ALAGAPPA UNIVERSITY

[Accredited with 'A+' Grade by NAAC (CGPA:3.64) in the Third Cycle and Graded as Category–I University by MHRD-UGC]

(A State University Established by the Government of Tamil Nadu)

KARAIKUDI – 630 003

Directorate of Distance Education

M.A. [Economics]

I - Semester

362 11

MICROECONOMICS - I

Reviewer	
Dr R. Suresh	Assistant Professor, Department of Economics and Rural Development, Alagappa University, Karaikudi

Authors:

Dr. D N Dwivedi, *Professor of Economics, Maharaja Agrasen Institute of Management Studies, Delhi*
Units (1.0-1.4, 2, 3.0-3.5, 6, 7.0-7.4, 8-10, 11.0-11.2.1, 12-14)

Dr. Suman Lata, *Lecturer, Ginni Devi Modi Girls College, Modinagar, Ghaziabad (UP)*
Units (4, 11.3)

Dr. Punithavathy Pandian, *Professor in the Department of Commerce, Madurai Kamaraj University, Madurai*
Unit (5)

MC Vaish, *Former Professor and Head, Department of Economics, University of Rajasthan, Jaipur*
Unit (7.6.2)

Vikas® Publishing House
Units (1.5-1.9, 3.6-3.10, 7.5-7.6.1, 7.7-7.11, 11.4-11.8)

SYLLABI-BOOK MAPPING TABLE

Microeconomics - I

CONTENTS

INTRODUCTION

Economics is fundamentally the study of choice-making behaviour of people. This choice-making behaviour is studied in a systematic or scientific manner, thereby giving economics the status of a social science. Although the study of economics is old, its scope has expanded vastly in the post-Second World War period. Modern economics is now divided into two major branches: microeconomics and macroeconomics. In this book, we will be focusing on microeconomics. Microeconomics is concerned with the microscopic study of

the various elements of the economic system and not with the system as a whole. As Russian economist Abba P. Lerner has put it, 'microeconomics consists of looking at the economy through a microscope, as it were, to see how the millions of cells in body economic—the individuals or households as consumers and the individuals or firms as producers—play their part in the working of the whole economic organism.'

This book, *Microeconomics-I* has been divided into fourteen units. The book has been written in keeping with the self-instructional mode or the SIM format wherein each Unit begins with an Introduction to the topic, followed by an outline of the Objectives. The detailed content is then presented in a simple and organized manner, interspersed with Check Your Progress questions to test the student's understanding of the topics covered. A Summary along with a list of Key Words, set of Self-Assessment Questions and Exercises and Further Readings is provided at the end of each Unit for effective recapitulation.

BLOCK - I

ANALYSIS OF MICROECONOMICS

UNIT 1 INTRODUCTION TO MICROECONOMICS

Structure

1.0 INTRODUCTION

Economics as a social science has two major branches-—*microeconomics* and *macroeconomics. Microeconomics* is the study of the economizing behaviour of the individual economic entities—individuals, households, firms, industries and factory owners. For example, *microeconomics* studies how *individuals* and *households* with limited income decide 'what to consume' and 'how much to consume' so that their total utility is maximized. In other words, microeconomics studies how individual consumers make choice of goods and services they want to consume and how they allocate their limited income between the goods and services of their choice to maximize their total economic welfare.

Similarly, in case of *business firms,* microeconomics studies how business firms take decision on what to produce, for whom to produce, how to produce, how much to produce, what price to charge, how to fight the competition, and how to promote sales to maximize their profit. In addition, microeconomics studies how demand for a product and its supply are affected by the change in price of the product and how price of a product is determined in the market. Obviously, decision making process in case business firms is a complex affair. Therefore, making a reasonable business decision requires a good knowledge and understanding of market conditions in respect of demand and supply conditions, cost of production, pricing system, nature and degree of competition, conditions in the financial market,

and so on. This requires an extensive and intensive analysis of the market conditions. This is subject matter of *microeconomics*.

Macroeconomics, on the other hand, studies the economic phenomena at the national aggregate level. Specifically, macroeconomics is the study of working and performance of the economy as a whole. It studies what factors and forces determine the level of national output or national income, rate of economic growth, employment, price level, and economic welfare. Besides, macroeconomics studies how government of a country formulates its macroeconomic policies—taxation and public expenditure policies (the fiscal policy), monetary policy, price policy, employment policy, foreign trade policy, etc., to resolve the problems of the country. Macroeconomics studies how these policies affect the economy. From *managerial economics* point of view, the study of macroeconomics gives basis for judging the *economic environment* of the country.

1.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Define microeconomics
- Interpret the scope of microeconomics
- Analyse economics as a positive science
- Discuss the Cobweb theorem

1.2 SCOPE OF MICROECONOMICS

Internal managerial issues refer to decision-making issues arising in the management of the firm. Internal managerial issues include problems that arise in operating the business organization. All such managerial issues fall within the purview and the control of the managers. Some of the basic internal management issues can be listed as follows.

- What to produce—choice of the business
- How much to produce—determining the size of the firm
- How to produce choice of efficient and affordable technology
- How to price the product—determining the price of the product
- How to promote sale of the product
- How to face price competition from the competing firms
- How to enlarge the scale of production—planning new investment
- How to manage profit and capital.

The microeconomic theories and tools of analysis that provide a logical basis and ways and means to find a reasonable solution to business problems constitute the

microeconomic scope of managerial economics. The main microeconomic theories that fall within the scope of managerial economics are following.

Theory of consumer demand: Theory of consumer demand analyses the decision--making behaviour of the consumers. The decision-making behaviour of the consumer relates to such questions as: how consumers decide what to consume; how much to consume; how much to buy, and how consumers react to change in price of the products they consume and price of their substitutes. Demand theory combined with quantitative tools helps in assessing the total demand for a product at different prices. Thus, the consumer demand theory helps in deciding 'what to produce'.

Theory of production: Theory of production analyses the nature of input-output relationship. It explains how output changes with change in inputs—labour or capital—given the technology. It provides guidance in the choice of technology and in maximizing the output from the resources of the firm. Thus, the knowledge and application of the theory of production helps in determining the optimum level of production, the size of the firm, and the employment of labour and capital.

Theory of cost: Theory of cost analysis the nature and pattern of change in cost of production with change in output. Specifically, theory of cost reveals the change in total, marginal and average cost of production. Application of cost theory helps in knowing the cost behaviour with increase in production and in determining the output that minimizes the average cost production. In view of profit-maximization objective, this theory helps in determining the profit maximizing output, give the price of the product.

Theory of price determination: The theory of price determination offers an analysis of how price is determined under different kinds of market conditions. Market conditions are determined on the basis of degree of competition between the firms of the industry—perfect competition, monopolistic competition, oligopoly and monopoly. The theory of price determination combined with the cost theory helps firms in determining the profit maximizing price of their product.

Theory of capital and investment decisions: Capital is the foundation of business firms. An efficient management of capital is one of the most important functions of the managers as it is the determinant of the success of the firm. The major issues in capital management are (*i*) the choice of investment avenues, (*ii*) assessing the efficiency and productivity of capital investment avenues, and (*iii*) making the choice of most efficient investment project. The theory of capital contributes a great deal in making appropriate investment decisions.

1.3 ECONOMICS AS A POSITIVE SCIENCE

Economics as a positive science seeks to analyze systematically and explain economic phenomena as they actually happen; find common characteristics of economic events; brings out the 'cause and effect' relationship between the economic

variables, if any; and generalizes this relationship in the form of a theoretical proposition. One of the main purposes of economic studies is 'to provide a system of generalization' in the form of economic theories that can be used to make predictions about the future course of related events. It means that economics has a positive character. Economics explains the economic behaviour of individual decision-makers under given conditions; their response to change in economic conditions; and it brings out the relationship between the change in economic conditions and economic decision of the people. In fact, *the main function of economics is to establish cause-and-effect relationship, if there is any, between two or more economic events and to provide basis for prediction.* Emphasizing the positive character of economics, Friedman says, "Economics as a positive science is a body of tentatively accepted generalizations about economic phenomena that can be used to predict the consequences of change in circumstances." One of the main tasks of economics is 'to provide a system of generalizations' or, more precisely, economic theories, capable of being used to predict economic phenomena. This makes economics a *positive science.* Here, 'positive' does not mean that theoretical statements are *positively* true: it means that it has a great possibility to occur if conditions are fulfilled.

1.3.1 Criteria for Choosing among Alternative Theories

When we take a decision, there are multiple options from which we can select the desired one. Among the alternatives available, we like to select the one (i) which has the highest probability of success or efficiency (ii) best fits our goals, desires, lifestyle and so forth. The primary important idea here is that first, there must be some genuine alternatives to choose from. Second, every decision must be taken while judging against some standard of judgement. This standard generally finds expression in the form of criteria, which conveys the values and preferences of the decision-maker. Also, these values and preferences are further influenced by the societal practices, corporate culture and so forth.

