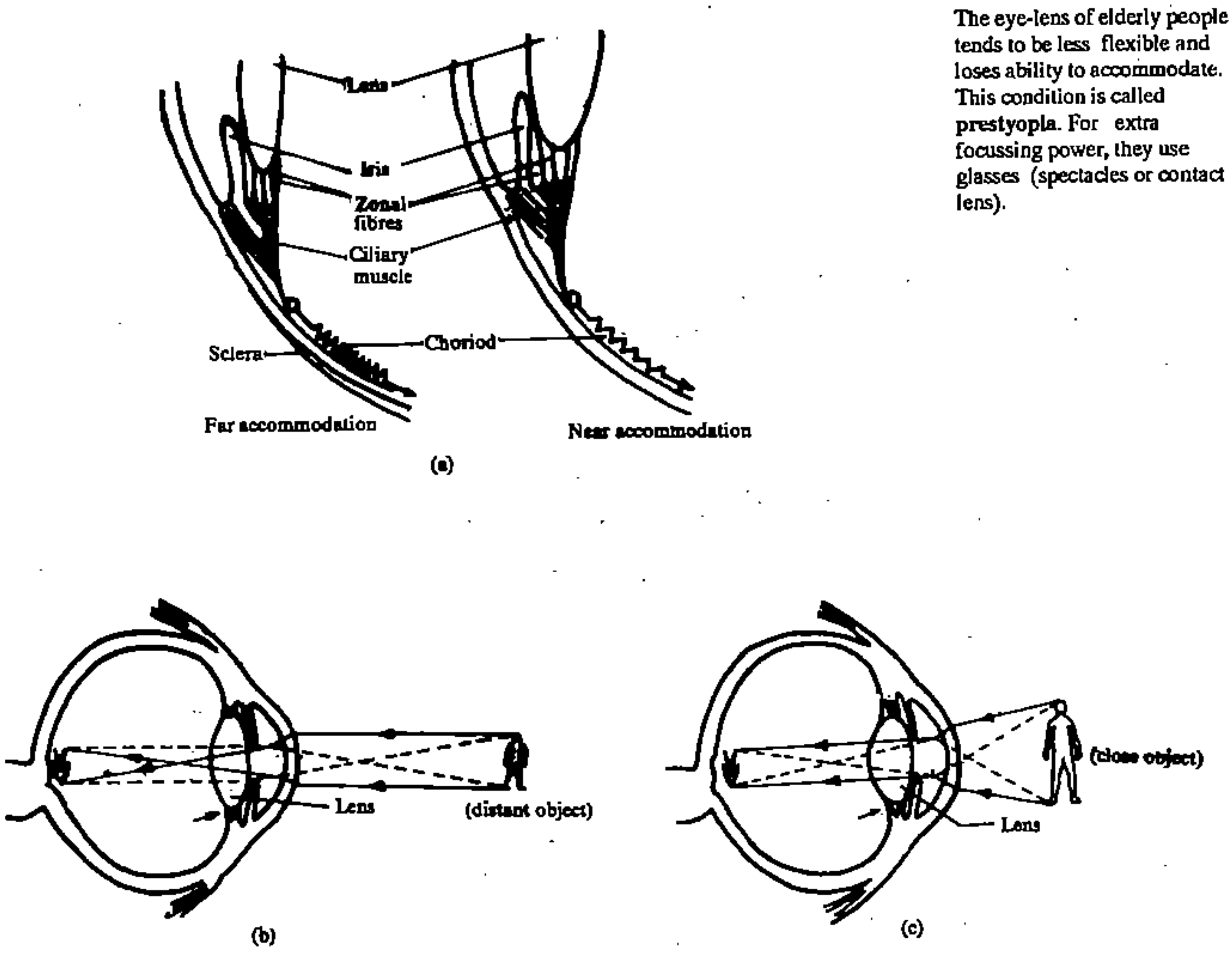


**Fig. 3.4: Far and near accommodation (a) In the diagram on the left, the ciliary muscle is relaxed. This causes the eye-lens to curve less. In the other diagram, the ciliary muscle is contracted. This causes the lens to curve more. (b) Accommodation for far vision (6m or more away). (c) Acommodation for**

### Constriction of Pupil

Constriction of the pupil means narrowing down of the diameter of the hole through which light enters the eye. This action occurs simultaneously with accommodation of the eye-lens and prevents the entry of light rays through the periphery of eye-lens, which can result in blurred vision. The pupil also constricts in bright light to protect the retina from sudden or intense stimulation. (When the level of illumination is low, the pupil dilates so that the retina can receive enough light.)

## Convergence

Human beings have single binocular vision. This signifies that both eyes focus on only one set of objects. When we stare straight ahead at a distant object, the incoming light rays are directed at both pupils, get refracted and are focussed at identical spots on the two retinas. Suppose that we move close to the object and keep our attention on the same stationary object. Our common sense suggests that even now images should form on the same points (in both retinas). It really does happen and our eyes automatically make adjustments by radial movement of two eyeballs. This is referred to as **convergence**.

Refer to Fig. 3.4 again. You will note that the images formed on the retina are inverted laterally as well as up-side-down. But in reality we do not see a topsy-turvy world. You may now ask: How does this happen? The solution to this apparent riddle lies in the capacity of the brain which automatically processes visual images. This suggests that though vision begins in the eye, perception takes place in our brain. Its proof lies in that severe brain injury can blind a person completely and permanently, even though eyes continue to function perfectly.

You may now like to reflect on what you have read. So you should answer the following SAQ before you proceed.

*Spend 5 min*

**SAQ 1**

Human beings are unable to see under water. Discuss why?

By now you must be convinced that mechanically speaking, human eye is an optical instrument resembling a camera. (A better analogy exists between the eye and a closed circuit c ɔur TV system.) The eye-ball has a light focussing system (cornea and lens), apert e (iris) and a photographic screen (retina). This is shown in Fig. 3.5. There are o: course very important differences between our eye and a camera. The engineering sophistication of human eye is yet to be achieved even in the costliest camera. The camera-man has to move the his camera lens for change of focus, whereas the eye-lens has automatic intensity and focal control. (The brain constantly analysis and perceives visual images. This is analogous to the development of a photograph.) The image on the retina is not permanent but fades away after 1/20th of a second and overlaps with the next image. This gives the impression of continuity. There is of course no film in the eye that records the images permanently as a photo film does.

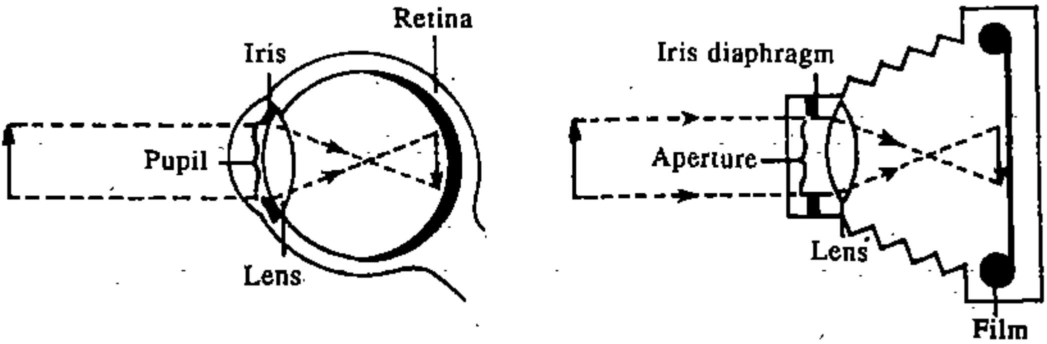


**Fig.3.5: The similarity between the eye and the camera**

### 3.2.3 Information Processing

As soon as light impulses impinge on the retina (and an image is formed), these are absorbed by rods and cones, which contain four kinds of photosensitive substances. These **visual pigment molecules** undergo structural (chemical) changes. It is believed that each rod cell contains about seventy million molecules of a purple-coloured photosensitive pigment, **rhodopsin**. Like rods, cones contain violet - coloured photosensitive pigment, **iodopsin**.

Each pigment molecule consists of two components: a colourless protein, **opsin**, and a coloured chromophore, **retinene**. Opsin is different for each of the four visual pigments and determines the frequency of light to which each pigment responds.

Rhodopsin has a molecular weight of about 4 x $10^4$ dalton. It consists of the **scotopsin** protein and the chromophore retinene, a derivative of vitamin A in the form called cis-retinene. Any deficiency of vitamin A causes **night-blindness**. Fig. 3.6 shows the absorption curve of rhodopsin.

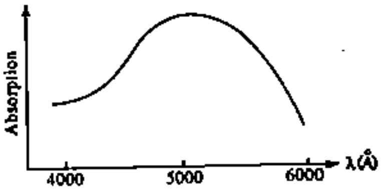


**Fig. 3.6 The absorption curve of rhodopsin**

Let us now understand as to what happens to rhodopsin in rods. (The same basic changes occur in the visual pigments in cones.) Refer to Fig. 3.7 which depicts the rhodopsin cycle. The first step in this process is the absorption of photon by rhodopsin, which then undergoes a chemical change. Its cis-retinene portion changes to **all-trans-retinene** On referring to Fig. 3.8 you will note the rotation that occurs around the carbon numbered 12. This change triggers decomposition of rhodopsin (into scotopsin and all-trans-retinene) by a multi-stage process known as bleaching action. The pigment loses colour and the visual excitory event is believed to occur. Then rhodopsin is resynthesized in the presence of vitamin A. In this process, an enzyme, **retinene isomerase**, plays the most vital role.

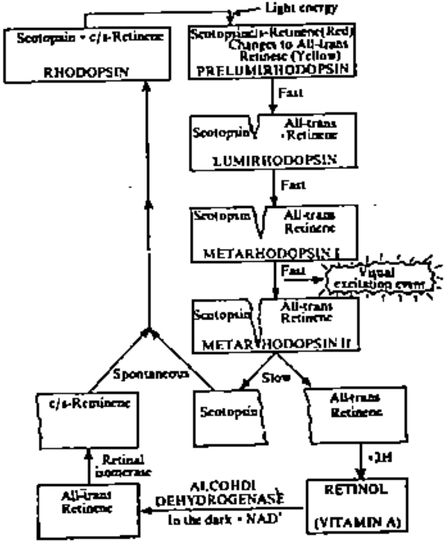


**Fig. 3.7: The rhodopin cycle: Bleaching action**

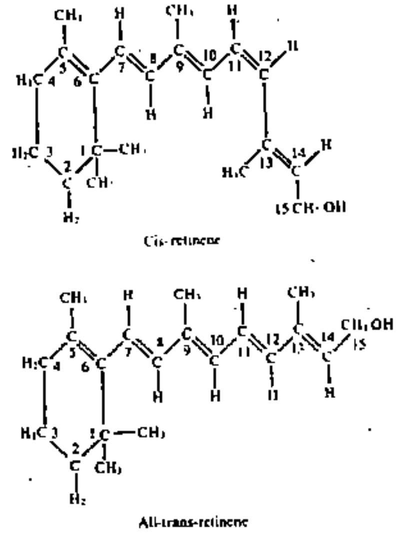


**Fig.3.8: Structures of cis-retinence and all-trans- retinence**

The rods respond even to poor illumination such as twilight. Rhodopsin is highly sensitive to even small amounts of light. Their responses to light generate colourless images and objects are seen only in shades of grey. It is for this reason that you see a red flower black in the evening. On the other hand, the pigments of the cones are much less sensitive to light and require bright illumination to initiate decomposition of chromophore. Visual acuity or ability to see clearly and to distinguish two points close together is very high and their responses produce coloured images.

The information received in terms of light is converted into electrical signals in the retina. The potential of the cell membranes of the photoreceptor cells undergoes a change even on brief illumination. This occurs through a complex chemical process involving a flow of calcium ions and sodium ions across the membrane. The change in membrane potential, $\Delta V_m$, is governed by the following equations in time and space :

$$\Delta V_m(t) = I_m R(1 - e^{-t/\tau}) \tag{3.1}$$

and

$$\Delta V_m(x) = V_0 e^{-x/L} \tag{3.2}$$

where $I_m$ is the membrane current, $R$ the membrane resistance, $\tau$ is the membrane time constant. $V_0$ is the change in the membrane potential at $x = 0$ ($x$ being the distance away from the site of current injection) and $L$ is the length constant. As can be seen, the spread of $\Delta V_m$ in space is governed by $L$ (whose values fall in the range of about 0.1 to 1 mm). It is important to note that while slow potentials are generated in most cells, action potentials are produced only in the ganglion cells. The signals generated in the retina are further transmitted to the higher centres in the visual pathway of the brain such as lateral geniculate nucleus and visual cortex. In this way, precise information about the image projected on the retina is conducted accurately to the brain. The transfer of visual information in a typical retinal circuit is shown (Fig. 3.9).

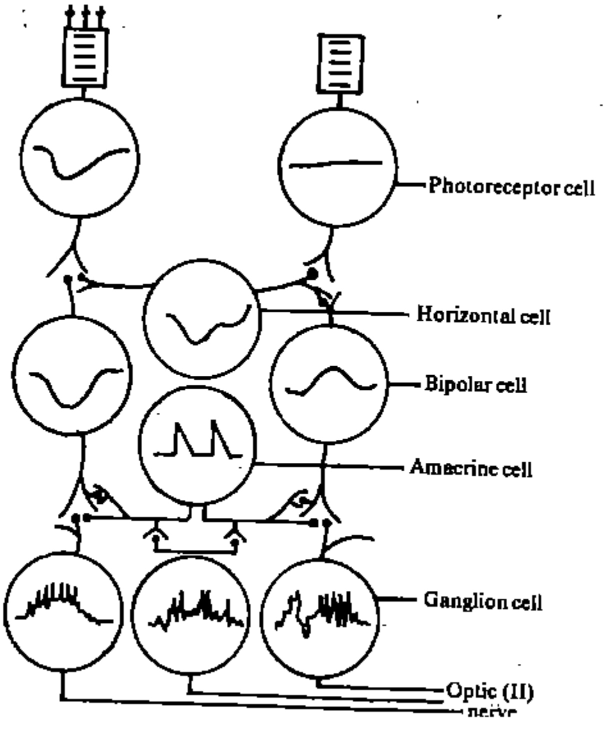


Fig.3.9: Retinal circuit showing the electrical links between cells of the retina: Action potential

We hope that now you have a reasonable idea of how we perceive the world around us. You may now like to know the factors that hamper vision.

### 3.2.4 Defects of Vision

Sometimes the eye loses its power of accommodation. When this happens, we are unable to see objects clearly and vision becomes blurred. These are corrected by using contact lenses or spectacles.

In one kind of such a defect, human beings can see nearby objects clearly but it is difficult to see objects at long distances. In such a (defective) eye, the image of distant objects is formed in front of the retina (Fig. 3.10a) rather than on the retina. This defect of the eye is known as **shortsightedness** or **myopia.** It is frequently observed in children and its occurrence is fast increasing in our country. In shortsightedness, the eyeball gets elongated. It can be corrected by using a concave (divergent) lens (Fig. 3.10b) of appropriate focal length which moves the image on to the retina.

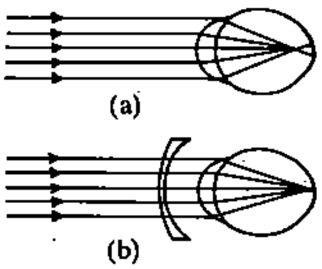


**Fig.3.10: (a) Short sightedness (b) its correction**

In another eye defect, eyeball gets shortened. Though distant objects are seen clearly, nearby objects look blurred. In this case the image is formed behind the retina (Fig. 3.11a). This defect is known as **longsightedness or hypermetropia.** It is normally observed in elderly people. It can be corrected by using a convex (convergent) lens of appropriate focal length (Fig. 3.11b).

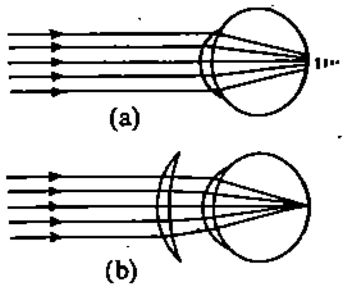


**Fig. 3.11: (a) Long sightedness (b) its correction**

Sometimes a person may suffer from both myopia and hypermetropia. Such people often use bifocal lenses, in which one part of the lens acts as a concave lens and the other part as a convex lens. The third type of defect of vision is called **astigmatism,** wherein distored images are formed. The corrective lenses are used to restore proper vision.

## 3.3 COLOUR VISION

You all know that human beings have remarkable sense to adore the varied creations of nature. This is particulary because colour is an automatic part of our perception. In fact, the phenomenon of colour vision has added real charm in life. Can you now realize what vision is like without colour? You will learn that the colour is a perceptual experience; a creation of the eye and the mind.

One of the earliest observations about colour perception was made in 1825 by Purkinje. He observed that at twilight, blue blossoms on flowers in his garden appeared more brilliant than the red. To understand this you must know the mechanism of colour vision. The process of colour perception is influenced by the physiology of the eye and the psychology of the person. Before we plunge into these details, it is important to know the dimensions of colour, i.e., the parameters with which colour may be defined.

### 3.3.1 The Dimensions of Colour

The most important physical dimension of colour is the wavelength of light. For most light sources, what we perceive is the dominant colour, which we call the **hue.** It is hue to which we give the names like red, blue or greenish yellow. In fact, the terms colour and hue are frequently used interchangeably. You may therefore conclude that **hue is the perceptual correlate for variations in wavelength.**

The second dimension relevant to colour vision is **illuminance,** which refers to the amount of light reaching the eye directly from the source. Illuminance, therefore, characterizes the perceived brightness of a coloured light. This relationship (between illuminance and brightness) is fairly complex because perceptual sensitivity varies with the wavelength of light. Every individual with normal eye possesses maximum sensitivity to light between the green and yellow parts of the spectrum (500nm - 600 nm). And the sensitivity to predominantly blue light (400 - 500 nm) is rather low.

Another physical dimension associated with colour is the degree of **purity** of spectral composition. That is, purity characterises the extent to which a colour (hue) appears to be mixed with white light. This is responsible for variations in the perceived saturations of the colour. For example, when we add white light in a spectrally pure blue, the light begins to look sky- blue. On progressive addition of white light you may eventually observe it as white.

We may therefore conclude that

> Colour, as a perceptual phenomenon, is three dimensional and is characterised by hue, saturation and brightness.

Intensity is defined as the amount of energy reaching a receiver of given cross-sectional area every second.

Thinking logically, you may now ask: Is there any other alternative expression for the dimensions of colour ? The answer to this question is: Yes, there is. It is based on the observation that colour depends on intensity of light. Let us now learn about it in some detail.

### Trichromacy

You must have realised sometimes that when intensity of light is low, we see no colours. You also know that by varying the wavelengths and/or intensities of lights of different colours, it is possible to produce light of a desired colour. In your school you must have learnt that all the colours of the visible spectrum can be produced by mixing lights of just three different wavelengths: red, green and blue. These are known as **primary** colours. The explanation for this trichromacy lies in the mechanism for colour vision. You will learn about it in the next sub- section.

Another important phenomenon associated with colour vision is complemantarity of colours i.e. pairs of colours, when mixed, seem to annihilate one another. For example, when we mix suitable proportion of a monochromatic blue light (λ~ 470 mm) with a monochromatic yellow light ( λ ~ 575 nm ), we obtain a colourless grey.

Reflecting on this observation, Hering suggested that complementary pairing is an indicator for pairing in the mechanisms responsible for signalling colour in the visual system. The complementary relationships among pairs of colours can be well represented as shown in Fig. 3.12. To locate the complementary colour in this figure all that you have to do is to choose any point and draw a line passing through the centre of the circle. A suitably adjusted mixture of two complementary colours will appear grey.

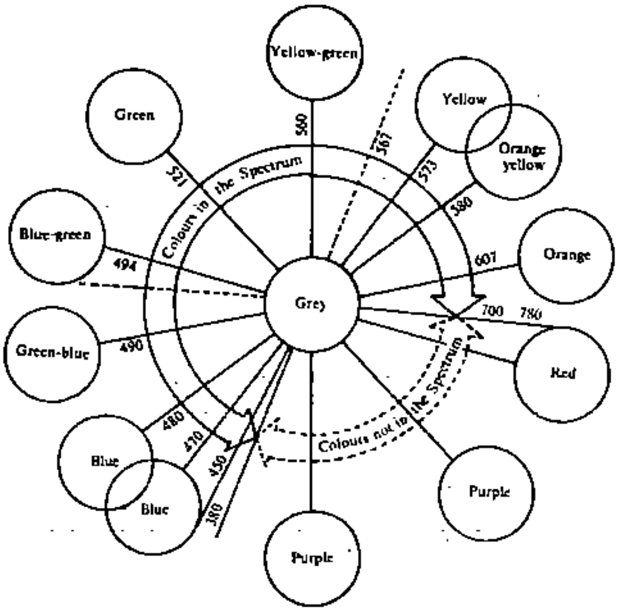


Fig. 3.12: The complementary colour circle

Before you proceed further, we want you to pause for a while and answer the following SAQ.

*Spend 5 min*

**SAQ 2**

How would you indicate brightness and saturation in Fig. 3.12?

Note the presence of 'purple' hues. You may recall that dispersion of white light by a prism does not reveal this hue. Then the question arises: What is their significance in the colour circle? The complementary circle will remain incomplete without them.

You may also note that though colour circle represents colours as a continuum, primary colours are perceptually quite distinct. The phenomena of primary colours and trichromacy led Young to propose three different types of receptors (cones) for colour vision. You will learn the details as you proceed.

### Colour Blindness

You now know that a single monochromatic light can be produced by combining two primary colours. The measurements made to know the amounts of these colours required to match a given monochromatic colour gave fairly standard results. That is, when we ask a group of people to match a test colour, experience tells that they mix the same proportions of primary colours. But colour-mixing requirements for some individual may be anomalous. In fact, some individuals may need only two, rather than three, primary colours to match all the monochromatic hues. These anomalies are indicative of varying degrees of **colour blindness.** People who show anomalous colour-matching requirement do not see the same colours as individuals having normal vision. The most common defect is in the proportions of red and green lights required to match a monochromatic yellow. The manifestation of this in everyday life is a limited ability to distinguish between red and green.

## 3.3.2 Colour Receptors

In the above paragraphs you have learnt that trichromatic theory led Young to propose that eye possesses three types of cones, each containing a different pigment. And three types of pigments in the cones correspond to three primary colours (three- dimensional colour vision). The absorption curves for these pigments are shown in Fig. 3.13. You will note that the curves show substantial overlap. Moreover, the blue mechanism is markedly less sensitive than the other two.

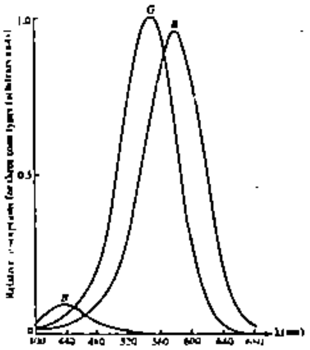


**Fig. 3.13: Spectral absorption curves for three different cone pigments**

The argument leading to this conclusion is rather subtle and needs closer analysis. To understand this, let us as:: How do humans distinguish such a large number of colours? Do we need a different type of receptor to discriminate each colour? Since the colours are numerous, the number of receptors available for a particular colour will be a small fraction of the total number of colour receptors. When monochromatic light reaches our eye, only the corresponding class of receptors will respond. And since the total number of responding receptors is comparatively small, the ability to see a monochromatic light will be much less than the ability to see white light. But in practice, this is not true. This led Young to,conclude that only a few different types of receptors are present, which by working in combination give rise to all the different colours we perceive. His experience with colour mixing led him to conclude that the number of receptor types is only three.

This theory was proposed even before very little was known about the physiology of the visual system. The outputs from the three types of receptors are transmitted separately to the brain which combines the information and constructs certain abstractions to which we give names like hue, saturation, yellow, blue etc.

We all know that yellow gives a sensation independent of red, blue and green, i.e. it seems as much of a primary colour. But no coding system is postulated for yellow in the trichromatic theory. Such feelings, though subjective, led Hering to propose an alternative theory of colour based on four colours: red, yellow, green and blue. This is known as **opponent-colour theory**. These colours are associated in pairs: red-green and blue-yellow. The members of a pair are thought to act in opposition adding upto white. Hering also specified a third pair of black and white to represent the varying brightness and saturation of colours. (The perception of brightness of the colour also depends on the mood of the perceiver.) You must appreciate that the most important difference between this theory and the trichromatic theory lies not in the number of postulated receptor types, but in the way their outputs are signalled to the brain. Fig. 3.14 depicts a simple version of the opponent-process theory. Three basic receptor types are indicated by *X*, *Y* and *Z*. Mixture of *Y* and *Z* is perceived as yellow. White is obtained by mixing *X*, *Y* and *Z*.

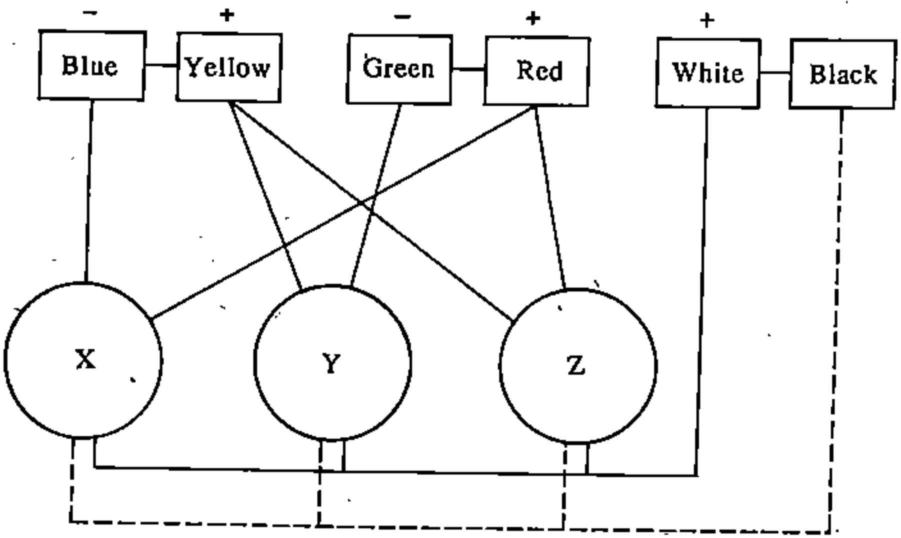


**Fig.3.14: Opponent-process theory based on Hering's postulate. *X*, *Y* and *Z* denote basic receptor types.**

According to the model shown in Fig. 3.14, three different receptor types are each sensitive to a range of wavelengths. The mode of operation is such that the activity level increases in response to a predominant input about one colour. You may ask: What happens in response to the input about the complementary colour? We expect it to decrease. To illustrate it, let us consider that the input to the blue-yellow system is predominantly in the yellow region of the spectrum. Then, there is an increase in activity (over a spontaneous level) about yellow colour. On the other hand, if the input is predominantly blue, there is a decrease in activity. Activity in the black-white mechanism is based on outputs from all three receptor types.

Even though trichromatic theory and the opponent- process theories appear conflicting, recent studies show evidences that they are compatible. Researches at the Johns Hopkins University (US) provide evidence in favour of trichromatic theory. However, the cones do not send 'color signals' directly to the brain. Cone signals pass through a series of neurons which are colour specific.

Vision is an endlessly fascinating area . We here conclude saying: Eye is not merely an instrument for survival; it is a means for enrichment of life.

We will now like you to answer the following SAQ.

SAQ 3

*Spend 5 min*

Name the regions of retina specialized for (a) colour and detailed vision at high levels of illumination and (b) non-colour vision at low levels of illumination.

## 3.4 SUMMARY

- Perception of light involves formation of sharp images in the eye and their interpretation in the brain. That is, vision involves a mix of physical and physiological phenomena.
- Human eyes are image making devices. They have striking similarities to a camera of automatic intensity and focal control. There are however differences in details.
- Cornea is the major converging element in the eye.
- The image of an object is formed on the retina. It consists of five types of neuronal cells: photoreceptors, bipolar, horizontal, acrine and ganglion neurons.
- The photoreceptor neurons are of two types: rods and cones. Rods are specially suited for vision in dim light and provide high-sensitivity. Colour vision and sharpness are possible due to cones.
- Image formation on the retina involves refraction of light, accommodation of eye-lens, constriction of pupil and convergence of the eyes.
- Information processing involves structural changes in photosensitive pigment rhodopsin by bleaching action.
- Two common defects of eye are Myopia (short sightedness) and hypermetropia (long sightedness). These are corrected by using a concave and a convex lens respectively.
- Colour, as a perceptual phenomenon, is three dimensional: hue, illuminance and purity.
- According to Young's trichromacy theory, colour vision requires three types of receptors (cones) for three primary colours.
- According to Hering's opponent - colour theory, colours are associated in pairs: red-green, blue-yellow and add up to white. The brightness and saturation are determined by a black-white pair.

## 3.5 TERMINAL QUESTIONS

1. List the differences between the human visual system and a camera.
2. When we enter a dark room, we feel blinded. Gradually we become dark adapted. The dark adaptation curve shown here shows a kink. Can you suggest an explanation in terms of rod and cone-adaptation ?

## 3.6 SOLUTIONS AND ANSWERS

**SAQs**

1. The refractive indices of water and cornea are 1.33 and 1.38, respectively. Due to small difference in these values, cornea is unable to bend light towards the retina. This is why humans are unable to see under water.
2. An arrow originating at the centre and directed towards the circumference would indicate increasing colour saturation. Brightness does not depend on hue and saturation. So a line drawn normally out of the page (towards you) would represent increasing brightness.
3. The first description applies to the fovea whereas the second description applies to the peripheral regions.

## TQ's

1. Some of the main differences are tabulated below. You can add more if you have thought of others.

| Camera | Eye |
|---|---|
| Lens is responsible for focussing. | Cornea as well as lens is involved in focussing; lens is responsible for fine adjustments |
| Lens is rigid and fixed. Fine focus is achieved by changing lens and/or by changing distance between lens and film. | Lens is soft and flexible. Fine focus is achieved by alterations in convexity of lens. |
| Only sophisticaled cameras have automatic aperture adjustments. | Pupil adjustment is an automatic response. |
| The brightness of a photograph depends directly on the level of illumination. | The brightness of a perceived scene depends on prevailing illumination as also the lighting level to which the eye has been previously exposed. |
| Light-sensitive substances do not regenerate once film has been exposed. | Light-sensitive substances are constantly regenerating. |
| Image is fixed. | Image is in constant motion (owing to eye movements). |
| Information stored in photographic form is not immediately transmitted. | Information on retina is automatically processed and the results are immediately transmitted to the brain. |

2. The principal mechanism for dark adaptation is regeneration of bleached visual pigments, in partcular rhodopsin. So, the first part of the curve signifies the foveal adaptation, due to cone cells. It levels off at the kink. And the second part of the curve represents the contribution of rods.

# UNIT 4 POLARISATION OF LIGHT

## Structure

## 4.1 INTRODUCTION

In Unit 1 of this block, you learnt that light is a transverse electromagnetic wave. In your school physics curriculum you have learnt that while every wave exhibits interference and diffraction, polarisation is peculiar only to transverse waves. You may even be familiar with basics of polarisation like: What distinguishes the polarised light from unpolarised light? Is light from an ordinary (or natural) source polarised? How do we get polarised light? and so on. In this unit we propose to build upon your this preliminary knowledge.

You must have seen people using antiglare goggles as also antiglare windshields for their cars. Do you know that polarisation of light has something to do with these? Polarisation of light also plays a vital role in designing sky light filters for cameras and numerous optical instruments, including the polarising microscope and polarimeter. You may get opportunity to handle some of these devices if you opt for physics laboratory courses PHE-08(L) and PHE-12(L).

In Sec. 4.2 we have discussed as to what is polarisation. In Sec. 4.3, you will learn about simple states of polarised light. Sec. 4.4 is devoted to ideal polarisers and Malus' law. In this section you will also learn about double refraction or optical birefringence - a property of materials helpful in producing polarised light. In Sec. 4.5, you will learn about some techniques of producing circularly and elliptically polarised light.

### Objectives

After going through this unit you should be able to

- explain what is linearly, circularly or elliptically polarised state of light
- describe how can light be polarised by reflection
- solve simple problems based on Malus' law and Brewster's law
- explain how optical birefringence helps in production of polarised light, and
- explain the production of linearly polarised light by dichroism.

## 4.2 WHAT IS POLARISATION?

What is polarisation? Why light, not sound, waves are known to polarise? These are some of the basic questions to which we must address ourselves. Polarisation is related to the orientation (oscillations) of associated fields (particles). Refer to Fig. 4.1 which depicts a mechanical wave (travelling along a string). From Fig. 4.1(a) you will note that the string vibrates only in the vertical plane. And vibrations of medium particles are confined to just **one** single plane. Such a wave is said to be (plane) polarised. How would you classify waves shown in Fig. 4.1(b) and (c)? The wave shown in Fig. 4.1(b) is plane polarised since vibrations are confined to the horizontal plane. But the wave in Fig. 4.1(c) is unpolarised because simultaneous vibrations in more than one plane are present. However, it can be polarised by placing a slit in its path as in Fig. 4.1(d). When the first slit is oriented vertically, horizontal vibrations are cut off. This means that only vertical vibrations are allowed to pass so that the wave is linearly polarised. What happens when a horizontal slit is placed beyond the vertical slit in the path of propagation of the wave? Horizontal as well as vertical components (of the incident wave) will be blocked. And the wave amplitude will reduce to zero.

Let us now consider visible light. The light from a source (bulb) is made to pass through a polaroid ( $P$ ), which is just like slit one in Fig. 4.1. The intensity of light is seen to come down to about 50%. Rotating $P$ in its own plane introduces no

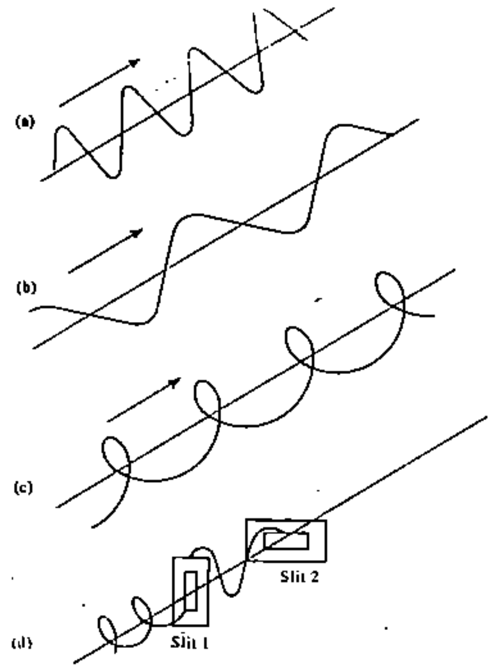


**Fig. 4.1: (a) A vertically plane polarised wave on a string (b) A horizontally plane polarised wave (c) an unpolarised wave. (d) The wave in (c) becomes plane polarised after passing through slit one; the wave amplitude reduces to zero if another slit oriented perpendicular to slit one is introduced.**

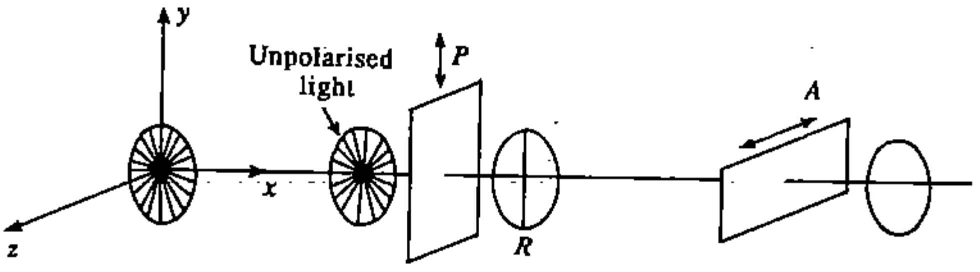


**Fig. 4.2: Schematics of the apparatus for observing polarisation of light**

further change in light intensity. Now if a second identical polaroid ($A$) is introduced in the path of light so that it is parallal to $P$, the intensity of light from the bulb remains unaffected. But rotating $A$ in its own plane has a dramatic effect! For 90° rotation, the light is nearly cut-off.

You can analyse this result in terms of electromagnetic theory, which demands complete description of associated electric vector and the way it oscillates with respect to the direction of propagation. For the arrangement shown in Fig. 4.2, the electric vector at the source has all orientations in the $yz$ plane. The wave propagates as such till it reaches the polaroid $P$, which allows essentially unhindered passage of electric vector oriented parallel to its transmission axis. If the transmission axis is along $y$-axis, the electric field along $y$-direction ($E_y$) passes through it unaffected. In addition, the $y$-components of electric field vectors inclined to $y$-axis can also pass through $P$. Thus, after passing through the polaroid $P$, the electric vectors oriented only along $y$-axis will be present. **When electric vector oscillates along a straight line in a plane perpendicular to the direction of propagation, the light is said to be plane polarised.** The plane polarised wave further travels to the polaroid $A$, which is identical to $P$. When $A$ is at 90° with respect to $P$, it can allow only the $z$-components of E to pass. Since only $y$-components of E are present in the wave incident on $A$, no light is transmitted by $A$.

We may now conclude that

1. No polarisation of longitudinal waves occurs as the vibrations are along the line of transmission only.
2. The transverse nature of light is responsible for their polarisation.

An important menifestation of this result arises in TV reception. You may have seen that the TV antenna on your roof-tops are fixed in horizontal position. Have you ever thought about it? This is because the **TV signal transmission in our country** is through horizontally oriented transmitting antenna. The explanation for this lies in the observation that the pick up by the receiving antenna is maximum when it is oriented parallel to the transmitting antenna. This is illustrated in Fig. 4.3 for a vertical (dipole) transmitting antenna.

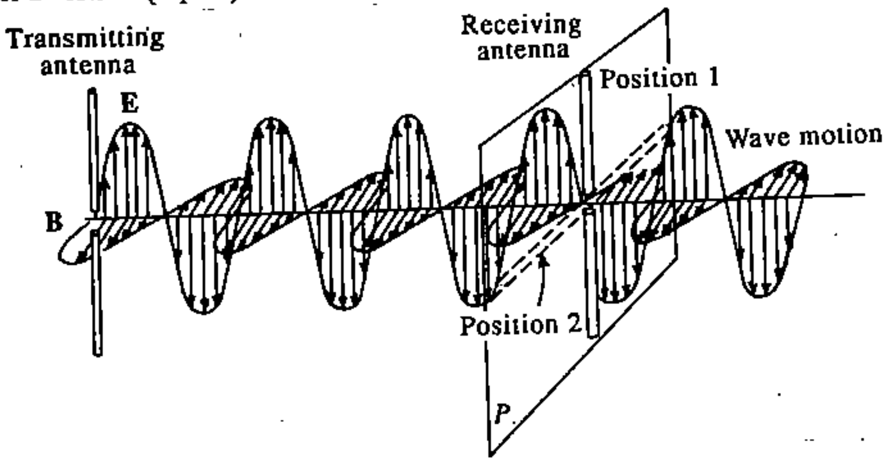


**Fig. 4.3: Polarisation of an electromagnetic wave. The antenna responds to the vertical electric field strength of the wave. Reception is maximum in Position 1 and minimum in Position 2.**

You may now like to know: Do natural (or ordinary) light sources emit polarised light? Answer to this question is 'yes' as well as 'no'! Is this answer not funny? You know that emission of light involves a large number of randomly oriented atomic (or molecular) emitters. Every individual excited atom radiates polarised waves for about $10^{-8}$ s. These waves form a resultant wave of given polarisation which persists for the lifetime of the excited atom. At the same time, other atoms (molecules) also emit waves, whose resultant states of polarisation may be quite different. Because of this randomness, every orientation of electric vector in space is equally probable. That is, electric vectors associated with light waves from a source are oriented in all directions in space and thus there is a completely unpredictable change in the overall polarisation. Moreover, due to such rapid changes, individual resultant polarisation states become almost indiscernible. The light is then said to be **unpolarised**.

In practice, visible light does not correspond to either of these extremes. The oscillations of electric field vectors are neither completely regular nor completely irregular. That is, light from any source is partially polarised. We ascribe a degree of polarisation to partially polarised light. The degree of polarisation is one for completely polarised light and zero for unpolarised light.

The next logical step perhaps would be to know various types of polarised light. Let us learn about this aspect now.

## 4.3 SIMPLE STATES OF POLARISED LIGHT

In a right handed coordinate system if a right handed screw is turned so that it rotates the $x$-axis towards the $y$-axis, the direction of advance of the screw represents the positive $z$-axis.

You now know that in e.m. theory, light propagation is depicted as evolution of electric field vector in a plane perpendicular to the direction of transmission. For unpolarised light, spatial variation of electric field at any given time is more or less irregular. For plane polarised light, the tip of electric vector oscillates up and down in a straight line in the same plane. The space variation of **E** for linearly polarised wave is shown in Fig. 4.4 (a). The diagram on the left shows the path followed by the tip of the electric vector as time passes. You will know that the tip of **E** executes one full cycle as one full wave length passes through a reference plane. There are two other states of polarisation: circular polarisation and elliptical polarisation. The path followed by the tip of E, as the time passes, for these is shown in Fig. 4.4 (b) and (c), respectively.

The $yz$-plane (or $x = 0$ plane) in Fig. 4.4 is the plane of polarisation of the wave. We can identify other states of polarisation by looking at the **trajectories of the tip of the electric field vector** as the wave passes through the reference plane. You should always look at the reference plane from the side away from the source (looking back at the source) for the definitions to be unique.

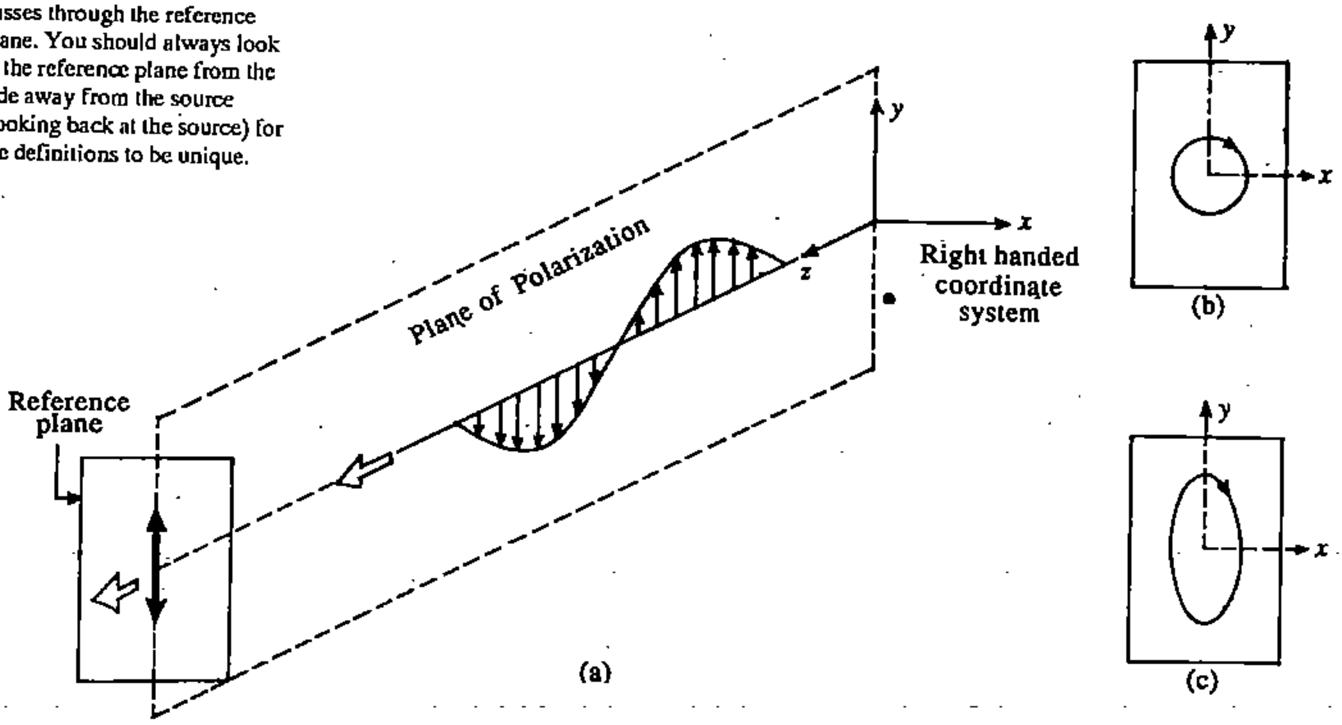


**Fig. 4.4: Spatial variation of electric field vector for (a) linearly polarised light. The diagram on the left show the path taken by the tip of the electric vector as time veries. (b) and (c) show the path taken by the tip of the electric vector for circularly and elliptically polarised light.**

Let us now mathematically analyse how superposition of two plane polarised light waves of same frequency moving in the same direction gives rise to linearly, circularly or elliptically polarised light.

### 4.3.1 Linear Polarisation

Suppose that two light waves are moving along the $z$-direction. Let their electric field vectors be mutually perpendicular, i.e. we choose these along the $x$ and $y$ axes and can represent them respectively in the form

$$\mathbf{E}_1(z,t) = \hat{e}_x E_{01} \cos(kz - \omega t) \tag{4.1}$$

$E_2(z,t)$ lags $E_1(z,t)$ for $\varphi > 0$ and vice-versa.

and

$$\mathbf{E}_2(z,t) = \hat{e}_y E_{02} \cos(kz - \omega t + \phi) \tag{4.2}$$

Here $\hat{\mathbf{e}}_x$ and $\hat{\mathbf{e}}_y$ are unit vectors along the $x$ and $y$-axes respectively. (These are also called polarisation vectors.) $\phi$ is the phase difference between the two waves.

We expect that the nature of the resultant wave will be determined by the phase difference between them and the value of the ratio $E_{02}/E_{01}$. Mathematically, we can write the vector sum of these as

$$\mathbf{E}(z,t) = \mathbf{E}_1(z,t) + \mathbf{E}_2(z,t)$$

$$= \hat{\mathbf{e}}_x E_{01} \cos(kz - \omega t) + \hat{\mathbf{e}}_y E_{02} \cos(kz - \omega t + \phi) \tag{4.3}$$

Resultant Amplitude

$$= \sqrt{(E_0)^2 + (E_0)^2}$$

$$= \sqrt{2}\, E_0$$

and

$$\tan\theta = \frac{E_{02}}{E_{01}} = \frac{E_0}{E_0} = 1$$

$$= \tan 45^\circ$$

or

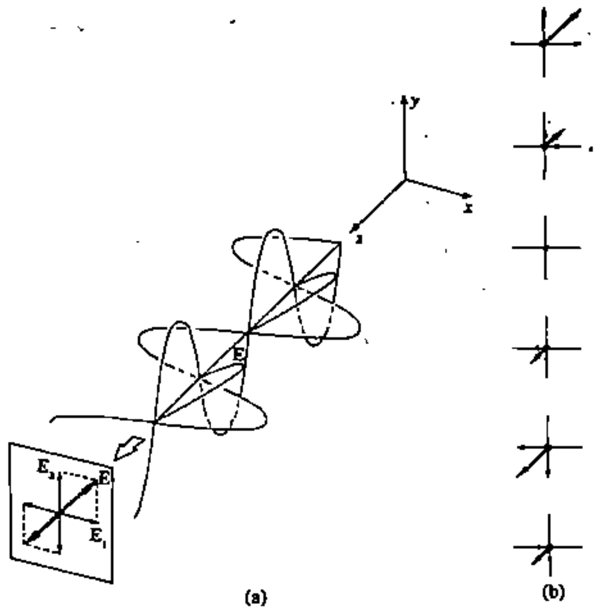
$$\theta = 45^\circ$$

Let us first take the simplest case where $\phi$ is zero or an integral multiple of $\pm\ 2\pi$. That is, when in - phase waves are superposed, Eq. (4.3) takes the form

$$\mathbf{E}(z,t) = (\hat{\mathbf{e}}_x E_{01} + \hat{\mathbf{e}}_y E_{02}) \cos(kz - \omega t) \tag{4.4}$$

The amplitude $\sqrt{E_{01}^2 + E_{02}^2}$ and the electric field oscillations in the reference frame make an angle $\theta = \tan^{-1}(E_{02}/E_{01})$ with the $x$-axis.

For the special case of in-phase waves of equal amplitude $(E_{01} = E_{02} = E_0)$, the resultant wave has amplitude equal to $\sqrt{2}\, E_0$ and the associated electric vector is oriented at 45° with the $x$- axis. So we may conclude that **when two in-phase linearly polarised light waves are superposed, the resultant wave has fixed**



(a) (b)

Fig. 4.5: Schematic representation of a plane polarised light wave

**orientation as well as amplitude.** That is , it too is linearly polarised, as depicted in Fig. 4.5 (a). In the plane of observation, you will see a single resultant **E** oscillating cosinusoidally in time along an inclined line (Fig. 4.5 (b) ). The **E** - field progresses through one complete cycle as the wave advances along the $z$- axis through one wavelength.

If we reverse this process, we can say that **any linearly polarised light can be visualised as a combination of two linearly polarised lights with planes of polarisation parallel to $x = o$ and $y = o$ planes.** (This is similar to resolving a vector in a plane along two mutually perpendicular directions.) In the subsequent sections, you will use this result frequently.

If the phase difference between two plane polarised light waves is an odd integral multiple of ± π, the resultant wave will again be linearly polarised:

$$\mathbf{E}(z,t) = (\hat{\mathbf{e}}_x E_{01} - \hat{\mathbf{e}}_y E_{02})\cos(kz - \omega t) \qquad (4.5)$$

What is the orientation of the resultant electric vector in the reference plane? To know the answer of this question, work-out the following SAQ.

*Spend 5 min*

SAQ 1

Depict the orientation of electric vector defined by Eq. (4.5) in the reference (observation) plane.

### 4.3.2 Circvlar Polarisation

$\cos(-\theta) = \cos\theta$

$\cos\left(\frac{\pi}{2} - \theta\right) = \sin\theta$

We now investigate the nature of the resultant wave arising due to superposition of two plane polarised waves whose amplitudes are equal ($E_{01} = E_{02} = E_0$) but phases differ by π /2, i.e. their relative phase difference $\phi = 2n - \frac{\pi}{2}$, $n = 0, \pm 1, \pm 2, \ldots$ For $n = 0$, we can rewrite Eqs. (4.1) and (4.2) as

$$\mathbf{E}_1(z,t) = \hat{\mathbf{e}}_x E_0 \cos(kz - \omega t) \qquad (4.6a)$$

$$\mathbf{E}_2(z,t) = \hat{\mathbf{e}}_y E_0 \sin(kz - \omega t) \qquad (4.6b)$$

The resultant wave is given by

$$\mathbf{E}(z,t) = E_0[\hat{\mathbf{e}}_x \cos(kz - \omega t) + \hat{\mathbf{e}}_y \sin(kz - \omega t)] \qquad (4.7)$$

You may note that the scalar amplitude of **E** is constant (= $E_0$ ) but its orientation varies with time. To determine the trajectory along which the tip of **E** moves, we can readily combine Eqs.(4.6a) and (4.6b) to yield

$$\left(\frac{E_1}{E_0}\right)^2 + \left(\frac{E_2}{E_0}\right)^2 = 1 \qquad (4.8)$$

which is the equation of a circle. That is, the orientation of resultant electric vector changes continuously and its tip moves along a circle as the wave propagates (time passes). This means that **E** is not restricted to a single plane. The question now arises: What is the direction of rotation? Obviously there are two possibilities: Clockwise and counterclockwise. To know which of these is relevant here, you should tabulate **E** at different space points at a given time, $t = 0$ say:

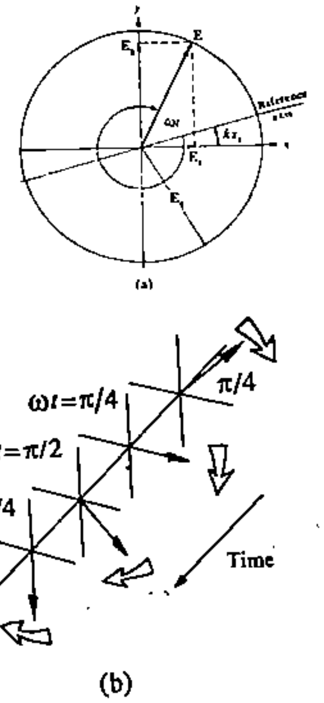
| Location in space | $z = 0$ | $z = \frac{\lambda}{8}$ | $z = \frac{\lambda}{4}$ | $z = \frac{3\lambda}{8}$ | $z = \frac{\lambda}{2}$ | $z = \frac{5\lambda}{8}$ | $z = \frac{3\lambda}{4}$ | $z = \frac{7\lambda}{8}$ | $z = \lambda$ |
|---|---|---|---|---|---|---|---|---|---|
| Electric field | $\hat{\mathbf{e}}_x E_0$ | $\frac{\hat{\mathbf{e}}_x + \hat{\mathbf{e}}_y}{\sqrt{2}} E_0$ | $\hat{\mathbf{e}}_y E_0$ | $\frac{\hat{\mathbf{e}}_y - \hat{\mathbf{e}}_x}{\sqrt{2}} E_0$ | $-\hat{\mathbf{e}}_x E_0$ | $-\frac{\hat{\mathbf{e}}_x + \hat{\mathbf{e}}_y}{\sqrt{2}} E_0$ | $-\hat{\mathbf{e}}_y E_0$ | $\frac{\hat{\mathbf{e}}_x - \hat{\mathbf{e}}_y}{2} E_0$ | $\hat{\mathbf{e}}_x E_0$ |

These are depicted in Fig. 4.6. If you position yourself in the reference plane and observe the evolution of **E** from $z = \lambda$ to $z = 0$ (backward towards source), you will find that the tip of **E** rotates clockwise. Such a light wave is said to be **right circular wave.** The electric field makes one complete rotation as the wave advances through one wavelength.

In case the phase difference $\phi = 2n + \frac{\pi}{2}$ with $n = 0, \pm 1, \pm 2, \ldots$, Eq. (4.7) is modified to

$$\mathbf{E}(z, t) = E_0 \left[ \hat{\mathbf{e}}_x \cos(kz - \omega t) - \hat{\mathbf{e}}_y \sin(kz - \omega t) \right] \qquad (4.9)$$

Alterntively, we may fix an arbitrary point $z = z_0$ and observe evolution of $E$ as time passes. The figure below depicts what is happening at some arbitrary point $z_0$ on the axis.



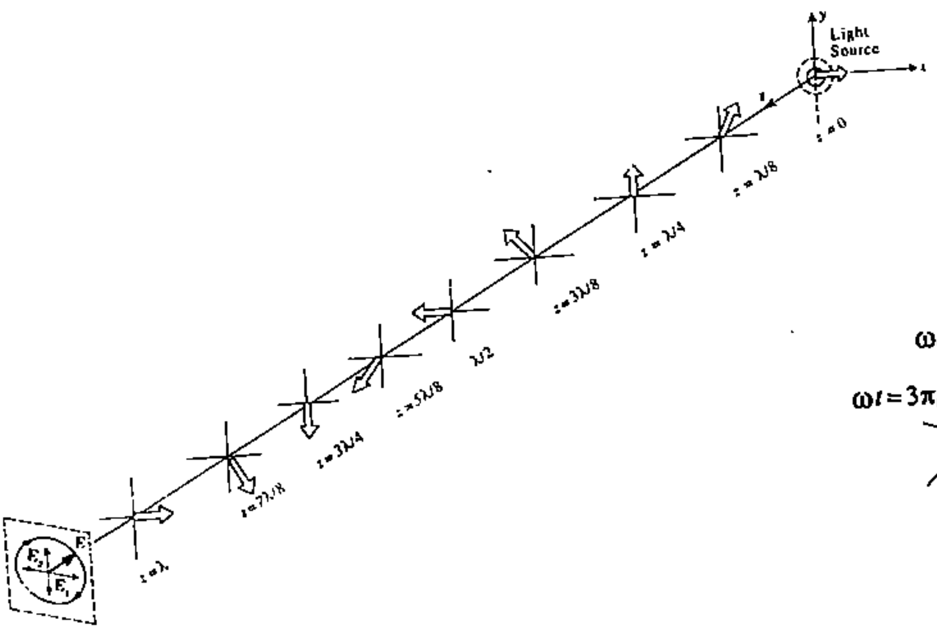


**Fig. 4.6: Rotation of the electic vector in a right-cirular wave. For consistency, we have used a right handed system.**

It shows that the E-vector rotates counter-clockwise in the reference frame. (Before proceeding further you should convince yourself by tabulating the values of **E** at $t = 0$ for different space point.) such a wave is referred to as **left-circular wave.**

Can you now guess as to what will happen if two oppositely polarised circular waves of equal amplitude are superposed? Mathematically, you should add Eqs. (4.8) and (4.9). Then you will find that

$$\mathbf{E} = 2E_0 \hat{\mathbf{e}}_x \cos(kz - \omega t) \qquad (4.10)$$

This equation is similar to Eq.(4.1) which represents a linearly polarised light wave. Thus, we may conclude that **superposition of two oppositely polarised circular waves (of same amplitude) results in a linearly or plane polarised light wave.**

### 4.3.3 Elliptical Polarisation

Let us now consider the most general case where two orthogonal linearly polarised light waves of unequal amplitudes and having an arbitrary phase difference $\phi$ are superposed. Physically we expect that beside its rotation, even the magnitude of resultant electric field vector will change. This means that the tip of **E** should trace out an ellipse in the reference plane as the wave propagates. To analyse this mathematically, we write the scalar part of Eq.(4.2) in expanded form:

$$\frac{E_2}{E_{02}} = \cos(kz - \omega t)\cos\phi - \sin(kz - \omega t)\sin\phi$$

On combining it with Eq. (4.1) we find that

$$\frac{E_2}{E_{02}} = \frac{E_1}{E_{01}}\cos\phi - \sin(kz - \omega t)\sin\phi$$

or

$$\frac{E_2}{E_{02}} - \frac{E_1}{E_{01}}\cos\phi = -\sin(kz - \omega t)\sin\phi \qquad (4.11)$$

It follows from Eq.(4.1) that

$$\sin(kz - \omega t) = \left[1 - (E_1/E_{01})^2\right]^{\frac{1}{2}}$$

so that Eq. (4.11) can be rewritten as

$$\frac{E_2}{E_{02}} - \frac{E_1}{E_{01}}\cos\phi = -\left[1 - (E_1/E_{01})^2\right]^{\frac{1}{2}}\sin\phi$$

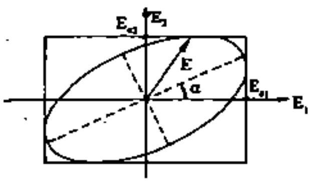


Fig 4.7: Schematics of elliptically polarised light

On squaring both sides and re-arranging terms, we have

$$\left(\frac{E_2}{E_{02}}\right)^2 + \left(\frac{E_1}{E_{01}}\right)^2 - 2\left(\frac{E_2}{E_{02}}\right)\left(\frac{E_1}{E_{01}}\right)\cos\phi = \sin^2\phi \qquad (4.12)$$

Do you recognise this equation? It defines an ellipse whose principal axis is inclined with the $(E_1, E_2)$ coordinate system (Fig. 4.7). The angle of inclination, say $\alpha$, is given by

$$\tan 2\alpha = \frac{2E_{01}E_{02}\cos\phi}{E_{01}^2 - E_{02}^2} \qquad (4.13)$$

For $\alpha = 0$ or equivalently $\phi = \pm\pi/2, \pm 3\pi/2, \ldots$, Eq. (4.12) reduces to

$$\left(\frac{E_1}{E_{01}}\right)^2 + \left(\frac{E_2}{E_{02}}\right)^2 = 1 \qquad (4.14)$$

which defines an ellipse whose principal axes are aligned with the coordinate axes. We would now like you to solve an SAQ.

*Spend 5 min*

SAQ 2

Starting from Eq. (4.12) show that linear and circular polarisation states are special cases of elliptical polarisation.

Now that you understand what polarised light is, the next logical step is to know techniques used to get polarised light. You will learn some of these now.

## 4.4 PRINCIPLES OF PRODUCING LINEARLY POLARISED LIGHT

The most impor.ant optical device in any polarised light producing arrangement is a polariser. It changes (input) natural light to some form of polarised light (output). Polarisers are available in several configurations. An ideal polariser is one which reduces the intensity of an incident unpolarised light beam by exactly 50 percent. When unpolarised light is incident on an ideal polariser, the outgoing light is in a definite polarisation state (P-state) with an orientation parallel to the transmission axis of the polariser. That is, the polariser somehow discards all except one particular polarisation state. How do we determine whether or not a device is a linear polariser? The law which provides us necessary tool is Malus' law. Let us learn about it now.

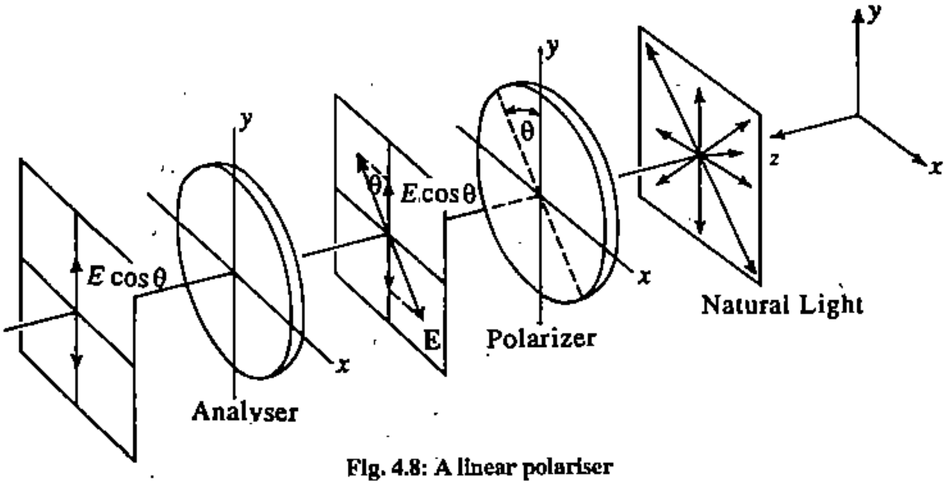


Fig. 4.8: A linear polariser

### 4.4.1 Ideal Polariser: Malus' Law

Refer to Fig. 4.8. Unpolarised light is incident on an ideal polariser, whose transmission axis makes an angle θ with $y$-axis. For this arrangement, only a P-state parallel to the transmission axis of the polariser will be transmitted. This light is incident on an identical ideal polariser, called analyser, whose transmission axis is vertical. Suppose that there is no absorption of light. Then, if **E** is the electric field transmitted by the polariser, only its component $E\cos\theta$ parallel to the transmission axis of the analyser would pass through. The intensity of the polarised light reaching the detector is given by

$$I(\theta) = I(0)\cos^2\theta \qquad (4.15)$$

where θ is the angle between the transmission axes of the polariser and the analyser. The maximum intensity I(0) occurs when the transmission axis of the polariser and the analyser are parallel.

Eq. (4.15) constitutes what is known as **Malus' law**. To use it to check whether an optical device is an ideal linear polariser or not, you may like to solve an SAQ.

**SAQ 3.**

*Spend 5 min*

Unpolarised light falls on two polarising sheets placed one over another. What must be the angle between their transmission axes if the intensity of light transmitted finally is one-third the intensity of the incident light? Assume that each polarising sheet acts as an ideal polariser.

So far we have confined to a linear ideal polariser. Polarisers are available in several configurations. (We can have circular or elliptical polarisers as well.) They are based on one of the following physical mechanisms: reflection, birefringence or double refraction, scattering and dichroism or selective absorption. You will now learn about some of these in detail.

### 4.4.2 Polarisation by Reflection: Brewster's Law

This effect was studied by Malus. One evening he was examining a calcite crystal while standing at the window of his house. The image of the Sun was reflected towards him from the windows of Luxembourg Palace. When he looked at the image through the calcite crystel, he was amused at disappearance of one of the double images as he rotated the crystal.

Reflection of light from a dielectric like plastic or glass is one of the most common methods of obtaining polarised light. You may have noticed the glare across a window pane or the sheen on the surface of a billiard ball or book jacket. It is due to reflection at the surface and the light is partially polarised. To understand its theoretical basis we will consider laboratory situations.

Suppose that an unpolarised light wave is incident on an interface between two different media at an angle $\theta_i$ as shown in Fig. 4.9.

The reflection coefficients when the electric vector of the incident wave is perpendicular to the plane of incidence or when it lies in the plane of incidence are given by Fresnel's equations (Eqs. (2.21a) and (2.21c)):

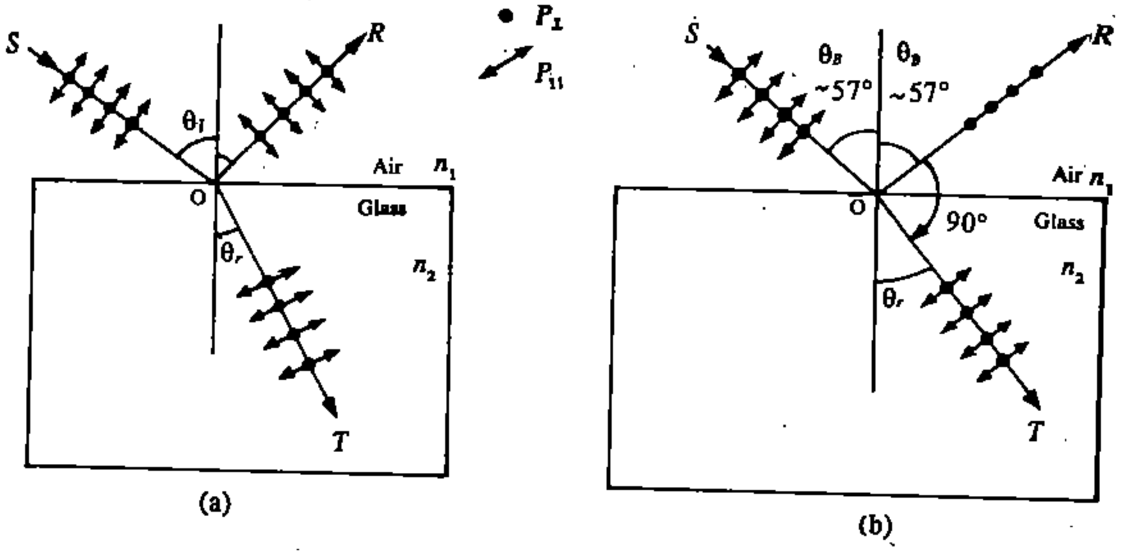


**Fig. 4.9: (a) Polarisation by reflection: the unpolarised light beam has been represented as ←•→ which indicate two electric field vibrations. '•' indicates electric field vibration perpendicular to the page ($P_\perp$) and '↔' indicates electric field vibration in the plane of the paper ( $P_\parallel$ ). (b) At Brewsfer's angle, the reflected light is plane polarised.**

$$R_\parallel = \frac{\tan^2(\theta_i - \theta_r)}{\tan^2(\theta_i + \theta_r)} \qquad (14.16a)$$

and

$$R_\perp = \frac{\sin^2(\theta_i - \theta_r)}{\sin^2(\theta_i + \theta_r)} \qquad (14.16b)$$

where $\theta_r$ is the angle of refraction. These equations show that whereas $R_\perp$ can never be zero, $R_\parallel$ will become zero when $\theta_i + \theta_r = \frac{\pi}{2}$. (The case $\theta_i = \theta_r$ is trivial as it implies continuity of optically identical media.) That is, there will be no reflected light beam with **E** parallel to the plane of incidence. The angle of incidence for which light is completely transmitted is called **Brewster's angle.** Let us denote it by $\theta_B$ . A plot of $R_\perp$ and $R_\parallel$ versus $\theta_B$ is shown in Fig. 4.10 for the particular case of air-glass interface.

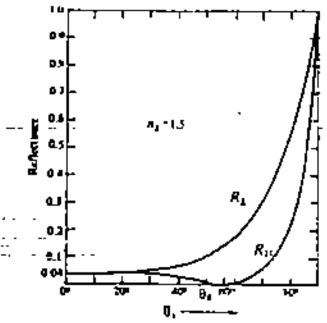


**Fig. 4.10: Variation of reflectance with angle of Incidence**

We can represent an incoming unpolarised light as made up oi two orthogonal, equal amplitude P-states with electric field vector parallel and perpendicular to the plane of incidence. Therefore, when the unpolarised wave is incident on an interface and the angle of incidence is equal to Brewster's angle, the reflected wave will be linearly polarised with E normal to the incidece plane: This provides us with one of the most convenient methods for production of polarised light. To elaborate, we recall from Snell's law that

$$n_i \sin \theta_B = n_2 \sin \theta_r$$

where $n_1$ and $n_2$ are the refractive indices of the media at whose interface light undergoes reflection. Since $\theta_r = \frac{\pi}{2} - \theta_B$, it readily follows that

$$n_1 \sin \theta_B = n_2 \cos \theta_B$$

or

$$\tan \theta_B = \frac{n_2}{n_1} \qquad (14.17)$$

That is, the tangent of Brewster angle is equal to the ratio of the refractive indices of the media at whose interface incident light is reflected. When the incident beam is in air ($n_1 = 1$) and the transmitting medium is glass ($n_2 = 1.5$), the Brewster angle is nearly 56°. Similarly, $\theta_B$ for air-water interface, like surface of a pond or a lake is 53°. This means that when the sun is 37° above the horizontal, the light reflected by a calm pond or lake should be completely linearly polarised.

We, however, encounter some problems in utilizing this phenomenon to construct an effective polariser on account of two reasons:

(i) The reflected beam , although completely polarised, is weak.

(ii) The transmitted beam, although strong, is only partially polarised.

These shortcomings are overcome using a **pile of plate** polarisers. You can fabricate such a device with glass plates for the visible, silver chloride plates for the infrared, and quartz for the ultraviolet region. It is an easy matter to construct a crude arrangement of this sort with a dozen or so microscope slides (Fig. 4.11).The beautiful colours that appear when the slides are in contact is due to interference, which you will study in the next block.

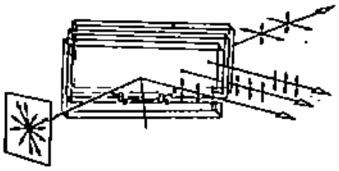


**Fig. 4.11: Polarisation of light by a pile of plates**

You may now like to solve an SAQ.

**SAQ 4**

*Spend 5 min*

A plate of flint glass is immersed in water. Calculate the Brewster angles for internal as well as external reflection at an interface.

Having studied as to how reflection of light can be used to produce polarised light, you may be tempted to know whether or not the phenomenon of refraction can also be used for the same? Refraction of light in isotropic crystals like NaCl or non-crystalline substances like glass, water or air does not lead to polarisation of light. However, refraction in crystalline substances like calcite or cellophane is optically anisotropic becouse it leads to what is known as double refraction or birefringence. This is because anisotropic crystals display two distinct principal indices of refraction, which correspond to the E-oscillations parallel and perpendicular to the optic axis. Let us now learn how birefringence can be used to produce polarised light.

### 4.4.3 Polarisation by Double Refraction

Mark a **black dot** on a piece of paper and observe it through a glass plate. You will see only one dot. Now use a calcite crystal. You will be surprised at the remarkable observation: instead of one, **two grey dots** appear, as shown in Fig. 4.12. Further, rotation of the crystal will cause one of the dots to remain stationary while the other appears to move in a circle about it. Similarly, if you place a calcite crystal on your

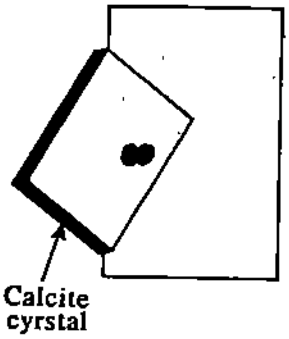


Fig. 4.12: Double refraction of a light beam by calcite crystal.

In some of the text books, you may find that ordinary and extraordinary rays are being denoted by bold letters O and E. We have used small letters (o- and e-) to avoid confusion with the notation for the electric field.

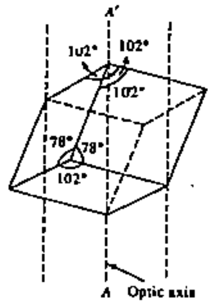


Fig. 4.13: A Calcite crystal. The line $AA'$ shows the direction of the optic axis. For the calcite crystal, the direction of the optic axis is determined by joining the two blunt corners of the crystal.

book, you will see two images of each letter . It is because the calcite crystal splits the incident light beam into two beams. This phenomenon of splitting of a light beam into two is known as **double refraction or birefringence.** Materials exhibiting this property are said to be birefringent. We bring you the excitement of Bartholinus, who discovered birefringentce, in his words:

> Greatly prized by all men is the diamond, and many are the joys which similar treasures bring, such as precious stones and pearls ... but he, who, on the other hand, prefers the knowledge of unusual phenomena to these delights, he will, I hope, have no less joy in a new sort of body, namely, a transparent crystal, recently brought to us from Iceland, which perhaps is one of the greatest wonders that nature has produced. As my investigation of this crystal proceeded there showed itself a wonderful and extraordinary phenomenon: objects which are looked at through the crystal do not show, as in the case of other transparent bodies, a single refracted image, but they appear double.

Before we discuss polarisation of light by double refraction in detail, you should familiarise yourself with some of the concepts related to this phenomenon. The two refracted beams into which incident light splits have different angles of refraction. The distinguishing feature of these two refracted light beams is that one of these obeys the Snell's law. It is called the **ordinary ray (o-ray)** in accordance with the nomenclature given by Bartholinus. The other beam does not obey Snell's law abd is called the **extraordinary ray (e-ray).** That is, a birefringent crystal displays two distinct indices of refranction. Another important concept is that of optic axis, which signifies some special direction in a birefringent crystal along which two refractive indices are equal (i. e. both o-and e-rays traval in the same direction with the same velocity). When unpolarised light is incident perpendicular to these special directions, both the o-and th e-rays travel in the same direction with different velocities. You may now like to know: Does optic axis refer to any particular line through the crystal? The answer to this question is: It refers to a direction. This means that for any given point in the crystal, an optic axis may be drawn which will be parallel to that for any other point. For example, $A\,A'$ and broken lines parallel to $A\,A'$ show the optic axis for a calcite crystal as shown in Fig. 4.13.

Birefringent crystals which posses only one optic axis are called **uniaxial crystals.** Similarly, crystals having two optic axes are called **biaxial crystals.** Calcite, quartz and ice are examples of uniaxial crystals and mica is a biaxial crystal. Most of the polarisation devices are made of uniaxial crystals. Further, the uniaxial crystal for which the refractive index o-ray ( $n_o$ ) is more than the refractive index for the e-ray ( $n_e$ ) is called **nigative uniaxial crystal.** On the other hand, if $n_e > n_o$, we have a **positive uniaxial crystal.** Values of $n_o$ and $n_e$ for some of the birefringent crystals are given in Table 4.1. The difference $\Delta\, n = n_e - n_o$ is a measure of birefringence.

**4.1: Refractive indices of some uniaxial birefringent crystals**

**for light of wavelength 5893 Å**

| Crystal | $n_o$ | $n_e$ |
|---|---|---|
| Tourmaline | 1.669 | 1.638 |
| Calcite | 1.6584 | 1.4864 |
| Quartz | 1.5443 | 1.5534 |
| Sodium Nitrate | 1.5854 | 1.3368 |
| Ice | 1.309 | 1.313 |

Let us now enquire how unpolarised light incident on uniaxial crystal gets polarised? We know that when unpolarised light beam enters a calcite crystal, it splits into the o-and the e-rays. The electric field vector of e-ray vibrates in the plane containing the optic axis and the electric field vector of o - ray vibrates perpendicular to it, as shown in Fig. 4.14. We may, therefore, concude that due to

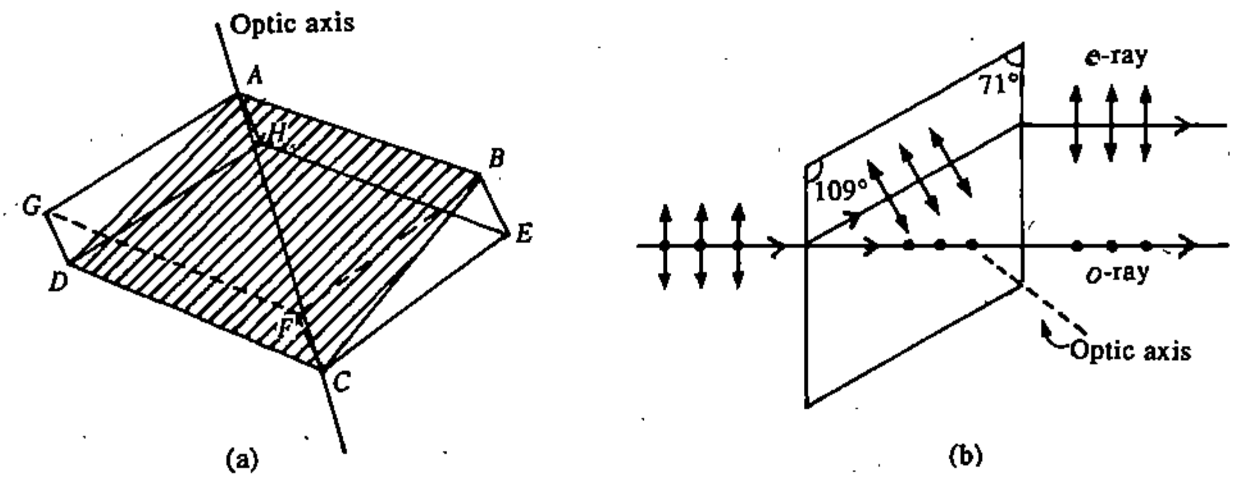


**Fig. 4.14: (a) *ABCD* is one of the principal sections of the calcite crystal; it contains the optic axis and is normal to the cleavage faces *BECF* and *AHDG*. (b) Unpolarised light beam passing through a prinicipal section of the calcite crystal.**

double refraction, the unpolarised light beam splits into two components which are **plane polarised.**

Huygens explained many aspects of double refraction in calcite on the basis of wave theory. Since the o-ray obeys Snell's law, it propagates with uniform velocity in all directions in the crystal. As a result, the wave surfaces are spherical. However, the e-ray propagates with different velocities in different directions in the crystal and hence the resulting wave surface is an ellipsoid of revolution, i. e. a spheroid. Further, to reconcile with the fact that both the o-and e-rays travel with the same velocity along the optic axis, both the wave surfaces ware assumed to touch each other at the two extremities of the optic axis. These features are depicted in Fig. 4.15. You may now like to know the nature of wave surfaces for o-and e-waves in positive uniaxial crystals. This is subject matter of TQ 1.

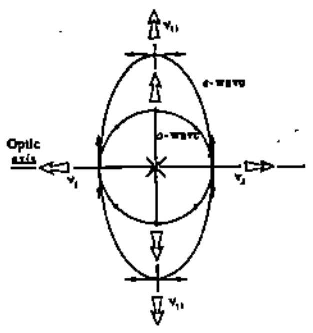


**Fig. 4.15: o-and e-wave surfaces in negative uniaxial crystal (calcite).**

From the above discussion it follows that in double refraction, an unpolarised light wave splits into o-and e-components with their E-vibrations perpendicular to each other. By selective absorption of one of the P-states, we can produce linearly polarised light. This is readily done by a device, called Nicol prism, by removing the o-ray through total internal reflection. It was designed by William Nicol in 1828. You will learn about it now.

### Nicol prism

Nicol prism is made from a naturally occurring crystal of calcite. The length of the crystal is three times its width and the smaller faces *PQ* and *RS* and ground from 71° to a more acute angle of 68° (Fig. 4.16). The crystal is then cut along *PS* by a plane passing through *P* and *S* and perpendicular to the principal section *PQSR*. The cut surfaces are polished to optical flatness and then cemented together with a layer of (nonrefringent material) Canada balsam.

Can you guess why Canada balsam is used as cementing material? Well, for sodium light, refractive index of Canada balsam is 1.552, which is midway between refractive indices for o-ray ($n_o = 1.658$) and the e-ray ($n_e = 1.486$) in calcite. Thus, it is an optically rarer medium with respect to ordinary ray and denser for extraordinary ray. The critical angle for total internal reflection of o-ray is $\sin^{-1}\dfrac{1.552}{1.658} = 69°$.

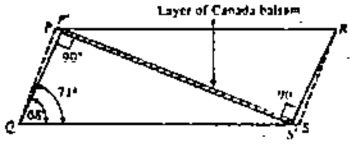


**Fig. 4.16: Nicol Prism**

So, when incident unpolarised light splits into two rays inside the crystal, the o-ray gets totally reflected at the canada balsam surface when it is incident on it at an angle of 69°. (It is for this reason that the end faces of the crystal are ground so as to make the angles 68° from 71°.) The emergent light will, therefore, be made up only of plane polarised e-component.

Some of the limitations of Nicol prism as polariser are:

1. It can be used for polarisation of visible light only.
2. e-ray also can get totally reflected by the Canada balsam surface if it is travelling along the optic axis. Why? It is so because in this situation the refractive index for e-ray will be same as for o-ray (i. e. greater than the refractive index for Canada balsam).

With time, a number of modifications have been incorporated in the basic design of the Nicol prism to overcome some of these limitations. However. we will not go into these details.

So far you have studied about production of linearly polarised light by reflection and double refraction. Other methods employed to produce linearly polarised light are selective absorption (or dichroism) and scattering. We will here discuss only dichroism and that too in brief.

### 4.4.4. Selective Absorption: Dichroism

As you know, unpolarised light wave can be regarded as made up of two orthogonal, linearly polarised waves. Many naturally occuring and man made materials have the property of selective absorption of one of these; the other passes through without much attenuation. This property is known as **dichroism**. Materials exhibiting this property are said to be **dichroic materials.** The net result of passing an unpolarised light through dichroic material is the production of linearly polarised light beam. A particularly simple dichroic device is the so-called Wire–Grid polariser. You will learn about it now.

### The Wire–Grid Polariser

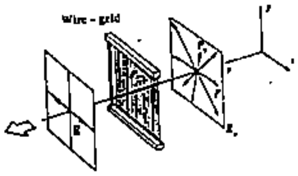


Fig. 4.17: The Wire-grid Polariser

The wire–grid polariser constists of a grid of parallel conducting wires, as shown in Fig. 4.17. Suppose that unpolarised light is incident on the grid from the right. It can be thought as made up of two orthogonal P–states: $P_x$ and $P_y$ in the reference plane $R_z$. The $y$–component of the electric field drives the electrons of each wire and generates a current. It produces (Joule) heating of the wire. The net result is that energy is transferred from the field to the wire grid. In addition, electrons accelerating along the $y$–direction radiate in the forward as well as backward directions. The incident wave tends to be cancelled by the wave re–radiated in the forward direction. As a result, transmission of $y$–component of field is almost blocked. However, the $x$–component of field is essentially unaltered as it propagetes through the grid and the light coming out of the wire–grid is linearly polarised. The wire–grid polariser almost completely attenuates the $P_y$ component when the spacing betwen the wires is less than or equal to the wavelength of the incident wave. You must realise that this restriction is rather stringent for the fabrication of a wire–grid polariser for visible light ( $\lambda \sim 5\times 10^{-7}$m ).

An easy way out of this difficulty in the fabrication of the grid polariser is to emplov long chain polymer molecules made up of atoms which provide high electrical conductivity along the length of the chain. These chains of polymer molecules behave similar to the wires in the wire–grid polariser. The alignment of these chains are almost parallel to each other. Because of high electrical conductivity, the electric vector of unpolarised light parallel to the chain gets absorbed. And the P–state perpendicular to these chains passes through. These chemically synthesized polarisers are fabricated in the form of plastic sheets and are known as **polaroids**. Since the spacing between these molecular chains in a polaroid is small compared to the optical wavelength, such polaroids are extremely effective in producing linearly polarised light.

### Dichroic Crystals

Some naturally occurring crystalline materials are inherently dichroic due to anisotropy in their structure. One of the best known dichroic materials is **tourmaline**, a precious stone often used in jewellery. Tourmalines are essentially boron silicates of differing chemical composition.The component of E perpendicular to the principal axis is strongly absorbed by the sample. Thicker the crystal, more complete will be the absorption. A plate cut from a tourmaline crystal parallel to its optic axis acts as a linear polariser. This is illustrated in Fig. 4.18.

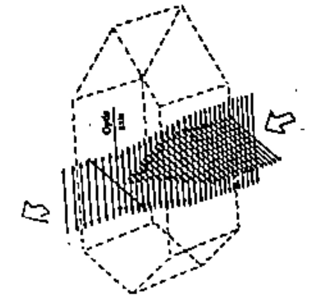


**Fig. 4.18: Tourmaline crystal polariser**

We shall now consider a class of optical elements known as wave plates which serve to change the polarisation of the incident wave. A wave plate introduces a phase lag between the two P–states by a predetermined amount. That is, the relative phase of the two emerging components is different from its initial value. This concept can be used to convert a given polarisation state into any other and in so doing it is possible even to produce circular or elliptic polarisation as well. This is the subject matter of the next section.

## 4.5 WAVE PLATES: CIRCULAR AND ELLIPTIC POLARISERS

Consider a plane wave incident on a calcite crystal. It splits in o-and e- waves. Since calcite is a negative uniaxial crystal, $n_o > n_e$ and $v_{\parallel}$ (velocity of e–wave) > $v_{\perp}$ (velocity of o–wave) implying that the e–ray travels faster than the o–ray. After traversing the calcite crystal of thickness $d$, the path difference between them is given by

In case of positive uniaxial crystals, $n_e > n_o$ and hence the path difference will be $d\,(n_e - n_o)$. In fact the general expressiopn for the path difference is $d\,(\,|n_o - n_e|\,)$

$$\Delta = d\,(n_o - n_e)$$

and the relative phase difference between o– and e–rays is

$$\delta = \frac{2\pi}{\lambda}\Delta = \frac{2\pi}{\lambda}(n_o - n_e)\,d \tag{4.18}$$

though while entering both the components were in phase.

The **state of polarisation** of emerging light depends on δ, apart from the amplitudes of incoming orthogonal field components. Let us now consider some specific cases:

(i) When the phase difference, $\delta = 2m\pi$, where $m$ is an integer, the ralative path difference is $m\lambda$. A device which induces a path difference between the two orthogonal field vibrations in integral multiples of λ is called the full wave plate. It introduces no observable effect on the polarisation of the incident beam. That is, the field vibrations of the emergent light will be identical with the field vibrations of the incident light.

(ii) When $\delta = (2m+1)\pi$, the relative path difference will be $(m + ½)\lambda$. Such crystals are called **half–wave plates.**

(iiii) When $\delta = (2m \mp 1)\frac{\pi}{2}$, the relative path difference will be $\left(m + \frac{1}{2}\right)\frac{\lambda}{2}$. Such a birefringent sheet is called **quarter–wave plate.** When linearly polarised light traverses a quarter–wave plate, the emergent light will, in general, be elliptical and the axes of the ellipse will coincide with the previleged directions of the thin plate. However, half–wave or full–wave plate leave the state of polarisation unchanged.

Thus, we may conclude that the **path difference between the o– and e–waves in a birefringent device depends on its thickness.**

You should now solve the following SAQ.

SAQ 5

*Spend 5 min*

Calculate the thickness of a quarter wave–plate for light of wavelength 5890 Å. The refractive indices for o – and e –rays are 1.55 and 1.50 respectively.

We now summarise what you have learnt in this unit.

## 4.6 SUMMARY

- Visible light can be linearly, circularly or elliptically polarised. All these polarisation states arise on superposition of two linearly (or plane) polarised light waves characterised by different amplitudes and phases.
- The electric field vectors of two linearly polarised light beams propagating along $z$–axis can be represented as

$$\mathbf{E}_1(z, t) = \hat{\mathbf{e}}_x E_{01} \cos(kz - \omega t)$$

$$\mathbf{E}_2(z, t) = \hat{\mathbf{e}}_y E_{02} \cos(kz - \omega t + \phi)$$

where $E_{01}$ and $E_{02}$ are the amplitudes of the two waves and $\phi$ is the phase difference between them. Superposition of these two polarised waves will result in

Linearly polarised light if $\phi = 0$ or an integral multiple of $\pm\ 2\pi$

Circularly polarised light if $\phi = \pi/2$ and $E_{01} = E_{02}$

Elliptically polarised light if $\phi = \pi/2$ and $E_{01} \neq E_{02}$

- According to Malus, when the transmission axes of polariser and the analyser are at an angle $\theta$, the intensity of the polarised light reaching the detector is givenby $I(\theta) = I(0)\cos^2\theta$ where, $I(0)$ is the intensity of the polarised light when $\theta = 0$.
- When natural light strikes an interface at Brewster's angle $\theta_B = \tan^{-1}(n_2/n_1)$, where $n_1$ and $n_2$ are the refractive indices of medium of incidence and transmission, the reflected light is linearly polarised.
- When light falls on a calcite crystal, it splits into two. The phenomenon is known as double refraction or birefringence. These two refracted beams are known as o– and e-rays. Snell's law holds for o–rays (ordinary rays).
- In a birefringent material, the o– and the e–rays travel in the same direction with same velocity along the optic axis. However, in a direction perpendicular to the optic axis, they travels with different velocities. The electric field vibrations for o– and the e–rays are mutually perpendicular.
- The phenomenon of double refraction produces linearly polarised light. Nicol prism works on this principle. In the Nicol prism, the o–ray undergo total internal reffection at the inferface and the transmitted beam consists of only electric field vibrations corresponding to e–ray and hence the transmitted beam is linearly polarised.
- Selective absorption (or dichroism) of the electric field component with particular orientations by material can also be used for producing linearly polarised light. Tourmaline is an example ofdichroic material.
- For a calcite crystal of thickness $d$ the path difference between o– and e– rays is given by $\Delta = d\,|n_o - n_e|$

The Corresponding phase difference

$$\delta = \frac{2\pi}{\lambda}\Delta = \frac{2\pi}{\lambda} d\left(|n_o - n_e|\right)$$

When the phase difference $\delta = 2m\pi$ where $m$ is an integer,the relative path difference between the o– and e–rays will be $m\lambda$. Such crystals are called full–wave plate. When $\delta = (2m + 1)\,\pi$, path difference will be $\lambda/2$ and such crystals act as full–wave plate. And when $\delta = (2m + 1)\,\frac{\lambda}{2}$, path difference will be $\lambda/4$ (for $m = 0$) and such crystals are called quarter–wave plate.

## 4.7 TERMINAL QUESTIONS

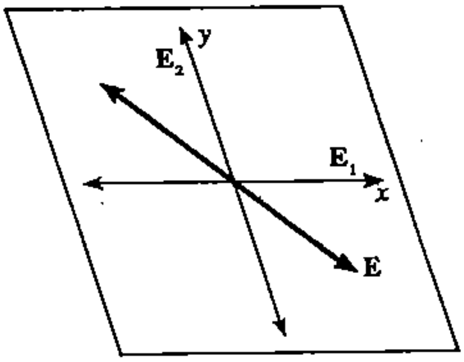
1. In sub–section 4.4.3, you studied about propagation of o– and e– waves in a negative uniaxial crystal (calcite). Draw a diagram and describe the propagation of o– and e–waves in a positive uniaxial crystal (quartz) for normal incidence.