1.4 DYNAMIC ECONOMIC ANALYSIS AND COBWEB THEOREM

Economists analyse stability and instability of market equilibrium under *static* and *dynamic conditions.* In this section, we discuss the stability and instability of market equilibrium under *static* and *dynamic* conditions, called *static* and *dynamic equilibrium.*

Static and Dynamic Equilibria

Static and Dynamic Economy: The concepts of static and dynamic equilibria are, generally, used as an analytical tool rather than as a realistic situation. The concept of static equilibrium is used to analyse a static economy. A static economy is one that remain in the state of motionlessness. The concept of dynamic equilibrium

is used to analyse equilibrium of a dynamic economy—an economy in which changes take place continuously. Let us, therefore, first know what is a static and what is a dynamic economy.

A **static economy** is one in which nothing is changing. It is like studying a "photograph" or a firm with constant inputs and output. To analyse a static economy, it is assumed that nothing is changing in the economy; the quantity of goods and services that are produced and consumed remain unchanged; the types and number of economic organisations remain the same; people's taste and preferences do not change; all prices are constant; and the output and employment are fixed—all at a given point of time. The equilibrium of an economy under these static conditions is called *static equilibrium.*

Close to static equilibrium is the concept of *stationary equilibrium.* When an economy remains in static equilibrium between any two points of time, it is said to be *stationary.* If the economy today is exactly what it was yesterday or over any period in the past, it is in the stationary state "in which nothing ever happens".

A **dynamic economy** may be defined as one which is changing continuously. The change may be autonomous or induced. A dynamic economy is in the process of continuous change. Such an economy passes through different stages of equilibrium and disequilibrium. When a dynamic economy is in the state of equilibrium, it is said to be *dynamic equilibrium.* The equilibrium may be stable, unstable or neutral under both static and dynamic conditions. This aspect is explained below.

Equilibrium under Static Conditions

Equilibrium under static conditions may be *stable* or *unstable*. In this section, we illustrate the conditions for stable and unstable equilibrium.

(i) **Stable Equilibrium under Static Conditions.** In static analysis, all economic variables bear the same time subscripts, i.e., all variables refer to the same point of time. There is no time-lag in the adjustment between the variables when a change takes place. Let us now see how equilibrium becomes stable under static conditions.

To illustrate *stability of equilibrium under static conditions,* let us consider a simple competitive market in which price and output are determined by a downward sloping demand curve and an upward sloping supply curve. As shown in Fig. 1.1, equilibrium price OP_2 and output OQ_2 are determined at point $P,$ the point of inter-section between demand curve DD' and supply curve SS'. Under static conditions, this equilibrium position would remain *stable.* For, if equilibrium is disturbed due to a change in any of the variables (viz., demand, supply and price), other things remaining the same, it will itself set forces to restore the equilibrium to its initial position. For example, if price shoots up from equilibrium price OP_2 to OP_3, supply will exceed by $AB = Q_1Q_3$. The excess supply will cause price to decrease to the equilibrium

price. Similarly, if prices falls, from OP_2 to OP_1, demand will exceed supply causing price to go up to point P and remain stable.
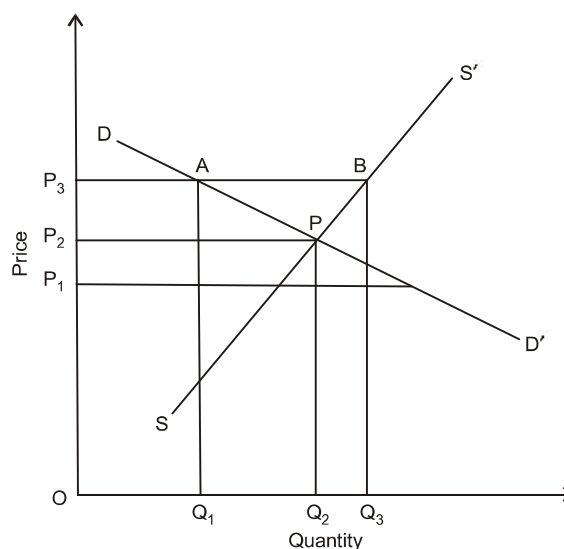
Fig. 1.1 *Stable Equilibrium*

(ii) Unstable Equilibrium under Static Conditions. Let us now look at the conditions of *unstable equilibrium under static conditions.* The cases of unstable equilibrium under static conditions, are extremely rare and temporary ones. Besides, the issue is not finally settled. Let us, consider a downward sloping demand curve and a *downward sloping* supply curve, as shown in Fig. 1.2. Note that the slope of supply curve is greater than that of the demand curve.

The negative slope of a supply curve indicates a peculiar behaviour of sellers in response to price change, that is, sellers sell less when price increases and sell more when price decreases. Such a behaviour of sellers may be attributed to *distress selling* by them. For example, suppose that a subsistence farmer capable of growing only one crop, sells only a part of his produce just sufficient to meet his non-food requirements. Given these conditions, he sells less when price is high and sells more when price is low so that supply curve has a negative slope. A similar supply curve may exist in the case of labourers, who prefer leisure to income and works only as many hours per day as necessary to earn a minimum income sufficient to meet their basic requirements. Under the conditions as shown in Fig. 1.2, if price increases for some reason and goes above the equilibrium price *PQ*, it goes on increasing without returning over to the equilibrium point *P*. So is the case with decline in price below the equilibrium price *PQ*. Price goes on decreasing without stabilizing ever.
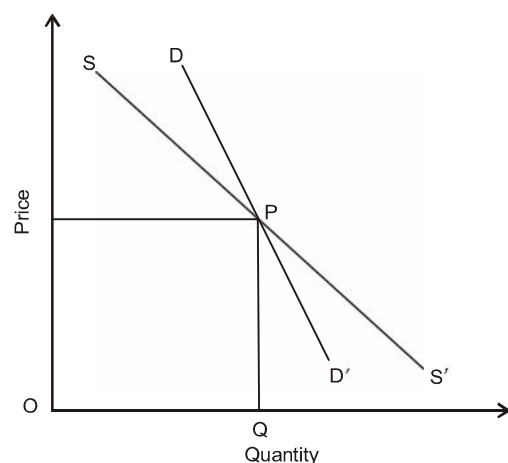
Fig. 1.2 *Unstable Equilibrium: Negative Slope of Supply Curve*

Equilibrium Under Dynamic Conditions: The Cobweb Theorem

Let us now discuss stability and instability of equilibrium under dynamic conditions. An important feature of economic dynamics is that it recognises the *time-lag* in the adjustment between the variables. Since we are considering (in our illustrations) the simple demand and supply theories under perfectly competitive conditions, we assume that *supply is a lagged function of price.* It implies that supply responds to change in price after a time-lag. With this basic condition in mind, we will now illustrate and discuss *stable* and *unstable equilibria* under dynamic conditions.

  (i) **Stable Equilibrium under Dynamic Conditions.**   As mentioned above, an equilibrium position is said to be stable if displacement of equilibrium itself sets forces of demand and supply into action that restore the initial equilibrium position. The simplest way to illustrate the stability and instability of equilibrium position under dynamic conditions is provided by 'Cobweb Theorem'. The name, 'Cobweb' Theorem, has been derived from the appearance of the diagram it produces.

  The Cobweb Theorem can be stated in the form of three theorems based on three different conditions.

   (*a*) **Theorem I:** If slope of demand curve ($\Delta P / \Delta Q$) is less than that of supply curve ($\Delta P / \Delta S$), the equilibrium is stable: the system is *convergent* to the original equilibrium.

   (*b*) **Theorem II:** If $\Delta P / \Delta D > \Delta P / \Delta S$ equilibrium is unstable. The adjustment process is *divergent* or *oscillatory* — demand, supply and price keep moving away and away from the original equilibrium point.

   (*c*) **Theorem III:** If $\Delta P / \Delta D = \Delta P / \Delta S,$ equilibrium is nondamped oscillating: it keeps circulating around the original equilibrium point with a constant change in price and quantity.