2. For a certain crystal, $n_o = 1.5442$ and $n_e = 1.5533$ for light of wavelength $6 \times 10^{-7}$ m. Calculate the least thickness of a quarter–wave plate made from the crystal for use with light of this wavelength.

## 4.8 SOLUTIONS AND ANSWERS

### SAQs

1. The plane of vibration of the electric vector defined by Eq. (4.5) is rotated with respect to that shown in the Fig. 4.5. This is signified by the negative sign before $\hat{\mathbf{e}}_y$ in the parentheses and is depicted below



2. We know from Eq. (4.12) that

$$\left(\frac{E_2}{E_{02}}\right)^2 + \left(\frac{E_1}{E_{01}}\right)^2 - 2\left(\frac{E_2}{E_{02}}\right)\left(\frac{E_1}{E_{01}}\right)\cos\phi = \sin^2\phi \qquad \text{(i)}$$

If we choose $\phi = \pi$ in (i), we get

$$\left(\frac{E_2}{E_{02}}\right)^2 + \left(\frac{E_1}{E_{01}}\right)^2 - 2\left(\frac{E_2}{E_{02}}\right)\left(\frac{E_1}{E_{01}}\right) = 0$$

which can be written in a compact form:

$$\left(\frac{E_2}{E_{02}} - \frac{E_1}{E_{01}}\right)^2 = 0$$

or

$$E_2 = \frac{E_{02}}{E_{01}} E_1 \qquad \text{(ii)}$$

This defines a straight line ($y = mx$) with slope $E_{02}/E_{01}$. In other words, elliptically polarised light reduces to linearly polarised light for $\phi = n\pi\,(n = 0, \pm 1, 2, \dots)$.

When $\phi = \pi/2$ and $E_{01} = E_{02} = E_0$, Eq. (4.12) reduces to

$$\left(\frac{E_2}{E_0}\right)^2 + \left(\frac{E_1}{E_0}\right)^2 = 1$$

which defines a circle ($x^2 + y^2 = a^2$) of radius $E_0$

3. Since both polarising sheets are ideal, thė intensityof the incident unpolarised beam, $I$, will reduce to half after passing throught one of them as shown in the Fig.4.19. After passing through the second polarising sheet, we are told that the intensity reduces to one third of original value.

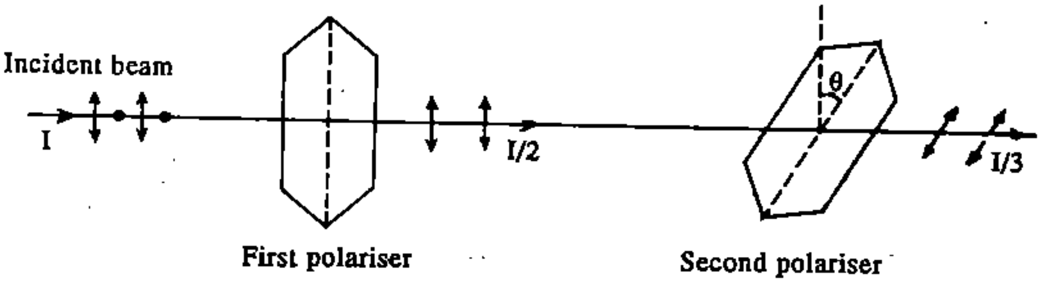


Fig. 4.19: Unpolarised light beam of intensity $I$ passing through two polarisers

From Malus' Law we know that

$$I(\theta) = I(0)\cos^2\theta$$

Here $I(\theta) = I/3$ and $I(0) = 1/2$. Therefore

$$\cos^2\theta = (2/3) = 0.666$$

or

$$\theta = \cos^{-1}(.666)^{1/2}$$

$$= 35.3^\circ$$

That is, the angle between the transmissio axes of two polarisers is about 35°

4. For external reflection

$$\tan i_B = \frac{n_2}{n_1} = \frac{1..67}{1.33}$$

$$\Rightarrow \qquad i_B = \tan^{-1}\left(\frac{1.67}{1.33}\right)$$

$$\text{or} \qquad i_B = 51.47^\circ$$

Fro internal reflection

$$\tan i_B = \frac{n_1}{n_2} = \frac{1.33}{1..67}$$

$$i_B = 38.53^\circ$$

5. The path difference produced between the o– and e– rays of birefringent crystal of thiickness $d$ is

$$\Delta = d\,(\,|\,n_0 - n_e\,|\,)$$

And corresponding relaltive phase difference is given by

$$\delta = \frac{2\pi}{\lambda}\Delta$$

$$= \frac{2\pi}{\lambda} d\,(\,|\,n_o - n_e\,|\,)$$

The phase difference produced bya quarter–wave plate

$$\delta = \pi/2$$

On comparing the above expressions for the phase difference, we have

$$d = \frac{\lambda}{4}(\,n_o - n_e\,)$$

$$= \frac{5890\,\mathring{A}}{4}(\,1.55 - 1.50\,)$$

$$= 73.63\,\mathring{A}$$

$$= 74\,\mathring{A}$$

## TQs

1. In case of negative uniaxial crystal (calcite), e–ray travels faster than the o–ray and hence $n_o > n_e$. Therefore, when a light beam is incident normally upon a calcite crystal, whose optic axis is parallel to the refracting surface and lies in the plane of incidence, o-wave has a spherical wavefront and the e–wave has an spheroidal wavefront.

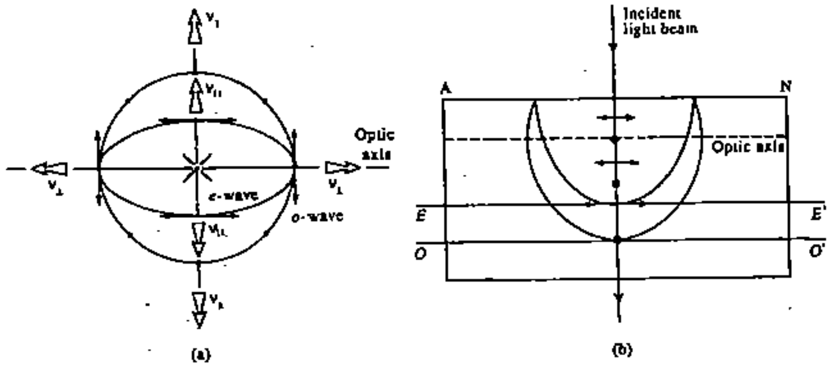


**Fig. 4.20: (a) o-and e-wave surfaces in a positive uniaxial crystal (quartz): (b) Propagation of o-and e-waves in quartz.**

In case of positive uniaxial crystals like quartz, the e–ray travels slower than the o–ray. Therefore, the spherical wavefront corresponding to o–ray will be outside the spheroidal wavefront corresponding to e–ray (Fig. 4.20a). Since the optical properties of birefringent crystal are symmetrical with respect to its optic axis, the axis of revolution of the sphroid must coincide with the optic axis of the crystal. When a light beam falls on a positive uniaxial crystal, with its optic axis in the plane of incidence and parallel to the refracting surface, the wavefront for o– and e–waves is shown in Fig. 4.20b.

In the above mentioned case, $E E'$ and $O O'$ are the refracted wave–fronts for e– and o–rays respectively at the same instant of time. They are parallel to each other and travel in the same direction which is perpendicular to the refracting surfacee *AN*.. These two wavefronts, however, will travel with different velocities. As a result, a path difference will be introduced between the o– and the e–ray on emergence, but there is no separation between the two beams. In principle, we can constructs quarter–wave plate, half–wave plate etc. using positive uniaxial crystal as well.

2. In the birefringent crystal of thickness *d*, the path difference between the o– and e–rays is $d\,|\,n_e - n_o\,|$.

In this problem, $n_e > n_o$, so that we can write $\Delta = d\,(n_e - n_o)$ and the corresponding relative phase difference

$$\delta = \frac{2\pi}{\lambda} d\,(n_e - n_o)$$

For constructing a quarter wave–plate, the path difference should be $\lambda/4$, which corresponds to phase difference of $\pi/2$. Thus, from above equation, we must have, for a quarter wave plate

$$\frac{2\pi}{\lambda} d\,(n_e - n_o) = \pi/2$$

or

$$d = \frac{\lambda}{4\,(n_e - n_o)}$$

We have

$$n_o = 1.5442, n_e = 1.5533, \text{and } \lambda = 6 \times 10^{-7}\,\text{m}.$$

Hence,

$$d = \frac{6 \times 10^{-7}\,\text{m}}{4 \times (1.5533 - 1.5442)}$$

$$= \frac{6 \times 10^{-1}\,\text{m}}{0.0367}$$

$$= 1.65 \times 10^{-5}\,\text{m}.$$

That is, the quarter–wave plate should be $1.65 \times 10^{-5}$ m thick.

# NOTES

# NOTES

Uttar Pradesh
Rajarshi Tandon Open University

# UGPHS-07
# Optics

Block
# 2
# INTERFERENCE

# BLOCK 2 INTERFERENCE

In Block 1 you studied the nature of light. There you studied that light is a wave motion. A very important characteristic of wave motion is the phenomenon of interference.

The term interference refers to the phenomenon that waves, under certain conditions, intensify or weaken each other. The phenomenon of interference is inseparably tied to that of diffraction. In fact, diffraction is more inclusive; it contains interference and, in a sense, even refraction and reflection. It is only because diffraction is mathematically more complex that we treat interference and diffraction in separate Blocks, and discuss interference first.

The prerequisite of all interference is the superposition of waves. If light from a source is divided by suitable apparatus into two beams and then superposed, the intensity in the region of superposition is found to vary from point to point between maxima, which exceed the sum of the intensities in the beams, and minima, which may be zero. This phenomenon is called interference.

There are two general methods of obtaining beams from a single beam of light, and these provide a basis of classifying the arrangements used to produce interference. In one method, the beam is divided by passage through apertures placed side by side. This method, which is also called division of wave front, is useful only with the sufficiently small sources. Alternatively, the beam is divided at one or more partially reflecting surfaces, at each of which, part of the light is reflected and part transmitted. This method is called division of amplitude. It can be used with extended sources, and so effects may be of greater intensity than with the division of the wavefront. In either case, it is convenient to consider separately the effects which result from the superposition of more than two beams (multiple beam interference).

Unit 5 begins with the study of wave motion. Being familiar to most students from their study of Oscillations and Waves, it will serve primarily as a review. With the help of the principle of superposition, we have explained the phenomenon of interference. In this unit, we discuss in detail the phenomenon of interference produced by the division of the wavefront of light wave. This unit would involve your seven hours of work.

In Unit 6, we will consider the formation of interference pattern by the division of amplitude. Such studies have many practical applications. Finally we briefly mention these applications. The study time required for this unit is about six hours.

Unit 7 is devoted to interferometry. It deals with Michelson interferometer, which is an example of two beam interference and Fabry-Perot interferometer which is an example of multiple beam interference. Finally, an appendix given at the end of the unit provides a brief introduction to complex amplitudes. You might like to read it to enrich your knowledge. However, you will not be examined on it. You should aim to finish this unit in about 5 hours.

# UNIT 5 INTERFERENCE BY DIVISION OF WAVEFRONT

## Structure

## 5.1 INTRODUCTION

Anyone with a pan of water can see how the water surface is disturbed in a variety of characteristic patterns, which is due to interference between water waves. Similarly, interference occurs between sound waves as a result of which two people who hum fairly pure tones, slightly different in frequency, hear beats. But if we shine light from two torches or flashlights at the same place on a screen, there is no evidence of interference. The region of overlap is merely uniformly bright. Does it mean that there is no interference of light waves? The answer is 'No'.

The interference in light is as real an effect as interference in water or sound waves, and there is one example of it familiar to everybody — the bright colours of a thin film of oil spread out on a water surface. There are two reason why the interference of light is observed in some cases and not in others? Firstly, light waves have very short wavelengths — the visible part of the spectrum extends only from 400 mm for violet light to 700 mm for red light. Secondly, every natural source of light emits light waves only as short trains of random pulses, so that any interference that occurs is averaged out during the period of observation by the eye, unless special procedures are used.

Like standing waves and beats, the phenomenon of interference depends on the superposition of two or more individual waves under rather strict conditions that will soon be clarified. When interest lies primarily in the effects of enhancement or diminution of light waves, these effects are usually said to be due to the interference of light. When enhancement (or constructive interference) and diminution (or destructive interference) conditions alternate in a spatial display, the interference is said to produce a pattern of fringes as in the double slit interference pattern. The same condition may lead to enhancement of one colour at the expense of the other colour, producing interference colours as in the case of oil slicks and soap film about which you will study in next unit.

In this unit, we will consider the interference pattern produced by waves originating from two point sources. However, in case of light waves, one cannot observe interference between the waves from two independent sources, although the interference does take place. Thus, one tries to derive the interfering waves from a single wave so that the constant phase difference is maintained between the interfering waves. This can be achieved by two methods. In the first method a beam is allowed to fall on two closely spaced holes, and the two beam emanating from the holes interfere. This method is known as division of wavefront and will be discussed in detail in this unit. In the other method, known as division of amplitude, a beam is divided at two or more reflecting surfaces, and the reflected beams interfere. This will be discussed in the next unit.

As the phenomenon of interference can be successfully explained by treating light as a wave motion, it is necessary to understand the fundamentals of wave motion. Although you have learnt about this in your class XII and also in the **PHE–02** course **"Oscillations and Waves"**, we will begin this unit with study of wave motion which will serve as a recapitulation.

In the next unit we will study how interference takes place by division of amplitude of light wave.

## Objectives

After studying this unit, you should be able to

- use the principle of superposition to interpret constructive and destructive interference,
- distinguish between coherent and incoherent sources of light,
- describe the origins of the interference pattern produced by double slit,
- describe the intensity distribution in interference pattern,
- express the fringe-width in terms of wavelength of light,
- describe various arrangements for producing interference by division of wavefront,
- appreciate the difference between Biprism and Lloyd's mirror fringes.

## 5.2 WAVE MOTION

**Study Comment**

You may find it useful to go through the Unit 6 of PHE–02 course in "Oscillation and Wave".

**Simple Harmonic Motion**

A simple harmonic motion is defined as the motion of a particle which moves back and forth along a straight line such that its acceleration is directly proportional to its displacement from a fixed point in the line, and is always directed towards that point.

The best and elementary way to represent a simple harmonic motion is to consider the motion of a particle along a reference circle (See Fig. 5.1). Suppose a particle $P$ travels in a circular path, counterclockwise, at a uniform angular velocity $\omega$. The point $N$ is the perpendicular projection of $P$ on the diameter $AOA'$ of the circle. When the particle $P$ is at point $B$, the perpendicular projection is at $O$. As the particle $P$ starts from $B$, and moves round the circle, $N$ moves from $O$ to $A$, $A$ to $A'$ and then returns to $O$. This back and forth motion of $N$ is simple harmonic. Let us obtain expressions for displacement, velocity and acceleration and define few terms.

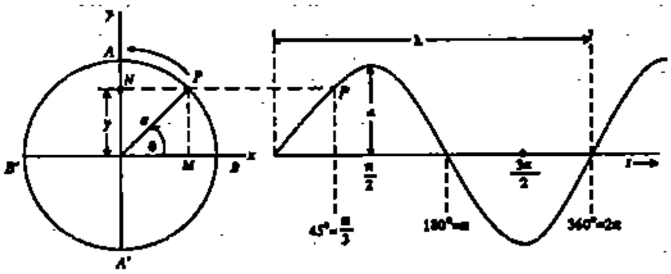


Fig. 5.1. Reference Circle (Left) and Simple Harmonic Motion (Right)

**Displacement**

Suppose the particle $P$ starts from $B$ and traces an angle $\theta$ in time $t$. Then its angular velocity $\omega$ is

$$\omega = \frac{\theta}{t}$$

where the angle $\theta$ is measured in radians. The displacement, $y$, of $N$ from $O$ at time $t$, is thus given by

$$y = ON = OP \sin NPO$$

$$= a \sin \theta \qquad [\because \angle NPO = \angle POB = \theta]$$

But $\quad \omega = \dfrac{\theta}{t}$, so that $\theta = \omega t$

$$\therefore \qquad \boxed{y = a \sin \omega t} \qquad \text{...(5.1)}$$

This is the equation of simple harmonic motion.

SAQ 1

See Fig. 5.1. If you have studied the motion of the point $M$, which is the foot of the perpendicular from the point $P$ on the $x$–axis, then write down the equation of simple harmonic motion.

**Velocity:** The velocity of $N$ is given by

$$\frac{dy}{dt} = a\omega \cos \omega t = \omega \sqrt{a^2 - y^2} \qquad \text{...(5.2)}$$

**Acceleration:** The acceleration of N is

$$\frac{d^2y}{dt^2} = -\,\omega^2 a \sin \omega t = -\,\omega^2 y \qquad \text{...(5.3)}$$

**Periodic Time:** The periodic time, $T$, of $N$ is time taken by $N$ to make one complete vibration. Thus

$$T = \frac{2\pi}{\omega} \qquad \text{...(5.4)}$$

**Amplitude:** Amplitude of vibration is equal to the radius of the reference circle i.e., $a$.

SAQ 2

A particle is executing simple harmonic motion, with a period of 3s and an amplitude of 6 cm. One-half second after the particle has passed through its equilibrium position, what is its (a) displacement, (b) velocity, and (c) acceleration?

**Phase:** The phase of a vibrating particle represents its state as regards

i) the amount of displacement suffered by the particle with respect to its mean position, and

ii) the direction in which the displacement has taken place.

In Fig. 5.1, we had conveniently chosen $t = 0$ as the time when $P$ was on the $x$-axis. The choice of the time $t = 0$ is arbitrary, and we could have chosen time $t = 0$ to be the instant when $P$ was at $P'$ (see Fig.5.2). If the angle $P'OX = \theta$ then the projection on the $y$-axis at any time $t$ would be given by

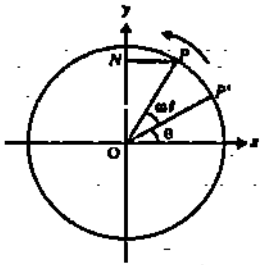


**Fig. 5.2: At $t = 0$, the point $P$ is at $P'$ and, therefore, the initial phase is $\theta$**

$$y = a \sin(\omega t + \theta) \qquad ...(5.5)$$

The quantity $(\omega t + \theta)$ is known as the phase of the motion and $\theta$ represents the initial phase. It is obvious from the discussion that the value of $\theta$ is quite arbitrary, and depends on the instant from which we start measuring time.

We next consider two particles, $P$ and $Q$ rotating on the circle with the same angular velocity $\omega$ and $P'$ and $Q'$ are their respective positions at $t = 0$. Let the angle $\angle P'OX$ and $\angle Q'OX$ be $\theta$ and $\phi$ respectively (see Fig. 5.3).

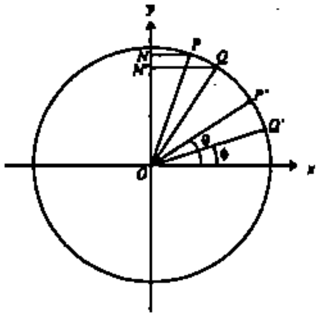


Fig.5.3: The points $N$ and $N'$ execute simple harmonic motion with the same frequency $\omega$. The initial phases of $N$ and N$'$ are $\theta$ and $\phi$ respectively.

Clearly at an arbitrary time $t$ the distance of the foot ot perpendiculars from the origin would be

$$y_P = a \sin(\omega t + \theta) \qquad ...(5.6a)$$

$$y_Q = a \sin(\omega t + \phi) \qquad ...(5.6b)$$

The quantity

$$(\omega t + \theta) - (\omega t + \phi) = \theta - \phi \qquad ...(5.7)$$

represents the phase difference between the two simple harmonic motions and if $\theta - \phi = 0$ (or an even multiple of $\pi$) the motions are said to be **in phase**, and if $\theta - \phi = \pi$ (or an odd multiple of $\pi$) the motions are said to be **out of phase**. If we choose a different origin of time, the quantities $\theta$ and $\phi$ would change by the same additive constant; consequently, the phase difference $(\theta - \phi)$ is independent of the choice of the instant $t = 0$

**Energy:** A particle performing simple harmonic motion possesses both types of energies; potential and kinetic. It possesses potential energy on account of its displacement from the equilibrium position and kinetic energy on account of its velocity. These energies vary during oscillation, however, their sum is conserved provided no dissipative forces are present. Since the acceleration of vibrating particle is $\omega^2 y$, the force needed to keep a particle of mass $m$ at a distance $y$ from $O$ is $m\ \omega^2 y$. If the particle is to be displaced through a further distance $dy$, the work to be done will be $\omega^2\ mydy$. Now the potential energy of the particle at a displacement $y$ is equal to the total work done to displace the particle from $O$ through a distance $y$.

$$\therefore \qquad \text{P.E.} = \int_0^y \omega^2 my\, dy = \frac{1}{2} m\ \omega^2 y^2 \qquad ...(5.8)$$

Using Eq. (5.2), the kinetic energy of the particle is given by

$$\text{K.E.} = \frac{1}{2} m \left(\frac{dy}{dt}\right)^2 = \frac{1}{2} m\omega^2 (a^2 - y^2) \qquad ...(5.9)$$

The total energy of the particle at any distance $y$ from $O$ is given by

$$\text{Total energy} = \text{K.E.} + \text{P.E.}$$

$$= \frac{1}{2} m\omega^2 (a^2 - y^2) + \frac{1}{2} m\omega^2 y^2$$

$$= \frac{1}{2} m\omega^2 a^2 \qquad \text{...(5.10)}$$

Therefore, total energy (intensity) is proportional to (amplitude)$^2$, and, since $\omega = 2\pi n$, $n$ being the frequency, the energy is also proportional to (frequency)$^2$.

If $I$ represents the intensity associated with a light wave then

$$I \alpha a^2$$

where $a$ represents the amplitude of the wave.

### Wave-motion

So far we considered a single particle, $P$, executing simple harmonic motion. Let us consider a number of particles which make a continuous elastic medium. If any one particle is set in vibration, each successive particle begins a similar vibration, but a little later than the one before it, due to inertia. Thus, the phase of vibration changes from particle to particle until we reach a particle at which the disturbance arrives exactly at the moment when the first particle has completed one vibration. This particle then moves in the same phase as the first particle. This simultaneous vibrations of the particles of the medium together make a wave. Such a wave can be represented graphically by means of a displacement curve drawn with the position of the particles as abscissae and the corresponding displacement at that instant as ordinate. If the particles execute simple harmonic motion, we obtain a sine curve as shown in Fig. 5.4.

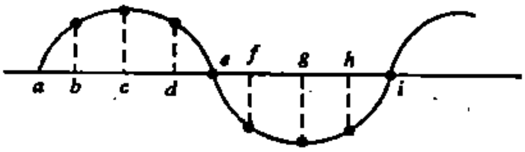


Fig. 5.4. Graphical representation of a wave

It will be seen that the wave originating at $a$ repeates itself after reaching $i$. The distance $ai$, after travelling which the wave-form repeats itself, is called the **wavelength** and is denoted by $\lambda$. It is also evident that during the time $T$, while the particle at $a$ makes one vibration, the wave travels a distance $\lambda$. Hence the velocity $v$ of the wave is given by

$$v = \frac{\lambda}{T}$$

If $n$ is the frequency of vibration then $n = 1/T$.

Hence, we have

$$v = n\lambda \qquad \text{...(5.11)}$$

### Particles in Same Phase

Particles $a$ and $i$ have equal displacements (= zero) and both are tending to move upwards. They are said to be in the same phase. The distance between them is one wavelength. Hence, wavelength is the distance between two nearest particles vibrating in the same phase. Two vibrating particles will also be in the same phase if the distance between them is $n\ \lambda$, where $n$ is an integer.

### Particles in Opposite Phase

Particles $a$ and $e$, both have the same displacement (= zero), but while $a$ is tending to go up, $e$ is tending to move downwards. They are said to be in opposite phase The distance between them is $\frac{\lambda}{2}$. The particles are out of phase if the distance between them is $(2n-1)\frac{\lambda}{2}$, where $n$ is an integer.

### Equation of a Simple Harmonic Wave

Fig. 5.5 shows the wave travelling in the positive $x$-direction. The displacement $y$ of the

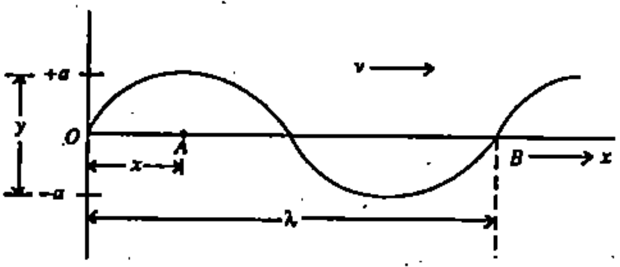


Fig. 5.5. A simple harmonic wave travelling towards right.

particle at $O$ at any time $t$ is given by

$$y = a \sin \omega t \qquad \ldots(5.1)$$

Let $v$ be the velocity of propagation of the wave. Then the wave starting from $O$ would reach at a point $A$, distant $x$ from $O$ in $x/v$ seconds. Hence the particle at $A$ must have started its vibration $x/v$ seconds later than the particle at $O$. Consequently, the displacement at $A$ at the time $t$ would be same as was at $O$ at time $\frac{x}{v}$ seconds earlier i.e. at time $t - \frac{x}{v}$. Substituting $t - \frac{x}{v}$ for $t$ in Eq. (5.1) we obtain the displacement at $A$ at time $t$, which is given by

$$y = a \sin \omega \left( t - \frac{x}{v} \right)$$

Using the relation $\omega = 2\pi/T$ and $v = \frac{\lambda}{T}$ we get

$$y = a \sin \frac{2\pi}{\lambda} (vt - x) \qquad \ldots(5.12)$$

This equation represents the displacement of a particle at a distance $x$ from a fixed point at a time $t$. This is, therefore, the equation of the wave. The wave shown in Fig. 5.5 is generated along a stretched string and in a rope. Such type of waves are called transverse waves. From Unit 4 of Block–1 , you already know that light travels in the form of transverse waves, therefore Eq. (5.12) represents a light wave.

### Relation between Phase Difference and Path Difference

The equation of simple harmonic wave is given by Eq. (5.12). If there are two particles $P_1$ and $P_2$ at distance $x_1$ and $x_2$ from the origin, then,

the phase angle of $P_1$ at a time $t = \frac{2\pi}{\lambda} (vt - x_1)$

and the phase angle of $P_2$ at a time $t = \frac{2\pi}{\lambda} (vt - x_2)$

$\therefore$ phase difference between $P_1$ and $P_2$

$$= \frac{2\pi}{\lambda} (vt - x_1) - \frac{2\pi}{\lambda} (vt - x_2)$$
$$= \frac{2\pi}{\lambda} (x_2 - x_1)$$

But $(x_2 - x_1)$ is the path difference between $P_2$ and $P_1$.

$$\therefore \text{Phase difference} = \frac{2\pi}{\lambda} \times \text{(path difference)} \qquad \ldots(5.13)$$

The expression phase difference $= \frac{2\pi}{\lambda}$ (path difference) can be obtained in a less formal manner by remembering that a difference in phase of $2\pi$ corresponds to a path difference of one wavelength and calculating the required phase difference by proportion.

When two or more sets of waves are made to overlap in some region of space, interesting effects are observed. For example, when two stones are dropped simultaneously in a quiet pool, two sets of waves are created. In the region of crossing, there are places where the disturbance is almost zero, and others, where it is greater than that given by either wave alone. These effects can be explained using a very simple law known as principle of superposition. We will use this principle in investigating the disturbance in regions, where two or more light waves are superimposed. Let us now briefly study this principle.

## 5.3 PRINCIPLE OF SUPERPOSITION

In any medium, two or more waves can travel simultaneously without affecting the motion of each other. Therefore, at any instant the resultant displacement of each particle of the medium is merely the vector sum of displacements due to each wave separately. This principle is known as "principle of superposition". It has been observed that when two sets of waves are made to cross each other, then after the waves have passed out of the region of crossing, they appear to have been entirely uninfluenced by the other set of waves. Amplitude, frequency and all other characteristics of the waves are as if they had crossed an undisturbed space.

As a simple example, we consider a long stretched string $AB$ (see Fig. 5.6). The end $A$ of the string is made to vibrate up and down. This vibration is handed down from particle to particle of the string. Suppose the string is vibrating in the form of a triangular pulse, which propagates to the right with a certain speed $v$. We next assume that from the end $B$ an identical pulse is generated which starts moving to the left with the same speed $v$.

Fig. 5.6(a) shows the position of pulse at $t = 0$. At a little later time, each pulse moves close to the other as shown in Fig. 5.6(b), without any interference. Fig. 5.6(c) represents the position at an instant when the two pulses interfere; the dashed curves represent the profile of the string, if each of the impulses were moving all by itself, whereas the solid curve shows the resultant displacement obtained by algebraic addition of each displacement. Shortly later in Fig. 5.6(d) the two pulses overlap each other and the resultant displacement is zero everywhere. At a much later time, the impulses cross each other (Fig. 5.6(e)) and move as if nothing had happened. This could hold provided the principle of superposition is true.

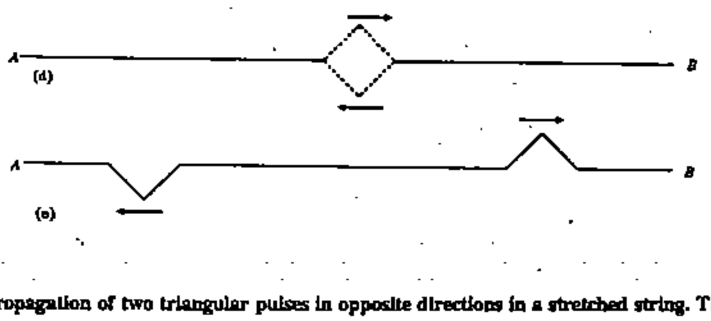


Fig.5.6: The propagation of two triangular pulses in opposite directions in a stretched string. The solid line gives the actual shape of the string; (a), (b), (c), (d) and (e) correspond to different instants of time.

Let us consider the following case of superposition of waves.

**Superposition of Two Waves of Same Frequency but having Constant Phase Difference**

Consider two waves of same frequency but having constant phase difference, say $\delta$. Since they have same frequency, i.e. same angular velocity, we write

$$y_1 = a_1 \sin \omega t$$

and $$y_2 = a_2 \sin (\omega t + \delta)$$

where $a_1$ and $a_2$ are two different amplitudes, and $\omega$ is common angular frequency of the two waves. By the principle of superposition, the resultant displacement is

$$\begin{aligned} y &= y_1 + y_2 \\ &= a_1 \sin \omega t + a_2 \sin (\omega t + \delta) \\ &= a_1 \sin \omega t + a_2 \sin \omega t \cos \delta + a_2 \cos \omega t \sin \delta \\ &= \sin \omega t(a_1 + a_2 \cos \delta) + \cos \omega t(a_2 \sin \delta) \end{aligned}$$

Let us write

$$a_1 + a_2 \cos \delta = A \cos \theta \qquad \text{... (5.14a)}$$

and $$a_2 \sin \delta = A \sin \theta \qquad \text{...(5.14b)}$$

where $A$ and $\theta$ are new constants. This gives

$$y = \sin \omega t \, A \cos \theta + \cos \omega t \, A \sin \theta$$

or $$y = A \sin (\omega t + \theta)$$

Hence the resultant displacement is simple harmonic and of amplitude $A$. Squaring and adding Eq. 5.14a and 5.14b, we get

$$A^2 \cos^2\theta + A^2 \sin^2\theta = (a_1 + a_2 \cos\delta)^2 + (a_2 \sin \delta)^2$$

or, $$A^2 = a_1{}^2 + a_2{}^2 + 2a_1 \, a_2 \cos\delta$$

Thus, the resultant intensity $I$ which is proportional to the square of the resultant amplitude, is given as

$$I = A^2 = a_1{}^2 + a_2{}^2 + 2a_1 \, a_2 \cos\delta \qquad \text{...(5.15)}$$

(Here we have taken the constant of proportionality as 1, for simplicity).

Thus, we find that the resultant intensity is not equal to the sum of the intensities due to separate waves i.e., $(a_1{}^2 + a_2{}^2)$. Since the intensity of wave is proportional to square of amplitude, $I_1 \alpha \, a_1{}^2$ and $I_2 \alpha \, a_2{}^2$ as before, taking the proportionality constant as 1, we can rewrite Eq. (5.15) as

$$I = I_1 + I_2 + 2 \sqrt{I_1 I_2} \cos \delta \qquad \text{...(5.16)}$$

In Example 1, see how Eq. (5.16) has been used to find the resultant intensity.

**Example 1**

Consider interference due to two coherent waves of same frequency and constant phase difference having intensities $I$ and $4I$, respectively. What is the resultant intensity when the phase difference between these two waves is $\pi/2$ and $\pi$?

**Solution**

According to Eq. (5.16)

$$I = I_1 + I_2 + 2 \sqrt{I_1 I_2} \cos \delta$$

Given: $I_1 = I$ and $I_2 = 4I$, so

$$\begin{aligned} I &= 5I + 2I \sqrt{4} \cos \delta \\ &= 5I + 4I \cos \delta \end{aligned}$$

Hence $$I_{\pi/2} = 5I + 4I \cos 90^0 = 5I$$

$$I_\pi = 5I + 4I \cos\pi = I$$

Thus there is a variation of intensity due to interference phenomenon.

Refer again to Eq. (5.16). The intensity $I$ is maximum when $\cos\delta = +1$, that is, when phase difference is given by

$$\delta = 2n\pi \text{ (even multiple of } \pi).$$

From Eq. (5.16)

$$I_{\max} = I_1 + I_2 + 2\sqrt{I_1 I_2}$$

The resultant intensity is, thus, greater than the sum of the two separate intensities. If

$$I_1 = I_2 \text{ then } I_{\max} = 4I_1$$

The intensity $I$ is minimum when $\cos\delta = -1$, i.e., when $\delta$ is given by

$$\delta = (2n+1)\,\pi \text{ (odd multiple of } \pi).$$

We have from Eq. (5.16)

$$I_{\min} = I_1 + I_2 - 2\sqrt{I_1 I_2}$$

The resultant intensity is thus less than the sum of two separate intensities. If $I_1 = I_2$, then $I_{\min} = 0$, which means that there is no light.

## SAQ 3

Two waves of same frequency and constant phases difference have intensities in the ratio 81:1. They produce interference fringes. Deduce the ratio of the maximum to minimum intensity.

In general, for the two waves of same intensity and having a constant phase difference of $\delta$, the resultant intensity is given by

$$I = 2I_1 + 2I_1\cos\delta \qquad (\because I_1 = I_2)$$

$$= 2I_1(1+\cos\delta)$$

$$= 4I_1\cos^2\frac{\delta}{2} \qquad \ldots(5.17)$$

Therefore, we find that when two waves of the same frequency travel in approximately the same direction and have a phase difference that remains constant with the passage of time, the resultant intensity of light in not distributed uniformly in space. The non-uniform distribution of the light intensity due to the superposition of two waves is called **interference**. At some points the intensity is maximum and the interference at these points is called constructive interference. At some other points the intensity is minimum and the interference at these points is called destructive interference.

**Usually, when two light waves are made to interfere, we get alternate dark and bright bands of a regular or irregular shape. These are called interference fringes.**

## SAQ 4

Fig. 5.7 shows two situations where waves emanating from two sources, $A$ and $B$, arrive at point $C$ and interfere. Which of the two situations indicate constructive interference and destructive interference? Give reasons. (Eq. (5.13) will help you in answering this question.)

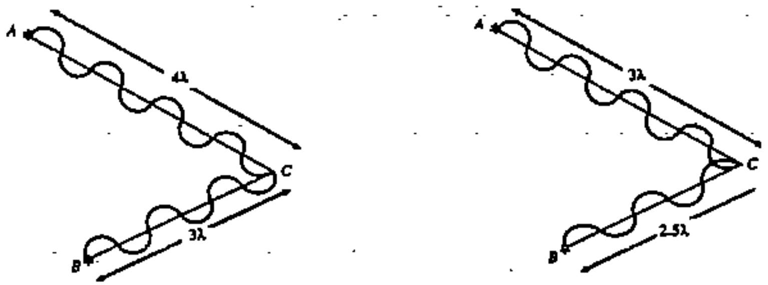


**Fig. 5.7:**

After solving the above SAQ one can infer that:

for constructive interference,

$$\boxed{\text{path difference} = n\lambda, \text{ where } n = 0, 1, 2, 3} \quad \ldots(5.18)$$

for destructive interference,

$$\boxed{\text{path difference} = m\frac{\lambda}{2}, \text{ where } m = 1, 3, 5, 7} \quad \ldots(5.19)$$

For the production of stationary interference patterns, i.e. definite regions of constructive and destructive interference, the interfering waves must have (1) the same frequency, and (2) a constant phase difference (and they must be travelling in the same or nearly the same direction). If these conditions are satisfied, we say the wave sources and the waves are coherent. Sources can readily be found with the same vibrating frequency; however, the phase relationship between the waves may vary with time. In the case of light, the waves are radiated by the atoms of a source. Each atom contributes only a small part to the light emitted from the source and the waves bear no particular phase relationship to each other; the atoms randomly emit light, so the phase "constant" of the total light wave varies with time. Hence, light waves brought together from different light sources are coherent over very short periods of time and does not produce stationary interference patterns. Light from two lasers (about this you will study in Block 4) can be made to form stationary interference patterns, but the lasers must be phase-locked by some means. How, then, was the wave nature of light originally investigated, since lasers are a relatively recent development ? In the following sections we will discuss the various arrangements, which provide coherent sources and enable us to observe interference phenomenon. Thomas Young had first demonstrated the interference of light. In the next section we will describe the experiment done by him.

## 5.4 YOUNG'S DOUBLE-SLIT EXPERIMENT

One of the earliest demonstrations of such interference effect was first done by Young in 1801, establishing the wave character of light. Young allowed sunlight to fall on a pinhole $S_0$, punched in a screen $A$ as shown in Fig. 5.8. The emergent light spreads out and falls on pinholes $S_1$ and $S_2$, punched in the screen $B$. Pinholes $S_1$ and $S_2$ act as coherent sources. Again, two overlapping spherical waves expand into space to the right of screen $B$. Fig. 5.8 shows how Young produced an interference pattern by allowing the waves from pinholes $S_1$ and $S_2$ to overlap on screen $C$.

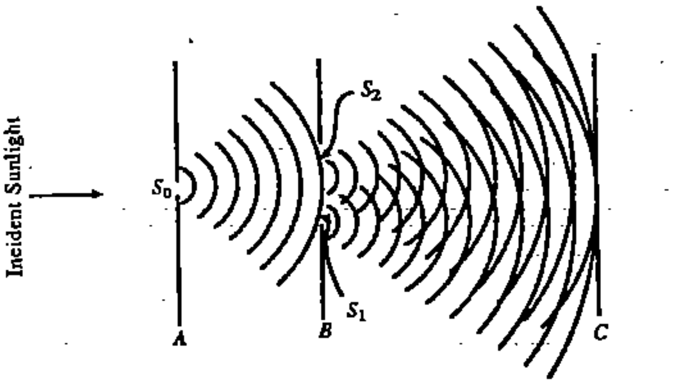


Fig.5.8: Young's double slit experiment. The pinholes $S_1$ and $S_2$ act as coherent sources and an interference pattern is observed on the screen $C$.

Fig. 5.9 shows the section of the wavefront on the plane containing $S_0$, $S_1$, and $S_2$. Since the waves emanating from $S_1$ and $S_2$ are coherent, we will see'alternate bright and dark curves of fringes, called interference fringes. The interference pattern is symmetrical about a bright central fringe (also called maximum), and the bright fringes decrease in intensity, the farther they are from the central fringe.

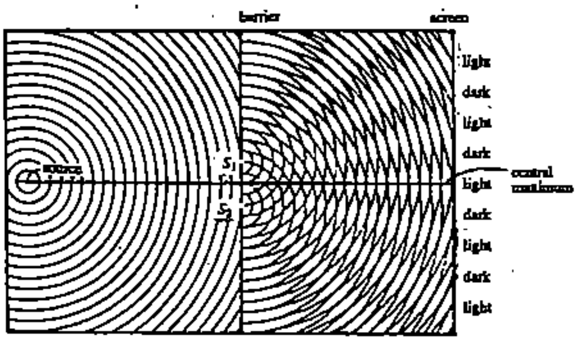


Fig. 5.9: Sections of the spherical wavefronts emanating from $S_0$, $S_1$ and $S_2$.

To analyze the interference pattern and investigate the spacing of the interference fringes, consider the geometry in Fig. 5.10. Let $S$ be a narrow slit illuminated by monochromatic light, and $S_1$ and $S_2$ two parallel narrow slits very close to each other and equidistant from $S$. The light waves from $S$ arrive at $S_1$ and $S_2$ in the same phase. Beyond $S_1$ and $S_2$, the waves proceed as if they started from $S_1$ and $S_2$ with the same phase because the two slits are equidistant from $S$.

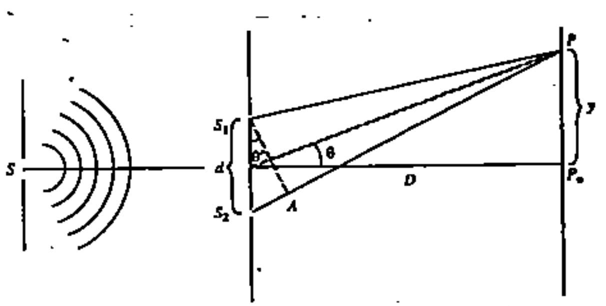


Fig.5.10: The geometry of Young's experiment: The path difference of the light from the slits arriving at $P$ on the screen is $d\ \sin\theta$.

It is assumed that the waves start out at the same phase, because the two slits $S_1$ and $S_2$ are equidistant from $S$. Furthermore, the amplitudes are the same, because $S_1$ and $S_2$ are the same size slits and very close to each other. (So the amplitude does not vary very much.) Hence these waves produce an interference pattern on a screen placed parallel to $S_1$ and $S_2$.

To find the intensity at a point $P$ on the screen, we join $S_1P$ and $S_2P$. The two waves arrive at $P$ from $S_1$ and $S_2$ having traversed different paths $S_1P$ and $S_2P$. Let us calculate this path difference $S_2P - S_1P$. Let,

$y$ = distance of $P$ from $P_0$, the central point on the screen

$d$ = separation of two slits $S_1$ and $S_2$,

$D$ = distance of slits from the screen.

The corresponding path difference is the distance $S_2A$ in Fig. 5.10, where the line $S_1A$ has been drawn to make $S_1$ and $A$ equidistant from $P$. As Young's experiment is usually done with $D >> d$ or $y$, the angle $\theta$ and $\theta'$ are nearly same and they are small.

Hence, we may assume triangle $S_1A\ S_2$ as a right-angled triangle and $S_2A = d \sin \theta' = d \sin \theta = d \tan\theta$, as for small $\theta$, $\sin\theta = \tan \theta$. As can be seen from the Fig. 5.10, $\tan \theta = y/D$.

$$\therefore \quad S_2P - S_1P = S_2A = d\frac{y}{D} \qquad \text{...(5.20)}$$

Now the intensity at the point $P$ is a maximum or minimum according as the path difference $S_2P - S_1P$ is an integral multiple of wavelength or an odd multiple of half wavelength (See Eq. 5.18 and Eq. 5.19). Hence, for bright fringes (maxima),

$$S_2P - S_1P = \frac{yd}{D} = 0, \lambda, 2\lambda, 3\lambda.... = m\lambda$$

where $m = 0, 1, 2.....$

$$\therefore \quad y = mD\ \lambda/d \quad \text{(bright fringes)} \qquad \text{...(5.21)}$$

The number $m$ is called the order of the fringe. Thus the fringes with $m = 0, 1, 2, ....$ etc. are called zero, first, second....etc. orders. The zeroth order fringe corresponds to the central maximum, the first order fringe ($m$ =1) corresponds to the first bright fringe on either side of the central maximum, and so on. For dark fringes (minima),

$$S_2P - S_1P = \frac{yd}{D} = \frac{\lambda}{2}, \frac{3\lambda}{2}, \frac{5\lambda}{2} ... = \left(m + \frac{1}{2}\right)\lambda$$

where $m$ =0, 1, 2, ...

$$y = \left(m + \frac{1}{2}\right)\frac{\lambda D}{d} \quad \text{(dark fringes)} \qquad \text{...(5.22)}$$

Eq. (5.21) or Eq. (5.22) can be used to find out the distance $y_n$ of the $n$th order bright (or dark) fringe. Try to solve the following SAQ.

SAQ 5

Monochromatic light passes through two narrow slits 0.40 mm apart. The third-order bright fringe of the interference pattern, observed on a screen 1.0 meter from the slits, is 3.6 mm from the centre of the central maximum. What is the wavelength of the light ?

**Fringe Width**

If $y_n$ and $y_{n+1}$ denote the distances of $n$th and ($n$+1)th bright fringes, then

$$y_n = \frac{D}{d}n\lambda$$

and

$$y_{n+1} = \frac{D}{d}(n+1)\lambda$$

The spacing between the $n$th and ($n$+1)th fringes (bright) is given by

$$y_{n+1} - y_n = \frac{D}{d}(n+1)\lambda - \frac{D}{d}n\lambda = D\lambda/d$$

It is independent of $n$. Hence, the spacing between any two consecutive bright fringes is the same. Similarly, it can be shown that the spacing between two dark fringes is also $\frac{D}{d}\lambda$. The spacing between any two consecutive bright or dark fringes is called the **fringe-width,** which is denoted by $\beta$. Thus

$$\beta = \frac{D}{d}\lambda \qquad \text{...(5.23)}$$

One also finds, by experiment, that fringe-width

i) varies directly as $D$,

ii) varies directly as the wave-length of the light used, and

iii) inversely as the distance $d$ between the slits

The fringe-widths are so fine that to see them, one usually uses magnifier or eye-piece.

To make certain that you really understand the meaning of the fringe width, try the following SAQs.

### SAQ 6

In a two-slit interference pattern with $\lambda = 6000$ Å, the zero order and tenth order maxima fall at 12.34 mm and 14.73 mm respectively. Find the fringe width.

### SAQ 7

If in the SAQ 6, $\lambda$ is changed to 5000 Å, deduce the positions of the zero order and twentieth order fringes, other arrangements remaining the same.

### Shape of the Interference Fringes

In Fig. 5.11, suppose $S_1$ and $S_2$ represent the two coherent sources. At the point $P$, there is maximum or minimum intensity according as

$$S_2P - S_1P = n\lambda$$

or

$$S_2P - S_1P = \left(n + \frac{1}{2}\right)\frac{\lambda}{2}$$

Thus for a given value of $n$, the locus of points of maximum or minimum intensity is given by

$$S_2P - S_1P = \text{constant},$$

which is the equation of a hyperbola with $S_1$ and $S_2$ as foci. In space, the locus of points of maximum or minimum intensity for a particular value of $n$ will be a hyperboloid of revolution, obtained by revolving the hyperbola about the line $S_1S_2$.

In practice, fringes are observed on a screen $XY$ in a plane normal to the plane of the figure and parallel to the line joining $S_1S_2$. Hence the fringes that are observed are simply the sections of the hyperboloids by this plane, i.e. they are hyperbolae. Since the wave-length of light is extremely small (of the order of $10^{-5}$cm), the value of $(S_2P - S_1P)$ is also of that order. Hence these hyperbolae appear, more or less, as straight lines.

### Intensity Distribution in the Fringe-System

To find the intensity, we rewrite Eq. (5.15), taking $a_1 = a_2$, as follows

$$I = A^2 = 2a^2\,(1 + \cos\delta)$$
$$= 4a^2 \cos^2\frac{\delta}{2}$$

If the phase difference is such that $\delta = 0, 2\pi, 4\pi$......, this gives $4a^2$ or 4 times the intensity of either beam. If $\delta = \pi, 3\pi, 5\pi$,......, The intensity is zero.

In between the intensity varies as $\cos^2 \delta/2$. Fig. 5.12 shows a plot of the intens'.y against the phase difference. When the two beams of light arrive at a point on the screen, exactly out of phase, they interfere destructively, and the resultant intensity is zero. One may well ask what becomes of the energy of the two beams, since the law of conservation of energy

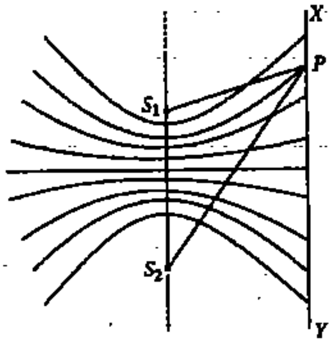


Fig. 5.11 : Shape of the fringes.

tells us that it cannot be destroyed. The answer to this question is that the energy, which apparently disappears at the minima, is actually still present at the maxima, where the intensity is greater than would be produced by the two beams acting separately. In other words, the energy is not destroyed, but merely redistributed in the interference pattern. The average intensity on the screen is exactly what would exist in the absence of interference. Thus, as shown in Fig. 5.12, the intensity in the interference pattern varies between $4A^2$ and zero. Now each beam, acting separately, would contribute $A^2$, and so, without interference, we would have a uniform intensity of $2A^2$, as indicated by the broken line. Let us obtain the average intensity on the screen for $\pi$ fringes. We have

$$I_{\text{average}} = \frac{\int_0^{\pi} I\, d\delta}{\int_0^{\pi} d\delta}$$

$$= \frac{\int_0^{\pi} \left(4A^2 \cos^2 \frac{\delta}{2}\right) d\delta}{\int_0^{\pi} d\delta}$$

$$= \frac{\int_0^{\pi} (2A^2 + 2A^2 \cos\delta)\, d\delta}{\int_0^{\pi} d\delta} \qquad \left(\because 1 + \cos\theta = \cos^2 \frac{\theta}{2}\right)$$

$$= \frac{[2A^2\, \delta + 2A^2 \sin\delta]_0^{\pi}}{[\delta]_0^{\pi}}$$

$$= \frac{2A^2\, \pi}{\pi}$$

$$= 2A^2$$

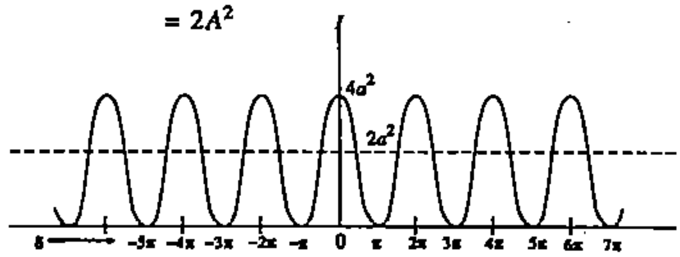


**Fig. 5.12: Intensity distribution for the interference fringes from two waves of the same frequency.**

Thus, the average intensity is equal to the sum of the separate intensities. That is whatever energy apparently disappears at the minima is actually present at the maxima. There is no violation of the law of conservation of energy in the phenomenon of interference.

Till now we have considered interference pattern produced when a monochromatic light from a narrow slit falls on two parallel slits. What happens if white light is used to illuminate slits? Read the following sub-section.

### 5.4.1 White-Light Fringes

If white light is used to illuminate the slits we obtain an interference pattern consisting of a central 'white' fringe, having on both sides a few coloured fringes and then a general illumination.

A pair of white light coherent sources is equivalent to a number of pairs of monochromatic sources. Each monochromatic pair produces its own system of fringes with a different fringe-width $\beta$, since $\beta$ depends on $\lambda \left(\beta = \frac{D\lambda}{d}\right)$.

At the centre of the pattern, the path difference between the interfering waves is zero. Therefore, the path difference is also zero for all wavelengths. Hence, all the different coloured waves of the white light produce a bright fringe at the centre. This superposition of the different colours makes the central fringe 'white'. This is the 'zero order fringe'.

As we move on either side of the centre, the path difference gradually increases from zero. At a certain point it becomes equal to half the wavelength of the component having the smallest wave-length, i.e., violet. This is the position of the first dark fringe of violet. Beyond this, we obtain the first minimum of blue, green, yellow and of red in the last. The inner edge of the first dark fringe, which is the first minimum for violet, receives sufficient intensity due to red, hence it is reddish. The outer edge of the first dark fringe, which is minimum for red, receives sufficient intensity due to violet, and is therefore, violet. The same applies to every other dark fringe. Hence, we obtain a few coloured fringes on both sides of the central fringe.

As we move further away from the centre, the path difference becomes quite large. Then, from the range 7500 – 4000 Å, a large number of wavelengths (colours) will produce maximum intensity at a given point, and an equally large number will produce minimum intensity at that point. For example, at any point $P$, we may have

$$\text{path difference} \begin{cases} = 11\ \lambda_1 = 12\lambda_2 = 13\lambda_3 = \ldots\text{etc. (maxima)} \\ \\ = \left(11 + \frac{1}{2}\right)\lambda_1' = \left(12 + \frac{1}{2}\right)\lambda_2' = \left(13 + \frac{1}{2}\right)\lambda_3' = \ldots \text{ etc. (minima)} \end{cases}$$

Thus, at $P$, we shall have 11th, 12th, 13th, etc., bright fringes of $\lambda_1$, $\lambda_2$, $\lambda_3$,...etc., and 11th, 12th, 13th,... etc., dark fringes of $\lambda'_1$, $\lambda'_2$, $\lambda'_3$, ...etc. Hence, the resultant colour at $P$ is very nearly white. This happens at all points, for which the path difference is large. Hence, in the region of large path difference uniform white illumination is obtained.

For maxima, path difference $= n\lambda$, where $n = 0, 1, 2, ..$

For minima, path difference $= \left(n + \frac{1}{2}\right)\lambda$, where $n = 0, 1, 2, \ldots$

**SAQ 8**

Let the path difference $S_1P - S_2P = 30 \times 10^{-5}$ cm. What are the $\lambda$'s for which the point $P$ is a maximum?

In the usual interference pattern with a monochromatic source, a large number of interference fringes are obtained, and it becomes extremely difficult to determine the position of the central fringe. Hence, by using white light as a source the position of central fringe can be easily determined.

### 5.4.2 Displacement of Fringes

We will now discuss the change in the interference pattern produced when a thin transparent plate, say of glass or mica, is introduced in the path of one of the two interfering beams, as shown in Fig. 5.13. It is observed that the entire fringe-pattern is displaced to a point towards the beam in the path of which the plate is introduced. If the displacement is measured, the thickness of the plate can be obtained provided the refractive index of the plate and the wavelength of the light are known.

Suppose a thin transparent plate of thickness $t$ and refractive index $\mu$ is introduced in the path of one of the constituent interfering beams of light (say in the path of $S_1P$, shown in Fig. 5.13). Now, light from $S_1$ travel partly in air and partly in the plate. For the light path from $S_1$ to $P$, the distance travelled in air is $(S_1P - t)$, and that in the plate is $t$. Suppose, $c$ and $v$ be the velocities of light in the air and in the plate, respectively. If the time taken by light beam to reach from $S_1$ to $P$ is, $T$, then

$$T = \frac{S_1P - t}{c} + \frac{t}{v}$$

or,

$$T = \frac{S_1P - t}{c} + \frac{\mu t}{c} \qquad \left(\because v = \frac{c}{\mu}\right)$$

$$= \frac{S_1P + (\mu - 1)\,t}{c}$$

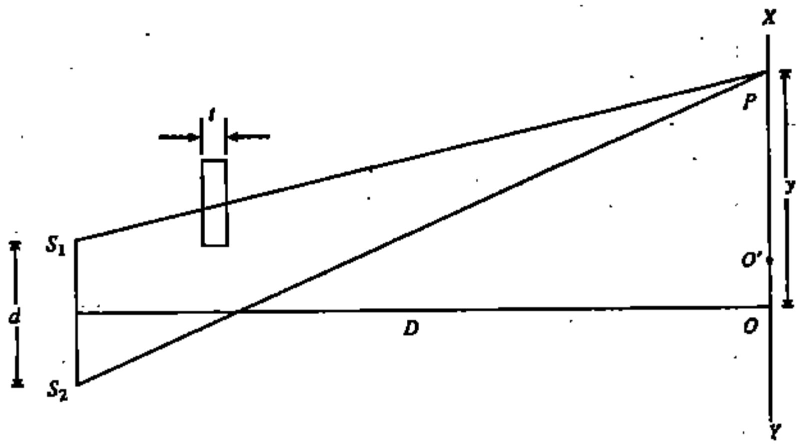


Fig. 5.13: If a thin transparent sheet (of thickness $t$) is introduced in one of the beams, the fringe pattern gets shifted by a distance $(\mu-1)\ tD/d$.

Thus the effective path in air from $S_1$ to $P$ is $[S_1P + (\mu-1)t]$, i.e., the air path $S_1P$ is increased by an amount $(\mu-1)t$, due to the introduction of the plate of material of refractive index, $\mu$.

Let $O$ be the position of the central bright fringe in the absence of the plate, the optical paths $S_1O$ and $S_2O$ being equal. On introducing the plate, the two optical paths become unequal. Therefore, the central fringe is shifted to $O'$, such that at $O'$ the two optical paths become equal. A similar argument applies to all the fringes. Now, at any point $P$, the effective path difference is given by

$$S_2P - [S_1P + (\mu-1)t]$$

$$= (S_2P - S_1P) - (\mu - 1)t$$

From Eq. (5.20), $S_2P - S_1P = \frac{d}{D}y$

$\therefore$ Effective path difference at $P = \frac{d}{D}y - (\mu - 1)\,t$

If the point $P$ is to be the centre of the $n$th bright fringe, the effective path difference should be equal to $n\lambda$ i.e.,

$$\frac{d}{D}y_n - (\mu - 1)\,t = n\lambda$$

or

$$\boxed{y_n = \frac{D}{d}[n\lambda + (\mu - 1)\,t]} \qquad \text{...(5.24)}$$

In the absence of the plate ($t = 0$), the distance of the $n$th bright fringe from $O$ is $\frac{D}{d}n\lambda$.

$\therefore$ Displacement $y_0$ of the $n$th bright fringe is given by

$$y_0 = \frac{D}{d}[n\lambda + (\mu - 1)\,t] - \frac{D}{d}n\lambda$$

$$\boxed{y_0 = \frac{D}{d}(\mu - 1)\,t} \qquad \text{...(5.25)}$$

The shift is independent of the order of the fringe, showing that shift is the same for all the bright fringes. Similarly, it can be shown that the displacement of any dark fringe is also given by Eq. (5.25). Thus, the entire fringe-system is displaced through a distance $\frac{D}{d}(\mu - 1)\,t$ towards the side on which the plate is placed. The fringe-width is given by:

$$\beta = y_{n+1} - y_n$$

$$= \frac{D}{d}[(n + 1)\,\lambda + (\mu - 1)\,t] - \frac{D}{d}[n\lambda + (\mu - 1)\,t] \qquad \text{(see Eq. (5.24))}$$

$$= \frac{D\lambda}{d},$$

which is the same as before the introduction of the plate.

Eq. (5.25) enables us to determine the thickness of extremely thin transparent sheets (like that of mica) by measuring the shift of the fringe system.

Now, apply this strategy yourself to SAQ 9.

SAQ 9

In a double slit interference arrangement one of the slits is covered by a thin micą sheet whose refractive index is 1.58. The distances $S_1S_2$ and $AO$ (see Fig. 5.13) are 0.1 cm and 50 cm, respectively. Due to the introduction of the mica sheet, the central fringe gets shifted by 0.2 cm. Determine the thickness of the mica sheet.

## 5.5 FRESNEL'S BIPRISM

With regard to Young's double-slit experiment, objection was raised that the bright fringes observed by Young were probably due to some compiicated modification of the light by the edges of the slits and not due to interference. Soon after, Fresnel devised a series of arrangements to produce the interference of two beams of light which was not subject to this criticism. One of the experimental arrangements, known as Fresnel's Biprism arrangement, is shown in Fig. 5.14.

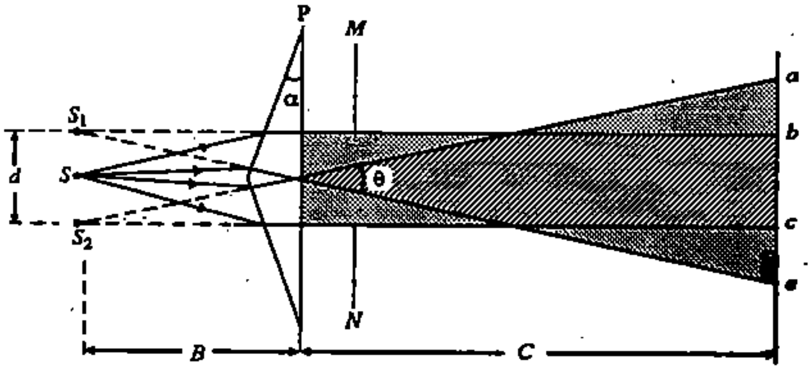


Fig. 5.14: Diagram of Fresnel's Biprism experiment.

$S$ is a narrow vertical slit illuminated by monochromatic light. The light from $S$ is allowed to fall symmetrically on the Biprism $P$, placed at a small distance from $S$ and having its refracting edges parallel to the slit. The light emerging from the upper and lower halves of the prism appears to start from two virtual images, $S_1$ and $S_2$ of $S$, which act as coherent sources. The cones of light $bS_1e$ and $aS_2c$, diverging from $S_1$ and $S_2$, are superposed and the interference fringes are formed in the overlapping region $bc$.

If screens $M$ and $N$ are placed, as shown in the Fig. 5.14, interference fringes are observed only in the region $bc$. When the screen $ae$ is replaced by a photographic plate, a picture like the upper one, in Fig. 5.15, is obtained.

The closely spaced fringes in the centre of the photograph are due to interference, while the wide fringes at the edge of the photograph are due to diffraction. These wider bands are due to the vertices of the two prisms, each of which act as a straight edge, giving a pattern of diffraction (about this you will learn in Block 3). When the screens $M$ and $N$ are removed from the light path, the two beams overlap over the whole region $ae$. The lower photograph in Fig. 5.15 shows for this case the equally spaced interference fringes superimposed on the diffraction pattern, of a wide aperture.

Fig. 5.15: Interference and diffraction fringes produced in the Fresnel Biprism experimental arrangement.

With such an experiment, Fresnel was able to show the interference effect without the diffracted beams through the two slits. Just as in Young's double slit experiment, this arrangement can also be used to determine the wavelength of monochromatic light. The light illuminates the slit $S$ and interference fringes can be easily viewed through the eye-piece. The fringe-width $\beta$ can be determined by means of a micrometer attached to the eye piece. If $D$ is the distance between source and screen, and $d$ the distance between the virtual images $S_1$ and $S_2$, the wave-length is given by

$$\lambda = \frac{\beta d}{D} \qquad ...(5.26)$$

The distances $d$ and $D$ can easily be determined by placing a convex lens between the Biprism and the eyepiece. For a fixed position of the eyepiece, there will be two positions of the lens, shown as $L_1$ and $L_2$ in Fig. 5.16 where the images of $S_1$ and $S_2$ can be seen at the eyepiece. Let $d_1$ be the distance between the two images, when the lens is at the position $L_1$ (at a distance $b_1$ from the eyepiece). Let $d_2$ and $b_2$ be the corresponding distances, when the lens is at $L_2$. Then it can easily be shown that

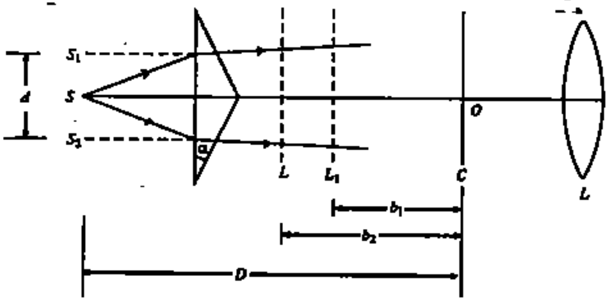


Fig. 5.16: Fresnel's biprism arrangement. $C$ and $L$ represents the position of cross wires and the eyepiece, respectively. In order to determine $d$ a lens is introduced between the biprism and cross wires. $L_1$ and $L_2$ represent the two positions of the lens where the slits are clearly seen.

$$d = \sqrt{d_1 d_2} \qquad ...(5.27a)$$

and

$$D = b_1 + b_2 \qquad ...(5.27b)$$

Use Eq. (5.26) and (5.27) to solve the following SAQ.

SAQ 10

In a Fresnel's Biprism experiment, the eyepiece is at a distance of 100 cm from the slit. A convex lens inserted between the Biprism and the eyepiece gives two images of the slit in two positions. In one case, the two images of the slit are 4.05 mm apart, and in the other case 2.10 mm apart. If sodium light of wavelength 5893 Å is used, find the thickness of the interference fringes.

## 5.6 SOME OTHER ARRANGEMENT FOR PRODUCING INTERFERENCE BY DIVISION OF WAVEFRONT

Two beams may be brought together in several other ways to produce interference. In **Fresnel's two-mirror** arrangement, light from a slit is reflected in two plane mirrors slightly inclined to each other. The mirror produces two virtual images of the slit, as shown in Fig. 5.17.

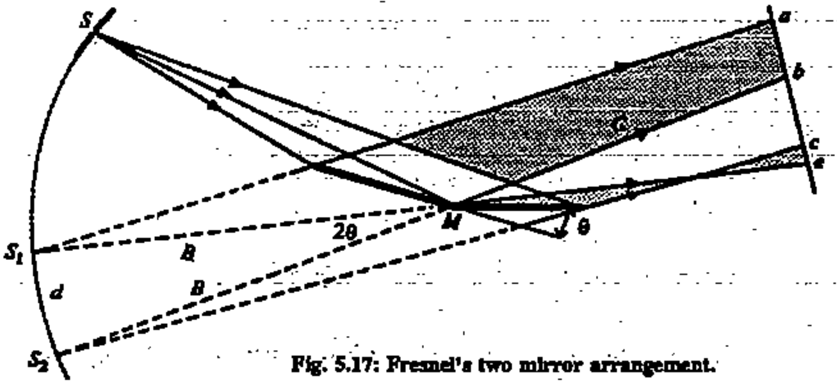


Fig. 5.17: Fresnel's two mirror arrangement.

They are like the images in Fresnel's biprism, and interference fringes are observed in the region *bc*, where the reflected beams overlap:

Even a simpler mirror method is available. This is known as **Lloyd's mirror**. Here the slit and its virtual image constitute the double source.

At grazing incidence, almost the entire incident light is reflected so that the direct and the reflected beam have nearly equal amplitudes. Hence the fringes have good contrast.

### Lloyd's Mirror

It is a simple arrangement to obtain two coherent sources of light to produce a stationary interference pattern. It consists of a plane mirror *MN* (Fig. 5.18) polished on the front surface and blackened at the back (to avoid multiple reflection). $S_1$ is a narrow slit, illuminated by monochromatic light, and placed with its length parallel to the surface of the mirror. Light from $S_1$ falls on the mirror at nearly grazing incidence, and the reflected beam appears to diverge from $S_2$, which is the virtual image of $S_1$. Thus $S_1$ and $S_2$ act as coherent sources. The direct cone of light $AS_1E$ and the reflected cone of light $BS_2C$ are superposed, and the interference fringes are obtained in the overlapping region *BC* on the screen.

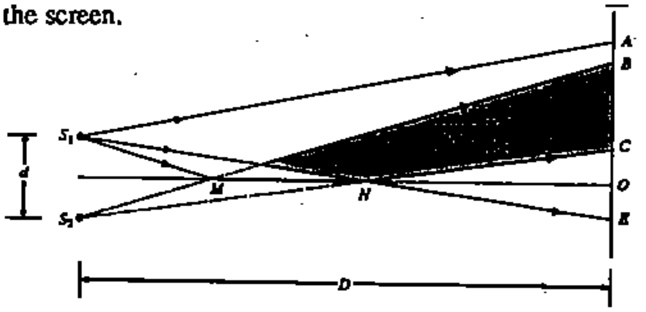


Fig. 5.18: Lloyd's mirror

### Zero-Order Fringe

The central zero-order fringe, which is expected to lie at *O* (the perpendicular bisector of $S_1S_2$) is not usually seen since only the direct light, and not the reflected light, reaches *O*. It can be seen by introducing a thin sheet of mica in the path of light from $S_1$, when the entire fringe system is displaced in the upward direction. (You could see this yourself while solving SAQ 11.)

---

SAQ 11

Interference bands are obtained with a Lloyd's mirror with light of wavelength $5.45 \times 10^{-5}$ cm. A thin plate of glass of refractive index 1.5 is then placed normally in the path of one of the interfering beams. The central dark band is found to move into the position previously occupied by the third dark band from the centre. Calculate the thickness of the glass plate.

---

With white light the central fringe is expected to be white, but actually it is found to be 'dark'. This is because the light suffers a phase change of $\pi$ or a path-difference of $\frac{\lambda}{2}$ when reflected from the mirror. Therefore, the path difference between the interfering rays at the position of zero-order fringe becomes $\frac{\lambda}{2}$ (instead of zero), which is a condition for a minimum. Hence the fringe is dark.

### Determination of Wavelength

Let *d* be the distance between the coherent sources $S_1$ and $S_2$, and *D* the distance of the screen from the sources. The fringe–width is then given by

$$\beta = \frac{D\lambda}{d}$$

Thus, knowing $\beta$, *D* and *d*, the wavelength $\lambda$ can be determined.

### Acromatic Fringes and their Production by Lloyd's Mirror

A system of white and dark fringes, without any colours, obtained by white light are known as 'achromatic fringes'.

Ordinarily, with white light, we obtain a central white fringe, having on either side of it a few coloured fringes (as you have studied in subsection 5.4.1). This is because the fringe-width $\beta = \frac{D\lambda}{d}$ is different for different wavelengths (colours). If however, the fringe-width is made the same for all wavelengths, the maxima of each order for all wavelengths will coincide, resulting into achromatic fringes. That is, for achromatic fringes, we must have

$$\frac{D\lambda}{d} = \text{constant}$$

$$\text{or} \qquad \frac{\lambda}{d} = \text{constant}$$

We can easily realise this condition with a Lloyd's mirror by using a slit illuminated by a narrow spectrum of the white light as shown in Fig. 5.19. The narrow spectrum $R_1\ V_1$ is produced by a prism, or, preferably, by a plane diffraction grating. The Lloyd's mirror is placed with its surface close to the violet end of the spectrum and such that $R_1\ V_1$ is perpendicular to its plane.

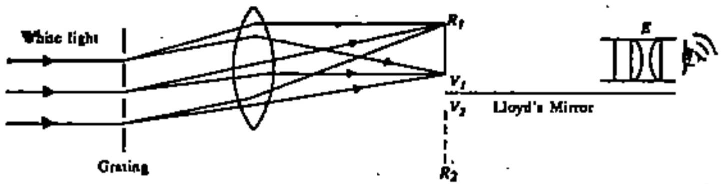


Fig. 5.19 : Achromatic fringes produced by Lloyd's mirror.

$R_1\ V_1$, and its virtual image, $R_2\ V_2$, formed by the mirror act as coherent sources. They are equivalent to a number of pairs of sources of different colours. Thus, the pair $R_1\ R_2$ produces a set of red fringes, and the pair $V_1\ V_2$ a set of violet fringes. The intermediate pairs produce the sets of fringes of intermediate colours. The red and violet fringes will be of the same width if

$$\frac{\lambda}{d} = \text{constant}$$

$$\text{i.e.} \qquad \frac{\lambda_r}{d_r} = \frac{\lambda_v}{d_v}$$

$$\text{or} \qquad \frac{d_r}{d_v} = \frac{\lambda_r}{\lambda_v}$$

where $d_r$ is the distance $R_1\ R_2$, and $d_v$ the distance $V_1\ V_2$.

Hence, the last expression gives

$$\frac{R_1 R_2}{V_1 V_2} = \frac{\lambda_r}{\lambda_v}$$

Therefore, if the distance of the violet end $V_1$ from the surface of the mirror is so adjusted by displacing the mirror laterally that the above condition is satisfied, the red and violet fringes will have the same width, and will exactly be superposed on each other. Since, in a grating spectrum, the dispersion is accurately proportional to the wavelength, the condition $(\lambda/d)$ = constant is simultaneously satisfied for all the wavelengths. Thus, when this adjustment is made, fringes of all colours are superposed on one another. Hence, achromatic fringes are observed in the eyepiece $E$ placed in the over-lapping region.

### Difference between Biprism and Lloyd's Mirror Fringes

The following are the main points of difference between the biprism and Lloyd's mirror fringes.

1) In biprism, the complete pattern of fringes is obtained. In Lloyd's mirror, ordinarily, only a few fringes on one side of the central fringe are visible, the central fringe itself being invisible.

2) In biprism the central fringe is bright, while in Lloyd's mirror it is dark.

3) The central fringe in biprism is less sharp than that in Lloyd's mirror.

The coherent sources in the biprism are $A_1 B_1$ and $A_2 B_2$ (Fig. 5.20a) the virtual images of a slit $AB$. In Lloyd's mirror, the coherent sources are a slit $A_1 B_1$ itself and its virtual image $B_2 A_2$ (Fig. 5.20 b). In both cases, $A_1$ and $A_2$ form one extreme pair of coherent point-sources, and $B_1$ and $B_2$ another extreme pair. In the biprism, the zero-order fringes corresponding to $A_1 A_2$ and $B_1 B_2$ are formed at $A_0$ and $B_0$, which lie on the right bisectors of $A_1 A_2$ and $B_1 B_2$, respectively. Hence, the zero-order fringe extends from $A_0$ to $B_0$. In Lloyd's mirror, on the other hand, all pair of coherent sources have a common perpendicular bisector, so that zero-order fringes due to all of these are formed in one and the same position. Hence the zero-order fringe is sharp in this case.

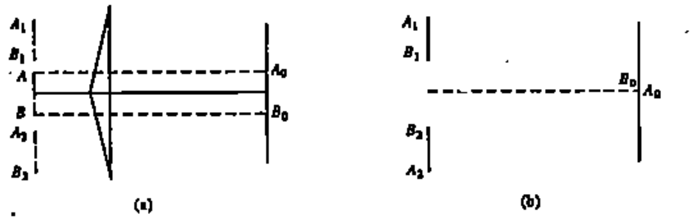


Fig. 5.20 : Showing the difference between biprism and Lloyd's mirror fringes.

4) In biprism $A_1 A_2 = B_1 B_2 = d$. Hence, the fringe-width $\beta = \frac{D\lambda}{d}$ is the same for all pairs of coherent sources. In Lloyd's mirror arrangement $d$ is different for different pairs of coherent sources, e.g., $A_1 A_2 > B_1 B_2$. Hence, the fringe-width is different for different pairs of coherent sources.

## 5.6 SUMMARY

- The relationship between phase difference and path difference is:

$$\text{phase difference} = \frac{2\pi}{\lambda}\ \text{(path difference)}$$

- If two waves of same frequency and of amplitudes $a_1$ and $a_2$ and phase difference $\delta$ are superposed then, according to principle of superposition, the amplitude $A$ of the resultant wave is given by

$$A^2 = a_1^2 + a_2^2 + 2a_1 a_2 \cos\delta$$

- Two sources are said to be coherent if they emit light waves with no or constant phase difference.

- When two waves of the same frequency travel in approximately the same direction and have a phase difference that remains constant with time, the resultant intensity of light is not distributed uniformly in space. This non-uniform distribution of the light intensity is due to the phenomenon of interference.

- For constructive interference

$$\text{path difference} = n\lambda,\ \text{where } n = 0, 1, 2, \ldots$$

and for destructive interference

$$\text{path difference} = m\frac{\lambda}{2},\ \text{where } m = 1, 3, 5, 7$$

- In an interference pattern, the distance between any two consecutive maxima or minima is given by

$$\beta = \frac{D\lambda}{d}$$

where $\beta$ is called the fringe-width, $\lambda$ is the wavelength of light used, $d$ is the distance between the two coherent sources, and $D$ is the distance between the sources and the screen.

- When a thin transparent plate of thickness $t$ and refractive index $\mu$ is introduced in the path of one of the constituent interfering beams of light, the entire fringe system is displaced through a distance $\frac{D}{d}(\mu - 1)\,t$.
- Just as in Young's double slit experiment, the wavelength of light can be determined from measurement of fringe-width produced by the biprism by the following relation:

  $$\lambda = \frac{\beta d}{D}$$

  where $d = \sqrt{d_1 d_2}$ and $D = b_1 + b_2$.

  $d_1$ is the distance between the two images, when the lens is at the position $L_1$ at a distance $b_1$ from the eyepiece. $d_2$ and $b_2$ are the corresponding distances when the lens is at $L_2$.
- Some other devices for producing coherent sources are : Fresnel's two mirror arrangement and Lloyd's mirror.
- Lloyd's mirror produces achromatic fringes.

## 5.7 TERMINAL QUESTIONS

1) Young's experiments is performed with light of the green mercury line. If the fringes are measured with a micrometer eye-piece 80 cm behind the double slit, it is found that 20 of them occupy a distance of 10.92 mm. Find the distance between two slits. Given that the wavelength of green mercury line is 5460 Å.

2) In a certain Young's experiment, the slits are 0.2 mm apart. An interference pattern is observed on a screen 0.5m away. The wavelength of light is 5000 Å. Calculate the distance between the central maxima and the third minima on the screen.

3) A Lloyd's mirror, of length 5 cm, is illuminated with monochromatic light ($\lambda$ = 5460 Å) from a narrow slit 0.1 cm from its plane, and 5 cm, measured in that plane, from its near edge. Find the separation of the fringes at a distance of 120 cm from the slit, and the total width of the pattern observed.

## 5.8 SOLUTIONS AND ANSWERS

SAQs

1) The distance $OM$ is given by $a \cos\theta$. Hence the equation is $x = a\cos\theta$ or $x = a\cos\omega t$.

2) $y = a\sin\omega t = a\sin\frac{2\pi}{T}t$

If we replace $\pi$ by 180°, and put $a$ = 6 cm = 0.06 m, and $T$ = 3 s, we get

$$y = (0.06)\sin\frac{2\times 180^\circ}{3}t$$

a) Thus displacement after 0.5 sec is,

$$y = 0.06\sin\frac{2\times 180^\circ}{3}\times 0.5$$
$$= 0.06\sin 60^\circ$$
$$= 0.052\text{ m}$$

b) Velocity, $v = a\omega\cos\omega t$

$$= a\frac{2\pi}{T}\cos\frac{2\pi}{T}t$$
$$= 0.06\times\frac{2\pi}{3}\cos\frac{2\times 180^\circ}{3}\times 0.5$$
$$= 0.06\times\frac{2\pi}{3}\times\cos 60^\circ$$
$$= 0.063\text{ ms}^{-1}.$$

c) Acceleration, $= \omega^2 y = \left(\frac{2\pi}{T}\right)^2 a \sin \omega t$

$$= \left(\frac{2\pi}{T}\right)^2 \times 0.06 \times \sin \frac{2\pi}{T} t$$

$$= \left(\frac{2\pi}{3}\right)^2 \times 0.06 \times \sin \frac{2 \times 180^\circ}{3} \times 0.5$$

$$= 0.228 \text{ ms}^{-2}.$$

3) We have

$$\frac{I_{\max}}{I_{\min}} = \frac{I_1 + I_2 + 2\sqrt{I_1 I_2}}{I_1 + I_2 - 2\sqrt{I_1 I_2}} = \frac{\left(\sqrt{I_1} + \sqrt{I_2}\right)^2}{\left(\sqrt{I_1} - \sqrt{I_2}\right)^2}$$

Now $\quad \frac{I_1}{I_2} = \frac{81}{1}$ or $\frac{\sqrt{I_1}}{\sqrt{I_2}} = \frac{9}{1}$ or $\sqrt{I_1} = 9\sqrt{I_2}$

Hence $\quad I_{\max}/I_{\min} = \frac{\left(9\sqrt{I_2} + \sqrt{I_2}\right)^2}{\left(9\sqrt{I_2} - \sqrt{I_2}\right)^2} = \frac{(10)^2 I_2}{(8)^2 I_2} = \frac{100}{64} = \frac{25}{16}$

4) The phase difference is related to the path difference by Eq. (5.13) as follows :

$$\text{phase difference} = \frac{2\pi}{\lambda} (AC - BC)$$

$$= \frac{2\pi}{\lambda} (4 - 3) \ \lambda$$

$$= 2\pi$$

This is the condition of maximum intensity. So the waves interfere, constructively, in Fig. 5.7(a).

In case of Fig. 5.7(b)

$$\text{phase difference} = \frac{2\pi}{\lambda} (AC - BC)$$

$$= \frac{2\pi}{\lambda} (3 - 2.5) \ \lambda$$

$$= \pi$$

This is the condition of minimum intensity.

Here the waves are completely out of phase and destructive interference occurs.

5) Given : $d = 0.40$ mm, $D = 10^3$ mm, $y = 3.6$ mm, and $m = 3$. Using Eq. (5.21), we get

$$\lambda = \frac{yd}{mD} = \frac{(3.6)(0.40)}{3 \times 10^3} = 4.8 \times 10^{-4} \text{ mm} = 4.8 \times 10^{-5} \text{ cm}$$

Hence, the light is in the blue-green region of the visible spectrum.

6) With $\lambda = 6000$ Å, the distance between zero-order and tenth order fringe is 14.73 mm – 12.34 mm = 2.39mm, so that the fringe width is 2.39 mm/10 = 0.239 mm.

7) $\beta = \frac{D\lambda}{d}$. Therefore

$$\frac{(\beta)_{6000}}{(\beta)_{5000}} = \frac{6000 \text{ Å}}{5000 \text{ Å}} = \frac{6}{5}$$

$\therefore \quad (\beta)_{5000} = \frac{5}{6} \times (\beta)_{6000} = \frac{5}{6} \times 0.239 = 0.199$ mm

Thus, with $\lambda = 5000$ Å, the zero-order fringe will still be at 12.34 mm, while the twentieth order fringe will be at

$$12.34 \text{ mm} + (0.199 \text{ mm} \times 20) = 16.32 \text{ mm}$$

8) For maxima, the path difference = $n\lambda$

or $\quad 30 \times 10^{-5}$ cm = $n\lambda$

$$\therefore \qquad \lambda = \frac{30 \times 10^{-5}}{n} \text{ cm}$$

where $n = 1, 2, 3, 4$ .........

9) $y_0 = 0.2$ cm; $d = 0.1$ cm; $D = 50$ cm

$$\text{Hence} \quad t = \frac{d\,y_0}{D\,(\mu - 1)} = \frac{0.1 \times 0.2}{50 \times 0.58}$$
$$= 6.7 \times 10^{-4} \text{ cm}$$

10) The fringe-width is given by

$$\beta = \frac{D\lambda}{2d}, \text{ where } d = \sqrt{d_1 \times d_2}$$

Here $d_1 = 4.05$ mm = 0.405 cm and $d_2 = 2.10$ mm = 0.210 cm.

$$\therefore \qquad d = \sqrt{0.405 \times 0.210} = 0.292 \text{ cm}$$

Also $D = 100$ cm and $\lambda = 5893$ Å = $5893 \times 10^{-8}$ cm.

$$\therefore \qquad \beta = \frac{100 \times 5893 \times 10^{-8}}{0.292} = 0.0202 \text{ cm}$$

11) By introducing a glass plate of thickness $t$ in one of the interfering beams, $t$ cm of air ($\mu = 1$) are replaced by $t$ cm of glass ($\mu = 1.5$). $t$ cm of glass are optically equivalent to $\mu t$ or 1.5 $t$ cm of air. The increase in the length of the path = $\mu t - t$ = $0.5t$. This produces a shift of 2 in the interference bands

$$\therefore \qquad 0.5t = 2\lambda = 2 \times 5.45 \times 10^{-5}$$

$$\text{and} \qquad t = \frac{2 \times 5.45 \times 10^{-5}}{0.5} = 21.8 \times 10^{-5} \text{ cm.}$$

**TQs**

1) The fringe width $\beta$ in Young's experiment is $\beta = \lambda D/d$

Since 20 fringes occupy a distance of 10.92 mm, the fringe width $\beta$ is

$$\beta = (10.92/20) \text{ mm} = (10.92 \times 10^{-3}/20) \text{ m}$$

Also $\quad D = 80$ cm = 0.8 m, and $\lambda = 5.460 \times 10^{-7}$ m

$$\therefore \qquad d = \frac{5.460 \times 10^{-7} \times 0.8 \times 20}{10.92 \times 10^{-3}} \text{ m} = 0.7912 \times 10^{-4} \text{ m}$$
$$= 0.07912 \text{ mm}$$

2) See Fig. (5.10). Suppose the required distance on the screen is $y$.

Here $\quad d = 2 \times 10^{-4}$ m (slit separation)

$\lambda = 5 \times 10^{-7}$ m (wave length)

$D = 5 \times 10^{-1}$ m (distance between slit to screen)

The minima is observed when the phase difference between the two waves is an odd multiple of $\pi$, i.e., when

$\delta = \pi, 3\pi, 5\pi, 7\pi, \ldots\ldots$

At the third minimum, $\delta = 5\pi$.

From Eq. (5.13), path difference $= \frac{2\pi}{\lambda}\,\delta = \frac{2\pi}{\lambda}\,(5\pi)$

But from Fig. 5.10, the path difference between the waves arriving at $P$ is $d \sin \theta$.

Hence $\quad 5\pi = \frac{2\pi}{\lambda}\,(d \sin \theta)$

or $\quad \sin \theta = \frac{\lambda}{2\pi}\,\frac{1}{d}\,(5\pi) = \frac{5\pi \times 5 \times 10^{-7}}{2\pi \times 2 \times 10^{-4}}$

$= 6.25 \times 10^{-3}$

From Fig. 5.10, the required distance on the screen $y = D \tan \theta$

$= D \sin \theta = 5 \times 10^{-1} \times 6.25 \times 10^{-3} \quad \because \tan \theta \approx \sin \theta$

$= 3.1$ m

3) Let $MM'$, (Fig. 5.21) the Lloyd's mirror be 5 cm long. The source $S_1$ is as shown in the figure. The interference pattern is observed in the region $AB$.

The fringe width $\beta$ is given by $\beta = \lambda\, D/d$

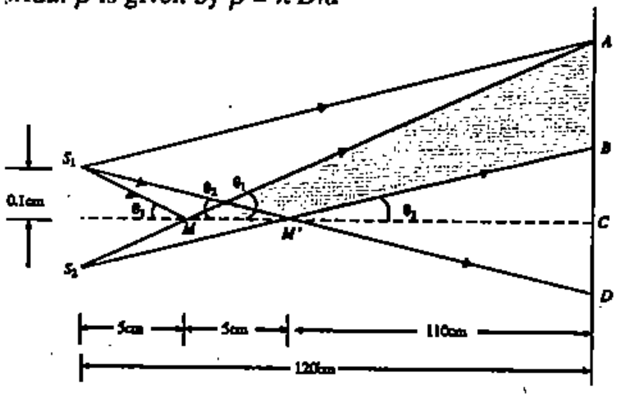


Fig. 5.21.

Given $\lambda = 5460$ Å $= 5.460 \times 10^{-7}$ m; $D = 120$ cm $= 1.20$ m and

$d = 0.2$ cm $= 2 \times 10^{-2}$ m

$$\therefore \qquad \beta = \frac{5.460 \times 10^{-7} \times 120}{2 \times 10^{-3}}\ \text{m} = 3276 \times 10^{-4}\,\text{m}$$

$$= 0.3276 \text{ mm.}$$

The total width of interference pattern is obviously $AB$. From Fig. (5.21),

$\tan \theta_1 = 0.1/5$, and $\tan \theta_2 = 0.1/10$

Also from rt angled $\Delta AMC$

$AC/MC = \tan \theta_1$ or $AC = 115 \times \tan \theta_1$

$= 115 \times 0.1/5 = 2.3$ cm

From rt angled $\Delta BM'C$

$BC/M'C = \tan \theta_2 = \tan \theta_2$ or $BC = 110 \times (0.1/10) = 1$ cm

$\therefore \quad AB = AC - BC = 2.3 - 1.1 = 1.2$ cm $= 1.2 \times 10^{-3}$ cm

# UNIT 6 INTERFERENCE BY DIVISION OF AMPLITUDE

## Structure

## 6.1 INTRODUCTION

We have all seen the marvellous rainbow colours that appear in soap bubbles and thin oil films. When a soapy plate drains, coloured reflections often occur from it. A similar effect occurs when light is reflected from wet pavements that has an oil slick on it. Have you ever wondered what causes the display of colours when light is reflected from such thin oil film or soap bubble?

All these effects are due to interference of light reflected from the opposite surfaces of the film. Thus the phenomenon owe its origin to a combination of reflection and interference.

In the last unit, we discussed the interference of light, but there, the two interfering light waves are produced by division of wavefront. For example, in Young's double slit experiment, light coming out of a pin hole was allowed to fall into two holes, and the light waves emanating from these two holes interfered to produce the interference pattern. But the interference of light waves, which is responsible for the colour of thin films, involves two light beams derived from a single incident beam by division of amplitude of the incident wave. When a light wave falls on a thin film, the wave reflected from the upper surface interferes with the wave reflected from the lower surface. This gives rise to beautiful colours. However, one must initially consider how the phase of a light wave is affected when it is reflected.

In the last unit, you noted that in Lloyd's mirror, the interference takes place between waves coming direct from the source and those reflected from an optically denser medium. As a consequence of this, the central fringe is found to be 'dark' instead of 'bright'. This was explained by assuming the fact that a phase change of $\pi$ takes place when light waves are reflected at the surface of a "denser" medium. We will begin this unit by giving proof of the statement made above; this proof will be based on the principle of reversibility of light.

It is also possible to observe interference using multiple beams. This is known as multiple beam interferometry, and it will be discussed in the next unit. It will be shown there that multiple beam interferometry offers some unique advantages over two beam interferometry.

### Objectives

After studying this unit, you should be able to

- prove that when a light wave is reflected at the surface of an optically denser medium, it suffers a phase change of $\pi$.

- describe the origin of the interference pattern produced by a thin film,
- describe the formation, shape and location of interference fringes obtained from a thin wedge-shaped film,
- describe how Newton's rings are used to determine the wavelength of light,
- explain why a thin coating of a suitable substance minimizes the reflection of light from a glass surface,.
- distinguish between fringes of equal inclination and fringes of equal thickness.

## 6.2 STOKES' ANALYSIS OF PHASE CHANGE ON REFLECTION

To investigate the phase change in the reflection of light at an interface between two media, Sir G.C. Stokes used the principle of optical reversibility. This principle states that a light ray, that is reflected or refracted, will retrace its original path, if its direction is reversed, provided there is no absorption of light.

Fig. 6.1(a) shows the surface *MN* separating media 1 and 2, the lower one being denser. Suppose medium 1 is air and medium 2 is glass.

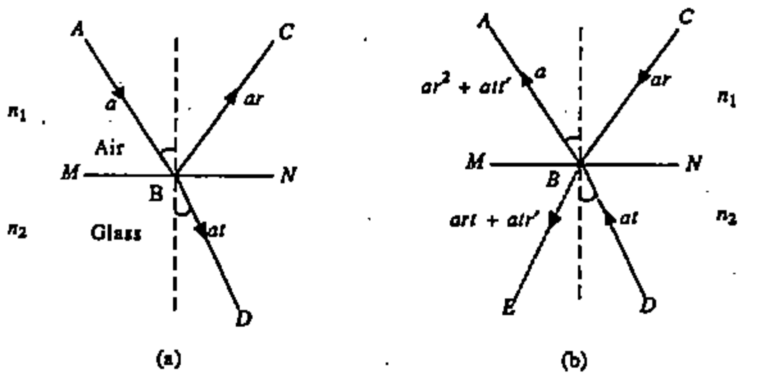


Fig. 6.1: (a) A ray is reflected and refracted at an air-glass interface. (b) The optically reversed situation; the two rays in the lower left must cancel. In both cases, $n_2 > n_1$ ($n_1$ and $n_2$ are the refractive indices of the media).

An incident light wave, *AB*, is partly reflected along *BC* and partly transmitted (refracted) along *BD*. Let *a* be the amplitude of the incident wave *AB*, *r* be the fraction of the amplitude reflected, and *t* be the fraction transmitted when the wave is travelling from medium 1 to 2. Then the amplitudes along *BC* and *BD* are *ar* and *at*, respectively..

Now, suppose the directions of the reflected and transmitted (refracted) waves are reversed. As shown in Fig. 6.1(b), the wave *BC*, on reversal, gives a reflected wave along *BA*, and a transmitted (refracted) wave along *BE*. The amplitude of reflected wave along *BA* is $ar.r = ar^2$ and the amplitude of transmitted wave along *BE* is *art*. Similarly, the wave *BD*, on reversal, gives a transmitted wave along *BA* and a reflected beam along *BE*. Let $r'$ and $t'$ be the fractions of amplitude reflected and transmitted when the wave is travelling from medium 2 to medium 1. Then the amplitude of the transmitted wave along *BA* is $att'$ and the amplitude of reflected wave along *BE* is $atr'$. But, according to principle of reversibility of light, the reflected and transmitted waves *BC* and *BD*, when reversed, should give the original ray of amplitude *a* along *BA* only. Hence, the component along *BE* should be zero and that along *BA* should be equal to *a*. That is

$$art + atr' = 0 \qquad ...(6.1)$$

and

$$ar^2 + att' = a \qquad ...(6.2)$$

From Eqs. (6.1) and (6.2), we get

$$r' = -r \qquad ...(6.3)$$

and

$$tt' = 1 - r^2 \qquad ...(6.4)$$

Eqs. (6.3) and (6.4) are known as **Stoke's-relations.**

You must be aware that a transverse wave in a spring undergoes a 180° phase change when reflected from a rigid support. A similar phase change occurs for the reflection of a light wave from the boundary of a medium, having a greater index of refraction. The optically denser medium corresponds to a rigid support. A light wave reflected from the boundary of a medium whose index of refraction is greater than that of the medium in which the incident wave travels undergoes a 180° phase change.

Now, observe carefully Eq. (6.3). Here $r$ is the fraction of amplitude reflected when incident wave is travelling from a rarer to denser medium, and $r'$ when incident wave is travelling from a denser to a rarer medium. The two fractions are numerically equal but have opposite signs. Hence, these are exactly out of phase with each other, i.e., their phase difference is '$\pi$'. If no phase change occurs when a light wave is reflected by a denser medium then there must be a phase change of $\pi$ when a light wave is reflected by a rarer medium—and conversely, if no phase change occurs when a light wave is reflected by a rarer medium then there must be a phase change of $\pi$ when a light wave is reflected by a denser medium. Now, out of the two alternatives mentioned above second one is correct because it has been experimentally observed (See sec 5.6 in connection with Lloyd's mirror) that the phase change of $\pi$ occurs when the light strikes the boundary from the side of rarer medium. Hence, light reflected by a material of higher refractive index than the medium in which the rays are travelling undergoes a 180° (or $\pi$) phase change.

Reflection by a material of lower refractive index than the medium in which the rays are travelling causes no phase change.

The following SAQ will provide a useful check of your understanding of this section.

SAQ 1

In Fig. 6.2, we have illustrated four situations. In the two examples on the left, the refractive index between the surfaces is higher than that outside; in the two examples on the right, it is lower. This determines whether or not there is a phase change. In Fig. 6.2(a) and (b), we have indicated the phase change taking place at the points marked by an arrow. Redraw the Fig. 6.2(c) and (d), indicating the phase change taking place at the points marked by an arrow.

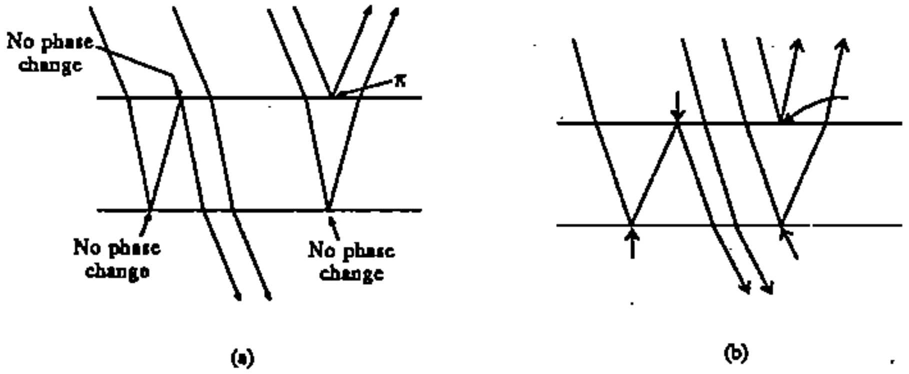


Fig. 6.2

## 6.3 INTERFERENCE IN THIN FILMS

Suppose a ray of light from a source $S$ strikes a thin film of soapy water, at $A$, see Fig. 6.3(a) . Part of this will be reflected as ray (1) and part refracted in the direction $AB$. Upon arrival at $B$, part of the latter will be reflected to $C$, and part refracted along $BT_1$. At $C$, the ray will again get partly reflected along $CD$ and refracted as ray (2) along $CR_2$. A continuation of this process yields two sets of parallel rays, one on each side of the film. In each of these sets, of course, the amplitude decreases rapidly from one ray to the next. Considering only the first two reflected rays (1) and (2) we find that these two rays are in a position to interfere. This is because, if we assume $S$ to be a monochromatic point source, the film serves as an amplitude-splitting device, so that ray (1) and (2) may be considered as arising from two coherent virtual sources $S'$ and $S''$ lying behind the film, that is, the two images of $S$ formed by reflection at the top and bottom surfaces of the film, as shown in Fig. 6.3 (b). If the set of parallel reflected rays is now collected by a lens, and focussed at $P$, each ray has travelled a different distance, and the phase relationship between them may be such as to produce destructive or constructive interference at $P$. It is such interference that produces the colours of this film when seen by naked eyes.

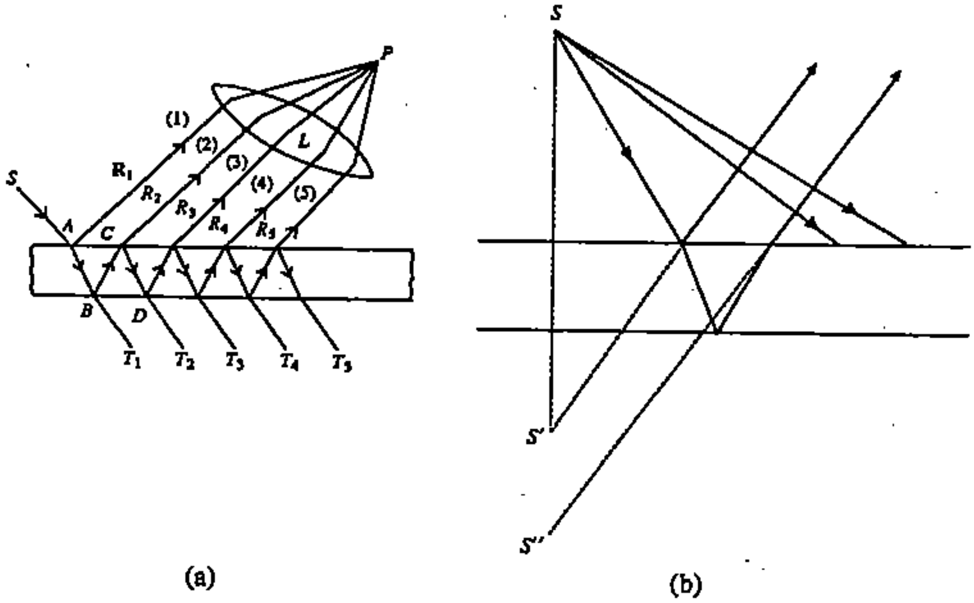


Fig. 6.3: (a) Multiple reflection in a soap film. (b) The interference pattern produced due to rays (1) and (2) is approximately the same as would have been produced by two coherent point sources $S'$ and $S''$.

Now, we know that the two rays reinforce each other, if the path difference between them is an integral multiple of $\lambda$, where $\lambda$ is the wavelength of light, which is being used to illuminate the film. Hence, let us first find out the path difference between the reflected rays (1) and (2).

### Path Difference in Reflected Light

Suppose the ray of light falling on the thin film of soapy water at $A$ be incident at an angle $i$, as shown in Fig. 6.4. Let the thickness of the film be $t$ and refractive index be $\mu$ (>1). At $A$ it is partly reflected along $AR_1$ giving the ray (1) and partly refracted along $AB$ at an angle $r$. At $B$ it is again partly reflected along $BC$ and partly refracted along $BT_1$. Similar reflections and refractions occur at $C$. Since, the rays $AR_1$ and $CR_2$ i.e. ray (1) and ray (2) have been derived from the same incident ray, they are coherent and in a position to interfere. Let $CN$ and $BM$ be perpendiculars to $AR_1$ and $AC$. As the paths of the rays $AR_1$ and $CR_2$ beyond $CN$ are equal, the path difference between ray (1) and (2) is given by

(path $ABC$ in film-path $AN$ in air)

$$\therefore \text{ path difference} = \mu\,(AB + BC) - AN \qquad ...(6.5)$$

Here $$AB = BC = \frac{BM}{\cos r} = \frac{t}{\cos r},$$

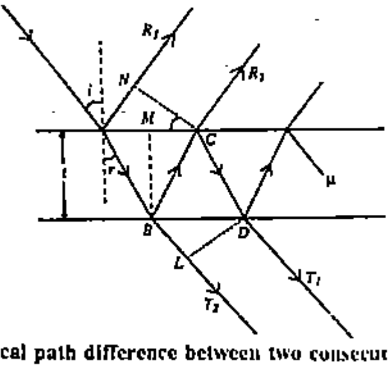


Fig. 6.4: Optical path difference between two consecutive rays in a multiple reflection.

$$AN = AC \sin i \quad \text{and}$$

Now,

$$\begin{aligned} AC &= AM + MC \\ &= BM \tan r + BM \tan r \\ &= 2t \tan r \end{aligned}$$

$$\begin{aligned} \therefore \quad AN &= 2t \tan r \sin i \\ &= 2t \frac{\sin r}{\cos r} (\sin i) \\ &= 2t \frac{\sin r}{\cos r} (\mu \sin r) \qquad \left[\because \frac{\sin i}{\sin r} = \mu\right] \\ &= 2\mu t \frac{\sin^2 r}{\cos r} \end{aligned}$$

Substituting these values of *AB*, *BC* and *AN* in Eq. (6.5) we get,

$$\text{path difference} = \mu\left(\frac{t}{\cos r} + \frac{t}{\cos r}\right) - 2\mu t \frac{\sin^2 r}{\cos r}$$

$$= \frac{2\mu t}{\cos r}(1 - \sin^2 r)$$

$$\therefore \quad \text{path difference} = 2\mu t \cos r \qquad \text{...(6.6)}$$

At *A*, the ray is reflected while going from a rarer to a denser medium and suffers a phase change of $\pi$. At *B*, the reflection takes place when the ray is going from a denser to a rarer medium, and there is no phase change.

However, we must take account of the fact that ray (1) undergoes a phase change of $\pi$ at reflection while ray (2) does not, since it is internally reflected (See SAQ 1). The phase change of $\pi$ is equivalent to a path difference of $\frac{\lambda}{2}$. Hence, the effective path difference between ray (1) and rays (2) is

$$2\mu t \cos r - \frac{\lambda}{2} \qquad \text{...(6.7)}$$

The sign of the phase cnange is immaterial. Here we have chosen the negative sign to make the equation a bit simpler in form.

As you know from Unit 5, if this path ainerence is an odd multiple of $\frac{\lambda}{2}$, we might expect rays (1) and (2) to be out of phase, and produce a minimum of intensity. Thus the condition

$$2\mu t \cos r - \frac{\lambda}{2} = (2n - 1)\frac{\lambda}{2}, \text{ where } n = 1, 2, \ldots$$

or

$$2\mu t \cos r = n\lambda \qquad \text{...(6.8)}$$

becomes a **condition for destructive interference** as far as rays (1) and (2) are concerned.

Next, we examine the phases of the remaining rays, (3), (4), (5),..... Since the geometry is the same, the path difference between rays (3) and (2) will also be given by Eq. (6.6). But, here, only internal reflections are involved, so the effective path difference will still be given by Eq. (6.6). Hence, if the condition given by Eq. (6.8) is fulfilled, ray (3) will be in the same phase as ray (2). The same holds true for all succeeding pairs, and so we conclude that, under the condition given by Eq. (6.8), rays (1) and (2) will be out of phase, but rays (2), (3), (4),....., will be in phase with each other. Now, since ray (1) has considerably greater amplitude than ray (2), we might think that they will not completely annul each other, that is, the condition given by Eq. (6.8) may not produce complete darkness. But it is not so. We will now prove that the addition of rays (3), (4), (5)....., which are all in phase with ray (2), will give a net amplitude, just sufficient to make up the difference and to produce complete darkness. Fig. 6.5 shows the amplitude of successive rays in multiple reflection.

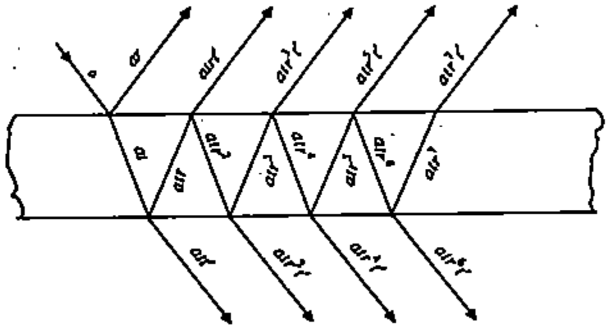


Fig. 6.5: Amplitude of successive rays in multiple reflection.

Adding the amplitudes of all the reflected rays but the first, on the upper side of the film we obtain the resultant amplitude:

$$\begin{aligned} A &= atrt' + atr^3t' + atr^5t' + atr^7t' + \ldots \\ &= atrt'(1 + r^2 + r^4 + r^6 + \ldots\ldots) \end{aligned}$$

Since $r$ is, necessarily, less than 1, the geometrical series in parentheses has a finite sum equal to $1/(1-r^2)$, giving

$$A = atrt' \frac{1}{(1-r^2)}$$

But from Stoke's treatment, Eq. (6.4), $tt' = 1 - r^2$, we obtain

$$A = ar \qquad \ldots(6.9)$$

This is just equal to the amplitude of the first reflected ray, hence, we conclude that under the condition of Eq. (6.8), there will be complete destructive interference. On the other hand, if the path difference given by Eq. (6.7) is an integral multiple of $\lambda$, i.e., when

$$2\mu t \cos r - \frac{\lambda}{2} = n\lambda, \text{ where } n = 0, 1, 2, \ldots \text{ etc.}$$

or

$$2\mu t \cos r = (2n+1)\frac{\lambda}{2} \qquad \ldots(6.10)$$

then ray (1) and (2) will be in phase with each other and gives a condition of **constructive interference**. But rays (3), (5), (7),.... will be out of phase with rays (2), (4), (6),.... Since (2) is more intense than (3), (4) is more intense than (5), etc., these pairs cannot cancel each other. As the stronger series combines with ray (1), the strongest of all, there will be maximum of intensity.

Thus, when a thin film is illuminated by monochromatic light, and seen in reflected light, it appears bright or dark according as $2\mu t \cos r$ is odd multiple of $\frac{\lambda}{2}$ or integral multiple of $\lambda$, respectively.

$$2\mu\, t \cos r = (2n+1)\frac{\lambda}{2} \quad \text{(condition of maxima)} \qquad \ldots(6.11a)$$

$$2\mu\, t \cos r = n\lambda \quad \text{(condition of minima)} \qquad \ldots(6.11b)$$

Before moving further, answer the following SAQ.

**SAQ 2**

Using Eq. (6.7), state whether the following statement is true or false. Give reasons.

**"An excessively thin film seen in reflected light appears perfectly black".**

Now we are in a position to know the reason of the production of colours in thin film of soap water.

### Colours in Thin Films

The eye looking at the film receives rays of light reflected at the top and bottom surfaces of the film. These rays are in a position to interfere. The path difference between the interfering rays, given by Eq. (6.7), depends upon $t$ (thickness of the film) and upon $r$, and, hence, upon inclination of the incident rays (the inclination is determined by the position of the eye relative to the region of the film, which is being looked at). The sunlight consists of a continuous range of wavelengths (colours). At a particular point of the film, and for a particular position of the eye (i.e., for a particular $t$ and a particular $r$), the rays of only certain wavelengths will have a path difference satisfying the condition of maxima. Hence, only these wavelengths (colours) will be present with the maximum intensity. While some others, which satisfy the condition of the minima will be missing. Hence, the point of the film being viewed will appear coloured.

We are working out an example so that the phenomenon of production of colours in thin film is clear to you.

---

**Example 1**

A thin film of $4 \times 10^{-5}$cm thickness is illuminated by white light normal to its surface ($r$ = 0°). Its refractive index is 1.5. Of what colour will the thin film appear in reflected light?

**Solution**

The condition for constructive interference of light reflected from a film is

$2\mu t \cos r = (2n + 1)\dfrac{\lambda}{2}$, where $n = 0,1,2,....$

Here $\mu = 1.5$; $t = 4 \times 10^{-5}$ cm and $r = 0°$ (since light falls normally) so that $\cos r = 1$.

$$\therefore \qquad 2 \times 1.5 \times 4 \times 10^{-5} = (2n + 1)\frac{\lambda}{2}$$

$$\text{or} \qquad \lambda = \frac{2 \times 2 \times 1.5 \times 4 \times 10^{-5}}{2n + 1}$$

$$\lambda = \frac{24 \times 10^{-5}\text{ cm}}{2n + 1} = \frac{24{,}000}{2n + 1}\text{Å}$$

Taking $n = 0, 1, 2, 3,$ ......... we get

$\lambda$ = 24000 Å, 8000 Å, 4800Å, 3431Å .........

These are the wavelengths reflected most strongly. Of these, the wavelength lying in the visible region is 4800Å (blue).

---

So far we have considered viewing of thin film in reflected light. Suppose the eye is now situated on the lower side of the film, shown in Fig. 6.3 and Fig. 6.5. The rays emerging from the lower side of the film can also be brought together with a lens and made to interfere.

Let us find out what colours will arise, when the film is viewed in this position. For this, we have to first calculate the path difference between the rays in transmitted light.

The path difference between the transmitted rays $BT_1$ and $DT_2$ is given by Eq. (6.6), i.e.,

$$(BC + CD) - BL = 2\mu t \cos r$$

In this case, there is no phase change due to reflection at $B$ or $C$, because in either case the light is travelling from denser to rarer medium (See SAQ 1). Hence, the effective path difference between $BT_1$ and $DT_2$ is also $2\mu\, t \cos r$.

The two rays $BT_1$ and $DT_2$ reinforce each other, if

$$2\ u\, t \cos r = n\lambda \text{ (condition of maxima)} \qquad ...(6.12a)$$

where $n = 1, 2, 3$.

In this case, the film will appear bright in the transmitted light.

The two rays will destroy each other if

$$2\ \mu\, t \cos r = (2n + 1)\ \frac{\lambda}{2} \text{ (condition of minima)} \qquad ..(6.12b)$$

where $n = 0, 1, 2,....$ and the film appears dark in transmitted light.

A comparison of Eqs. (6.11a), (6.11b), (6.12a) and (6.12b) shows that the conditions for the maxima and minima, in the reflected light are just the reverse of those in transmitted light. Therefore, only those colours will be visible in transmitted light, which were missed in reflected light. Hence, the film which appears bright in reflected light will appear dark in transmitted light and vice versa. In other words, the appearances of colours in the two cases is complimentary to each other.

Interference fringes produced by thin films can be classified into two: Fringes of equal inclination and fringes of equal thickness.

**Fringes of Equal Inclination**

If the lens used in Fig. 6.3 to focus the rays has a small aperture, interference fringes will appear on a small portion of the film. Only the rays leaving the point source that are reflected directly into the lens will be seen (see Fig. 6.6a). For an extended source, light will reach the lens from various directions, and the fringe pattern will spread out over a large area of the film, as shown in Fig. 6.6b.

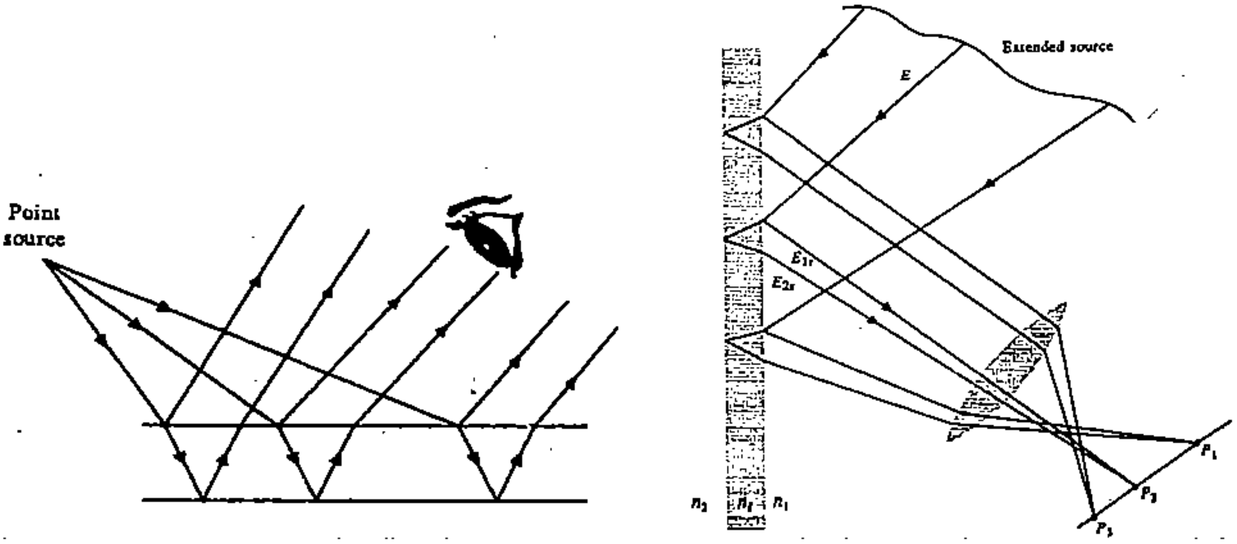


Fig.6.6: (a) Fringes seen in a small portion of the film. (b) Fringes seen on a large region of the film.

The angle $i$ or equivalently $r$, determined by the position $P$, will, in turn, control the path difference. The fringes appearing at points $P_1$ and $P_2$ in Fig. 6.7 are, accordingly, known as fringes of equal inclination.

Notice that as the film becomes thicker, the separation $AC$ in Fig. 6.4 between ray (1) and (2) also increases, since $AC = 2t \tan r$. When only one of the two rays is able to enter the pupil of the eye, the interference pattern will disappear. The larger lens of a telescope could then, be used to gather in both rays, making the pattern visible. The separation can also be reduced by reducing $r$, and, therefore, $i$, i.e., by viewing the film at nearly normal incidence.

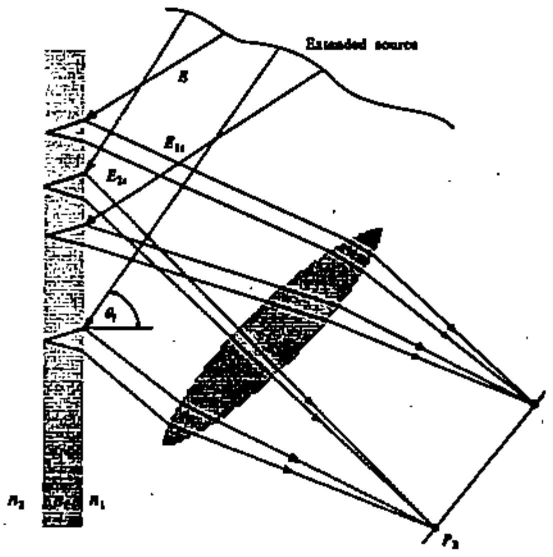


Fig. 6.7: All rays inclined at the same angle arrive at the same point.

The equal inclination fringes that are seen in this manner for thick plates are known as **Haidinger fringes.** With an extended source, the symmetry of the set up requires that the interference pattern consists of a series of concentric circular bands centered on the perpendicular drawn from the eye to the film, as shown in Fig. 6.8.

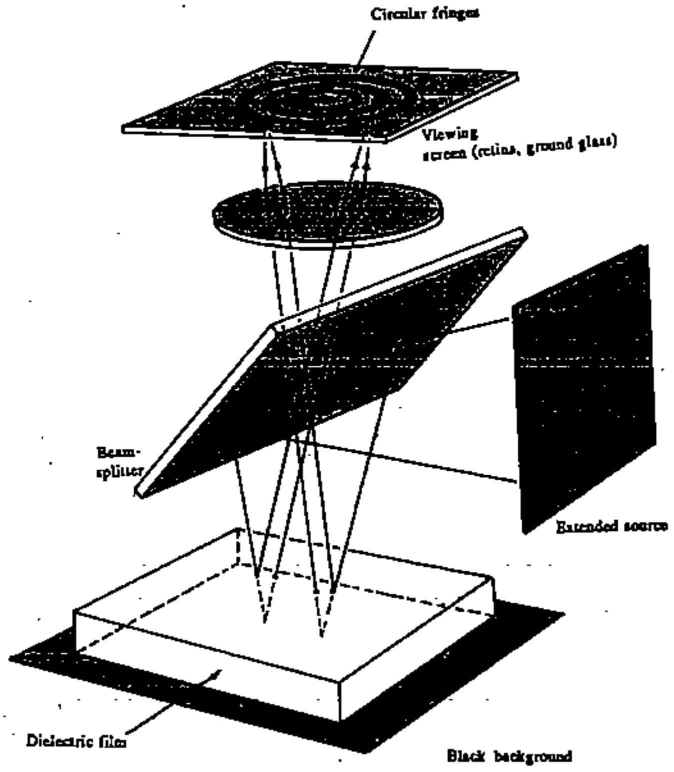


Fig.6.8: Circular Haidinger fringes centered on the lens axis.

Such fringes are formed at infinity, and are observed by a telescope focussed at infinity. These fringes are observed in Michelson interferometer, about which we will study in next unit.

### Fringes of Equal Thickness

Interference fringes, for which thickness $t$ is the dominant parameter rather than $r$, are referred to as **fringes of equal thickness.** Each fringe is the locus of all points in the film for which thickness is a constant. Such fringes are localised on the film itself, and are observed by a microscope focussed on the film. Fringes due to the wedge-shaped film belong to this class of fringes, which you will study in the next section.

Fringes of equal thickness can be distinguished from the circular pattern of Haidinger's fringes by the manner in which the diameters of the rings vary with order $n$. The central region in the Haidinger pattern corresponds to the maximum value of $n$, whereas just the opposite applies to fringes of equal inclination.

## 6.4 INTERFERENCE BY A WEDGE-SHAPED FILM

So far, we have assumed the film to be of uniform thickness. We will now discuss the interference pattern produced by a film of varying thickness, i.e., a film which is not plane-parallel. Such a film may be produced by a wedge, which consists of two non-parallel plane surfaces, as shown in Fig. 6.9a and 6.9b. Observe that the interfering rays do not enter the eye parallel to each other but appear to diverge from a point near the film.

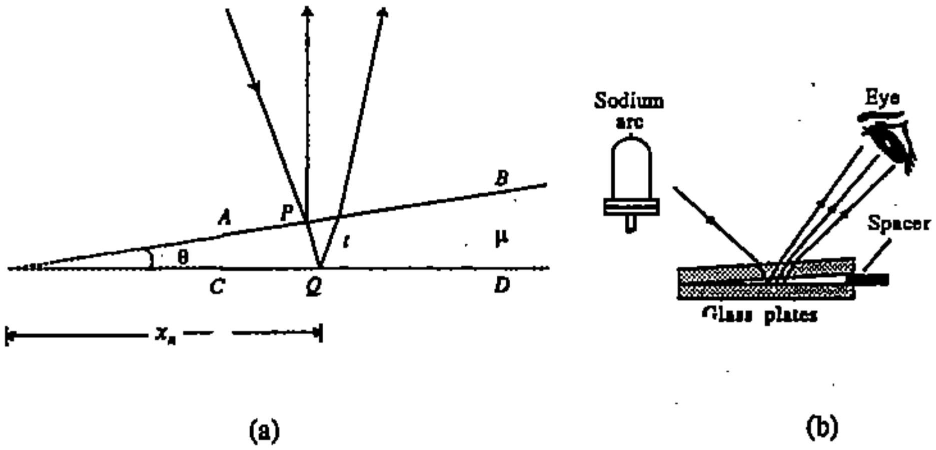


**Fig. 6.9: Fringes of equal thickness: (a) method of visual observations. (b) a parallel beam of light incident on a wedge.**

Let us consider a thin wedge-shaped film of refractive index $\mu$, bounded by two plane surfaces $AB$ and $CD$, inclined at an angle $\theta$ as shown in Fig. 6.9b. Let the film be illuminated by a monochromatic source of light from a slit held parallel to the edge of the wedge (the edge is the line passing through the point O and perpendicular to the plane of the paper). Interference occurs between the rays reflected at the upper and lower surfaces of the film. In this case the path difference for a given pair of rays is practically that given by Eq. (6.6). But, if it is assumed that light is incident almost normally at a point $P$ on the film, the factor cos $r$ may be considered equal to 1. Thus, the path difference between the rays reflected at the upper and lower surfaces is $2\mu t$, where $t$ is the thickness of the film at $P$. An additional path difference of $\frac{\lambda}{2}$ is introduced in the ray reflected from the upper surface. The effective path difference between the two rays is

$$2\mu t-\frac{\lambda}{2} \qquad \ldots(6.13)$$

Hence the condition for bright fringes becomes

$$2\mu t-\frac{\lambda}{2}=n\lambda$$

or $$2\mu t = (2n + 1)\frac{\lambda}{2} \qquad ...(6.14)$$

The condition for dark fringe is

$$2\mu t = n\lambda \qquad ...(6.15)$$

It is clear that for a bright or dark fringe of a particular order, $t$ must remain constant. Since in the case of a wedge-shaped film, $t$ remains constant along lines parallel to the thin edge of the wedge, the bright and dark fringes are straight lines parallel to the thin edge of the wedge. Such fringes are commonly referred to as "fringes of equal thickness". At the thin edge, where $t = 0$, path difference $= \frac{\lambda}{2}$, which is a condition for minimum intensity. Hence, the edge of the film is dark. The resulting fringes resemble the localized fringes in the Michelson interferometer (this you will study in next unit) and appear to be formed in the film itself.

**Spacing between Two Consecutive Bright (or Dark) Fringes**

For the $n$th dark fringe, we have

$$2\mu t = n\lambda$$

Let this fringe be obtained at a distance $x_n$ from the thin edge. Then $t = x_n \tan\theta = x_n\,\theta$ (when $\theta$ is small and measured in radians).

$$\therefore \qquad 2\mu\, x_n\theta = n\lambda \qquad ...(6.16)$$

Similarly, if the $(n + 1)$th dark fringe is obtained at a distance $x_{n+1}$ from the thin edge, then

$$2\mu\, x_{n+1}\,\theta = (n + 1)\,\lambda \qquad ...(6.17)$$

Subtracting Eq. (6.16) from Eq. (6.17), we get

$$2\mu\theta\,(x_{n+1} - x_n) = \lambda$$

Hence the fringe width $\beta$ is

$$\beta = x_{n+1} - x_n = \frac{\lambda}{2\mu\theta} \qquad ...(6.18)$$

where $\theta$ is measured in radians.

Similarly, it can be shown that the spacing between two consecutive bright fringes (fringe width) is $\frac{\lambda}{2\mu\theta}$.

**SAQ 3**

Using sodium light ($\lambda$ = 5893 Å), interference fringes are formed by reflection from a thin air wedge. When viewed perpendicularly, 10 fringes are observed in a distance of 1 cm. Calculate the angle of the wedge.

If the fringes of equal thickness are produced in the air film between a convex surface of a long-focus lens and a plane glass surface, the fringes will be circular in shape because the thickness of the air film remains constant on the circumference of a circle. The ring-shaped fringes, thus produced, were studied by Newton. In the next section, we will study Newton's ring.

## 6.5 NEWTON'S RINGS

When a plano-convex lens of large radius of curvature is placed with its convex surface in contact with a plane glass plate, air-film is formed between the lower surface of the lens (*LOL'*) and the upper surface of the plate (*POQ*), as shown in Fig. 6.10. The thickness of the air film is zero at the point of contact $O$, and it increases as one moves away from the point of contact. If monochromatic light is allowed to fall normally on this film, reflection takes place at both the top and bottom of the film. As a result of interference between the light waves reflected from the upper and lower surfaces of the air film, constructive or destructive interference takes place, depending upon the thickness of the film. The thickness of the air film increases with distance from the point of contact, therefore, the pattern of bright and dark fringe consists of concentric circles. In Fig. 6.10, 1 and 2, are the interfering rays corresponding to an incident ray $AB$. As the rings are observed in reflected light, the effective path difference between the interfering rays 1 and 2 is practically that given by Eq. (6.13).

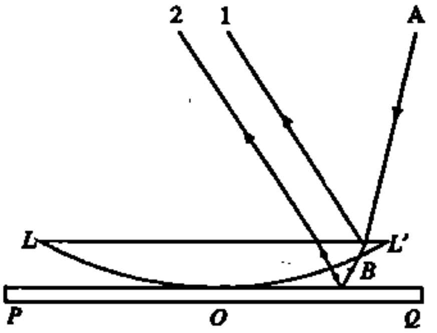


Fig.6.10: An arrangement for observing Newton's rings.

As we have considered an air-film, $\mu = 1$. The condition for the bright ring which is given by Eq. (6.14), is

$$2t = (2n-1)\frac{\lambda}{2} \qquad \ldots(6.19)$$

and the condition for the dark ring which is given by Eq. (6.15) is

$$2t = n\lambda \qquad \ldots(6.20)$$

Let us find out the relationship between the radii of the rings and the wavelength of the light. Consider Fig. 6.11, where the lens $LOL'$ is placed on the glass plate $POQ$. Let $R$ be the radius of curvature of the curved surface of the lens. Let $r_n$ be the radius of the $n$th Newton's ring corresponding to point $P$, where the film thickness is $t$. Draw perpendicular $PN$. Then, from the property of a circle, we have

$$PN^2 = ON \times NE$$

or

$$r_n^2 = t\,(2R-t)$$

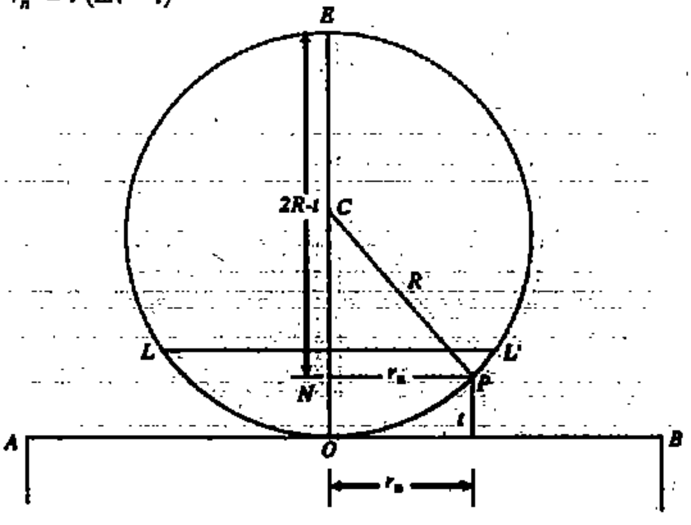


Fig. 6.11: $r_n$ represents the radius of the $n$th dark ring, the thickness of air film (where the $n$th dark ring is formed) is $t$.

Since $t$ is small compared to $R$, we can neglect $t^2$.

Hence, $$r_n^2 = 2Rt$$

or $$2t = \frac{r_n^2}{R} \qquad \text{...(6.21)}$$

The condition for a bright ring is

$$2t = (2n-1)\frac{\lambda}{2}$$

But from Eq. (6.21), $2t = \frac{r_n^2}{R}$

$$\therefore \qquad \frac{r_n^2}{R} = (2n-1)\frac{\lambda}{2}$$

or $$r_n^2 = (2n-1)\frac{\lambda R}{2} \text{ (Bright ring)}$$

If $D_n$ be the diameter of the $n$th bright ring, then $D_n = 2r_n$ or $r_n = \frac{D_n}{2}$. Substituting this in the last expression, we get

$$D_n^2 = 2\,(2n-1)\,\lambda R$$

or $$D_n = \sqrt{2\lambda\,R}\,\sqrt{2n-1}$$

or $$D_n \propto \sqrt{2n-1} \qquad (\lambda \text{ and } R \text{ being constant}) \qquad \text{...(6.22)}$$

This shows that the radii of the rings vary as the square-root of odd natural numbers. Thus the rings will be close to each other as the radius increases, as shown in Fig. 6.12.

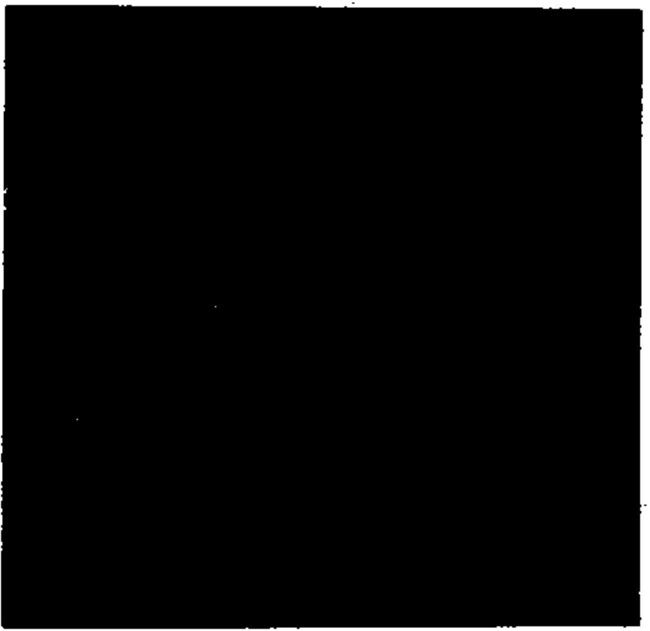


Fig. 6.12: Newton's rings as observed in reflected light.

Between the two bright rings there will be a dark ring whose radius will be proportional to the square-root of the natural numbers. Attempt the following SAQ and prove the above statement yourself.

### SAQ 4

Using Eqs. (6.20) and (6.21), prove that the radius of the dark ring is proportional to the square-root of the natural numbers.

The ring diameters depend on wavelength, therefore, the monochromatic light will produce an extensive fringe system such as that shown in Fig. 6.12.

When the contact between lens and glass is perfect, the central spot is black. **This is direct evidence of the relative phase change of $\pi$ between the two types of reflection, air-to-glass and glass-to-air, mentioned in Sec. 6.2. If there were no such phase change, the rays reflected from the two surfaces in contact should be in the same phase, and produce a bright spot at the centre.** However, the central spot can be made bright due to slight modification. In an interesting modification of the experiment, due to Thomas Young, if the lower plate is made to have a higher index of refraction than the lens, and the film in between is filled with an oil of intermediate index, then both reflections are at "rare-to-dense" surfaces. In this situation, no relative phase change occurs, and the central fringe of the reflected system is bright.

If $D_n$ is the diameter of the $n$th bright ring, then

$$D_n^2 = 2(2n - 1)\lambda R \qquad \text{...(6.23)}$$

If $D_{n+p}$ is the diameter of the $(n + p)$th bright ring, then

$$D_{n+p}^2 = 2\,[2(n + p) - 1]\lambda R \qquad \text{...(6.24)}$$

Subtracting Eq. (6.23) from Eq. (6.24), we get

$$D_{n+p}^2 - D_n^2 = 2\,[2(n + p) - 1]\lambda R - 2\,(2n - 1)\lambda R$$

$$= 4\,p\lambda R$$

$$\therefore \qquad \lambda = \frac{D_{n+p}^2 - D_n^2}{4pR} \qquad \text{...(6.25)}$$

It may be mentioned here, that the point of contact may not be perfect. As such the $n$th ring may not be the $n$th fringe but Eq. (6.25) is almost always valid. On measuring the diameters of the rings and the radius of curvature $R$, the wavelength $\lambda$ can be calculated with the help of the Eq. (6.25). In laboratory, the radius of curvature can be accurately measured with the help of a spherometer.

If a liquid of refractive index $\mu$ is introduced between the lens and the glass plate, then the expression for path difference between two interfering rays will also include $\mu$. Then the radii of the dark rings would be given by

$$r_n = \left(\frac{n\lambda R}{\mu}\right)^{1/2} \qquad \text{...(6.26)}$$

Thus, when a little water is introduced between the lens and the plate, the rings contract according to the relation

$$\boxed{\frac{\text{diameter of a ring in water-film}}{\text{diameter of the same ring in air-film}} = \frac{1}{\sqrt{\mu}}} \qquad \text{...(6.27)}$$

where $\mu$ is the refractive index of water..

A ring system is also observed in the light transmitted by Newton's ring plates. There are two differences in the reflected and transmitted systems of rings. (i) The rings observed in transmitted light are exactly complementary to those seen in the reflected light, so that the central spot is now bright. (ii) The rings in transmitted light are much poorer in contrast than those in reflected light.

Before moving to the next section, solve the following SAQ.

SAQ 5

If in a Newton's ring experiment, the air in the interspace is replaced by a liquid of refractive index 1.33, in what proportion would the diameters of the ring change?

## 6.6 APPLICATIONS OF THE PRINCIPLE OF INTERFERENCE IN THIN FILM

1. An important and simple application of the principle of interference within film is in the production of coated surfaces. To accomplish this, the glass lens is coated with the film of a transparent substance that has an index of refraction between the refraction indices for air and glass (See Fig. 6.13). The thickness of the film is one quarter of the wavelength of light in the film so that

$$t = \frac{\lambda}{4\mu_1}.$$

If we assume normal incidence, then the path difference between the light wave reflected from the upper surface of the film and the light wave reflected from the lower surface of the film is $2\mu_1 t = 2\mu_1 \times \frac{\lambda}{4\mu_1} = \frac{\lambda}{2}$. Both waves undergo a phase change of 180° as reflections at both surfaces are from "rare-to-dense". Thus, the two reflected waves are out of phase because of path difference and, therefore, these interfere destructively. Such a film is known as non-reflecting film, because it gives zero reflection. However, this does not mean that a non-reflecting film destroys light, but it merely redistributes light so that a decrease of reflection is accompanied by a corresponding increase of transmission.

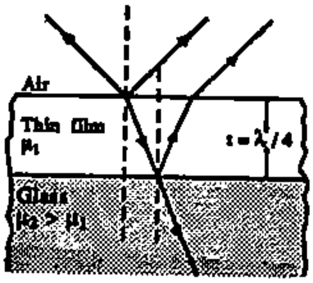


Fig. 6.13: A film coating on a glass lens makes the lens "non-reflecting" when the film thickness is $\lambda/4$ for normal incident. The total path difference of the reflected rays is then $\lambda/2$, and the waves interfere destructively, i.e., the incident light is totally transmitted.

The practical importance of these films is that by their use one can greatly reduce loss of light by reflection at the various surfaces of lenses or prisms used in binoculars, cameras, etc. Usually, glass is coated with a very thin layer of magnesium fluoride, the refractive index of which ($\mu = 1.38$) is intermediate between those of glass and air.

2. Another important application of thin film interference phenomenon is the converse of the procedure just discussed, viz., the glass surface is coated by a thin film of suitable material to increase the reflectivity. The film thickness is again $\lambda/4\mu_f$, where $\mu_f$ represents the refractive index of the film. The film is such that its refractive index is greater than that of the glass. This is because an abrupt phase change of $\pi$ occurs only at the air-film interface and the beams reflected from the air-film interface and the film-glass interface constructively interfere.

3. The fringes obtain by a wedge-shaped film has important practical applications in the testing of optical surfaces for flatness. An air-film is formed between a perfectly plane surface and the surface under test. If the latter surface is plane, the fringes will be straight and parallel, and, if not, these will be irregular in shape.

4. The accuracy of the grinding of a lens surface can be tested by observing the shape of Newton's rings formed between it and an accurately flat glass surface, using monochromatic light. If the rings are not perfectly circular, the grinding is imperfect.

You should be able to apply whatever you have learnt in this section to solve the following SAQ.

## 6.7 SUMMARY

- When the light wave is reflected from a boundary, there is an abrupt change of phase. When the light ray is reflected while going from a rarer to a denser medium, it suffers a phase change of $\pi$. But there is no phase change when the light ray is reflected while going from a denser to a rarer medium.

- Length $l$ in a medium of refractive index $\mu$ is optically equivalent to length $\mu l$ in a vacuum. $\mu l$ is called the optical path length of distance $l$ in the medium.

- For a thin film in reflected light, the conditions for constructive and destructive interference are:

$$2\,\mu\,t\cos r = (2n+1)\,\frac{\lambda}{2}\ \text{(maxima)}$$

$$2\,\mu\,t\cos r = n\lambda\ \text{(minima)}$$

  where $\mu$ is the refractive index of the film, $t$ is its thickness and $r$ is the angle of refraction in the film.

- For a thin film in transmitted light, the conditions for constructive and destructive interference are:

$$2\,\mu\,t\cos r = n\lambda\ \text{(maxima)}$$

$$2\,\mu\,t\cos r = (2n+1)\,\frac{\lambda}{2}\ \text{(minima)}$$

- The basic formula for the path difference between the interfering rays, obtained due to division of amplitude by a film of thickness $t$ and refractive index $\mu$, is $2\,\mu\,t\cos r$, where $r$ is the inclination of ray inside the film. If the thickness of the film is uniform, the path difference $2\,\mu\,t\cos r$ varies only with inclination $r$, and gives rise to the "fringes of equal inclination". On the other hand, if the thickness of the film is rapidly varying, the path difference $2\,\mu\,t\cos r$ changes mainly due to changes in $\mu$. This gives rise to the "fringes of equal thickness".

- The spacing $\beta$ between two consecutive bright (or dark) fringes produced by wedge-shaped film is given by

$$\beta = \frac{\lambda}{2\mu\theta}$$

  where $\lambda$ is the wavelength of light being used for illuminating the film, $\mu$ the refractive index of the film, and $\theta$ (measured in radians) the angle between the two plane surfaces, which form the wedge-shaped film.

- The diameters of the bright rings are proportional to the square-roots of the odd natural numbers, whereas the diameters of dark rings are proportional to the square-roots of natural numbers, provided the contact is perfect.

- On measuring the diameters of Newton's rings and the radius of curvature $R$, the wavelength can be calculated with the help of the following relation:

$$\lambda = \frac{D_{n+p}^2 - D_n^2}{4pR}$$

- The phenomenon of interference is used in the testing of optical surfaces and producing non-reflecting glasses of reflective coatings.

## 6.8 TERMINAL QUESTIONS

1) White light is reflected normally from a uniform oil film ($\mu = 1.33$). An interference maximum for 6000 Å and a minimum for 4500 Å, with no minimum in between, are observed. Calculate the thickness of the film.

2) Light ($\lambda = 6000$ Å) falls normally on a thin wedge-shaped film ($\mu = 1.5$). There are ten bright and nine dark fringes over the length of the film. By how much does the film thickness change over this length?

3) Two glass plates 12 cm long touch at one end, and are separated by a wire 0.048 mm in diameter at the other. How many bright fringes will be observed over the 12 cm distance in the light ($\lambda = 6800$ Å) reflected normally from the plates?

4) Newton's rings are formed in reflected light of wavelength $5895 \times 10^{-8}$cm with a liquid between the plane and curved surfaces. The diameter of the fifth ring is 0.3 cm and the radius of curvature of the curved surface is 100 cm. Calculate the refractive index of the liquid, when the ring is (i) bright, (ii) dark.

5) A Newton's rings arrangement is used with a source emitting two wave-lengths

$$\lambda_1 = 6.0 \times 10^{-5} \text{ cm and } \lambda_2 = 4.5 \times 10^{-5} \text{ cm}$$

and it is found that the $n$th dark ring due to $\lambda_1$ coincides with the ($n$+1)th dark ring due to $\lambda_2$. If the radius of curvature of the curved surface is 90 cm, find the diameter of the $n$th dark ring for $\lambda_1$.

## 6.9 SOLUTIONS/ANSWERS

SAQs

1) See Fig. 6.14

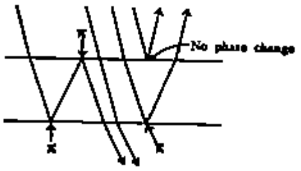


Fig. 6.14

2) According to Eq. (6.7) the path difference between the interfering rays in reflected light is $2\,\mu\,t \cos r - \dfrac{\lambda}{2}$. When the film is excessively thin, $t$ is very small, and $2\,\mu\,t \cos r$ is almost zero. Hence the path difference, in such a case becomes $\dfrac{\lambda}{2}$. This is a condition of minimum intensity. Hence, the film will appear black in the reflected light.

3) Let $\theta$ radian be the angle of the air-wedge. For normal incidence, the fringe-width is given by

$$\beta = \frac{\lambda}{2\theta} \qquad (\because \mu = 1 \text{ for air})$$

Here $\lambda = 5893 \times 10^{-8}$ cm and $\beta = 1/10$ cm.

$$\therefore \quad \theta = \frac{\lambda}{2\beta} = \frac{5893 \times 10^{-8}}{2 \times 1/10} = 2.95 \times 10^{-4} \text{ radian.}$$

4) According to Eq. (6.20), the condition for the dark ring is

$$2t = n\lambda$$

But from Eq. (6.19), $2t = \dfrac{r_n^2}{R}$

$$\therefore \quad \frac{r_n^2}{R} = n\lambda$$

If $D_n$ be the diameter of the $n$th dark ring, $r_n = \dfrac{D_n}{2}$

$$\therefore \quad \frac{D_n^2}{4R} = n\lambda$$

or $\quad D_n = \sqrt{4nR\lambda}$

or $\quad D_n = \sqrt{4R\lambda}\,\sqrt{n}$

or $\quad D_n \propto \sqrt{n}$

Thus, the diameters of the dark rings are proportional to the square root of the natural number.

5) $$\frac{(D_n)^2_{air}}{(D_n)^2_{liquid}} = \mu$$

or $$\frac{D_{liquid}}{D_{air}} = \frac{1}{\sqrt{\mu}} = \frac{1}{\sqrt{1.33}} = 0.867$$

The rings are contracted to 0.867 their previous diameters.

6) In this case of interference in thin films, the situation is somewhat different. The reflections at both the upper and lower surfaces of the material ($\mu$ =1.25) film take place under similar conditions, i.e., when light is going from a rarer to a denser medium. Thus, there is a phase change of $\pi$ at both reflections, which means no phase difference due to reflection between the two interfering beams.

The path difference between the two interfering beams is $2\,\mu\,t$ for normal incidence, where $t$ is the thickness and $\mu$ the refractive index of the film.

The two beams will destroy each other, if the path difference is an odd multiple of $\frac{\lambda}{2}$, i..e, when

$$2\,\mu\,t = (2n-1)\,\frac{\lambda}{2}; \text{ where } n = 1, 2, 3,\ldots..$$

This is the condition of minima.

Here $\mu = 1.25$ and $\lambda = 6000$ Å.

$$\therefore \qquad 2 \times 1.25 \times t = (2n-1) \times \frac{6000}{2}\,\text{Å}$$

Hence the required thickness is given by

$$t = (2n-1)\,\frac{6000}{2 \times 2 \times 1.25}\,\text{Å}$$

$$= (2n-1)\ 1200\text{Å}; \text{ where } n = 1, 2, 3,\ldots$$

TQs

1) The condition for an interference maximum in the light reflected normally from an oil film of thickness $t$ is

$$2\mu\,t = \left(n + \frac{1}{2}\right)\lambda\,; \text{ where } n = 0, 1, 2,\ldots$$

and that for an interference minimum is

$$2\,\mu\,t = n\lambda; \qquad \text{where } n = 1, 2, 3,\ldots\ldots..$$

Here $\mu = 1.33$. Now there is a maximum for $\lambda$ = 6000Å

We can write

$$2 \times 1.33 \times t = \left(n + \frac{1}{2}\right) 6000\ \text{Å.} \qquad \ldots\text{(i)}$$

$$2 \times 1.33 \times t = (n+1)\ 4500\text{Å} \qquad \ldots\text{(ii)}$$

In view of eq. (i) we have taken the integer $(n + 1)$ rather than $n$ in eq. (ii) Comparing eq. (i) and (ii), we get

$$\left(n + \frac{1}{2}\right) 6000 = (n+1)\ 4500$$

$$\therefore \quad n = 1.$$

Substituting $n = 1$ in eq. (i), we get

$$2 \times 133 \times t = \frac{3}{2} \times 6000\text{Å}$$

$$\therefore \quad t = \frac{3 \times 6000}{2 \times 2 \times 133} = 3383\text{Å}$$

2) The condition of destructive interference in light reflected from a film is

$$2\,\mu\,t \cos r = n\lambda.$$

Suppose the film thickness changes over this length by $\Delta t$. Let $n$ be the order of the dark fringe appearing at one end of the film. The order of the dark fringe at the other end will be $(n + 9)$. We, therefore, have

$$2\,\mu\,t \cos r = n\lambda,$$

and $$2\,\mu\,(t + \Delta t) \cos r = (n + 9)\lambda$$

Subtracting, we get

$$2\,\mu\,(\Delta t) \cos r = 9\lambda$$

$$\therefore \quad t = \frac{9\lambda}{2\mu \cos r}$$

If the fringes are seen normally, then $\cos r = 1$.

$$\therefore \quad t = \frac{9}{2\mu} = \frac{9 \times 6300}{2 \times 15} = 18900\ \text{Å}$$

$$= 1.89 \times 10^{-4}\ \text{cm}.$$

3) Let $t$ be the thickness of the wire and $l$ the length of the wedge, as shown in Fig. 6.15. The wedge angle is

$$\theta = \frac{t}{l}\ \text{radian.}$$

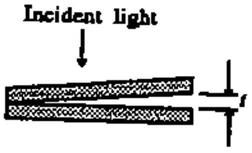


Fig. 6.15

Now, fringe-width $\beta = \dfrac{\lambda}{2\theta}$.

Putting value of $\theta$ from above we get

$$\beta = \frac{l\lambda}{2t}.$$

Since $N$ fringes are seen; $l = N\,\beta$. Thus

$$\beta = \frac{N\beta\lambda}{2t}$$

$$\therefore \quad N = \frac{2t}{\lambda}.$$

But $\lambda = 6800\text{Å} = 6800 \times 10^{-8}$ cm and $t = 0.048$ mm $= 0.0048$ cm.

$$\therefore \quad N = \frac{2 \times 0.0048}{6800 \times 10^{-8}} = 141.$$

4) i) The diameter $D_n$ of the $n$th bright ring is given by

$$D_n^2 = \frac{2\,(2n - 1)\,\lambda R}{\mu}$$

$$\therefore \qquad \mu = \frac{2\,(2n-1)\,\lambda R}{D_n^2}$$

Here $n = 5$, $\lambda = 5895 \times 10^{-8}$cm, $R = 100$ cm and $D_n = 0.3$ cm

$$\therefore \qquad \mu = \frac{2\,(10-1) \times 5895 \times 10^{-8} \times 100}{(0.3)^2} = 1.18$$

ii) The diameter of the $n$th dark ring is given by

$$D_n^2 = \frac{4n\,\lambda R}{\mu}$$

$$\therefore \qquad \mu = \frac{4n\,\lambda R}{D_n^2} = \frac{4 \times 5 \times 5895 \times 10^{-8} \times 100}{(0.3)^2} = 1.31.$$

5) $D_n^2 = 4nR\lambda$

where $D_n$ = diameter of $n$th ring, $R$ = the radius of curved surface and $\lambda$ = the wavelength of light.

If $D_n$ and $D_{n+1}$ be two diameters,.

$$D_n^2 = 4nR\lambda_1 \qquad \text{....(i)}$$

$$D^2{}_{n+1} = 4\,(n+1)\,R\lambda_2$$

But $\quad D_n = D_{n+1}$

$$\therefore \qquad 4nR\lambda_1 = 4\,(n+1)R\lambda_2$$

or $\quad 4nR\,(\lambda_1 - \lambda_2) = 4R\lambda_2$

or $\quad n = \dfrac{4R\lambda_2}{4R\,(\lambda_1 - \lambda_2)}$

$$= \frac{\lambda_2}{(\lambda_1 - \lambda_2)}$$

$$= \frac{4.5 \times 10^{-5}}{(6-4.5)\,10^{-5}} = 3$$

Putting $n = 3$ in (i)

$$D_3 = 4 \times 3 \times 90 \times 6 \times 10^{-5}$$

$$= 648 \times 10^{-4}$$

$$= 25.45 \times 10^{-2}\ \text{cm.}$$

# UNIT 7 INTERFEROMETRY

## Structure

## 7.1 INTRODUCTION

An instrument designed to exploit the interference of light and the fringe patterns that result from optical path differences, in any of a variety of ways, is called an optical interferometer. In this unit, we explain the functioning of the Michelson and the Febry-Perot interferometers, and suggest only a few of their many applications.

In order to achieve interference between two coherent beams of light, an interferometer divides an initial beam into two or more parts that travel diverse optical paths and then superpose to produce an interference pattern. One criterion for broadly classifying interferometers distinguishes the manner in which the initial beam is separated. Wavefront division interferometers sample portions of the same wavefront of a coherent beam of light, as in the case of Young's double slit, Lloyd's mirror or Fresnel's biprism arrangement. Amplitude-division interferometers, instead, use some type of beam-splitter that divides the initial beam into two parts. The Michelson interferometer is of this type. Usually the beam splitting is managed by a semi-reflecting metallic film. In this interferometer, the two interfering beams are widely separated, and the path difference between them can be varied at will by moving the mirror or by introducing a refracting material in one of the beams. Corresponding to these two ways of changing the optical path, there are two important applications of this interferometer, which we will study in this unit.

There is yet another means of classification that distinguishes between those interferometers that function by the interference of two beams, as in the case of the Michelson interferometer, and those that operate with multiple beams, as in the Fabry-Perot interferometer. In this unit, we will show that the fringes so formed are sharper than those formed by two beam interference. Therefore, the interferometers involving multiple beam interference have a very high resolving power, and, hence, find applications in high resolution spectroscopy.

### Objectives

After studying this unit, you should be able to

- understand how Michelson interferometer produces different types of fringes, viz., circular, localised (or straight) and white light fringes,

- describe few applications of Michelson interferometer,
- relate the intensity of the transmitted light to the reflectance of the plate surface in Fabry-Perot interferometer, and
- understand the difference between Michelson interferometer and Fabry-Perot interferometer.

## 7.2 MICHELSON INTERFEROMETER

It is an excellent device to obtain interference fringes of various shapes which have a number of applications in optics. It utilizes the arrangements of mirrors and beam splitter.

**Construction:** Its configuration is illustrated in Fig. 7.1.

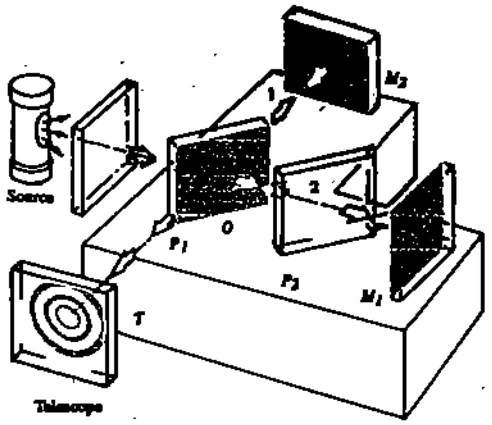


Fig. 7.1: Michelson Interferometer.

Its main optical parts are two plane mirrors $M_1$ and $M_2$ and two similar optically-plane parallel glass plates $P_1$ and $P_2$. The plane mirrors $M_1$ and $M_2$ are silvered on their front surfaces and are mounted vertically on two arms at right angles to each other. To obtain fringes, the mirrors $M_1$ and $M_2$ are made exactly perpendicular to each other by means of screws shown on mirror $M_1$. The mirror $M_2$ is mounted on a carriage which can be moved in the direction of the arrows. The plates $P_1$ and $P_2$ are mounted exactly parallel to each other, and inclined at $45^0$ to $M_1$ and $M_2$. The surface of $P_1$ towards $P_2$ is **partially silvered.** The plate $P_1$ is called beam splitter.

**Working:** An extended source (e.g., a diffusing ground glass plate illuminated by a discharge lamp) emits lightwaves in different directions, part of which travels to the right and falls on $P_1$. The light wave incident on $P_1$ is partly reflected and partly transmitted. Thus, the incident wave gets divided into two waves, viz., the transmitted wave 1 and the reflected wave 2. These two waves travel to $M_1$ and $M_2$ respectively. After reflection at $M_1$ and $M_2$ the two waves return to $P_1$. Part of the wave coming from $M_2$ passes through $P_1$ going downward towards the telescope, and part of the wave coming from $M_1$ gets reflected by $P_1$ toward the telescope. Since the waves entering the telescope are derived from the same incident wave, they are coherent, and, hence, in a position to interfere. The interference fringes can be seen in the telescope.

In contrast to the Young double slit experiment, which uses light from two very narrow sources, the Michelson interferometer uses light from a broad spreadout source.

You must be eager to know the purpose of the plate $P_2$, because till now we have not mentioned anything about $P_2$.

**Function of the plate $P_2$:** Note that if reflection at $P_1$ occurs at the rear surface at point $O$, as shown in Fig. 7.1, the light reflected at $M_2$ will pass through $P_1$ three times while the light reflected at $M_1$ will pass through only once. Thus, the paths of waves 1 and 2 in glass are not equal. Consequently, each wave will pass through the same thickness of glass only when a compensator plate $P_2$, of the same thickness and inclination at $P_1$, is inserted in the path of wave 1. The compensator plate is an exact duplicate of $P_1$ with the exception that it is not partially silvered. With the compensator in place, any optical path difference arises from the actual path difference.

**Form of fringes:** The form of the fringes depends on the inclination of $M_1$ and $M_2$. To understand how fringes are formed, refer to the Fig. 7.2, where the physical components are represented somewhat differently. An observer at the position of the telescope will, simultaneously, see both mirrors $M_1$ and $M_2$ alongwith the source $L$, formed by reflection in the partially silvered surface of the glass plate $P_1$. Accordingly, we can redraw the interferometer as if all the elements were in a straight line. Here $M_1'$ corresponds to the image of mirror $M_1$ formed by reflection at the silvered surface of the glass plate $P_1$ so that $OM_1 = OM_1'$. Depending on the positions of the mirrors, image $M_1'$ may be in front of, behind or exactly coincident with mirror $M_2$. The surfaces $L_1$ and $L_2$ are images of the source $L$ in mirrors $M_1$ and $M_2$ respectively. If we consider a single point $S$ on the source $L$, emitting light in all directions, then on reaching $O$, it gets split, and thereafter its segments get reflected by $M_1$ and $M_2$. In Fig 7.2, we represent this by reflecting the ray off both $M_1$ and $M_2$. Thus, the interference fringes may be regarded to be formed by light reflected from the surface of $M_1'$ and $M_2$. Here, $S_1$ and $S_2$ act as coherent point sources, because to an observer at $D$ the two reflected rays will appear to have come from the image points $S_1$ and $S_2$. The mirror $M_2$ and the virtual image of $M_1$ play the same roles as the two surfaces of the thin film, discussed in unit 6, and the same sort of interferences fringes result from the light reflected by these surfaces.

Now, let us discuss the various types of fringes, viz., circular fringes, localized fringes and white light fringes.

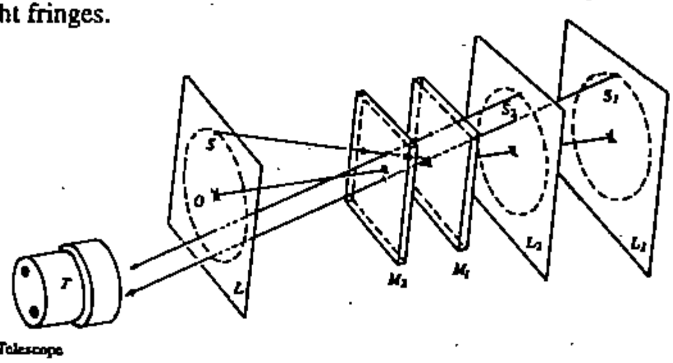


Fig. 7.2: A conceptual rearrangement of the Michelson Interferometer.

## 7.2.1 Circular Fringes

These fringes are observed when $M_1$ is exactly perpendicular to $M_2$. In this situation, the distance of the mirrors $M_1$ and $M_2$ from the plate $P_1$ can be varied.

Let us consider the various possible positions of the mirrors $M_1$ and $M_2$, and, eventually, see how it gives rise to circular fringes. (i) If the two mirrors have the same axial distance from the rear face of $P_1$, and if they are perpendicular to each other, the image $M_1$ is coincident with $M_2$. At the coincidence position, the two paths are of equal length. Thus, we expect the waves to reinforce each other and to form a maximum. But this is not so, because of $\pi$ phase change, which occurs on external (air-to-glass) reflection only. No phase change occurs on internal (glass-to-air) reflection, and none occurs on transmission or refraction. Look again at Fig. 7.1 and note that it is the light that comes from $M_1$ and goes to the observer that is reflected, air-to-glass, at $O$, and undergoes the $\pi$ change. This means that at the coincidence position there will be a minimum: **the centre of the field will be dark.**

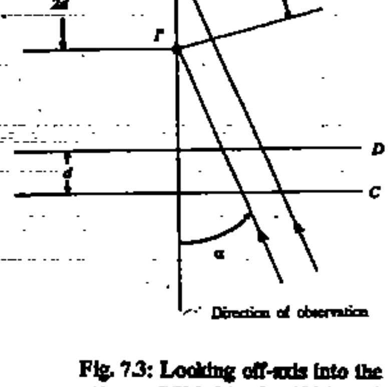


Fig. 7.3: Looking off-axis into the Michelson Interferometer.

(ii) Now, we move one of the mirrors. If the mirror is moved through a quarter of wavelength, $d = \lambda/4$, the path length (because if $d$ is separation between $M_1$ and $M_2$, then $2d$ is the separation between $S_1$ and $S_2$) changes by $\lambda/2$, the two waves getting out of phase by 180$^0$, the phase change compensates, and we have a maximum. Moving the mirror by another $\lambda/4$, gives minimum, another $\lambda/4$ another maximum and so on. Thus,

$$2d = m\lambda, \text{ where } m = 0, 1, 2, \qquad \ldots(7.1)$$

is the **Michelson's interferometer equation.**

**(iii) Next, assume that we look obliquely into the interferometer and that our line of sight makes an angle** $\alpha$ with the axis. Ordinarily, the two planes $M_1$ and $M_2$ are at a distance $d$ apart, and the two virtual images, $I$ and $I'$ separated by $2d$. But for oblique incidence, as we see from Fig. 7.3, the path difference between the two lines of sight becomes less and instead of Eq. (7.1), we get

$$2d \cos \alpha = m\lambda \text{ ; where } m = 0, 1, ... \qquad ...(7.2)$$

For a given mirror separation $d$, and a given order $m$, wavelength $\lambda$ and angle $\alpha$ is constant. The maxima will lie in the form of circles about the foot of the perpendicular from the eye to the mirrors. These circular fringes will look like the ones shown in Fig. 7.4. Fringes of this kind, where parallel beams are brought to interfere with a phase difference determined by the angle of inclination $\theta$, are referred to as fringes of equal inclination. These fringes are also known as Haidinger fringes. They differ from the fringes of equal inclination considered in Unit 6, only in that, here there are no multiple reflections so that the intensity distribution is in accordance with Eq. (5.17)

**Fig. 7.4: Fringes observed using (a) Michelson interferometer, (b) Fabry-Perot interferometer.**

**Fig. 7.5: Appearance of the various types of fringes observed in the Michelson interferometer. Upper row shows circular fringes whereas lower row shows, localized fringes. Path difference increases outward, in both directions, from the centre.**

The upper part of the Fig. 7.5 shows how the circular fringes look under different conditions. When $M_2$ is few centimeters beyond $M_1$, the fringe system will have the general appearance shown in (a) with the rings very closely spaced. If $M_2$ is now moved slowly toward $M_1$, so that $d$ is decreased, Eq. (7.2) shows that a given ring, characterized by a given value of the order $m$, must decrease its radius, because the product $2d \cos \alpha$ must remain constant. The rings, therefore, shrink and vanish at the centre, a ring disappearing each time $2d$ decreases by $\lambda$, or $d$ by $\lambda/2$. This follows from the fact at the centre $\cos \theta = 1$, so that Eq. (7.2) becomes

$$2d = m\lambda$$

which is Eq. (7.1).

To change $m$ by unity, $d$ must change by $\lambda/2$. Now as $M_2$ approaches $M_1$, the rings become more widely spaced as indicated in Fig. (7.5b), until we reach a critical position, where the central fringe has spread out to cover the whole field of view, as shown in Fig. 7.5 (c). This happens when $M_2$ and $M_1$ are exactly coincident, for it is clear that under these conditions the path difference is zero for all angles of incidence. If the mirror is moved still farther, it effectively passes through $M_1$, and new widely spaced fringes appear, growing out from the centre. These will gradually become more closely spaced, when the path difference increases, as indicated in (d) and (e) of the Fig. 7.5.

### 7.2.2 Localized Fringes (Straight Fringes)

If the mirrors $M_1$ and $M_2$ are not exactly parallel, the air film between the mirrors is wedge-shaped, as indicated in Fig. 7.6.

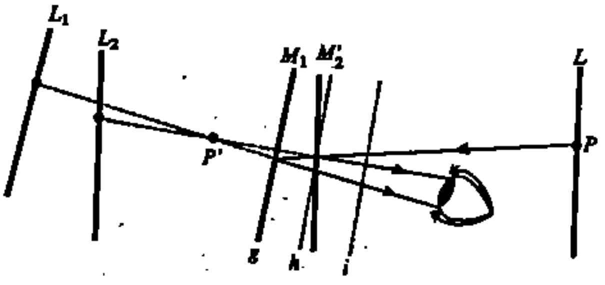


Fig. 7.6: The formation of fringes with inclined mirrors in the Michelson Interferometer.

The two rays reaching the eye from point $P$ on the source are now no longer parallel, but appear to diverge from point $P'$ near the mirrors. For various positions of $P$ on the extended source, the path difference between the two rays remains constant, but the distance of $P'$ from mirrors changes. If the angle between the mirrors is not too small, the latter distance is never great, and hence, in order to see these fringes clearly, the eye must be focused on or near the rear mirror $M_2$. The localized fringes are, practically, straight, because the variation of the path difference across the field of view is now due primarily to the variation of the thickness of the "air film" between the mirrors. With a wedge-shaped film, the locus of point of equal thickness is a straight line, parallel to the edge of the wedge. The fringes are not exactly straight, if $d$ has an appreciable value, because there is also some variation of the path difference with angle. They are, in general, curved and are always convex toward the thin edge of the wedge. Thus, with a certain value of $d$, we might observe fringes shaped like those of Fig. 7.5(g). $M_2$ could then be in position such as $g$ of Fig. 7.6. If the separation of the mirrors is decreased, the fringes will move to the left across the field, a new fringe crossing the centre each time $d$ changes by $\lambda/2$. As we approach the zero path difference, the fringes become straighter until the point is reached where $M_2$ actually intersects $M_1$, when they are perfectly straight, as in Fig. 7.5(h). Beyond this point, they begin to curve in the opposite direction, as shown in Fig. 7.5 (i). The blank fields shown in Fig. 7.5 (f) and (j) indicate that this type of fringe cannot be observed for large path differences. As the principle variation of path difference results from a change of the thickness $d$, these fringes have been termed fringes of equal thickness.

### 7.2.3 White Light Fringes

If a source of white light is used, no fringes will be seen at all except for a path difference so small that it does not exceed a few wavelengths. In observing these fringes, the mirrors are tilted slightly as for localized fringes, and the position of $M_2$ is found where it intersects $M_1$. With white light there will then be observed a central dark fringe, bordered on either side by 8 or 10 coloured fringes. This position is often rather troublesome to find, using white light only. It is best located approximately before hand by finding the place where the localized fringes in monochromatic light become straight. Then a very slow motion of $M_1$ through this region, using white light, will bring these fringes into view.

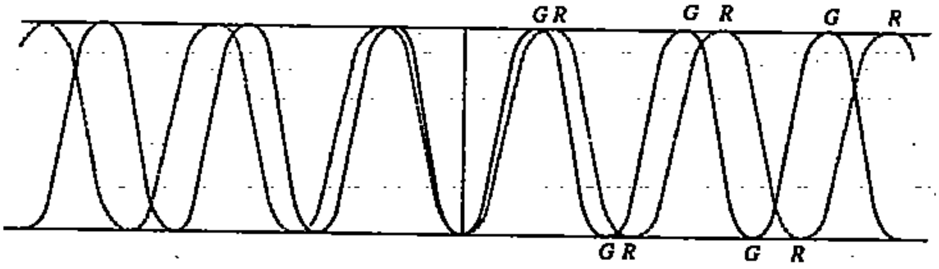


Fig. 7.7: The formation of white light fringes with a dark fringe at the centre.

The fact, that only a few fringes are observed with white light, is easily accounted for when we remember that such light contains all wavelengths between 400 and 750 mm. The fringes for a given colour are more widely spaced, the greater the wavelength. Thus, the fringes in different colours will only coincide for $d = 0$, as indicated in Fig. 7.7. The solid curve represents the intensity distribution in the fringes for the green light, and the broken curve for the red light. Clearly, only the central fringe will be uncoloured, and the fringes of different colours will begin to separate at once on either side. After 8 or 10 fringes, so many colours are present at a given point that the resultant colour is essentially white. White light fringes are, particularly, important in the Michelson interferometer, where they may be used to locate the position of zero path difference, as we shall see later.

## 7.2.4 Adjustment of the Michelson's Interferometer

i) **For Localised fringes:** The distance of the mirrors $M_1$ and $M_2$ from the silvered surface of $P_1$ are first made as nearly equal as possible by moving the movable mirror $M_2$. A pin-hole is placed between the lens and the plate $P_1$ (Fig. 7.8). If $M_1$ is not perpendicular to $M_2$, four images of the pin-hole are obtained, two by reflection at the semi-silvered surface of $P_1$ and the other two by reflection at the other surface of $P_1$.

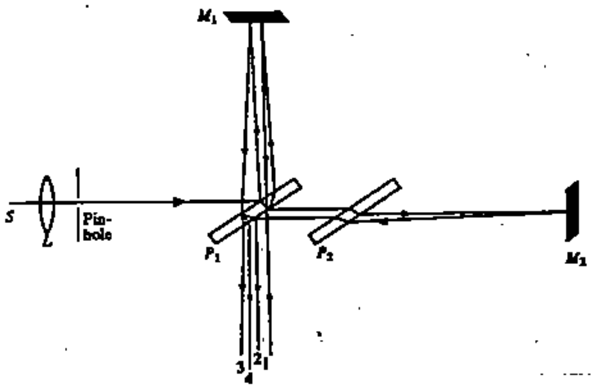


Fig. 7.8: Adjustment of Michelson interferometer.

The former pair is, naturally, brighter than the latter. The small screws at the back of the mirror, $M_1$, are then adjusted until the two bright images appear to coincide. The pin-hole is now removed. If the coincidence of the images was apparent, the air-film between $M_2$ and $M_1$ would be wedge-shaped, and the localised fringes would appear.

ii) **For White light Localised Fringes:** First, the localised fringes with monochromatic light are obtained. The mirror $M_2$ is then moved until the fringes become straight. Monochromatic light is replaced by white light. $M_2$ is further moved in the same direction until the central achromatic fringe is obtained in the field of view.

iii) **For Circular Fringes:** After localised fringes are obtained, the screws of $M_1$ are adjusted so that the spacing between these fringes increases. This happens when the angle of the wedge decreases. If this adjustment be continued, at one stage, the angle of the wedge will become zero, and the film will be of constant thickness. At this stage, circular fringes will appear. Finer adjustment is made until on moving the eye side ways or up and down, the fringes do not expand or contract.

## 7.2.5 Applications

There are three principal types of measurement that can be made with Michelson interferometer: (i) wavelengths of light (ii) width and fine structure of spectrum lines (iii) refractive indices. As explained in the sub-section 7.2.3, when a certain spread of wavelengths is present in the light source, the fringes become indistinct and, eventually, disappear as the path difference is increased. With white light they become invisible when $d$ is only a few wavelengths, whereas the circular fringes obtained with the light of single spectrum line can still be seen after the mirror has been moved several centimeters. Therefore, for making these measurements with this interferometer, it is adjusted for circular fringes.

a). **Determination of Wavelength of Monochromatic Light**

After having adjusted interferometer for circular fringes, adjust the position of $M_2$ to obtain a bright spot at the centre of the field of view. If $d$ be the thickness of the film and $n$ the order of the spot obtained, we have

$$2d \cos \alpha = n \lambda \tag{7.3}$$

But at the centre $\alpha = 0$; so that $\cos \alpha = 1$. Therefore

$$2d = n\lambda \tag{7.4}$$

If now $M_2$ be moved away form $M_1$ by $\lambda/2$, $2d$ increases by $\lambda$. Therefore $n + 1$ replaces $n$ in Eq. (7.4). Hence, $(n + 1)$th bright spot now appears at the centre (see sec. 7.2.1). Thus, each time $M_2$ moves through a distance $\lambda/2$, next bright spot appears at the centre. Suppose, during the movement of $M_2$ through a distance $x$, $N$ new fringes appear at the centre of the field. Then we have

$$x = N\frac{\lambda}{2}$$

$$\therefore \qquad \lambda = \frac{2x}{N} \tag{7.5}$$

Thus, by measuring the distance $x$ with the micrometer and counting the number $N$, the value of $\lambda$ can be obtained.

The determination of $\lambda$ by this method is very accurate, because $x$ can be measured to an accuracy of $10^{-4}$mm, and the value of $N$ can be sufficiently increased, as the circular fringes can be obtained up to large path differences.

**SAQ 1**

When the movable mirror of Michelson's interferometer is shifted through 0.0589 mm, a shift of 200 fringes is observed. What is the wavelength of light used? Give the answer in Angstrom units.

(b) **Determination of difference in Wavelength :** When the source of light has two wavelengths $\lambda_1$ and $\lambda_2$ very close together (like $D_1$ and $D_2$ lines of sodium), each wavelength produces its own system of rings. Let $\lambda_1 > \lambda_2$. When the thickness of the film is small, the rings due to $\lambda_1$ and $\lambda_2$ almost coincide, since $\lambda_1$ and $\lambda_2$ are nearly equal. The mirror $M_2$ is moved away. Then, due to different spacing between the rings of $\lambda_1$ and $\lambda_2$, the rings of $\lambda_1$ are gradually separated from those of $\lambda_2$. When the thickness of the air-film becomes such that dark rings of $\lambda_1$ coincides with bright rings of $\lambda_2$ (due to closeness of $\lambda_1$ and $\lambda_2$, the dark rings due to $\lambda_1$ will practically coincide with bright rings due to $\lambda/_2$ in the entire field of view), the rings have maximum indistinctness.

The mirror $M_2$ is moved further away through a distance, say, $x$ until the rings, after becoming most distinct, once again become most indistinct. Clearly, during this movement, $n$ fringes of $\lambda_1$ and $(n + 1)$ fringes of $\lambda_2$ have appeared at the centre (because then the dark rings of $\lambda_1$ will again coincide with the bright rings of $\lambda_2$). Now, since the movement of the mirror $M_2$ by $\lambda_2$ results in the appearance of one new fringe at the centre, we have

$$x = n\frac{\lambda_1}{2} = (n + 1)\frac{\lambda_2}{2}$$

$$\text{or} \qquad n = \frac{2x}{\lambda_1} \text{ and } (n + 1) = \frac{2x}{\lambda_2}$$

$$\therefore \qquad \frac{2x}{\lambda_2} - \frac{2x}{\lambda_1} = 1$$

$$\text{or} \qquad \frac{2x(\lambda_1 - \lambda_2)}{\lambda_1\lambda_2} = 1$$

$$\text{or} \qquad \lambda_1 - \lambda_2 = \frac{\lambda_1\lambda_2}{2x}$$

Since $\lambda_1$ and $\lambda_2$ are close together, $\lambda_1\ \lambda_2$ can be replaced by $\lambda^2$ where $\lambda$ is the mean of $\lambda_1$ and $\lambda_2$.

$$\therefore \qquad \lambda_1 - \lambda_2 = \frac{\lambda^2}{2x} \tag{7.6}$$

Thus if we measure the distance moved by $M_2$ between two consecutive positions of disappearance of the fringe pattern and the mean wavelength is known, we can determine the difference $(\lambda_1 - \lambda_2)$.

SAQ 2

In Michelson's interferometer, the reading for a pair of maximum indistinctness were found to be 0.6939 mm and 0.9884 mm. If the mean wavelength of the two components of light be 5893Å, deduce the difference between the wavelengths of the components.

### c) Determination of Refractive Index of a Thin Plate

If a thickness $t$ of a substance having an index of refraction $\mu$ is introduced into the path of one of the interfering beams in the interferometer, the optical path in this beam is increased because of the fact that light travels more slowly in the substance, and consequently, has a shorter wavelength. The optical path is now $\mu t$ through the medium, whereas it was practically $t$ through the corresponding thickness of air ($\mu = 1$). Thus, the increase in the optical path due to insertion of the substance is $(\mu - 1)t$.

In practice, the insertion of a plate of glass in one of the beams produces a discontinuous shift of the fringes so that the number of fringes cannot be counted. With monochromatic fringes, it is impossible to tell which fringe in the displaced set corresponds to one in the original set. With white light, the displacement in the fringes of different colours is very different. This illustrates the necessity of adjusting the interferometer to produce straight white light fringes. After having adjusted so, the cross-wire is set on the achromatic fringe, which is perfectly straight. The given plate is now inserted in the path of one of the interfering waves. This increases the optical path of the beam by $(\mu - 1)\ t$. Since the beam traverses the plate twice, an extra path difference of $2\ (\mu - 1)t$ is introduced between the two interfering beams. The fringes get shifted. The movable mirror $M_2$ is moved till the fringes are brought back to their initial positions so that the achromatic fringe again coincides with the cross wire. If the displacement of $M_2$ is $x$, then

$$2x = 2(\mu - 1)t$$

or

$$x = (\mu - 1)t. \tag{7.7}$$

Alternatively, if $N$ be the number of fringes shifted then

$$2(\mu - 1)t = N\lambda \tag{7.8}$$

Thus, measuring $x$, $t$, may be calculated if $\mu$ is known, or $\mu$ may be calculated if $t$ is known.

This method can be used to find the refractive index of a gas. The gas is introduced into an evacuated tube placed along the axis of one of the interfering beams, and the experiment is carried out as described above.

SAQ 3

A transparent film of glass of refractive index 1.50 is introduced normally in the path of one of the interfering beams of a Michelson's interferometer, which is illuminated with light of wavelength 4800Å. This causes 500 dark fringes to sweep across this field. Determine the thickness of the film.

There is yet another type of interferometer, called Fabry-Perot interferometer, which produces fringes much sharper than those produced by Michelson interferometer. In the next section, let us study this interferometer and see how it is used as a powerful spectrometer.

## 7.3 FABRY-PEROT INTERFEROMETER

It is based on the principle of multiple beam interference. It is a high resolving power instrument, which makes use of the 'fringes of constant inclination' produced by the transmitted light after multiple reflections between two parallel and highly-reflecting glass plates.

It consists of two optically-plane glass plates $A$ and $B$ (Fig. 7.9) with plane surfaces. The inner surfaces are coated with partially transparent films of high reflectivity and placed accurately parallel to each other. Screws are provided to secure parallelism if disturbed. The two uncoated surfaces of each plate are made to have a slight angle between them in order to avoid unwanted fringes formed due to multiple reflections in the plate itself.

One of the two plates is kept fixed, while the other can be moved to vary the separation of the two plates. In this configuration, the instrument is called a Fabry-Perot interferometer. Sometimes both the plates are at a fixed separation with the help of spacers. The system with fixed spacing is known as Fabry-Perot etalon. The Fabry-Perot interferometer (or etalon) is used to determine wavelengths precisely, to compare two wavelengths, to calibrate the standard metre in terms of wavelength, etc.

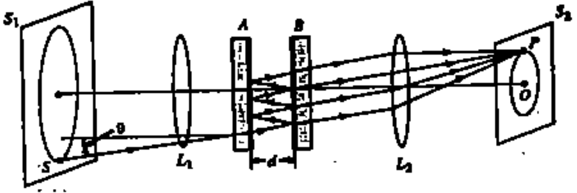


Fig. 7.9: Fabry-Perot Interferometer. $S$ is part of an external light source.

$S_1$ is a broad source of monochromatic light and $L_1$ a convex lens which makes the beam more collimated. An incident ray suffers a large number of internal reflections successively at the two silvered surfaces, as shown. At each reflection a small fractional part of the light is also transmitted. Thus, each incident ray produces a group of coherent and parallel transmitted rays with a constant path difference between any two successive rays. A second convex lens, $L_2$, brings these rays together at a point $P$ in its focal plane, where they interfere. Hence, the rays from all points of the source produce an interference pattern on a screen $S_2$ placed in the focal plane of $L_2$.

**Formation of the Fringes:** Let $d$ be the separation between the two silvered surfaces, and $\theta$ the inclination of particular ray with the normal to the plates. Then the path difference between any two successive transmitted rays corresponding to the incident ray is $2d \cos\theta$. The medium between the two silvered surfaces is usually air. As you saw, while solving SAQ 1 in Unit 6, that $\pi$ phase changes occur on both of these (air-to-glass) surfaces, hence, the condition

$$2d \cos\theta = n\lambda,$$

holds for maximum intensity.

Here, $n$ is an integer, called the order of interference, and $\lambda$ the wavelength of light. The locus of points in the source which give rays of a constant inclination $\theta$ is a circle. Hence, with an extended source, the interference pattern consists of a system of bright concentric rings on a dark background, each ring corresponding to a particular value of $\theta$. Fig. 7.4(b) shows the fringes obtained using a Fabry-Perot interferometer. Also shown, in the figure for comparison, are fringes obtained by using Michelson interferometer (see Fig. 7.4a). It can readily be seen that the Fabry-Perot interferometer, which employs the principle of multiple beam interference, produces much sharper fringes, and could, hence, be used to study hyperfine structure of spectral lines. The intensity distribution of the circular fringes of Fig. 7.4b is not in accordance with Eq. (5.17). To determine how much light is reflected and transmitted at the two surfaces, let us read the following section.

### 7.3.1 Intensity Distribution

**Comment:** You are advised to go through the Appendix carefully given at the end of this unit.

We return now to the problem of reflections from a parallel plate, already considered in a two-beam approximation in Unit 6. Fig. 7.10 shows the multiple reflections and transmissions through a plane parallel plate of "air" enclosed between two glass plates of Fabry-Perot interferometer. Here, $n'$ is the refractive index of glass plate and $n$ the refractive index of air enclosed. Suppose a wave is incident at an angle $\theta$, as shown in Fig. 7.10. This incident wave will suffer multiple reflections. Let the reflection and transmission amplitude co-efficient be $r$ and $t$ at an external reflection and $r'$ and $t'$ at an internal reflection.

If the amplitude of incident ray is expressed as $a\ e^{i\omega t}$, the successive transmitted rays can be expressed by appropriately modifying both the amplitude and phase of the initial wave. Referring to Fig. 7.10, these are

$$A_1 = (tt'\,a)\,e^{i\omega t}$$
$$A_2 = (tt'\,r'^2\,a)\,e^{i(\omega t-\delta)}$$
$$A_3 = (tt'\,r'^4\;\;a)\,e^{i(\omega t-2\delta)} \text{ and so on.}$$

A little inspection of these equations shows that

$$A_N = tt'\,r'^{2(N-1)}\,ae^{i\omega t}\,e^{-i(N-1)\delta}$$

The quantities $r, r', t, t'$, are given in terms of $n, n'\ \theta, \theta'$ by the Fresnel formulae. For our present purpose we do not need these explicit expressions but only relations between them. We have

$$tt' = T \qquad \text{...7.9(a)}$$

and
$$r^2 = r'^2 = R \qquad \text{...7.9(b)}$$

where $R$ and $T$, respectively are the reflectivity and transmissivity of the plate surfaces. Then, using Eq. 7.9, we have

$$A_1 = aTe^{i\omega t},$$
$$A_2 = aTRe^{i(\omega t-\delta)},$$
$$A_3 = aTR^2e^{i(\omega t-2\delta)}, \text{ and so on.}$$

By the principle of superposition, the resultant amplitude is given by

$$A = aT + aTRe^{-i\delta} + aTR^2e^{-2i\delta} + aTR^3e^{-3i\delta} + ....$$

Here, we have ignored $e^{i\omega t}$, as it is of no importance in combining waves of the same frequency. Hence,

$$A = aT\,(1 + Re^{-i\delta} + R^2e^{-2i\delta} + R^3e^{-3i\delta} + ...)$$

The infinite geometric series in the parentheses has the common ratio $Re^{i\delta}$ and has a finite sum because $r^2 < 1$. Summing up the series, we obtain

$$A = aT\frac{1}{1 - Re^{-i\delta}}$$

The complex conjugate of $A$ is therefore

$$A^* = aT\frac{1}{1 - Re^{+i\delta}}$$

Hence the resultant intensity $I$ is given by

$$I = AA^* = \frac{a^2T^2}{(1 - Re^{-i\delta})\,(1 - Re^{+i\delta})}$$

$$= \frac{a^2T^2}{1 - R^2 - 2R\,(e^{i\delta} + e^{-i\delta})} = \frac{a^2T^2}{1 + R^2 - 2R\cos\delta}$$

$$= \frac{a^2T^2}{(1 - R^2) + 2R\,(1 - \cos\delta)} = \frac{a^2T^2}{(1 - R^2) + 4R\sin^2\dfrac{\delta}{2}}$$

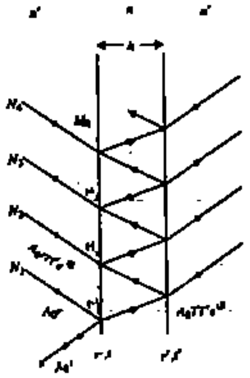


Fig. 7.10: Multiple reflection and transmission in a parallel "air" plate enclosed between the two plates of Fabry-Perot interferometer.

$$= \frac{a^2T^2}{(1-R)^2} \frac{1}{1 + \frac{4R}{(1-R)^2}\sin^2\frac{\delta}{2}} \qquad \text{...(7.10)}$$

The intensity will be a maximum when $\sin^2 \frac{\delta}{2} = 0$, i.e. $\delta = 2n\pi$ where $n = 0, 1, 2, ...$.Thus

$$I_{max} = \frac{a^2T^2}{(1-R)^2} \qquad \text{...(7.11)}$$

Similarly, the intensity will be a minimum when $\sin^2 \frac{\delta}{2} = 1$, i.e. $\delta = (2n+1)\pi$ where $n = 0, 1, 2, ...$Thus

$$I_{min} = \frac{a^2T^2}{(1-R)^2} \frac{1}{1 + \frac{4R}{(1-R)^2}} = \frac{a^2T^2}{(1+R)^2} \qquad \text{...(7.12)}$$

Eq. (7.10) can now be written as

$$I = \frac{I_{max}}{1 + \frac{4R}{(1-R)^2}\sin^2\frac{\delta}{2}} \qquad \text{...(7.13)}$$

or

$$I = \frac{I_{max}}{1 + F\sin^2\frac{\delta}{2}} \qquad \text{...(7.14)}$$

Here, $F = \frac{4R}{(1-R)^2}$ is called the coefficient of Finesse. Eq. (7.14) is the intensity expression for the Fabry-Perot fringes.