Figure 1.3 illustrates the *stable equilibrium,* the theorem I. As the figure shows, demand and supplycurves intersect each other at point *P* determining equilibrium price at OP_3 and equilibrium output at OQ_4. If price rises for some reason in, say period *t,* to OP_5, equili-brium will be disturbed. For, at the new price, demand falls to OQ_1 and supply is expected to rise to OQ_6 with a time-lag, of course. Note that de-mand is much less than the expected supply OQ_6 at this price. In period *t* + 1, supply rises to OQ_6 exceeding demand by Q_1Q_6. Consequently, price falls to OP_1 causing a rise in demand to OQ_6. But in response to fall in price, supply decreases in period *t* + 2 to OQ_2. The demand now exceeds supply by Q_2Q_6. Therefore, price rises to OP_4, causing an increase in supply by Q_2Q_5 in period *t* + 2. It is now the turn of price to adjust itself to the existing demand and supply conditions. This whole process is repeated period after period. Note that each time the process of adjustment is repeated, the magnitude of change in supply, price and demand goes on decreasing. For example, in period *t* + 1, supply increases by Q_1Q_6 in period *t* + 2 it decreases by Q_2Q_6, and in period *t* + 3 it increases by Q_2Q_5 that $Q_1Q_6 > Q_2Q_5 > Q_3Q_4$. So is the case with price and demand. As a result of decreasing magnitude in changes, the system in Fig. 1.3 converges to the original equilibrium point *P.* Thus, the equilibrium once displaced sets the forces which restore the original equilibrium position. The equilibrium position is therefore *stable.*
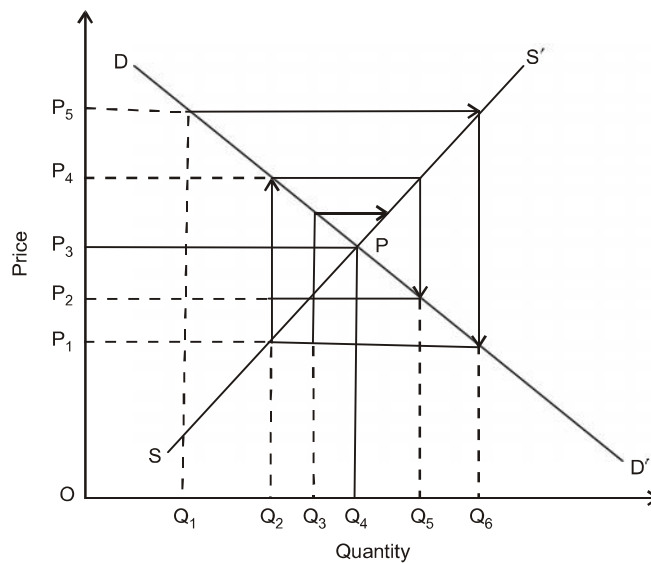
Fig. 1.3 *Stable Equilibrium*

(ii) Unstable Equilibrium under Dynamic Conditions. Let us now discuss the *unstable equilibrium,* i.e., theorem II. If slope of the supply curve is less than the slope of the demand curve, then the process of adjustment makes the price and quantity diverge away and away from the equilibrium position. In this case, the amplitude of fluctuations in price and output around

the equilibrium point, goes on increasing. The new equilibrium position is therefore unattainable. This case is illustrated in Fig. 1.4. Note that the slope of the supply curve, SS', is greater than that of the demand curves, DD_1 and DD_2, i.e., $\Delta P / \Delta S = \Delta P / \Delta D$. Suppose that the original equilibrium position is at point P and demand curve, DD_1 shift upward to DD_2. Owing to upward shift in demand curve, price in period, say t, rises from OP_2 to OP_3. Since there is a supply lag of one period, supply increases in period $t + 1$ from OQ_2 to OQ_3. Now, supply exceeds the existing demand, and hence price falls to OP_1. In the following period, $t + 2$, supply decreases by Q_1Q_3 ($= NM$). Now demand exceeds supply and, therefore, price rises to OP_4. Note that in each period of adjustment, the amplitude of rise and fall in price and quantity goes on increasing causing movement of price-quantity combination further and further away from the original equilibrium point, P. Therefore, equilibrium becomes *unstable*.

Fig. 1.4 Unstable Equilibrium: Dynamic Analysis

This process, however, cannot continue indefinitely. The explosive conditions force the producers to restrict the supply so that the slope of supply curve is reduced. When slope of the supply curve becomes less than that of the demand curve, price and output converge to a new equilibrium position (as shown in Fig. 1.3).

Check Your Progress

1. List the major issues in capital management.
2. What is a dynamic economy?

1.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The major issues in capital management are (*i*) the choice of investment avenues, (*ii*) assessing the efficiency and productivity of capital investment avenues, and (*iii*) making the choice of most efficient investment project. The theory of capital contributes a great deal in making appropriate investment decisions.

2. A dynamic economy may be defined as one which is changing continuously. The change may be autonomous or induced. A dynamic economy is in the process of continuous change.

1.6 SUMMARY

- Economics as a social science has two major branches—microeconomics and macroeconomics. Microeconomics is the study of the economizing behaviour of the individual economic entities—individuals, households, firms, industries and factory owners.

- Similarly, in case of business firms, microeconomics studies how business firms take decision on what to produce, for whom to produce, how to produce, how much to produce, what price to charge, how to fight the competition, and how to promote sales to maximize their profit.

- Macroeconomics, on the other hand, studies the economic phenomena at the national aggregate level. Specifically, macroeconomics is the study of working and performance of the economy as a whole.

- Internal managerial issues refer to decision-making issues arising in the management of the firm. Internal managerial issues include problems that arise in operating the business organization.

- Theory of consumer demand analyses the decision- making behaviour of the consumers.

- Theory of cost analysis the nature and pattern of change in cost of production with change in output. Specifically, theory of cost reveals the change in total, marginal and average cost of production.

- Economics as a positive science seeks to analyze systematically and explain economic phenomena as they actually happen; find common characteristics of economic events; brings out the 'cause and effect' relationship between the economic variables, if any; and generalizes this relationship in the form of a theoretical proposition.

- The concepts of static and dynamic equilibria are, generally, used as an analytical tool rather than as a realistic situation. The concept of static equilibrium is used to analyse a static economy.

- A static economy is one in which nothing is changing. It is like studying a "photograph" or a firm with constant inputs and output.

1.7   KEY WORDS

- **Macroeconomics:** It studies what factors and forces determine the level of national output or national income, rate of economic growth, employment, price level, and economic welfare.

- **Consumer demand:** It is defined as *demand* for goods and services that comes from individual people rather than from companies.

1.8   SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. Write a short note on the scope of microeconomics.
2. What is static equilibrium?
3. What is the criteria for choosing among alternative theories?

Long Answer Questions

1. Discuss the Cobweb theorem with the help of diagrams.
2. Compare and contrast the subject matter of macroeconomics and microeconomics.
3. Analyse the significance of the study of microeconomics.

1.9   FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.

UNIT 2 PARTIAL AND GENERAL EQUILIBRIUM ANALYSIS

Structure

2.0 INTRODUCTION

A fundamental feature of an economic system is the interdependence and interrelatedness of economic activities—production and consumption—of its various constituents—individuals, households, firms, banks and other kinds of financial institutions. The working mechanism of economic system is unimaginably complex. It is not possible to trace the behaviour of each economic element and its interaction with the rest of the economy and trace equilibrium of each and every element of the economy. The economists, therefore, adopt two kinds of approaches to economic analysis : (i) Partial equilibrium approach, and (ii) General equilibrium approach.

2.1 OBJECTIVES

After going through this unit, you will be able to:

- Differentiate between partial equilibrium and general equilibrium approach
- Discuss Walrasian approach to general equilibrium

2.2 PARTIAL AND GENERAL EQUILIBRIUM ANALYSIS: AN OVERVIEW

Partial equilibrium approach ignores the interdependence of the various segments of the economy. It isolates the segment or the phenomenon of the study from the

other segments and assumes non-existence of influences of the changes occurring outside the area delimited for the study. For example, in the analysis of utility-maximisation behaviour of the households, their incomes are assumed to remain constant even if incomes change due to change in factor prices in factor markets; prices of related goods (substitutes and complements) are assumed to remain constant even if they change due to change in demand and supply conditions; and the consumer's taste and performance are assumed to be given even if they are not. Similarly, in the analysis of profit maximising behaviour of the firms, the factor prices, technology, and commodity-prices are assumed to remain constant even if these variables continue to change.

The **general equilibrium approach**, on the other hand, recognises the interdependence of constituent parts of the economic system. It recognises the interrelations and interdependence of economic variables and seeks to answer the question how all the segments of the economy reach an equilibrium position simultaneously. General equilibrium shows, by using the tools of partial equilibrium analysis, how prices and outputs are simultaneously determined in all segments of the economy.

In this unit, we present a simple analysis of general equilibrium system using a simplified general equilibrium model. Let us begin with a brief description of the general equilibrium system.

2.2.1 General Equilibrium Approach

We have noted that various parts of the economy are mutually interdependent and function in close relationship with each other. In fact, in an economy everything depends on everything else. In such a system, price of a single commodity or factor cannot, in principal, be determined unless all other prices are known. Furthermore, prices are not determined one by one. If at all, they are determined simultaneously. The general equilibrium approach seeks to answer such questions as: Does the market mechanism produce a general equilibrium solution wherein each market or segment of the economy is in equilibrium? Is the equilibrium in product markets necessarily consistent with the equilibrium in factor markets? Is the behaviour of each consumer consistent with that of every other consumer, with that of every producer, and with that of each factor supplier? If so, is this solution unique, or are there several other set of prices that will satisfy an equilibrium solution? In other words, does there exist a unique equilibrium solution? Even if it exists, will it be stable in the sense that a disturbance which causes a departure from equilibrium sets up automatic forces that bring the system back again to equilibrium?