If we plot $I$ against $\delta$ for different values of $R$ (the reflectivity of the plates), a set of curves is obtained (Fig. 7.11). They show that the larger the value of $R$, the more rapid is the fall of intensity on either side of a maximum. (That is, higher the reflectivity of the plates, sharper are the interference bright fringes.) Further, as Eq. (7.11) and (7.12) show, larger the value of $R$, greater is the difference between $I_{max}$ and $I_{min}$. In fact, we obtain a system of sharp and bright rings against a wide dark background.

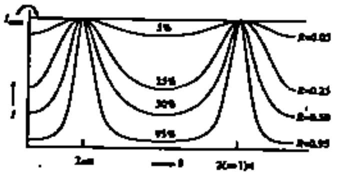


Fig. 7.11: The transmitted intensity as a function of δ showing how the sharpness depends on reflectance. Percentages refer to reflectance of surfaces.

As mentioned in the beginning of the sec. 7.3, Fabry-Perot interferometer is a high resolving power instrument. Its resolving power $\frac{\lambda}{\Delta\lambda}$ is given by

$$\frac{\lambda}{\Delta\lambda} = \frac{4\pi h \cos r \sqrt{F}}{4.147\,\lambda}$$

where $h$ is the thickness of the film enclosed between the two silvered surfaces, $r$ is the angle of refraction inside the film, $\lambda$ the wavelength of incident light and $F$ is the coefficient of Finesse.

To have an idea of the numerical value of resolving power, let us consider a Fabry-Perot etalon with $h$ = 1cm and $F$ = 80. The resolving power for normal incidence in the wavelength region around $\lambda$ = 5000 Å would be

$$\frac{\lambda}{\Delta\lambda} = \frac{4\pi\sqrt{80}}{5 \times 10^{-5} \times 4.147} = 5.42 \times 10^5$$

that is, two wavelengths separated by 0.0092 Å can be resolved at $\lambda$ = 5000Å.

### 7.3.2 Superiority over Michelson's Interferometer

When the light consists of two or more close wavelengths (such as $D_1$ and $D_2$ lines of sodium), then in a Fabry-Perot interferometer each wavelength produces its own pattern, and the rings of one pattern are clearly separated from the corresponding rings of the other pattern. Hence the instrument is very suitable for the study of the fine structure of spectral lines. In Michelson's instrument separate patterns are not produced. The presence of two close wavelengths is judged by the alternate distinctness and indistinctness of the rings when the optical path difference is increased.

## 7.4 SUMMARY

- The Michelson interferometer uses an extended monochromatic source.
- When $M_1$ and $M_2$ are perpendicular to each other, i.e., when $M_1$ and $M_2$ are parallel, the fringes given by a monochromatic source are circular and localized at infinity.
- When the mirrors of the interferometer are inclined with respect to each other, i.e., when $M_1$ and $M_2$ are not perpendicular to each other, a pattern of straight parallel fringes are obtained.
- Whether $M_1$ and $M_2$ are parallel or inclined, any fringe shift seen in an interferometer may be due to either a change in thickness or a change in refractive index.
- As the movable mirror is displaced by $\frac{\lambda}{2}$, each fringe will move to the position previously occupied by an adjacent fringe. If $N$ is the number of fringes that have moved past a reference point, when the mirror is moved a distance $x$, then

$$x = N\frac{\lambda}{2}$$

- Michelson interferometer can be used in the measurement of two closely spaced wavelengths.
- Fabry-Perot interferometer, which employs the principle of multiple beam interference, produces much sharper fringes than those produced by Michelson interferometer.
- In the Fabry-Perot interferometer it is the fringe pattern formed by transmitted light that is observed and as such that intensity distribution would be given by

$$I = \frac{I_{max}}{1 + \frac{4R}{(1-R)^2}\sin^2\frac{\delta}{2}}$$

- Resolving power of Fabry-Perot interferometer is given by

$$\frac{\lambda}{\Delta\lambda} = \frac{4\pi h \cos r\sqrt{F}}{4.147\,\lambda}$$

## 7.5 TERMINAL QUESTIONS

i) When one leg of a Michelson interferometer is lengthened slightly, 150 dark fringes sweep through the field of view. If the light used has $\lambda = 480$ mm, how far was the mirror in that leg moved?

2) Circular fringes are observed in a Michelson interferometer illuminated with light of wavelength 5896 Å. When the path difference between the mirrors $M_1$ and $M_2$ is 0.3 cm, the central fringe is bright. Calculate the angular diameter of the 7th bright fringe.

## 7.6 SOLUTIONS AND ANSWERS

**SAQs**

1) The distance, $x$, moved by the mirror when $N$ fringes cross the field of view is given by

$$x = N\frac{\lambda}{2}$$

$$\therefore \quad \lambda = \frac{2x}{N}$$

Here, $x = 0.00589$ cm, and $N = 200$.

$$\therefore \qquad \lambda = \frac{2 \times 0.00589}{200} = 0.0000589 \text{ cm} = 5890 \text{ Å}$$

2) If $x$ be the distance moved by the movable mirror between two consecutive positions of maximum indistinctness (or distinctness), we have

$$\Delta\lambda = \frac{\lambda_1 \times \lambda_2}{2x} = \frac{\lambda^2}{2x},$$

Where $\lambda$ is the average of $\lambda_1$ and $\lambda_2$.

Here $\lambda = 5893$Å $= 5893 \times 10^{-8}$cm and $x = 0.9884 - 0.6939 = 0.2945$ mm $= 0.02945$ cm.

$$\therefore \qquad \Delta\lambda = \frac{(5893 \times 10^{-8})^2}{2 \times 0.02945} = 5.896 \times 10^{-8} \text{cm} = 5.896\text{Å}.$$

3) Let $t$ be the thickness of the film. When it is put in the path of one of the interfering beams of the Michelson's interferometer, an additional path difference of $2(\mu - 1)t$ is introduced. If $N$ be the number of fringes shifted, we have

$$2\,(\mu - 1)\,t = N\lambda$$

$$\therefore \qquad t = \frac{N\lambda}{2\,(\mu - 1)}$$

Here $N = 500$; $\lambda = 4800 \times 10^{-8}$cm, $\mu = 1.50$.

$$\therefore \qquad t = \frac{500 \times 4800 \times 10^{-8}}{2\,(1.50 - 1)}$$

$$= \frac{500 \times 4800 \times 10^{-8}}{2 \times 0.50}$$

$$= 0.024 \text{ cm.}$$

**TQs**

1) Darkness is observed when the light beams from the two legs are 180° out of phase. As the length of one leg is increased by $\frac{\lambda}{2}$, the path length increases by $\lambda$, and the field of view changes from dark to bright to dark. When 150 fringes pass, the leg is lengthened by an amount

$$(150)\left(\frac{\lambda}{2}\right) = (150)\,(240 \text{ nm}) = 36{,}000 \text{ nm} = 0.036 \text{ mm}$$

2) The expression for the bright circular fringe is

$$2d \cos r = n\lambda$$

At the centre $r = 0$, so that

$$2d = n\lambda \qquad \text{(i)}$$

$n$ now stands for the order of the central bright fringe. The order of fringes decreases as we move outwards from the centre. Thus the second bright fringe is of $(n - 1)$th order,...., seventh bright fringe is of $(n - 6)$th order. Hence if $\theta$ be the angular radius of 7th bright fringe, we have

$$2d \cos \theta = (n - 6)\lambda \qquad \text{(ii)}$$

Eq. (i) and (ii) give

$$2d\,(1 - \cos \theta) = 6\lambda$$

or $\cos\theta = 1 - \dfrac{6\lambda}{2d}$

Putting the given values:

$$\cos\theta = 1 - \frac{6 \times (5896 \times 10^{-8}\ \text{cm})}{2 \times 0.3\ \text{cm}}$$

$$= 1 - 0.0005896 = 0.9994$$

$\therefore$ $\theta = \cos^{-1}(0.9994) = 2°.$

$\therefore$ angular diameter = 4°.

## 7.7 APPENDIX

### Method of Complex Amplitudes

In place of using the sine or the cosine to represent a simple harmonic wave, one may write the equation in the exponential form as

$$y = ae^{(i\omega t - kx)} = ae^{i\omega t}\, e^{-i\delta}$$

where $\delta = kx$ is constant at a particular point in space and represents phase of the wave. The presence of $i = -1$ in this equation makes the quantities complex. We can nevertheless use this representation, and at the end of the problem take either the real (cosine) or the imaginary (sine) part of the resulting expression. The time–varying factor exp $(i\omega t)$ is of no importance in combining waves of the same frequency, since the amplitudes and relative phases are independent of time. The other factor, $a$ exp $(-i\delta)$, is called the complex amplitude. It is a complex number whose modulus $a$ is the real amplitude, and whose argument $\delta$ gives the phase relative to some standard phase. Negative sign merely indicates that the phase is behind the standard phase. In general, the vector $a$ is given by

$$a = ae^{i\delta} = x + iy = a(\cos\delta + i\sin\delta)$$

Then it will be seen that

$$a = \sqrt{x^2 + y^2}, \quad \tan\delta = \frac{y}{x}$$

Thus, if $a$ is represented as in Fig. (7.12), plotting horizontally its real part and vertically its imaginary part, it will have the magnitude $a$ and will make the angle $\delta$ with the $x$ axis, as we require for vector addition.

The advantage of using complex amplitudes lies in the fact that the vector addition of real amplitudes can be written more easily in the form of an algebraic addition of complex amplitudes. For example, consider the real parts of two waves that follow the equations

$$A_1 = A_1 e^{i(\omega t + \delta_1)}$$

and

$$A_2 = A_2 e^{i(\omega t + \delta_2)} \qquad \text{...(7.15)}$$

Adding these two equations gives

$$A = A_1 + A_2 = A_1 e^{i(\omega t + \delta_1)} + A_2 e^{i(\omega t + \delta_2)} \qquad \text{...(7.16)}$$

We can now take out the common exponent $i\omega t$:

$$A = e^{i\omega t}\,(A_1 e^{i\delta_1} + A_2 e^{i\delta_2}) \qquad \text{...(7.17)}$$

The square of the resultant, A², is found by multiplying the complex terms by their complex conjugates:

$$A^2 = (A_1 e^{i\delta_1} + A_2 e^{i\delta_2})(A_1 e^{-i\delta_1} + A_2 e^{-i\delta_2})\cdot$$

$$= A_1{}^2 + A_2{}^2 + A_1 A_2\ e^{i(\delta_1 - \delta_2)} + e^{-i(\delta_1 - \delta_2)} \qquad \ldots(7.18)$$

Then, from Euler's formula,

$$e^{i\delta} + e^{-i\delta} = \cos\,\delta + i\,\sin\,\delta + \cos\delta - i\,\sin\delta = 2\cos\,\delta \qquad \ldots(7.19)$$

ánd therefore, Eq. (7.18) becomes

$$A^2 = A_1{}^2 + A_2{}^2 + 2A_1A_2\cos\,(\delta_1 - \delta_2) \qquad \ldots(7.20)$$

the same as Eq. (5.15).

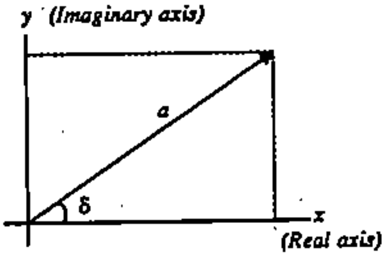


Fig. 7.12: Representation of a vector in the complex plane.

Thus, in obtaining the resultant intensity as proportional to the square of the real amplitude, we multiply the resultant complex amplitude by its complex conjugate, which is the same expression with $i$ replaced by $-i$ throughout.

Uttar Pradesh
Rajarshi Tandon Open University

# UGPHS-07
# Optics

Block

# 3

## DIFFRACTION

# BLOCK INTRODUCTION

In Bolck 2 you learnt that when light from two coherent sources is made to superpose, redistribution of energy manifests in the formation of fringes. This phenomenon, known as interference, was explained on the wave model of light. What may puzzle you is the fact that light casts shadows of objects, i.e., light appears to travel in straight lines rather than bending around obstacles. This apparent contradiction was explained by Fresnel. You will learn that the ease with which a wave bends around corners is determined by the size of the obstacle relative to wavelength of light. The wavelength of light is about $10^{-7}$m and the obstacles used in ordinary experiments are about $10^5$ times bigger. However, a large number of obstacles, whose sizes are comparable to the wavelength of light, do exhibit diffraction of light.

The phenomenon of diffraction was first observed by Grimaldi and a systematic explanation is due to Fresnel. According to him, in diffraction phenomenon, interference takes place between secondary wavelets from different parts of the same wavefront. Diffraction is classified in two categories: Fresnel diffraction and Fraunhofer diffraction. For Fresnel diffraction, discussed in Unit 8, the experimental arrangement is fairly simple. The source or the observation screen or both are at a finite distance from the obstacle. But theoretical analysis of Fresnel diffraction, being essentially based on geometrical construction, is somewhat cumbersome. Nevertheless, Fresnel diffraction is more general; it includes Fraunhofer diffraction as a special case.

In Fraunhofer diffraction, the source of light and the observation screen (or human eye) are effectively at infinite distance from the obstacle. The Fraunhofer diffraction from a single slit is of particular interest in respect of the general theory of optical instruments. This is discussed in detail in Unit 9. You will learn that when a narrow vertical slit is illuminated by a distant point source, the diffraction pattern consists of a series of spots along a horizontal line and situated symmetrically about a central spot. For a circular aperture, the diffraction pattern consists of concentric rings with a bright central disc.

In Unit 10 you will learn about double slit and *N*-slit diffraction patterns. A distinct feature of double slit pattern is that it consists of bright and dark fringes similar to those observed in interference experiments. The *N*-slit diffraction pattern shows will-defined interference maximum. The sharpness of interference maximum increases as *N* increases. For a sufficiently large value of *N*, interference maxima become narrow lines. This is why diffraction gratings are an excellent tool in spectral analysis.

An important point to learn is that fringed (diffracted) image of a point source is not a geometrical point. And diffraction places an upper limit on the ability of optical devices to transmit perfect information about any object. That is, all optical systems are diffraction limited. In Unit 11 you will learn to characterise the ability of an optical instrument to distinguish two close but distinct diffraction images of two objects or wavelengths based on Rayleigh criterion.

**Acknowledgements**

We thank Shri Sunder Singh and Gopal Krishna Arora for secretarial assistance.

# UNIT 8 FRESNEL DIFFRACTION

## Structure

## 8.1 INTRODUCTION

We know from our day-to-day experience that we can hear persons talking in an adjoining room whose door is open. This is due to the ability of sound waves to bend around the corners of obstacles in their way. You are also familiar with the ability of water waves to propagate around obstacles. You may now ask: Does light, which is an electromagnetic wave, also bend around corners of obstacles in its path? In the previous block you have learnt menifestation of wave nature of light in the form of interference: Light from two coherent sources interferes to form fringed pattern. But what may puzzle you is the fact that light casts shadows of objects, i.e. appears to travel in straight lines rather than bending around corners. This apparent contradiction was explained by Fresnel who showed that the ease with which a wave bends around corners is strongly influenced by the size of the obstacle (aperture) relative to its wavelength. Music and speech wavelengths lie in the range 1.7 cm to 17m. A door is about 1 m aperture so that long wavelength waves bend more readily around the door way. On the other hand, wavelength of light is about $10^{-7}$ m and the obstacles used in ordinary experiments are about $10^5$ times bigger. For this reason, light appears to travel along straight lines and casts shadows of objects instead of bending around their corners. However, it does not mean that light shows no bending, it does so under suitable conditions where size of obstacles is comparable with the wavelength of light. You can get a feel for this by closely examining shadows cast by objects. You will observe that the edges of shadows are not sharp. **The deviation of waves from their original direction due to an obstruction in their path is called diffraction.**

You may have seen TV tower in Delhi. It 235m high and almost three times taller than Qutub Minar. Have you ever thought: Why TV transmission is beamed from a height? The TV transmission involves short wavelength signals, $\lambda \sim 1$cm. These are bilcked by hills, buildings and the curvature of Earth. It is only to avoid blockage that TV signals are transmitted from high towers. The radio signals are reflected by the ionospheric layers before reaching us. In contrast to this, the TV signals, which are microwaves, do not get reflected by the ionosphere. Their transmission takes place along the line of sight. To get the TV signals transmitted over long distances, geostationary satellites are employed, which when placed at suitable height, reflect these signals.

The phenomenon of diffraction finds great use in our daily life. The music from Vividh Bharti – an AM station — comes via long waves ($f$ ~500 kHz – 20MHz and hence λ from $10^3$m to 10m). You will learn that diffraction places a fundamental restriction on optical instruments including human eye, in respect of resolution of objects. In this block you will learn a lot of good physics involved in diffraction limited system.

The phenomenon of diffraction was first observed by Grimaldi, an Italian mathematician. And a systematic explanation of diffraction was given by Fresnel on

Fraunhofer diffraction and fresnel diffraction are also called far field diffraction and near field diffraction, respectively.

the basis of Huygens' principle. According to him, diffraction is attributed to mutual interference of secondary wavelets from a single wave. (The interference phenomenon involves two coherent wave trains.) This means that in diffraction phenomenon, interference takes place between secondary wavelets from different parts of the same wavefront.

For mathematical convenience and ease in understanding, diffraction is classified in two categories: Fraunhofer diffraction and Fresnel diffraction. In Fraunhofer class of diffraction, the source of light and the observation screen (or human eye) are effectively at infinite distance from the obstacle. This can be done most conveniently using suitable lenses. It is of particular practical importance in respect of the general theory of optical instruments. You will learn about it in the next unit.

In Fresnel class of diffraction, the source or the observation screen or both are at finite distance from the obstacle. You will recognise that for Fresnel diffraction, the experimental arrangement is fairly simple. But its theoretical analysis is more difficult than that of Fraunhofer diffraction. Also, Fresnel diffraction is more general; it includes Fraunhofer diffraction as a special case. Moreover, it has importance in historical perspective in that it led to the development of wave model of light. You will learn some of these details in this unit.

You may be aware of the preliminaries of diffraction phenomenon from your school physics curriculum. Or you may have opted PHE-02 course on Oscillations and Waves. In whatever situation you are placed, you should refresh your knowledge.

### Objectives

After studying this unit you will be able to

- state simple experiments which illustrate diffraction phenomenon
- describe an experimental set-up for diffraction at a circular aperture
- explain that Fraunhofer diffraction is a special case of Fresnel diffraction
- discuss the concept of Fresnel half-period zones and apply it to zone plate
- discuss diffraction pattern due to a circular aperture and a straight edge, and
- solve numerical problems.

## 8.2 OBSERVING DIFFRACTION: SOME SIMPLE EXPERIMENTS

As you know, the wavelength of visible light is very small (about $10^{-7}$ m). And to see diffraction, careful observations have to be made. We will now familiarise you with some simple situations and experiments to observe diffraction of light. The prerequisits for these are: (i) a source of light, preferably narrow and monochromatic, (ii) a sharp edged obstacle and (iii) an observation screen, which could be human retina as well.

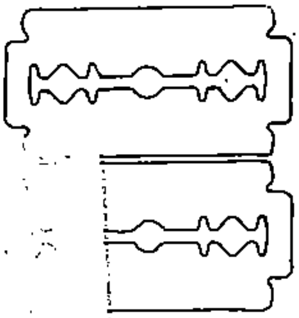
1. Look at a distant street light at night and squint. The light appears to streak out from the bulb. This is because light has bent around the corners of your eyelids.

2. Stand in a dark room and look at a distant light bulb in another room. Now move slowly until the doorway blocks half of the light bulb. The light appears to streak out into the umbra region of the dark room due to diffraction around the doorway.

3. Take a piece of fine cloth, say fine handkerchief or muslin cloth. Stretch it flat and keep it close to the eye. Now focus your eye on a distant lamp (atleast 100 m away) through it. Do you observe an enlarged disc surrounded by a regular

pattern of spots arranged along a rectangle? On careful examination you will note that the spots on the outer part of the pattern appear coloured. Now rotate the handkerchief in its own plane. Does the pattern rotate? You will be excited to see that the pattern rotates about the central disc. Moreover, the speed of rotation of the pattern is same as that of the handkerchief.

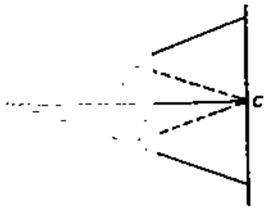
We are now tempted to ask: Do you know why this pattern of spots is obtained? You will agree that the handkerchief is a mesh (criss-cross) of fine threads in mutually perpendicular directions. Obviously, the observed pattern



bserving diffraction
ir of razor blades

s a member of the which was appointed snel's dissertation. Fresnel, and hence , Poisson argued l bright spot should e shadow of a tacle. His logic, tio ad absurdum ws: Consider the perfectly round cast by a point own below.

wave theory, all the periphery will This is because they the same distance ce. So the waves the rim *PP'* and ould all be in entre of the implies that there right spot at the hadow. This was surd by Poisson. tely not aware that in question had liscovered by t a century ago. sson's objection, out the ng a disk of 2mm is surprise, he ie central bright



e distance (about 2 m) from ... was mounted on a movable stand so that its distance from the obstacle could ... ed. They used steel ball bearings of radii 1.58 mm, 1.98 mm, 2.37 mm and 3.17 mm as spherical obstacles. (As such, you should not attach much significance to the exactness of these sizes.) These four spheres were mounted on a glass plate, which was kept at a distance of about 2 m from the pinhole.

The photographic plate was kept at distances of 5cm, 10cm, 20cm, 40cm and 180cm from the mounted glass plate (obstacle). For the last case, the diffraction patterns obtained from these spheres are shown in Fig. 8.3 (a). These patterns essentially characterize the distribution of light intensity in the region of geometrical shadow of the obstacles.

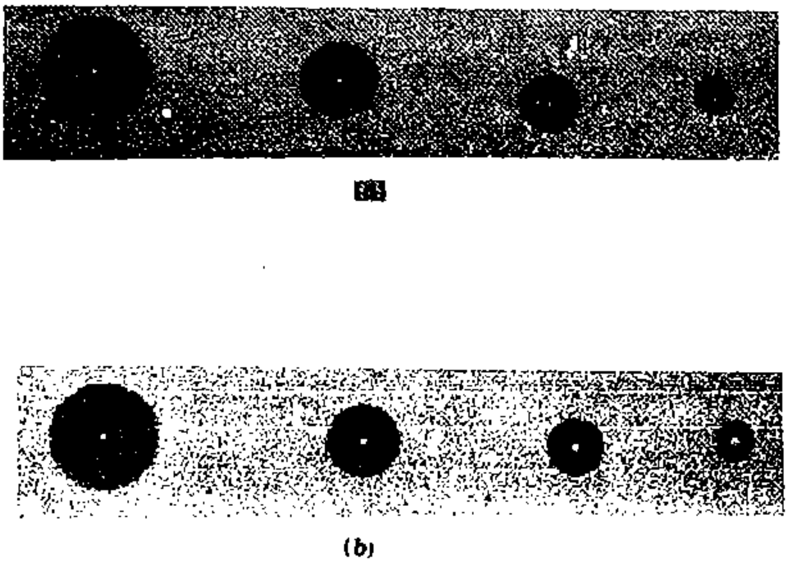


Fig. 8.3: Fresnel diffraction patterns: Kathvate experiments with (a) spheres and (b) circular discs of four sizes

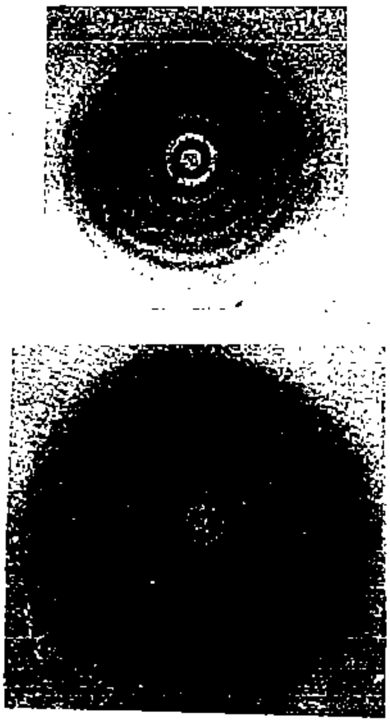


Fig. 8.4: Enlarged view of fringe pattern for the sphere of radius 1.58mm

The diffraction patterns for circular discs of the same size are illustrated in Fig.8.3(b).You will find that these patterns are almost similar to those for spheres. Moreover, the diffraction patterns on the left half of this figure, which correspond to bigger spheres and discs (radii 3.17mm and 2.37mm), show the geometrical shadow and a central bright spot within it. On the other hand, in the diffraction pattern corresponding to the smaller sphere (or disc) of radius 1.98mm, the geometrical image is recognizable but has fringes appearing on its edges. The fringe pattern around the central spot becomes markedly clearer for the sphere of radius 1.58mm. An enlarged view of this pattern is shown in Fig. 8.4. The formation of the bright central spot in the shadow and the rings around the central spot are the most definite indicators of non-rectilinear propagation of light. Instead, light bends in some special way around opaque obstacles. These departures from rectilinear propagation come under the heading of diffraction phenomenon.

Let us pause for a minute and ask: Are these diffraction patterns unique for a given source and obstacle? The answer to this question is: Fresnel patterns vary with the distance of the source and screen from the obstacle. Let us now learn how this transition evolves.

### 8.3.1 Spatial Evolution of a Diffraction Pattern:Transition from Fresnel to Fraunhofer Class

To observe transition in the Fresnel diffraction pattern with distance, we have to introduce a small modification in Kathvate's experimental arrangement, as shown in Fig. 8.5 (a). The point source is now located at the focal point of a converging lens *L*. The spherical waves originating from the source *O* are changed into plane waves by this lens and the wavefront is now parallel to the diffracting screen with a narrow opening in the form of a long narrow slit (Fig.8.5 (b)). These waves pass rough the slit. The diffracted waves are also plane and may have an angular read. You may now like to know the shape, size and intensity distribution in the ffraction pattern on a distant screen.

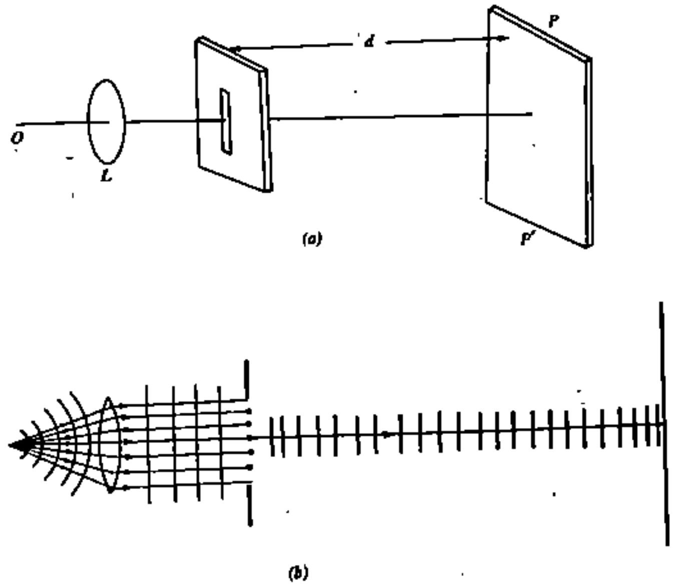


Fig. 8.5(a): Arrangement to observe transition in Fesnel diffraction pattern (b) Cross-sectional view of the geometry shown in (a) above.

When the incident wavefront is strictly parallel to the diffracting screen, we get a vertical patch of light when the screen is immediately behind the aperture. That is, a region $A'B'$ of uniform illumination on the screen. The size of this region is equal to the size of the slit both in width and height. The remaining portion of the screen is absolutely dark. A plot of this intensity distribution is shown in Fig. 8.6 (a). From $P$ to $A'$, the intensity is zero. At $A'$, it abruptly rises to $I_0$, and remains constant from $A'$ to $B'$. At $B'$, it again drops to zero. We can say that $A'B'$ represents the edges of the geometrical shadow (and the law of rectilinear propagation holds).

A slit is a rectangular opening whose width (0.1mm or so) is much smaller than its length 1 cm or more.

2. As the screen is moved away from the aperture, a careful observation shows that the patch of light seen in (1) above begins to lose sharpness. If the distance between the obstacle and the observation screen is large compared to the width of the slit, some fringes start appearing at the edges of the patch of light. But this patch resembles the shape of the slit. The intensity distribution shows diffraction rippling effect somewhat like that shown in Fig. 8.6(b). From this we can say that the intensity distribution in the pattern depends on the distance at which the observation screen is placed.

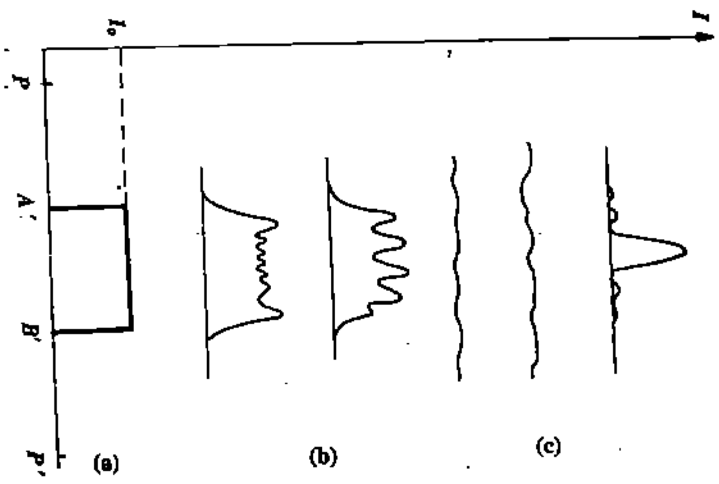


Fig. 8.6: Spatial evolution of a diffraction pattern

3. When $d$ (~1m) is much greater than the width of the slit (~0.1 mm), the fringes seen in (2) above — close to edge of the patch — now spread out and the geometrical image of the slit can no longer be recognized. As distance is increased further, diffraction effects become progressively more pronounced.

4. When $d$ is very large, i.e. once we have moved into the Fraunhofer region, ripples no longer change character. You can observe this pattern by putting a convex lens after the slit. The observation screen should be arranged so that it is at the second focal plane of the lens. These variations in Fraunhofer diffraction are shown in Fig. 8.6(c).

From this we may conclude that Fresnel diffraction can change significantly as the distance from the aperture is varied.

You must now be interested to understand physical basis of these observations . The first systematic effort in this direction was made by Fresnel. Let us learn about it now.

## 8.4 FRESNEL CONSTRUCTION

Let us consider a plane wave front represented by $WW'$ propagating towards the right, as shown in Fig. 8.7(a). We want to calculate the effect of this plane wavefront at an external point $P_0$ on the screen at a distance $d$. Then we will introduce an obstacle like a straight edge and see how intensity at $P_0$ changes.

We know that every point on the plane wavefront may be thought of as a source of secondary wavelets. We wish to compute the resultant effect at $P_0$ by applying Huygens-Fresnel principle. One way would be to write down the equations of vibrations at $P_0$ due to each wavelet and then add them together. This is a cumbersome proposition. The difficulty in mathematical calculation arises on two counts: (i) There are an infinite number of points which act as sources of secondary wavelets and (ii) Since the distance travelled by the secondary wavelets arriving at $P_0$is different, they reach the point $P_0$ with different phases. To get over these difficulties, Fresnel devised a simple geometrical method which provided useful insight and beautiful explanation of diffraction phenomenon from small obstacles. He argued that it is possible to locate a series of points situated at the same distance from $P_0$ so that all the secondary wavelets originating from them travel the same distance. We can, in particular, find the locus of those points from where the wavelets travel a distance $b + \frac{\lambda}{2}, b + \frac{2\lambda}{2}, b + \frac{3\lambda}{2}$, and so on.

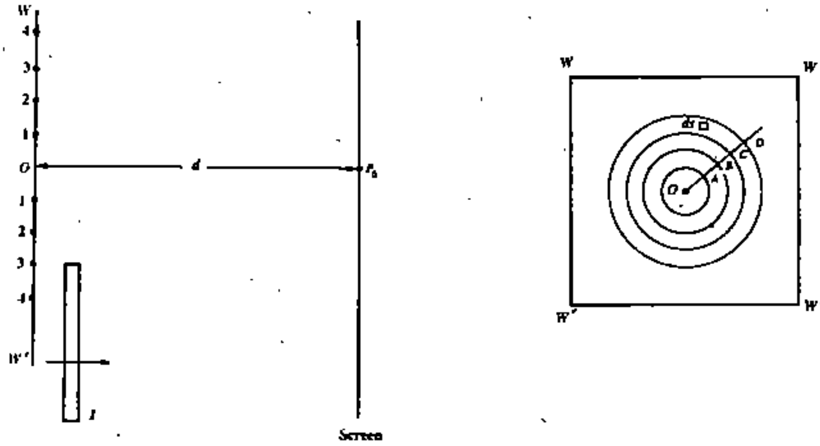


Fig. 8.7: Fresnel construction (a) Propagation of a plane wavefront and (b) division of wavefront into annular spaces enclosed by concentric circles

he Fresnel construction consists of dividing the wavefront into annular spaces iclosed by concentric circles (Fig. 8.7(b)). The net effect at $P_0$ will be obtained by umming contributions of wavelets from these annular spaces, called **half period lements**. When an obstacle is inserted in between the wavefront $WW'$ and the oint $P_0$, some of these half period elements will be obstructed depending upon the ze and shape of the obstacle. The wavelets from the unobstructed parts only will each $P_0$ and their resultant can be calculated easily by Fresnel's method. Let us ow learn about Fresnel's construction, half period elements and the method of ummation of the contributions of secondary wavelets.

## .4.1 Half Period Elements

'o discuss the concept of Fresnel's half-period elements we assume, for simplicity, ıat light comes from infinity so that the wavefront passing through the aperture is lane. Refer to Fig. 8.8. It shows a plane wavefront $WW'F'F$ of monochromatic ight propagating along the $z$-direction. We wish to calculate the resultant amplitude f the field at an arbitrary point $P_0$ due to superposition of all the secondary luygens' wavelets originating from the wavefront at the aperture. To do so, we livide the wavefront into half-period zones using the following construction: From the point $P_0$ we drop a perpendicular $P_0O$ on the wavefront, which cuts it at $O$. The point $O$ is called the pole of the wavefront with respect to the point $P_0$. Suppose that $b$ is the distance between the foot of the perpendicular to $P_0$, i.e. $OP_0 = b$. Now with $P_0$ as centre, we draw spheres of radii $b + \frac{\lambda}{2}$, $b + \frac{2\lambda}{2}$, $b + \frac{3\lambda}{2}$, and so on. You can easily visualise that these spheres will intersect the plane wavefront in a series of concentric circles with centre $O$ and radii $OQ_1$, $OQ_2$, $OQ_3$, ... as shown in Fig. 8.8. This geometrical construction divides the wavefront into circular strips called zones. The first zone is the space enclosed by the circle of radius $OQ_1$, the second zone is the annular space between the circles of radii $OQ_2$ and $OQ_1$. The third zone is annular space between the circles of radii $OQ_3$ and $OQ_2$ and so on. These concentric circles or annular rings are called **Fresnel zones** or **half period elements**. This nomenclature has genesis in the fact that the path difference between the wavelets reaching $P_0$ from corresponding points in successive zones is $\lambda/2$.

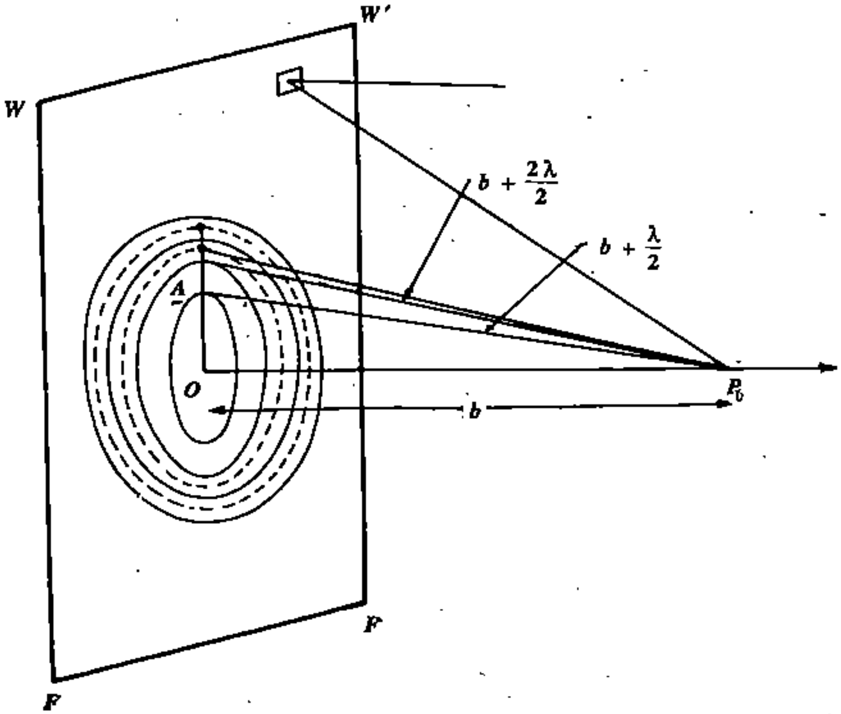


Fig. 8.8: Half-period zones on a plane wavefront: A schematic construction

To compute the resultant amplitude at $P_0$ due to all the secondary wavelets. emanating from the entire wavefront, we first consider an infinitesimal area $dS$ of the wayefront. We assure that the amplitude at $P_0$ due to $dS$ is (i) directly proportional to the area $dS$ since it determines the number of secondary wavelets, (ii) inversely proportional to the distance of $dS$ from $P_0$ and (iii) directly proportional to the obliquity factor $(1 + \cos\theta)$, where $\theta$ is the angle between the normal drawn to the wavefront at dS and the line joining dS to $P_0$. $\theta$ is zero for the central point $O$. As we go away from $O$, the value of $\theta$ increases until it becomes 90° for a point at infinite distance on the wavefront (Fig. 8.9). Physically, it ensures that wavefront moves forward. That is, there is no reverse (or backward) wave.

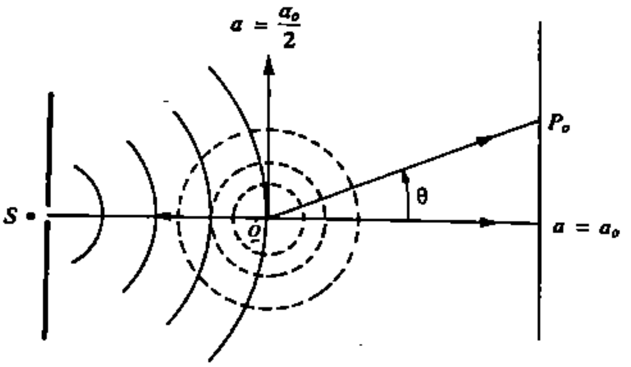


Fig. 8.9: The obliquity factor for Huygens' secondary wavelets

If we denote the resultant amplitudes at $P_0$ due to the first, second, third, fourth, ..., nth zone by $a_1, a_2, a_3, a_4, \ldots$, then we can write

$$a_n = \text{Const} \times \frac{A_n}{b_n}(1 + \cos\theta) \tag{8.1}$$

where $A_n$ is the area of the $n$th zone and $b_n$ is the average distance of the $n$th zone from $P_0$.

Eq. (8.1) shows that to know the amplitude of secondary wavelets arriving at $P_0$ from any zone, we must know $A_n$. This in turn requires knowledge of the radii of the circles defining the boundaries of the Fresnel zones. To calculate the radii of various half period zones in terms of known distances, let us denote $OQ_1 = r_1$, $OQ_2 = r_2$, $OQ_3 = r_3, \ldots, OO_n = r_n$. From pythagoras' theorem we find that the radius of the first circle (zone) is given by

$$r_1 = \left[ (b + \frac{\lambda}{2})^2 - b^2 \right]^{1/2} = \sqrt{b\lambda + \frac{1}{4}\lambda^2}$$

$$\cong \sqrt{b\lambda}$$

The approximation $\lambda << b$ holds for practical systems using visible light. Similarly, the radius of the $n$th circle (zone) is given by

$$r_n = \left[ (b + \frac{n}{2}\lambda)^2 - b^2 \right]^{1/2}$$

$$= \left[ nb\lambda + \frac{n^2\lambda^2}{4} \right]^{1/2}$$

$$\cong \sqrt{nb\lambda} \tag{8.2}$$

where we have neglected the term $\frac{n^2\lambda^2}{4}$ in comparison to $nb\lambda$. This approximation holds for all diffraction problems of interest to us here.

It readily follows from Eqs. (8.1) and (8.2) that the radii of the circles are proportional to the square root of natural numbers, i.e. $\sqrt{1}, \sqrt{2}, \sqrt{3}, \sqrt{4}, \ldots$ Therefore, if the first zone has radius $r_1$, the successive zones have radii 1.41 $r_1$, 1.73 $r_1$, 2 $r_1$and so on. For He-Ne laser light ( $\lambda$ = 6328 Å ), if we take $P_0$ to be 30 cm away ($b$ = 30 cm), the radius of the first zone will 0.436 mm.

Let us now calculate the area of each of the half-period zones. For the first zone

$$A_1 = \pi r_1^2 = \pi [ ( b + \frac{\lambda}{2} )^2 - b^2 ]$$

$$= \pi b \lambda + \frac{\pi}{4}\lambda^2$$

$$\cong \pi b \lambda \tag{8.3a}$$

The area of the second zone, i.e. the annular region between the first and the second circles is

$$\pi\left(r_2^2 - r_1^2\right) = \pi[ ( b + \lambda )^2 - b^2 ] - \pi b \lambda$$

$$\cong 2\pi b \lambda - \pi b \lambda = \pi b \lambda \tag{8.3b}$$

Similarly, you can readily verify that the area of the $n$th zone

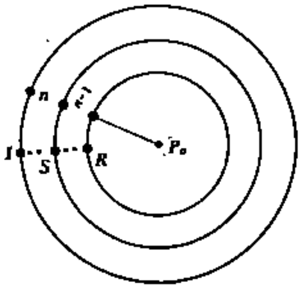
$$A_n = \pi ( r_n^2 - r_{n-1}^2 ) \cong \pi b \lambda \tag{8.3c}$$

That is, all individual zones have the same area. The physical implication of the equality of zone areas is that the secondary wavelets starting from every zone will be very nearly equal. You must however remember that the result contained in Eq.(8.3) is approximate and is valid for cases where $b >> n\lambda$. A more regorous calculation shows that the area of a zone gradually increases with $n$:

$$A_n = \pi \lambda [ b + ( n - 1 ) \frac{\lambda}{2} ] \tag{8.3d}$$

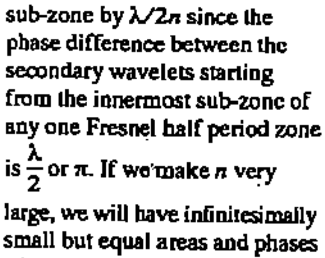
However, the effect of this increase is almost balanced by the increase in the average distance of the $n$th zone from $P_0$. That is, the ratio $A_n/b_n$ in Eq. (8.1) remains $\pi\lambda$, which is a constant, independent of $n$. This means that the amplitude due to any zone will be influenced by the obliquity factor, which is actually responsible for monotonic decrease in the amplitudes of higher zones ( $a_1 > a_2 > a_3, \ldots > a_n$ ). Also, it is important for our computation to note that consecutive zones differ by one-half of a wavelength. Therefore, the secondary waves from any two corresponding points in successive zones [$n$th and ($n$ -1)th or ($n$ + 1)th] reach $P_0$ out of phase by $\pi$ or half of a period.

Suppose that the contribution of all the secondary wavelets in the $n$th zone at $P_0$ is denoted by $a_n$. Then, the contribution of ($n$-1)th zone $a_{n-1}$, will tend to annihilate the effect of $n$th zone. Mathematically we write the resultant amplitude at $P_0$ due to the whole wavefront as a sum of an infinite series whose terms are alternately positive and negative but the magnitude of successive terms gradually diminishes:



Refer to figure above and consider the contributions from the ( $n$ – 1 )th and $n$th zones. Firstly, the areas of the two annular regions are approximately equal, i.e., the secondary wavelets starting from both the zones are equal. Secondly, the points on the innermost circle of ( $n$ – 1 )th zone e.g. points like $R$ are situated at a distance of $d + ( n - 2 ) \lambda/2$ from $P_0$ whereas the points on the innermost circle of $n$th zone e.g. points like $S$ are situated at a distance of $d + ( n - 1 ) \lambda/2$ from $P_0$. The difference in path between the secondary wavelets to reach $P_0$ from $R$ and $S$ is $\lambda/2$. This means that the waves reaching $P_0$ are out of phase by $\pi$ and cancel each other. Similarly for every point between $R$ and $S$ in the ( $n$ – 1 )th zone we have a corresponding point between $S$ and $T$ in the $n$th zone with a path difference of $\lambda/2$ or phase difference of $\pi$ and hence cancel each other. Since the areas of the two zones are approximately equal, we arrive at the result that **for every point in the ( $n$ – 1 )th zone we have a point in the $n$th zone which is out of phase by $\pi$ or half of a period.**

So far we have considered the effect of a whole number of half period elements at a given point. The sum of the amplitudes due to all the secondary wavelets starting from the $n$th zone was represented by $a_n$. But so far we have not computed the magnitude and phase of the amplitude vector $a_n$. An obvious related probem is to calculate the effect at $P_0$ due to a fraction of agiven half period element. We can do this easily by the following vector summation method. We divide a Fresnel zone into a series of $n$ sub-zones of equal areas. Refer to figure below. It shows such a division for the annular space between $(n-1)$th and $n$th circles. $O$ is taken as centre and circles of slightly differing radi have been drawn such that the annular space between two consecutive circles encloses equal area. Now within the area covered by a sub-zone, we can neglect variation in inclination factor. Since all these sub-zones have been drawn so that they have equal areas, the amplitude at $P_0$ due to these small equal areas will be the same. But the phases will change continuously from one sub-zone to the next sub-zone by $\lambda/2n$ since the phase difference between the secondary wavelets starting from the innermost sub-zone of any one Fresnel half period zone is $\frac{\lambda}{2}$ or $\pi$. If we make $n$ very large, we will have infinitesimally small but equal areas and phases of wavelets from these may be taken to vary continuously and uniformly.



Thus we have a set of disturbances of equal amplitude but uniformly changing phase such that the phase difference between the two extreme disturbances is $\pi$. These extreme vectors are represented by $AA'$ and $BB'$ in the figure shown. We know that in such a case the vector diagram is a semicircle and the resultant of the summation of amplitudes is the diameter $AB$.

$$\xi = a_1 + a_2 e^{i\pi} + a_3 e^{i2\pi} + a_4 e^{i3\pi} + \ldots$$

$$= a_1 - a_2 + a_3 - a_4 + \ldots + (-1)^{n+1} a_n + \ldots = \frac{a_1}{2} \quad (8.4)$$

There are several methods of arriving at this result. Here we will describe a simple graphical construction. (The mathematical method is given as TQ). Let us denote the amplitudes of resultant vectors **AB, CD, EF, GH**, ... respectively by $a_1, a_2, a_3, a_4$, ...due to the first, second, third, fourth, ... zone. (We know that $a_1, a_2, {}_3, a_4, \ldots a_n$ are alternately positive and negative). These vectors are shown separately in Fig. 8.10(a) to show their magnitudes and positions. But their true positions are along the same line, as shown in Fig. 8.10 (b). The resultant of the first two zones will be the small vector **AD**. But the resultant of the first three zones is the large vector **AF**; of the four zones the smaller vector **AH** and so on. Refer to Fig. 8.10(a) again. You will note that the resultant of infinitely large number of zones is equal to $a_1/2$.

If we consider a finite number of zones, say $n$, the resultant is given by

$$\xi(n) = \frac{a_1}{2} + \frac{a_n}{2} \quad (8.5)$$

where $n$ is any number (odd or even).

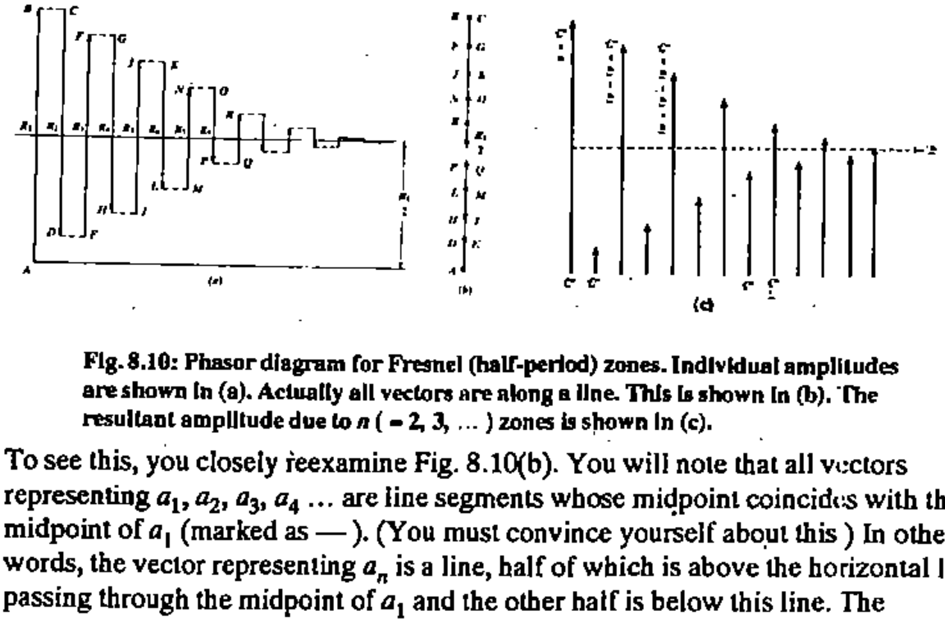


Fig. 8.10: Phasor diagram for Fresnel (half-period) zones. Individual amplitudes are shown in (a). Actually all vectors are along a line. This is shown in (b). The resultant amplitude due to $n$ ( = 2, 3, ... ) zones is shown in (c).

To see this, you closely reexamine Fig. 8.10(b). You will note that all vectors representing $a_1, a_2, a_3, a_4$ ... are line segments whose midpoint coincides with the midpoint of $a_1$ (marked as —). (You must convince yourself about this ) In other words, the vector representing $a_n$ is a line, half of which is above the horizontal line passing through the midpoint of $a_1$ and the other half is below this line. The resultant of $n$ zones is a vector joining $A$ to the end of the vector representing $a_n$. When $n$ is odd, the end point of the vector representing $a_n$ will be above the horizontal line by $a_n/2$, which proves the required result.

If $n$ is even, the end point will be below this horizontal line by $a_n/2$. Added vectorially, we have the same result. We thus see that the resultant amplitude at $P_0$ due to $n$ zones is half the sum of amplitudes contributed by the first and the last zone. $\xi$ will be numerically greater than $a_1/2$ when $n$ is odd and smaller than $a_1/2$ when $n$ is even. For example, the resultant contribution due to 7 zones is $AO$, which is equal to $\frac{a_1}{2} + \frac{NO}{2}$. On the other hand, for 8 zones, the resultant is $AQ = \frac{a_1}{2} - \frac{PQ}{2}$.

It may be emphasized that in this graphical method of summation of the series, we have used three properties: (i) vectors representing $a_1, a_2$, ... are all along the same straight line (ii) alternate vectors are oppositely directed and (iii) the magnitudes of $a_1, a_2, \ldots$ decrease gradually.

We now consider a simple example to illustrate these concepts.

### Example 1

Consider a series with $n = 100$ in which each term is equal to the arithmetic mean of the preceding and the following terms. Calculate the resultant.

### Solution

As a special case, we can take the terms of the series as 100, 99, 98, … 3, 2, 1.

$$\therefore\ \xi = (100 - 99) + (98 - 97) + (96 - 95) + \ldots (4 - 3) + (2 - 1)$$

$$= 1 + 1 + 1 \ldots 50 \text{ terms}$$

$$= 50$$

which is half of the first term. Now consider the relation

$$\xi = \frac{a_1}{2} + \frac{a_n}{2}.$$

and take different number of terms in this arithmatic series. If we have only one term, $a_1 = 100$ we take the first term as 100 and also the last term as 100. Then we get

$$\xi = \frac{a_1}{2} + \frac{a_n}{2} = 100$$

Next we take two terms. Then

$$\xi = (100 - 99) = 1$$

Also

$$\frac{a_1}{2} + \frac{a_n}{2} = \frac{100}{2} - \frac{99}{2}$$

$$= 50 - 49.5 = 0.5$$

For three terms, $\xi = (100 - 99) + 98 = 99$

and $\dfrac{a_1}{2} + \dfrac{a_3}{2} = 50 + 49 = 99$

For four terms, $\xi = (100 - 99) + (98 - 97) = 2$

and $\dfrac{a_1}{2} + \dfrac{a_4}{2} = 50 - 48.5 = 1.5$

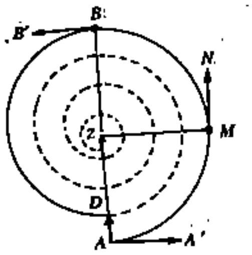
For five terms $\xi = (100 - 99) + (98 - 97) + 96 = 98$

and $\dfrac{a_1}{2} + \dfrac{a_5}{2} = 50 + 48 = 98$

For six terms $\xi = (100 - 99) + (98 - 97) + (96 - 95) = 3$

and $\dfrac{a_1}{2} + \dfrac{a_6}{2} = 50 - 47.5 = 2.5$

and so on. Thus we see that $\xi$ is given by $\dfrac{a_1}{2} + \dfrac{a_n}{2}$ to a fairly good degree of accuracy.



Now we will compute magnitude and phase of the resultant **AB**. If all the equal disturbances from the sub-zones were in the same phase, the resultant would have been a line along $AA'$ and equal to the length of the arc of the semicircle $AB$ ( $= \pi r$ ) of radius $r$. But we find that the actual resultant amplitude is $AB = 2r$. Thus the resultant amplitude is $\frac{2r}{\pi r} = \frac{2}{\pi}$ times the value which would be obtained if all the wavelets within a Fresnel half period element had the same phase. Since the line $AB$ is parallel to the line $MN$, we see that the resultant phase of vector **AB** is the same as that of the vector **MN** representing the disturbance starting from the middle point ($M$) of the zone. In other words, AB is peerpendicular to $AA'$. That is, it is a quarter-period behind the wavelet starting from the innermost sub-zone. We can find, in a similar manner, the resultant contribution due to the next half-period zone. It is given by $CD$ and differs from $AB$ by $\pi$. The resultant of the sum of thses two zones is the small vector **AD**. The magnitudes of vectors and their phases for succeeding zones are shown in figure below. The resultant curve is the vibration spiral with gradually smaller and smaller semicircles until eventually it coincides with Z. The resultant when all the half-period elements are considered is $AZ$ which is half of that which would be produced by the first zone alone. It is equal to $\frac{1}{2} \times \frac{2}{\pi} = \frac{1}{\pi}$ times that which would be produced by all the wavelets from the first zone acting together in the same phase.

### 8.4.2 Rectilinear Propagation

Refer to Fig. 8.11. It shows several collinear apertures $A, B, C, \ldots$ Light originates from a point source and propagates towards the right. Suppose that the source is 1m away. We may take the spherical wave falling on the obstacle as nearly a plane wave. (The radius of curvature of the incident spherical wave will not qualitatively change the argument.) Let us work out the sizes of Fresnel half period elements for the typical case where the screen is 30 cm away from the aperture. Taking $\lambda = 5 \times 10^{-5}$ cm, we get $r_1 = \sqrt{(30\text{ cm})}\sqrt{\times\ (5 \times 10^{-5}\text{ cm})}$ $= 3.87 \times 10^{-2}$ cm. This means that the diameter of the first zone is less than 1 mm. Let us consider the 100th zone. Its radius $r_{100} = \sqrt{30\text{ cm} \times 100 \times 5 \times}$ $\sqrt{10^{-5}\text{ cm}} = 3.87 \times 10^{-1}$ cm so that the diameter will be a little less than 1 cm. Therefore, if the aperture is about 1 cm in diameter, the amplitude at $P_0$ due to the whole wavefront is $\frac{a_1}{2} + \frac{a_{100}}{2}$. $a_{100}$ will be fairly small, so that the intensity is essentially half of that due to the first half period zone, which is the intensity expected at $P_0$ when the aperture is completely removed. We may say that light travels to $P_0$ from a region nearly 0.4 mm in radius around $O$. That is, **light travels in a straight line.**

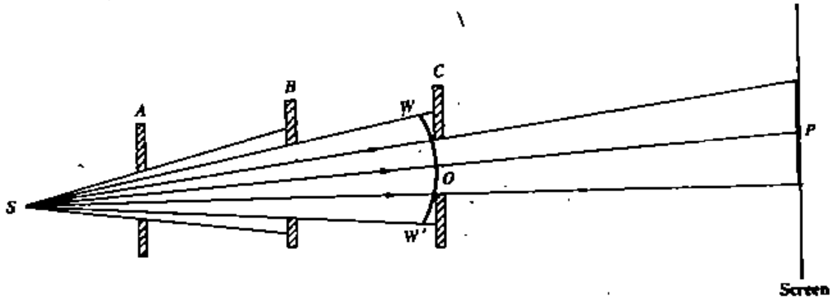


Fig. 8.11: Fresnel construction and rectilinear propagation of light

Let us now understand the formation of shadows and illuminated regions due to an obstacle (Fig. 8.12). Consider the point $P_2$ *whose pole is* $O_2$. If the distance between $O_2$ and the edge $A$ of the obstacle is nearly 1 cm, over 100 half period elements will be accomodated in it. And as seen above, the intensity at $P_2$ will be nearly equal to $\frac{a_1}{2}$. *In other words, the obstacle* $T$ will have no effect at the point $P_2$. Similarly, at $P_1$, which is taken 1 cm inside the geometrical edge of the shadow, over 100 half period elements around O1 are obstructed and the intensity at $P_1$ will be less than $\frac{a_{100}}{2}$, which is almost negligible. This implies almost complete darkness at $P_1$. In other words, the obstacle has completely obstructed the light from the source and

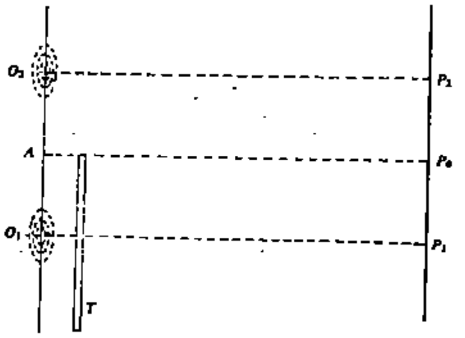


Fig. 8.12: Fresnel construction and formation of shadows/illuminated regions

ıe region around point $P_1$ is in the shadow. Only around $P_0$, which signifies the eometrical edge of the shadow, we find fluctuations in intensity depending upon ow many half period elements have been allowed to pass or have been obstructed. 'his explains the observed rectilinear propagation of light since Fresnel zones are bstructed or allowed through by obstacles of the size of a few mm for these typical istances.

ı special optical device, designed to obstruct light from alternate half- period lements is known as **Zone plate.** It provides experimental evidence in favour of 'resnel's theory. Let us learn about it now.

### ;.4.3 The Zone Plate

'he zone plate is a special optical device designed to block light from every other ıalf-period zone. You can easily make a zone plate by drawing concentric circles ɔn a white paper, with their radii proportional to the square roots of natural numbers ınd shading alternate zones. Fig. 8.13 shows two zone plates of several Fresnel ;ones, where all even numbered or odd numbered zones are blacked out. Now ɔhotograph these pictures. The photographic transparency (negative) in reduced size acts as a Fresnel zone plate. (Recently, Gabor has proposed a zone plate in which zones change transmission according to a sinusoidal wave.) Lord Rayleigh made the first zone plate in 1871. Today zone plates are used to form images using X-rays and microwaves for which conventional lenses do not work.

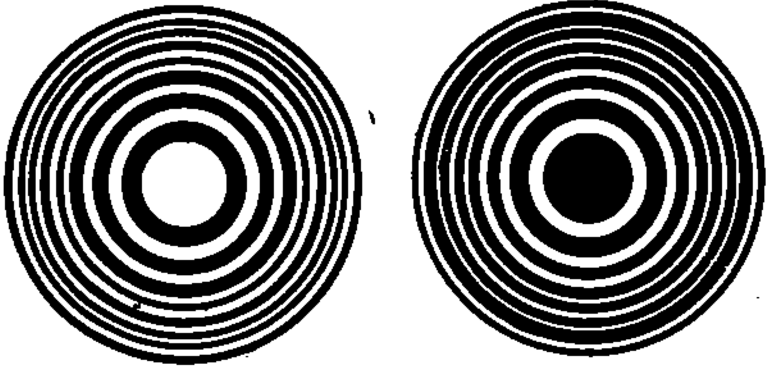


Fig. 8.13: Zone plates: (a) positive (b) negative

If you now pause for a while and logically reflect upon the possible properties of a Fresnel zone plate, you will arrive at the following conclusions:

1. A zone plate acts like a converging lens (see Example 2) and produces a very bright spot. To understand the formation of the spot let us suppose that the first ten odd zones are exposed to light. Then, Eq.(8.4) tells us that the resultant amplitude at $P_0$ is given by

$$\xi_{20} = a_1 + a_3 + a_5 + \dots + a_{19} \qquad (8.5)$$

If obliquity factor is not important, we may write $\xi_{20} = 10a_1$, which means that the amplitude for an aperture containing 20 zones is twenty times and intensity is 400 times that due to a completely unobstructed wavefront.

**Example 2**

Show that a zone plate acts like a converging lens.

**Solution**

Refer to Fig. 8.14. It shows the section of the zone plate perpendicular to the plane of the paper. $S$ is a point source of light at a distance $u$ from the zone plate. A bright image is formed at $P_0$ at a distance $v$ from the plane of the zone plate.

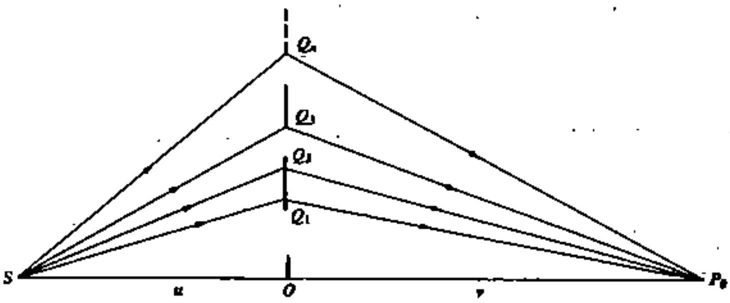


**Fig. 8.14: Action of a Zone Plate as a converging lens**

You can easily write

$$SQ_1 + Q_1P_0 = u + v + \frac{\lambda}{2}$$

$$SQ_2 + Q_2P_0 = u + v + \frac{2\lambda}{2}$$

$$SQ_n + Q_nP_0 = u + v + \frac{n\lambda}{2}$$

By Pythagoras' theorem we can write

$$SQ_n = \sqrt{SO^2 + OQ_n^2}$$

$$= \sqrt{u^2 + r_n^2} = u + \frac{r_n^2}{2u} + \ldots$$

where $r_n$ is the radius of the nth zone.

Similarly, you can convince yourself that

$$Q_nP_0 = v + \frac{r_n^2}{2v} + \ldots$$

If $r_n << u$ or $v$, we can ignore terms higher than $\frac{r_n^2}{2u}$ or $\frac{r_n^2}{2v}$. Hence

$$SQ_n + Q_nP_0 = u + \frac{r_n^2}{2u} + v + \frac{r_n^2}{2v} = u + v + \frac{n\lambda}{2}$$

If we identify $f_n = \frac{r_n^2}{n\lambda}$ as the focal length of the zone plate, we find that

$$\frac{1}{u} + \frac{1}{v} = \frac{n\lambda}{r_n} = \frac{1}{f_n}$$

which is identical to the lens equation.

2. The zone plate has several foci. To understand this, we assume that the observation screen is at a distance of one focal length from the diffracting aperture. Then it readily follows from the above example that the most intense (first order) focal point is situated at $f_1 = r_1^2/\lambda$. To give you a feel for numerical values, let us calculate f1 for a zone plate with radii $r_n = 0.1\sqrt{n}$ cm and illuminated by a monochromatic light of wavelength $\lambda = 5500$ Å. You can easily see that

$$f_1 = \frac{r_1^2}{\lambda} = \frac{(0.1\text{ cm})^2}{5500 \times 10^{-8}\text{ cm}} = 182\text{ cm}$$

To locate higher order focal points, we note from Eq. (8.2) that for $r_n$ fixed, $n$ increases as $b$ decreases. Thus for $b = f_1/2$, $n = 2$. That is, as $P_0$ moves towards the zone plate along the axis, the same zonal area of radius $r$ encompasses more half-period zones. At this point, each of the original zones covers two half-period zones and all zones cancel. When $b = f_1/3$, $n = 3$. That is, three zones contribute from the original zone of radius $r_1$. Of these, two cancel out but one is left to contribute. Thus other maximum intensity points along the axis are situated at

$$f_n = \frac{r_1^2}{n\lambda} \qquad \text{for } n \text{ odd} \tag{8.9}$$

For above numerical example, $f_3 = \frac{182}{3}$ cm, $f_5 = \frac{182}{5}$ cm, $f_7 = \frac{182}{7}$ cm and so on. Between any two consecutive foci, there will be dark points.

## 8.5 DIFFRACTION PATTERNS OF SIMPLE OBSTACLES

From Sec. 8.3 you will recall that by utilizing Kathvate's experimental arrangement, the Fresnel diffraction pattern of various apertures and obstacles could be photographed by varying distances between the source, the object and the photographic plate. We will now use results derived in Sec. 8.4 to explain the observed diffraction pattern of simple obstacles like circular aperture and straight edge.

We begin by studying the Fresnel diffraction pattern of a circular aperture.

### 8.5.1 A Circular Aperture

Refer to Fig. 8.15. It shows a sectional view of the experimental arrangement in which a plane wave is incident on a thin metallic sheet with a circular aperture. You will note that the plane of the wavefront is parallel to the plane of the metal plate; both being perpendicular to the plane of the paper.

Let us calculate the intensity at a point $P_0$ lying along the line passing through the centre of the circular aperture and perpendicular to the wavefront. Suppose that the distance between the point $P_0$ and the circular aperture is $b$. As discussed earlier, the intensity at the observation point due to the entire uninterrupted plane wavefront is

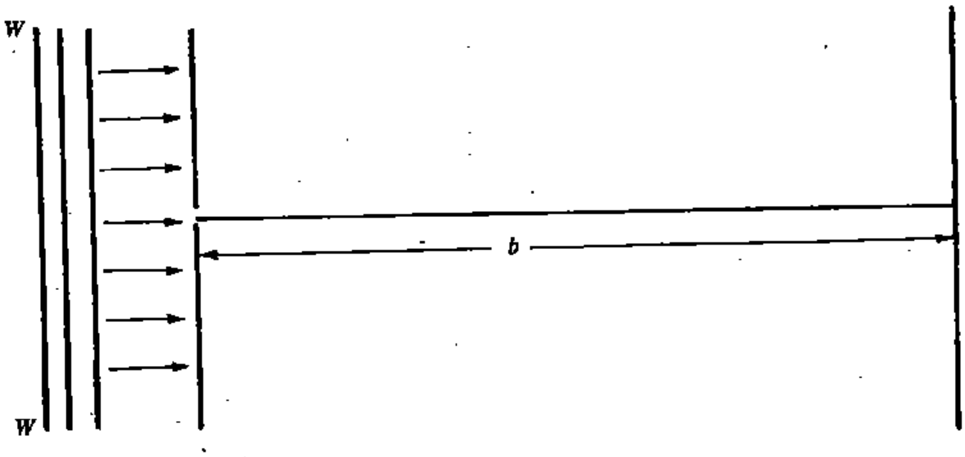


Fig. 8.15: Diffraction by a circular aperture: A cross-sectional view of the experimental arrangement

given by Eq. (8.4) where $a_1$, $a_2$, ... etc. give the contributions due to successive Fresnel zones. Our problem here can be solved by constructing appropriate Fresnel zones and finding out as to how many of these half period elements are transmitted by the aperture. However, it is important to note that for an aperture of a given size, the number of half period elements transmitted may not always be the same. This is because the radii of the Fresnel zones depend upon the distance of point $P_0$ from $O$ ( $r_n = \sqrt{n\lambda b}$ ). You can easily convince yourself that if the point $P_0$ is far away from the aperture ($b$ is very large), the radii of the first zone, equal to $\sqrt{\lambda b}$, may be larger than the radius of the aperture. In such a situation, all the secondary wavelets starting even from the entire first zone alone may not be transmitted. That is, the wavelets from a small portion of the first Fresnel zone only are transmitted.

The next question we have to address to is: How to calculate the amplitude at $P_0$ when the aperture has transmitted only a fraction of the first Fresnel zone? As a first approximation, we assume that the wavelets arrive at $P_0$ in phase. (This is quite justified because the path difference between the extreme wavelets within anyone half period elements is $\lambda/2$. If only a fraction of the first zone transmits here, the net phase difference will be correspondingly less.) Further, the inverse square law for intensity tells us that the amplitude at $P_0$ will be inversely proportional to $b$. Hence, the effect at $P_0$, which is at a large distance, will be small.

As the point $P_0$ moves towards the aperture ($b$ becomes smaller), the zone size shrinks and a greater part of the central zone is transmitted. As a result, the intensity increases gradually. As the observation point comes closer and closer, with the shrinking of the sizes of zones, a stage may reach when the first zone exactly fills the aperture. Then $\sqrt{b\lambda}$ , the radius of the first zone is also the radius of the aperture. We know that the first zone contributes $a_1$ to the amplitude at $P_0$. Compare it with the situation where the obstacle with circular aperture is not present. The entire wavefront contributes but the amplitude at $P_0$ is $\frac{a_1}{2}$. Since intensity is proportional to the square of amplitude, the intensities at $P_0$ with and without the aperture are respectively $a_1^2$ and $\frac{a_1^2}{4}$. That is, the intensity at a given point is four times as large when the aperture is inserted in the path than when it is completely removed. This surprising result is not apparent in the realm of everyday experience dominated by rectilinear propagation of light.

As the observation point $P_0$ comes still closer, the circular aperture may transmit first two zones. The amplitude will then be ( $a_1 - a_2$ ) which is expected to be very small. The additional light produces practically zero amplitude, hence darkness, at $P_0$. Bringing the point $P_0$ gradually closer will cause the intensity to pass through maxima and minima along the axis of the aperture depending on whether the number of zones transmitted is odd or even. If we continue to bring the point $P_0$ closer to $O$, the number of Fresnel zones transmitted by the aperture goes on increasing. The value $\frac{a_1}{2}$ is finally reached when the point $P_0$ is so close that an infinitely large number of zones contribute to the amplitude.

The same variation in intensity should be experienced if the point $P_0$ is kept fixed and the radius of the aperture is varied continuously. This can be done experimentally but is somewhat more difficult.

We have calculated the intensity at points on the axis but the above considerations do not give any information about the intensity at points off the axis. A detailed and complex mathematical analysis which we shall not discuss here, shows that $P_0$ is surrounded by a system of circular diffraction fringes. Photographs of these fringe patterns have been taken by several workers and we referred to Kathvate's experiments earlier in this unit.

We now illustrate the concepts developed here by solving an example.

**Example 3**

In an experiment a big plane metal sheet has a circular aperture of diameter 1 mm. A beam of parallel light of wavelength $\lambda = 5000$ Å is incident upon it normally. The shadow is cast on a screen whose distance can be varied continuously. Calculate the distance at which the aperture will transmit 1,2,3,... Fresnel zones.

## Solution

Let $b_1, b_2, b_3 \ldots b_n$ be the distances at which 1,2,3,... $n$ zones are transmitted by an aperture of fixed radius $r$. From Eq. (8.2) we can write

$$n\, b_n\, \lambda = r_n^2$$

so that

$$b_n = \frac{r_n^2}{n\lambda}$$

$$\therefore \qquad b_1 = \frac{r_1^2}{\lambda} = \frac{(.05\ \text{cm})^2}{5 \times 10^{-5}\ \text{cm}} = 50\ \text{cm}$$

Similarly, we find that

$$b_2 = \frac{r_2^2}{2\lambda} = \frac{50\ \text{cm}}{2} = 25\ \text{cm},\ b_3 = \frac{50}{3}\ \text{cm} = 16.7\ \text{cm},\ b_4 = \frac{50\ \text{cm}}{4} = 12.5\ \text{cm}$$

$$b_5 = 10\ \text{cm},\ b_6 = 8.3\ \text{cm},\ b_7 = 7.1\ \text{cm},\ b_8 = 6.2\ \text{cm}.$$

The amplitudes corresponding to these distances are plotted in Fig. 8.16.

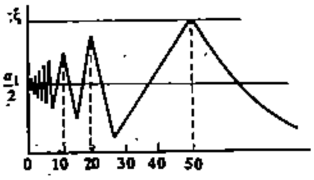


Fig. 8.16: Variation of amplitudes when a circular aperteure transmits integral multiple Fresnel zones

Another conclusion of some historic interest follows if we substitute the aperture by a circular disc or a round obstacle just covering the first Fresnel zone. The light reaching the point of observation $P_0$ will be due to all zones except the first. The second zone is therefore the first contributing zone and the intensity of light spot at the centre of the shadow of the obstacle will be almost equally bright as when the first zone was unobstructed.

You may now ask: Why is the bright spot at the centre only? This is because there is no path difference and hence phase difference between waves reaching an axial point. At any other off-axis point, waves will reach with different phases and may tend to cancel mutually. The existence of this spot was demonstrated by Arago, though Poisson gave his theoretical arguments to disprove wave theory of light.

You may now like to answer an SAQ.

SAQ 1

*Spend 2 min*

A 25 paise coin has a diameter of 2 cm. How many Fresnel zones does it cut off if the screen is 2 m away? Do you expect to see a bright spot at the centre? If we move the screen to a distance of 4 m, how many zones will it cut off? Will the bright spot now look brighter? Why? Take $\lambda = 5 \times 10^{-7}$m.

So far we have discussed diffraction patterns which had axial symmetry: the object or aperture was circular and the plane wavefront originated from a point source. We now wish to consider the case wherein source is a slit source. This source will emit cylindrical waves with the slit as axis. Let us now study the diffraction pattern of a straight edge.

## 8.5.2 A Straight Edge

The slit has a very small width compared to its length. Or we may say that in comparison to its width, it has an infinite length.

Let S be a slit source perpendicular to the plane of the paper. This sends a cylindrical wavefront towards the obstacle which is a straight edge perpendicular to the paper. You can take a thin metal sheet or a razor blade with the sharp edge parallel to the slit: Fig. 8.17(a) shows a section perpendicular to the length of the slit.

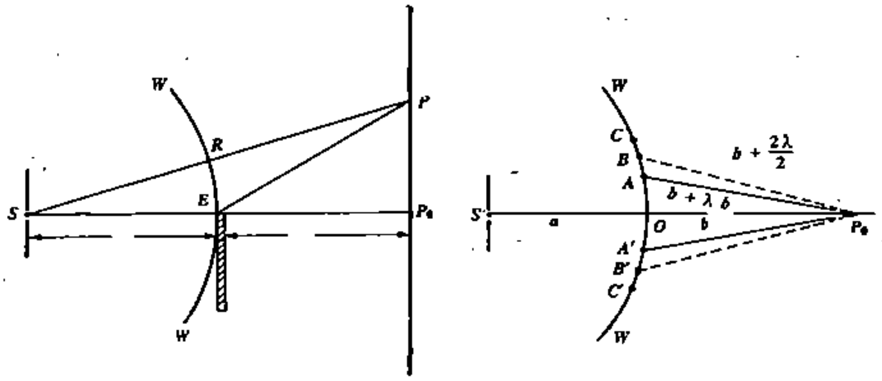


**Fig. 8.17: (a) Cross sectional view of the geometry to observe diffraction due to a straight edge and (b) Fresnel construction divides the cylindrical wavefront in half period strips**

The line joining $S$ and $E$, the point on the wavefront,when produced meets the screen at $P_0$,which is the geometrical boundary of the shadow. Consider any point $P$ on the screen. A line joining it to $S$ cuts the wavefront at $R$. We wish to know how intensity varies on the screen. This calculation is somewhat complicated because we now have a cylindrical wavefront. Moreover, the obstacle does not have an axial symmetry.

For a plane wave and obstacles with axial symmetry you know how to construct Fresnel zones. To construct half period elements for a straight edge, we divide the cylindrical wavefront into strips. As before, we make sure in the construction that the amplitudes of the wavelets from these strips arrive at $P_0$ out of phase by $\pi$ so that alternate terms are positive and negative. This is achieved by drawing a set of circles with $P_0$ as centre and radii $b$, $b + \frac{\lambda}{2}$, $b + \frac{2\lambda}{2}$, ... etc. cutting the circular section of the cylindrical wave at points $O, AA', BB', CC'$, ... Fig. (8.17 b). If lines are drawn through $A, A', B, B'$ etc. normal to the plane of the paper, the upper as well as the lower half of the wavefront gets divided into a set of **half-priod strips**. These half period strips stretch along the wavefront perpendicular to the plane of the paper and have widths $OA, AB, BC$ ... in the upper half and $OA', A'B', B'C'$, ... in the lower half. You may recall that Fresnel zones are of equal area. For half period strips, this does not hold. The areas of half-period strips are proportional to their widths and these decrease rapidly as we go out along the wavefront from $O$.

From the geometry of the arrangement it is obvious that on the screen there will be no intensity variation along the direction parallel to the length of the slit. Therefore, the bright and dark fringes will be straight lines parallel to the edge.

A plot of theoretically calculated intensity distribution on the screen is shown in Fig. 8.18. You will note the following salient features:

(i) As we go from the point $P'$ deep inside the shadow towards the point $O$ defining the edge of the shadow, the intensity rises gradually. At $P'$, the intensity is almost zero.

(ii) At $O$, the intensity is one-fourth of what would have been the intensity on the screen with the unobstructed wavefront.

(iii) On moving further towards $P$, the intensity rises sharply and goes through a alternating series of maxima and minima of gradually decreasing magnitude and approach the value for the unobstructed wave. This is expected since effect of the edge at far off distances will be almost negligible.

(iv) The intensity of first maxima is greater than the intensity of unobstructed wave, i.e. it is greater than 4 times the intensity at $O$. Beyond these alternate maxima and minima, there is uniform illumination.

(v) The diffraction fringes are not of equal spacing (as in interference experiments); the fringes gradually come closer together as we move away from the point $O$.

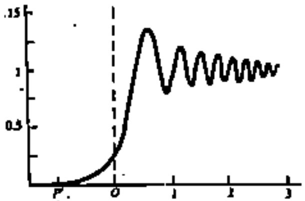


Fig. 8.18: Intensity distribution in the diffraction pattern due to a straight edge

You may now like to know atleast qualitative explanation of these results. To do so, we first consider the illumination at a point $P$ outside the geometrical shadow. The line joining $P$ and $S$ cuts the wavefront at $R$ so that the wavefront is divided in two parts. The amplitude of light at $P$ is due to the part $WE$ of the wavefront, which is completely unaffected by the straight edge. The amplitude at $P$ will be maximum if $RE$ contains odd number of half strips. This will happen if $EP - RP = (2n + 1) \lambda /2$. (When $EP - RP = n \lambda$, the portion $RE$ will contain even number of strips.) As pointed out earlier, the amplitudes due to strips are alternately positive and negative. Therefore, as point $P$ moves away from $O$, the illumination on the screen will pass alternately through maxima and minima when the number of half period strips in $RE$ is 1,2,3,4,...

It is worthwhile to ponder as to what pattern the geometry of the experimental configurations throws ? We expect dark and bright bands parallel to the edge. However, the dark bands will not be completely dark, since the upper half of the wavefront $RW$ always contributes light to this part of the screen.

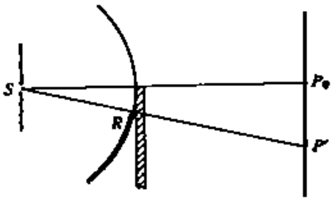


Fig. 8.19: The observation point is in the geometrical shadow of the straight edge

Let us now consider the situation for the point $P'$ inside the geometrical shadow. Refer to Fig. 8.19. You will note that the corresponding point $R$ is shifed below the edge so that the illumination at $P'$ is due entirely to the wavelets from the upper half of the wavefront; the lower portion having been blocked by the edge. Even the upper half is exposed only in part. If the edge cuts off $r$ strips of the upper half of the wavefront, the effect at $P'$ will be due to $(r + 1)$, $(r + 2)$, $(r + 3)$ etc. strips which may be taken to be equal to one-half of that due to $(r + 1)$th strip. This will rapidly diminish to zero as shown in Fig. 8.18, because the effectiveness of higher order strips goes on decreasing.

Let us now deduce the width of the diffraction bands. Again Refer to Fig. 8.17(a). Suppose that we have the $n$th dark band at $P$. Then

$$EP - RP = n \lambda \tag{8.6}$$

From the $\Delta\, EPO$, we have

$$EP = (b^2 + x^2)^{1/2} = b\left(1 + \frac{x^2}{b^2}\right)^{1/2}$$

$$\approx b\left(1 + \frac{1}{2}\frac{x^2}{b^2}\right) = b + \frac{1}{2}\frac{x^2}{b} \tag{8.7}$$

where we have retained only first two terms in the binomial series.

From the $\Delta\, SPO$, we can similarly write

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

Hence,

$$RP = SP - SR = b + \frac{1}{2}\frac{x^2}{(a + b)} \tag{8.8}$$

and

$$EP - RP = \left(b + \frac{1}{2}\frac{x^2}{b}\right) - \left(b + \frac{1}{2}\frac{x^2}{a + b}\right)$$

$$= \frac{1}{2}\left(\frac{x^2}{b} - \frac{x^2}{a + b}\right) = \frac{x^2 a}{2b(a + b)} \tag{8.9}$$

For the $n$th dark band, we get

$$\frac{x^2 a}{2b(a + b)} = n\lambda$$

or

$$x = \sqrt{n\frac{2b(a + b)}{a}\lambda} \tag{8.10}$$

We therefore find that the distances of the dark bands from the edge of the geometrical shadow are proportional to the square root of natural numbers. Consequently the bands will get closer together as we go out from the shadow. This fact distinguishes the diffraction bands from the interference bands, which are equidistant.

To enable you to grasp these ideas, we now give a solved example.

---

**Example 4**

In the above experiment if $a = 30$ cm, $b = 30$ cm and $\lambda = 5 \times 10^{-5}$ cm, calculate the position of the 1st, 2nd, 3rd and 4th minima from the edge of the shadow.

## Solution

From Eq. (8.10) we know that the distance of nth minima from the edge of the shadow is given by

$$x = \sqrt{n\frac{2b(a + b)}{a}\lambda}$$

If we substitute given values of $a$, $b$ and $\lambda$ and take $n = 1,2,3,4$ we find that

$$x_1 = \left[\frac{2 \times (30\,\text{cm}) \times (60\,\text{cm})}{30\,\text{cm}} \times (5 \times 10^{-5}\,\text{cm})\right]^{1/2}$$

$$= 7.75 \times 10^{-2}\,\text{cm}$$

$$x_2 = \sqrt{2}\,x_1 = 1.09 \times 10^{-1}\,\text{cm}$$

$$x_3 = \sqrt{3}\,x_1 = 1.34 \times 10^{-1}\,\text{cm}$$

$$x_4 = 2\,x_1 = 1.55 \times 10^{-1}\,\text{cm}.$$

From these values we find that the distance between consecutive minima decreases continuously as we move away from the edge of the shadow.

You may now like to answer an SAQ.

SAQ 2

*Spend 2 min*

Instead of the straight edge we keep a narrow obstacle, say a wire of diameter 1 mm. What will be the intensity on the screen?

Let us now summarise what you have learnt in this unit.

## 8.6 SUMMARY

- When the distance between the source of light and the observation screen or both from the diffracting aperture/obstacle is finite, the diffraction pattern belongs to Fresnel class.
- When the screen is very close to the slit, the illumination on the screen is governed by rectilinear propagation of light.
- The Fresnel diffraction pattern represents fringed images of the obstacle. Depending on the distance, there can be an infinite number of Fresnel diffraction patterns of a given obstacle/ aperture.
- When plane wavefronts are incident on a diffracting slit and the pattern is observed on a screen effectively at an infinite distance, the diffraction pattern belongs to Fraunhofer type. Unlike the Fresnel diffraction, there is only one Fraunhofer diffraction pattern.
- Fresnel construction for the diffraction pattern from any obstacle on which a plane wavefront is incident consists of dividing the wavefront into half period zones.
- The area of each Fresnel half-period zone is equal to $\pi b \lambda$.
- The resultant amplitude due to $n$th zone at any axial point is given by

$$a_n = \frac{A_n}{P_n}(1 + \cos\theta)$$

- The magnitude of resultant amplitude $AB$ due to the first half period element is $\frac{2}{\pi}$ times the value which would be obtained if all the wavelets within the half- period element had the same phase.
- The phase of the resultant vector of the first half period zone is $\frac{\pi}{2}$ behind the phase of light from the centre of the zone.
- A zone plate is an optical device in which alternate half-period zones are blackened.
- The diffraction pattern due to a circular aperture consists of a central bright spot.
- The diffraction pattern of a straight edge consists of alternate bright and dark bands. The spacing between minima (or maxima) decreases as we move away from the edge of the shadow:

$$x = \sqrt{n\frac{2b(a+b)}{a}\lambda}$$

## 8.7 TERMINAL QUESTIONS

1. Starting from Eq. (8.4) establish Eqs. (8.6) and (8.7). Assume that the obliquity factor is such that each term in Eq. (8.4) is less than the arithmetic mean of its preceding and succeeding terms.)
2. The eighth boundary of a zone plate has a diameter of 6mm. Where is its principal focal point located for light of wavelength 5000 Å.
3. How many fresnel zones will be obstructed by a sphere of radius 1 mm if the screen is 20cm away ? Take = 5000 Å. If the distance of the screen is increased to 200 cm, What will be the size of the sphere which will cut off 10 zones.

## 8.8 SOLUTIONS AND ANSWERS

### SAQs

1. The radius of the coin is equal to 1 cm. To know the number of zones being obstructed, we use the relation

$$n = \frac{r_n^2}{b\lambda}$$

where $r_n = 1$ cm, $b = 200$ cm and $\lambda = 5 \times 10^{-5}$ cm.

$$\therefore \qquad n = \frac{(1\text{ cm})^2}{(200\text{ cm}) \times (5 \times 10^{-5}\text{ cm})}$$

$$= 10$$

You should definitely expect to see a dim spot at the centre because eleventh zone is the first contributing zone.

When the screen is 4 m away, the number of zones being obstructed is given by

$$n = \frac{(1\text{ cm})^2}{(400\text{ cm}) \times (5 \times 10^{-5}\text{ cm})}$$

$$= 5$$

That is, only five zones are obstructed now and the first contributing term in Eq. (8.4) is $a_6$, which will contribute more than $a_{11}$. Therefore, the central spot is expected to be brighter. Does it not contradict the inverse square law?

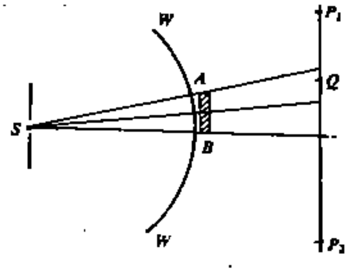


Fig. 8.20: A cross-sectional view of the arrangement for producing diffraction due to a narrow obstacle

2. Refer to Fig. 8.20. A point $P_1$ outside the geometrical shadow is similar to such a point in the straight edge. So we will have unequally spaced bright and dark fringes parallel to the wire on each side of the shadow. What is the intensity at $Q$ inside the shadow ? It is simply half the effect of the first half period strip on either side of the thin wire. It will show equally spaced fringes inside the shadow.