Thus, the task of general equilibrium theory is to find out whether there exists a general equilibrium in an economy. A general equilibrium is defined as a state in which all economic units maximise their respective objective function and all prices are simultaneously in equilibrium, and all markets are cleared. General equilibrium theory explains how this state can, if ever, be attained. If attained, whether it remains stable.

Leon Walras (1834–1910), a French economist, was the first to attempt to answer these questions in his book *Elements of Pure Economics* (1874). Although long before Walras, Cournot had realised that "for a complete and precise solution of the partial problems of the economic system, it is inevitable that one must consider the system as a whole." In their opinion, the problem of general equilibrium was beyond the resources of mathematical analysis. However, Walras showed, by using a system of simultaneous equations, that all prices and quantities in all markets are simultaneously determined through their interaction with each other. In this unit, we present a simple model of Walrasian system.

2.2.2 The Walrasian Approach to General Equilibrium

In the Walrasian system of general equilibrium, the behaviour of each decision-maker is presented by a set of equations. Since each decision-maker functions simultaneously in two different capacities—as a buyer and as a seller, his behavioural equations consists of two subsets of equations. One subset describes his demand for different commodities (or factors); it contains as many equations as the number of commodities (or factors) supplied. Thus, demand side of the commodity market is described by as many equations as the number of commodities multiplied by the number of consumers demanding the commodities. Similarly, supply side of the market is described by as many equations as the number of commodities multiplied by the number of firms supplying the commodities. Factor market is similarly described, in Walrasian model, by two sets of equations—one each on demand and supply sides. In this system of describing working of an economy through equations, there are as many 'unknown' variables to be determined as independent equations. The 'unknowns' are the quantities of all commodities and factors purchased and sold by each individual, and prices of all commodities and factors.

To illustrate Walrasian system, let us consider a simple two-consumer-two-commodity-two-factor model for general equilibrium analysis. Assume that there are only two consumers, *A* and *B*; only two commodities, *X* and *Y* and only two factors, *K* and *L*. Assume also that factors *K* and *L* are owned by the consumers, and commodities *X* and *Y* are produced by the two firms. Let us now specify the number of equations and of 'unknowns', assuming the existence of perfect competition in both commodity and factor markets.

Note that the number of equations (20) is the same as the number of unknowns (20). It is necessary, though not sufficient, condition for the general equilibrium solution that the number of independent equations must be the same as the number of unknowns. Another requirement of general equilibrium solution is that all equations must be simultaneously solved. The above example satisfies this condition of general equilibrium solution. The fulfillment of this condition however does not necessarily guarantee the existence of a general equilibrium solution. To this point, we will return later. First let us formally describe the Walrasian general equilibrium model.

No. of equations		No. of unkowns	
1. Demand functions of two goods by two consumers,	$2 \times 2 = 4$	1. Quantities of 2 goods demanded by 2 consumers	$2 \times 2 = 4$
2. Supply functions of two goods by two firms	$2 \times 2 = 4$	2. Quantities of 2 goods supplied by 2 firms	$2 \times 2 = 4$
3. Demand functions of two factors by two firms	$2 \times 2 = 4$	3. Quantities of 2 factors demanded by 2 firms	$2 \times 2 = 4$
4. Supply functions of two factors by 2 suppliers (*A* and *B*)	$2 \times 2 = 4$	4. Quantities of 2 factors supplied by 2 firms	$2 \times 2 = 4$
5. Market clearing equations of commodities	2	5. Prices of 2 commodities	2
6. Market clearing equation of factors	2	6. Price of 2 factors	2
Total No. of Equations	20	Total no. of unknowns	20

The Walrasian General Equilibrium Model

This section presents a formal summary of Walrasian general equilibrium model. Let us suppose that an economy has n commodities, h households (or individuals) and m inputs (or factors) and describe the commodity and input sectors.

Commodity Sector. The demand for each commodity is expressed by a demand function which depends on prices of all commodities, $P_1, P_2, \ldots, P_n$, and on the level and distribution of consumer incomes $M_1, M_2, \ldots, M_n$, which consumers earn by supplying their factor services. Thus, the demand function for each commodity may be expressed as

$$Q_i^d = D_i (P_1, P_2, \ldots, P_n, M_1, M_2, \ldots, M_n) \qquad \ldots(2.1)$$

There are $n \times h$ demand functions in the general system. The supply of each commodity is similarly expressed through supply functions. The quantity supplied of a commodity depends on the prices of all commodities, $P_1, P_2, \ldots, P_n$, and prices of all inputs $V_1, V_2, \ldots, V_3$. Thus, supply function is given as

$$Q_i^s = S_i (P_1, P_2, \ldots, P_n, V_1, V_2, \ldots, V_3) \qquad \ldots(2.2)$$

There are $n \times f$ supply functions of n commodities for f firms.

Input Sector. Resources (or inputs) are owned and supplied by the households and demanded by firms. Let R represent the amount of resource K owned by an individual i. The actual amount supplied R_k of a resource K will depend on all input prices and the level and distribution of ownership. Thus supply function of a resource is given as

$$R_k^s = S_k (V_1, V_2, \ldots, V_n; R_{k1} R_{k2}, \ldots, R_{kn}) \qquad \ldots(2.3)$$

There will be $m \times h$ equations.

The actual amount demanded (R_k^d) of each resource will depend on output levels, output prices, and input prices. Thus,

$$R_k^d = D_k (Q_1, Q_2, \ldots, Q_n; P_1, P_2, \ldots, P_n;$$
$$V_1, V_2 \ldots V_n) \qquad \ldots(2.4)$$

where $Q_1, Q_2, \ldots, Q_n$ represent output levels.

There will be $n \times m$ equations.

Besides, resource constraints should also be incorporated into the model. It may be expressed as

$$\sum_{i=1}^{n} \overline{R}_k^d \leq R_{k_1} \qquad \qquad ...(2.5)$$

Identities. An important identity which emerges from the circular flows of incomes is that values of all outputs, i.e., P_iQ_i must equal the total income of the society, i.e., $M_1 + M_2, ..., M_h$. That is,

$$\sum_{i=1}^{n} P_iQ_i = \sum_{j=1}^{h} M_j \qquad \qquad ...(2.6)$$

Secondly, the total expenditure equals total income. Income of each individual is calculated by multiplying the amount of resource K supplied by an individual j, which equals $\overline{R}_{kj}^s$ time the resource price V_k. That is,

$$M_j = \sum_{k=1}^{m} V_k \overline{R}_{kj}^s \qquad \qquad ...(2.7)$$

Finally, the fundamental identity for the economy as a whole can thus be expressed as

$$\sum_{i=1}^{n} P_iQ_i = \sum_{j=1}^{h} \sum_{k=1}^{m} V_k \overline{R}_{kj}^s \qquad \qquad ...(2.8)$$

Equation (2.8) shows that the prices of resources are directly linked to the prices of output. Prices and quantities of resources supplied cannot be determined without determining the price of commodities. The Walrasian model therefore requires that all the equations must be solved simultaneously. A general equilibrium occurs at $n + m$ prices when all the equations are simultaneously solved.

Graphical Illustration of Tendency Towards General Equilibrium

Assuming a $2 \times 2 \times 2$ model, we show in this section that the model economy has a tendency towards general equilibrium under the following assumptions.

Assumptions

1. There exists perfect competition in both commodity and factor markets.

2. There are only two commodities, *X* and *Y,* which are substitutes for each other, and two firms produce one commodity each.

3. Consumers' utility functions are given and they maximise their utility subject to income constraint.

4. There are only two factors of production, *L* and *K,* which are available in fixed supply. Factors are homogeneous and perfectly divisible.