### TQs

1. We rewrite Eq. (8.4) as

$$a(P_0) = \frac{a_1}{2} + \frac{a_1}{2} - a_2 + \frac{a_3}{2} + \frac{a_3}{2} - a_4 + \frac{a_5}{2} + \ldots \qquad \text{(i)}$$

When $n$ is odd, the last term would be $\frac{a_n}{2}$. We are told that the obliquity is such that each term is less than the arithmatic mean of its preceding and succeeding terms i.e., $a_n > \frac{1}{2}(a_{n-1} + a_{n+1})$. Then, the quantities in the parentheses in (i) will be positive. So when $n$ is odd, the minimum value of the amplitude of the fields produced by consecutive zones is given by

$$a(P_0) > \frac{1}{2}(a_1 + a_n) \tag{ii}$$

To obtain the upper limit, we rewrite Eq. (8.4) as

$$a(P_0) = a_1 - \frac{a_2}{2} - \frac{a_2}{2} - a_3 + \frac{a_4}{2} - \frac{a_4}{2} - a_5 + \frac{a_6}{2} - \dots - \frac{a_{n-1}}{2} + a_n.$$

Following the argument used in obtaining the lower limit on the amplitude, we find that the upper limit is

$$a(P_0) < a_1 - \frac{a_2}{2} - \frac{a_{n-1}}{2} + a_n \tag{iii}$$

Since the amplitudes for any two adjacent zones are nearly equal, we can take $a_{n-1} = a_n$. Within this approximation

$$a(P_0) < \frac{a_1 + a_n}{2} \tag{iv}$$

The results contained in (ii) and (iv) suggest that when $n$ is odd, the resultant amplitude at $P_0$ is given by

$$a(P_0) = \frac{a_1 + a_n}{2} \tag{v}$$

Following the same method, you can readily show that if $n$ were even,

$$a(P_0) = \frac{a_1 - a_n}{2} \tag{vi}$$

2. $D_8 = 0.6$ cm so that $r_8 = 0.3$ cm. We know that

$$f_n = \frac{r_n^2}{n\lambda}$$

$$\therefore \quad f_8 = \frac{r_8^2}{8\lambda} = \frac{(0.3\text{ cm})^2}{8 \times (5 \times 10^{-5}\text{ cm})}$$

$$= 2.25 \times 10^2\text{ cm}$$

$$= 225\text{ cm}$$

3.a) The radius of a Fresnel zone is given by

$$r_n = \sqrt{n\lambda b}$$

Here we are told that $r_n = 0.1$ cm, $b = 20$ cm and $\lambda = 5 \times 10^{-5}$cm.

$$\therefore\ n = \frac{r_n^2}{b\lambda} = \frac{10^{-2}\ \text{cm}^2}{(20\ \text{cm}) \times (5 \times 10^{-5}\ \text{cm})} = 10$$

(b) In this part we have to calculate $r_n$ for given values of $n = 10$, $b = 200$ cm and $\lambda = 5 \times 10^{-5}$ cm.

$$r_n = \sqrt{10 \times (200\ \text{cm}) \times (5 \times 10^{-5}\ \text{cm})}$$

$$= 0.32\ \text{cm}$$

# UNIT 9 FRAUNHOFER DIFFRACTION

## Structure

## 9.1 INTRODUCTION

In the previous unit you studied Fresnel diffraction and learnt that the diffraction pattern depends on the distance between aperture and screen as well as the source. As the observation screen is moved away from the aperture, the diffraction pattern passes from the forms predicted in turn by geometrical optics, Fresnel diffraction and Fraunhofer diffraction. When plane wavefront is incident at the diffracting aperture, the transition from Fresnel to Fraunhofer pattern is determined by the ratio of the size of the diffracting obstacle to its distance from the source and/or the observation screen. You will now learn about Fraunhofer diffraction in detail.

In Sec. 9.2 we have described the experimental arrangement and salient features of the observed Fraunhofer diffraction pattern from a single slit illuminated by a point source. This is followed by a simple discussion on theoretical analysis of the observed results. Since we deal with plane wavefronts, you will find that theoretical analysis is fairly simple. In Sec. 9.3 we have described Fraunhofer diffraction by a circular aperture because of its importance for optical devices. You will learn that the diffraction pattern consists of a central bright disc (called Airy disc) surrounded by concentric dark and bright rings. As a corollary, you will see that a random array of small and closely circular obstacles gives overlapping diffraction patterns called halos. You may have observed brilliant halos while deriving a car whose fogged window is illuminated by motorcycle at the back. We shall discuss the physical basis for diffraction halos at the end of this unit.

### Objectives

After going through this unit you will be able to

- describe experimental arrangement for observing Fraunhofer diffraction pattern from a narrow vertical slit and a circular aperture
- explain observed irradiance on the basis of simple theoretical analysis
- solve numerical problems, and
- explain formation of diffraction halos.

## 9.2 DIFFRACTION FROM A SINGLE SLIT: POINT SOURCE

From the previous unit, you may recall that to observe Fraunhofer diffraction pattern, we require a point source, which is far-away (almost at infinity) from the diffracting aperture (a single slit in the present discussion). The wavefronts of light approaching the diffracting aperture can be assumed to be essentially plane. The observation screen should also be at infinite distance from the aperture. You may now like to ask: Is it practical to put the source of light and the observation screen at infinite distance from the diffracting aperture? This definitely is not practical because (i) the intensity of diffracted light reaching the observation screen would be reduced infinitesimally (inverse square law) and (ii) we will require infinitely big laboratory rooms. Do these limitations suggest that we cannot observe Fraunhofer diffraction? These difficulties are readily overcome by using converging lenses in an actual experiment.

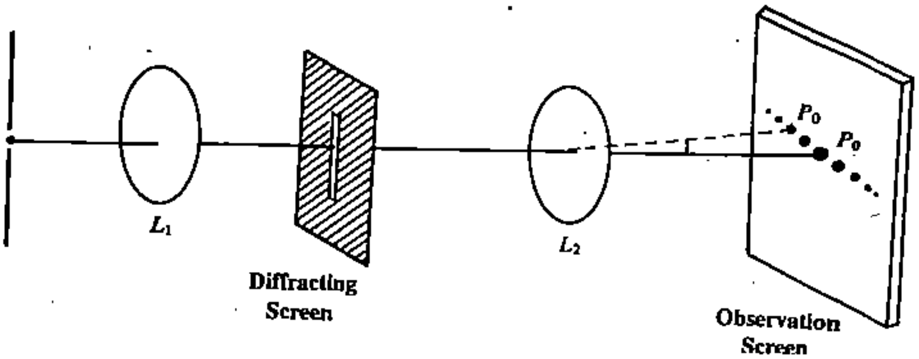


Fig. 9.1: Producing Fraunhofer diffraction pattern

The experimental arrangement for producing Fraunhofer diffraction pattern is shown in Fig. 9.1. The source of light is placed in the focal plane of a converging lens $L_1$, so that a plane wave is incident on a long narrow slit. Another convergent lens $L_2$ is placed on the other side of the slit. The observation screen is placed at the second focal point of this lens. Then light reaching any point on the observation screen is due to parallel diffracted wavelets from different portions of the wavefront at the slit. You must note that the observation screen and diffraction screen are kept parallel. Moreover, both the screens are perpendicular to the common axis of $L_1$ and $L_2$. The slit is so adjusted that the common axis of these lenses is perpendicular to the length of the slit and passes through the middle of the slit both in height and width.

In a physics laboratory this arrangement is easily achieved by using an ordinary spectrometer. We hope that you got an opportunity to work with a spectrometer in your second level laboratory course. To observe the diffraction from a point source, the slit of the collimator should be replaced by a fine pinhole, which should be carefully positioned at the focal point of the collimator lens. The observation screen can be placed at the second focal plane in the back focal plane of the telescope. Alternatively, we may observe the back focal plane of lens $L_2$ with an eyepiece. The diffracting screen with slit aperture, is placed between the two lenses suitably on the turn table.

### 9.2.1 Observed Pattern

Let us pause for a minute and think how wonld diffraction pattern of the vertical slit appear ? Or what would be the distribution of intensity in this pattern ? You may think that the diffraction pattern would be a single vertical line or a series of vertical lines on the observation screen. This line of thought is wildly off-target. The actual diffraction pattern is astonishingly different; it consists of a horizontal streak of light composed of bright elongated spots connected by faint streaks. In other words, after passing through the vertical slit, light spreads along a horizontal line. This neans that diffraction pattern is along a line perpendicular to the length of the liffracting slit. You may interpret this horizontal diffraction as a spread out image of the point source. The extent of horizontal spreading is controlled by the width of he slit; as the width increases, the spreading decreases. And in the extreme case of ι very wide slit, the (horizontal) diffraction streak reduces to a bright point. ?hysically, very wide slit means that the slit has effectively been removed.

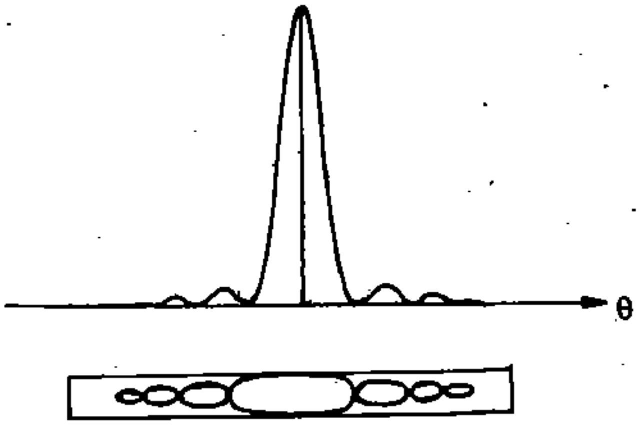


Fig. 9.2: Observed Fraunhofer diffraction pattern of a diffracting slit

The salient features of the observed Fraunhofer diffraction pattern of a single vertical slit from a point source are shown in Fig. 9.2. These are summarised below:

i) The diffraction pattern consists of a horizontal streak of light along a line perpendicular to the length of the slit.

ii) The horizontal pattern is a series of bright spots. The spot at the central point $P_0$, which lies at the intersection of the axis of $L_1$ and $L_2$ with the observation screen, is the brightest. On either side of the brightest spot we observe many more bright spots symmetrically situated with respect to $P_0$.

iii) The intensity of the central spot is maximum. The peak intensities of other spots, on either side of the central spot, decrease rapidly as we move away from $P_0$. The central maximum is called principal maxima and the others as secondary maxima.

The width of an image is specified by the distance between two consecutive minima.

iv) The width of the central spot is double of the width of other spots.

v) A careful examination of the diffraction pattern shows that the central peak is symmetrical. But on either side of the central maximum, secondary maxima are asymmetrical. In fact, the positions of the maxima are slightly shifted towards the observation point $P_0$.

Let us now learn the theoretical basis of these results.

### 9.2.2 Calculation of Intensity Ditribution

The first step in the calculation of intensity distribution is to realise that the observed diffraction pattern is focussed on the observation screen placed at the back focal plane of lens $L_2$. We know that only parallel rays are brought to focus in the back focal plane of the lens. The beam of rays parallel to the axis of the lens are focussed at the focal point. However, the beam inclined to the axis of the lens is brought to focus on the back focal plane but away from the focal point. We can as well describe this observation in terms of the wavefront; the two being perpendicular to each other. Since diffraction pattern lies on a horizontal line (which is at right angles to the common axis of $L_1$ and $L_2$), the diffracted wavefronts will be vertical planes perpendicular to the plane of the paper. That is, after passing

We take the plane of the paper as horizontal. The plane of the paper is defined by the diffraction streak and the axis of the lens $L_2$.

through the vertical slit, the incident plane waves are replaced by a system of vertical plane waves which proceed in different directions. Therefore for our theoretical analysis it is sufficient to assume that when a plane wavefront falls on the diffracting slit, each point of the aperture such as $A\,A_1\,A_2\,A_3 ...\,B$ (Fig. 9.3) becomes a source of secondary wavelets, which propagate in the direction of the point $P_\theta$ under consideration. These are diffracted plane waves. (You should realize that diffracted waves have no existence in the domain of geometrical optics. The diffracted waves arise due to interaction between light and matter. In the present case, the interaction is between light and the jaws of the slit.)

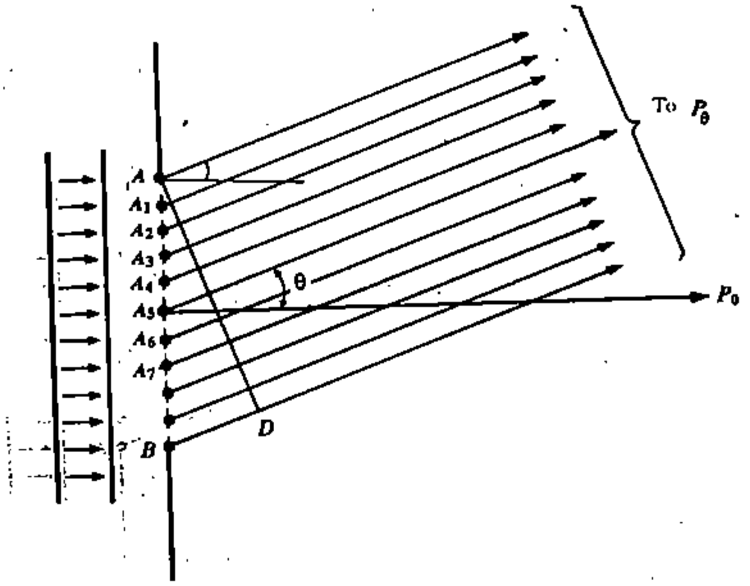


Fig.9.3: Geometry of single slit diffraction

Refer to Fig. 9.3 which shows the geometry for the irradiance at the point $P_\theta$ (on the distant screen) which makes an angle θ with the axis. In order to sum up the contributions of different wavelets at $P_\theta$, we must know their amplitudes and phases. The amplitudes of the disturbances from $A$ , $A_1$ , $A_2$ , ... will be very nearly equal. Do you know why? This is because the distance of point $P_\theta$ from the diffracting screen is very large compared to the width ($b$) of the aperture.

Two sources are said to be coherent if they emit in-phase waves of the same frequency.

Now let us consider the phases of the disturbances reaching the point $P_\theta$. You will agree that the points $A$ , $A_1$ , $A_2$ , $A_3$ , ... $B$ within the aperture form a series of coherent sources since they have originated from the same point source. Also points $A$ , $A_1$ , $A_2$ ,... $B$ are in the same phase since they lie on the same plane wavefront. The phase difference between different diffracted rays reaching $P_\theta$ arises due to the difference in path lengths travelled by them to reach this point. To know the phase difference, we draw a plane normal to the parallel diffracted rays. The trace of this plane in the plane of the paper is $AD$ (Fig. 9.3). Though the disturbances are in phase at points $A$ , $A_1$ , $A_2$ , ... $B$ when they start, they reach the trace $AD$ in different phases because of the unequal path lengths travelled by them. The optical paths of diffracted waves from the plane $AD$ to the focal point $P_\theta$ are equal. The optical paths of all rays between perpendicularly intersecting planes containing the parallel beam of light and the point where rays converge after traversing the lens are equal. Therefore, the wavelets arrive at $P_\theta$ with the same relative phase difference as the ones existing at the trace $AD$.

Let us consider the aperture $AB$ to be divided into $n$ equal parts so that $AA_1 = A_1A_2 = A_2A_3 = b/n = \Delta$. It means that the number of point sources is $(n+1)$. Actually, the aperture contains a continuous distribution of points from $A$ to $B$, and therefore in the limiting case, $n \to \infty$ and $\Delta \to 0$ such that $n\Delta \to b$. Consider two rays starting from two neighbouring points $A$ and $A_1$. The path difference between them is $AA_1 \sin\theta$ where $\theta$ is the angle between the diffracted rays and the normal to the slit. Hence the corresponding phase difference is given by

$$\phi = \frac{2\pi}{\lambda}(AA_1\sin\theta) = \frac{2\pi}{\lambda}\left(\frac{b}{n}\sin\theta\right) = \frac{2\pi}{\lambda}\Delta\sin\theta \qquad (9.1)$$

Let the field at $P_\theta$ due to the disturbance originating from $A$ be $a_0 \cos \omega t$. Then, the field due to the disturbance from $A_1$ is $a_0 \cos(\omega t - \phi)$. Here we have assumed that the amplitudes of disturbances from different points are equal. The fields due to disturbances from successive points $A_2, A_3, \ldots B$ are $a_0 \cos(\omega t - 2\phi)$, $a_0 \cos(\omega t - 3\phi)$ ..., $a_0 \cos(\omega t - n\phi)$, respectively. The magnitude of resultant field $E$ at $P_\theta$ is equal to the sum of these disturbances. Hence

$$E = a_0 \cos \omega t + a_0 \cos(\omega t - \phi) + a_0 \cos(\omega t - 2\phi) + \ldots + a_0 \cos(\omega t - n\phi)$$

In Unit 2 of PHE-03 course on Oscillations and Waves, we summed up this series (Eq. (2.38)). We will just quote the result here:

$$E = a_0\left[\frac{\sin\frac{n\phi}{2}}{\sin\left(\frac{\phi}{2}\right)}\right]\cos\left(\omega t - \frac{n\phi}{2}\right)$$

$$= E_\theta \cos\left(\omega t - \frac{n\phi}{2}\right) \qquad (9.2)$$

where $E_\theta$ is the amplitude of the resultant field at $P_\theta$:

$$E_\theta = a_0 \frac{\sin\left(\frac{n\phi}{2}\right)}{\sin(\phi/2)} \qquad (9.3)$$

In the limit $n \to \infty$ and $\Delta \to 0$, $n\Delta \to b$. Then from Eq. (9.1) we have

$$\frac{n\phi}{2} = \frac{n}{2}\frac{2\pi}{\lambda}\Delta\sin\theta = \frac{\pi}{\lambda}(n\Delta)\sin\theta = \frac{\pi}{\lambda}b\sin\theta$$

so that $\phi = \frac{2\pi}{\lambda}\frac{b\sin\theta}{n}$ will be very small for $n \to \infty$. We may therefore write

$$\sin\left(\frac{\phi}{2}\right) = \frac{\phi}{2} = \frac{\pi b\sin\theta}{n\lambda}$$

Substitute this result in Eq. (9.3). On simplification you will find that

$$E_\theta = a_0\frac{\sin\left(\frac{n\phi}{2}\right)}{\sin(\phi/2)} = a_0\frac{\sin(n\phi/2)}{(\phi/2)} = na_0\frac{\sin\left(\frac{\pi b\sin\theta}{\lambda}\right)}{\left(\frac{\pi b\sin\theta}{\lambda}\right)}$$

$$= na_0\left(\frac{\sin\beta}{\beta}\right) = A\left(\frac{\sin\beta}{\beta}\right) \qquad (9.4)$$

where we have written

$$A = n\,a_0$$

and

$$\beta = \pi \frac{b \sin\theta}{\lambda} \tag{9.5}$$

You will note that for a given wavelength, β signifies half of the phase difference between disturbances originating from the extreme points $A$ and $B$. The expression for resultant field at $P_\theta$ takes the form

$$E_\theta = A \frac{\sin\beta}{\beta} \cos(\omega t - \beta) \tag{9.6}$$

The corresponding intensity distribution at $P_\theta$ is given by

$$I_\theta = A^2 \left( \frac{\sin\beta}{\beta} \right)^2 \tag{9.7}$$

Let us pause for a while and ponder as to what have we achieved. This result suggests that the intensity is maximum at $\theta = 0$. This readily follows by noting that when we substitute $\theta = 0$ we have both β and sin β equal to zero but

$$\lim_{\beta \to o} \frac{\sin\beta}{\beta} = 1$$

Therefore,

$$I_{\theta = 0} = A^2$$

This result is expected on geometrical considerations. In the limits of a distant screen, the central point becomes equidistant from each point on the slit. All diffracted waves arrive in phase at $P_0$ and interfere constructively. $A^2$ is then the value of the maximum intensity at the centre of the pattern. This maximum is also termed **principal maximum**.

For brevity we write $I_{\theta = 0} = A^2 = I_0$. Then intensity at any point at an angle θ with the horizontal axis, is given by

$$I_\theta = I_0 \left( \frac{\sin\beta}{\beta} \right)^2$$

**Positions of maxima and minima**

A plot of Eq. (9.7) for intensity distribution is shown in Fig.9.4. You will note that the intensity is maximum for $\theta = 0$: $I_{\theta = 0} = I_0 = A^2$. The intensity gradually falls on either side of the principal maximum and becomes zero when $\beta = +\pi$ or $\beta = -\pi$ since $\sin(\pm\pi)$ is zero. This is the first minimum. So we can say that the angular half width of principal maximum is from 0 to π. The second minimum on either side occurs at $\beta = \pm 2\pi$. Thus we get the minima when

$$\beta = \pm \pi, \pm 2\pi, \pm 3\pi \ldots$$

$$= m\pi \quad m = \pm 1, \pm 2, \pm 3, \ldots \tag{9.8}$$

Note that the value $m = 0$ is excluded because it corresponds to the principal maximum (for $\beta = 0$). Substituting the value of β from Eq. (9.8) in Eq. (9.5) we find that the condition for minima is given by

$$b \sin\theta = \pm\lambda, \pm 2\lambda, \pm 3\lambda \ldots$$

$$= m\lambda, m = \pm 1, \pm 2, \pm 3, \ldots \qquad (9.9)$$

You may now conclude that the angular width of the principal maximum ($m = 1$) is defined by $b \sin\theta = \lambda$ or $\theta = \frac{\lambda}{b}$.

That is θ depends upon the wavelength of light and the slit width. For a given slit width, the spread in diffraction pattern depends directly on the wavelength. Accordingly you should expect that red light will be diffracted through a larger angle than the blue or violet light.

You may now like to know: What will happen when white light illuminates a single slit? We expect that each wavelength will be diffracted independently. This gives rise to a white central spot surrounded by coloured fringes. The outer part of this pattern would tend to be reddish. You can easily observe this diffraction pattern by looking through the tines of a dinner fork at a candle in a dimly illuminated room. On twisting the fork about its handle, you will observe the diffraction pattern as soon as the cross-sectional area becomes small enough.

The expression $I_\theta = I_0\left(\frac{\sin\beta}{\beta}\right)^2$ gives the diffraction intensity in different directions. In order to determine the directions (and positions) of **secondary maxima**, we differentiate this equation with respect to β and equate the result to zero. This gives

$$\frac{dI_\theta}{d\beta} = 2I_0\left(\frac{\sin\beta}{\beta}\right)\left[\frac{\beta\cos\beta - \sin\beta}{\beta^2}\right]$$

$$= 2I_0 \sin\beta\left[\frac{\cos\beta}{\beta^2} - \frac{\sin\beta}{\beta^3}\right] = 0$$

or $$\sin\beta\,(\beta - \tan\beta) = 0$$

From this we get the conditions $\sin\beta = 0$ and $\beta - \tan\beta = 0$

The condition $\sin\beta = 0$ implies that $\beta = \pm m\pi$, where $m$ is any integer. This is a trivial condition as it signifies minima and is of no interest.

The condition $\beta = \tan\beta$ therefore gives the positions of secondary maxima. This is a transcendental equation. The roots of this equation can be found by a graphical method. All you have to do is to recall that an angle equals its tangent at intersections of the straight line

$$y = \beta$$

and the curve

$$y = \tan\beta \qquad (9.10)$$

Plots of these curves are also shown in Fig. 9.4. The points of intersection excluding $\beta = 0$ (which corresponds to principal maximum) occur at $\beta = 1.43\pi, 2.46\pi, 3.47\pi$ etc. and give the position of the first, second, third maxima on either side of the central maximum. You should note that these maxima do not fall midway between the two minima. For instance, the first maximum occurs at 1.43 π rather than 1.50 π. Similarly the second maxima occurs at 2.46 π rather 2.50 π and so on. This means that the intensity curves are asymmetrical. The plot clearly shows that the positions of maxima are slightly shifted towards the centre of the pattern. You may recall that this is observed experimentally as well.

A very clear idea of the single slit pattern can be obtained from the following simple qualitative argument. The path difference between waves diffracted by extreme points in the slit is $BD = b \sin\theta$ (see figure below). If $BD$ is an integral multiple of λ, we will show that the resultant intensity at $P_\theta$ will be zero. For $m = 1$, the angle θ satisfies the equation $b \sin\theta = \lambda$. We divide the slit into two equal halves $AM$ and $MB$ as shown in the figure below. Consider the waves starting from the two point sources $A$ and $M$. The path difference between them is $AM \sin\theta = (b/2)\sin\theta = \lambda/2$. The corresponding phase difference will be π. Therefore the two waves on superposition lead to zero intensity at $P_1$.Similarly, for a point $A_1$, just below $A$, there will be a corresponding point $M_1$ just below M such that the path difference between disturbances generated by them is again λ/2. On superposition, this pair also leads to zero intensity at $P_1$. We can thus pair off all the points in the upper half ($AM$) with corresponding points in the lower half ($MB$) and the disturbances due to upper half of the slit will be cancelled by disturbances due to the lower half. So the resultant intensity at $P_1$ will be zero. This explains why we get a minimum intensity at $P_1$ when the path difference between the rays form extremes equal to λ.

Let us now consider the case $m = 2$ so that the path difference $b \sin\theta$ between the extreme rays is equal to 2λ. You can now imagine that the slit is divided into four equal parts and by similar pairing show that the first and second quarters have a path difference of λ /2 and cancel each other. Third and fourth quarters cancel each other by the same argument so that the resultant intensity in the focal plane at $P_1$ is again zero. For $m = 3$ the path difference between the two extreme rays $b \sin\theta = 3\lambda$. In this case, the slit should be divided into six equal parts to show similar pairing and cancellation and then leading to zero intensity. By this simple qualitative argument, we have shown that when the path difference between the extreme parallel diffracted rays in a particular direction is an integral multiple of λ, the resultant diffracted intensity in that direction is zero.

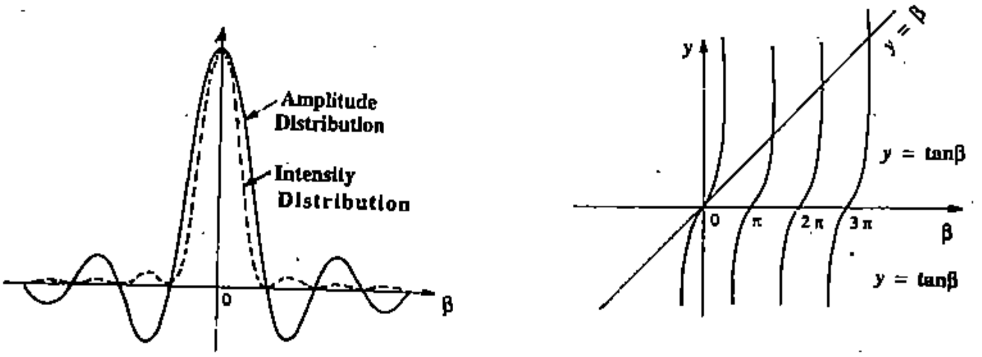


**Fig. 9.4: Amplitude and intensity contours for Fraunhofer diffraction of a single slit showing positions of maxima and minima**

Let us now calculate the intensities at these positions of maxima. The intensity of first maximum is given by

$$\left(\frac{\sin 1.43\,\pi}{1.43\,\pi}\right)^2 = 0.0496$$

$y = \beta$ is a straight line passing through the origin. $y = \tan\beta$ is represented by a family of curves having for asymptotes $\beta = \frac{\pi}{2}, \pm\frac{3\pi}{2}, \pm\frac{5\pi}{2}, \ldots$

This means that the intensity of the first secondary peak (nearest to the central peak) is about 4.96% of the central peak. Similarly, you can calculate and convince yourself that the intensities of the second and third maxima are about 1.68% and 0.83% of the central maximum. We call these maxima as secondary maxima.

The intensities of the secondary maxima can be calculated to a fairly close approximation by finding the values of β at halfway positions i.e. at $\beta = \frac{\pi + 2\pi}{2}, \frac{2\pi + 3\pi}{2}, \frac{3\pi + 4\pi}{2}, \cdots$ etc. The intensities at these positions are $7\frac{4}{9\pi^2}, \frac{4}{25\pi^2}, \frac{4}{49\pi^2}, \cdots$ or $\frac{1}{22.1}, \frac{1}{61.7}, \frac{1}{121}, \cdots$ of the central maximum which are very close to the above calculated values. From this you may conclude that most of the light is concentrated in the central maximum.

Another important characteristic of the principal maximum is that its width is double of the width of secondary maximum. We have left its mathematical proof as an exercise for you. Before you proceed, you should solve SAQ 1.

*Spend 5 min*

**SAQ 1**

Show that the principal maximum is twice as wide as the secondary maxima.

To give you a feel for numerical values and fix up the ideas developed in this section, we now give a few solved examples. You should go through these carefully.

**Example 1**

In the experimental set up used to observe Fraunhofers diffraction of a vertical slit (width 0.3mm), the focal length of lens $L_2$ is 30 cm. Calculate (a) the diffraction angles and positions of the first, second and third minima, and (b) the positions of the first, second and third maxima on either side of the central spot. The slit is illuminated with yellow sodium light which is a doublet. You may take $\lambda = 6000$ Å.

**Solution**

You have seen that the conditions for minima are given by $b \sin \theta = m\lambda$; $m = \pm 1, \pm 2, \pm 3, \ldots$ For small values of $\theta$, we may write $\sin \theta \cong \theta$. Then

$$\theta = m \frac{\lambda}{b}$$

and the distance $P_0 P_0$ is $f\theta$ where $f$ is the focal length. Therefore, the diffraction angles $\theta_1$, $\theta_2$, $\theta_3$ for the first, second and third minima are $\frac{\lambda}{b}$, $2\frac{\lambda}{b}$, and $3\frac{\lambda}{b}$, respectively.

On substituting the numerical values of $\lambda$ and $b$ we find that

$$\theta_1 = \frac{6000 \times 10^{-8}\ \text{cm}}{0.3 \times 10^{-1}\ \text{cm}} = 2 \times 10^{-3}\ \text{rad}$$

$$\theta_2 = 2\theta_1 = 4 \times 10^{-3}\ \text{rad}$$

$$\theta_3 = 3\theta_1 = 6 \times 10^{-3}\ \text{rad}$$

The distances $d_1$, $d_2$, $d_3$ of these minima from the central spot are

$$d_1 = f\theta_1 = (30\ \text{cm}) \times 2 \times 10^{-3} = 60 \times 10^{-2}\ \text{cm} = 0.06\ \text{cm}$$

$$d_2 = 2f\theta_1 = 2 \times 0.06\ \text{cm} = 0.12\ \text{cm}$$

$$d_3 = 3f\theta_1 = 2 \times 0.06\ \text{cm} = 0.18\ \text{cm}$$

You will note that these minima are separated by a distance of 0.06 cm on the focal plane of the lens. We know that the first three secondary maxima occur at $\beta = 1.43\,\pi$, $2.46\,\pi$ and $3.47\pi$, respectively. The corresponding diffraction angles for these three maxima are

$$(\theta_1)_{max} = 1.43\frac{\lambda}{b}, (\theta_2)_{max} = 2.46\frac{\lambda}{b} \text{ and } (\theta_3)_{max} = 3.47\frac{\lambda}{b}$$

$$\therefore \qquad (\theta_1)_{max} = (1.43)(2 \times 10^{-3}), (\theta_2)_{max} = (2.46)(2 \times 10^{-3}),$$

and

$$(\theta_3)_{max} = (3.47)(2 \times 10^{-3})$$

and the corresponding distances from the central point ($P_0$) are

$$d_1 = f(\theta_1)_{max} = (30\ \text{cm}) \times 1.43 \times 2 \times 10^{-3} = 0.86\ \text{cm}$$

$$d_2 = f(\theta_2)_{max} = (30\ \text{cm}) \times 2.46 \times 2 \times 10^{-3} = 0.16\ \text{cm}$$

$$d_3 = f(\theta_3)_{max} = (30\ \text{cm}) \times 3.47 \times 2 \times 10^{-3} = 0.21\ \text{cm}$$

### Example 2

In the above experiment, we change slit widths to 0.2mm, 0.1mm, and 0.6mm. Calculate the positions of the first and second minima.

**Solution**

For slit width $b = 0.2$ mm, we have

$$d_1 = f\theta_1 = (30\text{ cm}) \times \frac{6000 \times 10^{-8}\text{ cm}}{0.2 \times 10^{-1}\text{ cm}} = 0.09\text{ cm}$$

Similarly

$$d_2 = f\theta_2 = 2 \times 0.09\text{ cm} = 0.18\text{ cm}$$

These minima are separated by 0.09 cm. Recall that the corresponding value for a slit of width 0.03 cm was 0.06 cm. This means that for a given wavelength, the spread of secondary maximum increases as slit width decreases. This conclusion is brought out in the following calculations as well.

For a slit of width $b = 0.1$ mm, we have

$$d_1 = (30\text{ cm}) \times \frac{6000 \times 10^{-8}\text{ cm}}{0.1 \times 10^{-1}\text{ cm}}$$

$$= 0.18\text{ cm}$$

and

$$d_2 = 2 \times 0.18\text{ cm} = 0.36\text{ cm}$$

For slit width $b = 0.06$ mm, we have

$$d_1 = (30\text{ cm}) \times \frac{6000 \times 10^{-8}\text{ cm}}{0.6 \times 10^{-1}\text{ cm}}$$

$$= 0.3\text{ cm}$$

and

$$d_2 = 2 \times 0.3\text{ cm} = 0.6\text{ cm}$$

We thus find that for slits of widths 0.3mm, 0.2mm 0.1mm, and 0.06mm, the first minimum on either side of the principal maximum occurs at distances of 0.06 cm, 0.09 cm, 0.18 cm, and 0.3 cm. In these four cases, the corresponding principal maximum extends over 0.12 cm, 0.18 cm, 0.36 cm, and 0.6 cm.

This shows that as the slit becomes narrower, the spread of central maximum increases. Conversely, the wider the slit width, the narrower is the central diffraction maximum.

We now consider an interesting case where width of the slit is varied in comparison to the wavelength of light.

### Example 3

Consider a slit of width $b = 10\lambda$, $5\lambda$ and $\lambda$. Calculate the spread of the central maximum.

**Solution**

From Eq. (9.9), we note that for a slit of width $b = 10\,\lambda$, the first minimum is located at

$$10\,\lambda \sin\theta = \lambda$$

or

$$\sin\theta = 0.10$$

and

$$\theta = 5.7^\circ$$

For a slit of width $5\lambda$, we have

$$5\lambda \sin\theta = \lambda$$

or

$$\theta = 11.5^\circ$$

That is, as the aperture of the slit changes from $10\lambda$ to $5\lambda$, the diffraction pattern spreads out about twice as far. For $b = \lambda$,

$$\sin\theta = 1$$

or

$$\theta = 90^\circ$$

The first minimum falls at 90°. That is, the central maximum spreads out and the diffraction pattern shows no ripple. These features are shown in Fig. 9.5.

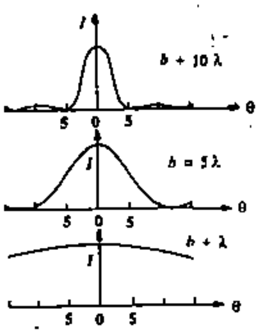


**Fig. 9.5: Single- slit diffraction irradiances as the slit width varies**

You may now like to answer an SAQ.

SAQ 2

*Spend 5 min*

We illuminate the slit of Example 1 with violet light of wavelength 4358 Å from a mercury lamp. Show that the diffraction pattern shrinks correspondingly.

## Diffraction Pattern of a Rectangular Aperture

So far we have described Fraunhofer diffraction pattern of a slit aperture. Let us now consider as to what will happen if both dimensions of the slit are made comparable. We now have a rectangular aperture of width $b$ and height $a$ as shown in Fig 9.6 (a). We expect that the emergent wave will spread along the length as well as the width of the slit. Can you depict the diffraction pattern? It is shown in Fig. 9.6 (b). Mathematically, the intensity is given by $I = \dfrac{I_0 \sin^2\beta \sin^2}{\alpha^2\,\beta^2}$ where $\beta = b\,\pi \sin\theta/\lambda$ and $\alpha = \pi\, a \sin\theta/\lambda$.

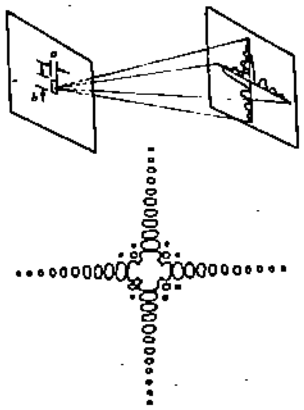


**Fig.9.6: Single-slit diffraction. Both dimensions of the rectangular aperture are small and a two- dimensional diffraction pattern is discernible on the screen (b) Diffraction image of a single square aperture.**

## Slit Source

The experimental arrangement shown in Fig. 9.1 is modifed as shown in Fig. 9.7. Here instead of the point source we use a slit source (Fig. 9.7(a)).

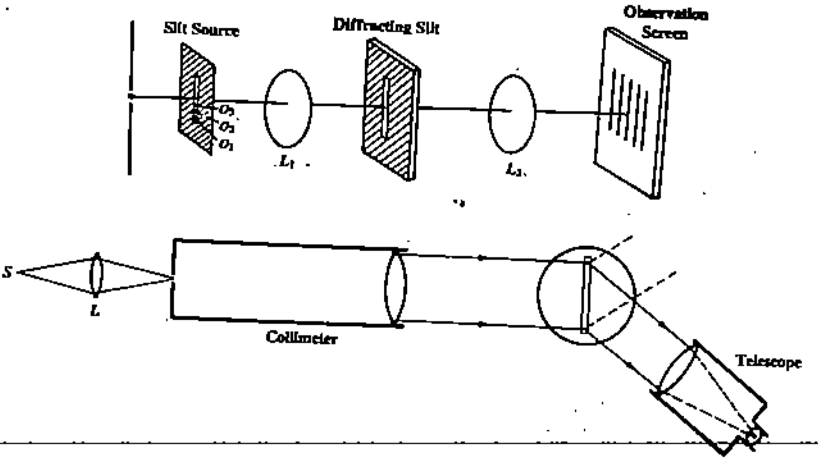


**Fig.9.7: (a) Experimental arrangement for diffraction from a vertical narrow single slit illuminated by a slit source (b) Experimental arrangement in a physics laboratory.**

As a matter of fact, the experimental arrangement, which is commonly employed in most experiments, uses a spectrometer (Fig. 9.7(b)). The slit of the collimator arm is illuminated so that each point of the slit source acts as an independent source. You know that a point source gives a horizontal streak of light as the diffraction pattern of a vertical slit. Now when we substitute a slit as a source, we can imagine a series of point sources $O_1$, $O_2$, $O_3$, ... etc, one above the other to form the slit source (Fig. 9.7(a)). Each point source wil give its own diffraction pattern since each point is to be regarded as an independent point source. With the same diffracting slit and the same lenses $L_1$ and $L_2$, the central diffraction maximum due to all point sources will lie above one another and give a central bright vertical fringe. Similarly from secondary maxima and minima points, we will obtain a series of vertical fringes, which will be situated at equal intervals on either side of the central fringe. The resulting pattern arises by superposition of a series of horizontal diffraction streaks stacked on each other in a vertical direction. The intensity along any horizontal line will be the same as in Fig. 9.2. We should note that each point of the slit source acts as an independent and effectively as a non-coherent source.

You will observe that clear fringes are obtained only when the width of the source slit is small. Suppose that the width of the source slit is gradually increased. This will lead to an increase in the width of its image on the observation screen. A stage will come when the width of the image, i.e. the fringe width, becomes comparable with the distances between successive vertical fringes. This will gradually make the vertical fringes less clear and indistinct. For a similar reason, we obtain clear fringes only when the source slit is parallel to the diffraction slit.

## 9.3 DIFFRACTION BY A CIRCULAR APERTURE

Fraunhofer diffraction by a circular aperture is of particular interest because a lens in an optical device (microscope,telescope, eye) can be regarded as a circular aperture. For this case, the experimental arrangement is shown in Fig. 9.8(a). A plane wave is incident normally on the aperture and a lens whose diameter is much larger than that of the aperture is placed close to it. The Fraunhofer diffraction pattern is observed on the back focal plane of the lens. Because of the rotational symmetry of the system,we expect that the diffraction pattern will consist of concentric dark and bright rings. Fig. 9.8(b) shows the diffraction pattern which is

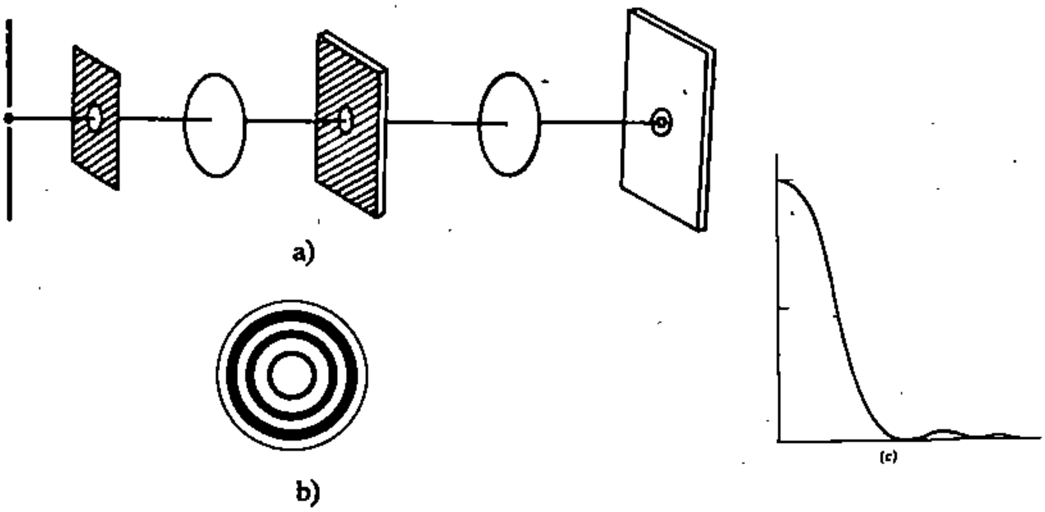


**Fig.9.8: (a) Experimental arrangement for observing the Fraunhofer diffraction pattern by a circular aperture. (b) The Airy pattern: The circle of light at the center corresponds to the zeroth order. (c) The corresponding intensity distribution.**

known as the Airy pattern. The detailed derivation of the diffraction pattern for a circular aperture involves complicated mathematics. So we just quote the final result for the intensity distribution:

$$I = I_0 \left[ \frac{2J_1(\gamma)}{\gamma} \right]^2 \tag{9.11}$$

where

$$\gamma = \frac{\pi D}{\lambda} \sin\theta \tag{9.12}$$

Here $D$ is the diameter of the aperture, $\lambda$ is the wavelength of light and $\theta$ is the angle of diffraction, $I_0$ is the intensity at $\theta = 0$ (which represents the central maximum) and $J_1(\gamma)$ is the Bessel function of the first order. (We know that you are not very familiar with Bessel functions.) We may just mention that the variation of $J_1(\gamma)$ is somewhat like a damped sine curve. Moreover, the intensity is maximum at the centre of the pattern since

$$\lim_{\gamma \to 0} \frac{2J_1(\gamma)}{\gamma} \to 1$$

similar to the relation

$$\lim_{\beta \to 0} \frac{\sin\beta}{\beta} \to 1$$

Other zeros of $J_1(\gamma)$ occur at $\gamma = 3.832, 5.136, 7.016, \ldots$ which correspond to the successive dark circles in the Airy pattern. Thus the first dark ring appears when

$$\sin\theta = \frac{3.832\lambda}{\pi D} \simeq \frac{1.22\,\lambda}{D} \tag{9.13}$$

Let us compare this result with the analogous equation for the narrow slit. We find that the angular half-width of the central disc, i.e. the angle between the central maximum and the first minimum of the circular aperture, differs from that for the slit pattern through the weird number 1.22. The intensity distribution of Eq. (9.11) is plotted in Fig. 9.8(c). The pattern is similar to that for a slit, except that the pattern for circular apertures now has rotational symmetry about the optical axis. The central maximum is consequently a circular disc of light, which may be regarded as the diffracted "image" of the circular aperture. It is called the **Airy disc.** It is surrounded by a series of alternate dark and bright fringes of decreasing intensity. However, the pattern is not sharply defined. If you consider any section through the circular aperture, intensity distribution is very much the same as obtained from a point source with a single slit. Indeed, the circular aperture pattern will be obtained if you rotate the single slit pattern about an axis in the direction of the light and passing through the central point of the principal maximum.

We now give an example to enable you to have a feel for the numerical values.

**Example 4**

Plane waves from a helium-neon laser with wavelength 6300 Å are incident on a circular aperture of diameter 0.5 mm. What is the angular location of the first minimum in the diffraction pattern? Also calculate the diameter of Airy disc on a screen 10m behind the aperture.

**Solution**

We know from Eq. (9.13) that

$$D \sin\theta = 1.22\,\lambda$$

On substituting the given values, we get

$$(0.5 \times 10^{-3}\mathrm{m})\sin\theta = 1.22 \times 630 \times 10^{-9}\,\mathrm{m}$$

or

$$\sin\theta = \frac{1.22 \times 630 \times 10^{-9}\,\mathrm{m}}{0.5 \times 10^{-3}\,\mathrm{m}}$$

$$= 1.54 \times 10^{-3}$$

In the small angle approximation, $\sin\theta \cong \theta$ so that

$$\theta = 1.54 \times 10^{-3}\,\mathrm{rad} = 0.087^\circ$$

On the screen placed 10m away, the linear location of the first minimum is

$$x = D\tan\theta \simeq D\sin\theta \cong D\theta$$

Hence

$$x = (10\,\mathrm{m}) \times (1.54 \times 10^{-3}\,\mathrm{rad})$$

$$= 15.4 \times 10^{-3}\,\mathrm{m} = 1.54\,\mathrm{cm}$$

This value of $x$ signifies the radius of the Airy disc so that the diameter is about 3 cm.

You can observe a white light circular diffraction pattern by making a small pinhole in a sheet of aluminum foil. Then look through it at a distant light bulb or a candle standing in a poorly illuminated (dark) room.

Another important result of above analysis is that the angular width of a beam is diffraction-limited. When a perfectly plane wave from a distant point source is incident on a diffracting aperture (of width or diameter $b$), the angular width of the diffracted beam is $\lambda / b$. This is illustrated in Fig. 9.9. The angular width can be zero if $b$ is infinite (1mm or so). At large distances from the diffracting aperture, beam width $W = L\,(\lambda/b)$. It has important implications for laser beams which are known to be highly directional. To have an ideas about it, let us consider a diffraction-limited laser beam ($\lambda$ = 6000 Å) of 2 mm diameter. The angular spread of the beam is

$$\theta = \frac{\lambda}{b} = \frac{6 \times 10^{-5}\ \text{cm}}{0.2\ \text{cm}} = 3 \times 10^{-4}\ \text{rad}$$

It means that in an auditorium (of lenght 15 m), the spatial spread $W$ = (1500cm) × ($3 \times 10^{-4}$) = 5 mm, which is very small. For a typical penlight type flash light, the transverse dimensions of the filament should be of the order of a micrometer, which is really hard to make.

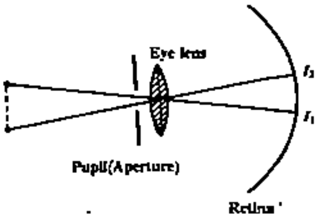


Fig. 9.9: Schematics of a diffraction limited system

Imagine that a random array of small circular apertures is illuminated by plane waves from a white point source. We know that each aperture will generate an Airy type diffraction pattern. If the apertures are small and close together, the diffraction patterns are large and overlap. The overlapping diffraction patterns produce a readily visible **halo**, namely, a central white disc surrounded by circular coloured rings. Which colour do you expect to be at the outermost rim? Should it not be red? Similar halos are also observed when the diffraction is due to a random array of circular obstacles.

Suspended water ($n$ = 1.33) droplets in air ($n$ = 1.00) give rise to diffraction halos. When observed through a light cloud cover around the sun or moon, these diffraction halos are referred to as coronas. We can distinguish between diffraction halos and ice crystal halos. Ice crystal halos are due to refraction and dispersion by the ice crystals; they have red on the inside of the rings.

While driving a car at night, you may have seen brilliant halos through fogged up car windows on which light of a motorcycle following you is incident. These are diffraction halos. You can easily produce such halos by breathing on the side of a clear glass and then looking through the fogged area at a small source (e.g., match, penlight, or distant bulb).

When the cornea swells (becomes edematous), small droplets of fluid form randomly between the stromal fibers. These random droplets produce a diffraction halo that the person sees when looking at light. Such halos are one of the warning signs of high ocular pressures. These halos can also be produced by epithelial damage due to poorly fitting contact lenses.

## 9.4 SUMMARY

- To observe Fraunhofer diffraction pattern, the distance of the diffracting screen from the source and/or observation screen should be almost infinite. Experimentally this condition is achieved by using convergent lenses.
- The diffraction pattern of a vertical slit consists of a horizontal streak of light. This horizontal diffraction pattern may be regarded as a spreadout image of the point source and consists of a series of diffraction spots symmetrically situated with respect to the central point.
- The central spot has a maximum intensity and its width is twice compared to other spots which are of equal width. Their intensities decrease rapidly. In fact, most of the light is concentrated in the central maximum.

- The plane wavefront incident on the slit gives rise to a system of vertical plane wavefronts which originate from each point of the diffracting aperture.
- The intensity at any point $P_\theta$ on the screen is computed by taking the phase difference between the successive diffracted waves into account. The intensity at a point $P_\theta$ is given by

$$I = I_0\left(\frac{\sin\beta}{\beta}\right)^2$$

where $\beta = \pi\dfrac{b\sin\theta}{\lambda}$ and $b$ is width of the slit.

- If the path difference $b\sin\theta$ between waves diffracted by extreme ends of the slit is an integral multiple of $\lambda$, we obtain zero intensity.
- The diffraction pattern of a thin slit source consists of a series of vertical fringes. In this pattern, the central vertical fringe is the brightest and the intensity of other fringes decreases rapidly. The width of central fringe is double of that for other fringes.
- The diffraction pattern of a circular aperture consists of concentric rings with a central bright disc. The first dark ring appears when $\sin\theta = 1.22\,\lambda/D$.

## 9.5 TERMINAL QUESTIONS

1. A single slit has a width of 0.03 mm. A parallel beam of light of wavelength 5500 Å, is incident normally on it. A lens is mounted behind the slit and focussed on a screen located in its focal plane, 100 cm away. Calculate the distance of the third minimum from the centre of the diffraction pattern of the slit.

2. A helium-neon laser emits a diffraction-limited beam ( $\lambda = 6300$ Å ) of diameter 2 mm. What diameter of light patch would the beam produce on the surface of the moon at a distance of $376 \times 10^3$km from the earth? You may neglect scattering in earth's atmosphere.

## 9.6 SOLUTIONS AND ANSWERS

SAQs

1. We know that angular spread of the central maximum is from

$$\theta = \sin^{-1}\left(\frac{\lambda}{b}\right) \text{ to } \theta = -\sin^{-1}\left(\frac{\lambda}{b}\right).$$

For small θ, we have $\sin\theta = \theta$ and we find that principal maximum is spread from $\theta = \dfrac{\lambda}{b}$ to $\theta = -\dfrac{\lambda}{b}$.

Similarly, you can show that the first secondary maximum on the positive side extends from $\theta_1 = \dfrac{\lambda}{b}$ to $\theta_1 = \dfrac{2\lambda}{b}$ and on the negative side from $\theta = -\dfrac{\lambda}{b}$ to $\theta = -\dfrac{2\lambda}{b}$.

Thus we see that the central maximum is twice as wide as a secondary maxima.

2. We know that

$$d\sin\theta_1 = \lambda$$

$$\therefore \qquad (0.3 \times 10^{-1}\ \text{cm})\sin\theta_1 = 4358 \times 10^{-8}\ \text{cm}$$

or

$$\sin\theta_1 = 1.45 \times 10^{-3}$$

In the small angle approximation we can lake

$$\theta_1 = 1.45 \times 10^{-3}\ \text{rad}$$

and

$$\theta_2 = 2.90 \times 10^{-3}\ \text{rad}$$

On comparing these values with those given in Example 1 for the first and second minima you will note that violet light is diffracted about 27% less.

## TQs

1. From Eq. (9.9) we know that the conditions for minima are given by

$$b\sin\theta = n\lambda \qquad n = \pm 1, \pm 2, \ldots$$

Here $b = 0.03$ mm $= 3 \times 10^{-3}$ cm, $n = 3$ and $\lambda = 5500$ Å

$$\therefore \qquad \sin\theta = \frac{n\lambda}{b} = \frac{3 \times (5500 \times 10^{-8}\ \text{cm})}{3 \times 10^{-3}\ \text{cm}} = 5.5 \times 10^{-4}$$

In the small angle approximation, $\sin\theta \approx \theta \approx \tan\theta$

$$\therefore \qquad x = 5.5 \times 10^{-4} \times (100\ \text{cm})$$

$$= 5.5 \times 10^{-2}\ \text{cm}$$

2. Suppose that the light patch on the Moon is taken to be Airy disc of diameter $x$ of a diffraction limited beam of initial diameter 2 mm. Then using Eq. (9.13) we can write

$$\sin\theta = \frac{1.22\,\lambda}{D} = \frac{1.22 \times (6300 \times 10^{-8}\ \text{cm})}{(0.2\ \text{cm})}$$

$$= 384.3 \times 10^{-6}$$

In the small angle approximation, $\sin\theta \simeq \theta = 384 \times 10^{-6}$ rad. Since $x = 2r\ \theta$, we find on substituting the numerical values that

$$x = 2 \times (376 \times 10^{3}\ \text{km}) \times (384.3 \times 10^{-6})$$

$$= 289\ \text{km}$$

# UNIT 10 DIFFRACTION GRATING

## Structure

## 10.1 INTRODUCTION

You have learnt about Fraunhofer diffraction produced by a single slit aperture. When a narrow vertical slit is illuminated by a distant point source, the Fraunhofer diffraction pattern consists of a series of spots situated symmetrically about a central spot, along a horizontal line. The intensity of the central spot is maximum and it decreases rapidly as we move away from the central spot. For a circular aperture, the diffraction pattern consists of concentric rings with a bright central disc. You also learnt that diffraction phenomenon limits the ability of optical devices to form sharp and distinct images of distinct objects. This restriction at one time hampered the spectroscopic work particularly for substances whose spectrum consisted of doublets. (Sodium doublet wavelengths correspond to 5890 Åand 5896 Å. Because of their proximity, these wavelengths seem to overlap.) But you will recall that diffraction pattern is sensitive to wavelength of light as well as the slit width. To take advantage of these it was thought that the problem could be overcome by increasing the number of diffracting slits. And the idea really worked. For simplicity, we have first discussed diffraction pattern by a double slit.

In Sec. 10.2 we have listed qualitative features of the observed double slit diffraction pattern and compared these with those of a single slit pattern. A distinct feature of double slit pattern is that it consists of bright and dark fringes similar to those observed in interference experiments. In Sec. 10.3 we have derived the equation for the resultant intensity distribution. This mathematical analysis is extension of what you have already learnt for single slit. You will learn that the intensity of the central maximum is four times the intensity due to either slit at that point. However, the interference maxima are diffused (broader). These results are generalised for the case of *N* equally spaced, identical slits in Sec. 10.4.

You will observe that as the number of slits increases, interference maxima get narrower (sharper). For sufficiently large value of *N*, interference maxima become narrow lines. For this reason, diffraction gratings are an excellent tool in spectral analysis. The occurrence of diffraction grating effects in nature is surprisingly common. Do you know that the green on the neck of a male mallard duck, blue appearance of wings of Morpho butterflies and the beautiful colours of the '*eye*' of the peacock's feathers are also due to diffraction grating effects? The layered structure in cat's retina acts as reflection grating and is responsible for mettalic green reflection at night.

### Objectives

After studying this unit, you should be able to

- state salient features of the double slit diffraction pattern
- qualitatively compare single-slit diffraction
- pattern with double and N-slit patterns
- derive equation for the intensity distribution for the double slit pattern
- extend the double-slit calculation for N equally spaced slits.
- describe the use of a diffraction grating in spectral analysis, and
- solve numerical examples.

## 10.2 OBSERVING DIFFRACTION FROM TWO VERTICAL SLITS

Refer to Fig. 10.1. It shows the experimental arrangement for observing diffraction from two vertical parallel slit - apertures in an opaque screen. Both the slits have same width($b$) and height($h$). The width of the intervening opaque space between the two slits is $a$. Therefore, the distance between two similar points in these

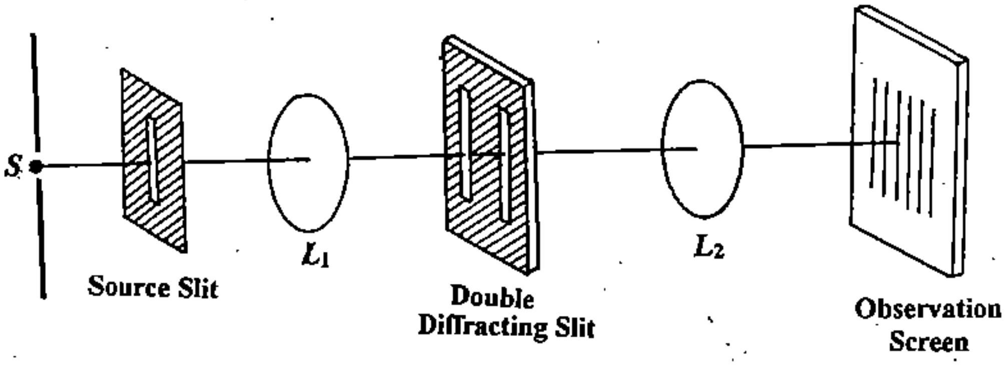


**Fig. 10.1: Experimental arrangment for observing diffraction from two identical vertical slits**

apertures $d = b + a$. Have you noticed that diffracting apertures are illuminated by a slit source rather than a point source of light? We have used this arrangement because this corresponds more nearly to the actual conditions under which an experiment is performed. That is, the diffraction pattern from a slit source is of greater practical importance than from a point source. The ray geometry of Fig. 10.1 for observing Fraunhofer diffraction from a double slit illuminated by a slit source is shown in Fig. 10.2. The length of the source slit in the arrangement should be adjusted to be parallel to the lengths of the diffracting slits.

Suppose we block one of the diffracting slits, say slit 1, shown in Fig. 10.1 and observe the diffraction pattern on the screen. Obviously, you should expect the

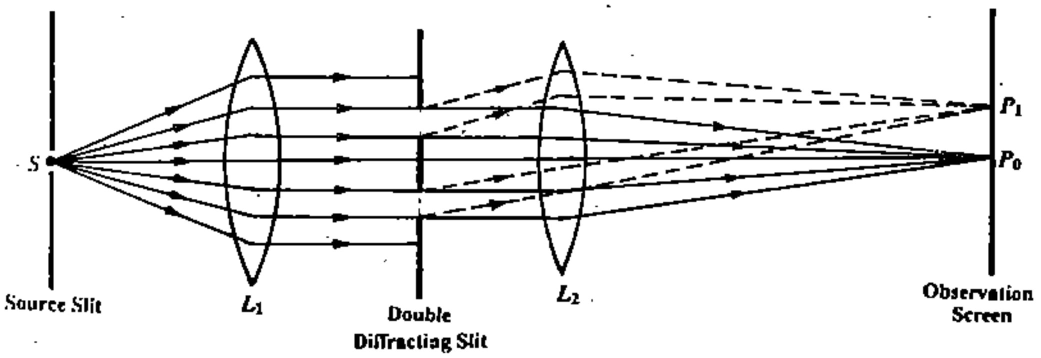


Fig.10.2: Ray geometry of experimental arrangement shown in Fig. 10.1

single slit diffraction pattern (due to slit number 2 which has not been blocked). Next, uncover slit 1 and block the other. You should again expect single slit diffraction pattern with exactly the same intensity distribution. But what may surprise you at the first glance is that both diffraction patterns are not only identical, they are located at the same position. Were you not expecting these diffraction patterns to be laterally displaced? These patterns are not laterally shifted with respect to one another because of the (well corrected) lens $L_2$. This is true even for $N$ identical vertical slits. The diffracted wavefronts originating from any slit, and travelling along the axis of lens $L_2$ are focussed at $P_0$, which forms the peak of the central spot. The diffracted wavelets moving at an angle θ are focussed at $P_\theta$.

In a well corrected lens consider parallel beams of light travelling in a direction inclined to the axis different parts of the lens. They are all brought to focus on the back focal plane at a poit which is located by the beam passing though the optical centre of the lens.

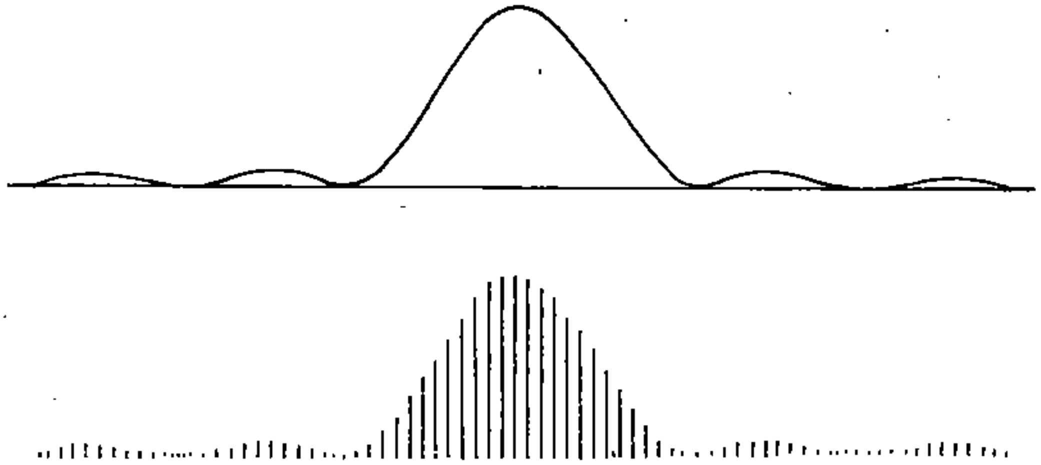


Fig. 10.3: Observed double slit diffraction pattern

Now uncover both the slits so that each slit gives its own diffraction pattern. The salient features of the resultant diffraction pattern, shown in Fig. 10.3, are summarised below:

(i) The double slit diffraction pattern consists of a number of equally spaced fringes similar to what is observed in interference experiments.

(ii) The intensities of all fringes are not equal. The fringes are the brightest in the central part of the pattern.

(iii) As we move away on either side of the central fringe, the intensity gradually falls off to zero.

(iv) The fringes reappear with reduced intensity three or four times and become too faint to observe thereafter.

(v) The intensity at the maximum of double slit pattern is greater than the intensity of principal maximum in single slit pattern.

What is responsible for this pattern? How bright are double slit fringes compared to those in the single slit pattern? You will discover answers to these and other related questions in the following section.

## 10.3 INTENSITY DISTRIBUTION IN DOUBLE SLIT PATTERN

For calculating the intensity distribution for the arrangement shown in Fig. 10.1 it is sufficient for us to consider a point source. This is because a point source gives the intensity distribution along a section perpendicular to the vertical fringes formed from a slit source. For deriving the equation for intensity of double slit pattern, we extend the procedure used for the single slit (Unit 9). Slit 1 acts as a source of diffracted plane wavefronts originating from points $A_1, A_2, A_3$, ... in it. We represent these by $a_0 \cos \omega t$, $a_0 \cos(\omega t - \phi)$, $a_0 \cos(\omega t - 2\phi)$, ..., where $\phi$ is the constant phase difference. The magnitude of electric field $E_1$ produced by this slit at the point $P_\theta$ is given by (Eq. 9.6):

$$E_1 = A\left(\frac{\sin\beta}{\beta}\right)\cos[(\omega t - \beta] \tag{10.1}$$

where $\beta = \dfrac{\pi\, b \sin\theta}{\lambda}$

For every point like $A_1$ in slit 1, we have a corresponding point $B_1$ in slit 2 at a distance $d$. The phase difference between diffracted wavefronts reaching $P_\theta$ from $A_1$ and $B_1$ is given by

$$\delta = \frac{2\pi}{\lambda}(a+b)\sin\theta = \frac{2\pi}{\lambda}d\sin\theta \tag{10.2}$$

Therefore, the diffracted plane wavefronts starting from points $B_1, B_2, B_3$, ... may be represented as $a_0\cos(\omega t - \delta)$, $a_0\cos(\omega t - \delta - \phi)$, $a_0 \cos(\omega t - \delta - 2\phi)$, ... And the field $E_2$ produced by slit 2 at $P_\theta$ is given by

$$E_2 = A\left(\frac{\sin\beta}{\beta}\right)\cos[(\omega t - \delta) - \beta] \tag{10.3}$$

Since the sources $A_1, A_2, A_3$, ... and $B_1, B_2, B_3$,... are coherent, the magnitude of resultant field at $P_0$ due to the double-slit is obtained by the superposition of magnitudes of individual fields:

$$E = E_1 + E_2$$

$$= A\frac{\sin\beta}{\beta}[\cos(\omega t - \beta) + \cos(\omega t - \beta - \delta)]$$

Using the trigonometric identity $\cos A + \cos B = 2\cos\left(\dfrac{A+B}{2}\right)\cos\left(\dfrac{A-B}{2}\right)$, we can rewrite the above expression as

$$E = 2A\left(\frac{\sin\beta}{\beta}\right)\cos\left[(\omega t - \beta) - \frac{\delta}{2}\right]\cos\left(\frac{\delta}{2}\right)$$

$$= 2A\left(\frac{\sin\beta}{\beta}\right)\cos(\omega t - \beta - \gamma)\cos\gamma \tag{10.4}$$

where $\gamma = \frac{\delta}{2} = \frac{\pi}{\lambda} d \sin\theta$.

The intensity is proportional to the square of the amplitude. So

$$I_\theta = 4A^2 \left(\frac{\sin\beta}{\beta}\right)^2 \cos^2\gamma \tag{10.5}$$

For $\theta = 0$, both $\beta$ and $\gamma$ vanish so that

$$I_{\theta=0} = 4A^2 = 4I_\theta$$

and the expression for intensity of double slit diffraction pattern can be written as

$$I_\theta = 4I_0 \left(\frac{\sin\beta}{\beta}\right)^2 \cos^2\gamma \tag{10.6}$$

Since the maximum value of $I_0$ is $4I_0$, we see that the double slit provides **four times** as much intensity in the central maximum as the single slit. This is exactly what you should have expected since the incident beams are in phase and amplitudes superpose.

If you closely examine Eq.(10.6) you will recognise that the term $(\sin^2\beta)/\beta^2$ represents the diffraction pattern produced by a single slit of width $b$. The $\cos^2\gamma$ term represents the interference pattern produced by two diffracted beams (of equal intensity) having phase difference $\delta$. That is, the intensity of double slit diffraction pattern is product of the irradiances observed for the double-slit interference and single slit diffraction. For $a > b$, the $\cos^2\gamma$ factor will vary more rapidly than the $(\sin^2\beta)/\beta^2$ factor. Then **we obtain Young's interference pattern for slits of very small widths.** In general, the product of sine and cosine factors may be considered as a modulation of the interference pattern by a single slit diffraction envelope. We shall discuss it in detail a little later.

Before we investigate the positions of maxima and minima, let us understand the physical phenomenon that takes place. Diffracted light emerging from these two slits constitutes two coherent beams. These interfere leading to the formation of fringes on the screen. But the intensity of a fringe depends upon the intensities of interfering beams and the phase difference between them when they reach the point under observation. We know that the intensities of diffracted beams are controlled by the diffraction conditions and the direction of observation. Consequently, the intensities of interference fringes is not the same at different points of the screen. In particular, in those directions in which the intensities of diffracted beams are large, the constructive interference will lead to brighter fringes whereas in directions where the two diffracted beams themselves have lower intensities, even their constructive interference will lead to faint fringes.

You should note that we have described the phenomenon as **interference between two diffracted beams.** How do we distinguish between the two words interference and diffraction which we have used? When secondary wavelets originating from different parts of the same wavefront are made to superimpose, we call it diffraction. Such a case arises when we consider all the wavelets arising from the various points situated in the aperture between the two jaws of a slit. But when two separate beams coming from two different slits are superimposed, we call it interference. It should be clear that in all cases where we apply the principle of superposition, the wavelets have to be coherent in nature to produce an observable pattern.

Before you proceed, you may like to answer an SAQ.

**SAQ 1**

If instead of a monochromatic source we use a source emitting two wavelengths, $\lambda_1$ and $\lambda_2$ ( $< \lambda_1$ ), how will the double slit diffraction pattern get influenced?

*Spend 2 min*

### 10.3.1 Positions of Minima and Maxima

To study the position of minima and maxima in the double slit pattern, we use the equation

$$I_\theta = 4I_0\left(\frac{\sin\beta}{\beta}\right)^2\cos^2\gamma$$

We note that the intensity $I_\theta$ will be zero when either ( $\sin\beta/\beta$ )$^2$ or $\cos^2\gamma$ is zero.

From Unit 9 you will recall that the factor ( $\sin\beta/\beta$ )$^2$ will be zero for

$$\beta = \frac{\pi b\sin\theta}{\lambda} = \pi, 2\pi, 3\pi, \ldots m\pi\ (\ m \neq 0\ )$$

or

$$b\sin\theta = \lambda, 2\lambda, 3\lambda, \ldots m\lambda \qquad (10.7)$$

This equation specifies the directions along which the available intensity of either beam is zero by virtue of diffraction taking place at each slit.

The second factor ($\cos^2\gamma$) will be zero when

$$\gamma = \frac{\pi d\sin\theta}{\lambda} = \frac{\pi}{2}, \frac{3\pi}{2}, \frac{5\pi}{2}, \cdots\left(n + \frac{1}{2}\right)\pi$$

or

$$d\sin\theta = \frac{\lambda}{2}, \frac{3\lambda}{2}, \frac{5\lambda}{2}, \cdots\left(n + \frac{1}{2}\right)\lambda \qquad (10.8)$$

This gives the angles for the intensity to be zero by virtue of destructive interference between two beams. You may recall that this is the same as the condition for the minimum of the interference pattern between two point sources. Eqs. (10.7) and (10.8) specify the direction when the intensity is zero.

We cannot obtain the exact positions of the maxima by any simple relation. This is because we have to find the maximum of a function which is product of two terms. But we can find their approximate positions if we assume that ( $\sin\beta/\beta$ ) does not vary appreciably over a given region. We are quite justified in making this approximation if the slits are very narrow. Note that we observe the maxima near the centre of the pattern. Under these conditions, the positions of maxima are solely determined by the $\cos^2\gamma$ factor. You know that this factor defines maxima for

$$\gamma = 0, \pi, 2\pi, \ldots n\pi$$

or

$$d\sin\theta = 0, \lambda, 2\lambda, \ldots, n\lambda \qquad (10.9)$$

We know that $d\sin\theta$ represents the path difference between the corresponding points in the two slits. When this path difference is a whole number of wavelengths, constructive interference occurs between the two beams. Then we get a maximum which leads to the formation of a series of bright fringes. The central fringe corresponds to $d\sin\theta = 0$. The $n$th fringe (on either side) occurs when $d\sin\theta = n\lambda$. We therefore say that $n$ represents the **order** of interference.

### 10.3.2 Missing Orders

In the intensity expression $I_\theta = 4I_0\left(\frac{\sin\beta}{\beta}\right)^2 \cos^2\gamma$, we have $\beta = \frac{\pi b \sin\theta}{\lambda}$ and $\gamma = \frac{\pi d \sin\theta}{\lambda}$

Thus we see that β and λ are not independent. These are connected to each other through the relation

$$\frac{\gamma}{\beta} = \frac{\pi d \sin\theta}{\pi b \sin\theta} = \frac{d}{b} = \frac{a+b}{b} \tag{10.10}$$

Cases of special interest arise when $d$ is an integral multiple of $b$, say it is an integer $P$ so that $d = pb$. This will happen when the opaque portion a is an integral multiple of the transparent part $b$. The possibilities are: $a = b$, $a = 2b$, or $a = 3b$ etc, so that $d/b = p = 2,3,4, \ldots$ etc in these cases. Under these conditions, the directions of diffraction minimum and interference maximum will necessarily coincide. To show this, let us assume that a direction of diffraction minimum is given by

$$b\sin\theta = m\lambda$$

We will automatically have the interference maximum in this direction since

$$d\sin\theta = (pb)\sin\theta = p\,b\sin\theta$$

$$= pm\,\lambda = n\lambda$$

where $n = pm$. The possible values of $p$ are 2,3,4, … and those of $m$ are 1,2,3, … Thus the $n$th order interference fringes for which $n = pm$ will have zero intensity since the intensity of both beams is zero by virtue of diffraction condition. As a result, their constructive interference also leads to net zero intensity. These are usually known as **missing orders**. For example, when $p = 2$, we will have 2,4,6, 8 … orders missing for $m$ values of 1,2,3,… etc. Similarly, when $p = 3$, we will have 3,6,9 … orders missing and so on.

The special case when $d/b = 1$, means that the opaque part $a = 0$ and the two slits exactly join one another. Then we find that all the interference orders are missing. Actually this means that we now have a single slit of double width and what we get is a single slit diffraction pattern and (with no interfernce fringes).

These ideas are illustrated in the following example.

**Example 1**

Consider a double slit arrangement with $b = 7.0 \times 10^{-3}$cm, $d = 3.5 \times 10^{-2}$cm and $\lambda = 6300$ Å. How many interference minima will occur between the diffraction minima on either side of the central maximum? If a screen is placed at a distance of 5m from the diffracting aperture, what is the fringe width?

**Solution**

The first diffraction minima on either side will occur when $b\sin\theta = \pm\lambda$. That is, for $\sin\theta = \pm\lambda/b = 9 \times 10^{-3}$. The interference b minima will occur when Eq. (10.8) is satisfied, i.e. when

$$1\,\mathrm{d}\,\mathrm{sin}\,d\sin\theta = \left(n + \frac{1}{2}\right)\lambda$$

On substituting the given values, we find that

$$\sin\theta = \left(n + \frac{1}{2}\right)\frac{\lambda}{d} = \left(n + \frac{1}{2}\right)1.8 \times 10^{-3} \qquad n = 0, 1, 2, \ldots$$

i.e.

$$\sin\theta = 0.9 \times 10^{-3}, 2.7 \times 10^{-3}, 4.5 \times 10^{-3}, 6.3 \times 10^{-3} \text{ and } 8.1 \times 10^{-3}$$

Thus there will be ten minima between the two first order diffraction minima. If $\theta$ is small we may write $\theta_1 = 0.9 \times 10^{-3}$ rad, $\theta_2 = 2.7 \times 10^{-3}$ rad, $\theta_3 = 4.5 \times 10^{-3}$ rad, $\theta_4 = 6.3 \times 10^{-3}$rad, $\theta_5 = 8.1 \times 10^{-3}$ rad and the angle between successive minima is $1.8 \times 10^{-3}$ rad.

The angular separation between two intereference maxima is given by

$$\Delta\theta = \frac{\lambda}{d} = \frac{6.3 \times 10^{-5}\text{cm}}{3.5 \times 10^{-2}\text{cm}} = 1.8 \times 10^{-3} \text{ rad.}$$

Note that this is the same as the angle between successive minima. Thus the fringe width $f$. $\Delta\theta$ d is

$$(500\text{ cm}) \times 1.8 \times 10^{-3} = 0.9\text{ cm}$$

### 10.3.3 Graphical Representation

We will now plot $\cos^2\gamma$, $(\sin^2\beta/\beta^2)$, and their product separately to study the double slit pattern. Before doing that we must decide on the relative scale of the abscissas $\gamma$ and $\beta$ since the shape of the pattern will depend upon this choice. You already know that $\gamma/\beta$ is equal $d/b$. Let us say that in a particular case $\gamma/\beta = d/b = 4$. We must then plot the proposed curves for $\gamma = 4\beta$. In Fig. 10.4, the curves (a) and (b) are plotted to the same scale of $\theta$.Fig. 10.4(a) depicts the curve for $\cos^2\gamma$ which given a set of equidistant maxima of equal intensity located at $\beta = 0, \pm\pi, \pm 2\pi, \pm 3\pi$ ...

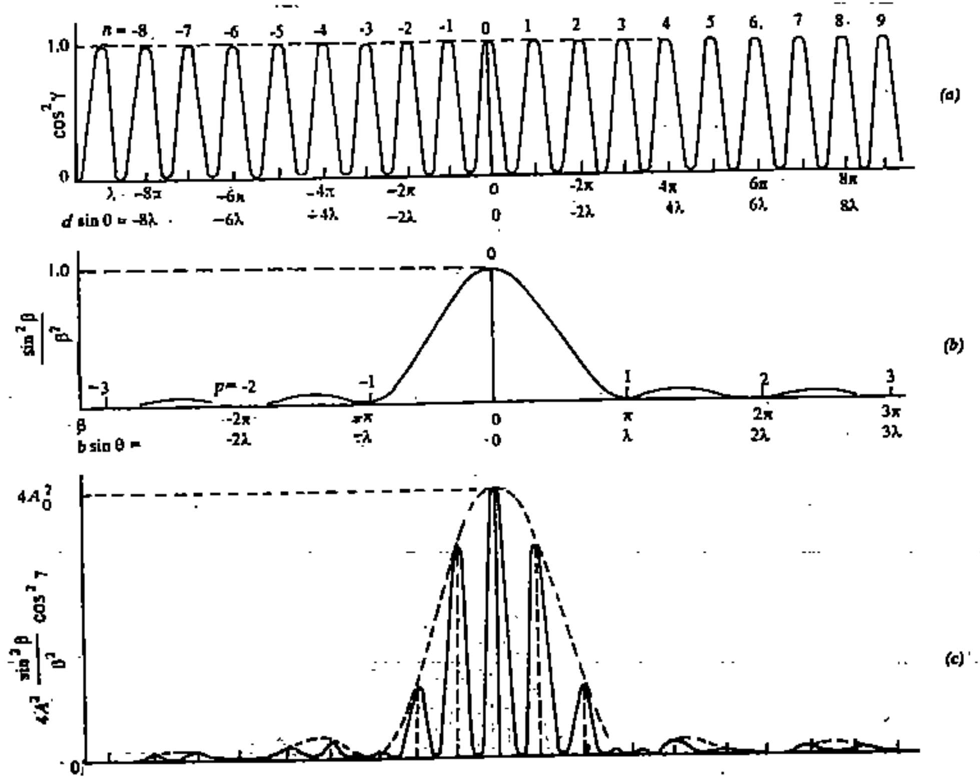


Fig. 10.4: Intensity curves for double slit. We have taken $\gamma = 4\beta$

In Fig. 10.4(b) we have plotted $(\sin\beta/\beta)^2$ which gives a maximum at $\beta = 0$ and minima at $\beta = \pm\pi, \pm 2\pi$, ...In Fig. 10.4(c) we have plotted their product. What do you observe? The intensity of the fringes in the resultant pattern is not the same as it was in Fig.10.4(a). It is modulated (reduced) by the factor $\frac{\sin^2\beta}{\beta^2}$. This means that the central fringe or the zeroth fringe is the brightest, and the successive three fringes are of decreasing intensity until we reach the point $\beta = \pi$ where the intensity is zero. Thus the fourth fringe corresponding to $\cos^2 \gamma = \pm 4\pi$ falls at $\beta = \pm \pi$ or $- \pi$ and their product is zero. Therefore, the fourth fringe on either side of the central maxima has zero intensity and its location at the angle satisfics simultaneously

$$B = \pm \pi \text{ and } \gamma = \pm 4\pi$$

or

$$b \sin\theta = \pm \lambda \text{ and } d \sin\theta = \pm 4\lambda$$

This fourth fringe will therefore be missing. We will observe the 5th, 6th and 7th fringes. We can argue in a similar manner that for 8th fringe

$$\beta = \pm 2\pi \text{ and } \gamma = \pm 8\pi$$

which will therefore have zero intensity and thus be missing.

You may now like to answer the following SAQ.

*Spend 2 min*

**SAQ 2**

Write down the general condition for missing orders in terms of the ratio *d*/*b*.

## 10.4 FRAUNHOFER PATTERN FROM *N* IDENTICAL SLITS

You now know that interference of waves diffracted by individual slits determines the intensity distribution in the double slit pattern. Let us now consider the diffraction pattern produced by *N* vertical slits. We use the same experimental arrangement as shown in Fig. 10.1 for two slits. For simplicity we assume that (i) each slit is of width *b* and has the same length (ii) all slits are parallel to each other and (iii) the intervening opaque space between any two successive slits is the same, equal to *a*. Therefore the distance between any two equivalent points in two consecutive slits is *a* + *b*. Let us denote it by *d* which we call the **grating element.** As before, we take the source of light to be in the form of a slit and adjust the length of this source slit to be vertical and parallel to the length of *N* slits. As arrangement consisting of a large number of parallel, equidistant narrow rectangular slits of the same width is called **diffraction grating.** As discussed in the double slit pattern, the diffraction pattern will consist of vertical fringes parallel to the slit source. Let us now study the intensity distribution in this pattern.

### 10.4.1 Intensity Distribution

To derive an expression for the intensity distribution we will follow the procedure and arguments similar to those used for the double slit. Consider a point source of light which sends out plane waves. That is, a plane wavefront is incident on the arrangement shown in Fig. 10.5. (Speaking in terms of ray-optics, we may say that light rays fall normally on the grating). You may recall that the intensity distribution along any section perpendicular to the vertical fringes formed from a slit source will be the same as obtained from a point source. Physically, light emerging from *N* slits after diffraction at each slit results in *N* diffracted beams. Since these are coherent, interference takes place between them resulting in the formation of fringes. It is important to note that diffraction controls the intensity from each slit in a give direction.

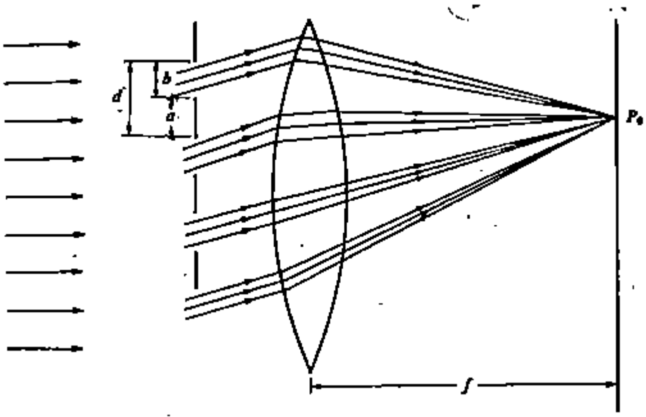


Fig. 10.5: Fraunhofer diffraction of a plane wave incident normally on a multiple slit aperture

As before, we consider the diffracted rays proceeding towards $P_\theta$, where θ is the angle between the diffracted rays and the normal to the grating. Let $E_1, E_2, E_3, \ldots E_N$ denote the fields produced by the first, the second,the third … and the $N$th slit at the point $P_\theta$. Then we have

$$E_1 = A\frac{\sin\beta}{\beta}\cos(\omega t - \beta)$$

$$E_2 = A\frac{\sin\beta}{\beta}\cos(\omega t - \beta - \delta)$$

$$E_3 = A\frac{\sin\beta}{\beta}\cos(\omega t - \beta - 2\delta)$$

$$\vdots$$

$$E_N = A\frac{\sin\beta}{\beta}\cos[\omega t - \beta - (N-1)\delta]$$

Where various symbols have the same meaning as in Sec. 10.3. Also, we have assumed that the phase changes by equal amount δ from one slit to the next.

The field $E$ at $P_\theta$ is obtained by summing these $N$ terms:

$$E = A\frac{\sin\beta}{\beta}\cos(\omega t - \beta) + A\frac{\sin\beta}{\beta}\cos(\omega t - \beta - \delta)$$

$$+ A\frac{\sin\beta}{\beta}\cos(\omega t - \beta - 2\delta) + \ldots$$

$$+ A\frac{\sin\beta}{\beta}\cos[\omega t - \beta - (N-1)\delta] \tag{10.11}$$

You can write it as

$$E = A\frac{\sin\beta}{\beta}\Big\{\cos(\omega t - \beta) + \cos(\omega t - \beta - \delta) + \ldots$$

$$+ \cos[\omega t - \beta - (N-1)\delta]\Big\}$$

In com.plex notation,

$\exp(i\theta) = \cos\theta + i\sin\theta$ (i)

so that

$\mathrm{Re}[\exp(i\theta)] = \cos\theta$ (ii)

It mears that

$\cos(\omega' - \beta) = \mathrm{Re}[e^{i(\omega t - \beta)}]$

$\cos(\omega' - \beta - \delta)$
$= \mathrm{Re}[e^{i(\omega t - \beta - \delta)}]$

$\cos(\omega' - \beta - (N-1)\delta)$
$= \mathrm{Re}[e^{i(\omega t - \beta - (N-1)\delta)}]$

$\therefore \cos(\omega t - \beta) +$
$\cos(\omega' - \beta - \delta) + \ldots =$
$\mathrm{Re}[e^{i(\omega t - \beta)} +$
$e^{i(\omega t - \beta - \delta)} + \ldots +$
$e^{i(\omega t - \beta - (N-1)\delta)}]$ (iii)

The RHS can be written as

$\mathrm{RHS} = e^{i(\omega t - \beta)}[1 + e^{-i\delta}$
$+ e^{-2i\delta} + \ldots + e^{-i(N-1)\delta}]$ (iv)

This is a geometric series with commc n factor $e^{-i\delta}$ and can be summed up easilyl using

$\left(S = \frac{1 - r^n}{1 - r}\right)$:

$\mathrm{RHS} = e^{i(\omega t - \beta)} \times \frac{1 - e^{-iN\delta}}{1 - e^{-i\delta}}$

$= e^{i(\omega t - \beta)} \times$
$\frac{e^{-iN\delta/2}}{e^{-i\delta/2}}\left(\frac{e^{iN\delta/2} - e^{-iN\delta/2}}{e^{i\delta/2} - e^{-i\delta/2}}\right)$
$= e^{i(\omega t - \beta - (N-1)\delta/2)}$

$\frac{\sin N\delta/2}{\sin\delta/2}$

Hence LHS of (iii) is recovered by the Real part, which is Eq. (10.12).

You have learnt to sum the series given here [Unit 2, Block 1 of the PHE–02 course on Oscillations and Waves Eq. (2.38)]. We have reproduced it in the Margin. The result is

$$E = A\left(\frac{\sin\beta}{\beta}\right)^2 \frac{\sin N\gamma}{\sin\gamma}\cos\left[\omega t - \beta - \frac{1}{2}(N-1)\delta\right] \qquad (10.12)$$

where $\gamma = \frac{\delta}{2} = \frac{\pi}{\lambda} d\sin\theta$. $\sin\gamma$ is referred to as the grating term.

The intensity of the resultant pattern is obtained by squaring the amplitude of the resultant field in this expression. Therefore,

$$I_\theta = A^2 \frac{\sin^2\beta}{\beta^2}\frac{\sin^2 N\gamma}{\sin^2\gamma} \qquad (11.13)$$

Let us pause for a while and ask: What have we achieved so far? We have obtained an expression for the resultant intensity of diffraction pattern from $N$-slits. We expect it to be true for any number of slits.

For a single slit, Eq. (11.13) reduces to

$$I_\theta = A^2 \frac{\sin^2\beta}{\beta^2}$$

which is the same as Eq. (9.7).

*Spend 2 min*

SAQ 3

Show that for $N = 2$, Eq. (10.13) reduces to Eq. (10.6) for the double slit.

### 10.4.2 Positions of Principal Maxima

For obtaining the positions of maxima (as well as minima), let us re-examine Eq. (11.13). We note that the intensity distribution is a product of two terms; the first terms ( $\sin^2\beta/\beta^2$ ) represents the diffraction pattern produced by a single slit whereas the second the term ( $\sin^2 N\gamma/\sin^2\gamma$ ) represents the interference pattern of $N$ slits. The interference term controls the width of interference fringes, while the diffraction term governs their intensities.

As in case of the double slit, we cannot locate the exact positions of maxima; their approximate positions can however be obtained by neglecting the variation of $\sin^2\beta/\beta^2$. This is quite justified for very narrow slits. Therefore, for obtaining the positions of maxima we consider only the interference term.

The maximum value of $\frac{\sin^2 N\gamma}{\sin^2\gamma}\,(=N^2)$ occurs for $\gamma = 0, \pi, 2\pi, \ldots n\pi$. At the first glance, you will note that the quotient becomes indeterminate at these values. In such a situation, we compute the first derivative of the numerator as well as the denominator separately before inserting the value of argument. Following this procedure you will readily obtain

$$\lim_{\gamma\to n\pi}\frac{\sin N\gamma}{\sin\gamma} = \lim_{\gamma\to n\pi}\frac{N\cos N\gamma}{\cos\gamma} = \pm N$$

so that

$$\left(\frac{\sin N\gamma}{\sin\gamma}\right)^2 = N^2$$

The expression for intensity now takes the form

$$I_\theta = A^2 \frac{\sin^2 \beta}{\beta^2} N^2 = N^2 A^2 \frac{\sin^2 \beta}{\beta^2} \qquad (10.14)$$

where $\beta = \frac{\pi\, b \sin\theta}{\lambda}$.

We therefore conclude that the positions of maxima are obtained when

$$\gamma = 0, \pi, 2\pi, \ldots\, n\pi \text{ or } N\gamma = 0, N\pi, 2N\pi, \ldots\, Nn\pi \qquad (10.15)$$

Physically, at these maxima the fields produced by each of the slits are in phase and the resultant field is N times the field due to each of the slits.

When $N$ is large, the intensity, being proportional to $N^2$, is very large and we will obtain intense maxima, if only $\sin^2 \beta/\beta^2$ is not too small. Such maxima are known as **principal maxima.**

We can rewrite the condition of principal maxima as

$$d \sin\theta_{max} = n\lambda \qquad (10.16)$$

which is identical to Eq. (10.9). It implies that

1. The principal maxima in $N$-slit pattern correspond in position to those of the double slit.
2. The relative intensities of different orders are modulated by the single slit diffraction envelop.
3. $n$ cannot be greater than $d/\lambda$ since $|\sin\theta| \le 1$. Can you imagine the implications of this condition? If you ponder for a while, you will realise that this condition suggests existence of only a finite number of principal maxima, which are designated as the first, second, third, ... **order** of diffraction. Moreover, there will be as many first order principal maxima as the number of wavelengths in the incident wave.
4. The relation between $\beta$ and $\gamma$ obtained for double slit in terms of slit width and slit separation does not change. That is, Eq. (10.10) hold for $N$-slits as well.

### 10.4.3 Minima and Secondary Maxima

To be able to find the minima in the diffraction pattern, we locate the minima of the interference term. We note that the numerator in $\sin^2 N\gamma/\sin^2\gamma$ will become zero more often than the denominator. The numerator becomes zero for $N\gamma = 0, \pi, 2\pi, \ldots\, p\pi$, or $\gamma = \frac{p\pi}{N}$. Therefore, $\sin\gamma \left(= \sin \frac{p\pi}{N}\right)$ will not become zero for all integral values of $p$. It will become zero only for **special** cases when $p = 0, N, 2N, \ldots$ and $\gamma$ assumes values which are integral multiple of $\pi$. But you will recall that for these special values of $\gamma$, both $\sin N\gamma$ and $\sin\gamma$ vanish and the interference term defines the positions of principal maxima already discussed. However, for all other values of $p$, the numerator vanishes but not the denominator. That is, **intensity vanishes** when $p$, though an integer, is not an integral multiple of $N$. Hence, the condition for minimum is $\gamma = p\,\pi/N$ except when $p = nN$; $n$ being the order. These values correspond to

$$\gamma = \left[\frac{\pi}{N}, \frac{2\pi}{N}, \ldots, \frac{(n-1)\pi}{n}, \frac{(N+1)\pi}{N}\right], \left[\frac{(N+2)\pi}{N}, \ldots \frac{(2N-1)\pi}{N}\right], \left[\frac{(2N+1)\pi}{N}, \ldots\right]$$

These values of $\gamma$correspond to path difference

$$d\sin\theta_{min} = \left[\frac{\lambda}{N}, \frac{2\lambda}{N}, \frac{3\lambda}{N}, \frac{(N-1)\lambda}{N}\right] \cdots \left[\frac{(N+1)\lambda}{N} \cdots\right]$$

You should note that the values 0, $\frac{N\lambda}{N}$, $\frac{2N\lambda}{N}$, $\cdots$, which correspond to $d\sin\theta_{max} = n\lambda$ and represent principal maxima, are omitted.

Let us now summarise what you have learnt in this unit so far.

**The condition for principal maxima:**

$$\gamma = 0, \pi, 2\pi, \ldots, n\pi$$

and therefore

$$N\gamma = 0, N\pi, 2N\pi, \ldots, nN\pi$$

We may write

$$\gamma = \frac{\pi d}{\lambda}\sin\theta_{max} = n\pi \text{ where } n = 0, 1, 2, \ldots$$

In terms of path difference,

$$d\sin\theta_{max} = n\lambda$$

**The conditions for minima:**

$$N\gamma = nN\pi \pm \pi, nN\pi \pm 2\pi, \ldots, nN\pi \pm q\pi$$

where $q$ is not an integral multiple of $N$. We can rewrite it as

$$\gamma = n\pi \pm \frac{\pi}{N}, n\pi \pm \frac{2\pi}{N}, \cdots$$

In terms of path difference

$$d\sin\theta_{min} = n\lambda \pm \frac{\lambda}{N}, n\lambda \pm \frac{2\lambda}{N}, \cdots, n\lambda \pm \frac{q\lambda}{N}$$

where $q \neq 0, N, 2N, \ldots$

If you write all possible values of $N\gamma$, you will find that we have $(N - 1)$ positions of minima between any two successive principal maxima. Further, we know that between any two consecutive minima, there has to be a maxima. Such maxima are said to be **secondary maxima.** There will be $(N - 2)$ positions of secondary maxima between two consecutive principal maxima. The secondary maxima are not symmetrical, as in the two slit pattern. Moreover, the intensity of secondary maxima is very small and are therefore of little practical importance. Typical diffraction patterns and the corresponding intensity distributions predicted by Eq. (10.13) for $N = 4$ are shown in Fig. 10.6.

You may now like to answer the following SAQ.

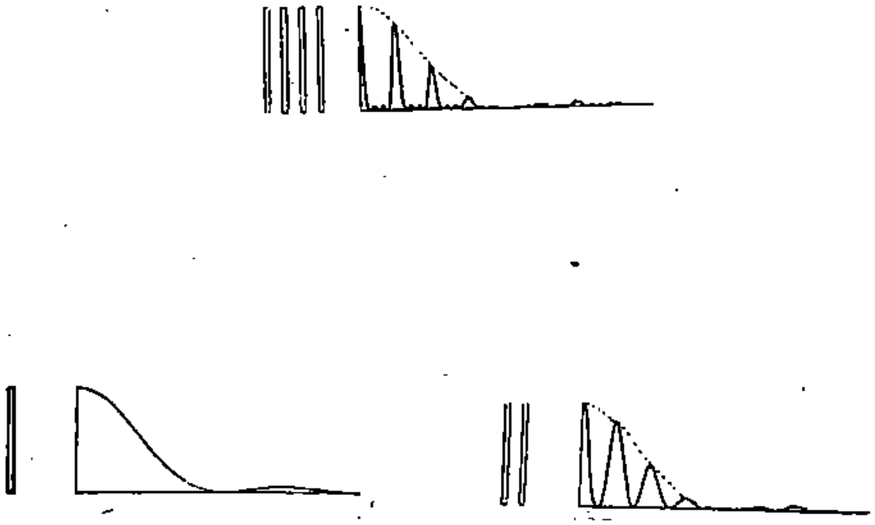


**Fig. 10.6: Fraunhofer diffraction pattern for four slits. For comparison, patterns for one and double slits are also shown. The intensity distribution predicted by Eq. (10.13) is also shown.**

SAQ 5

Show schematically the positions of principal maxima, secondary maxima and secondary minima for a diffraction grating with 6 slits.

*Spend 2 min*

**Hint:** We expect 5 minima between two consecutive principal maxima. Also we have 4 secondary maxima between the two principal maxima.

**Example 2**

**Calculate the maximum number of principal maxima that can be formed with a grating 5000 lines per cm for light of wavelength 5000 Å.**

$$\text{Grating element } d = \frac{1}{5000 \times 10^{-8}\text{cm}} = 2 \times 10^{-4}\text{cm}$$

**The condition for the formation of principal maxima is $d \sin_{max} = n\lambda$. Since $|\sin\theta| < 1$ we cannot have $n$ greater than $\frac{d}{\lambda}$ In this specific case**

$$n = \frac{2 \times 10^{-4}\text{cm}}{5000 \times 10^{-8}\text{cm}} = 4$$

**Therefore, it will be able to show 1st, 2nd, 3rd and 4th orders of principal maxim..**

**If, on the other hand, we have a grating with 15000 lines $\text{cm}^{-1}$**

$$n = \frac{(1/15000^{-1}\text{cm})}{5 \times 10^{-5}\text{cm}} = \frac{6.6 \times 10^{-5}\text{cm}}{5 \times 10^{-5}\text{cm}}$$

which is less than 2. Such a grating will show only 1st order of spectrum with $\lambda = 5000$ Å. You can verfing this result while observing grating spectrum in your second level physics laboratory course.

### 10.4.4 Angular Half-Width of Principal Maxima

You now know that for $N$ slits

1. The principal maxima occur when $\gamma = n\pi$ and thertefore $N\gamma = Nn\pi$.
2. On wither side of the principal maxima, we have a minimum when $N\gamma = nN\pi \pm \pi$ or when $\gamma = n\pi \pm \frac{\pi}{N}$. In terms of path difference and angle of diffraction, these conditions for principal maxima and the adjacent minimum can be rewritten as

$$d\sin\theta_{max} = n\lambda \tag{10.16}$$

and

$$d\sin\theta_{min} = n\lambda \pm \frac{\lambda}{N} \tag{10.17}$$

The angle between $\theta_{max}$ and $\theta_{min}$ is called the anular half width of principal maxima, let us denote it by $\delta\,\theta$. We now proceed to calculate this angle. We can calculate $\delta\theta\ (= |\,\theta_{max} - \theta_{min}\,|)$ by computing $\theta_{max}$ and $\theta_{min}$ from Eqs. (10.16) and (10.17). Alternatively, by choosing $\theta_{min} > \theta_{max}$, we substitute $\theta_{min} = \theta_{max} + \delta\,\theta$ in Eq. (10.17) to obtain

$$d\sin(\theta_{max} + \delta\,\theta) = n\lambda + \frac{\lambda}{N}$$

or

$$d\sin\theta_{max}\cos\delta\,\theta + d\cos\theta_{max}\sin\delta\,\theta = n\lambda + \frac{\lambda}{N}$$

For $\delta\,\theta \to 0$, $\cos\delta\,\theta \to 1$ and $\sin\delta\,\theta \to \delta\,\theta$. Hence

$$d\sin\theta_{max} + d\cos\theta_{max}\,\delta\,\theta = n\lambda + \frac{\lambda}{N}$$

Using Eq. (10.16), we find that it takes a compact form:

$$d\cos\theta_{max}\,\delta\,\theta = \frac{\lambda}{N}$$

so that

$$\delta\,\theta = \frac{\lambda}{N\,d\cos\theta_{max}} \tag{10.18}$$

which shows that the principal maximum becomes sharper as $N$ increases. It is for this reason that grating spectrum is so sharp. You will now tearn about it in detail.

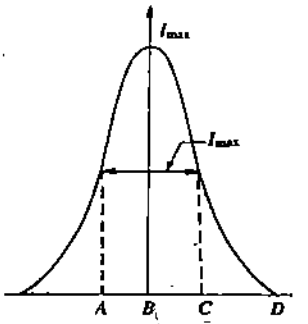
You may now question as to why is $\delta\,\theta$ called **angular half width**. It is quite simple. You know that the principal maximum extends from minimum on one side to minimum on the other side and $\delta\,\theta$ is half of it. While solving SAQ 4 you have seen that for 6 slits the principal maximum extends from

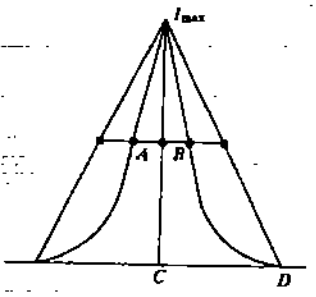
$$N\gamma = 5\pi \text{ to } N\gamma = 7\pi$$

or

$$d\sin\theta_{max} = \frac{5\lambda}{6} \text{ to } \frac{7\lambda}{6}$$

You must note that the term **half width** of a spectrum line (or a diffraction curve) has a slightly different meaning. The diagram whown represents the Intensigy vs $\theta$ curve. The half width gives the width of the curve at $\frac{I_{max}}{2}$. It is equal to $AB$ in the diagram. The angular half width, on the other hand, is equal to $CD$. Obviously you can convince yourself that $AB$ is not equal to $CD$. Only in the extreme case when the curve is a triangle, $AB = CD$.





## 10.5 DIFFRACTION GRATING

You have learnt about the diffraction pattern produced by a system of parallel equidistant slits. An arrangement of a large number of equidistant narrow vertical slits is known as diffraction grating. The first gratings were made by Fraunhofer. He stretched fine silver wire on a frame. His grating had nearly 200 wires to a centimeter. Afterwards gratings were made by ruling fine lines with a diamond pen on a glass plate. The transparent part between the lines acted as a slit while the ruling itself acted effectively as the opaque part. Rowland was among the first to rule gratings on a metallic surface. He produced plane as well as concave gratings with nearly 5000 lines per centimeter. These gratings are difficult to make and are expensive but celluloid replicas can be made fairly cheaply and are commonly used in the physics laboratory for spectral analysis. You can make a simple coarse grating for demonstration purposes on a plate by drawing equidistant and parallel scratches on the photographic emulsion. Now-a-days it is possible to produce gratings holographically. Holographic gratings have greater rulings per cm and are definitely better than ruled gratings. You will get an opportunity to learn holographic details in Block-4.

### 10.5.1 Formation of Spectra

We have seen that for a monochromatic light of wavelength $\lambda_1$, the principal maxima are given by the grating equation

$$d \sin \theta_1 = n \lambda_1 \qquad n = 0, 1, 2, 3, \ldots$$

With the experimental arrangement described above we will get these principal maxima as one line in each order. Now if another source of light emits a longer wavelength $\lambda_2$, we will get a corresponding line in each order at a larger angle $\theta_2$:

$$d \sin \theta_2 = n \lambda_2 \qquad n = 0, 1, 2, 3, \ldots$$

However if the same source of light emits both the colours corresponding to wavelengths $\lambda_1$and $\lambda_2$, we will get two lines simultaneously in each order. These two lines will be seen as two spectrum lines separated from each other. This is because except the central maximum (zeroth order), the angles of diffraction for $\lambda_1$ and $\lambda_2$ are different in various other orders. In the central maxima $\theta = 0$ for all wavelengths and this is why different colours are not separated from each other. What do you expect to observe when we have a white light source? The central image will be white while all other orders will show colours.

We note that in the grating equation, if we know $d$, $\theta$ and $n$, we can calculate the wavelength of light. Since the grating element ($d$) is known for a grating and $\theta$ can be measured, this arrangement provides a simple and accurate method of measuring $\lambda$. This is discussed in the following section.

### 10.5.2 Observing Grating Spectra

In your second level physics laboratory course, you must have observe grating spectra using a simple spectrometer. This arrangement is depicted in Fig. 10.7. The light from the given source is focussed (with the help of a lens) on the slit of the collimator which sends out a parallel beam of light.

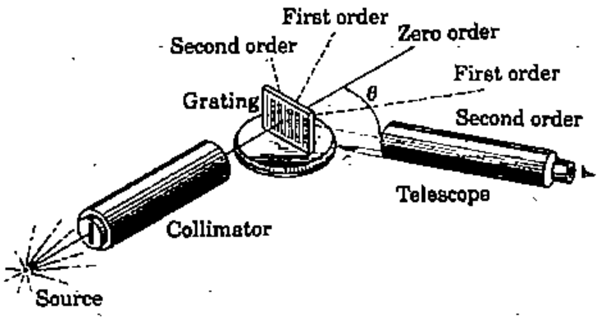


**Fig. 10.7: A schematic diagram of experimental arrangement for observing grating spectra**

The telescope arm is rotated and brought in line with the collimator. This ensures that the parallel beam of light falling on the objective of telescope is focussed at the crosswires, which is in the focal plane of the eye piece. The position of the source of light should be adjusted to get the brightest image. We mount the diffraaction grating on the turntable and adjust it so that the light is incident normally on the grating. Now we rotate the telescope arm to the left or right to get the first order spectrum in the field of view. If the source of light is a discharge tube containing sodium, mercury or argon the spectrum will consist of a series of spectrum lines: Each spectrum line is a diffracted image of the slit, formed by different wavelengths present in the source. To get sharp line images, we adjust the grating so that the diffracting slits are parallel to the collimator slit. This can be done by rotating the grating in its own plane.

To measure the wavelength of each line, we set the vertical crosswires at the centre of each spectrum line and note the position of the telescope in each case. The difference between the position of the telescope and the direct position gives the angle of diffraction for each of the lines. To reduce error, the position of the telescope is noted on both sides of the direct position and half of this angle gives the angle of diffraction.

Light from a molecule gives a band like appearance and is oftem called band spectum, while an incandescent lamp of similar sources will give a continuous spectrum, where various colours merge into one another.

You must have observed that

1. The spectrum exists on both sides of the direct beam
2. Apart from the first order, the second or even third order spectrum (depending upon the grating element) are also present.
3. Different spectrum lines are not equally bright or sharp. This depends upon the energy levels and the transitions of the atom giving the spectrum. These concepts are further illustrated in the following example.

**Example 3**

Rowland ruled 14438 lines per inch in his grating. (i) Calculate the angles of diffraction for violet ( $\lambda = 4000$ Å ) and red ( $\lambda = 8000$ Å ) colours in the first order of spectrum. What is the largest wavelength which can be seen with this grating in the third order?

**Solution**

(i) The grating element $d = \dfrac{2.54 \text{ cm}}{14438} = 0.0001759$ cm

$$= 1.759 \times 10^{-4} \text{cm}$$

Suppose that the violet colour ( $\lambda = 4000$ Å ) is diffracted through angle $\theta_\nu$. Recall the condition for maximum:

$$d \sin \theta_\nu = n\lambda$$

For first order on substituting the given values, you will get

$$\sin \theta_\nu = \frac{4 \times 10^{-5}}{1.759 \times 10^{-4} \text{cm}} = 0.2274$$

Therefore $\theta_\nu \approx 13^\circ$

Similarly, for red colour ( $\lambda = 8000$ Å ), we have

$$\sin \theta_r = \frac{8 \times 10^{-5} \text{cm}}{1.759 \times 10^{-4} \text{cm}} = 0.4548$$

so that

$$\theta_r \doteq 27^\circ$$

This means that the entire visible spectrum in the first order extends from nearly $\theta = 13^\circ$ to $\theta = 27^\circ$, i.e. covers an angle of about 14°.

(ii) $$d \sin\theta = 3\,\lambda_{max}$$

According to the given condition, $\theta = 90^\circ$ so that $\sin\theta = 1$ and $d = 3\lambda_{max}$

or

$$\lambda_{max} = \frac{d}{3} = \frac{1.759 \times 10^{-4}}{3}\,\text{cm} = 5860\,\text{Å}$$

This calculation suggests that in the third order spectrum, the sodium doublet consisting of 5890 Å and 5896 Å will not be visible. Do you recall this from your observations on spectral analysis using a diffraction grating? If you have so far not opted for the second level physics, it will be worthwhile to verify this result.

If you calculatesin $\theta_v$ and sin $\theta_r$ for 1st, 2nd and 3rd orders, you will find that for

$$\text{1st order} \quad \left.\begin{array}{l} \sin\theta_v = 0.2274 \Rightarrow \theta_v \sim 13^\circ \\ \sin\theta_r = 0.4548 \Rightarrow \theta_r \sim 27^\circ \end{array}\right] \Rightarrow 14^\circ \text{ spread}$$

$$\text{2nd order} \quad \left.\begin{array}{l} \sin\theta_v = 0.4548 \Rightarrow \theta_v \sim 27^\circ \\ \sin\theta_r = 0.9096 \Rightarrow \theta_r \sim 65^\circ \end{array}\right] \Rightarrow 38^\circ \text{ spread}$$

$$\text{3rd order} \quad \left.\begin{array}{l} \sin\theta_v = 0.6822 \Rightarrow \theta_v \sim 43^\circ \\ \sin\theta_{max} = 1 \text{ for } \lambda_{max} \\ \qquad = 5860\,\text{Å and } \theta_{max} = 90^\circ \end{array}\right] \Rightarrow 47^\circ \text{ for } 4000\,\text{Å} - 60000\,\text{Å}$$

$\sin\theta_r > 1$ and cannot be observed. $\Rightarrow$ entire visible spectrum is not available.

Schematically it is shown below:

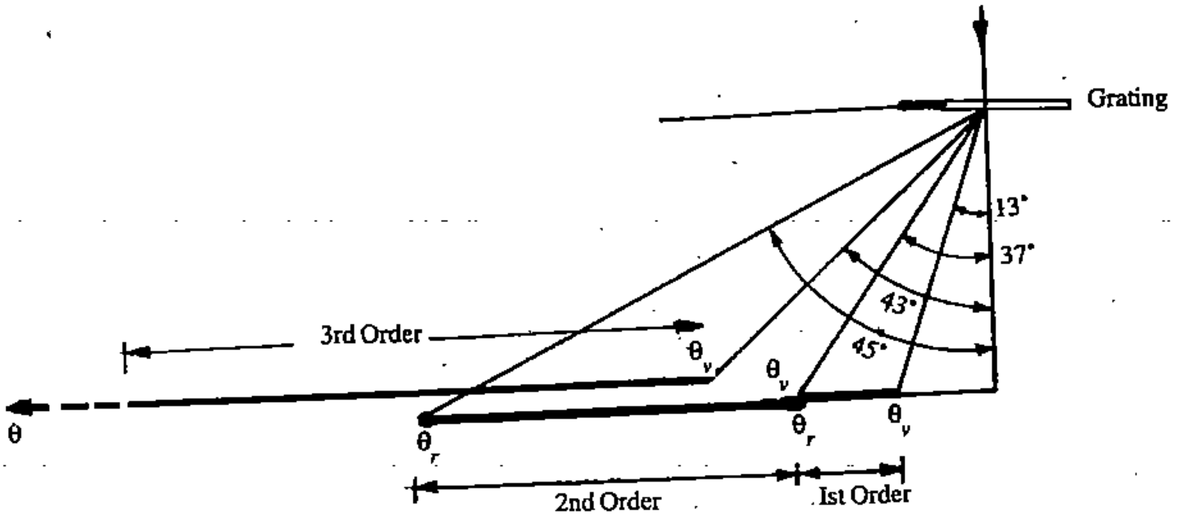


Fig. 10.8: Schematics of angles for over-all spread of various orders of spectrum

Thus we find that in 1st order red just touches second order violet. (This is because we have selected $\lambda = 4000$ Å and $\lambda = 8000$ Å ). It means that there is essentially no overlapping of first and second order spectra. The third order $\lambda_v$ begins at $\theta \simeq 43°$ If you calculate wavelength $\lambda_x$ of 2nd order present there you will find that

$$d \sin 43° = 3\lambda_v = 2\lambda_x \Rightarrow \lambda_x = \frac{3 \times 4000\,\text{Å}}{2} = 6000\,\text{Å}.$$

Therefore $\lambda = 6000$ Åof the 2nd order occurs at the same place as $\lambda = 4000$ Å of third order. Therefore, from 6000 Ato 8000 A will have overlapping colours. This difficulty is usually avoided by using suitable colour filters.

We now summarise what you have learnt in this unit.

## 10.6 SUMMARY

- The double slit diffraction pattern consists of a number of equally spaced fringes similar to what is observed in interference experiments. These fringes are the brightest in the central part of the pattern.
- In double slit pattern fringes reappear three or four times before they become too faint to observe.
- The central maximum in double slit pattern is four times brighter than that in single slit pattern.
- The intensity of double slit diffraction pattern at an angle θ is given by

  $$I_\theta = 4I_0 \frac{\sin^2 \beta}{\beta^2} \cos^2 \gamma$$

  Here $I_0 = A^2$, $\beta = \dfrac{\pi b \sin\theta}{\lambda}$ and $\gamma = \dfrac{\pi}{\lambda} d \sin\theta$; where $b$ is slit width and $d$ is distance between two similar points in these apertures. It is equal to $a + b$, where $a$ is the width of the intervening opaque space between two slits.
- The intensity of double slit diffraction pattern is product of the irradiances observed for the double slit interference and single slit diffraction. Physically, it arises due to interference between two diffracted beams.
- For slits of very small widths, the double slit diffraction pattern reduces to Young's interference pattern.
- The conditions of maxima and minima in double slit pattern are:

  $$d \sin\theta = n\lambda \qquad (\text{maxima})$$

  and

  $$b \sin\theta = m\lambda \qquad (\text{minima})$$
- The intensity distribution in $N$-slit diffraction pattern is given by

  $$I_\theta = A^2 \frac{\sin^2 \beta}{\beta^2} \frac{\sin^2 N\gamma}{\sin^2 \gamma}$$

The term sin $\gamma$ is referred to as the grating term.

- As the number of slits increases, the maxima get narrower and for sufficiently large values of $N$, they become to sharp lines. The angular half width of principal maximum $\delta\theta$ is given by

$$\delta\theta = \frac{\lambda}{N\, d \cos\theta_{max}}$$

The principal maximum is sharp for large values of $N$.

## 10.7 TERMINAL QUESTIONS

1. If we use a white light source in the arrangement shown in Fig. 10.6, how will if affect the fringes?

2. Can there be principal maxima of zero intensity because of diffraction at each slit? If yes, discuss.

## 10.8 SOLUTIONS AND ANSWERS

**SAQs**

1. $\lambda_1$ will give its diffraction pattern within which we will get the interference fringes. The pattern for $\lambda_2$ will be smaller if $\lambda_2 < \lambda_1$. They will both be supperimposed on one another coinciding at $\theta = 0$.

2. The general conditions for missing orders in terms of $\gamma$ and $\beta$ are $\gamma = \pm m\pi$ or $d \sin\theta = \pm m\lambda$ and $\beta = \pm p\pi$ or $b \sin\theta = p\lambda$. Therefore

$$\frac{d}{b} = \frac{m}{p}$$

both $m$ and $p$ are integers, the missing orders occur when $d/b$ is a ratio of two integers. When $d/b = 1$, i.e. the two slits exactly join, all the interference orders are missing. Physically it means that we have a single slit of double width and consequently no interference.

For $\frac{d}{b} = 2$, second, fourth, sixth, ... orders will be missing. What do you say about $\frac{d}{b} = 3$ ?

3. For $N = 2$, Eq. (10.13) takes the form

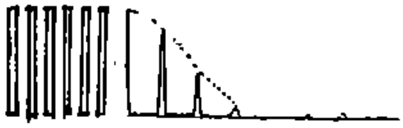
$$I_\theta = A^2 \frac{\sin^2\beta}{\beta^2} \frac{\sin^2 2\gamma}{\sin^2\gamma}$$

$$= A^2 \frac{\sin^2\beta}{\beta^2} \frac{(2\sin\gamma\ \cos\gamma)^2}{\sin^2\gamma}$$

$$= 4A^2 \frac{\sin^2\beta}{\beta^2} \cos^2\gamma$$

which is the required result for the double slit.

4. See figure given below:



## TQs

1. As before, each wavelength will give its interference fringes. The central fringe for all wavelengths will coincide and hence the central fringe will be white. Fringes of order $n = 1, 2, 3, \ldots$ located on either side of the central fringe, at different $\theta$ values given by $d \sin \theta = n \lambda$ for different wavelengths will be coloured.

2. There can be a principal maxima whose intensity is zero because of the diffraction at each slit. There are called **missing orders** or **absent spectra.** We know that the relationship between $\beta$ and $\gamma$ in terms of slit width and slit separation for $N$ slits is the same as for the double slit. Therefore, the conditions for missing orders remain unaltered. And a particular maximum will be absent if it is formed at the same angle as the minimum of single slit diffraction pattern. This occurs at an angle which satisfies Eqs. (10.16) and (10.17).

# UNIT 11 DIFFRACTION AND RESOLUTION

### Structure

## 11.1 INTRODUCTION

In the preceding two units of this block you have learnt that due to diffraction, the image of an object is fringed even if an aberration-free converging lens is used. That is, image of a point object is spread over a small area on the observation screen. Does this mean that no optical device can form a perfect image? The answer to this question is: The image of a point source is not geometrical point. And diffraction does place a limit on the ability of optical devices to transmit perfect information (quality image) about any object. Such optical systems are said to be diffraction limited.

Broadly speaking, diffraction limited systems can be classified into two categories: (i) Human eye, microscope and telescope which enable us to see two objects (near or distant) distinct and (ii) Grating and prism which form a spectrum and enable us to see two distinct wavelengths (colours). In principle, in both types of instruments two close fringed (diffraction) images are formed on the screen. The question that should logically come to your mind is: How to characterise the ability of an optical instrument to distinguish two close but distinct diffraction images of two objects or two wavelengths? This ability is measured in terms of resolving power. You may now like to know: What criterion enables us to compute resolving power? The most widely used criterion is due to Rayleigh. According to this, two diffraction images are said to be just resolved when the first minimum of diffraction pattern of one object falls at the same position where the central maximum of the diffraction pattern of the other lies. When the patterns come closer than this, the objects are not resolvable. When the patterns overlap less than this, the images are distinct and hence objects are resolvable. It is also important to know whether the same criterion applies to both types of optical devices? How can we improve resolution and see deeper in space even during the day? We have addressed all these aspects in this unit.

### Objectives

After studying this unit, you should be able to

- explain how diffraction limits image forming ability of optical devices
- use Rayleigh criterion to compute expressions for resolving power of a telescope, a microscope and a diffraction grating

- solve numerical problems based on resolution, and
- describe how Michelson stellar interferometer helps in improving resolution.

## 11.2 DIFFRACTION AND IMAGE FORMATION

You may recall from Unit 9 that when the size of pupil is greater than 2.4 mm, the human eye does not form a perfect point image (due to aberrations). However, for pupil sizes smaller than 2.4 mm, the human eye appears to be a diffraction-limited system. To gain some quantitative measure of visible acuity, let us estimate the size of image formed on our retina. If we approximate the pupil in human eye by a circular aperture, we have to consider how it influences the image formed by eye-lens on the retina (Fig. 11.1). From Unit 9 you may recall that the diffraction image of a point source due to a circular aperture is a bright central disc surrounded by a series of alternate dark and bright rings of decreasing intensity. The angular half-width of the central disc is given by $\theta = 1.22\,\lambda/D$ where $D$ is the diameter of the aperture. And the lateral width of this image will be $f\theta$, where $f$ is the focal length of eye-lens. This means that the size of an image formed on retina depends on the wavelength of light and diameter of the aperture. If we take the pupil diameter to be 2 mm, then for middle of visible spectrum ( $\lambda = 5500$ Å )

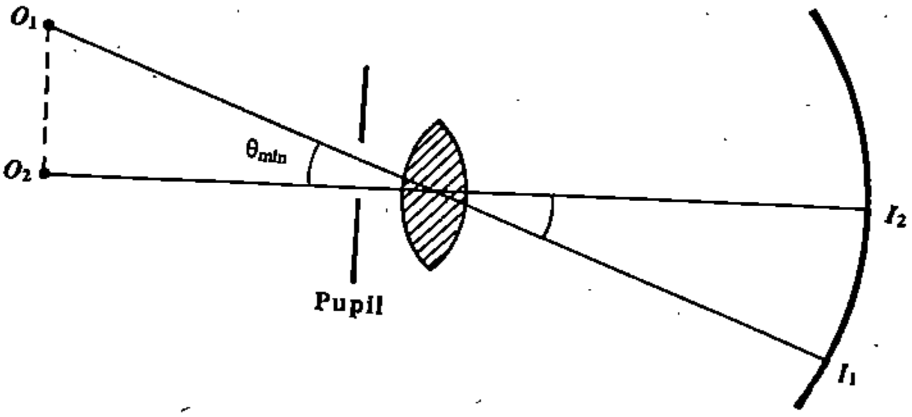


Fig. 11.1: Visible acuity and image formation on retina

$$\theta = \frac{1.22\lambda}{D} = \frac{1.22 \times (5.5 \times 10^{-5}\,\text{cm})}{(2 \times 10^{-1}\,\text{cm})} = 3.35 \times 10^{-4}\,\text{rad} \simeq 1\ \text{minute of arc}$$

Thus if the object is at a distance of 2 m, the size of image formed in a normal unaided human eye should be $(2 \times 3.35 \times 10^{-4}\text{rad}) \times 2\text{m} = 1.34 \times 10^{-3}\text{m}$.

Now refer to Fig. 11.2. It shows the image of a point source, luminous star say, formed by an astronomical telescope whose objective acts as circular aperture and produces Airy pattern. The image essentially is a bright circular disc of angular diameter $2\theta\left(= \frac{2.44\,\lambda}{D}\right)$, which depends on $\lambda$ and $D$. Larger the aperture, truer is the image, i.e. smaller is the Airy disc. On the other hand, if the aperture size is small, the size of Airy disc increases. That is, no matter how free from aberrations an astronomical telescope objective be, what is observed at best is not a point image of a star. For similar reasons we find that the image of a point object formed by a microscope is of finite size. We may therefore conclude that **diffraction constrains an optical device in the formation of a sharp point-like image of a point source due to the finite sizes of its components.**

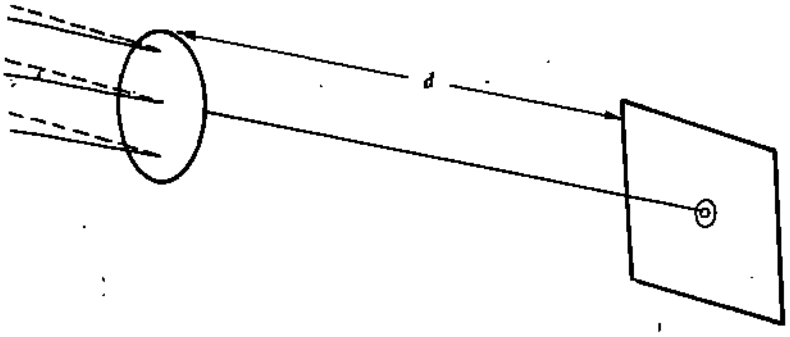


Fig. 11.2: Image of a luminous star formed by an astronomical telscope

An actual manifestation of this restriction arises in imaging when we observe two point sources or two spectrum lines. Since the objective of every optical instrument acts as a circular aperture and the point sources are mutually incoherent, the image consists of two independent Airy patterns. When the Airy discs are small and distinct, the two sources are said to be well resolved. The question now is how close can we bring these two discs so that they are just resolved. You will learn the answer to this question now.

## 11.3 RESOLVING POWER OF OPTICAL INSTRUMENTS

There are several criteria for the resolution limit. But we will confine ourselves to the conventional specification, **the Rayleigh criterion,** which however arbitrary, has the virtue of being particularly simple. According to this, the two patterns are resolved when the first minimum of diffraction pattern of one coincides with the central maximum of the diffraction pattern of the other. In Rayleigh's own words;

> This rule is convenient on account of its simplicity and it is sufficiently accurate in view of the necessary uncertainty as to what exactly is meant by resolution.

We will now consider the specific cases of an astronomical telescope, a microscope and a diffraction grating.

### 13.3.1. Astronomical Telescope

Imagine that a telescope points towards two close luminous stars, which subtend an angle $\alpha$ on the objective. The plane waves from these stars reach the objective and give rise to Airy diffraction patterns (Fig.11.3). Since the stars are effectively at an infinite distance from us, the diffraction patterns (images) are formed in the back focal plane of the telescope objective, where it is examined with the aid of the eye piece.The angle between mid points of central discs is equal to the angle subtended by the stars at the objective. For these stars to be just resolved, Rayleigh's criterion demands that maximum (centre) of the Airy disc due to one star should fall on the minimum (periphery) of the disc due to the other star, as shown in Fig. 11.4. (The

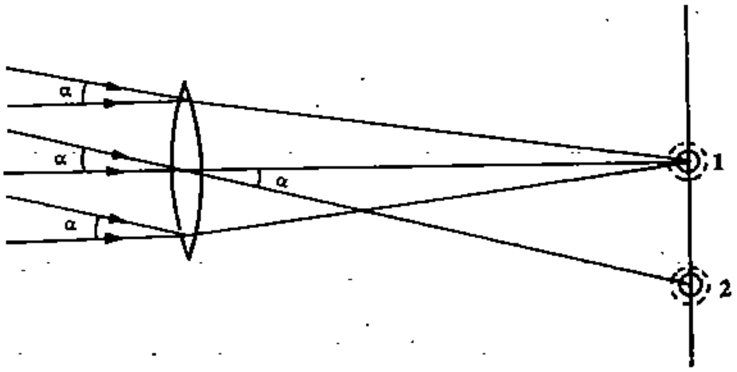


Fig. 11.3: Formation of Airy patterns in imaging of two stars by a telescope

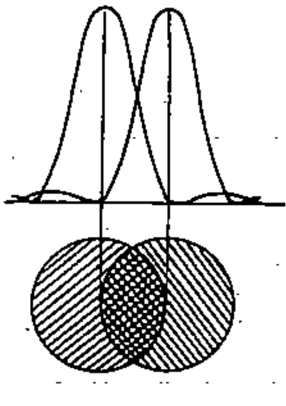


Fig. 11.4: Rayleigh criterion for imaging of two stars of small angular separation

corresponding intensity curves are also shown.) Mathematically, we demand that for the two stars to be just resolved, the angle subtended by the two stars at the objective should be equal to the angular half width of the Airy disc. Recall Eq. (9.13). It suggests that the **minimum resolvable angular separation or angular limit of resolution** for two close stars which can be resolved by a telescope is

$$\theta_{min} = \frac{1.22\,\lambda}{D} \tag{11.1}$$

Two stars subtending an angle $\alpha$ at the objective will be resolved for $\alpha > \theta_{min}$ and unresolved for $\alpha < \theta_{min}$. The intensity plot for more than resolved, just resolved (Rayleigh limit), and unresolved stars are shown in Fig. 11.5.

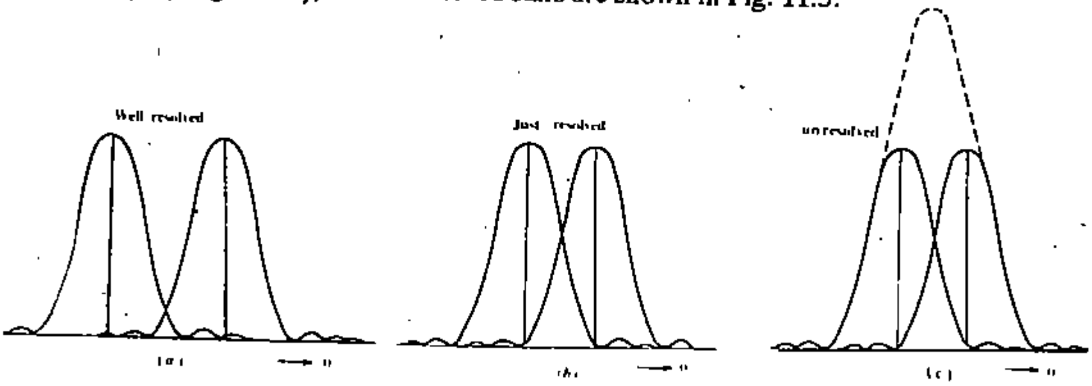


**Fig. 11.5: Plot of intensities of two resolved, Just resolved and unresolved stars**

The centre-to-centre linear separation of two just resolved stars, **the limit of resolution**, is given by

$$s = f\theta_{min} = \frac{1.22\,\lambda f}{D} \tag{11.2}$$

where $f$ is the focal length.

The **resolving power** for an optical device is generally defined as the reciprocal of resolving limit, i.e., as $\theta_{min}^{-1}$ or $s^{-1}$. This means that resolution ability of diffraction-limited systems depends on the size of the aperture and the wavelength. For a given wavelength, the resolving power of a telescope can be increased by using objectives of larger diameter. To give you some appreciation of numerical values, we now give a solved example. You should go through it carefully.

### Example 1

An astronomical observatory has a 40 inch telescope. Calculate the minimum angle of resolution for this telescope. Take $\lambda = 6000$ Å.

**Solution**

Before the S.I. system of units was adopted, the objective sizes were expressed in inches.

From Eq. (11.1) we recall that

$$\theta_{min} = 1.22\,\lambda/D$$

On substituting the given data, you will find that

$$\theta_{min} = \frac{1.22 \times (6 \times 10^{-5}\ \text{cm})}{40 \times 2.54\ \text{cm}}$$

$$= 7.2 \times 10^{-7}\ \text{rad}$$

$$= 0.15\ \text{seconds of arc}$$

The diameter of the largest telescope is about 80 inch ( ~2m ) and the corresponding angular separation of the objects it can resolve is 0.07 seconds of arc. This very low limit is not achieved in ground based telescopes due to turbulence in lower atmosphere.

For human eye, $\theta_{min} = 3.35 \times 10^{-4}$rad. Therefore, the actual lateral width of the image of a distant point formed on your retina is

$$s = f\theta_{min}$$

If we take $f = 3$ cm, we find that

$$s = (3\,\text{cm}) \times 3.35 \times 10^{-4}$$

$$= 10.05 \times 10^{-4}\,\text{cm}$$

$$= 10\ \text{micron}$$

This is roughly three times the mean spacing between photoreceptors (cones) at the centre of the retina. Therefore, for a normal unaided human eye, the linear separation between two point objects at a distance of 3m subtending this angle will be equal to $(3.35 \times 10^{-4} \times 3\text{m} = 1 \times 10^{-3}\text{m}) = 1\text{mm}$. This means that the unaided eye will resolve two point objects 1mm apart at a distance of about 3m.

You can easily verify this result atleast qualitatively. You should just draw two lines one millimeter apart and view these from a distance. (Alternatively, you can see marks on a millimetre scale or some news print). Move forward or backward till these become blurred and just merge into one another. Experience tells us that 1 mm is barely resolved at 2 m. The difference is due to optical defects in the eye or the structure of retina.

You may now like to answer an SAQ.

**SAQ 1**

*Spend 5 min*

An astronaut orbiting at an height of 400 km claims that he could see the individual houses of his city as they passed beneath him. Do you believe him. If not, why?

You now know that a 40 inch telescope has a minimum angle of resolution equal to $7.2 \times 10^{-7}$rad. The minimum angle of resolution of the eye is about $3.35 \times 10^{-4}$ rad. An important question that should come to our mind is: What should be the magnifying power of the telescope to take full advantage of the large diameter of the objective? The telescope must magnify about $\dfrac{3.35 \times 10^{-4}\,\text{rad}}{7.2 \times 10^{-7}\,\text{rad}} = 465$ times. Note that any further magnification will only make the image bigger but it would not be accompanied by increase in details which are not available in the primary image. (The resolution is determined by diffraction at the objective, i.e. the magnitude of $\theta_{min}$.) To get some idea about these details, you should carefully go through the following example.

**Example 2**

Compare the performances of two telescopes with objectives of apertures 100 cm and 200 cm. Take their focal lengths to be equal.

**Solution**

We know that for a telescope, the minimum angle of resolution

$$\theta_{min} = \frac{1.22\,\lambda}{D}$$

For the first telescope $\theta_{min} = \dfrac{1.22\,\lambda}{100\text{ cm}}$, where $\lambda$ is in cm. Therefore, the radius of central diffraction disc $r = f\,\theta_{min} = f\dfrac{1.22\,\lambda}{100\text{ cm}}$ and the area of Airy disc

$$A_1 = \pi r^2 = \pi\left(f\frac{1.22\,\lambda}{100\text{ cm}}\right)^2$$

The area of the telescope objective which collects light is $\pi\left(\dfrac{100\text{ cm}}{2}\right)^2$. This light is largely concentrated in the central maximum and gradually decreases as $\left(\dfrac{\sin N\gamma}{\sin\gamma}\right)^2$. If we assume that light is uniformly distributed over the disc, its brightness, i.e. light per unit area

$$I_1 \propto \pi\left(\frac{100\text{ cm}}{2}\right)^2 \div \pi\left(f\frac{1.22\,\lambda}{100\text{ cm}}\right)^2$$

$$= (50)^2 \times \frac{(100)^2}{f^2(1.22)^2\lambda^2}\text{cm}^4$$

$$= \frac{100^2 \times 100^2}{4f^2(1.22)^2\lambda^2}\text{cm}^4$$

For the second telescope $\theta_{min} = \dfrac{1.22\,\lambda}{200\text{ cm}}$. That is, the minimum angle of resolution for the second telescope is half of that for the first telescope. In other words, the R.P of 200 cm telescope is twice as large. To compare their relative performances, let us compare the brightness. As before, the area of central diffraction disc

$$A_2 = \pi\left(f\frac{1.22\,\lambda}{200\text{ cm}}\right)^2$$

and brightness

$$I_2 \propto \frac{(200^2)(200^2)}{4f^2(1.22)^2\lambda^2}\text{cm}^4$$

$$= 16I_1$$

In words, the area of the central diffraction disc of second telescope is four times more. And the of the image of the star will be proportional to fourth power of its area.

So we may conclude that

(i) The ability of a telescope to resolve two close stars depends on the diameter of its objective.

(ii) The intensity of the image is sixteen times since the objective collects four times more light and concentrates it over an area which is only one fourth. This means that a distant star, which is too faint to be observed by a smaller objective (of the first telescope), becomes visible by a larger telescope. That is, a bigger telescope can see farther in the sky. Therefore, **the deeper we want to penetrate the space, the greater should be the aperture of the objective of telescope.**

You may now like to pause and ponder for a while. Then you should answer SAQ 2.

### SAQ 2

*Spend 5 min*

We can see the stars at night but as sun-rises they gradually fade away and are not visible during the day. What measure would you suggest to enable researchers to make astronomical observations in the day time itself?

### Example 3

Calculate the dip in the resultant intensity of two $\left(\frac{\sin\beta}{\beta}\right)^2$ curves according to Rayleigh's criterion, i.e., when the maximum of one curve falls on the minimum of the other curve.

### Solution

We assume that the two curves have equal intensity. These curves are symmetrical and will cross at $\beta = \pi/2$, as shown in Fig. 11.6.

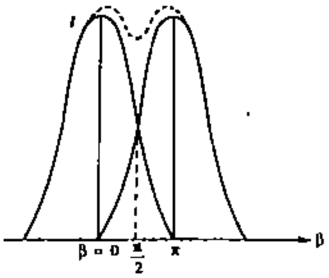


Fig. 11.6: Resolution of two single slit patterns: Rayleigh's criterion

At the point of intersection, both curves have equal intensity:

$$I = \left(\frac{\sin\frac{\pi}{2}}{\pi/2}\right)^2 = \frac{4}{\pi^2} = 0.4053$$

At this point the resultant intensity will be equal to the sum of the two intensities and therefore equal to 0.8106. This means that according to Rayleigh's criterion, the resultant intensity will show a dip of about 20%. And this dip is easily visible to even unaided human eye. If these two curves are brought closer, the dip will gradually decrease and it becomes difficult to resolve the images. Moreover, if these intensities were unequal, the dip will not be 20%.

In the above example we have taken the intensity of both the curves to be equal. This essentially means that in Rayleigh criterion we take both the stars to be equally luminous. Another important point to note is that the curves are of finite angular (or lateral) width. In the case of grating (or prism), two spectrum lines, though assumed to be of equal intensity, are very sharp. Now the question arises: Can we use the same criterion even for a grating? From your second level physics laboratory you may recall answer to this question; we do use the same criterion. Is the dip 20% or so even in this case? To discover, answer to this question, you should answer the following SAQ.

### SAQ 3

*Spend 5 min*

What is the dip in the resultant intensity of two $\left(\frac{\sin N\gamma}{\sin\gamma}\right)^2$ curves according to Rayleigh criterion?

A more realistic criterion for resolving power has been proposed by Sparrow. We know that at the Rayleigh limit there is a central dip or saddle point between adjacent peaks. As the distance between two point sources is less than the Rayleigh limit, the central dip will grow shallower and may ultimately disappear (Fig. 11.7) The angular separation corresponding to that configuration is said to be Sparrow's limit. Note that the resultant maximum has a broad flat top; there is no change in slope. However, we will not discuss it any further.

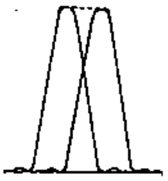


Fig. 11.7: Sparrow's resolution criterion

Another useful image forming device is a microscope. Let us now learn to calculate its resolving power.

### 13.3.2 Microscope

We know that an astronomical telescope is used to view far off objects whose exact distances are usually unknown. However, we were chiefly interested in their smallest permissible angular separation at the objective. In case of an optical microscope, the objects being examined are very close to the objective and subtend

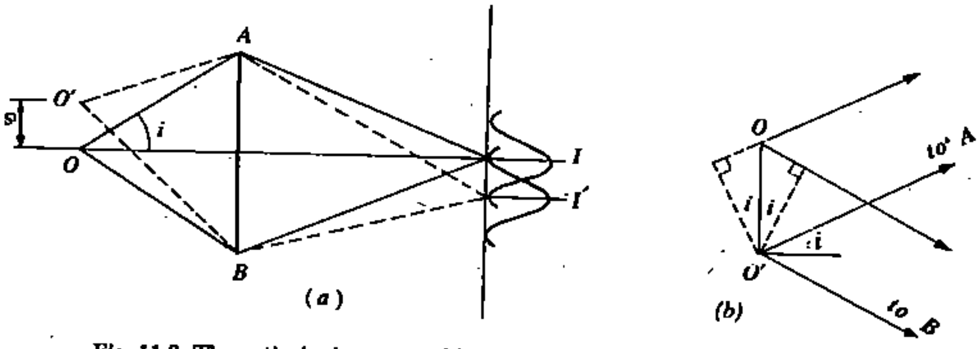


**Fig. 11.8: The optical microscope. (a) Airy pattern images of two objects $O$ and $O'$ separated through a distance s (b) Ray diagram for computation of path difference $O'B - O'A$**

a large angle. For this reason by **resolving power of a microscope we mean the smallest distance, rather than the minimum angular separation, between two point objects ( $O$ and $O'$ ) when their fringed images ( $I$ and $I'$ ) are just resolved.** Each image consists of a central Airy disc (surrounded by a system of rings which are very faint and not considered.) According to Rayleigh criterion, the first maximum of $I$ should be at the same position where the first minimum of $I'$ lies. The angular separation between the two discs on the limit of resolution $\theta_{min} = \frac{1.22\lambda}{D}$. When two images are just resolved, the wave from $O'$ diffracted to $I'$ has zero intensity (first dark ring) and the path difference $O'B - O'A = 1.22\,\lambda$ (Fig.11.8 (a)). We show an enlarged part in Fig. 11.8 (b) from which we see that $O'B$ is longer than $OB$ by $s \sin i$, and $O'A$ is shorter by the same amount. Here the point $O$ subtends an angle $2\,i$ at the objective of the microscope. Thus the path difference of the extreme rays from $O'$ to the objective is $2\,s \sin i$. Upon equating this to $1.22\,\lambda$ we find that the minimum separation between two points in an object that can be resolved by a microscope is given by

$$2\,s \sin i = 1.22$$

or

$$s = \frac{1.22\lambda}{2 \sin i} = = \frac{0.61\lambda}{\sin i}$$

In high power microscopes, the space between the object and objective is filled with oil of refractive index $\mu$. For an oil immersed objective, the above expression becomes

$$s = \frac{0.61\,\lambda}{\mu \sin i} \qquad (11.3)$$

You may now like to answer an SAQ.

**SAQ 4**

*Spend 5 min*

In the above discussion we assumed that the two point objects were self-luminous. Suppose two objects are illuminated by the same source. Will Eq. (11.3) still hold?

Abbe investigated this problem of image formation in detail and found that the resolving power depends on the mode of illumination of the object. In the above treatment both $O$ and $O'$ were treated as self-luminous objects and thus the light given out by these had no constant phase relationship. For all practical modes of illumination, the resolving power may be taken simply as

$$R.P = \frac{0.61\,\lambda}{\mu \sin i}$$

The term $\mu \sin i$ is termed as **numerical aperture** ($N.A$) of the microscope objective. The maximum value of $i$ is 90°. This gives the **microscopic** limit on $R.P$ approximately as $\frac{\lambda}{2\mu}$. This shows that smaller the $N.A$, greater will be the $R.P$.

In practice, good objectives have $N.A \approx 1$ so that the smallest distance that can be resolved by a microscope is of the order of the wavelength of light used. Obviously, with light of shorter wavelength, say ultraviolet rather than visible light, microscopy allows for perception of finer details. (We may have to take the photographs and then examine the images.)

In your school physics curriculum you have learnt that electrons exhibit diffraction effects. The deBroglie wavelength of an electron is given by

$$\lambda(\text{Å}) = \frac{12.3}{\sqrt{V}} \qquad (11.4)$$

For electrons accelerated to 100 kV, the wavelength is

$$\lambda(\text{Å}) = \frac{12.3}{\sqrt{10^5}} = 0.039 \times 10^{-10}\,\text{m} \qquad (11.5)$$

This wavelength is $10^5$ times smaller than that for visible light. The resolving power of an **electron microscope** will therefore be very high. This makes it possible to examine objects that would otherwise be completely obscured by diffraction effects in the visible spectrum. In this connection we may mention tremendous utility of electron microscope in the study of minute objects like viruses, microbes and finer details of crystal structures. It is better than even ultraviolet microscope for high resolution applications.

The deBroglie wavelength of an electron is given by

$$\lambda = \frac{h}{m_e v}$$

where $h$ is Planck's constant, $m_e$ is electronic mass and $v$ is electron speed. When an electron beam is accelerated through a potential difference $V$, we can write

$$v = \sqrt{\frac{2 V_e}{m_e}}$$

On combining these relations we find that

$$\lambda = \frac{h}{\sqrt{2 m_e e}} \cdot \frac{1}{\sqrt{V}}$$

Substituting the values

$h = 6.6 \times 10^{-34}$ Js,

$m_e = 9.11 \times 10^{-31}$ kg

and $e = 1.6 \times 10^{-19}$ C,

you will find that

$$\lambda(\text{Å}) = \frac{12.3}{\sqrt{V}}$$

### 11.3.3 Diffraction Grating

You are familiar with a sodium lamp. It gives out two close spectral lines, the so-called $D_1$ and $D_2$ lines with wavelengths $\lambda_1 = 5890$ Å and $\lambda_2 = 5896$ Å. For such lines, the resultant peak may become somewhat ambiguous. The problem we now wish to consider is: What is the smallest difference $\Delta\lambda$, that a diffraction grating can resolve? The resolving power of a grating is defined as

$$R.P = \frac{\lambda}{(\Delta\lambda)_{min}} \qquad (11.6)$$

where $(\Delta\lambda)_{min}$ is the least resolvable wavelength difference or limit of resolution and $\lambda$ is the mean wavelength. It is sometimes also called chromatic resolving power.

We know that the grating forms a principal maximum corresponding to wavelength $\lambda$ at the diffraction angle $\theta$. Similarly, the principal maxima at corresponding to $\lambda + \Delta\lambda$ will be at $\theta + \Delta\theta$. At first thought you may argue that the two colours will be separated and always appear to be resolved since the two angles are different. This could be so if the principal maxima, i.e. the spectrum lines in the experimental arrangement, were truly sharp like an ideal geometric line. But we know that the principal maximum has a finite angular width. Therefore, the question is: How close can these be brought so that they are seen distinct? Obviously, sharper the lines, the closer these can be brought and still be seen as two.

This question was also carefully examined by Rayleigh. In Fig. 11.9 (a) we show plots of two widely separated principal maxima. In Fig. 11.9 (b) we have brought these closer so that the principal maximum of $\lambda + \Delta\lambda$ is situated at the position where the minimum of $\lambda$ falls. The dotted line defines resultant intensity, which shows a dip. You will recall that according to Rayleigh criterion, this is the closest that we can bring these curves and still regard them as separate. If we bring them still closer as in Fig. 11.9 (c), the resultant intensity (shown by the dotted line) signifies a single enhanced principal maxima.

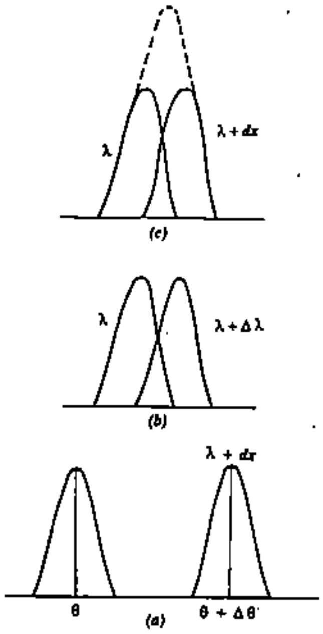


Fig. 11.9: Resolution of two spectral lines

According to Rayleigh criterion, the condition for resolution of two spectral lines by a diffraction grating is obtained by noting that for the common diffraction angle $\theta$, the following two equations should be satisfied simultaneously:

$$d\sin\theta = n(\lambda + \Delta\lambda)$$

for principal maxima of $\lambda + \Delta\lambda$ and

$$d\sin\theta = n\lambda + \frac{\lambda}{N}$$

for first minimum adjacent to the principal maximum for wavelength $\lambda$. On simplifying these we get

$$\frac{\lambda}{\Delta\lambda} = nN \qquad (11.7)$$

We note that in a given order $n$, the $R.P$ is proportional to the total number of slits. Does this mean that $R.P$ increases indefinitely with $N$? It is not so. Think why? Does it have some connection with the width of the grating? You will also note that the resolving power is independent of grating constant. It means that resolving powers of two gratings having equal number of lines but different grating constants will be equal.

To enable you to grasp these concepts and appreciate the numerical values, we now give some more solved examples.

**Example 4**

For $D_1$ and $D_2$ sodium lines, $\lambda_{D_1} = 5890$ Å and $\lambda_{D_2} = 5896$ Å. Calculate the minimum number of lines in a grating which will resolve the doublet in the first order.

**Solution**

Let us take the average wavelength as 5893 Å. From Eq. (11.6) we find that the resolving power is

$$\frac{\lambda}{\Delta\lambda} = \frac{5893 \times 10^{-8}\,\text{Å}}{6 \times 10^{-8}\,\text{Å}} = 982.2$$

Therefore, we must have a grating with more than 983 lines to resolve sodium doublet in first order. A grating of 1000 lines will serve the purpose.

**Example 5**

Suppose that to observe sodium doublet we use a grating having $d = 10^{-3}$cm and a lens of focal length 2 m. Let us calculate the linear separation of the two lines in the 1st and 2nd order.

**Solution**

We know that

$$d\sin\theta = n\lambda$$

For the $D_1$ line

$$\sin\theta_1 = \frac{5890 \times 10^{-8}\,\text{cm}}{10^{-3}\,\text{cm}} = 5890 \times 10^{-5}$$

or

$$\theta_1 \cong 5890 \times 10^{-5}\,\text{rad}$$

Similarly for the $D_2$ line

$$\sin\theta_2 = \frac{5896 \times 10^{-8}\,\text{cm}}{10^{-3}\,\text{cm}} = 5896 \times 10^{-5}$$

or

$$\theta_2 \simeq 5896 \times 10^{-5}\,\text{rad}$$

$$\therefore \qquad \Delta\theta = (\theta_2 - \theta_1) = 6 \times 10^{-5}\,\text{rad}$$

With a lens of focal length 200 cm, we find that linear separation between $D_1$ and $D_2$ lines is

$$l = f\Delta\theta$$

$$= (200\,\text{cm}) \times (6 \times 10^{-5}\text{rad})$$

$$= 12 \times 10^{-3}\,\text{cm} = 0.12\,\text{mm}$$

This shows that 6 Å are separated by 0.12 mm in 1st order. Alternatively we may say that linear separation is nearly 50 Å per millimeter in the first order. You can readily check that in the second order this linear separation will be 25 Å per millimeter.

## 11.4 IMPROVING RESOLUTION

You now know that with the help of a telescope, we can view a faint star, resolve two close stars and measure the angle subtended by the double star at the objective of the telescope. However, it is worth noting that based on Fraunhofer diffraction image of a star, we cannot measure its angular diameter. To overcome this limitation, Fizeau suggested a slight modification in that we should use a two slit adjustable aperture (with provision for lateral adjustment), in front of the objective of the telescope. As a result, the plane wavefront falling on the double slit is diffracted and collected by the objective. The Fraunhofer diffraction pattern of the double slit is formed in the back focal plane of the objective. The measurements to determine angular diameter are made from the observations on these interference fringes.

Refer to Fig. 11.10. Two slit apertures $S_1$ and $S_2$ are at a distance $d$ apart. The telescope is first pointed towards the double stars, which act as two point sources $O$ and $O'$. The two point sources are separated by an angle $\theta$ in a direction at right angles to the lengths of the slits. Such objects emit white light and because of intensity considerations, the observations have to be made with white light fringes. It is therefore customary to assume an effective value of the wavelength emitted by the source. This depends upon the distribution of intensity of the light and the colour response of the eye. The interference patterns due to $O$ and $O'$ have the same fringe spacing since this spacing depends upon separation between slit apertures and the focal length of the objective. Moreover, these fringe patterns are shifted

The intensity of the double slit pattern is given by

$$I = 4R^2 \frac{\sin^2\beta}{\beta^2} \cos^2\gamma$$

where $\beta = \dfrac{\pi a \sin\theta}{\lambda}$ and $\gamma = \dfrac{\pi d \sin\theta}{\gamma}$ in which $a$ is the slit width and $d$ is the slit separation. The positions of the maxima are given by

$$d \sin\theta = n\lambda$$

where $n = 0, 1, 2, 3, \ldots$ When $\theta$ is small, the successive maxima occur at

$$\theta = 0, \frac{\lambda}{d}, \frac{2\lambda}{d}, \frac{3\lambda}{d} \ldots$$

so that the angular separation between successive maxima is given be $\theta_1 = \frac{\lambda}{d}$. Further, if $a$ is small, the interference pattern will be essentially a $\cos^2\gamma$ curve near the centre.

with respect to each other by an angle θ. Therefore, as shown in the figure the central maximum of the pattern due to $O$ is at $P$ and that due to $O'$ is at $P'$. If $O$ and $O'$ are two incoherent sources, the combined pattern is formed by summing the intensities of these two patterns at each point. Assuming that both $O$ and $O'$ have equal brightness, we can plot two $\cos^2\gamma$ curves on the same scale and shift them suitably to obtain the resultant curve.

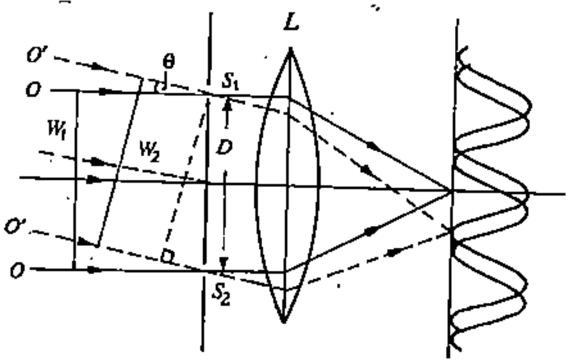


Fig. 11.10: Principle of measurement of angular diameter of stellar objects by Interferometry

We can show graphically that if this shift is a small fraction of the angular separation θ, the resultant intensity distribution resembles a $\cos^2\gamma$ curve. However, the intensity does not fall to zero at the minimum. The net result is a fringe pattern shown in Fig. 11.10(b). By successive adjustments a stage can come when the maximum of one pattern, say due to $O$, coincides with the minimum of $O'$. Then we have $\frac{\theta_1}{2} = \frac{\lambda}{2d}$ And the paths from the two sources differ by $\frac{\lambda}{2}$ We can show graphically that the resultant curve 2 shows a uniform intensity and the fringes have disappeared. If we displace the two curves further, the fringes reappear and become sharp when the fringes are displaced by a whole fringe width, i.e. $\theta = \theta_1$. They disappear again when $\theta = \frac{3\theta_1}{2}$ or $\frac{5\theta_1}{2}$. Therefore, with two point sources subtending an angle θ at the double slit, the condition for the disappearance of fringes is

$$\theta = \frac{\lambda}{2d}, \frac{3\lambda}{2d}, \frac{5\lambda}{2d}$$

To measure angular separation of a double star, the double slit is mounted in front of the objective of the telescope which points towards the double star. (We should remember that the line joining the stars should be perpendicular to the length of the slits.) We expect interference pattern due to the double slit. If on adjusting the separation between the slits, the interference fringes can be made to disappear, we can infer that the star is a double star. The first disappearance should take place when the angular separation is $\frac{\lambda}{2d}$. Let us compare this with the expression for the resolving power of a telescope ( $\theta = 1.22\frac{\lambda}{a}$, where $a$ is the diameter of the objective). If the double slits are $a$ apart and the first disappearance occurs for $d = a$, the angle θ between the double stars is $\theta = \frac{\lambda}{2d} = \frac{\lambda}{2a}$. This angle is effectively half of the $R.P$ of the telescope. It explains the genesis of the statement: **The $R.P$ of a telescope may be doubled by placing a double slit in front of it.** You must however note that with a double slit, we can only infer the presence of a double star (from the disappearance of the fringes); we neither get the images of the stars nor resolve them. Indeed, even before the disappearance of the fringes, a blurring of fringes starts. This angle is only a small fraction of $\theta_1$. You may have realised that this method enables us to measure the **angular diameter** of the disc of the star and Michelson successfully used it in 1920.

## Angular Diameter of a Star

For measuring the angular diameter of the disc of a star we should first know the condition for the disappearance of fringes for a double slit placed in front of a telescope. In contrast to two point sources, the disc of a star consists of a series of points extending from one end $O_1$ to another end $O_2$. In Fig. 11.10, we see that when $O_1$ and the central point $O$ satisfy the condition for disappearance of fringes, the point just next to $O_1$ will have a similar point next to $O$ and so on. Thus all the points between $O_1$ and $O$ will have corresponding points lying between $O$ and $O_2$ satisfying the condition for disappearance of fringes. Since the angle between $O_1$ and $O$ for the first disappearance of fringes is $\frac{\lambda}{2d}$, the angle between $O_1$ and $O_2$ (which is for the total disc) equals $\frac{\lambda}{d}$. Thus the angular disc $\theta$ of the star, computed from the first disappearance of fringes, is given by $\theta = \frac{\lambda}{d}$. For successive disappearances $\theta$ is given by $\theta = \frac{2\lambda}{d}, \frac{3\lambda}{d}, \cdots$ If the source is a circular disc, the condition for the first disappearance is $\theta = 1.22\,\frac{\lambda}{d}$. This method was successfully used to measure angular diameters of planetary satellites. But attempts to apply it for single stars failed because of their small angular diameters. Even with the largest slit separation possible with the available telescopes, the fringes remained distinct; no disappearance was achieved. To overcome this difficulty, Michelson devised stellar interferometer in 1890. We will discuss it now.

### 11.4.1 Michelson Stellar Interferometer

The principle of Michelson's Stellar Interferometer is illustrated in the Fig.11.11. The slit apertures $S_1$ and $S_2$ in front of the telescope are fixed. Light reaches them after reflection from a symmetrical system of mirrors $M_1$, $M_2$, $M_3$ and $M_4$ mounted on a rigid girder in front of the telescope. The inner mirrors $M_3$ and $M_4$ are fixed but

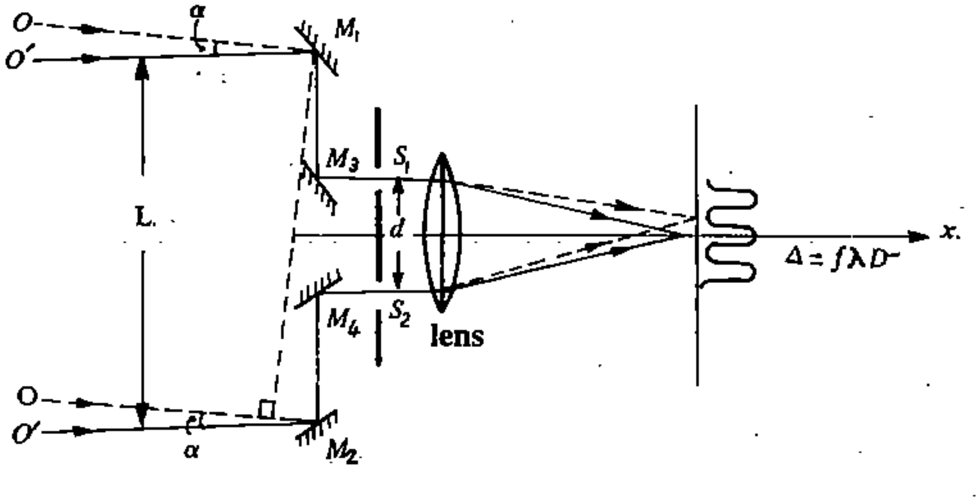


Fig. 11.11: Schematics of Michelson Stellar Interferometer

the outer mirrors $M_1$ and $M_2$ can be separated out symmetrically in a direction perpendicular to the lengths of the slit apertures. Therefore light from one edge of the star (shown as solid line) reaches the point $P$ in the focal plane via the paths $OM_1\,M_3\,S_1\,P$ and $OM_2\,M_4\,S_2\,P$. This will form interference fringes with the angular separation equal to $\frac{\lambda}{d}$. The other edge of the star sends light along the dotted lines and produces a similar system displaced slightly with the central fringe at $P'$. You now know that when two extreme fringe systems are displaced by a whole fringe width, the resultant intensity pattern will show uniform intensity and the fringes will disappear. The angular diameter of the star $\alpha = 1.22\,\frac{\lambda}{D}$, where $D$ is the separation of outer mirrors $M_1$ and $M_2$. You can easily convince yourself by noting that the optical paths $M_1\,M_3\,S_1$ and $M_2\,M_4\,S_2$ have been maintained equal so that the optical path difference for light from the two edges of the star is the same at $S_1$ and $S_2$ as at $M_1$ and $M_2$. If the path difference at $M_1$ and $M_2$ is one whole wavelength, the path difference at $S_1$ and $S_2$ is also one whole $\lambda$ and fringe shift is equal to one fringe width. This leads to disappearance of fringes. As shown in the diagram, the dotted lines inclined at an angle $\alpha$ will have a path difference of $\lambda$ when $\alpha = \frac{\lambda}{D}$. In this arrangement the smallest angular diameter that can be measured is determined by the separation of the outer mirrors $M_1$ and $M_2$ rather than the diameter of the objective of the telescope. Therefore, the stellar interferometer magnifies the effective resolving power of the telescope in the ratio $\frac{D}{d}$. We may emphasize that for a circular star disc, the fringes will disappear when $\alpha = 1.22\,\frac{\lambda}{D}$. This implies that the outer mirrors have to be moved out somewhat.

The interferometer was mounted on the large reflecting telescope (diameter 100 inch) of the Mount Wilson observatory, which was used because of its mechanical strength. The first star whose diameter was measured by this method was Betelegeuse ($\alpha$ -orions) whose fringes disappeared when the separation between $M_1$ and $M_2$ was equal to 121 inches. Assuming $\lambda = 5700$ Å, we find that

$$\alpha = \frac{1.22\lambda}{D} = \frac{1.22 \times 5700 \times 10^{-8}\,\text{cm}}{121 \times 2.54\,\text{cm}}$$

$$= 22.7 \times 10^{-8}\,\text{rad}$$

$$= 0.047\ \text{seconds of arc}$$

The distance of Betelegeuse was measured by parallex method. Its linear diameter was then found to be $4.1 \times 10^8$ km, which is about 300 times the diameter of the sun. The maximum separation of the outer mirrors was 6.1m so that the smallest measurable angular diameter with $\lambda = 5500$ Å was about 0.02 seconds of arc. This is insufficient for most of the stars. The smallest star for which measurements were made was Arcturus. Its actual diameter is 27 times that of the sun.

At the surface of the earth, the sun disc has an angular diameter of about $32' \simeq 0.018$ rad. If we imagine the sun to be at a distance of the nearest star, its disc would subtend an angle only 0.007 seconds of arc. This will require a mirror separation of 20 m for disappearance of fringes. It is difficult to achieve this since we require a rigid mechanical connection between mirrors and eye piece.

Let us now summarise what you have learnt in this unit.

## 11.5 SUMMARY

- Diffraction constrains an optical device in the formation of a sharp point-like image of a point source.
- Rayleigh criterion for resolution of two images demands that the first minimum of diffraction pattern of one object and the central maximum of the diffraction pattern of the other should fall at the same position.
- The minimum resolvable angular separation or angular limit of resolution of two close objects by a telescope is given by

$$\theta_{min} = \frac{1.22\lambda}{D}$$

where λ is the wavelength and $D$ is diameter of the objective of the telescope.

- The resolving power of a telescope is inverse of angular limit of resolution. The deeper we want to penetrate the space, the greater should be the aperture of the objective of telescope.
- The resolving power of a microscope is defined as the smallest **distance** between two point objects when their fringed images are just resolved:

$$R.P = \frac{0.61\lambda}{\mu \sin i} = \frac{0.61\lambda}{N.A}$$

where i is the angle of incidence. sin $i$ is known as numerical aperture and is approximately equal to one for good objective.

- The resolving power of a diffraction grating is defined as

$$R.P = \frac{\lambda}{(\Delta\lambda)_{min}} = nN$$

where Δ λ is the least resolvable wavelength difference, $n$ is order of spectrum and $N$ is the total number of slits.

## 11.6 TERMINAL QUESTIONS

1. A diffraction limited laser beam ( $\lambda = 6300$ Å ) of diameter 5 mm is directed at the earth from a space laboratory orbiting at an altitude of 500km. How large an area would the central beam illuminate?

2. The resolving power of a prism is given by

$$\frac{\lambda}{d\lambda} = t\frac{d\mu}{d\lambda}$$

where $t$ is the length of the base of the prism, μ is the refractive index of the material of prism for wavelength λ. A prism is made of dense flint glass for which refractive indices for λ = 6560 Å and 4860 Å are 1.743 and 1.773 respectively. Calculate the length of the base of the prism.

## 11.7 SOLUTIONS AND ANSWERS

### SAQs

1. The minimum angle of resolution of eye

$$\theta = \frac{1.22\,\lambda}{D} = \frac{1.22 \times (5.5 \times 10^{-5}\text{cm})}{0.2\text{ cm}} = 3.36 \times 10^{-4}\text{rad}$$

The lateral width for resolution

$$l = r\theta = (4 \times 10^{5}\text{m}) \times (3.36 \times 10^{-6}\text{rad}) = 1.34\text{ m}$$

Since it is must less thaln the width of individual houses, it is not wise to believe the astro nant.

2. As we increase the aperture of the telescope, the light collected by it from a star gradually increases and gets concentrated in the image (the diffraction disc). Ultimately a stage will come when the image of the star becomes brighter than the background and is visible (This is because the intensity of the image of a star is proportional to fourth power while the background sky light increases as the square of the area of the aperture.) This means that you can see stars during the day by using a telescope of sufficient aperture!

3. The maximum is at $Nn\pi$ and minimum at $(Nn + 1)\pi$. The two curves are symmetrical and if they are of equal intensity, they will cross at $N\gamma = Nn\pi + \frac{\pi}{2}$. Therefore, if you evaluate the function $\left(\frac{\sin N\gamma}{\sin\gamma}\right)^2$ at $N\gamma = Nn\pi$ and $N\gamma = Nn\pi + \frac{\pi}{2}$, i. e. $\gamma = n\pi$ and $\gamma = n\pi + \frac{\pi}{2N}$, you will find that

$$\left(\frac{\sin Nn\pi}{\sin n\pi}\right)^2 = N^2$$

and

$$\left[\frac{\sin\left(Nn\pi + \frac{\pi}{2}\right)}{\sin\left(n\pi + \frac{\pi}{2N}\right)}\right]^2 = \frac{1}{\sin^2\left(\frac{\pi}{2N}\right)} = \frac{1}{\left(\frac{\pi}{2N}\right)^2}$$

$$= \frac{4N^2}{\pi^2}$$

Hence the required ratio is $\frac{4}{\pi^2} = 0.4053$

Therefore the resultant intensity will show a dip of about 20% as in the case of a telescope.

4 The waves given out by each self-luminous object bear no constant phase relationship so that the intensities can be added up. The objects viewed with microscopes are illuminated by the same source and there will be some phase relationship between the waves emanating from these. Strictly speaking the intensities will not be additive. But Abbe found that Eq. (11.3) gives the correct order for the limit of resolution.

### TQs

1. We know that angular spread of light beam is given by

$$\theta = \frac{1.22\lambda}{D} = \frac{1.22x\,(6300 \times 10^{-8}\,\text{cm})}{(0.5\,\text{cm})}$$

$$= 1.54 \times 10^{-4}\,\text{rad}$$

Since the diameter of light patch

$$x = 2r\,\theta$$

the area of the earth illuminated by the beam focussed from the space laboratory at an altitude of 500 km is

$$A = \frac{\pi x^2}{4} = \pi r^2 \theta^2$$

$$= \frac{22}{7} \times (25 \times 10^{10}\,\text{m}^2) \times (1.54 \times 10^{-4})^2$$

$$= 10934\,\text{m}^2 = 0.01\,\text{km}^2$$

2. $$d\mu = 1.773 - 1.743 = 0.03$$

$$d\lambda = 6560 - 4860 = 1700\,\text{Å} = 1700 \times 10^{-8}\,\text{cm}$$

Note that spectral spread is very wide whereas $d\lambda$ should be a small change. Assuming that $\mu$ changes linearly between these two colours, we have

$$\frac{d\mu}{d\lambda} = -\frac{0.03}{1700 \times 10^{-8}\,\text{cm}} = -\frac{3}{17} \times 10^4\,\text{cm}^{-1} \approx -1765\,\text{cm}^{-1}$$

The negative sign signifies inverse value of relationship between $\lambda$ and $\mu$. The prism is made of dense flint glass and to just resolves $D_1$ and $D_2$ lines find that

$$R.P = \frac{5893\,\text{Å}}{6\,\text{Å}} = 982$$

so that

$$\frac{\lambda}{d\lambda} = 982 = 1765\,t$$

and

$$t = \frac{982}{1765\,\text{cm}^{-1}} = 0.556\,\text{cm} \approx 0.6\,\text{cm}$$

Uttar Pradesh
Rajarshi Tandon Open University

UGPHS-07
Optics

Block

# 4

# LASERS AND THEIR APPLICATIONS

# BLOCK INTRODUCTION

In the preceding blocks, you have learnt that light is an electromagnetic wave. It exhibits polarisation, interference, diffraction, etc. As you have seen, these phenomena are well understood in terms of the wave theory of light. This block, as the title indicates, essentially deals with lasers - a source of coherent light - and their applications, particularly in the areas of photography and optical communication. Lasers owe its invention to the quantum theory according to which, light energy consists of minute packets or quanta. The invention of lasers and related developments has once again brought the field of optics at the forefront of basic research and technological applications. Widespread use of light from laser is because of its high degree of coherence, high directionality, unprecedented brightness, etc. Without fear of exaggeration, we may say that the present period in optics may be called the Laser Age. In this block, we intend to give you a flavour of the basic physical principles involved in the design and operation of lasers and also about some of its important applications.

Laser is a coherent source of light. But what is coherence? You learnt about coherent source of light in Block 2 in connection with Young's double-slit interference experiment. It was emphasized there that for obtaining observable interference fringe pattern, the light from the slits must be coherent. In Unit 12, the first unit of this block, you will learn about the concept of coherence of waves. If the phases of two waves have a definite phase relationship, they are said to be coherent. This phase relationship between waves, which can be in time or space, gives rise to temporal coherence and the spatial coherencee. You will learn that temporal coherence of electromagnetic waves menifests as monochromaticity and the visibility of the interference fringe pattern indicates the extent of spatial coherence between the intefering waves.

In Unit 13, you will learn the working principle of lasers. In particular, we have discussed concept of stimulated emission of radiation and the prerequisits for obtaining laser light. Though the first laser used ruby (solid) as active medium, now lasers which employ liquids and gases as the active medium are available. You will learn about different type of lasers. Coherence (Monochromaticity), high directionality and brightness are some of the properties of lasers which are responsible for their so many and so varied applications. In this unit, you will also learn about some of these applications.

Holography is a technique of three-dimensional photography. This technique was invented by Dennis Gabor much before the invention of lasers. However, the full potential of this technique could be realised only after this invention. In Unit 14, you will learn details of this novel technique. Some of the applications of holography have also been discussed.

The use of lasers has revolutionised communication technology. The monochromaticity of laser light makes it an efficient carrier of information. Communication - transmission of speech, data, etc. - at optical frequency is much faster and reliable compared to radio and microwave communication. However, optical communication suffers from the drawback that signals get attenuated by dust particles, rains, etc. Thus, for efficient terrestrial optical communication, optical fibres are used. How is light transmitted through optical fibres? What are the characteristics of such fibre? These and other related questions form the subject matter of Unit 15.

The units are not of equal length. On an average, Unit 12 should take 4h, Unit 13 should take 7h, Unit 14 should take 5h and Unit 15 should take 6h. We hope you will enjoy studying this block.

We wish you success.

**Acknowledgement**

We thank Shri Sunder Singh and Gopal Krishna Arora for secretarial assistance.

# UNIT 12 COHERENCE

## Structure

## 12.1 INTRODUCTION

In Unit 5 of this course, you studied about Young's double-slit interference experiment. We emphasized that for observing interference fringe pattern, the light from two sources must be coherent. By coherence, we mean that the light waves from two slits have a constant phase relationship. Can you recall how this condition of coherence is achieved? If you are unable to do so, you should refer back to relevant pages. Now the question arises why coherence is a prerequisite for observing interference? You will learn about coherence in detail now.

In Sec. 12.2, we elaborate the concept of coherence as applied to waves in general. Further, the most elementary definition of coherence says that the phases of the coherent waves have a predictable relationship at different points and at different times in space. This space and time predictability of the phase relationship of waves gives rise to two types of coherence, namely, spatial coherence and temporal coherence. The concept of temporal coherence, which refers to the phase relationship at different times at a point, has been discussed in Sec. 12.3. You will also learn about the correlation between the width of a spectral line and temporal coherence. In Sec. 12.4, we have discussed spatial coherence which relates to the coherence of two waves travelling side by side. The relationship between the visibility of fringe pattern with spatial coherence is also discussed in detail.

### Objectives

After going through this unit, you should be able to

- explain the concept of coherence
- distinguish temporal coherence from spatial coherence
- relate temporal coherence with the width of spectral lines
- relate spatial coherence with the visibility of fringe pattern, and
- solve numerical problems based on coherence.

## 12.2 WHAT IS COHERENCE?

If you are asked what is coherence, you may say that it is the condition necessary to produce observable interference of light. And if you are asked what is interference, you may say it is connected with interaction of waves that are coherent. Well, nothing definite follows from such circular arguments! In fact, coherence is a property of light whereas interference is the effect of interaction of light waves. The crucial consideration in interference phenomenon is the relative phase of waves arriving at a given point from two or more sources. That is, in order to observe interference fringes, there must exist a definite phase relationship between the light waves from two sources. Hence, we may say that the necessity of having coherent sources for observing interference fringes essentially implies that the waves from the two sources must have a constant and predictable phase relationship. It is the absence of a definite phase relationship between light waves from ordinary sources that we do not obtain any observable interference fringe pattern.

Now, you may ask: Why there is no definite phase relationship between light waves from two ordinary light sources? Well, the basic mechanism of emission of light involves atoms radiating electromagnetic waves in the form of photons. Each atom radiates for a small time (of the order of $10^{-9}$s) . Meanwhile, other atoms begin to radiate. The phases of these emitted electromagnetic waves are, therefore, random; if there are two such sources, there can be no definite phase relationship between the light waves emitted from them.

In general, sources, and the waves they emit, are said to be coherent if they

(i) have equal frequencies,

(ii) maintain a phase difference that is constant in time.

If either of these properties is lacking, the sources are incoherent and the waves do not produce any observable interference.

Let us pause for a while and ask ourselves: Why it is a prerequisit for observing interference fringe pattern? To answer this question, let us consider the origin of the bright and dark fringes in the Young's experiment (Fig. 12.1). Let $E_1$ and $E_2$ be the electric fields associated with the light waves emanating from slits $S_1$ and $S_2$. These waves superpose and the combined electric field at any point on the screen is given by,

$$E = E_1 + E_2 \tag{12.1}$$

You may recall that in the interference pattern, we observe the intensity of light, not the electric field. Since the average intensity of light is prorportional to the time - averaged value of the associated electric field, we have

$$I \alpha < E^2 > \tag{12.2}$$

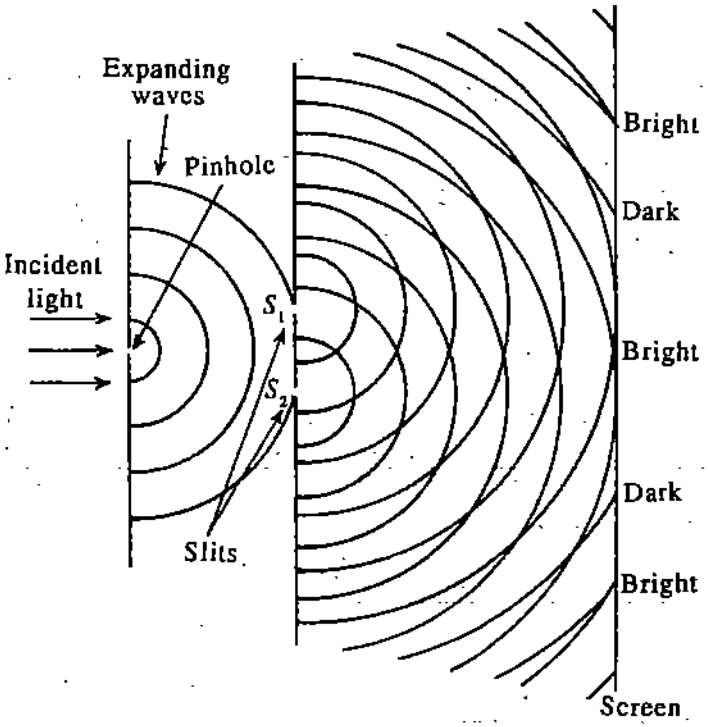


Fig.12.1: Young's interference experiment.

Thus, we have, from Eq. (12.1) and (12.2),

$$I = < E_1^2 > + < E_2^2 > + 2 < E_1 E_2 >$$

$$= I_1 + I_2 + 2 < E_1 E_2 > \quad (12.3)$$

Eq. (12.3) shows that the resultant intensity on the screen is the sum of intensities $I_1$ and $I_2$ (due to individual slit sources) and an interference term $2 < E_1 E_2 >$. The interference term is crucial because it determines whether the resultant intensity is an uniform illumination or a fringe pattern on the screen. The contribution of the interference term to the resultant intensity depends primarily on the phase relationship between the light waves emanating from the two slits.

Let us first consider the case when the light waves are in phase at one instance and are out of phase at another instance. In such a situation, the product $E_1 E_2$ will be positive at one instance and negative at the other. As a result, the time average of $E_1 E_2$ will be zero, i.e.

$$< E_1 E_2 > = 0$$

Waves having this kind of phase relationship (varying with time) are said to be **incoherent** and the resultant intensity will be

$$I = I_1 + I_2 \quad (12.4)$$

Thus, when light waves from two incoherent sources interfere, the resultant intensity will be the sum of individual intensities and the screen will be uniformly illuminated. To give you a simple example, when the headlights of a car illuminate the same area, their combined intensity is simply the sum of two separate intensities. The headlights are incoherent sources and there is no contribution of the interference term.

Now, what will happen if the light waves from two slits have a definite phase relationship i.e. a phase relationship which is constant in time. Source of light emiting such waves are coherent. When light sources are coherent, the resultant intensity is not simply the sum of individual intensities. It is so because in that situation, the interference term in equation (12.3) is non-zero. Let us see what is the form of the interference term when two coherent light waves superpose. There are two cases;

(a) When $E_1 = E_2$, that is, the two waves have same amplitude, frequency and phase. Thus,

$$I_1 = < E_1^2 > = < E_2^2 > = I_2$$

and

$$2 < E_1 E_2 > = 2 < E_1^2 > = 2I_1$$

The resultant intensity,

$$I = I_1 + I_2 + 2 < E_1 E_2 >$$

$$= I_1 + I_1 + 2I_1$$

$$= 4I_1 \quad (12.5)$$

Thus, the points on the screen where two interfering waves are in phase, the resultant intensity is four times that due to an individual source. These points will, therefore, appear bright on the screen.

(b) $E_1 = -E_2$, that is, the two waves have same amplitude and frequency but their phases differ by 180° which remains constant in time. In that case, the two waves are completely out of phase and the resultant wave amplitude and intensity will be zero.

$$E = E_1 + E_2 = 0$$
$$\Rightarrow I = 0.$$

The points on the screen where the interfering light waves satisfy above condition will have zero intensity and hence they will appear dark.

Thus, the constant phase relationship between superposing light waves i.e. coherence, is a necessary condition for obtaining interference fringe pattern. When the phase relationship is not constant, the points where superposing light waves arrive in phase at one instant may receive light waves which are completely out of phase at another instance. This results in uniform illumination of the screen and no interference fringe pattern can be observed.

In the above discussion, you have studied about the necessity of having coherent sources for observing interference fringe pattern. As mentioned earlier, coherence, which is essentially a correlation phenomenon between two waves, can be with respect to time and/or space. Thus, for expediency, we distinguish two types of coherence: **Temporal Coherence and Spatial Coherence**. Temporal coherence, or the longitudinal spatial coherence (often called monochromaticity) applies to waves travelling along the same path. It refers to the constancy and predictability of phase relationship as a function of time. Spatial coherence, or transverse spatial coherence refers to the phase relationship between waves travelling side by side, at a certain distance from one another. The further apart are the two waves, less likely they are to be in phase, and less coherent the light will be. You will study these two types of coherences in the following sections.

## 12.3 TEMPORAL COHERENCE

While studying interference and diffraction of light in the previous two blocks of this course, we assumed that electromagnetic waves remained perfectly sinusoidal for all time. This kind of electromagnetic waves are, however, practically impossible to obtain from ordinary light sources. Why is it so? It is because light emitted from an ordinary source consists of finite size wave trains. Each wave train is sinusoidal in itself and has a characteristic frequency (or wavelength) and phase. However, the collection of wave trains is not sinusoidal. Thus, light waves coming from an ordinary source can not have one single frequency (monochromatic). Instead, it has a range of frequencies; that is, it has a frequency bandwidth. For these reasons, the so called monochromatic light, such as from gas discharge tube, is more appropriately called quasi- monochromatic.

This aspect of light (i.e. monochromaticity) refers to its temporal coherence. The temporal coherence can be identified qualitatively as the interval of time during which the phase of the wave motion changes in a predictable manner as it passes through a fixed point in space. And in wave motion corresponding to light from ordinary sources, a predictable phase relationship can be observed only within the average length of the wave trains on time scale.

To elaborate the concept of temporal coherence, let us consider a typical time variation of the amplitude of an electromagnetic wave as shown in Fig. 12.2.

You may notice from the figure that the electric field at time $t$ and $t + \Delta t$ will have a definite phase relationship if $\Delta t << \tau_c$ and will not have any phase relationship if $\Delta t >> \tau_c$ where $\tau_c$ represents the average duration of the wave trains. The time $\tau_c$ is known as **coherence time** of the radiation and the wave is said to be coherent for time $\tau_c$.

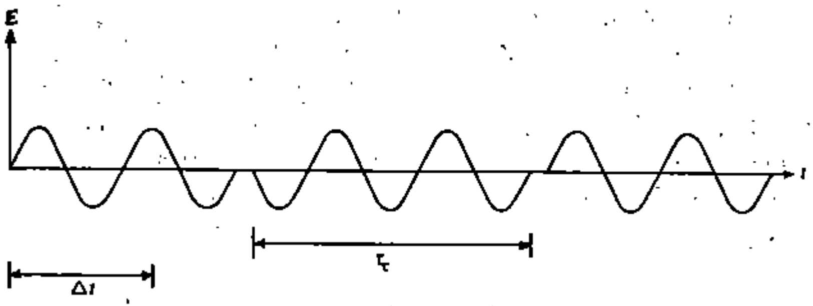


Fig.12.2: Typical variation of the amplitude of an electromagnetic wave with time. Three typical wave trains have been shown. The coherence time $\tau_c$ is the average duration of the wave trains.

And the path length corresponding to $\tau_c$, given as $L_c = c\,\tau_c$ is called the coherence **length** of the radiation, where $c$ is the velocity of light.

In order to study the time-coherence of the radiation, let us re-consider Michelson's interferometer experiment. For completeness, we have reproduced the experimental arrangement in Fig. 12.3. A nearly **monochromatic** light source is used in the investigation.

For the source ($S$) we may use a neon lamp in front of which we place a filter ($F$) so that radiation corresponding to $\lambda = 6328$ Å is allowed to fall on the beam-splitter $G$. Glass plate $G'$ is the compensating plate. You may recall from Unit 7, if the eye is in the

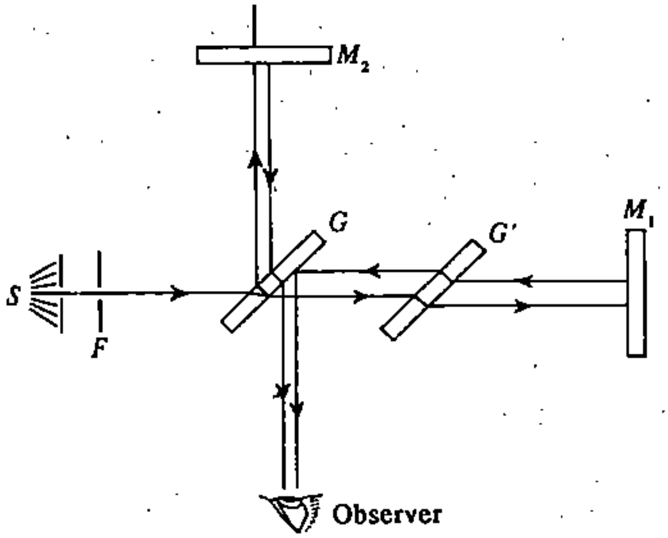


Fig.12.3: Light paths for Michelson Interterometer.

position as shown in the figure, circular fringes are observed due to interference of the beams reflected from mirrors $M_1$ and $M_2$. You may also recall that for obtaining these circular fringes, the mirrors should be at right angle to each other and the path difference $(GM_2 - GM_1)$ should be small. If mirror $M_2$ is moved away from the beam splitter $G$, the visibility and hence the contrast of the interference fringes will become poorer, and, eventually, the fringe pattern will disappear. Why does it happen? Does disappearance of interference fringes has to do something with temporal coherence of light waves from neon lamp? Yes, it is so. The disappearance of the fringes is due to the following phenomenon. When mirror $M_2$ is moved through a distance $d$, an additional path $2d$ is introduced for the beam which gets reflected by $M_2$. As a result, the beam reflected from $M_2$ interferes with the one reflected from $M_1$ which had originated $(2d/c)$ s, where $c$ is the velocity of light, earlier from the light source. Clearly, if this time delay $(2d/c)$ is greater than the coherence time $(\tau_c)$ of the radiation from the source, the waves reaching the eye after reflection from mirors $M_1$ and $M_2$ will not have any definite phase relationship. In other words, the waves reflected from mirrors $M_1$ and $M_2$ are incoherent. Thus, no interference fringes will be seen. On the other hand, if $(2d/c) < < \tau_c$, a definite phase relationship exists between the two reflected waves and hence interference fringes with good contrast will be seen. It is so because in this case, we are superposing two wave trains (after reflection from mirrors $M_1$ and $M_2$) which are derived from the same wave train (from the source) and hence they are temporally coherent.