5. Production functions show diminishing marginal rate of technical substitution (MRTS) and decreasing returns.

6. Firms maximise their profits subject to resource constraint.

To begin with, let us assume that both commodity and factor markets are in equilibrium. Prices in both the markets are in equilibrium. Demands for commodities, *X* and *Y*, are equal to their respective supplies. Similarly, demand for each factor is equal to its supply.
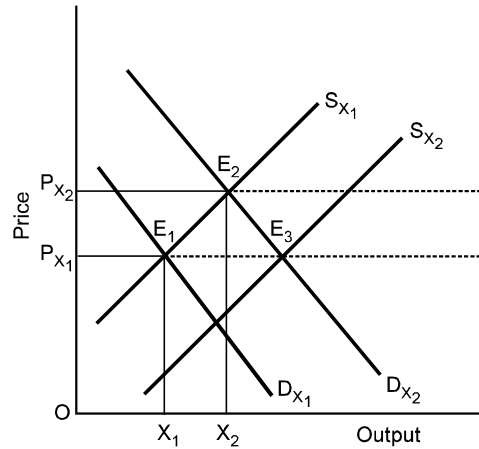
Fig. 2.1 *Market for Commodity X*

The equilibrium in commodity *X* market is illustrated in Fig. 2.1. The initial demand and supply curves for commodity *X* are represented by D_{x1} and S_{x1} respectively. The demand and supply curves intersect at point E_1 determining price of *X* at OP_{x1}. At this price, demand for X (i.e., OX_1) equals its supply. The market for commodity *X* being in equilibrium, the one-firm industry *X* would also be in equilibrium. The equilibrium of firm *X* is illustrated in Fig. 2.2. The firm (or industry) produces OX_1 at which $AC = MC = \text{Price} = MR$.
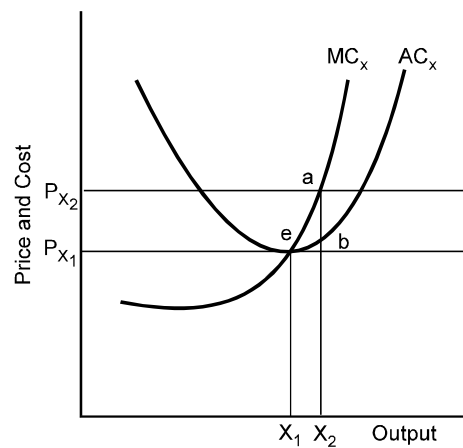
Fig. 2.2 *Industry* X

Similarly, the initial equilibrium positions of commodity market Y and of the firm producing Y are illustrated in Figs. 2.3 and 2.4, respectively. The commodity market Y is in equilibrium at price OP_{y1} at which demand for Y equals its supply, OY_1. Industry Y is in equilibrium at output OY_1. At this output, $AC = MC =$ Price $= MR$ in industry Y.
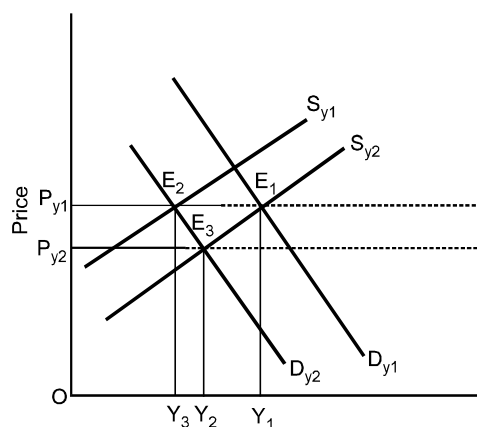
Fig. 2.3 *Market for Commodity* Y

Let us now suppose that, due to some exogenenous factor, consumers' taste changes in favour of commodity X. As a result, demand curve for X, i.e., D_{x1} shifts upward to the position of D_{x2} (see Fig. 2.1). Consequently, price of X rises from P_{x1} to P_{x2}. The output of X rises to OX_2 and the industry makes an abnormal profit of ab per unit of output (see Fig. 2.2). The supernormal profits attracts firms from industry Y to industry X and the existing ones increase their output. As a result, demand for factors increases. This causes a rise in demand for factors L and K, in industry X. Since factors are fully employed, where do the factors come from? To find an answer to this question, let us examine what is happening in industry Y.
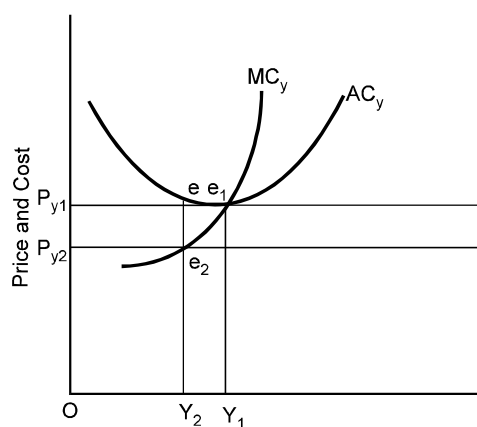
Fig. 2.4 *Industry* Y

Since we have assumed a shift in consumer's taste other things remaining the same, the additional demand for *X* comes only from a shift in demand from *Y* to *X*. This shift occurs because *X* and *Y* are substitutes for each other. Due to shift in demand from *Y* to *X*, the initial demand curve for *Y,* i.e., D_{y1} shifts downward to the position of D_{y2}. Output of *Y* falls to OY_2 and price falls from OP_{y1} to OP_{y2}. (Fig. 2.3). As a result, the equilibrium of industry *Y* shifts from E_1 to E_2 and firms incur a loss of ee_2 per unit (Fig. 2.4).
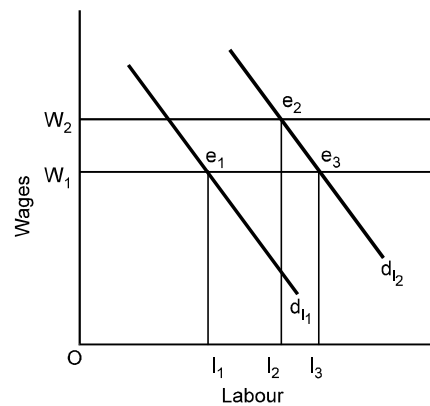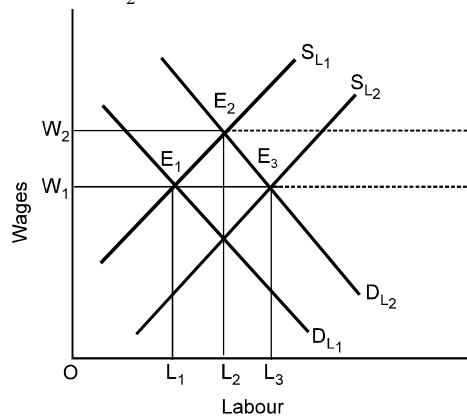
Fig. 2.5 *Labour Market for Industry* X

Fig. 2.6 *Demand for Labour by a Firm in Industry* X

2.2.3 The Effect of Change in Factor Demand

Let us now examine the effect of change in consumer demand on factor demand and changes in factor market. In order to analyse the effects in a somewhat wider framework, let us drop the assumption that there is only one firm in each industry and assume, instead, that there are several firms in each industry. Recall that the firms in industry *X* are making supernormal profits while firms is industry *Y* are incurring losses. Some firms in industry *Y* are therefore forced to quit the industry and some are induced to transfer their resources to industry *X*. Besides, the demand factors in industry *X* would increase. This tendency in the commodity markets affects the factor markets with respect to each industry. Consider first the increase on demand for factors in industry *X* and its effect on factor prices.

The entry of new firms to industry *X* and expansion of production by the existing firms increases demand for labour and capital in this industry. The effect of increase in demand for labour is illustrated in Figs. 2.5 and 2.6. Suppose that the labour market for industry *X* was initially in equilibrium at point E_1. Due to the increase in demand for labour, the demand curve D_{L1} shifts to D_{L2} causing increase in the employment of labour in industry *X* from OL_1 to OL_2 and increase in wage rate for the industry from OW_1 to OW_2. The increase in demand for labour by an individual firm of the industry is illustrated in Fig. 2.6. It shows that the demand curve for labour by an individual firm shifts rightward from d_{l1} to d_{l2}. At new wage rate OW_2, an individual firm employs Ol_2 workers or $l_1 l_2$ additional workers at the ruling wage rate (Note that $l_1 l_2$ multiplied by the number of firms in the industry equals $L_1 L_2$ in Fig. 2.5).
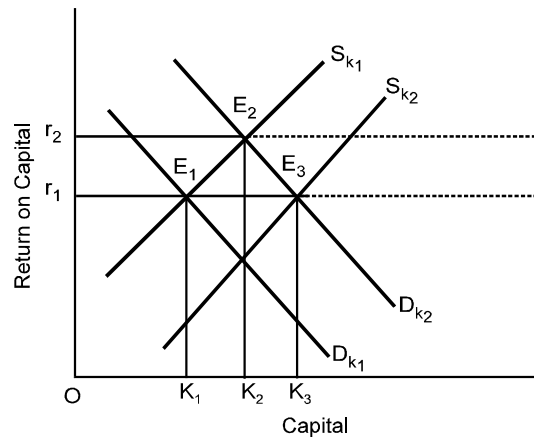
Fig. 2.7 *Capital Market for Industry* X