For the neon light ($\lambda = 6328$ Å), the disappearance of fringes occurs when path difference between the reflected waves from mirrors $M_1$ and $M_2$ is about a few cm. This path difference, $L_c = c\ \tau_c$ is known as **coherence length.** Hence for neon line, $\tau_c \sim 10^{-10}$ s. For commercially available lasers, the coherence length exceeds a few kilometers. Thus, if light beam from a laser be used in the above experiment, we can observe interference fringes for $d$ as long as a few kilometers (provided, of course, we have such a big laboratory!).

In short, if the two paths, $GM_1$ and $GM_2$ in Fig. 12.3 are equal in length, the fringes have maximum contrast, hence a maximum temporal coherence. If they are not of equal length then the contrast is less. Hence temporal coherence is less. Temporal coherence is, therefore, inversely proportional to the magnitude of the path difference and directly proportional to the length of the wave train. The wave trains are of finite length; each containing only a limited number of waves. The length of a wavetrain is, therefore, the product of the number of waves, $N$, contained in a wave train and of its wave length $\lambda$, so $L_c = \lambda N$. Since visibility or the contrast of the interference fringes is directly proportional to the length of the wave train, it can also be taken as proportional to the product of $N$ and $\lambda$. Further, for a given source of light, you can have some idea about its temporal coherence in terms of the path difference between two interfering waves of Michelson interferometer. You should now work out the following SAQ.

*Spend 2 min*

SAQ 1

If light of 660 nm wavelength has a wavetrain $20\lambda$ long, what is its (a) coherence length and (b) coherence time.

### 12.3.1 Width of Spectral Line

You might have studied in school physics course about the origin of spectral lines. You may recall that when an atom undergoes a transition from an excited state to the ground state, it emits electromagnetic radiation. The energy (and hence frequency) of the radiation is equal to the difference in energies of the excited and the ground state. Each substance has a unique set of energy state to which its atoms can be excited. Each substance, therefore, has a characteristic set of energy values (and hence frequencies) for the emitted radiations. This set of frequency values constitutes the spectrum of the substance.

Due to one of the fundamental principles of quantum mechanics, namely, the uncertainty principle, about which you would learn in the course on Modern Physics (PHE-11), the lines in the spectrum are not sharp i.e. corresponding to each spectral line, there is a continuous distribution of frequency in a narrow frequency interval. This narrow frequency or wavelength interval is known as width of the spectral line. For example, for Cd red line, width of this interval is about 0.007 Å.

You may now be interested in knowing what determines the width of the spectral lines? Is width of spectral lines related to temporal coherence? Yes, temporal coherence of the source of light is intimately related to the width of its spectral lines. To see how, let us again consider the interference fringes obtained by Michelson interferometer. You may recall from Unit 7 that Michelson's Interferometer can be used for the measurement of two closely spaced wavelengths. Let us consider a sodium lamp source which emits predominantly two closely spaced wavelengths, $\lambda_1 = 5896$ Å and $\lambda_2 = 5890$ Å. Now, you may recall from Unit 7 that near $d = 0$, the fringe patterns corresponding to both the wavelengths will overlapp. If the mirror is moved away from the plate $G$ by a distance $d$, Fig. 12.3, the maxima corresponding to the wavelength $\lambda_1$ will not, in general, occur at the same angle as for $\lambda_2$. It is so because the spacing between the fringes for $\lambda_1$ and $\lambda_2$ will be different. Indeed, if the distance $d$ is such that the bright fringe corresponding to $\lambda_1$ coincides with the dark fringe corresponding to $\lambda_2$, we have

$$2d = m\lambda_1 \qquad \text{(bright fringe)} \qquad (12.4a)$$

and

$$2d = \left(m + \frac{1}{2}\right)\lambda_2 \qquad \text{(dark fringe)} \qquad (12.4b)$$

and the fringe system will disappear. The condition for disappearence of fringe pattern can, therefore, be expressed as (see margin remark)

$$\frac{2d}{\lambda_2} - \frac{2d}{\lambda_1} = \frac{1}{2}$$

$$\Rightarrow \qquad 2d = \frac{\lambda_1\lambda_2}{2(\lambda_1 - \lambda_2)} = \frac{\lambda^2}{2(\lambda_1 - \lambda_2)} \qquad (12.5)$$

since $\lambda_1 \approx \lambda_2$.

If 'd' is the distance through which one of the mirrors has been moved, the effective path difference will be 2d. And the condition for bright fringe is

$$2d = m\lambda$$

and the condition for dark fringe is

$$2d = (m + 1/2)\lambda$$

where m is an integer.

$2d = m\lambda_1 \Rightarrow m = 2d/\lambda_1$

and

$2d = (m + 1/2)\lambda_2$

$\Rightarrow m = (2d/\lambda_2) - 1/2$

$\therefore \frac{2d}{\lambda_2} - \frac{2d}{\lambda_1} = \frac{1}{2}$

Now, if we assume that the light beam consists of all wavelengths lying between $\lambda$ and $\lambda + \Delta\lambda$, instead of two discrete values $\lambda_1$ and $\lambda_2$, fringes will not be observed if

$$2d \leq \frac{\lambda^2}{\Delta\lambda} \qquad (12.6)$$

To arrive at equation (12.6) you should solve the following SAQ.

**SAQ 2**

Starting from Eq. (12.5) which gives the path difference ($2d$), in terms of two distinct wavelengths $\lambda_1$ and $\lambda_2$, for which fringes will disappear, derive Eq. (12.6) which is for all wavelengths lying between $\lambda$ and $\lambda + \Delta\lambda$.

*Spend 5 min*

Now, can you see the basic reason why fringe pattern disappears? Is it somehow related to the nonmonochromaticity of the light beam? Yes, it is so. If fact, the moment we consider that the light beam consists of all wavelengths lying between $\lambda$ and $\lambda + \Delta\lambda$, we are essentially considering interference pattern produced by non-monochromatic light beam. You may notice from equation (12.6) that as the spread in the wavelength ($\Delta\lambda$) becomes small (more and more monochromatic), the path difference ($2d$) for disappearance of fringes becomes large. And as mentioned earlier, larger the value of path difference for which fringe pattern does not disappear, more temporally coherent the light beam is. In other words, monochromaticity or the sinusoidal nature of light beam is strongly related to its temporal coherence. The temporal coherence of the beam is, therefore, directly associated with the width of the spectral line. Since no fringe pattern is observed if the path difference, $2d$, exceeds the coherence length, $L_c$, we may assume that the beam consists of all the wavelengths lying between $\lambda$ and $\lambda + \Delta\lambda$ with

$$\Delta\lambda = \frac{\lambda^2}{L_c} \qquad (12.7)$$

This gives the relation between coherence length and spread in wavelength of a light beam. Further, since $\nu = c/\lambda$, the spread in frequency $\Delta\nu$ is

$$\Delta\nu = \frac{c}{\lambda^2}\Delta\lambda$$

$$= c/L_c$$

And, the coherence time is defined as, $\tau_c = L_c/c$. Therefore, we have

$$\Delta\nu \sim 1/\tau_c \qquad (12.8)$$

Thus the frequency spread of a spectral line is of the order of the inverse of the coherence time.

In this section, we discussed about temporal (or longitudinal spatial) coherence which relates the predictability or constancy of the phase relationship between two waves arriving at the same point after traversing different optical paths. In other words, we talked about the constancy of phases of waves travelling along the same line. Light beam was considered as a series of wave trains. As per requirement of temporal coherence, if these wave trains are to produce observable interference fringe pattern, they must (a) have the same frequency and (b) overlap at the point of observation (i.e. path difference should be less than the coherence length). Now, what about the phase relationship between two waves travelling side by side at a certain distance from each other? Well, the constancy of the phase relationship of such waves relates to another type of coherence called spatial (or transverse spatial) coherence. This is the subject matter of the next section.

## 12.4 SPATIAL COHERENCE

In unit 6, you studied about Young's double-slit experiment for obtaining interference fringe pattern. You may recall that one of the prerequisits for observing the interference pattern was that the source of light should be a point source. Can you say why this condition was imposed? What will happen if, instead of a point source, an extended source of light is used? These are some of the issues which relate to the spatial coherence about which we will study now.

You are aware that in an extended conventional source of light, the radiations emitted from different parts are independent of each other, and in that sense, such sources may be thought of as incoherent. But our interest is not so much in the nature of the source itself as in the quality of the illumination field it produces, for example, in a plane at some distance from the source. Thus, in Young's experiment we are interested in the extent to which there is a constant phase relationship between $S_1$, and $S_2$, Fig. 12.4a, so that interference effects can be observed. In other words, we are interested in examining the effect of the finite size of the source, $S$, on the interference pattern.

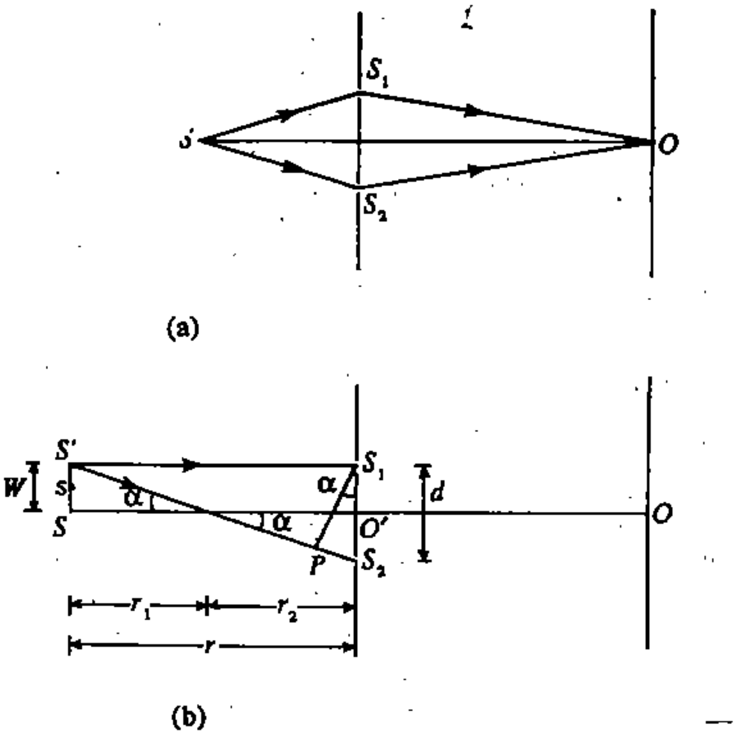


**Fig. 12.4: (a) Young's double slit experiment with a point source, S, (b) Young's double slit experiment with an extended source S ' S,**

In order to understand the effect of an extended source (and hence of spatial coherence) on the interference fringes, let us consider Young's double-slit experiment with an extended source. Fig. 12.4b shows schematically the two slits $S_1$ and $S_2$ with an extended light source $S'\,S$ of width $W$ at a distance $r$. Light from some point $s$ in the source illuminates the slits, and interference fringes are produced on the screen. If the source consisted of just this single point $s$ (as in an idealised Young's experiment, Fig. 12.4a), the fringes of maximum visibility would have been observed. A real source (such as $S'\,S$ in Fig. 12.4b) is, however, of finite size and the fringes produced by illumination from other points of the source are displaced relative to those due to $s$. Light from the extended source, therefore, produces a spread in fringes with a consequent reduction in the visibility of the fringe pattern.

In order to have some quantitative idea about the spatial coherence, let us assume that the two extreme points of the extended source (Fig. 12.4b), $S'$ and $S$ act as two independent sources. Each source will produce its own interference pattern. Let us assume that $SS_1 = SS_2$ and the point $O$ is such that $S_1O = S_2O$. Clearly the point source $S$ will produce a maximum around $O$. On the other hand, intensity at $O$ due to $S'$ will depend on the path length $(S'\,S_2 - S'\,S_1)$. You may recall from Unit 6, that if this path difference

$$S'\,S_2 - S'\,S_1 = \lambda/2 \tag{12.9}$$

the minima of interference pattern due to $S$ will fall on the maxima of that due to $S'$. As a result, there will not be any observable interference pattern. From Fig. 12.4b, we have

$$S'\,S_2 - S'\,S_1 = S_2P = \alpha d$$

But $$\alpha = \frac{d/2}{r_2} = \frac{W}{r_1}$$

Thus, $$r = r_1 + r_2 = \frac{1}{\alpha}\left(W + \frac{d}{2}\right)$$

or

$$\alpha = \frac{W + (d/2)}{r}$$

Therefore,

$$S' S_2 - S' S_1 = \alpha d = \left( W + \frac{d}{2} \right) \frac{d}{r} = \frac{Wd}{r}$$

$$(\text{ neglecting } d^2 \text{ term })$$

Thus, no fringes will occur if

$$\frac{Wd}{r} = \frac{1}{2}\lambda$$

$$W = \frac{r\lambda}{2d} \qquad (12.10)$$

We may, therefore, conclude that if we have an extended source whose linear dimension is $\sim\lambda r/d$, no interference fringe pattern will be observed. Equivalently, for a given source of width $W$, interference fringes will not be observable if the separation, $d$, between slits $S_1$ and $S_2$ is greater than $\lambda\, r/W$. If $\theta$ denotes the angle subtended by the source ( $S' S$ ) at the point $O'$ (midpoint of slits $S_1 S_2$), then $\theta = W/r$. So,

For every point on an extended source of extension r/d, there is a point at a distance r/2d which produces interference fringes separated by half a fringe width. Thus, for sources of such an extension, the visibility of the fringes would be poor.

$$d = \frac{\lambda}{\theta} \qquad (12.11)$$

which gives the maximum lateral distance between slits $S_1$ and $S_2$ such that the light beam from the extended source may be assumed to have some degree of coherence (i.e., the light waves from an extended source, after passing through slits $S_1$ and $S_2$ are able to produce interference fringes). The quantity $\lambda/\theta$ is known as **Lateral (or transverse) Coherence Width** and is denoted by $l_w$. You may note that the coherence width is linear in dimension and is approximately perpendicular to the direction of wave propagation. By contrast, the coherence length, introduced in relation to temporal coherence, is along the direction of wave propagation. For this reason, temporal coherence is sometimes called longitudinal coherence and spatial coherence is sometimes called lateral coherence.

Further, closely related to coherence width is a parameter called **coherence area** given as

$$\begin{aligned} a_c &= \pi\,(l_w/2)^2 \\ &= \pi\,(\lambda/2\theta)^2 \qquad (12.12) \end{aligned}$$

The waves at any two points within coherence area are coherent. You may have noticed that Eqs. (12.11) and (12.12) apply to the case in which the extended source is essentially a uniform linear source. If the source is in the form of an uniform circular disc, the lateral coherence width is given as

$$l_w = 1.22\,\lambda/\theta \qquad (12.13)$$

Well, in order to recaptulate what you have studied in this section, how about solving an SAQ!

SAQ 3

*Spend 2 min*

Suppose we set up Young's experiment with a small circular hole of diameter 0.1 mm in front of a sodium lamp ($\lambda \sim 589.3$nm) source. If the distance from the source to the slits is 1m, how far apart will the slits be when the fringe pattern disappears?

### 12.4.1 Angular Diameter of Stars

Now, let us consider an application of the concept of spatial coherence. In the preceding paragraphs, we have seen that the angle subtended by the extended source at the midpoint of the slit separation is related to the lateral coherence width ($l_w$). Also for a critical value of $l_w$, the interference fringes will disappear. If, instead of an ordinary extended source of light, we consider a terrestrial extended source such as a star, you may like to know: Is it be possible to know its angular diameter (i.e. the angle subtended by the star on the slits) by observing the disappearence of fringes? Indeed, it is possible. For measuring the angular diameter of a star, Young's double slit experiment set-up needs modification. Modification in the experimental set-up is necessitated because, for such an arrangement, if we take a typical value of the angular diameter of a star as $\sim 10^{-7}$ radians, the distance $d$ between the slits for which fringes disappear will be

$$d = \frac{1.22\lambda}{\theta} = \frac{1.22 \times 5 \times 10^{-7}}{10^{-7}}$$

$$\simeq 6\text{m}.$$

And for such a large value of $d$, the fringe width will be too small.

To overcome this difficulty, Michelson used an ingeneous technique. He achieved an effectively large value of $d$ by using two movable mirrors $M$ and $M'$ as shown in Fig. 12.5. This modified interferometer is known as **Michelson's Stellar Interferometer.**

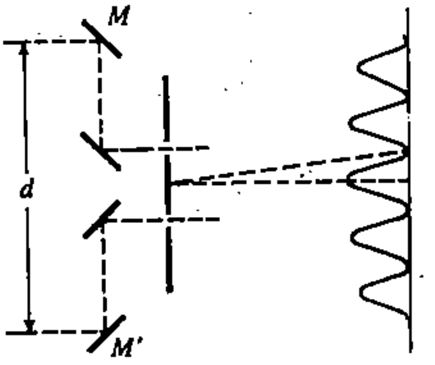


**Fig. 12.5: Michelson's stellar interferometer.**

Since you have studied in detail about the Michelson Steller interferometer in Unit 7, we would just mention here the results of one typical experiment. In a typical experiment the first disappearance of fringes occured when the distance between mirrors $M$ and $M'$ was about 7.3m which gave the angular diameter, $\theta$ as (taking $\lambda \sim 5 \times 10^{-7}$m)

$$\theta = 1.22\lambda/d$$

$$\theta = \frac{1.22 \times 5 \times 10^{-7}}{7.3}$$

$$= 8.4 \times 10^{-8} \text{ rad}$$

From the known distance of star and the value of its angular diameter, $\theta$, we can estimate its diameter.

### 12.4.2 Visibility of Fringes

Till now, we have been discussing coherence and its importance for observing interference fringes. We have been talking about the disappearence of fringes under different circumstances. For example, in the Young's double slit experiment, interference fringes are seen on a screen with highly spatially coherent light. The fringes are rather distinct, their visibility is high. As the two slits are moved further apart the fringes are more closely spaced and will lose visibility. The degree of visibility, therefore, is the measure of spatial coherence.

The amount of radiation power incident per unit area is called areana.

Assume that two wave trains of light, each of finite length $\Delta l$, overlap to their full extent. Such complete overlap will result in distinct maxima and minima of highest degree of visibility. Even if the wave trains overlap partially, as in Fig. 12.6, interference is possible. However, the degree of visibility of the fringes will diminish depending on the extend of overlap. The question, therefore, is not how much the wave trains must overlap to produce interference; rather, the question is how much visibility we need to see a fringe pattern?

The definition of visibility is essentially a matter of comparison. Visibility, $V$, can be defined as the ratio of the difference between the maximum areana $E_{max}$, and minimum areana $E_{min}$, to the sum of the areanes; i.e.

$$V = \frac{E_{max} - E_{min}}{E_{max} + E_{min}}; \qquad (12.14)$$

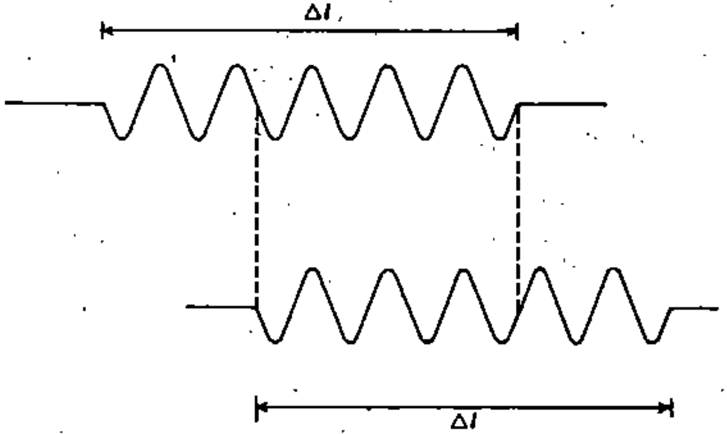


Fig.12.6: Partial overlap of two wave trains.

Let us assume that $E_{max}$ can take any arbitrary value but $E_{min} = 0$. Then visibility, $V = 1$. On the other hand, if $E_{max} = E_{min}$, $V = 0$, fringes cannot be seen. Thus, the visibility may assume any value between 0 and 1. Generally, a visibility of 0.8 is considered high, but a value 0.2 is barely visible.

Now, you may like to know whether visibility is related to coherence? Yes, it is. To see how, let two points on a distant screen be illuminated by two light sources that produce equal areanas $E_o$. Light waves from each consists of two parts-coherent ($A$) and incoherent ($B$). Areana due to the coherent part, $A$ can be expressed as

$$E_A = \rho E_o$$

where, $\rho$ is the degree of coherence,

and, the areana due to incoherent part is,

$$E_B = (1 - \rho) E_o$$

Interference fringes are observed because of part $A$. The coherent part forms fringes whose maxima have intensities. Areana of the maxima is four times as high as the individual contribution. Thus, the maximum areana, $(E_A)_{max}$, is $4\rho E_o$ and minimum is zero. Moreover, on this interference fringe pattern, due to the coherent part $A$, a **uniform distribution** due to incoherent part $B$, is superimposed. The areana of this distribution will be twice as high as the contribution $E_B$, because it comes from two sources. Hence,

$$(E_B)_{max} = 2E_B = 2(1-\rho)E_o = (E_B)_{min}$$

As a result, the areana in the maxima is

$$E_{max} = (E_A)_{max} + (E_B)_{max}$$

$$E_{max} = 4\rho E_o + 2(1-\rho)E_o$$

$$= 2(1+\rho)E_o$$

and the areana in the minima

$$E_{min} = (E_A)_{min} + (E_B)_{min}$$

$$= 0 + 2(1-\rho)E_o$$

Therefore, Eq. (12.14) for visibility of the fringes can be written as,

$$V = \frac{2(1+\rho)E_o - 2(1-\rho)E_o}{2(1+\rho)E_o + 2(1-\rho)E_o}$$

$= \rho$, the degree of coherence.

Thus, the degree of visibility (or the contrast) of the fringes producéd by two light waves is equal to degree of coherence between them.

The highest visibility and hence highest degree of coherence will occur when the minimum areana in the expression for $V$ is zero. In that case, both the visibility and the degree of coherence are unity. Although concievable in theory, but this is not possible in practice. Complete coherence is merely a theoretical result. However, with the development of laser, about which you would study in the next unit, it is now possible to have light beam of extremely high degree of coherence.

## 12.5 SUMMARY

- Coherence is a property of light. A predictable phase relation exists between light waves passing through a point at different times.
- Temporal coherence or the longitudinal spatial coherence refers to the predictability of the phase of radiation as a function of time. In other words, temporal coherence can be identified as the interval of time during which the phase of the wave changes in a predictable manner as it passes through a fixed point in space. This time interval is known as **coherence time**, $\tau_c$. And the path length corresponding to $\tau_c$, given as $L_c = c\,\tau_c$ is called the **coherence length** of the radiation.

- Temporal coherence is related with the width of the spectral lines. The spread in wavelength is given as,

$$\Delta\lambda = \frac{\lambda^2}{L_c}$$

and the corresponding spread in the frequency of the spectral line is

$$\Delta\nu \sim 1/\tau_c$$

- Spatial coherence or transverse spatial coherence refers to the correlation between the phases of two light waves travelling side by side. Use of point source in Young's double slit experiment is essentially to meet the requirement of spatial coherence.
- If an extended source of light of width $W$ is used in Young's interference experiment, for observing interference fringe pattern following condition must be satisfied

$$W = \lambda/\theta$$

where, $\lambda$is the wavelength of the light, and $\theta$ is the angle subtended by the extended source on the slits.

The quantity $(\lambda/\theta)$ is known as **lateral (or transverse) coherence width, $l_w$.**

- For a circular extended source, the coherence width $l_w$ is given as

$$l_w = \frac{1.22\lambda}{\theta}$$

- Visibility of an interference pattern is given as

$$V = \frac{E_{max} - E_{min}}{E_{max} + E_{min}}$$

where, $E_{max}$ is maximum areana and $E_{min}$ isminimum areana.

In terms of the degree of coherence, $\rho$, the visibility is given as

$$V = \frac{2(1+\rho)E_o - 2(1-\rho)E_o}{2(1+\rho)E_o + 2(1-\rho)E_o}$$

where, $E_0$ is the areana produced on the screen by individual light source.

## 12.6 TERMINAL QUESTIONS

1. The sodium line at $\lambda = 5890$ Å, produced in a low-pressure discharge, has spread in wavelength, $\Delta\lambda = 0.0194$ Å. Calculate (a) the coherence length and (b) line width in hertz.

2. If the visibility in an interference fringe pattern is 50 percent and the maxima receive 15 units of light, how much light does the minima receive?

## 12.7 SOLUTIONS AND ANSWERS

**SAQs**

1. The wavelength of the light, $\lambda = 660$ nm and $N$,the number of waves in the wave train is 20.

   (a) So, the coherence length

$$L_c = N\lambda$$

$$= 20 \times 660 \text{ nm}$$

$$= 13200 \text{ nm} = 13.2 \times 10^{-12} \text{m}$$

   (b) Coherence time

$$\tau_c = L_c/c \text{ ; where } c = \text{velocity of light} = 3 \times 10^8 \text{ ms}^{-1}$$

$$\tau_c = \frac{13200 \times 10^{-9} \text{ m}}{3 \times 10^8 \text{ ms}^{-1}}$$

$$= 4400 \times 10^{-17} \text{s}$$

$$= 4.4 \times 10^{-20} \text{s}$$

2. Eq. (12.5), $2d = \lambda^2/2\,(\lambda_1 - \lambda_2)$ gives the path difference for the disappearance of fringe pattern due to light of wavelengths $\lambda_1$ and $\lambda_2$. When this expression is to be used for the disappearance of the fringe pattern due to the light beam consisting of all wavelengths lying between $\lambda$ and $\lambda + \Delta\lambda$, we must divide the interval (width) into two equal parts of $\Delta\lambda/2$. Thus, the fringe pattern will be produced by wavelength values

$$\lambda_1 = \lambda + (\Delta\lambda/2)$$

$$\lambda_2 = \lambda$$

   With these values, Eq. (12.5) reduces to

$$2d = \frac{\lambda^2}{2((\lambda + \Delta\lambda/2) - \lambda)} = \frac{\lambda^2}{2(\Delta\lambda/2)} = \frac{\lambda^2}{\Delta\lambda}$$

   which is Eq. (12.6)

   Now, for each wavelength lying between $\lambda$ and $\lambda + \Delta\lambda/2$, there will be a corresponding wavelength lying between $\lambda + \Delta\lambda/2$ and $\lambda + \Delta\lambda$ such that the minima of one falls on the maxima of the other. Therefore, the fringe pattern will disappear.

3. Width (or the diameter) of the source

$$W = 0.1 \text{ mm} = 1 \times 10^{-4} \text{m}$$

And distance between the source and slits

$$r = 1\text{m}$$

Hence the angle subtended by the source on slits

$$\theta = \frac{W}{r} = \frac{1 \times 10^{-4}\text{m}}{1\text{m}} = 10^{-4}\,\text{radian}$$

Wavelength of the light

$$\lambda = 589.3 \times 10^{-9}\text{m}$$

The lateral coherence width for a circular extended source

$$l_w = \frac{1.22\lambda}{\theta} = \frac{1.22 \times 589.3 \times 10^{-9}\text{m}}{10^{-4}\text{rad}}$$

$$= 0.72\text{ cm}$$

Thus, if the separation between the slits is more than 0.72 cm, the fringe pattern will disappear.

TQs

1. $\lambda = 5890\,\text{Å} = 5890 \times 10^{-10}\text{m}$

$\Delta\lambda = 0.0194\,\text{Å} = 0.0194 \times 10^{-10}\text{m}$

(a) From equation (12.7), we have

$$\Delta\lambda = \frac{\lambda^2}{L_c}\text{; where, } L_c = \text{ coherence length}$$

$$\Rightarrow \qquad L_c = \frac{\lambda^2}{\Delta\lambda} = \frac{(5890 \times 10^{-10})^2\,\text{m}^2}{0.01194 \times 10^{-10}\,\text{m}}$$

$$= 0.18\text{m}$$

(b) The spread in frequency $\Delta\nu$ (line width in hertz) and the coherence time $\tau_c$ is related as (equation (12.8))

$$\Delta\nu = \frac{1}{\tau_c} = \frac{1}{L_c/c} = \frac{c}{L_c};$$

where $c$ = velocity of light = $3 \times 10^8$ m/s

$$\Rightarrow \Delta\nu = \frac{3 \times 10^8\,\text{m/s}}{0.18\text{ m}}$$

$$= 1.6 \times 10^9\,\text{Hz.}$$

2. The visibility of an interference fringe pattern is given as

$$V = \frac{E_{max} - E_{min}}{E_{max} + E_{min}}$$

where

$E_{max}$ is the maximum areana i.e. the amount of radiation power contained in the maxima of the fringe pattern; and $E_{min}$ is the minimum areana.

From the problem, we have

$$V = 50 \text{ percent} = \frac{1}{2}; \; E_{max} = 15 \text{ units}; \; E_{min} = ?$$

So, from above equation for visibility, we have

$$\frac{1}{2} = \frac{15 - E_{min}}{15 + E_{min}}$$

$$\Rightarrow \quad (15 + E_{min}) = 2(15 - E_{min})$$

$$\Rightarrow \quad E_{min} = 5 \text{ units}$$

Hence, 5 units of light will be received in the minima of the fringe pattern.

# UNIT 13 PHYSICS OF LASERS

## Structure

## 13.1 INTRODUCTION

In the previous unit, you learnt about coherence and coherent sources of light. It was explained there why conventional thermal sources of light emit radiation which have very low degree of coherence. However, phenomenon like interference which requires coherent light sources, can indeed be observed with conventional light sources. The quest for obtaining a light source with high degree of coherence led to the invention of lasers. As you know, a useful indicator of the degree of coherence is the coherence length. For ordinary light, the coherence length is of the order $10^{-2}$m, whereas the coherence length for a laser light can be as long as $10^{5}$m! So, you may appreciate the difference in the degree of coherence between an ordinary light and the laser light. In the present unit, we will discuss about this source of highly coherent light beam-the LASER.

The name laser is an acronym for *L*ight *A***mplification by** *S***timulated** *E***mission of** *R***adiation.** You must realise that the key words here are amplification and stimulated emission. The existence of stimulated emission of radiation, when radiation interacts with matter, was predicted by Einstein in 1916. His theoretical prediction was realised by *C.H.* Townes and co-workers in 1954 when they developed microwave amplification by stimulated emission of radiation (**maser**). The principle of maser was adapted for light in visible range by *A*. Schawlow and *C.H.* Townes in 1958 but the first laser device was developed by *T.H.* Maiman in 1960. Once the laser was invented, it has found applications in such diverse fields as basic research, industry, medicine, space, photography, communication, defence, etc.

In Sec. 13.2, you will learn about the quantum mechanical description of the emission and absorption of light. In particular, you will learn about spontaneous emission and stimulated emission of radiation. In Sec. 13.3, the physical principles involved in the operation of lasers viz.excitation (or pumping), the need of an active medium and the feedback mechanism have been explained. Since the invention of laser by Maiman using small ruby rod as active medium, Lasers have come a long way. Presently, lasers are built using solid or liquid or gas as active media. Apart from these, now semi-conductor based lasers are finding wide applications. These different types of lasers have been briefly discussed in Sec. 13.4. The applications of lasers are so many and so varied that their detailed account will take us too far. In Sec. 13.5, we have, however, briefly discussed applications of lasers in industry, medicine, communication and basic research. In the next unit, you will study about holography, which would not have been possible without laser light. And in Unit 15, you will study about optical fibres-a medium of transporting light-which is a very active area of research and development for long distance optical communication purposes.

### Objectives

After going through this unit, you should be able to

- explain the concept of stimulated emission of radiation and differentiate it from spontaneous emission
- describe the need and methods of pumping
- list the characteristics of the active medium for lasers
- describe different types of lasers, and
- describe the important applications of lasers.

## 13.2 LIGHT EMISSION AND ABSORPTION

As you are aware, most of the man-made sources of light are the solids and gases heated to high temperatures. For example, in case of incandesent bulb, the tungsten filament is heated, and in case of murcury tube light, the gas is heated. The energy of the heating source is absorbed by the atoms or molecules of the solid or the gas, which, in turn, emit light. The basic mechanism of the origin of light from within gas molecules, liquids and solids is similar in many respect to that from an individual atom. And the process of emission and absorption of light from atoms can be understood in terms of Bohr's atomic model. Though you might have studied Bohr's model in your school physics course, we briefly discuss it here for the sake of completeness.

### 13.2.1 Quantum Theory: A Brief Outline

You may recall from your school physics course that according to Bohr's theory, the energy of an atom or a molecule can take on only definite (discrete) values. These are known as the energy levels of the atom. The transition of an atom from one energy level to another energy level occurs in quantum jump. This was one of the basic assumptions of Bohr's theory. On the basis of this presumption, Bohr postulated that light is not emitted by an electron when it is revolving in one of its allowed orbits (and hence has a fixed value of energy). Light emission takes place when the atom makes a transition from an excited state (of energy $E_i$) to a state of lower energy $E_f$. The frequency of the emitted radiation is given by

$$h\nu = E_i - E_f \qquad (13.1)$$

where $E_i$ is the energy of the initial orbit, $E_f$ is the energy of the final orbit, $\nu$ the frequency of the emitted light and $h$ is the Planck's constant. The quantized orbits of the electron and the energy level diagram of the simplest atom-the hydrogen atom-are shown in Fig. 13.1.

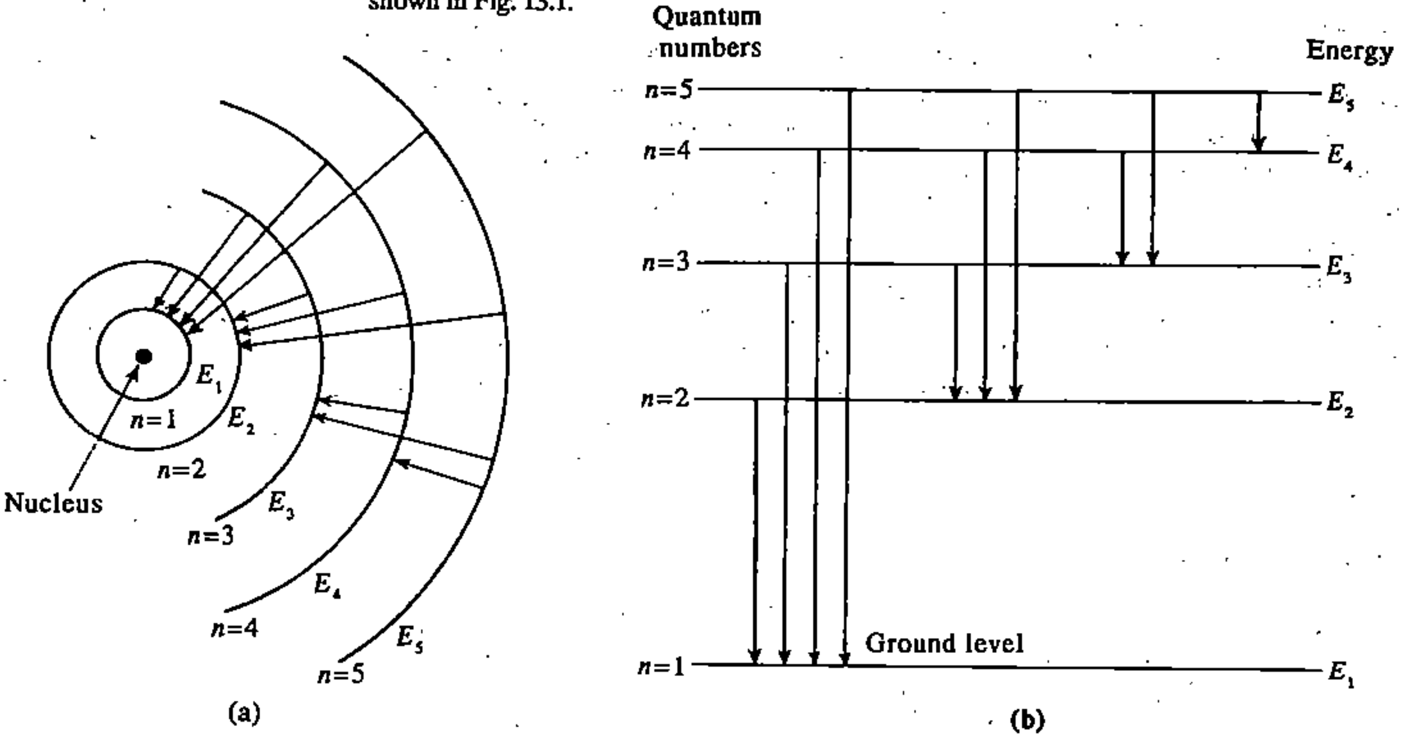


**Fig. 13.1: (a) Bohr circular orbits for the revolving electron of hydrogen atom, showing transitions, giving rise to the emitted light waves of different frequencies; (b) Energy level diagram for the hydrogen atom.**

The quantum mechanical explanation about the origin of light, as discussed above, applies to all the known light sources. To focus our attention on the atomic processes involved in the emission and absorption of light, let us consider only two energy levels of an atom. Let the energy of the lower level be $E_1$ and that of the upper level be $E_2$. An atom lying in level $E_2$ will tend to make a transition to level $E_1$ so that it occupies a state of lower energy. Such emission process is known as **spontaneous emission** because it occurs in the absence of any external stimulus. The process of spontaneous emission is shown in Fig. 13.2(a). The photon emitted in spontaneous emission will have the energy ($E_2$-$E_1$), while its other characteristics such as momentum, polarisation, will be arbitrary. The light emitted by ordinary sources results due to spontaneous emission. **Absorption of light** is the converse process of emission. The atom in a lower energy state can absorb a photon of energy $h\nu$ $(= E_2 - E_1)$ and get excited to the upper level $E_2$. The absorption process is depicted in Fig. 13.2(b).

Now, can you guess what will happen if an atom is in the higher energy level, $E_2$, and a photon of energy $h\nu$ $(= E_2 - E_1)$ interacts with it? Well, in such a situation, the photon may trigger the atom in the upper level to emit radiation. This emission process is known as **stimulated emission**. When the atom is already in the higher energy level, the photon, instead of being absorbed, may play the role of a trigger, and induce the transition from $E_2$ to $E_1$. As a result, the atom falls into lower energy level and an **additional** photon of energy $h\nu = E_2 - E_1$ is emitted. **In this process of stimulated emission, shown in Fig. 13.2(c), both the inducing and the induced photons have the same energy.** The light from laser is due to the stimulated emission of radiation.

It is worth mentioning here that of the three processes mentioned above, only the first two, that is, the spontaneous emission and the absorption of light were postulated on the basis of Bohr's theory. It was only when Einstein considered the whole idea of emission

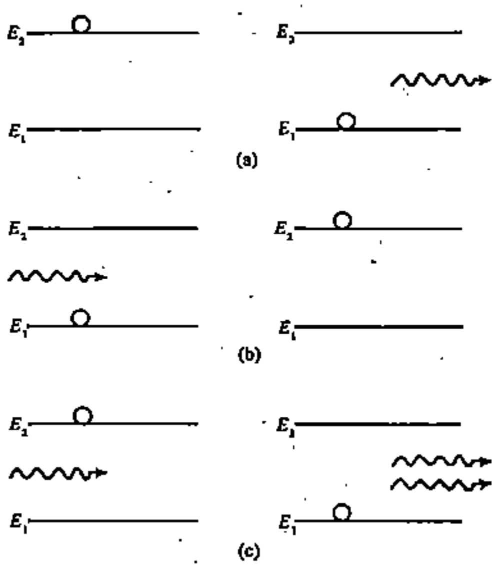


Fig. 13.2: (a) Spontaneous Emission (b) Absorption and (c) Stimulated Emission of light

and absorption of radiation in terms of thermodynamic equilibrium between matter and radiation that stimulated emission of radiation could be predicted. What were Einstein's theoretical arguments for the prediction? Let us learn these now.

### 13.2.2 Stimulated Emission of Radiation: Einstein's Prediction

Stimulated emission, as mentioned above, is the reverse of the process in which electromagnetic radiation or photons are absorbed by the atomic systems. When a photon is absorbed by an atom, the energy of the photon is converted into the internal energy of the atom. The atom is then raised to an excited (higher energy) state and it may radiate this energy spontaneously, emitting a photon and reverting to the ground (or some lower energy) state. However, during the period the atom is in the excited state, it can be stimulated to emit a photon if it interacts with another photon. This stimulating photon should have precisely the energy of the one that would otherwise be emitted spontaneously. Let us look at the theoretical arguments put forward by Einstein for the existence of stimulated emission.

Refer to Fig. 13.3 which shows a system of two energy levels $E_1$ and $E_2$ with population

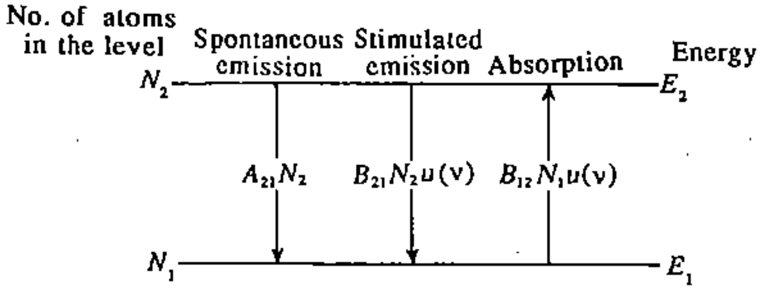


Fig. 13.3: An atomic system of two energy levels showing different emission and absorption processes.

of atoms $N_1$ and $N_2$ respectively. Let $E_1 < E_2$. You may recall from Unit 13 of PHE-06 that according to Maxwell-Boltzmann distribution, the ratio of population of atoms in different levels for the system in thermal equilibrium is given as

$$\frac{N_2}{N_1} = e^{-(E_2 - E_1)/k_B T}$$

$$\text{or} \quad N_2 = N_1 e^{-h\nu/k_B T} \tag{13.2}$$

where, $k_B$ is the Boltzmann constant and $T$ is the absolute temperature.

Now what will be the ratio of the population of the energy levels if radiation of energy $h\nu$ is introduced into the system? Einstein proposed that if this system of energy levels and the radiations is to remain in thermal equilibrium, the rate of downward transition (due to spontaneous and stimulated emission) must be equal to the rate of upward transition (due to absorption). He, therefore, arrived at the relation (see box below),

$$\frac{N_2}{N_1} = \frac{B_{12}u(\nu)}{A_{21} + B_{21}u(\nu)} \tag{13.3}$$

where, $u(\nu)$ is the energy density of radiation at frequency $\nu$ and $B_{12}, A_{21}, B_{21}$ are Einstein's co-efficients. $A_{21}$ is associated with spontaneous emission, $B_{21}$ is associated with stimulated emission and $B_{12}$ is associated with absorption.

Following Einstein, let us write down the rates of spontaneous and stimulated emission and the rate of absorption of radiation. The rate of spontaneous emission will be independent of the energy density of the radiation field because for this process to occur, presence of photon is not required. This emission process will be proportional to the number of atoms, $N_2$, in the higher energy state. So, we may write the rate of spontaneous emission as

$$P_{21} = N_2A_{21} \tag{i}$$

where $A_{21}$ is constant of proportionality.

Assume next that the system of atoms is subject to some external radiation field. In that case, as mentioned earlier, one of the two processes, namely, the stimulated emission and absorption, may occur. The probability of their occurrence depends on the energy density of radiation at the particular frequency separating the two levels and the population of states from which transition takes place. Therefore, the rate of stimulated emission will be proportional to the energy density of the radiation and the population of higher energy state, $N_2$. Thus, the rate of stimulated emission

$$P_{21} = N_2B_{21}u(\nu) \tag{ii}$$

where $B_{21}$ is another constant of proportionality and $u(\nu)$ is energy density of radiation at frequency $\nu$.

On the other hand, the rate of absorption will depend on $u(\nu)$ and the population of the lower energy state, $N_1$. Thus, the rate of absorption

$$P_{12} = N_1B_{12}u(\nu) \tag{iii}$$

where $B_{12}$ is the constant of proportionality. The constants $A_{21}$, $B_{12}$ and $B_{21}$ are known as **Einstein's coefficients.**

With the system in thermal equilibrium, the net rate of downward transition must be equal to the net rate of upward transition. Thus, we may write

$$N_2A_{21} + N_2B_{21}u(\nu) = N_1B_{12}u(\nu) \tag{iv}$$

Dividing both side by $N_1$, we get

$$\frac{N_2}{N_1}A_{21} + \frac{N_2}{N_1}B_{21}u(\nu) = B_{12}u(\nu)$$

or

$$\frac{N_2}{N_1}(A_{21} + u(\nu)B_{21}) = B_{12}u(\nu)$$

so that

$$\frac{N_2}{N_1} = \frac{B_{12}u(\nu)}{A_{21} + u(\nu)B_{21}} \tag{v}$$

Form Eqs. (13.2) and (13.3), we have

$$\frac{B_{12}\,u\,(\nu)}{A_{21}\,+\,u\,(\nu)\,B_{21}} = e^{-\,h\,\nu/k_B\,T}$$

or

$$u\,(\nu) = \frac{A_{21}}{B_{12}}\,\frac{1}{e^{h\,\nu/k_B\,T}\,-\,(B_{21}/B_{12})} \tag{13.4}$$

Now, you may recall from unit of PHE-06 that the energy density of black body radiation is given by Planck's radiation law:

$$u\,(\nu) = \frac{8\pi\,h\nu^3}{c^3}\,\frac{1}{e^{h\,\nu/k_B\,T}-1} \tag{13.5}$$

Equation (13.5) must be same as Eq. (13.4). So we must have

$$B_{21} = B_{12} \tag{13.6}$$

and

$$A_{21}/B_{21} = \frac{8\pi\,h\nu^3}{c^3} \tag{13.7}$$

These are **Einstein's relations**. On the basis of Einstein's relations, we can conclude the following:

(a) Eq. (13.6) indicates that the probabilities of absorption and stimulated emission are the same. In other words, when an atomic system is in equilibrium, absorption and emission take place side by side. Normally, $N_2 < N_1$, and absorption dominate stimulated emission. An incident photon is more likely to be absorbed than to cause stimulated emission. But, if we could find a material that could be induced to have a majority of atoms in the higher state than in the lower state, i.e. $N_2 > N_1$, the stimulated emission may dominate absorption. This condition of the atomic system (where $N_2 > N_1$) is known as **population inversion**. And when the stimulated emission dominates over absorption in the atomic system, it is said to lase.

(b) If we substitute $B_{12} = B_{21}$ in equation (13.4), we get the ratio of the number of spontaneous emission to stimulated emission

$$\frac{A_{21}}{B_{21}\,u\,(\nu)} = e^{h\,\nu/k_B\,T}-1 \tag{13.8}$$

When the system is in thermal equilibrium at temperature $T$, for $h\nu << k_BT$, Eq. (13.8) suggests that stimulated emission will dominate spontaneous emission. On the other hand, when $h\,\nu >> k_BT$, spontaneous emission will dominate stimulated emission. Now which of these two processes will dominate for ordinary thermal sources of light? To know that, you should do the following SAQ.

SAQ 1

The absolute temperature, $T$, for an ordinary source of light is typically of the order of $10^3$K. With the help of Eq. (13.8), show that in such sources, the process of spontaneous emission will dominate over the stimulated emission.

*Spend 5 min*

## 13.2.3 Einstein's Prediction Realised

You now know that when matter and radiation are in thermal equilibrium, besides spontaneous emission and absorption of radiation by matter, there must be a third process, called stimulated emission. This prediction did not attract much attention untill

1954, when Townes and coworkers developed a microwave amplifier (MASER) using $NH_3$. In 1958, Shawlow and Townes showed that the maser principle could be extended into visible region. In 1960, the prediction was realised by Maiman who built the first laser using Ruby as an active medium. Maiman found that a suitable active component for a laser could be made from a single crystal of pink ruby: aluminium oxide ($Al_2O_3$), coloured pink by addition of about 0.5 percent chromium. For any laser action to take place, a condition of population inversion must be met. By population inversion we mean that the number of atoms in higher energy state is larger than the ground (or some lower energy) state. The energy states of the chromium atom, as shown in Fig. 13.4, are ideal for obtaining population inversion. The chief characteristics of energy levels of a Chromium atom is that the levels labelled as $E_1$ and $E_2$ have a life time $10^{-8}$s. whereas the state marked $M$ has a life time $3 \times 10^{-3}$s. The energy state $M$ with such a long life time (as compared to other excited states) is called a **metastable state.**

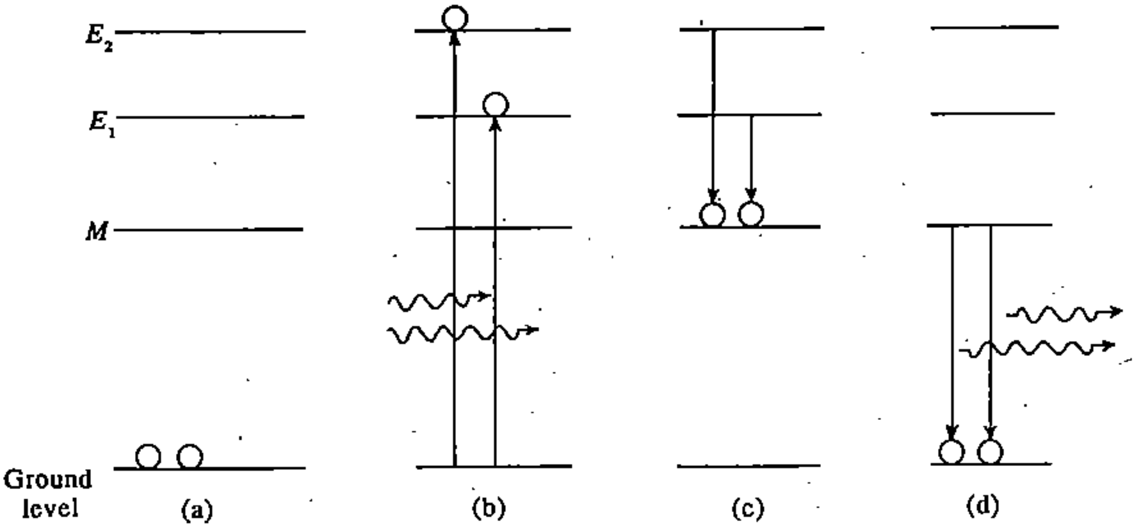


**Fig.13.4: Energy levels of chromium atom: (a) atoms in the ground state (b) on absorbing photons, atoms are excited to one of the two energy levels $E_1$ and $E_2$. (c) atoms give up some of its energy to the crystal lattice and fall to a metastable level, M, (d) When stimulated by photons, the atoms in metastable level emit photon and fall to ground state.**

A chromium atom in its ground state can absorb a photon ($\lambda = 6600$ Å) and make a transition to the level $E_1$; it could also absorb a photon of $\lambda$ ~4000 Å and make a transition to the level $E_2$. In either case, it subsequently makes a non-radiative transition, in time $10^{-8}$s, to the metastabe state $M$. Since the state $M$ has a very long life, the number of atoms in this state keeps on increasing and we may achieve a population inversion between the state $M$ and $G$ (the ground state). Thus, we may have larger number of atoms in the level $M$ compared to those in the state $G$. Once population inversion is achieved, light amplification can take place.

When an atom undergoes a non-radiative transition the energy is not released in the form of photons; rather, the energy is transferred via atomic collisions, collision with the crystal lattice etc.

In the original set up of Maiman, the pink ruby was machined into a rod of length nearly four centimeter and diameter half a centimeter. Its ends were polished optically flat and parallel and were partially silvered. The rod was placed near an electronic flash tube (filled with xenon gas) that provided intense light for pumping chromium atoms to higher energy states. The set up of ruby laser is shown in Fig. 13.5. When the required population inversion was achieved with the help of electronic flash tube, the first few photons released (at random) by atoms dropping to the ground state stimulated a cascade of photons, all having the same frequency.

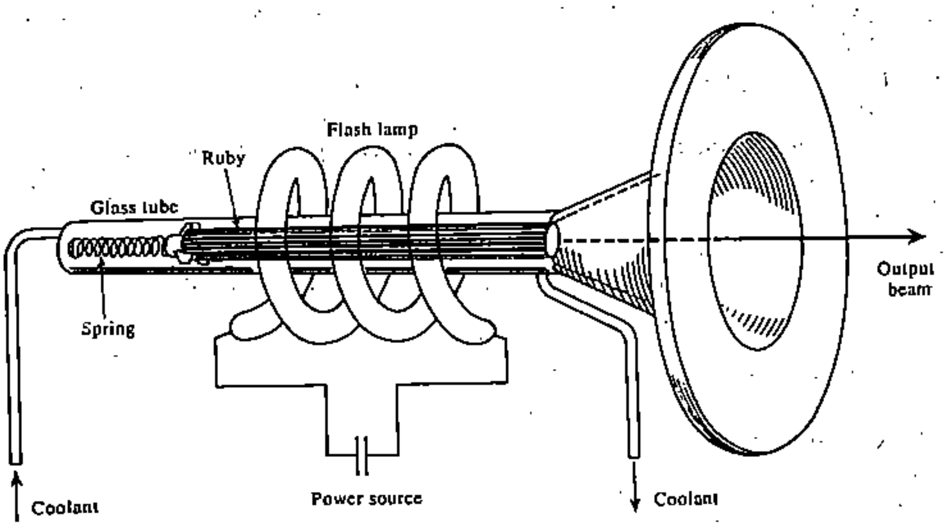


Fig.13.5: The Ruby Laser

You now know how a ruby laser, developed by Maiman, works. You will appreciated that production of laser light demands that certain conditions must be met beforehand. (We deliberately avoided reference to these in above paragraphs.) Firstly, is it possible to achieve laser light from any medium? If not, what are the chahracteristics of the medium which can produce laser light after proper excitation? (The media capable of producing laser light are called active media.) Secondly, how do we achieve population inversion? Further, for sustained laser light, it is necessary to feed some of the output energy back into the active medium. This is known as feedback and is achieved by resonant cavity. What is the nature of this resonant cavity for lasers? These are some of the important aspects of laser operation and design about which you will learn now.

## 13.3 PREREQUISITS FOR A LASER

A laser requires three prerequisites for operation. Firstly, there should be an **active medium** which, when excited, supports **population inversion** and subsequently lases. Secondly, we should ensure **pumping** mechanism, that raises the system to an **excited state.** And lastly, in most cases, there is an **optical cavity** that provides the feedback necessary for laser oscillation. These are shown schematically in Fig. 13.6.

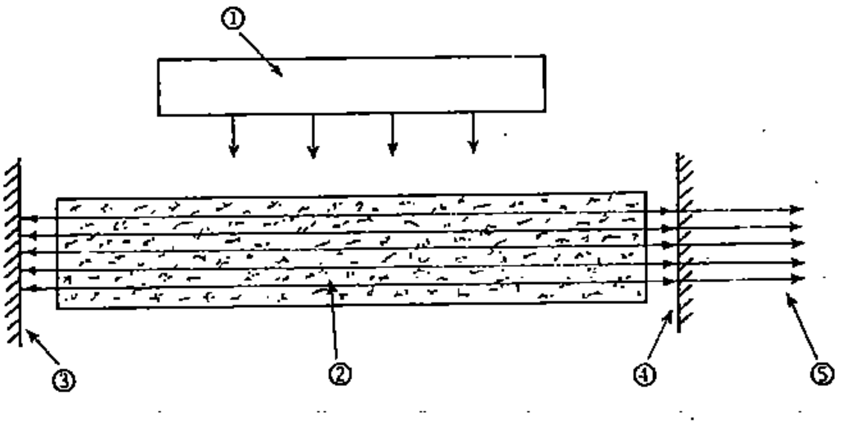


Fig.13.6: Basic components of a laser oscillator: Energy source (1) supplies energy to active medium (2). Medium is contained between two mirrors (3 and 4). Mirror 3 is fully reflective while mirror 4 is partially transparent. Laser radiation (5) emerges through partially transparent mirror.

In a typical laser operation, energy is transferred to the active material, which is raised to the excited state, and ultimately lases in various ways. The medium may be a solid, liquid or gas and it may be one of the thousands of materials that have been found to lase. The process of raising the medium to the excited state is called **pumping**, in analogy to pumping of water from lower to a higher level of potential energy. Some lasers are built as laser amplifier. They need no optical cavity. Most lasers, however, are laser oscillators. For sustained laser oscillations, some kind of feedback mechanism is needed. The feedback mechanism is provided in the form of optical resonant cavity. In both laser amplifiers and oscillators, the first few quanta of radiation will probably be emitted spontaneously and will trigger stimulated emission.

Let us now discuss the above mentioned three components of a laser.

## 13.3.1 The Active Medium

The heart of the laser is a certain medium – solid, liquid or gaseous – called an active medium. Since Maiman's discovery of ruby, many new laser materials have been discovered. They include crystals other than ruby, glasses, plastics, liquids, gases and even plasma (the state of matter in which some of the atomic electrons are dissociated from the atoms). What should be the characteristics of an active medium? The only general requirement for an active medium is that it provides an upper energy state into which atoms can be pumped and a lower state to which they will return with the spontaneous emission of photons. The medium must also allow a population inversion between the two states. It may happen that the active species or centres, which provide lasing levels, constitute a small fraction of the medium. For example, in case of ruby, which is $Al_2O_3$ with some of the Al atoms replaced by Cr atoms, only the latter (Cr) is the active centre. Typical number of active species per cubic centimeter in solids and liquids is $10^{19}$ to $10^{20}$ and that for gaseous media their number is about $10^{15}$ to $10^{17}$. How the light beam gets amplified when it passes through an active medium? To get the answer we examine the process of population inversion now.

## Population Inversion

Why is the condition of population inversion between the lasing level necessary for operation of lasers, i.e. for amplification of light to occur? We can investigate this by calculating the change in intensity of the light beam passing through an active medium. Refer to Fig. 13.7. A collimated beam of light having intensity $I_\nu$ travels along the $x$-axis through an active medium of thickness $dx$.

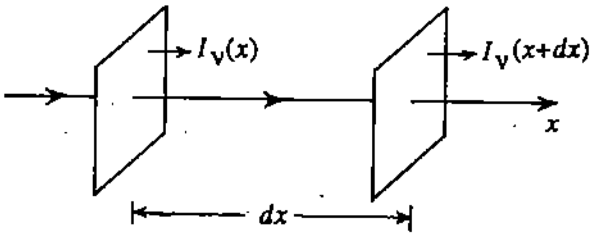


Fig. 13.7: Light beam of intensity $I_\nu$ passing through an active medium along the x-axis

If the cross-sectional area of each of the planes is $S$, volume of the layer will be $Sdx$. Let $N_1(\nu)\,d\nu$ represent the number of atoms per unit volume which are capable of absorbing radiation whose frequency lies between $\nu$ and $\nu + d\nu$. *The number of upwardly transmitted* $(E_1 \rightarrow E_2)$ atoms per unit time in the layer of volume $Sdx$ would be (refer to box on page 26)

$$N_1(\nu)\, d\nu\, B_{12}\, u(\nu)\, Sdx$$

In each transition, a photon of energy $h\nu$ is absorbed. Thus, energy lost per unit time from the incident radiation is

$$h\nu\,[\,N_1(\nu)\, d\nu\, B_{12}\, u(\nu)\,]\, Sdx$$

Similarly, let $N_2(\nu)\, d\nu$ represent the number of atoms per unit volume which are capable of undergoing stimulated emission by falling down to level $E_1$. The frequency of these photons lie between $\nu$ and $\nu + d\nu$. Then the number of stimulated photons emitted per unit time in the layer is

$$N_2(\nu)\, d\nu\, B_{21}\, u(\nu)\, Sdx$$

In each transition, photon of energy $h\nu$ is emitted and this reinforces the propagating beam. Thus the energy gain by the incident radiation per unit time is

$$h\nu\,[\,N_2(\nu)\, d\nu\, B_{21}\, u(\nu)\,]\, Sdx$$

You may have noticed that we have neglected spontaneous emission. It is so because a photon, emitted via spontaneous process, is in a random direction. And, as such, it does not contribute appreciably to the intensity of the beam.

As a result of above processes, will the intensity of the light beam increase or decrease with time? Since u ($\nu$) $d\nu\, Sdx$ represents the energy in the layer within frequency range $\nu$ and $\nu + d\nu$, we can write the rate of change of the energy with time as

$$\frac{\partial}{\partial t}(u(\nu)\, d\nu\, Sdx) = h\nu\,[\,-N_1(\nu)\, B_{12}\, u(\nu) + N_2(\nu)\, B_{21}\, u(\nu)\,]\, d\nu\, Sdx$$

or

$$\frac{\partial u(\nu)}{\partial t} = -\, h\nu\,[\,B_{12} N_1(\nu) - B_{21} N_2(\nu)\,]\, u(\nu) \qquad (13.9)$$

If $I_\nu$ represents intensity, $I_\nu\, d\nu$ signifies the energy crossing a unit area per unit time whose frequency lies between $\nu$ and $\nu + d\nu$. Then

$$[\,I_\nu(x + dx)\, d\nu - I_\nu(x)\, d\nu\,]\, S$$

denotes the rate at which the energy flows out of the layer. Since $u(\nu)\, d\nu\, Sdx$ represents radiation energy contained in the layer with frequency in the range $\nu$ and $\nu + d\nu$, we will have

$$[\,I_\nu(x + dx) - I_\nu(x)\,]\, d\nu\, S = \frac{\partial}{\partial t}[\,u(\nu)\, d\nu\, Sdx\,]$$

or

$$\frac{\partial u(\nu)}{\partial t} = \frac{I_\nu(x + dx) - I_\nu(x)}{dx} = \partial I_\nu / \partial x \qquad (13.10)$$

From Eq. (13.9) and (13.10), we have

$$\frac{\partial I_\nu}{\partial x} = -\, h\nu\,[\,B_{12} N_1(\nu) - B_{21} N_2(\nu)\,]\, u(\nu)$$

But

$$I_\nu = u(\nu)\, \nu$$

(13.11)

where $v$ = velocity of light in the active medium (= $c/n$; $n$ = refraction index of the medium). Thus, we get

$$\frac{\partial I_\nu}{\partial x} = -\frac{h\nu B}{v}(N_1 - N_2)I_\nu$$

where $B$ $(= B_{12} = B_{21})$ denotes either Einstein's coefficient. Hence

$$\frac{\partial I_\nu}{I_\nu \partial x} = -\frac{h\nu n}{c}(N_1 - N_2)B \qquad (13.12)$$

If the light beam is propagating in absorbing media, the loss of intensity, $-dI_\nu$, will be proportional to $I_\nu$ and $dx$;

$$dI_\nu = -\alpha_\nu I_\nu dx$$

where $\alpha_\nu$ is absorption coefficient. We can rewrite it as

$$\frac{\partial I_\nu}{\partial x} = -\alpha_\nu I_\nu \qquad (13.13)$$

On integration we find that

$$I_\nu = I_\nu(x = 0)\, e^{-\alpha_\nu x} \qquad (13.14)$$

If we compare Eqs. (13.12) and (13.13), we get the expression for absorption co-efficient:

$$\alpha_\nu = \frac{h\nu n}{c}(N_1 - N_2)B \qquad (13.15)$$

At thermal equilibrium, $N_1 > N_2$, that is, the population of ground state is greater than the population of the excited state and as can be seen from Eq. (13.15), $\alpha_\nu$ is positive. Positive $\alpha_\nu$ implies, (from equation 13.14) that the intensity of the beam decreases as it propagates through the material. The lost energy is used up in the excitation of atoms to higher energy states.

On the other hand, if we have a situation in which $N_2 > N_1$, $\alpha_\nu$ will be negative and intensity of the light beam would increase, that is, get amplified as it propagates through the material. This process is **light amplification**. Since this occurs when there is a higher population in excited state than in the ground (or lower energy) state, the material is said to be in the state of population inversion. Thus, the condition of population inversion is necessary for amplification of intensity of light beam.

### 13.3.2 Excitation (or Pumping)

In the previous sub-section, you have learnt about the necessity of population inversion in the active medium for obtaining laser light. The process of obtaining population inversion is known as pumping or excitation. The aim of the pumping is to see that upper energy level is more intensely populated than the lower energy level. Alternatively, we can obtain the population inversion by depopulating lower energy level (other than ground state) faster than the upper energy level. There are several ways of pumping a laser and achieving the population inversion necessary for stimulated emission to occur.

Most commonly used are the following:

1. Optical Pumping
2. Electric Discharge
3. Inelastic Atomic Collision
4. Direct Conversion

**In Optical Pumping**, a source of light is used to supply energy to the active medium. Most often this energy comes in the form of short flashes of light, a method first used in Maiman's Ruby Laser and widely used even today in **Solid-State Lasers**. The laser material is placed inside a helical xenon flash lamp of the type customary in photography. The xenon flash lamp for pumping is shown in Fig. 13.5.

Another method of pumping is by direct electron excitation as it occurs in an **electric discharge**. This method is preferred for pumping Gas lasers of which the argon laser is a good example. The electric field (typically several KV $m^{-1}$) causes electrons, emitted by the cathode, to be accelerated towards the anode. Some of the electrons will impinge on the atoms of the active medium (electron impact), and raise them to the excited state. As a result, the population inversion is achieved in the active medium.

In the **inelastic atomic collision** method of pumping, the electric discharge provides the initial excitation which raises one type of atoms to their excited state or states. These atoms subsequently collide inelastically with another type of atoms. The energy transferred inelastically raises the later type of atoms to the excited states and these are the atoms which provide the population inversion. An example is Helium-Neon Laser, to be discussed later, in which such a pumping process is employed.

A direct conversion of electrical energy into radiation occurs in light emitting diodes. Such light emitting diodes (LED) are used for pumping by **direct conversion** in semi-conductor lasers.

These are some of the processes used for pumping atoms of the active medium to achieve population inversion. Atoms (or molecules) used as active centres often exhibit rather complex system of energy levels. However, for all the variety of these structures, the actual pumping schemes may be narrowed down to a few rather simple diagrams correctly showing the pumping process. Typically, these pumping schemes involve three to four levels. We think you would like to know about them.

Let us consider some of the pumping schemes. To do so, let us identify different energy states necessary to explain the pumping scheme as: the ground state as 0; the lower lasing state as 1; the upper lasing state as 2; and the pumping state as 3. We shall indicate pumping transition by upward arrow, the lasing transition by downward arrow and non- radiative fast decay by slanted arrows. Now let us consider a **three-level** pumping scheme shown in Fig. 13.8a. Let us assume that by one of the pumping methods, more than half the number of atoms of active species have been pumped from ground state to pumping state 3. The pumped atoms in state 3 decay non-radiatively to upper lasing state 2. This decay is very fast, (life time is typically of the order to $10^{-8}$ s). The upper lasing state 2 is generally a metastable state i.e. the life time of this state ($\sim 10^{-3}$ s) is much higher than the pumping state (or the excited state). Therefore, we have a situation of population inversion between lasing states 2 and 1 and hence lasing may take place. You may note that in this pumping scheme, the ground state (0) and the lower lasing state (1) are the same state. This feature of the pumping scheme proves too demanding for the pumping process because in normal circumstances, the ground state

Atoms or molecules tend to occupy lowest energy state. Therefore, the population of the ground state (lowest energy state) is high.

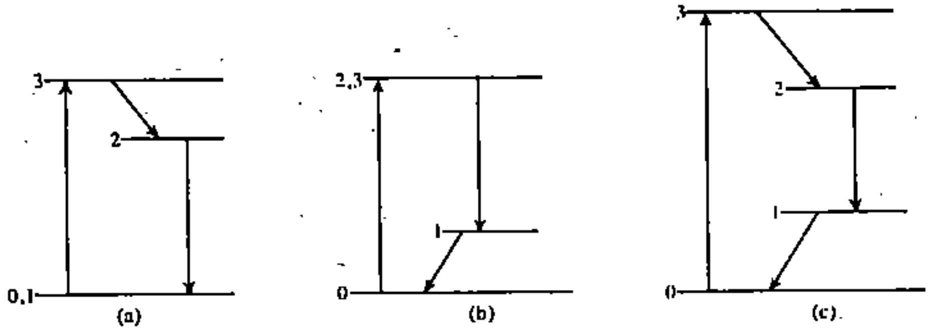


Fig.13.8: Three level pumping schemes, (a) the ground state (0) and lower lasing state (1) are the same, (b) pumping state (3) and upper lasing sitate (2) are the same. (c) Four level pumping scheme.

is highly populated. And, as you can appreciate, an ideal lower lasing state (1) should be empty or very thinly populated. How to get rid of this problem?

According to the uncertainty principle, about which you will study in the PHE-11 course on Modern Physics, an energy state with longer life time will have narrow frequency band.

This problem can be taken care of if the pumping scheme is as shown in Fig. 13.8 b. As you can see, the atoms in the lower lasing state undergo non-radiative transition to the ground state (0). Since this transition is very fast ( $\sim 10^{-8}$s) ,the lower lasing level is empty for all practical purposes. You may, however, note that the same energy state acts as pumping state (3) and the upper lasing state (2). This state of affairs has its own shortcoming. If the pumping state has to act as upper lasing state, it must have a longer life time (metastable state) which implies that it must have very narrow frequency width. On the other hand, for proper utilisation of pumping energy, this state must have a wide frequency width so that more and more atoms get accomodated there. So, you see, it is a kind of conflicting requirements put on a single energy states.

The pumping scheme free from the shortcomings mentioned above with reference to three-level pumping scheme is what we call **four-level** pumping scheme shown in Fig. 13.8c. In this case, the pumping state (3) and the upper state (2) are seperate; atoms in the pumping state undergo non-radiative transition to the upper lasing state. The four level pumping scheme, however, has some limitations. Substantial energy is lost during non-radiative transitions between pumping state (3) and the upper lasing state (2) and between the lower lasing state (1) and the ground state (0).

You may now ask: Which pumping scheme is better and preferred? Each pumping scheme has its own advantages and disadvantages. The choice of the pumping scheme in designing a laser depends upon the active media, the kind of use we want to put the laser light to, etc. We will discuss these aspects in the following sections. You may now like to answer an SAQ

**SAQ 2**

*Spend 2 min*

If laser action occurs by the transition from an excited state to the ground state and it produces light of 693nm wavelength, what is the energy of the excited state. Take the energy of the ground state to be zero.

### 13.3.3 Feedback Mechanism: Optical Resonant Cavity

On the basis of the discussion in the previous sections, you now know that when a state of population inversion exists in an active medium, a light beam of particular frequency passing through it would get amplified. It happens because in such a situation, stimulated emission dominates spontaneous emission. This is the basic principle of optical amplifier. But a laser is much more than a simple optical amplifier. The laser, which produces a highly coherent beam of light, does not include a coherent light beam to initiate stimulated emission. Instead, it is the spontaneously emitted photon from upper lasing state which stimulates the emission of new photons. Each spontaneous photon can initiate many other stimulated transitions which, in turn, may cause light amplification. Well, in this way, we do get amplification of light by stimulated emission. But, how is coherence of this amplified light ascertained? In other words, how can we ensure that the laser light has a very narrow band width (monochromaticity) and a high degree of phase correlation? As such, the amplified light from laser is not coherent. It is because the spontaneous photons are independent of each other and travel in different directions. Therefore, the corresponding stimulated photons will also travel in different directions.

Can you suggest as to what should we do for obtaining a highly coherent laser beam? For obtaining a coherent light beam, we need to have a mechanism by which a condition is created such that spontaneous emission only in certain selected direction can develop stumulated emission. This mechanism is known as feedback mechanism. The spontaneous photons emitted in other directions leave the active medium without initiating much stimulated emission.

Now, you may ask; how do we actually achieve this favourable condition for spontaneously emitted photons in some preferred direction to further stimulate emission? Well, this is accomplished by means of an **optical resonator**-an essential component of a laser. Let us understand how an optical resonator works.

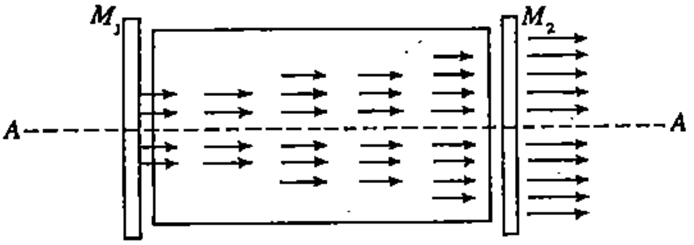


Fig.13.9: Optical resonator consisting of two mirrors $M_1$ and $M_2$; $M_1$ is totally reflecting whereas $M_2$ is semitransparent; the axis of the mirrors is aligned with that of the active material.

Optical cavity resonator can have many configurations. The schematic arrangement of a simple resonator is shown in Fig. 13.9. It consists of a pair of plane mirrors, $M_1$ and $M_2$, set on an optic axis which defines the direction of the laser beam. The active material is placed in between these mirrors. The photons emitted spontaneously along the $AA$ direction or sufficiently close to it travel a relatively longer distance within the active material. It is so because photons travelling along $AA$ will be reflected back and forth by the mirrors $M_1$ and $M_2$. You may notice that the direction of travel of these photons is quite fixed. Now, as a result of spending more time in the active material, these spontaneous photons will interact with more and more atoms in upper lasing level. Thus, the stimulated emission will add identical photons in the same direction, providing an ever-increasing population of coherent photons that bounce back and forth between the mirrors. On the other hand, spontaneous photons and the corresponding stimulated emission in other directions will traverse relatively shorter distances (and hence spend lesser time) in the active medium. Hence they will soon die out. Thus the optical resonant cavity provides the desired selectivity of propagation direction and thereby ensures the spatial coherence of the laser beam.

You may recall that the spatial coherence is a measure of the uniformity of the phase across the optical wavefront. And the temporal coherence is a measure of the monochromaticity of the light.

Now, what about monochromaticity of the laser light? Well, the laser light is highly monochromatic due to very nature of its origin - the stimulated emission. It is so because the spontaneously emitted photons whose frequency do not match with the frequency difference between lasing levels will not give rise to stimulated emission. Thus, the band of wavelengths emitted during spontaneous emission is narrowed down. The monochromaticity of the laser light can further be enhanced by the optical resonant cavity. Suppose there are more than one upper lasing levels in a particular active medium. In that case, the laser output will consist radiations of more than one frequency. Now, if the mirrors of the resonant cavity are such that their reflectivity is a function of frequency, the radiations due to undesired lasing between levels will be damped out. Therefore, resonant cavity is the most vital component of the laser to obtain highly coherent light beam as output.

In this section, you learnt basic constituents of a laser. Since the invention of ruby laser by Maiman in 1960, the research and development in this field has produced a variety of lasers. It is not possible to discuss all of them in detail here. However, we will discuss some of them now.

## 13.4 TYPES OF LASERS

As such, lasers can be classified in a variety of ways. One of these is in terms of their active media. As mentioned earlier, materials in all the three states of matter, namely, solid, liquid and gas, have been used as active medium to produce laser beam. Further, lasers have also been constructed using semi- conductors and plasma as active medium. In the following, let us know about some of them with particular reference to the physical properties of the active medium and the pumping methods employed.

### 13.4.1 Solid State Lasers

These lasers use an active material which is essentially an insulator doped with ions of impurity in the host structure. These lasers invariably use optical pumping to obtain the condition of population inversion. The sources for optical pumping may be discharge flashtubes, continuously operating lamps or even an auxiliary laser. The active centres in these lasers are transition element ions doped in the dielectic crystal. The host material for these active centres are generally oxide crystals. The most popular type of solid-state lasers are the **ruby laser** and **Nd:YAG** (neodymium: yttrium, aluminium, garnet) **laser**. Ruby is $Al_2O_3$ crystal (corundum) doped with triply ionized chromium atom ($Cr^{3+}$). You have learnt the functioning of this laser in section 13.2.

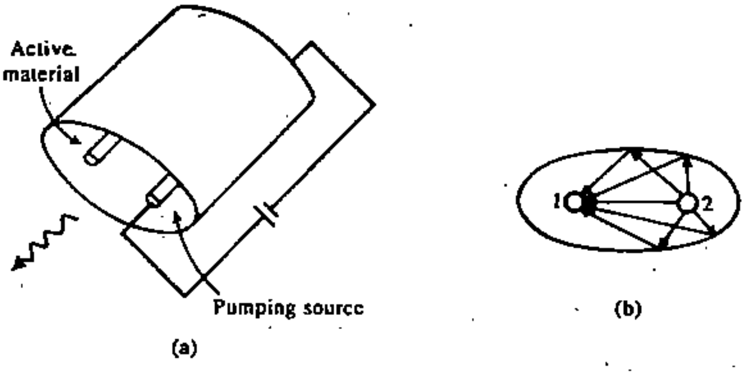


**Fig. 13.10: Pumping arrangement for solid-state lasers.**

In solid-state lasers, the optical pumping is done by placing the active material (in the form of rod) at one focus and the pumping source (in the shape of a right cylinder) at another focus of an elliptical reflector as shown in Fig. 13.10a. The advantage of such an arrangement is that any light leaving one focus of the ellipse will pass through the other focus after reflection from the silvered surface of the pump cavity. All of the pump radiation, therefore, is maximally focussed on the active material, as shown in Fig. 13.10b.

### The Nd: YAG Laser

This laser, unlike ruby laser, employs four level pumping scheme. The energy levels of the neodymium (the active material) is shown in Fig. 13.11. In order to keep the discussion simple, we have not used the spectroscopic notations for different energy levels in Fig. 13.11. Rather, energy levels have been marked $E_0$, $E_1$, and so on. The optical pumping raises the Nd atoms in the ground state ($E_0$) to a few excited states ($E_7$, $E_8$). The energy levels marked $E_4$ and $E_1$ are the lasing levels. The pumped atoms in the excited states undergo non-radiative transition to the upper lasing level, $E_4$. Out of the group of lower lasing levels, the major portion of energy is emitted in the transition $E_4 \rightarrow E_1$. The Nd: YAG laser is an example of four-level laser.

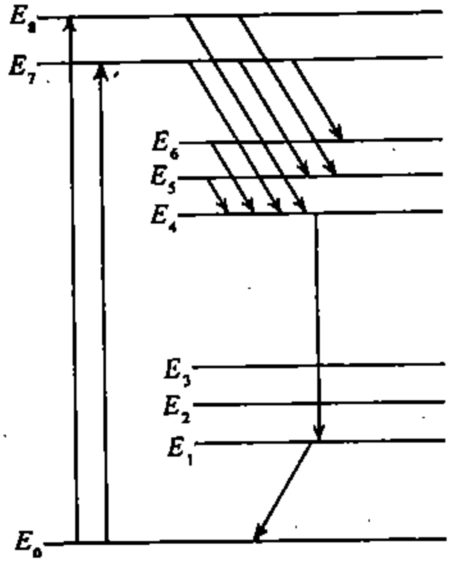


Fig. 13.11: Energy level diagram of Nd (neodymium) ion in Nd: YAG.

This solid-state laser has two advantages: (a) it has a low excitation threshold and (b) has a high thermal conductivity. Due to high thermal conductivity, it can be used for generating light pulses at a high repetition rate or for continuous operation.

## 13.4.2 Liquid Lasers

In this class of lasers, as the name indicates, the active media are either the liquid solutions of organic dyes or specially prepared liquids doped with rare-earth ions viz $Nd^{3+}$. However, majority of liquid lasers use a solution of an organic dye as active medium and hence are also called **organic dye lasers**. Solvents used for the purpose are water, methanol, benzen, aceton etc. The liquid lasers are optically pumped. The energy states taking part in the lasing transition are the different vibrational energy states of different electronic energy states of the dye molecule. Since you may not be familiar with the vibrational energy states of molecules, we do not discuss the pumping scheme of this class of lasers.

In contrast to solids, liquid do not crack or shatter and can be made in sizes almost unlimited. Another advantage of liquid lasers is due to their (that of organic dyes) wide obsorption bands in the visible and near ultravoilet portion of the electromagnetic spectrum. Therefore, liquid lasers are an ideal candidate for **tunable laser** i.e. the frequency and hence energy of the output laser beam can be selected with ease.

### 13.4.3 Gas Lasers

The attractive feature of gas lasers in which rarified gases are the active media, is that they can be designed to produce output beams over a wide range of wavelengths. Except for the cesium- vapour laser, gas lasers are pumped electrically rather than optically. Can you say why? It is because the condition for amplification by stimulated emission, at one wavelength or another, are satisfied by an electrical discharge through almost any gas. Another reason for employing electrical pumping for gas lasers is that, unlike solids and liquids, the absorption lines of active centres in gaseous media exhibit substantially narrow widths. Therefore, optical pumping would prove very inefficient for gas lasers because the pump radiation obtained from optical sources do not have line spectrum of very narrow lines. In other words, the energy of optical pump radiation has a considerable spread in its value and since the gaseous active media will absorb radiation of almost single energy, most of the pump energy will go waste. Hence, optical pumping is not used for gas lasers. Further, gas lasers have advantage over solid state and liquid lasers in that they are free from local irregularities. Most gaseous systems have a high degree of optical perfection simply because the density of the gas is uniform.

We will now briefly describe a typical gas laser-the Helium-Neon gas laser. This was the first gas laser operated successfully.

#### The Helium-Neon Laser

In the helium-neon laser, a mixture of helium (He) and neon (Ne) gases is used as active medium. **Lasing levels are provided by the exited states of the Ne atoms,** whereas the He atoms play an important role in pumping Ne atoms to the excited states. The He-Ne

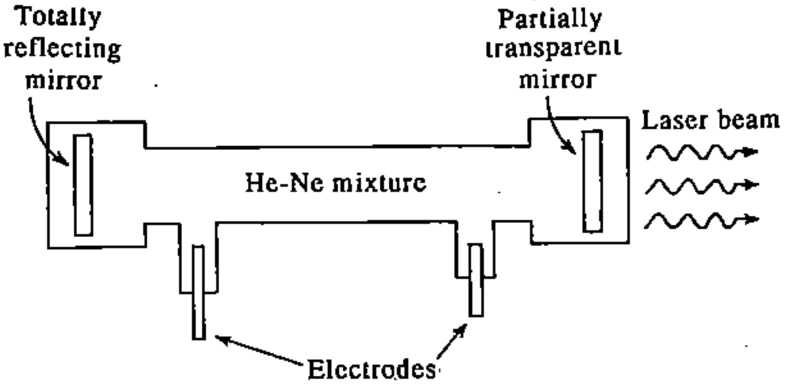


Fig.13.12: The He-Ne Laser

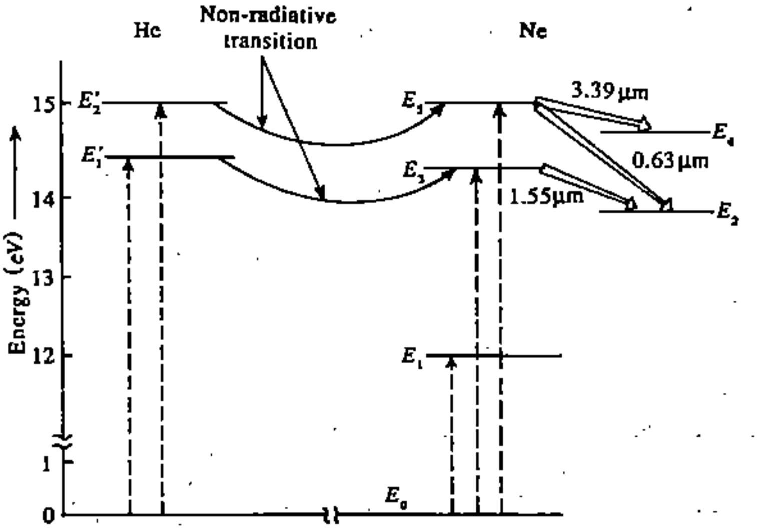


Fig.13.13: Energy level diagram of helium-neon laser. Arrows (⇒) indicate the lasing transition

laser is shown schematically in Fig. 13.12. The pumping is done by a stationary glow discharge fired by a direct current. When the potential difference between the anode and cathode is about 1000V, a glow discharge is initiated in the working capillary (containing He-Ne mixture) of a few millimeter diameter.

Now, let us look at the pumping scheme of the He-Ne laser. Refer to Fig. 13.13 which shows the energy level diagram of He-Ne laser. When free electrons produced during the gas discharge pass through the He-Ne mixture, they collide with the He and Ne atoms and excite them by impact energy transfer. Such absorptive transitions due to electron impacts are shown by dashed arrows in Fig. 13.13. These excited states of He ( i. e. $E_1'$ and $E_2'$ ) are metastable. Thus, He-atoms excited to these states stay there for a long time before losing energy by collision. The interesting feature of the He-Ne energy diagram is that the excited states of Ne, namely $E_3$ and $E_5$ have approximately same energy as that of $E_1'$ and $E_2'$ of He atom. Therefore, when He-atoms in $E_1'$ and $E_2'$ collide with Ne-atoms in ground state, the He-atoms transfer their energy to Ne-atoms and raise them to the states $E_3$ and $E_5$. Such an exchange of energy is known as **resonant collision energy transfer.** Due to this energy transfer, He-atoms fall back to ground state. As a result, the excited states $E_3$ and $E_5$ of Ne-atoms have a sizable population which is much more than that of states $E_2$ and $E_4$. Thus a condition of population inversion is achieved between the upper lasing levels $E_5$ (or $E_3$) and lower lasing levels $E_4(E_2)$. In such a situation, any spontaneously emitted photon can trigger laser action between these levels. The Ne atoms then drop down from the lower lasing levels $E_2$ and $E_4$, to the level $E_1$ through spontaneous emission.

The wavelength of transition between levels $E_5 \rightarrow E_4, E_5 \rightarrow E_2, E_3 \rightarrow E_2$ are 3.39 $\mu$m, 0.63 $\mu$m and 1.55 $\mu$m respectively. As you can easily make out, radiations corresponding to 3.39 $\mu$m and 1.55 $\mu$m fall in the infrared region of the electromagnetic spectrum. The radiation corresponding to 0.63 $\mu$m, however, gives the red light - characteristic light of He-Ne laser. Proper selection of different frequencies may be made by choosing end mirrors of the resonant cavity which has high reflectivity over only the desired wavelength range.

Before we conclude our discussion about types of lasers, you must know that apart from those mentioned above, there are many other types of gas lasers. We may particularly mention **molecular lasers** (carbon dioxide laser), **chemical lasers, plasma lasers, semiconductor lasers,** etc. We have not discussed these here since for an understanding of their pumping schemes, you need to know molecular spectroscopy, semiconductor physics etc. It is, however, worth mentioning here that the essential principles, in so far as laser action is concerned, remain the same in all types of lasers.

The importance of lasers in contemporary physics lies in their so many and so varied applications. To give you a glimse of these we now discuss some of the important applications of lasers.

## 13.5 APPLICATIONS OF LASERS

Applications of any device essentially stem from its unique features. What are the unique features of a laser? First and the formost, laser light is highly coherent. This characteristic has enabled us to use lasers for data transmission and processing, precision measurements, photography (holography), etc. Secondly, laser light has unprecedented brightness (energy per unit area). Brightness of laser light, a by-product of its coherence, can be many orders of magnitude greater than the brightest of the light produced by conventional sources. Further, laser beams are highly directional.

In a typical laser, this directionality is limited only by the diffraction of the emerging beam by the laser aperture itself. The brightness and directionality of laser beam are exploited to produce targetted effects in materials. These applications include material working (such as heat treatment, welding, cutting, hole burning etc.), isotope seperation, medical diagnostics, etc. In the following, you will learn some of these applications of lasers.

### 13.5.1 Communication

You may be aware that in a typical communication system, information is communicated (between the transmitter and the receiver) through electromagnetic waves, which are known as carrier wave. These are modulated by the desired signal (the oscillations of the information proper). Normally the signal frequency is appreciably lower than the frequency of the carrier wave. Moreover, higher the carrier frequency, wider frequency range it can modulate. In other words, the capacity of a communication channel is proportional to the frequency of the carrier wave. The frequency in the centre of the visible spectrum is about 100,000 times greater than the frequency of 6 cm waves used in microwave-radio relay systems. Consequently, the theoretical information capacity of a typical light wave is about 100,000 times greater than that of a typical microwave.

Long distance communication systems rely on the principle of multiplexing-the simultaneous transmission of many different messages (information) over the same pathway. The ordinary human voice (conversation) requires a frequency band from 200 to 4000 Hz, a band 3800 Hz wide. A telephone call, therefore, can be transmitted on any band that is 3800 Hz wide. It can be carried by a coaxial cable in the frequency band between 1,000,200 and 1,004,000 Hz, in the MHz range, or a He-Ne laser beam (638.8 nm, $4.738 \times 10^{14}$Hz) in the frequency range between 473,800,000,000,200 and 473,800,000,004,000 Hz. You may note here that the telephone message requires about 0.4 percent of the available co-axial carrier frequency. And, the same telephone message requires less than **one-billionth** of 1 percent of the available laser-beam frequency. Thus, the information carrying capacity could be enhanced tremendously if laser beams are employed as carriers. So, wait for some more time till laser trunk lines come into use in a big way and you may be saved from listening *"All the lines in this route are busy. Please dial after some time"*!!

Now you may ask: Light, as such, was available to us from time immemorial, then why is it that we are using (or planning to use!) it for communication purposes now? Is it related to the discovery of a laser in any way? Yes, it is. As we mentioned earlier, light from conventional sources may not be pure (that is, it may be non-monochromatic) and hence cannot be used for transmitting signals. Radio waves from an electromagnetic oscillator are confined to fairly narrow region of electromagnetic spectrum (i.e. it has a well defined frequency). These radio waves are, therefore, free from "noise" (considerable spread in frequency values) and hence can be used for carrying a signal. In contrast, all conventional light sources are essentially 'noise' generators i.e. they simultaneously emit electromagnetic radiations of different frequencies and hence are not suitable as carrier waves. With the invention of lasers, however, the situation changed. As you know, the light produced by lasers is highly monochromatic and coherent which enable them to act as carrier waves in the communication systems.

Now, what is the medium through which laser beam travels while it carries information? The signal carrying laser beams can be transmitted through free (unguided) space, and by light guides. Light guides in the form optical fibres have found wide use in optical communication. You will learn about the details of fibre optics in Unit 15 of this course.