Let us now see what happens in the capital market. The changes in the capital market for industry *X* is illustrated in Figs. 2.7 and 2.8. Demand for capital increases in this industry, and the initial capital demand curve D_{k1} shifts upward to the position of D_{k2} causing equilibrium point of capital market to shift to E_2 and return on capital to rise to O_{r2} (Fig. 2.7). The capital-demand curve for an individual firm in industry *X* shifts from d_{k1} to d_{k2} as return on capital increases and the employment of capital by an individual firm increases from O_{k1} to O_{k2} (Fig. 2.8). The total demand for capital in industry *X* increases by K_1K_2 (Fig. 2.7) which equals k_1k_2 multiplied by the of firms in the industry.
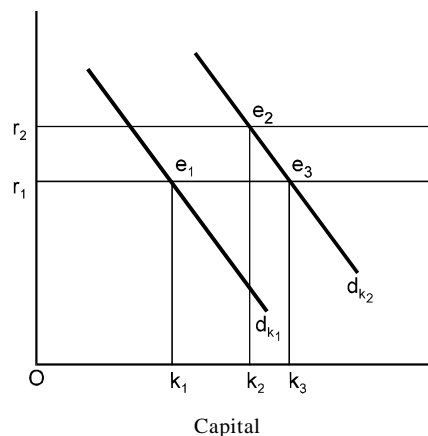
Fig. 2.8 *Demand for Capital by a Firm in Industry* X

Let us now see what has happened in the factor markets in respect of industry *Y*. First let us consider the labour market. The changes in the labour market in respect of industry *Y* are illustrated in Figs. 2.9 and 2.10. Let the labour market for industry *Y* to be in equilibrium at point E_1. Recall from Fig. 2.4 that firms in industry *Y* incur losses. Therefore, the demand for labour decreases and labour demand curve shifts downward from its initial position D_{L2} to D_{L1}. Consequently, the wage rate decreases from OW_2 to OW_1 and employment of

labour in the industry decreases from OL_3 to OL_1. The decrease in demand for labour by an individual firm of industry Y is shown in Fig. 2.10 by a downward shift in labour demand curve from d_{l2} to d_{l1}. Each firm employs less of labour even though wage rate has gone down from OW_2 to OW_1.
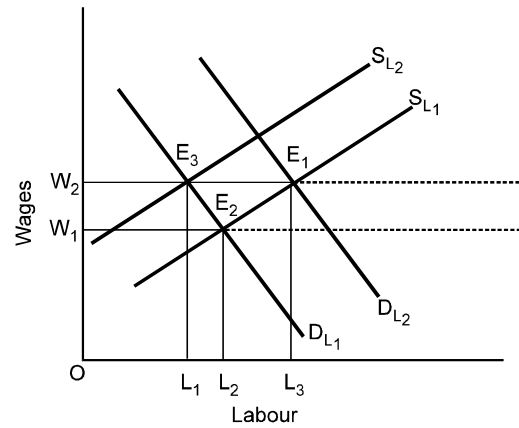
Fig. 2.9 *Labour Market for Industry* Y
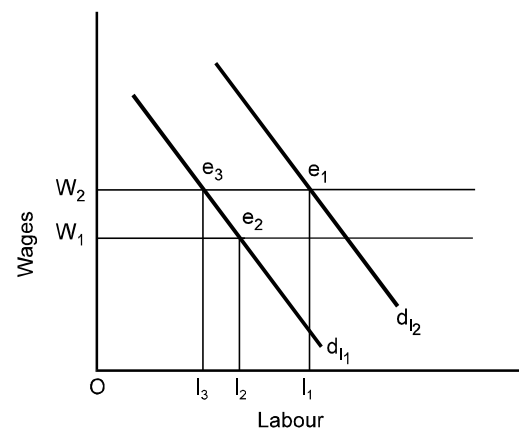
Fig. 2.10 *Demand for Labour by a Firm in Industry* Y

Let us now turn to the capital market for industry Y. A condition similar to the labour market for the industry Y takes place in the capital market too for industry Y. Demand for capital decreases as shown by the downward shifts of capital demand curve of both individual firms (Fig. 2.12) and industry (Fig. 2.11) because return on capital in the industry decreases.

To sum up, due to the change in consumer's preference in favour of commodity X caused by an exogenous factor, demand for commodity X has increased and for commodity Y decreased. As a result, price of X increases and that of Y decreases. Factor price remaining the same, profitability of industry X increases while that of industry Y decreases. This leads to increase in demand for L and K in industry X and to decrease in demand for L and K in industry Y. These changes in demand for factors have led to disequilibrium in the system since firms

in industry *X* are earning supernormal profits which is in consistent with perfect competition. In a perfectly competitive system, however, the disequilibrium is self-correcting. Let as now see how the process of automatic adjustment begins and where it ends.
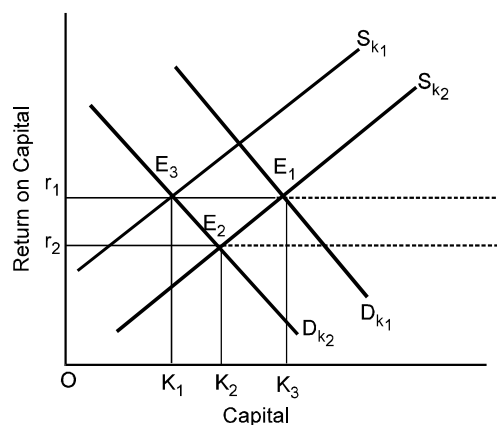
Fig. 2.11 *Capital Market for Industry* Y
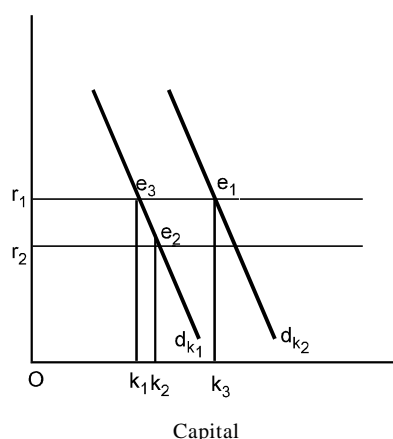
Fig. 2.12 *Demand for Capital by a Firm in Industry* Y

2.2.4 The Process of Automatic Adjustment

We have noted that both wages and returns on capital increase in industry X and decrease in industry *Y*. This will cause factors (labour and capital) to move from industry *Y* to industry *X*. Consequently, in the long-run, factor supply to industry *X* would increase and to industry *Y,* it would decrease. The increase in supply of labour and capital to industry *X* is shown by a rightward shift in the labour supply curve from SL_1 to SL_2 (Fig. 2.5) and in capital supply curve from S_{k1} to S_{k2} (Fig. 2.7).

With increased supply of labour and capital to industry *X*, the supply of commodity *X* increase causing supply curve to shift to S_{x2} and new equilibrium is reached at point E_3 (Fig. 2.1). As shown in Fig. 2.1, new equilibrium is gained at the original price OP_{x1} but at a greater output of *X*. In industry *Y* reverse happens.

Since factors move out of industry *Y,* the factor supply to the industry is reduced as shown by downward shift in factor supply curves in Figs. 2.9 and 2.11. Besides, since firms of this industry have a tendency to move out, industry's production declines. Consequently, the market supply curve of commodity *Y* shifts backwards to S_{y1}. A new equilibrium is reached at point E_2 at original price OP_{y1} (Fig. 2.3) and level of output (OY_3) falls much below the original output OY_1.

Thus, in the long-run, markets of both the commodities, *X* and *Y*, return to a stable equilibrium at the original price level, though at different levels of output: while output of *X* increases, that of *Y* decreases. An important point to be borne in mind is that new equilibrium is not necessarily gained at the original price. Whether new equilibrium is gained at original price or not depends on the extent of increase (decrease) in the supply of the commodity, i.e., the extent to which supply curve shifts forward (backward). The original equilibrium is regained only if supply of commodity increases (decreases) and supply curve shifts forward (backward) exactly to the extent of excess (shortfall) in demand. If shifts in the supply curve are greater or smaller than excess of deficit in demand, the new equilibrium price will be different from the original price.

Once the commodity markets reach new equilibrium and stabilise, the inward and outward flows of factors ends. For example, as shown in Figs. 2.1 and 2.2, industry *X* and its firms are in equilibrium. Price is refixed at OP_{x1} at which all firms are earning only normal profits, since price $= AC = MC = MR$ (Fig. 2.2). There is no incentive for the existing firms to expand their output. Nor is there any incentive for new firms to enter the industry. Under these conditions, there is no incentive for the factors to move to this industry. This leads to saturation in the factor markets for industry *X.* It simultaneously stops the flow of resources out of industry *Y.* This leads to saturation in factor markets for both the industries.

The new equilibria in labour and capital markets for industry *X* are presented in Figs. 2.5 and 2.7. The labour supply curve for industry *X* finally shifts to S_{L_2} and a new equilibrium is set at point E_3 (Fig. 2.5). Similarly, the capital supply curve for industry *X* shifts to Sk_2 (Fig. 2.7) and a new equilibrium is set in capital market for industry *X* at original wage rate Or_1. Thus, both labour and capital markets reach a new equilibrium.