### 13.5.2 Basic Research

The discovery of a laser gave birth to an entirely new branch of optics known as **nonlinear optics**. Even at ordinary laser intensities, transparent materials (which are usually nonconductors), respond in an unusual manner. You may recall, for example, that the dielectric constant of material depends on its nature as well as on the frequency of the light passing through it. But, it has been observed that when the ordinary light beam is replaced by a laser beam, the dielectric constant also depends on the instantaneous magnitude of the electric field component of the laser beam. In other words, the response of a material to high electric fields is non-linear. It is just one of the several non linear effects that a laser beam produces when it interacts with matter. In fact, almost all the laws of optics are modified to some extent at high intensities produced by pulsed lasers.

Another important application of lasers in basic research and development is in the field of **thermonuclear fusion**. As you know, for effective fusion to take place, extremely high temperature ( ~ $10^8$K) must be maintained. In principle, such high temperatures can be achieved by powerful laser beams.

Yet another remarkable application of lasers is in **isotope separation**. You may recall that one of the basic requirements of harnessing nuclear energy from uranium is to have 2-3% of uranium isotope ($^{235}$U) in the fuel. In natural uranium, however, the percentage of $^{235}$U is only 0.7. (The major constituent of natural uranium is $^{238}$U.) Therefore, to have fuel enriched in $^{235}$U, we can use laser beams. Each of these isotopes absorbs radiation of different frequency. So when a laser beam of particular frequency is passed through the mixture of $^{235}$U and $^{238}$U , the atoms of $^{235}$U absorb the radiation and get excited. The excited atoms of the desired isotope are further excited so that they get ionized. Once ionised, it can easily be separated by applying a dc electric field. This is one of the several methods of using laser beam for isotope seperation.

### 13.5.3 Medicine

A properly focussed laser beam, is an excellent tool for surgery. The advantage of laser surgery is that it is bloodless since the beam not only cuts, it also "welds" blood vessels. It has a high sterility as no contact of tissues with surgical tools takes place. Also, the laser surgery is painless and operations are very fast. In fact there is not enough time for the patient to respond to the incision and sense pain. Laser beams are being widely used for performing eye and stone surgery.

A word of caution. As such, any light can cause damage. Laser, in particular, can be highly damaging because it has spatial coherence, i.e., it can be focussed down to a high power densities. The **maximum permissible exposure** (MPE) is 0.0005 mJ cm$^{-2}$. For exposure time from $2 \times 10^{-5}$s to 10 s, the limit is MPE = $1.8t^{3/4}$mJ cm$^{-2}$.

### 13.5.4 Industry

Invention of lasers has made it possible to develop sophisticated tools of material working (such as drilling, welding, etc) processes used in industry. With appropriate choice of lasers, a laser beam can be focussed into a light spot of diameter 10-100$\mu$m! Can you imagine this dimension-it will be smaller than the dot you mark with your pen on a piece of paper! Due to this sharp focussing, a very high concentration of energy is available within a small spot on the surface of the material. For example, when a 1kW output of a continuous wave (cw) laser is focussed a spot of 100$\mu$ m diameter, the resultant irradiance (intensity) will be 10 W cm . This makes laser an effective tool for drilling very fine hole through the materials.

Laser **cutting**, as compared to other cutting processes, offers several advantages e.g. possibility of fine and precise cuts, minimal amount of mechanical distortion and thermal damage introduced in the material being cut, chemical purity of the cutting process, etc. Laser cutting is extensively used in industry. For example, in high-tech garment factories, $CO_2$ laser capable of 100W of continuous output is used for cutting cloth. The laser cuts 1m cloth in a second! And, laser cutting is also employed in the fabrication of spacecraft to cut the sheets of titanium, steel and aluminium. In cutting and most of the industrial applications, carbon-dioxide ($CO_2$) laser is used.

### 13.5.5 Environmental Measurements

You may be aware of the conventional technique of determining the concentration of various atmospheric pollutants such as gases (carbon monoxide, sulphur dioxide, oxides of nitrogen, etc) and a variety of material particles (dust, smoke, flyash etc). In this method, the nature and concentration of pollutants is determined by chemical analysis. The major deficiency of this method is that it does not provide real-time data. The technique developed with lasers for measuring the concentration of pollutants is essentially the 'remote-sensing' technique which does not require sample to be analysed in laboratory. Since it provides information about the change in atmospheric composition with time, it can serve well for monitoring the environmental pollution.

For determination of pollutants in the form of material particles, the technique is based on the scattering of light. The technique is known as LIDAR (*l*ight *d*etection *a*nd *r*anging) and its operations are similar to those of a radar. In brief, a pulsed laser is passed through the location under investigation and the back scattered light in detected by a photodetector. The time taken by the back scattered light to be detected gives information about the concentration of pollutant matter.

For the determination of gaseous pollutants, the basic principle involved is the absorption of light by the gaseous atoms or molecules. As different gas absorbs at different wavelengths, passing laser beams of different wavelengths provides information about the gaseous constituents of the environment.

### 13.5.6 Photography: Holography

The conventional photographic process, as you know, consists of recording an illuminated three-dimensional object or scene as a two-dimensional image on a photosensitive surface. The light reflected from the object is focussed on the photo sensitive surface by some kind of image forming device, which can be a complex series of lenses or simply a pin-hole in an opaque screen.

The coherent nature of the laser beam has brought about a qualitatively new method of photography without lens system. This new method, called holography, allows three-dimensional (that is, complete), pictures of a given object or a scene to be taken. Holography (also known as photography by wave-front reconstruction) does not, as such, record an image of the object being photographed; rather, it records the reflected light waves themselves. The photographic record so obtained is called hologram. The hologram bears no resemblence to the original object. It, however, contains - in a kind of optical code - all the information about the object that would be contained in an ordinary photograph. In addition, the hologram also contains information about the object that cannot be recorded by any other photographic process. Holography is the subject matter of the next unit (i.e. Unit 14).

## 13.6 SUMMARY

- According to the Bohr's theory, if an atom makes a transition from an excited state (of energy $E_i$) to a state of lower energy $E_f$, emission of electromagnetic radiation (photons) take place. The energy of the emitted photon is

$$h\nu = E_i - E_f$$

- When electromagnetic radiation interacts with matter, three type of processes may occur

  (i) Spontaneous Emission

  (ii) Absorption

  (iii) Stimulated Emission

- Light emitted by ordinary sources is due to spontaneous emission.
- The existence of stimulated emission of radiation was predicted by Einstein on the basis of thermodynamic considerations. If the population of the energy level $E_1$ be $N_1$ and that of $E_2$ be $N_2$ ($E_1 < E_2$), then, the ratio of the population of the two states is given as

$$\frac{N_2}{N_1} = \frac{B_{12}u(\nu)}{A_{21} + B_{21}u(\nu)}$$

  where, $u(\nu)$ is energy density of radiation at frequency $\nu$ and $B_{12}$, $B_{21}$ and $A_{21}$ are Einstein co-efficients.

- Einstein coefficients are related to each other through the relations

$$B_{21} = B_{12}$$

$$\frac{A_{21}}{B_{21}} = \frac{8\pi h\nu^3}{c^3}$$

- Einstein's relation clearly indicates that stimulated emission may dominate spontaneous emission provided the condition of population inversion exists. And in a atomic system where a condition of population inversion exists, one may have amplification of light, that is, laser light.
- Einstein's prediction was first realised in the optical frequency range by Maiman who developed a laser using a ruby rod.
- There are three prerequisits for laser operation:

  (i) Active medium

  (ii) Pumping

  (iii) Optical resonant cavity

- The change in intensity of a light beam passing through an active medium is given by

$$\frac{\partial I_\nu}{\partial x} = -I_\nu \frac{h\nu n}{c}(N_1 - N_2)B$$

where $n$ is refractive index

$B$ is Einstein's coefficient.

This relation clearly indicates that for enhancement in the intensity of the light beam as it traverses the active medium, $N_2 > N_1$, i.e. a condition of population inversion must exist.

- There are variety of methods for pumping, such as, optical pumping, electronic discharge, inelastic atomic collisions etc. The choice of pumping process mainly depends upon the nature of the active medium.
- There are two types of pumping schemes:three level and four-level.
- Optical resonant cavity helps in obtaining sustained laser light.

## 13.7 TERMINAL QUESTIONS

1. Assume that an atom has two energy levels seperated by an energy corresponding to a frequency 4.7 × $10^{14}$Hz, as in the He-Ne laser. Let us assume that all the atoms are located in one or the other of these two states. Calculate the fraction of atoms in the upper state at room temperature T = 300K.

2. A pulsed laser used for welding produces 100 W of power during 10 m. Calculate the energy delivered to the weld.

## 13.8 SOLUTIONS AND ANSWERS

SAQs

1. The ratio of the number of spontaneous to stimulated emission is given as

$$\frac{A_{21}}{B_{21}\,u\,(\nu)} = e^{h\,\nu/k_B\,T} - 1$$

The absolute temperature of an ordinary source of light has been given as

$$T = 10^3\text{K}$$

Let us take the wavelength of light, $\lambda = 6000$ Å. Hence the corresponding frequency,

$$\nu = \frac{c}{\lambda} = \frac{3 \times 10^8\,\text{ms}^{-1}}{6000 \times 10^{-10}\text{m}} = 0.5 \times 10^{15}\ \text{Hz}$$

Planck's constant $h = 6.6 \times 10^{-34}$ J s

Boltzmann constant $k_B = 1.38 \times 10^{-23}$ JK$^{-1}$

Hence,

$$\frac{A_{21}}{B_{21}\,u\,(\nu)} = \exp\left[\frac{6.6 \times 10^{-34}\,(\text{J . s}) \times 0.5 \times 10^{15}\,(\text{s}^{-1})}{1.38 \times 10^{-23}\,(\text{JK}^{-1}) \times 10^3\,(\text{K})}\right] - 1$$

$$= \exp[23] - 1$$

$$= 10^{10}$$

Thus, for ordinary sources of light, the number of spontaneous emission is much, much greater than the number of stimulated emission.

2. Let the energy of the excited state (upper lasing state) be $E_2$ and that of the ground state (lower lasing state) be $E_1$. The laser light is due to the atomic transitions from $E_2$ to $E_1$. Thus, the frequency of the laser light will be

$$\nu = \frac{E_2 - E_1}{h}$$

Now, as per the given problem,

$$E_2 = ?\ E_1 = 0 \text{ and } \lambda = 693\text{nm} = 693 \times 10^{-9}\text{m}$$

Hence,

$$\nu = \frac{c}{\lambda} = \frac{3 \times 10^{8}(\text{m/s})}{693 \times 10^{-9}(\text{m})} = 3.1 \times 10^{14}\text{s}^{-1}$$

$$E_2 - E_1 = h\nu$$

$$E_2 = 6.6 \times 10^{-34}(\text{J.s}) \times 3.1 \times 10^{14}(\text{s}^{-1})$$

$$= 20.46 \times 10^{-20}\text{J}$$

$$= 12.77\,\text{eV}$$

**TQs**

1. Let the two energy levels be $E_1$ and $E_2$ (such that $E_1 < E_2$) and their population be $N_1$ and $N_2$ respectively. According to the Boltzmann distribution

$$\frac{N_2}{N_1} = e^{-(E_2 - E_1)/k_B T}$$

We know that

$$(E_2 - E_1) = h\nu$$

$$= 6.62 \times 10^{-34}(\text{J.s}) \times 4.7 \times 10^{14}(\text{s}^{-1})$$

$$= 31.114 \times 10^{-20}\text{J}$$

and

$$k_B T = 1.38 \times 10^{-23}(\text{J/K}) \times 300(\text{K})$$

$$= 4.14 \times 10^{-21}\text{J}$$

Hence,

$$\frac{N_2}{N_1} = e^{-h\nu/k_B T}$$

$$= e^{-\left[\frac{31.114 \times 10^{-20}}{4.14 \times 10^{-21}}\right]}$$

$$= e^{-(75.1)}$$

$$= 2.29$$

2. Power = Energy per unit time

$$= \frac{\text{Energy}}{\text{Time}}$$

Given, Power = 100W = 100 (J/s)

Time = 10ms = $10 \times 10^{-3}$ (s)

$\therefore$ Energy = Power × Time

$$= 100\,(\text{J/s}) \times 10 \times 10^{-3}\,(\text{s})$$

$$= 1\text{J.}$$

$\therefore$ Energy delivered to the weld is 1 joule.

# UNIT 14 HOLOGRAPHY

## Structure

## 14.1 INTRODUCTION

In the previous Unit, we pointed out that one of the revolutionary applications of lasers is in the development of a novel technique of photography, known as holography. This word is the combination of two Greek words - **holos** (complete) and **graphos** (writing). That is, holography is the technique of obtaining complete picture (as true as the object itself) of an object or a scene. In other words, it is a three- dimensional recording of an object or a scene. Well, you may be wondering as to what essentially differentiates this technique from the normal photography! In normal photography, a two-dimensional image of a three-dimensional object is recorded on a photosensitive surface. The photosensitive surface records the intensity distribution of light falling on it after reflection from the object. As a consequence, we obtain a permanent record of the intensity distribution that existed at the plane occupied by the photographic plate when it was exposed. Since the photosensitive surface is sensitive only to the intensity variation, the phase distribution existing in the plane of the photographic plate is completely lost and is responsible for the absence of the three-dimensional character in it. Holography is that technique of photography where not only the amplitude (and hence the intensity) but also the phase distribution can be recorded. As a result, pictures obtained by holographic technique possess three-dimensional form and are visually rich.

Holography was introduced by Dennis Gabor in 1948. He showed that one could indeed record both the amplitude and the phase of a wave by using interferometric principles. In Sec. 14.2, you will learn the basic concepts involved in the holographic technique. You will be able to appreciate the similarity between the hologram and the diffraction grating. The process of holography i.e. how to obtain a hologram, how to obtain images from the hologram etc. has been explained in Sec. 14.3. Due to high cost of lasers, (an essential requirement for holography) this technique is not being used extensively. The technique, however, has tremendous potential and some of the important applications have been explained in Sec. 14.4.

### Objectives

After going through this unit, you should be able to

- differentiate between normal photography and holography
- explain the basic principle of holography
- describe how holograms are obtained, and
- state some of the applications of holography.

## 14.2 HOLOGRAPHY: THE BASIC PRINCIPLE

Holography is the process of recording the interference pattern produced by light waves reflected by an object and reference waves. This interference pattern of the object is unique and is called **hologram** (total recording). If you look at a hologram, you will realise that it does not even remotely resemble the object. However, when this recorded pattern is illuminated by a suitably chosen **reconstruction wave**, out of the many component waves emerging from the hologram, one wave completely resembles the object wave in both amplitude and phase. Thus, when you look at this wave, you perceive the object still being in position even though the object may not be present there. Since during reconstruction (that is, image production), the object wave itself is emerging from the hologram, the image has all the effects of three-dimensionality. You can indeed shift your viewing position and "look behind" the objects.

Reference wave is the light wave falling directly on the photosensitive plate.

Object wave is the light wave reflected from the object and received at the photosensitive surface at the time of recording the hologram.

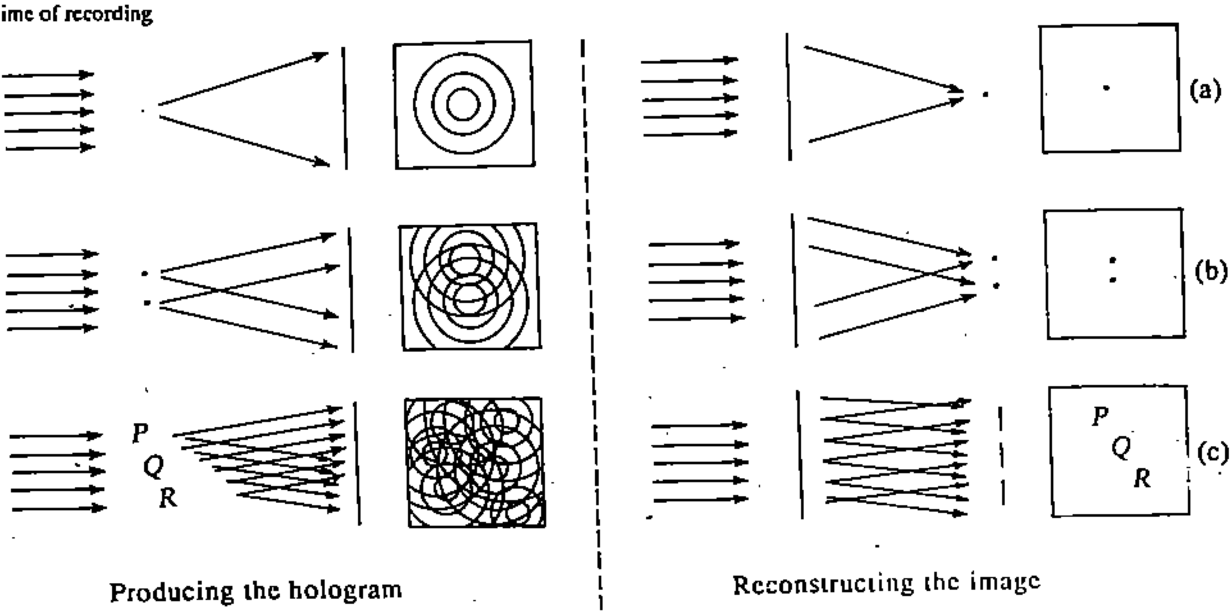


**Fig. 14.1: The principle of holography: (a) Point object forming concentric diffraction rings as in zone plate; reconstruction of zone plate gives point image (top right). For two points and a more complex object, these features are shown in (b) and (c) respectively.**

Let us understand the basic concept involved in holography with the help of a simple example. Incident light, shown in Fig. 14.1 (a), is diffracted by a point object. It gives rise to a series of bright and dark concentric rings. The pattern is recorded photographically and made into a transparency. This pattern, called a **Gabor zone plate,** is similar to a Fresnel zone plate. In the second step (top right) light is incident on the ring pattern (i.e. the Gabor zone plate) and focussed by it into a point, as focussed by a zone-plate.

Now, refer to Fig. 14.1(b) in which the object consists of two points (pixels). The diffraction pattern then consists of two sets of concentric rings. When the pattern is illuminated, each of the two sets focus, and the image consists of two points. As the object is an aggregate of many pixels, its diffraction pattern is shown in Fig. 14.1(c). The intermediate recording is a continuum of superposed zone plates- an unrecognizable multiplicity of lines and rings. Each pixel in the object forms its own set of fringes. Within each set, the light interfers but between sets, there is no fixed phase relationship and hence no interference. In order to make the different signals compatible in phase, another wave called reference is added. Refer to Fig. 14.2 where the effect of adding a sufficiently strong reference beam to the random - phase signal is shown. As a result, the phase of the resultant of reference and the signal becomes similar to that of the reference alone. Thus contributions from different pixels produce an interference fringe pattern.

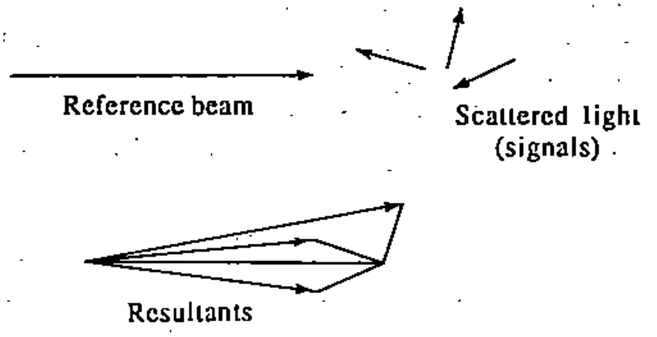


Fig. 14.2: Addition of a strong coherent reference beam (top left) with random-phase signals (top right) gives similar resultant (bottom).

The essence of holography is that the process of image formation is being interrupted and splitted into two. In the first step, the object is transformed into a photographic record, called the hologram and in the second step called reconstruction, the hologram is transformed into image. No lens is needed in either step. You may now like to answer an SAQ.

SAQ 1

*Spend 3 min*

Using the size of the amplitude vectors drawn in Fig. 14.2 calculate

(a) the ratio of intensities, and (b) the contrast resulting from these intensities.

At this stage, you may say that in photography, what we essentially record is the light reflected from the object and not its diffraction pattern. Well, it is easy to extend the basic idea of holography, explained above in terms of Gabor's zone- plate, to the actual photography situations. Reflected light waves, like other waves, are described by their amplitude (or intensity) and their phase (or frequency). To capture the wave pattern completely (that is, to obtain the hologram) both the amplitude and the phase of the wave must be recorded at each point on the recording surface. As you are aware, recording of the amplitude portion of the wave is achieved in normal photography by converting it to corresponding variation in the opacity of the photographic emulsion. The photographic emulsion is, however, insensitive to phase relations. In holography (also known as **wave-front reconstruction**), the phase relations are rendered visible to the photographic plate through the technique of interferometry. You may recall from block-2 of this course that interferometry converts phase relations into corresponding amplitude relations.

When two plane waves derived from a common source impinge at different angles on a screen, they produce a set of uniform, parallel interference fringes. The spacing of the fringes depends solely on the angle between the impinging waves (that is, on the path difference between them). A photographic recording of such a fringe pattern results in a grating-like structure. In case of holography, one of these waves is the one reflected from the object (called the object wave) and hence need not be a plane wave. The wavefront of the reflected wave will be highly irregular because of the unevenness of the object surface. When this irregular reflected wave pattern interferes with the reference wave, the resulting interference pattern will not be uniform. Rather, it will have irregular interference pattern– the irregularity of the impinging wave fronts. At places where the signal bearing waves (**the object wave**) have maximum amplitude, the interference fringes have the greatest contrast and vice-versa. Thus, variations in the amplitude of the object wave menifest as the variation in contrast of the recorded fringe pattern. Can you recall the implications of the spacing of the interference fringes? It is related to the path difference (and hence the phase difference) between the two interfering waves. And the path difference, in turn, depends on the angle between them. Larger the angle between the two interfering waves, more closely spaced will be the fringes and vice-versa. Therefore, variations in the phase of the object wave menifest as the variations in the spacing of the fringes on the photographic record (the hologram). Thus, in a hologram, both the amplitude and the phase of the signal-bearing wave (the object wave) are preserved as variations in the contrast and spacing of the recorded interference fringes respectively. The hologram obtained in this manner has many properties similar to the diffraction grating about which we will discuss in the next section. When this hologram is illuminated by light of appropriate wavelength, a three- dimensional image of the object can be obtained.

## 14.3 HOLOGRAPHY: THE PROCESS

As mentioned earlier, the process of image formation by holography is a two step process. In the first step, the waves reflected from the object are recorded in such a way that complete information regarding the amplitude and phase variations is preserved. This recording of wave-front is called the hologram. The second step involves the reconstruction of an image of the object by illuminating the hologram by light wave called reconstruction wave (which is identical to the reference wave). In the following, we discuss these two steps and also mention some of the practical considerations about the holographic technique.

### 14.3.1 Production of a Hologram

Holograms can be produced in several ways depending upon the relative orientation of the reflected (or scattered) and the reference waves. For example, Gabor's zone-plate, which is nothing but a hologram, is the record of interference between the two waves travelling more-or-less in the same direction. This is easily done with objects that have enough open spaces between them, such as a wire mesh or opaque letters on a clear background (Fig. 14.1c). Signal and reference, in other words, travel in the same direction. Such a hologram is called Gabor hologram or **in-line hologram.** It was only after the invention of laser that this novel technique of photography became truly practical. With the help of lasers, *N*. Leith and Juris Upatnicks' produced what is known as **off-axis hologram.** In the off-axis hologram, the reference beam and the object beam arrive at the recording plate from substantially different directions. This made possible holography of solid three dimensional objects. Now, the question arises: How holograms are recorded? To understand this, refer to Fig. 14.3. A beam of coherent laser light (in which all points on the wavefront are in phase) is split into two beams. One beam illuminates the object to be recorded and the light reflected from this object falls on a photographic plate. The other beam, called the reference beam, is reflected from a mirror to the same photographic plate. Due to superposition of wavefronts of these two beams, an interference pattern is recorded on the photographic plate. The record on the photographic plate (hologram) is simply a pattern of interfering wavefronts and shows no resemblence to the recorded object. The hologram, however, contains "all the information" about the object.

Ordinarily, these interference fringes are very closely spaced and cannot be seen by unaided eye. Hence the hologram appears to be uniformly gray. When seen by

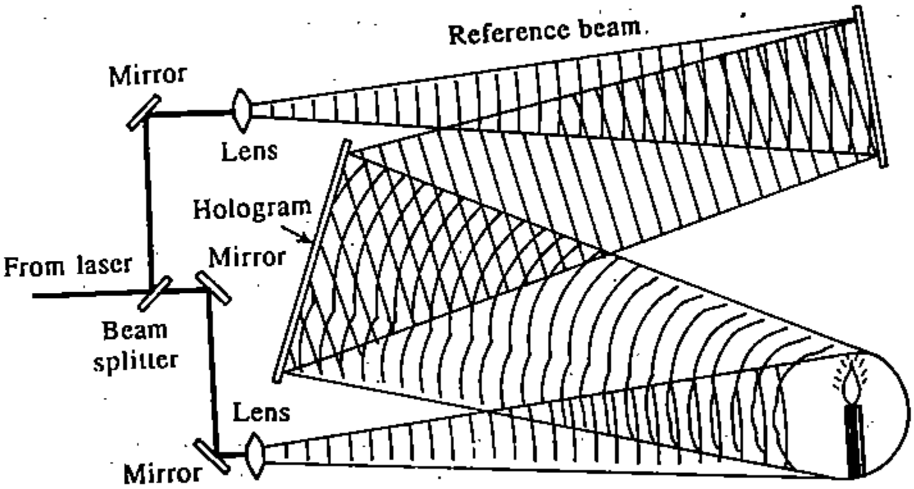


Fig. 14.3: Recording the hologram; microscope lenses broadens both beams without affecting their coherence.

microscope, however, a hologram is found to consist myriad of tiny "cells", each cell containing a series of fringes of various lengths and spacing. Further, a laser is needed for holography, merely because its coherence length exceeds the path difference due to unevenness of the object.

Now, having learnt how holograms are recorded, let us pause for a moment and think about the fundamental difference-in terms of technique as well as characteristics-of a hologram and a conventional photograph. This is the subject matter of TQ 1.

## 14.3.2 Reconstruction of Image

As mentioned above, hologram of an object is the recording of the interference pattern, on a photographic plate, produced by the object and the reference waves. The hologram, when viewed with unaided eye, does not even remotely resemble the object photographed. The process of obtaining image of the object is known as **reconstruction.** In the reconstruction process, as shown in Fig. 14.4, the hologram is illuminated by the light beam (which is similar to reference beam) alone and the reconstructed wavefronts appear to diverge from the image of the object. Let us investigate the process analytically.

Let us represent the wave reflected (or scattered) from the object when it reaches the photographic plate as (Fig. 14.5)

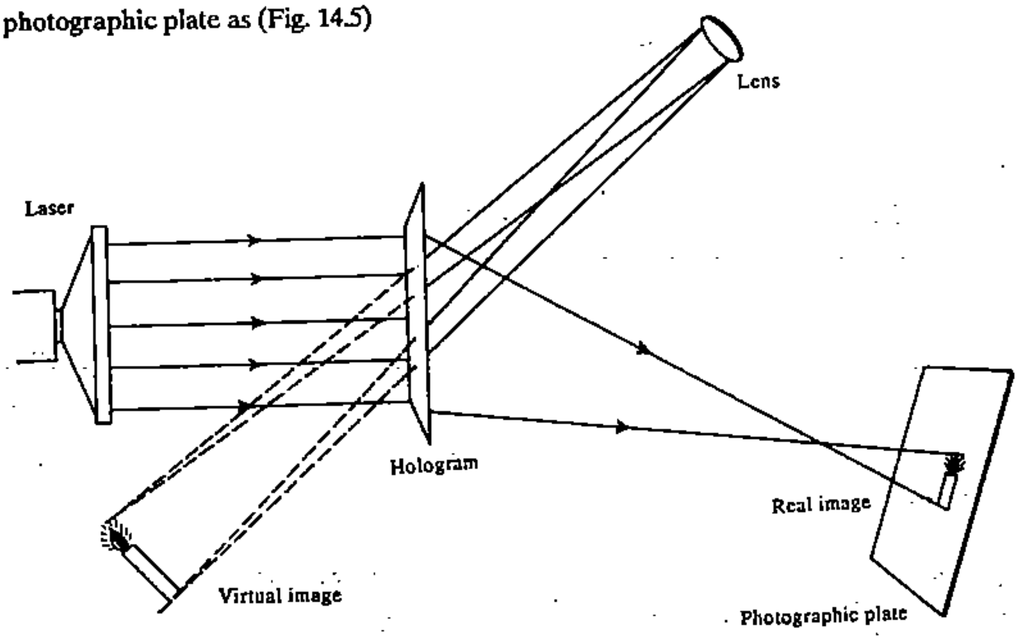


Fig. 14.4: Reconstruction process of an image in holography

$$\psi_1 = A_1(x,y)\cos[\omega t + \phi_1(x,y)] \tag{14.1}$$

and the reference wave as

$$\psi_2 = A_2\cos[\omega t + \phi_2(x,y)] \tag{14.2}$$

You may notice that the amplitude of the reference wave is not a function of $x$ or $y$ (the photographic plate is in $xy$ plane) indicating, therefore, that it is constant at all points on the photographic plate. On the other hand, the amplitude of the object wave, $A_1$, is a function of $x$ and $y$ because it will vary from point to point on the photographic plate due to reflection from the object. Similarly, the phase of the reference wave $\phi_2$ will be constant if it (the reference wave) falls normally on the photographic plate and will be function of $x,y$ if the incidence is at some angle. The phase of the object wave $\phi_1$ will be, however, a function of $x$ and $y$. When these two waves arrive at the photographic plate, the total field distribution will be

$$\begin{aligned}\psi_{\text{total}} &= \psi_1 + \psi_2\\ &= A_1(x,y)\cos\Big[\omega t + \phi_1(x,y)\Big] + A_2\cos\Big[\omega t + \phi_2(x,y)\Big]\end{aligned} \tag{14.3}$$

As you know, the photographic plate responds only to the intensity. Thus, to get the intensity distribution on the photographic plate, we must take the time average of $(\psi_{\text{total}})^2$ i.e.

$$\begin{aligned}I(x,y) &= \langle(\psi_{\text{total}})^2\rangle\\ &= \langle\Big[A_1(x,y)\cos[\omega t + \phi_1(x,y)] + A_2\cos[\omega t + \phi_2(x,y)]\Big]^2\rangle\\ &= A_1^2\langle\cos^2(\omega t + \phi_1)\rangle + A_2^2\langle\cos^2(\omega t + \phi_2)\rangle\\ &\quad + 2A_1A_2\langle\cos(\omega t + \phi_1).\cos(\omega t + \phi_2)\rangle\\ &= \frac{A_1^2}{2} + \frac{A_2^2}{2} + 2A_1A_2\cdot\frac{1}{2}\langle\cos(2\omega t + \phi_1 + \phi_2) + \cos(\phi_2 - \phi_1)\rangle\end{aligned}$$

$$(\because \cos(A+B) + \cos(A-B) = 2\cos A\cos B)$$

$$= \frac{A_1^2}{2} + \frac{A_2^2}{2} + A_1A_2\cos(\phi_2 - \phi_1) \tag{14.4}$$

Eq. (14.4) indicates that the phase information of the object wave is also recorded in the intensity pattern on the photographic plate.

Now, as mentioned earlier, during the reconstruction process, the interference pattern on the photographic plate (called hologram) is illuminated by a reconstruction wave. Let this reconstruction wave, $\psi_3$ has the same phase as that of the reference wave, $\phi_2$. So,

$$\psi_3(x,y) = A_3\cos[\omega t + \phi_2(x,y)] \tag{14.5}$$

What will be the nature of the transmitted wave when the reconstruction wave falls on the hologram? Well, the hologram is exposed in such a manner that the amplitude transmittance is linearly related to $I(x,y)$, the incident intensity at the time of recording. So, we have, the transmitted wave

$$\psi_4 \, \alpha \, \psi_3 (x,y) \, I(x,y)$$

$$\psi_4 = \left[ \frac{(A_1^2 + A_2^2)}{2} \psi_3 + \frac{A_1 A_2 A_3}{2} \cos(\omega t + \phi_1) + \frac{A_1 A_2 A_3}{2} \cos(\omega t + 2\phi_2 - \phi_1) \right] \quad (14.6)$$

**SAQ 2**

Starting from the relation

$$\psi_4 \, \alpha \, \psi_3 (x,y) \, I(x,y)$$

derive Eq.(14.6) using Eqs. (14.4) and (14.5)

*Spend 5 min*

The transmitted wave represented by Eq. (14.6) consists of three terms. What do these term signify physically? The first term is the reconstruction wave ($\psi_3$) with its amplitude modulated by the amplitude of the object wave ($A_1$). It is so because $A_1$ is a function of $x$ and $y$ whereas the reference wave amplitude $A_2$ is a constant. As a result, this part of the transmitted wave will travel, with slight attenuation, in the direction of the reconstruction wave. The second term is identical to the object wave ($\psi_1$) except for the constant term $(A_2 A_3)/2$. Here lies the beauty of holography. **The hologram and the reconstruction wave have generated a wave which is in every way identical to the wave which originated from the real object itself while recording the hologram.** This part of the transmitted wave forms a virtual image of the object. The third term which is similar to the object wave forms a real image of the object. As a result, a three-dimensional picture of the object can be obtained by placing a camera in the position of real image. The reconstruction process alongwith various parts of the transmitted wave is shown in Fig. 14.5. You may note that the object is not present when image is reconstructed. However, one of the evolving beam, resulting due to reconstruction process, is identical to the beam reflected by the object at the time of recording the hologram.

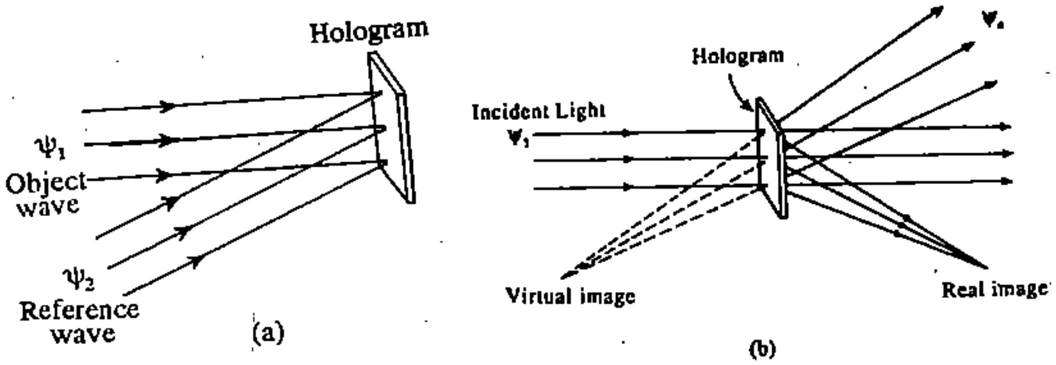


**Fig. 14.5: (a) Recording the hologram: Wave reflected from an object interferes with the reference wave. (b) Reconstruction: The hologram diffracts the reconstruction wave, resulting in transmitted wave which produces a real and a virtual image.**

## 14.3.3 Practical Considerations

So far we have discussed physical principles and the experimental arrangements of holography. Suppose you are in the actual process of producing holograms and its subsequent reconstruction to obtain a three-dimentional image of the object. What are the important aspects of the process, and components used therein, about which you should be careful? Well, there are several practical considerations in holography which are essentially related to the photographic film, the stability and the coherence condition. Let us have a closer look on these practical considerations.

So far as the photographic film is concerned, hologram must be recorded on films of high resolvance. Look again at Fig. 14.3. You may notice that the reference wave, (the light reflected by the mirror), and the signal (the light reflected by the object) subtend certain angle at the photographic plate. If this angle is too large, more than a few degrees, the fringes formed between the signal and reference are very closely spaced and even the best emulsion cannot resolve them. To obtain high resolution, extremely fine-grain film has to be used. But fine-grain films are very slow and hence require larger exposure time (a few minutes). And, if during this exposure time object moves, the recording of hologram will not be proper. What is the way out of this problem? The way out of this situation is to use high power laser beam to compensate for the exposure time.

Further, the whole system of recording the hologram should be highly stable i.e. it should be completely free from vibration. Can you say why? It is because the density of the fringes on the photographic film is extremely high. For example, if the angle between the signal and the reference wave is 30° (Refer Fig. 14.3) and the wavelength of the laser light is 633nm, the fringe frequency (Refer to Block-2)

$$= 1/d\,;\ \text{where } d \text{ is fringe width}$$

$$= \frac{\sin\theta}{\lambda} = \frac{\sin 30}{633 \times 10^{-9}} \text{ lines per meter}$$

$$= 7 \times 10^{5} \text{ lines per meter.}$$

Can you imagine the smallness of this seperation! The fringe width will typically be a thousandth of a millimeter. Therefore, if any component of the holographic set-up moves during recording, the whole fringe pattern will disappear. To meet this stability requirement, the film exposure time should be kept minimum (by using very high power laser) and the holographic system should be isolated from outside vibrations.

The most important and obvious consideration in holography is to use coherent illumination. The coherence length of the laser used for illuminating the object must be greater than the path difference between the reference wave and the object wave. The practical problem is that as power of laser increases (which we use for minimising the exposure time), its coherence length is reduced. Similarly, the coherence area (spatial coherence) of illumination from a laser must be greater than the transverse size of the object to be photographed.

Having learnt about various aspects of holography, you may now be interested to know about its applications. This is the subject matter of the next section.

## 14.4 APPLICATIONS OF HOLOGRAPHY

There are many aspects of holography. Its influence on interferometry, photography, microscopy, astronomy, pattern recognition and even art has only begun to bear fruit. We will now discuss these in brief.

### Holograaphic Interferometry

You will appreciate that, in most of the cases, one of the first areas to benefit from the new technique was the area that gave rise to it. Similar was the case with holography which introduced a new range of powerful methods to interferometry. Interferometry is generally used for precise measurement and comparison of wavelengths, for measuring very small distances or thicknesses (of the order of wavelengths of light) etc. Testing for stresses, strains and surface deformation is one of the most useful practical applications of holographic interferometry.

In the double-exposure technique of holographic interferometry for measuring deformation in object due to strain, two exposures are made of the object, - one before loading, and the other after (i.e. under strain). The original object and the object after deformation are recorded holographically on the same photographic plate. The hologram thus obtained is a double exposure, with the second pattern of wave fronts superposed on the first. When this hologram is reconstructed by illuminating it with the reference wave, both images are viewed simultaneously. Since they are slightly different due to deformation, the two images interfere. Thus, any distortion of the object will show in the form of fringes. Like other kinds of interferometry, the technique readily detects changes that produce optical-path difference of the order of a fraction of the wavelength of light. And unlike normal interferometry, however, it is possible to perform experiment quite readily with almost any type of material.

## Holographic Microscope

Microscopy has been the primary area of application of holography. In fact, Gabor's discovery of this tecnique was the outcome of his attempt to enhance the resolving power of an electron microscope. In contrast to a conventional high power microscope, a holographic microscope has an appreciable depth of field and it need not be focussed at all. To see how a holographic microscope functions, refer to Fig. 14.6. The light beam from laser is split into two. One beam is passing through the specimen and through the microscope, and the other beam is led around it. The two beams interfere on the film, producing hologram. The reconstructed image can be viewed in any desired cross-section. The observer merely looks at the cross-section, he or she wishes to see, moving back and forth throughout the depth of the image without the object being present at all.

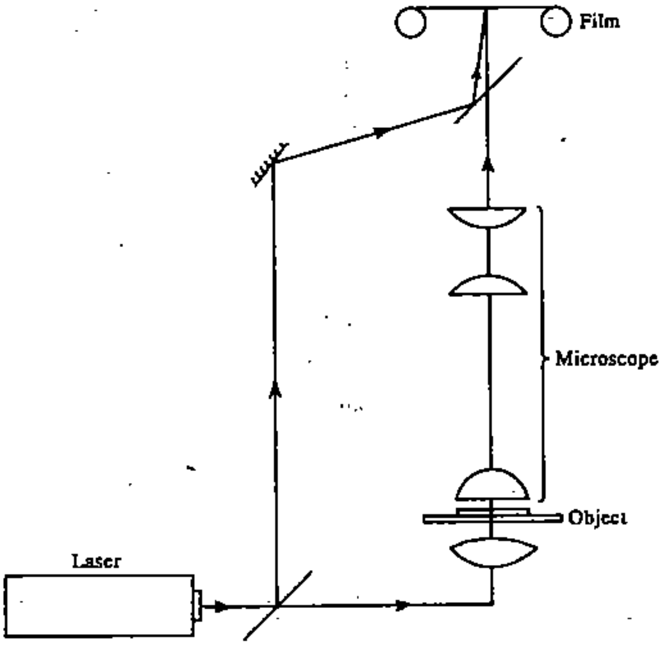


Fig.14.6: Holographic microscope.

## Information Storage

Information can be stored and retrieved more efficiently in the form of holograms than in the form of real images. Further, it is the characteristic of the hologram that it will only reconstruct the holographic image if the reconstruction beam is incident on the hologram at the correct angle. Due to this property, several holograms can be recorded on the same holographic plate by using a slightly different angle between the object and the reference beams for each hologram. Thus, on reconstruction, depending upon the angle of incidence of reconstruction beam, a particular holographic image will be visible. Perhaps this is how information is stored in the brain. If that is the case, it would help explain why attempts to locate certain centres in the brain never met with much success and why brain injury often does not lead to predictable circumscribed defects.

## Pattern Recognition

One of the most exciting applications of holography is the pattern recognition, also called the character recognition. Early pattern recognition systems, before holography came on the scene, were based on geometrical optics. Consider, for example, that we want to read the letter $A$ (Fig. 14.7). $A$ set of characters $A, B, C$ ... are printed on a strip of film and this film is moved through the image plane. If the character to be read matches the character on the film, the output from a photo detector is zero, triggering a printer. But, in reality, this does not work. The character and the negative must be aligned perfectly, both in position and size, which is an unrealistic requirement.

Modern pattern recognition systems are based on holography. In place of a mark containing the real imagge of the letter $A$ we may use the hologram of the letter $A$.

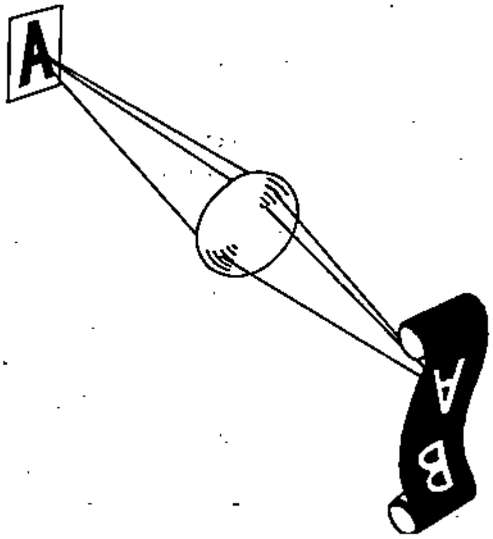


Fig.14.7: Pattern recognition based on geometric optics

As in holography, hologram of the letter $A$ is the superposition of two sets of wavefronts, the signal and the reference. The signal is diffracted by an original $A$ and the reference is a beam of collimated light. Subsequently, when the hologram of $A$ is illuminated with light from another $A$, plane wavefronts arise that can be focussed into a bright spot (Fig. 14.8 top). The spot can easily be recognised by eye or photoelectrically. On the other hand, if the wavefronts are coming from $B$ or from other characters, they do not transform into perfectly plane wavefronts and do not produce a focussed spot. Instead, a diffuse patch of light (centre) is produced. Hence, we can scan a given matrix of characters and determine whether or not a particular character is present (bottom).

The holograms shown in Fig. 14.8 appear to be amplitude filters. But because they are generated by interference between signal and reference, they in fact represent both amplitude and phase of light. They are called "complex", "matched", or "vander Lugt filters".

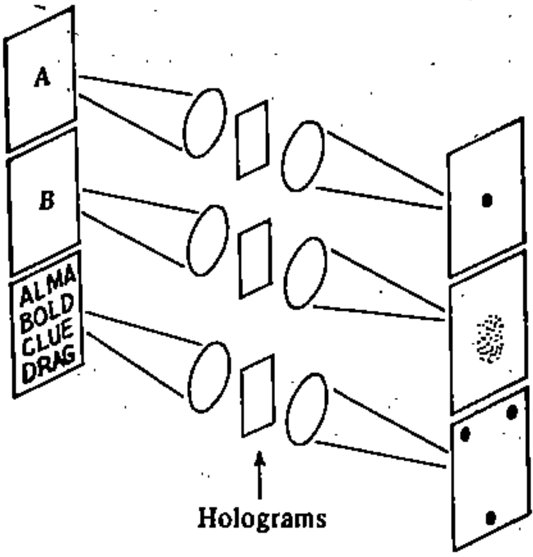


**Fig. 14.8:Pattern recognition by holographic vander Lugt filter. (The holograms are seen between the lenses.)**

Form reading machine are a distant reality. Some letters and words are "inside" others. For example *F* is inside *E*, *P* is inside *R* and *B*, *T*, *L* have the same horizontal and vertical lines and 'arc' is inside 'search'. Clearly, more alike are the two characters, the less will be the power of discrimination. Another problem will be to 'teach' the machine to recognise the "meaning" of a letter set in different typeface. The letter *A* can be written in an infinite number of variations possible when it comes to handwriting. However, pattern-recognition using holographs is being extensively used in developing fingeprints library which stores the fingerprints of individuals with dubious character.

## 14.5 SUMMARY

- Holography, discovered by Gabor, is a novel technique of photography by which a three-dimensional picture of an object or a scene can be obtained. In holography, the interference pattern produced by the light reflected from the object and a reference beam is recorded. Such recording on the photographic plate is called the hologram.
- The three-dimensional picture of the object is obtained by illuminating the hologram by a reconstruction light beam, which in most cases, is identical to the reference beam. Holography is, therefore, also known as wavefront reconstruction photography.
- Hologram is produced by splitting a beam of coherent light from laser into two. One beam is directed, with the help of mirror(s), towards the object and the other is made to fall directly on the photographic plate. The light reflected from the object reaches the photographic plate and interfers with the reference beam. The recorded interference pattern on the photographic plate is the hologram.

- If $\psi_1\ (=A_1(x,y)\cos[\omega t+\phi_1(x,y)]$ and $\psi_2\ (=A_2\cos[\omega t+\phi_2(x,y)]$ respectively represents the object wave (wave reflected from the object being photographed) and the reference wave, the intensity distribution on the photographic plate is given as

$$I(x,y) = \frac{A_1^2}{2} + \frac{A_2^2}{2} + A_1A_2\cos(\phi_2 - \phi_1)$$

- During reconstruction of image, when the hologram is illuminated by the reconstruction wave, $(\psi_3 = A_3\cos[\omega t + \phi_2(x,y)]$ the transmitted wave through the hologram is

$$\psi_4 = \left[\frac{(A_1^2 + A_2^2)\psi_3}{2} + \frac{A_1A_2A_3}{2}\cos(\omega t + \phi_1) + \frac{A_1A_2A_3}{2}\cos(\omega t + 2\phi_2 - \phi_1)\right]$$

The second term on the right hand side has the same form as the object wave and it represents the three-dimensional virtual image of the object. The third term is also similar to the object wave and represents the real image of the object which can be recorded on a photographic plate.

- In order to obtain a hologram, the photographic plate on which the hologram is to be obtained must be of high resolution. This is required because the density of interference fringes in the hologram is extremely high. Also, the whole arrangement of holography- recording the hologram as well as its subsequent reconstruction-must be highly stable, i.e. it should be free from even a slightest mechanical vibration. And of course, we must use coherent light for recording the hologram as well as reconstructing the image.

- Holography has varied applications. Holographic interferometry is a distinct improvement over normal interferometry because the former can be used for any kind of material. Holographic microscopy has enourmous magnification and it also offers appreciable depth of field. Holography find extensive use in information storage and pattern recognition.

## 14.6 TERMINAL QUESTIONS

1. (a) How is the process of holography different from ordinary photography?

   (b) Discuss some of the salient features of a hologram?

2. Following Gabor, assume that amplitudes of signals and reference are in ratio 1:10. Suppose that the two beams when they combine may be completely out of phase or in phase. What is the maximum ratio of their intensities?

3. If the angle subtended at the hologram by the signal and the reference beam is 15°, what is the spacing of the fringes provided the wavelength is 492 nm?

## 14.7 SOLUTIONS AND ANSWERS

### SAQs

1. (a) The least possible amplitude (when signal and reference are out of phase, pointing in opposite direction) is $4.36 - 1 = 3.36$.

   This is because, measuring the lengths of vectors Fig.14.2, we find that the ratio of signal versus reference is $1:4.36$.

The highest possible amplitude (when signal and reference are in phase) and pointing in same direction is 4.36 + 1 = 5.36. The ratio of the amplitudes = 3.36/5.36 . Thus, the ratio of intensities is

$$= (3.36/5.36)^2 = 0.39$$

(b) The contrast is given as

$$\frac{I_{max} - I_{min}}{I_{max} + I_{min}}$$

$$\frac{(5.36)^2 - (3.36)^2}{(5.36)^2 + (3.36)^2} = 0.44$$

which is high enough to make the reconstruction visible.

2. The transmitted wave is linearly proportional to the incident intensity $I(x, y)$ at the time of recording the hologram and the reconstruction wave, i.e.

$$\psi_4 \,\alpha\, \psi_3 (x, y) I (x, y)$$

$$\alpha\, \psi_3 \left[ \frac{A_1^2}{2} + \frac{A_2^2}{2} + A_1 A_2 \cos(\phi_2 - \phi_1) \right]$$

(using egn 14.4)

$$\alpha \frac{(A_1^2 + A_2^2)}{2} \psi_3 + \psi_3 \left[ A_1 A_2 \cos(\phi_1 - \phi_1) \right]$$

$$\alpha \frac{(A_1^2 + A_2^2)}{2} \psi_3 + \left[ A_3 \cos(\omega t + \phi_2) \right] \left[ A_1 A_2 \cos(\phi_2 - \phi_1) \right]$$

(using equaqtion 14.5)

$$\alpha \frac{(A_1^2 + A_2^2)}{2} \psi_3 + A_1 A_2 A_3 \left[ \cos(\omega t + \phi_2) \cos(\phi_2 - \phi_1) \right]$$

$$\alpha \frac{\psi_3 (A_1^2 + A_2^2)}{2} + A_1 A_2 A_3 . 1/2 \Big[ \cos(\omega t + \phi_2 + \phi_2 - \phi_1)$$

$$+ \cos(\omega t + \phi_2 - \phi_2 + \phi_1) \Big]$$

( using $\cos(A + B) + \cos(A - B) = 2 \cos A \cos B$ )

$$\psi_4 = \frac{(A_1^2 + A_2^2)}{2} \psi_3 + \frac{A_1 A_2 A_3}{2} \cos(\omega t + \phi_1) + \frac{A_1 A_2 A_3}{2} \cos(\omega t + 2\phi_2 - \phi_1)$$

which is equation (14.6)

## TQs

1. (a) The technique of holography (photography by wave front reconstruction) differs from that of ordinary photography in three aspects. Firstly, in ordinary photography, the light reflected from the object is received on the photographic plate with the help of lenses or other image forming device. Amplitude of the light wave, reflected from each point of the object, is recorded at corresponding point on the photographic plate. On the other hand, in holography, no lens or other

image forming device is needed and hence, as such, no image is formed on the hologram. What essentially is obtained is the interference pattern due to the light reflected from the object and the reference beam. Secondly, for obtaining hologram, coherent light is used whereas in case of normal photography, no such source of light is needed. The requirement of coherent light is due to the fact that the hologram is an interference pattern. Thirdly, in holography, a set of mirrors is used to render the reference and object beam on the photographic plate.

(b) Hologram has several interesting properties. Some of them are given below:

(i) The image obtained from the hologram has three-dimensional character unlike normal photographs which are two-dimensional. Due to the three-dimensional character of the image obtained in holography, you can observe different perspective of the object by changing the viewing position. Also, if a scene has been recorded, you can focus at different depths.

(ii) We do not obtain negative in holography. Hologram itself, however, can be considered as negative in so far as obtaining the positive is concerned. Otherwise, there is no similarity between the typical negative of the ordinary photographs and the hologram. You may have noticed that when the negative of an ordinary photograph is seen through, we do get a feel of the object or the scene photographed. On the other hand, when we look at a hologram we observe a hodgepodge of speaks, blobs and whorls; it has no resemblance whatsoever with the original object.

2. Let amplitude of the signal (or the object wave) be $A_1$ and that of the reference wave be $A_2$, then, as per the problem

$$\frac{A_1}{A_2} = \frac{1}{10}$$

When these two waves are out of phase, their resultant amplitude will be (10 -1) = 9. On the other hand, when they are in phase, the resultant amplitude will be (10 + 1) = 11. Thus, the ratio of their intensities,

$$\frac{I_{min}}{I_{max}} = \frac{(9)^2}{(11)^2} = 0.67$$

3. The spacing of the fringes is given as

$$d = \frac{\lambda}{\sin\theta}$$

$$= \frac{492 \times 10^{-9}}{\sin 15^{\circ}} \text{m}$$

$$= 1.8\,\mu\text{m}$$

# UNIT 15 FIBRE OPTICS

## Structure

## 15.1 INTRODUCTION

You might have seen advertisement dispalys (made of glass or plastic rods) and illuminated fountains. While looking at these, you might also have noticed that light seems to travel along curved path. In the above mentioned cases, most of the incoming light is contained within the boundaries of the medium (glass or plastic or water) due to the phenomenon of **total internal reflection**. And since the medium itself has a curved shape, the light travelling through it appears to travel along a curved path. Optical fibre, which is made of transparent glass or plastic, also transmit light in a similar fashion. These fibres are thread like structure and a bundle of it can be used to transmit light around corners and over long distances. Since optical fibre can transmit light around corners, it is being used for obtaining images of inaccessible regions e.g. the interior parts of human body. The real potential of the optical fibres was, however, revealed only after the discovery of lasers.

You may recall from Unit 13 of this course that the discovery of lasers- a source of coherent and monochromatic light - raised the hope of realising communication at optical frequencies. Since increase in frequency of the carrier wave enables it to carry more information, communication at optical frequencies ( $\sim 10^{15}$ Hz) has obvious advantages over communication at radio wave ( $\sim 10^{6}$ Hz) and microwave ( $\sim 10^{9}$ Hz) frequencies. But, early attempts at communication at optical frequencies faced a major problem. When optical radiation travles through the Earth's atmosphere, it is attenuated by dust particles, fog, rain etc. Thus, a need for an **optical waveguide** was felt and the answer was the optical fibres. Optical fibres are an integral part of optical communication — transmission of speech, data, picture or other information — by light. In this unit, you will study about the optical fibres, especially in the context of optical communication.

In Sec.15.2, you will learn the physical principles involved in transmission of light through fibres. Types of fibres used in optical communication has also been explained. General considerations about the optical communication through fibres has been discussed in Sec. 15.3. In the same section, you will also learn about the requirements which must be met by optical fibres so that efficient optical communication may take place. The area of optical fibre is relatively new and an exciting field of activity. A thorough understanding demands rather sophisticated mathematical background on the part of the student. It has, therefore, been attempted here to keep the mathematical aspects to a bare minimum and the underlying physical principles have been highlighted.

**Objectives**

After going through this unit, you should be able to

- explain light transmission through fibre
- distinguish between step-index and GRIN fibres
- derive expression for pulse dispersion in fibres, and
- solve simple problems on optical fibres.

## 15.2 OPTICAL FIBRES

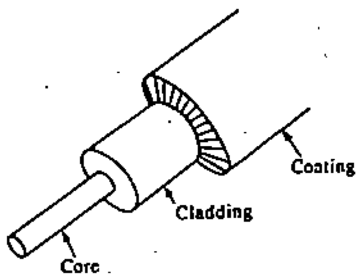


**Fig. 15.1: Optical Fibre**

An optical fibre consists of a cylindrical glass core surrounded by a transparent cladding of lower refractive index. This assembly is further covered by a plastic coating to protect it against chemical attack, mechanical impact and other handling damages. Fig. 15.1 shows the geometry of a typical optical fibre. The core diamter is in the range 5 $\mu$m to 125 $\mu$m with the cladding diameter usually in the range 100 $\mu$m to 150 $\mu$m. The plastic coating diameter is around 250 $\mu$m.

In order to understand why the incoming light does not come out through the cylindrical surface of the fibre, you should recall the phenomenon of total internal reflection. You are aware that when light travels from an optically denser medium to a rarer medium, it bends away from the normal as shown in Fig.15.2(a). If the refractive indices of the two media are $n_1$ and $n_2$ such that $n_1 > n_2$, and $\theta_1$ and $\theta_2$ are the angle of incidence and angle of refraction respectively, then, from Snell's law

$$\frac{n_1}{n_2} = \frac{\sin\theta_2}{\sin\theta_1} \qquad (15.1)$$

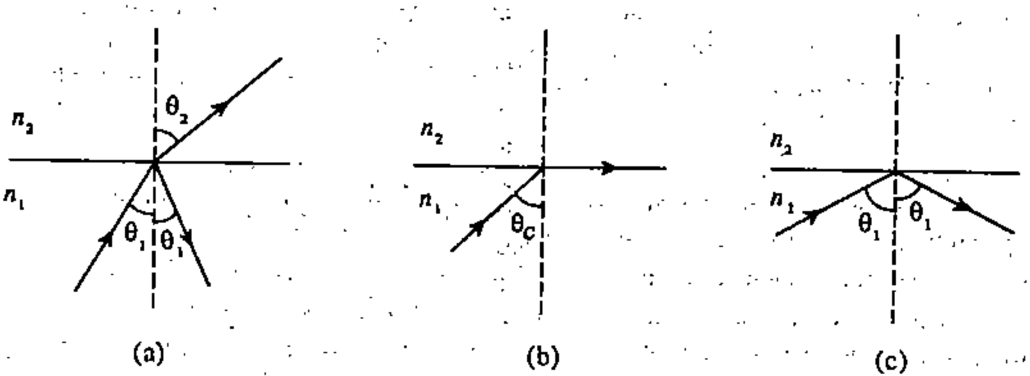


**Fig. 15.2: Total internal reflection**

As the angle of incidence is increased, the refracted ray will further bend away form the normal. Ultimately, when the angle of incidence reaches the critical value - known as critical angle, $\theta_c$ - the refracted ray travels along the interface separating the two media, as shown in Fig. 15.2 (b). And, when the angle of incidence is increased beyond $\theta_c$, there is no refracted ray and the incident ray undergoes total internal reflection into the optically denser medium, Fig.15.2(c). This phenomenon is known as total internal reflection and the critical angle, $\theta_c$ is given as, from Eq.(15.1)

$$\frac{n_1}{n_2} = \frac{\sin(\pi/2)}{\sin\theta_c} \Rightarrow \theta_c = \sin^{-1}(n_2/n_1) \qquad (15.2)$$

Transmission of light, based on above principle, through an optical fibre of core refractive index $n_1$ and cladding refractive index $n_2$ with $n_1 > n_2$ is shown in Fig.15.3(a).

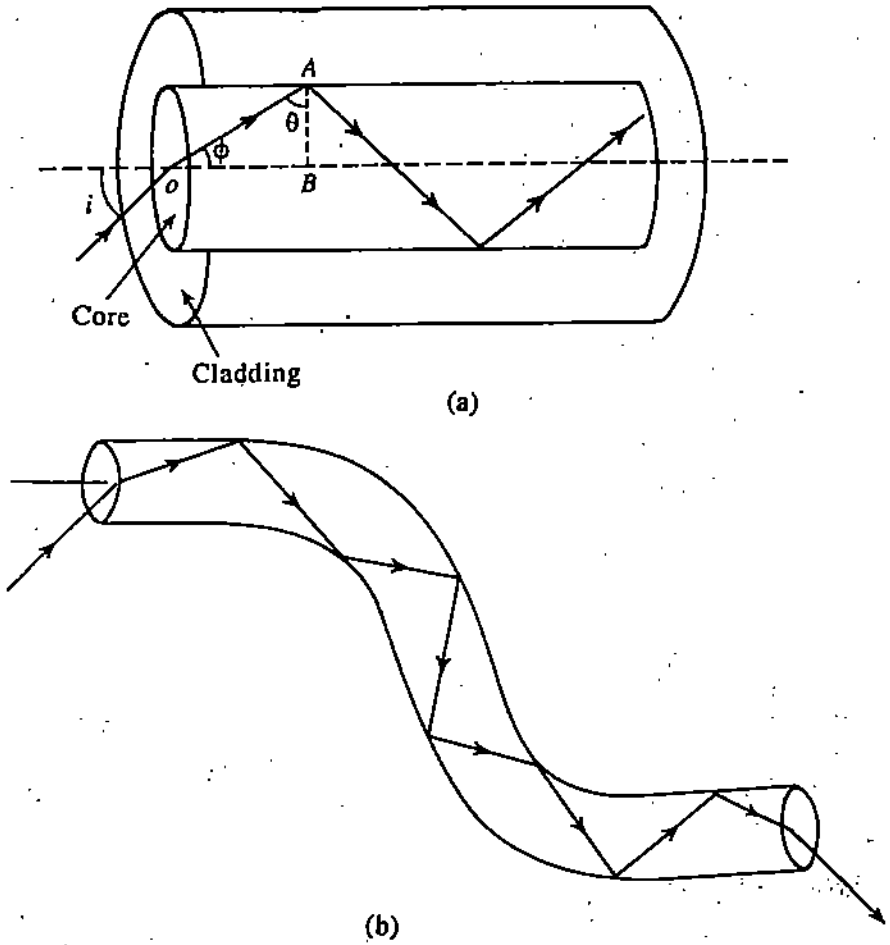


**Fig.15.3: (a) Light propagation through a fibre by total internal reflection.**
**(b) Light propagation through a bent fibre.**

When the ray of light is incident at angle $\theta(> \theta_c)$ at the core - cladding interface, it undergoes total internal reflection. Due to cylindrical symmetry of the fibre, the ray undergoes total internal reflection at subsequent incidences at the core -cladding interface and hence gets trapped inside the fibre. Due to this "guiding" property, optical fibres are also called **"Optical Waveguides"**. Fibres in the bent form can also guide the light, as indicated in Fig.15.3(b), provided that, even at curved portion, the angle of incidence is greater than $\theta_c$. Do you know why cladding material is needed? The need for a cladding material of lower refractive index is due to two reasons. Firstly, to achieve total internal reflection at the core -cladding interface. Secondly, when light undergoes total internal reflection, a part of it penetrates into the cladding material (region of lower refractive index). This may lead to leakage of light, and it may also couple with the light travelling in adjacent fibres. The use of sufficiently thick cladding material prevents this type of loss.

You may note, from Eq. (15.2), that the critical angle for the incident ray depends on the refractive indices of the core and the cladding material. In Fig. 15.3(a), $\theta$ is the angle at which incident light falls on the core - cladding interface and this angle is different from the angle, $i$, at which light is incident at the entrance aperture of the fibre. It is so because the entrance aperture is an air (refractive index $n_0 \sim 1$) - glass (refractive index $n_1$) interface. Thus, according to Snell's law, (refer to Fig.15.3(a))

$$n_0 \sin i = n_1 \sin \phi \qquad (15.3)$$

Now, if this ray has to undergo total internal reflection at the core - cladding interface, from Eq.(15.2)

$$\sin\theta \geq n_2/n_1$$

from Δ *OAB*,

$$\sin\phi = \sin(90 - \theta) = \cos\theta$$
$$= (1 - \sin^2\theta)^{1/2}$$
$$= [1 - (n_2/n_1)^2]^{1/2}$$

Hence, Eq. (15.3), taking $n_0 = 1$, may be written as,

$$\sin i_{max} = n_1 \sin\varphi$$

$$= n_1 \left[\frac{n_1^2 - n_2^2}{n_1^2}\right]^{1/2}$$

$$= \left(n_1^2 - n_2^2\right)^{1/2}$$

$$i_{max} = \sin^{-1}\left[n_1^2 - n_2^2\right]^{1/2}$$

(15.4)

The angle of incidence, $i_{max}$, given by Eq.(15.4) is a measure of the light gathering capacity of the fibre. You should convince yourself that if the incidence angle is greater than $i_{max}$, the light will be refracted into the cladding material. All the light incident on the fibre aperture along the core formed by $i = 0$ to $i = i_{max}$ will undergo total internal reflection in the fibre. The quantity $(n_{12} - n_{22})_{1/2}$ in Eq. (15.4) is called the numerical aperture of the fibre.

### 15.2.1 Types of Fibres

As mentioned above, in its simplest form, an optical fibre consists of a glass core and a cladding (also of glass) of lower refractive index. This type of fibre in which there is a sudden change in the refractive index at the core-cladding interface is called **Step-index fibre**. The variation of the refractive index with the radius of such a fibre is shown in Fig. 15.4.

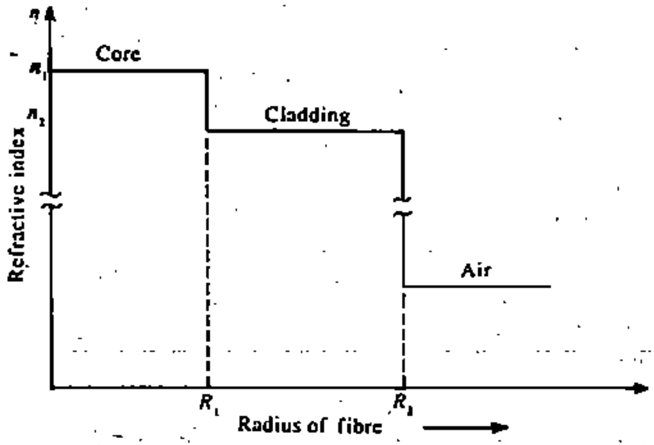


Fig.15.4: Refractive-index profile of a step-index fibre.

In optical communication, signal is transmitted through the fibre in the form of pulses.

Further, when light travels through the optical fibres, there are different types of losses as well as a broadening of the pulse. These aspects of the optical fibres are of vital importance for optical comunication and have been discussed in the next section. In order to overcome some of the inherent deficiencies of the step - index fibres, another type of fibre in use is called **GRadient - INdex Fibre (or GRIN - fibre)**. In the GRIN

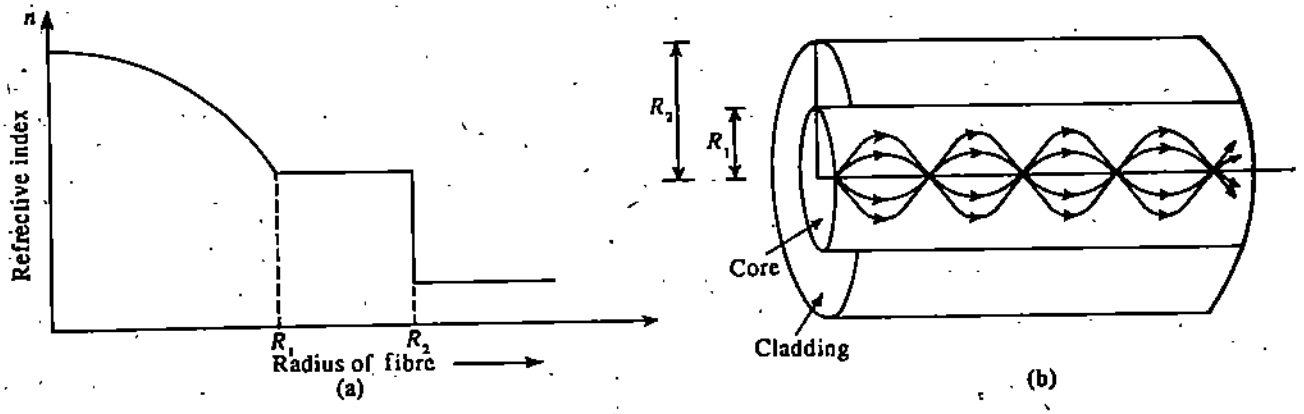


Fig.15.5: (a) The refractive index profile of a Gradient - Index fibre; (b) Ray paths in such a fibre.

fibre, the refractive index of the core material decreases continuously along the radius, nearly in parabolic manner, from a maximum value at the center of the core to a constant value at the core - cladding interface. The variation of the refractive index, with radius, of a GRIN - fibre is shown in Fig.15.5(a).

Since the refractive index gradually decreases as one moves away from the axis of the fibre, a ray that enters the fibre is continuously bent towards the axis of the fibre as shown in Fig.15.5(b). Can you explain why does this happen? This smooth bending of the ray towards the axis is again a consequence of Snell's law. As the ray moves away from the centre, it encounters media of lower and lower refractive indices and hence bends towards the axis of the fibre. Can you name a natural phenomenon which results due to the atmospheric gradient of refractive index? You guessed rightly - **the Mirage**, which is observed while looking across on expanse of hot desert is one such example.

**SAQ 1**

What will happen if the refractive index of the cladding material is higher than that of the core?

*Spend 2 min*

Having learnt about the basic principles involved in transmission of light in optical fibres, let us study some of its important features as a component of optical communication system. But before we do that, let us see what are the uses to which optical fibres has been put to.

## 15.2.2 Applications of Optical Fibres

The most elementary application of the optical fibres is the transmission of light either

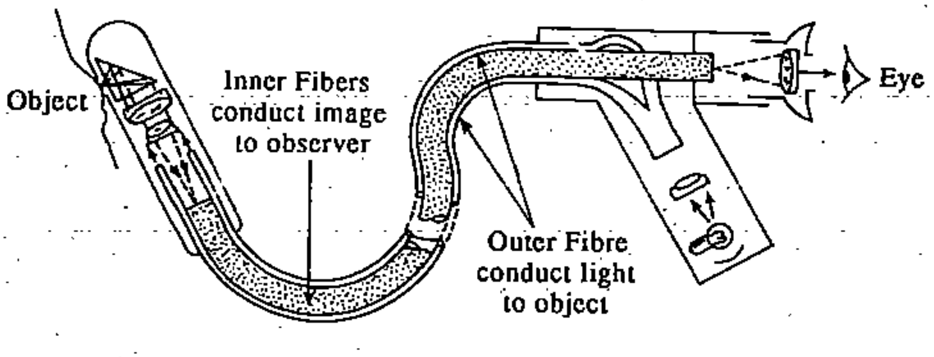

Object
Inner Fibers conduct image to observer
Eye
Outer Fibre conduct light to object

Fig.15.6: Flexible fibrescope

example of transmission of images using optical fibres is the flexible **fibrescope**. As shown in Fig. 15.6, some of the fibres conduct light into the cavity to be examined, while the others carry the image back to the observer. The image conducting fibres, upto 140000 of them, are by necessity very thin, often no more than 10 $\mu$m in diameter, and the entire fibre-bundle has diameter of the order of a few mm. Fibrescope are used extensively in medicine and engineering. They make it possible to inspect a cavity in the human body and to look inside the heart while it beats.

Of increasing interest is the use of fibre guides for communication. Compared to electrical conductors, optical fibres are lighter in weight, less expensive, equally flexible, not subject to electrical interferences and more secure to interceptions. Fibres can now be made which has losses as low as 0.2dBkm$^{-1}$. This is a remarkable achievement considering that only a decade ago the best fibres had losses in excess of 1000 dBkm$^{-1}$ and 20 dBkm$^{-1}$ was thought to be the limit.

## 15.3. OPTICAL COMMUNICATION THROUGH FIBRE

As mentioned earlier, optical communication refers to the transmission of speech, data, picture or other information by light. You may recall from Unit 13 that the replacement of radiowaves and microwaves by light waves is especially attractive because of the enhanced information carrying capacity of the latter. Optical frequencies are some five orders of magnitude higher than, say, microwave frequencies. Therefore, larger volume of information can be transmitted through fibre cable compared to that through copper coaxial cable (used for microwave communication) of similar size. Further, in contrast with metallic conduction techniques (e.g. through copper cables), communication by light offers the possibility of complete electrical isolation, immunity to electromagnetic interference and freedom from signal leakage. In a typical optical communication system, the information carrying signal originates in a transmitter, passes through an optical fibre link or an optical channel and enters a receiver which reconstruct the original information. In order to minimize the distortion, the signal is encoded into digital form before transmission. In this way, retrieval of the signal at some distance down the line depends only on the recognition of either the presence or the absence of a pulse representing a binary (0 or 1) digit. Minor distortion and noise may therefore be tolerated as long as pulses can be detected and regenerated, free from distortion.

You may be wondering that with above advantages, why light was not used for communication purposes. It is not as if these advantages of using light as carrier of information were not known. Rather, it was the unavailability of a suitable source of light which could be modulated. Light from lasers, being highly monochromatic, can effectively be modulated by the information carrying signals. The laser light, acting as the carrier wave, respond, either directly or indirectly, to the electrical signal say, from telephone. These signals can, therefore, modulate the carrier wave which then travel through the optical fibre (the optical waveguide). At the receiving end of the fibre, a photodetector receives and demodulate these optical signal into sound waves. For long distance optical transmission line, yet another component, called **repeater** is used in optical communication system. Repeater essentially amplify and reshape the signal and retransmit it along the fibre.

Optical communication, as such, can be carried out thorugh open space. Then why do we need fibres to carry optical signals? The reason lies in substantial attenuation (or damping) of the signal while it travels in open space between the information source and information use. For example, communication between one sattelite to another is carried out through open space because the intervening region is essentailly vacuum. However, similar open space optical communication will not be feasible between a satallite and the earth or between two places on the earth because earth's atmosphere strongly influences the light transmission. Hence, the need for an optical waveguide (fibres) for terestrial optical communication.

Well, you have learnt in the previous section how light is transmitted through optical fibres. But, is this property of fibres enough for transmitting information carrying signals from one point to another? No, the optical fibre must have some additional characteristics if at all it has to serve as an effective optical signal carrying medium. The optical fibre should be, as much as possible, free from pulse dispersion in order to carry large volume of information. Pulse dispersion arises because different light rays take different times to travel a fixed length in the fibre. Secondly, as we know, even the light from lasers may have a spread in its wavelength. That is, even laser light is not completely monochromatic. And since the refractive index of fibre material is a function of wavelength, light of different wavelengths will travel with different velocities. This inherent property of material is yet another cause of pulse dispersion and is known as material dispersion. Further, the optical radiation will be attenuated by the material of the fibre due to scattering and other phenomenon. In the following you will learn how these problems can be tackled.

### 15.3.1 Pulse Dispersion in Fibres

You may recall from Sec. 15.2 that rays of light incident at the core - cladding interface at an angle greater than the critical angle $\theta_c$ undergoes total internal reflection and propagate through the fibre as shown in Fig. 15.7.

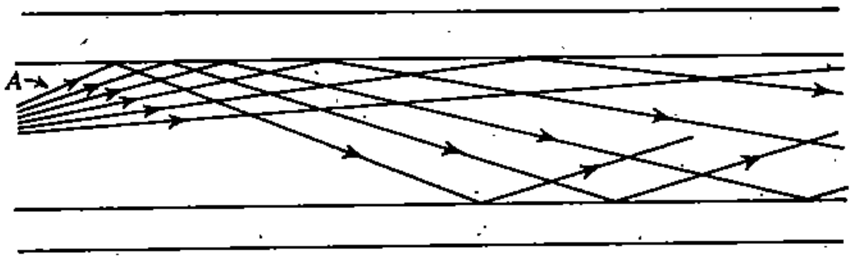


Fig.15.7: Rays of light passing through a fibre

However, the ray, marked *A* in Fig. 15.7, which is incident on the core-cladding interface at the largest angle will travel a longer optical path as compared to other rays incident at smaller angles. As a result, different rays will take different times in traversing a given length of the fibre. This causes broadening of the information carrying pulses, as shown in Fig. 15.8. What effect the pulse broadening has on the signal transmission capacity of

The transmission capacity of the fibre is determined by the number of pulses transmitted per unit time. For correct information retrieval, the pulses must remain resolvable i.e. they should not overlap each other.

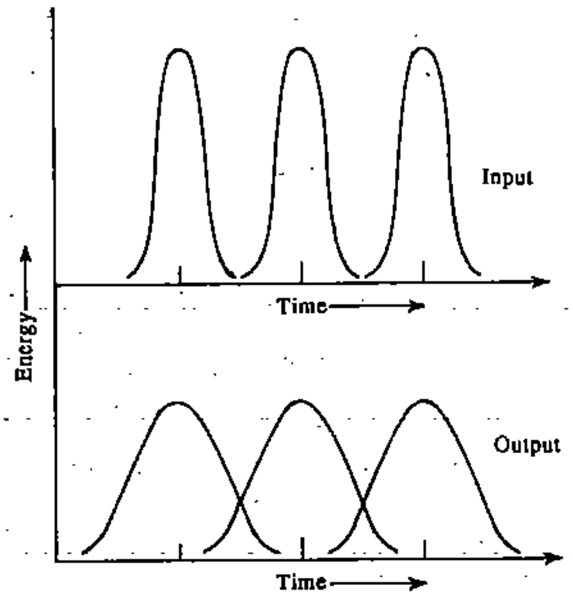


Fig.15.8: Pulse dispersion: (a) At the input, the information carrying pulses are well resolved. (b) At the output, due to broadening, pulses overlap and are unresolvable.

the fibre? Well, pulse broadening severely restricts the transmission capacity of the fibre. It is so because the pulses which are well resolved (Fig. 15.8a) at the input may overlap at the output (Fig. 15.8b) due to pulse broadening. To avoid this overlap, the time delay between two consecutive pulses must be increased. Therefore, the number of pulses that can be transmitted per unit time through the fibre will go down, that is, the transmission capacity of the fibre will be reduced.

To have a quantitative idea about the pulse dispersion in case of propagation through step - index fibre, refer to Fig. 15.9. Let a ray of light be incident at an angle $i$ with the axis of the fibre. The time taken by this ray to travel a distance $PR$

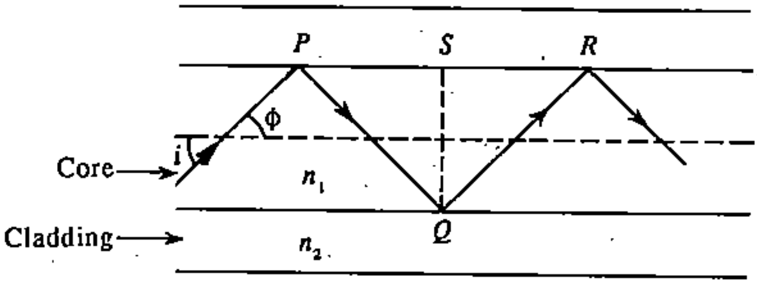


Fig.15.9: Ray of light passing through step - index fibre.

$$t = \frac{PQ + QR}{c/n_1} \quad \text{where } c/n_1 = \text{ velocity of light in the core medium (refractive index n}_1)$$

$$= \frac{n_1}{c}\,\frac{1}{\cos\varphi}\,(PS + QR)$$

$$= \frac{n_1\,(PR)}{c\cos\varphi}$$

What does this relation indicate? It indicates that the time taken by the ray of light in travelling a distance through the fibre depends on the angle it makes with the axis of the fibre. Thus, for a fixed length $L$ of the fibre, minimum time will be taken by a ray which travel along the axis of the fibre ($\phi = 0$) i.e.

$$t_{\min} = n_1 L/c$$

and the maximum time will be taken by the ray for which $\phi$ is equal to $(\pi/2 - \theta_c)$; where $\theta_c$ is the critical angle at the core - cladding interface. Thus, $\phi = \cos^{-1}(n_2/n_1)$ and the maximum time

$$t_{\max} = \frac{n_1 L}{c\,(n_2/n_1)} = \frac{n_1^2 L}{c\,n_2}$$

Thus, if all the input rays travel along the fibre simultaneously, the spread in time in traversing a distance $L$ will be

$$\Delta t = t_{\max} - t_{\min}$$

$$= \frac{n_1 L}{c\,n_2}\,(n_1 - n_2) \qquad (15.5)$$

**SAQ 2**

*Spend 3 min*

If the core and cladding refractive indices for a step - index fibre is 1.47 and 1.46 respectively, what will be the broadening of a pulse after a distance of 5 km?

Due to the pulse dispersion represented by Eq. (15.5), the signal transmission capacity of optical fibres is severly restrained. Therefore, an efficient optical fibre should have least possible pulse dispersion so that it can carry larger number of pulses per unit time.

Now the question is: Do we have any method to minimize the pulse broadening in optical fibres? Yes, there are methods by which we may minimize the pulse broadening. One of them is to use gradient - index (GRIN) fibre. In the following, you will learn how GRIN-fibres help in reducing the pulse broadening.

### 15.3.2 Pulse dispersion: GRIN Fibres

You may recall form Sec. 15.2 that core of the GRIN-fibre offers gradually decreasing refractive index environment to light rays as it moves away from the axis of the fibre. Let us see how this parabolic refractive index profile of the GRIN - fibre (Fig.15.5(a)) helps in reducing the pulse dispersion. Refer to Fig.15.10 in which two rays $A$ and $B$ are shown to enter the core axis at differnt angles. As the rays move towards the core - cladding interface, they encounter decreasing refractive index environment. As a result, both of them will bend away from the normal and hence towards the axis of the core. The paths taken by rays are not straight lines as in the case of step,- index fibre; rather, it is sinusoidal. It is because in the core, refractive index is a continuously decreasing function of the core radius. Now, ray $A$ which makes the smaller angle with the axis travels smaller distance through the core whereas ray $B$ travels a longer distance. However, the time taken by both of them, seperately, in traversing a fixed distance along the fibre will be same. Can you say why? It is so because ray $A$ which travels a shorther distance, does so in the region of higher refractive index. Hence the velocity of light along the path taken by ray $A$ will be smaller (velocity of light $= c/n$). On the other hand, ray $B$ which travels a relativly larger distance, does so in the region of lower refractive index and hence with higher velocity. The net result is that the rays making different angles with the core axis take equal time in propagating through the fibre. Due to this reason, the pulse broadening is reduced in GRIN-fibre.

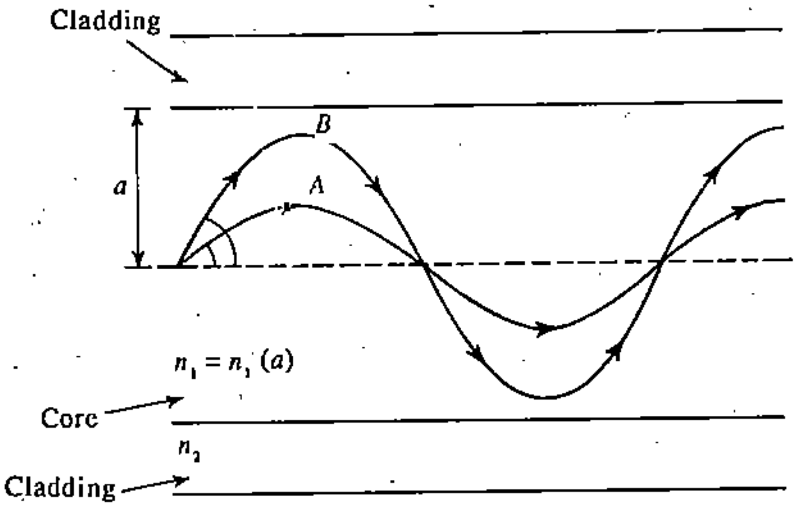


Fig.15.10: Two rays A and B travelling through a GRIN- fibre.

The volume of information which may be transmitted through GRIN - fibre is more or less free from pulse broadening due to above reason. The information carrying capacity of such fibre is only limited by material dispersion about which you will learn in the following.

### 15.3.3 Material Dispersion

Above we discussed about the pulse broadening in optical fibres arising because of the fact that light rays incident at different angles at the core-cladding interface take different times to traverse a fixed length of fibre. We also discussed how to reduce this dispersion by using GRIN-fibre. Now, suppose that the light beam travelling through the fibre is free from the pulse broadening due to above mentioned reason. Does it mean that the beam is free from pulse broadening? No, there is yet another source of pulse broadening known as material dispersion. Material dispersion arises due to the variation of refractive index with wavelength, i.e. the velocity of light in the medium is dependent upon its wavelength. You are aware that light even from a highly pure source (like laser which give highly monochromatic light) will have a spread in its wavelength. Therefore, different wavelengths, with in the range, will travel with different velocities and hence will arrive at the end of the fibre at different times and cause broadening of the pulse. You may note that the material broadening is an intrinsic physical property of the fibre material.

Although glass is transparent to electromagnetic radiation in the visible range, it does absorb a part of it due to several processes. As a result, the input power of the light beam will suffer a loss while traversing the length of the fibre. In the following, we briefly discuss some of these processes causing power loss in fibres.

### 15.3.4 Power Loss

When electromagnetic radiation interacts with matter, it may lose energy via different mechanisms. In case of optical fibre material, silica, major loss in energy or power is caused due to absorption of photons by impurity atoms. Therefore, to minimize this loss, the fibre material should be of high purity. Secondly, the photons may also lose energy by exciting the atoms of oxygen and silicon (the building blocks of silica, $SiO_2$). Thirdly, silica being amorphous material, it offers randomely varying refractive index. Due to this, the propagating light beam may get scattered and its direction of propagating may change drastically. These loss causing mechanisms are taken care of by proper design and synthesis technique of the fibres.

The power loss we are talking about is expressed in terms of bel or decibel which are comparable units. One bel means that power in one channel or at one time is 10 times that in another channel, or at another time. 2 bel means 100x, 3 bel means 1000x and so on. For practical use, the unit bel is too large. Hence the decibel, dB, is used. 1 bel = 10 dB. A **decibel** (dB) is equal to $10\log_{10}(p_2/p_1)$ where $p_1$ and $p_2$ are input and output powere levels. Thus, if the power level of an optical signal reduces by half, the power loss in decibels will be $10\log_{10}(1/2) = -3$dB. In optical fibre communication, the power loss is expressed as dBkm$^{-1}$. In long distance optical communication through fibres, the permissible loss is 20dBkm$^{-1}$. With modern techniques of synthesis, optical fibres with power loss as low as $\sim 0.2$dBkm$^{-1}$ can be produced.

## 15.4 SUMMARY

- An optical fibre consists of a transparent glass core and a cladding of lower refractive index. Since the refractive index of the cladding material is lower than that of the core, much of the light launched into one end will emerge from the other end due to a large number of total internal reflections.
- In the step-index fibre, the refractive index changes suddenly at the core-cladding interface. On the other hand, in the gradient-index (GRIN-) fibres, the refractive index decreases continuously from the core axis as a function of radius.
- The maximum entrance core angle, also known as acceptance angle, is a measure of the light gathering capacity of the fibre and is given as

$$\sin i_{\max} = \frac{1}{n_0}\left[n_1^2 - n_2^2\right]^{1/2}$$

The term $\left(n_1^2 - n_2^2\right)^{1/2}$ is known as the **numerical aperture** of the fibre.

- In optical communication, information is transmitted in the form of pulses. While travelling through the fibres, these pulses broaden because rays incident at different angles at the core- cladding interface take different times in traversing a fixed length of the fibre. Pulse broadening due to this reason in a step-index fibre of length $L$ is given as,

$$\Delta t = \frac{n_1 L}{c\, n_2} [\, n_1 - n_2 \,]$$

- Pulse broadening can be greatly reduced if, instead of step-index fibre, we use a GRIN- fibre. It is so because in GRIN-fibre, though different rays traverse different optical paths in the core, they all take same time in travelling through a given length of the fibre.
- Material dispersion is yet another cause of pulse broadening. Material dispersion arises because the refractive index (and hence the velocity of light) a medium is a function of wavelength of light. And, even highly monochromatic light has a spread in its wavelength.

## 15.5 TERMINAL QUESTIONS

1. Suppose you have two optical fibres $A$ and $B$. The refractive indices of the core ($n_1$) and the cladding ($n_2$) materials is

   $(n_1)_A = 1.52$, $(n_2)_A = 1.41$, $(n_1)_B = 1.53$, $(n_2)_B = 1.39$

   Which of the two fibres will have higher light gathering capacity?

2. A step-index fibre $6.35 \times 10^{-5}$m in diameter has a core of refractive index 1.53 and a cladding of refractive index 1.39. Determine (a) the numerical aperture for the fibre; (b) the acceptance angle (or maximum entrance cone angle).

## 15.6 SOLUTIONS AND ANSWERS

**SAQs**

1. If the refractive index of the cladding material is higher than that of the core material of the fibre, the incoming light will not undergo total internal reflection. It is so because when the light travels from a rarer to denser medium, it bends towards the surface normal. Thus, the light ray incident on the core-cladding interface will, instead of coming inside the core, get refracted in the cladding material (refer to Fig. 15.2).

2. The pulse broadening is given as

$$\Delta t = \frac{n_1 L}{c\, n_2} (n_1 - n_2)$$

As per the problem,

$L = 5 \times 10^3$m, $n_1 = 1.47$, $n_2 = 1.46$ and $c = 3 \times 10^8$ms $^{-1}$

So,

$$\Delta t = \frac{1.47 \times 5 \times 10^3\,(\text{m})}{3 \times 10^8\,(\text{ms}^{-1}) \times 1.46}\,(1.47 - 1.46)$$

$$= \frac{7.35 \times 10^3(\text{m})}{4.38 \times 10^8\,(\text{ms}^{-1})}\,(0.01)$$

$$= 0.17\,\mu\text{s}$$

**TQs**

1. Refer to Fig. 15.3. The maximum angle of incidence, $i_{max}$, of the light beam at air-core interface is the measure of the light gathering capacity of the fibre. The sine of this angle of incidence is given as

$$\sin i_{max} = \frac{1}{n_0}\left[n_1^2 - n_2^2\right]^{1/2}$$

where, $n_o$, $n_1$ and $n_2$ are the refractive indices of air, core and cladding respectively.

$n_o$ = refractive index of air = 1.

For the fibre $A$,

$$n_1 = 1.52 \text{ and } n_2 = 1.41$$

$$\sin i_{max} = [(1.52)^2 - (1.41)^2]^{1/2}$$

$$= [0.3223]^{1/2}$$

$$(i_{max})_A = \sin^{-1}[0.57] \cong 35^\circ$$

For the fibre $B$,

$$n_1 = 1.53 \text{ and } n_2 = 1.39$$

$$\sin i_{max} = [(1.53)^2 - (1.39)^2]^{1/2}$$

$$(i_{max})_B = \sin^{-1}[0.64] \cong 40^\circ$$

Hence, the light gathering capacity of fibre $B$ is greater than fibre $A$.

2.a) The numerical aperture of the fibre is given as,

$$\text{N.A.} = [n_1 - n_2]^{1/2}$$

$$= [(1.53)^2 - (1.39)^2]^{1/2}$$

$$= 0.64$$

b) The acceptance angle or the maximum entrance angle, $i_{max}$, corresponds to $\theta_c$, the critical angle for total internal reflection at the core-cladding interface.

$$\sin i_{max} = \frac{1}{n_0}[\text{N.A.}]$$

$$= 0.64$$

$$\Rightarrow i_{max} = \sin^{-1}[0.64]$$

$$\cong 40^\circ$$