The new equilibria in labour and capital markets are presented in Figs. 2.9 and 2.11. The labour supply curve for industry *Y* shifts backward to S_{L_2} and a new equilibrium is set at point E_3 (Fig. 2.9). As to capital market for industry *Y*, capital supply curve shifts backward to S_{k1} and new equilibrium is set at E_3 at original rate of return, Or_1. Thus, both factor markets for industry *Y* reach a new equilibrium, though at a much lower level of employment of both labour and capital. In industry *Y* labour employment decreases from O_{L_3} to O_{L_1} and capital employment decreases from O_{K_3} to O_{K_1}.

As in case of commodity markets, whether factor markets reach new equilibrium at the original level of factor prices or not depends on the extent to which factor supply curve shift forward (or backward).

We may now sum up the above discussion. We started by assuming the whole system to be in equilibrium. The system was then assumed to be disturbed by an exogenous factor, i.e., change in consumer's taste. This led to a chain of actions and reactions in commodity and factor markets. These actions and reactions led the system to stablise at a new equilibrium. The attainment of new equilibrium is however certain only under perfect competition and continuous production function with diminishing returns to scale.

The above illustration does not provide a formal proof of existence of a stable general equilibrium solution. It simply describes the tendency towards a general equilibrium under perfectly competitive conditions. Whether there exists a unique and stable general equilibrium solution is the question which we will discuss in the next section.

2.3 EXISTENCE, UNIQUENESS AND STABILITY OF GENERAL EQUILIBRIUM

In this section, we answer the questions (i) does there exist a general equilibrium solution? (ii) if it does, is it unique? (iii) is the solution stable?

1. The Existence of General Equilibrium

If number of equations and the number of 'unknowns' are equal, it may sometimes make one think that there exists a general equilibrium solution. But, the equality of number of equations with that of unknowns is neither a sufficient nor a necessary condition for the existence of a general equilibrium solution. That it is not a sufficient condition is easy to prove. It is possible to find a system of two equations with two unknowns that has no solution in the realm of real numbers, for only real numbers have economic meaning. For instance, suppose we have two equations.

$$x^2 + y^2 = 0$$
$$x^2 - y^2 = 1$$

Solving for *x* and *y,* we get $x = \sqrt{1/2}$ and $y = i\sqrt{1/2},$ where the imaginary number *i* satisfies $i^2 = -1$.

It can also be shown that equality of equations and unknowns is not a necessary condition. Consider the equation

$$x^2 + y^2 = 0$$

This single equation with two unknowns offers a unique solution for *x* and *y* in the domain of real number, i.e., $x = 0$ and $y = 0$.

The example suggests that a unique general equilibrium may exist at zero prices and even at negative prices. These may be the cases of certain 'free goods' or 'nuisance goods'. The problems of zero or negative price could be solved by eliminating free goods or nuisance goods. But, as Menger pointed out, there may be a tendency of free goods to decrease as economic development takes place. Therefore, all kinds of goods must be included in Walrasian system. This is something which Walras did not realise. Hence, his demonstration of existence of general equilibrium solution is unsatisfactory.

Furthermore, it is mathematically possible to show the existence of general equilibrium solution involving zero and negative prices. But, while negative prices and quantities of consumer goods is understandable, it is difficult to imagine zero or negative factor prices and quantities. One can hardly imagine a worker paying for his employment.

2. Uniqueness of General Equilibrium Solution

The uniqueness of general equilibrium solution requires that, at all partial equilibrium levels, demand and supply schedules intersect at only one point giving a positive price. At any other price higher than the price so determined, $S > D$, and at any lower price $D > S$, as shown in Fig. 2.13. But if demand schedule of a commodity is backward bending, as in case of inferior goods, there will be no unique equilibrium. Instead there will be multiple equilibria. As shown in Fig. 2.14, there are two equilibrium points, e_1 and e_2.
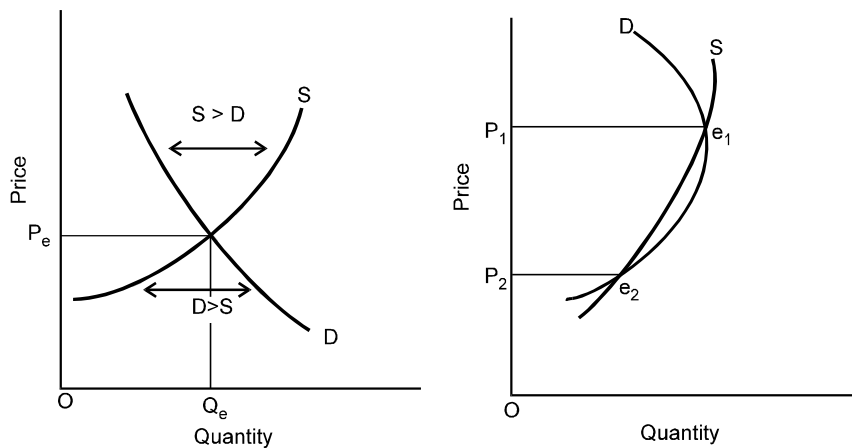
Fig. 2.13 *Unique and Stable Equilibrium*   **Fig. 2.14** *Multiple Equilibria*

However, Wald and, later, Arrow and Debreu have shown that "the Walrasian system does possess a unique and economically meaningful solution, provided returns to scale are constant or diminishing and there are no joint products or external effects either in production or in consumption." Obviously, the unique solution exists under restrictive assumptions.

Stability of General Equilibrium Solution

Walras also tried to show that general equilibrium is stable. "Walras' stability analysis was based on the assumption that the rate of price changes varies directly with the amount of excess demand. Walras, like Marshall, treated instability in the context of multiple equilibria; the unstable position is invariably found between two stable positions. But unstable equilibria in Walras arise from the intersection of a backward bending supply curve of a productive service with a more steeply falling demand curve. This implies the possibility but certainly not the necessity of multiple equilibria because the supply curve may never bend back again no matter how high factor price rise."

Walras attempted to show not only stability in a single market but also a multimarket stability. (For details see above the illustration of tendency towards general equilibrium in all markets). Hicks has also attempted to show, in his *Value and Capital,* that the multimarket does not exist provided there are no strong income effects. It is however difficult to establish that the general equilibrium solution is determinate and stable.

Evaluation of General Equilibrium approach

Walrasian general equilibrium model has many shortcomings. Many of its assumptions are highly restrictive and unrealistic. The uniqueness and stability of solution that it offers are doubtful. It is also alleged sometimes that the Walrasian general equilibrium model has little economic content.

Despite its shortcomings, the Walrasian general equilibrium model has its own merits.

First, Walras was the first to recognise and formalise the mutual interdependence of various prices and quantities in an economic system. Although it is widely known that in economics, every thing depends on everything else, the full implications of this generalisation were not grasped before Walras.

Second, general equilibrium approach has a wide applicability to the analysis of various economic phenomena. Modern theories of money, international trade, employment, and economic growth are general equilibrium theories in a simplified form. Also, the 'new' welfare economics is an outgrowth of general equilibrium theory. The modern macroeconomics and micro-economics can be viewed as different ways of giving operational relevance to general equilibrium analysis.

Check Your Progress

1. Name the two kinds of approaches adopted by economists for economic analysis.
2. Define general equilibrium.

2.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The economists, therefore, adopt two kinds of approaches to economic analysis: (i) Partial equilibrium approach, and (ii) General equilibrium approach.

2. A general equilibrium is defined as a state in which all economic units maximise their respective objective function and all prices are simultaneously in equilibrium, and all markets are cleared.

2.5 SUMMARY

- A fundamental feature of an economic system is the interdependence and interrelatedness of economic activities—production and consumption—of its various constituents—individuals, households, firms, banks and other kinds of financial institutions.

- Partial equilibrium approach ignores the interdependence of the various segments of the economy. It isolates the segment or the phenomenon of the study from the other segments and assumes non-existence of influences of the changes occurring outside the area delimited for the study.

- The general equilibrium approach, on the other hand, recognises the interdependence of constituent parts of the economic system.

- Leon Walras (1834–1910), a French economist, was the first to attempt to answer these questions in his book *Elements of Pure Economics* (1874).

- In the Walrasian system of general equilibrium, the behaviour of each decision maker is presented by a set of equations.

- "Walras' stability analysis was based on the assumption that the rate of price changes varies directly with the amount of excess demand.

- Walrasian general equilibrium model has many shortcomings. Many of its assumptions are highly restrictive and unrealistic. The uniqueness and stability of solution that it offers are doubtful.

2.6 KEY WORDS

- **Economic development:** It is the process by which a nation improves the economic, political, and social well-being of its people.

- **Equilibrium:** It is a state in which opposing forces or influences are balanced.

2.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. List the shortcomings of the Walrasian approach to general equilibrium.

2. What is the main objective of the general equilibrium approach?

3. Mention any two differences between partial equilibrium approach and general equilibrium approach.

Long Answer Questions

1. Describe the Walrasian approach to general equilibrium.

2. Evaluate the effect of change in consumer demand on factor demand and changes in the factor market.

3. 'The partial equilibrium approach ignores the interdependence of the various segments of the economy.' Elucidate the statement.

2.8 FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.

UNIT 3 ANALYSIS OF CONSUMER DEMAND

Structure

3.0 INTRODUCTION

In the previous unit, you studied about general and partial equilibrium analysis. This unit will introduce you to the concept of demand analysis, ordinal utility theory, revealed preference theory and the theory of consumer's surplus. *Demand* is an economic principle alluding to the consumer's desire and inclination to pay a price for a particular good or service. Holding all other factors constant, an increase in the price of a good or service will decrease *demand*, and vice versa.

3.1 OBJECTIVES

After going through this unit, you will be able to:

- Interpret demand analysis
- Describe the ordinal utility theory
- Discuss the revealed preference theory
- Explain the theory of consumer's surplus

3.2 DEMAND ANALYSIS

Conceptually, the term 'demand' implies a 'desire' or 'want' for a commodity backed by the ability and willingness to pay for it. Unless a person has adequate purchasing power or resources and the preparedness to spend his resources, his desire for a commodity is not considered as his demand. For example, if a man wants to buy a car but he does not have sufficient money to pay for it, his want is not his demand for the car. And, if a rich miserly person wants to buy a car but is not willing to pay, his desire too is not his demand for a car. But if a man, has sufficient money and is willing to pay, his desire to buy a car is an effective demand.

The desires without adequate purchasing power and willingness to pay do not affect the market, nor do they generate production activity. A want with *three attributes – desire to buy*, *willingness to pay and ability to pay –* becomes effective demand. Only an effective demand figures in economic analysis and business decisions.

The term 'demand' for a commodity (i.e., quantity demanded) always has a reference to 'a price', 'a period of time' and 'a place'. Any statement regarding the demand for a commodity without reference to its price, time of purchase and place is meaningless and is of no practical use. For instance, to say 'the demand for TV sets is 50,000' carries no meaning for a business decision, nor does it have any use in any kind of economic analysis. A meaningful statement regarding the demand for a commodity should, therefore, contain the following information:

(*a*) the quantity demanded,

(*b*) the price at which a commodity is demanded,

(*c*) the time period over which a commodity is demanded, and

(*d*) the market area in which a commodity is demanded.

For example, saying, 'the annual demand for TV sets in Delhi at an average price of ₹ 15,000 a piece is 50,000' is a meaningful statement.

3.3 ORDINAL UTILITY THEORY

In this section, we will discuss the ordinal utility theory of consumer behaviour. The technique economists use under ordinal utility approach is called "indifference curve". Therefore the ordinal utility approach is also known as indifference curve analysis of consumer behaviour.

The indifference technique was invented and used by Francis Y. Edgeworth(1881) to show the possibility of exchange of commodities between two individuals. About a decade later, Irving Fisher (1892) used indifference curve to explain consumer's equilibrium. Both Edgeworth and Fisher, however, believed in cardinal measurability of utility. It was Vilfred Pareto who introduced, in 1906, the ordinal utility hypothesis to the indifference curve analysis. In the subsequent

decades, many significant contributions were made by Eugen E. Slutsky, W.E. Johnson, and A.L. Bowley. Yet, indifference curve technique could not gain much ground in the analysis of consumer behaviour till early 1930s. In 1934, John R. Hicks and R.G.D. Allen developed systematically the ordinal utility theory as a powerful analytical tool of consumer analysis. Later, Hicks provided a complete exposition of indifference curve technique in his *Value* and *Capital*. Though in his later work, *A Revision of Demand Theory*, he has dropped some of his earlier assumptions, indifference analysis is regarded as the most powerful tool of consumer analysis.

The fundamental departure that indifference curve analysis makes from the Marshallian marginal utility analysis is the hypothesis that utility can be measured only ordinally, not cardinally. Recall that 'cardinalists' assumed that utility is cardinally measurable, and that utility of one commodity is independent of other commodities. In contract, the 'Ordinalists' believe that *cardinal measurement of utility is neither feasible nor necessary to analyse consumer's behaviour.* According to ordinalists, all that is required to analyse consumer's behaviour is that the consumer should be able to order his preferences. In fact, the consumer is able to express his preference for the quantity of a commodity to that of others. For example, a consumer can always say that he prefers to 10 kg of wheat to 5 kg of rice.

3.3.1 Assumptions of Ordinal Utility Theory

The indifference curve analysis of consumer's behaviour makes, at least implicitly, the following assumptions:

1. **Rationality**. The consumer is a rational being. He aims at maximising his total satisfaction, given his income and prices of goods and services he consumes. Furthermore, he has full knowledge of his circumstances.

2. **Ordinal Utility.** Indifference curve analysis assumes that utility can be expressed only ordinally. That is, the consumer is able to tell only the order of his preferences.

3. **Transitivity and Consistency of Choice**. Consumer's choices are *transitive.* Transitivity of choice means that if a consumer prefers A to B and B to C, he must prefer A to C. Or, if he treats $A = B$ and $B = C$, he must treat $A = C$, Consistency of choice means that, if he prefers A to B in one period, he will not prefer B to A in another period or treat them as equal. The transitivity and consistency in consumer's choices may be symbolically expressed as follows.

   Transitivity. If $A > B$, and $B > C$, then $A > C$, and

   Consistency. If $A > B$, in one period, then $B \not> A$ or $B \neq A$ in another.

4. **Nonsatiety**. It is also assumed that the consumer is not oversupplied with goods in question and that he has not reached the point of saturation in case of any commodity. Therefore, a consumer always prefers a larger quantity of all the goods.

5. **Diminishing Marginal Rate of Substitution.** The marginal rate of substitution means the rate at which a consumer is willing to substitute one commodity (X) for another (Y), i.e., the units of Y he is willing to give up for one unit of X so that his total satisfaction remains the same. This rate is given by $\Delta Y/\Delta X$. The assumption is that $\Delta Y/\Delta X$ goes on decreasing, when a consumer continues to substitute X for Y.

3.3.2 The Meaning and Nature of Indifference Curve

An indifference curve may be defined as *the locus of points, each representing a different combination of two goods but yielding the same level of utility or satisfaction. Since each combination of two goods yields the same level of utility, the consumer is indifferent between any two combinations of goods when it comes to making a choice between them.* A consumer is very often confronted with such a situation in real life. Such a situation arises because he consumes a large number of goods and services, and often he finds that one commodity serves as a substitute for another. It gives him an opportunity to substitute one commodity for another, and to make various combinations of two substitutable goods. It may not be possible for him to tell *how much* utility a particular combination gives, but it is always possible for him to tell which one between any two combinations is preferable to him. It is also possible for him to tell which combinations give him equal satisfaction. If a consumers is faced with equally good combinations, he would be indifferent between the combinations. When such combinations are plotted graphically, the resulting curve is known as **indifference curve**. Indifference curve is also called **Iso-utility Curve** and **Equal Utility Curve.**

For example, let us suppose that a consumer consumes only two commodities X and Y and he makes five combinations which he calls *a, b, c, d* and *e*. All these combinations yield him equal utility. Therefore, he is indifferent between the combinations *a, b, c, d,* and *e* of two commodities, X and Y. His combinations are presented in Table 3.1, which may be called as **indifference schedule**—a schedule of various combinations of two goods, between which a consumer is indifferent. The last column of the table shows an undefined utility (u) derived from each combination of X and Y. If combinations *a, b, c, d,* and *e* given in Table 3.1 are plotted and joined to form a smooth curve (as shown in Fig. 3.1), the resulting curve is known as **indifference curve**. On this curve, one can locate many other points showing many other combinations of X and Y which yield the same satisfaction. Therefore, the consumer is indifferent between the combinations which may be located on the indifference curve.

Table 3.1 *Indifference Schedule of Commodities* X *and* Y

Combination	Commodity X	Commodity Y	Utility
a	25	5	u
b	15	7	u
c	10	12	u
d	6	20	u
e	4	30	u

The Indifference Map

We have drawn a single indifference curve in Fig. 3.1 on the basis of an indifference schedule given in Table 3.1. The combinations of the two commodities, *X* and *Y*, given in the indifference schedule or those indicated by the indifference curve are by no means the only combinations of the two commodities. The consumer may be faced with many other combinations with *less* of one or both the goods—each combination yielding the same level of satisfaction but less than the level of satisfaction indicated by the indifference curve *IC* in Fig. 3.1. Therefore, another indifference curve can be drawn, say, through points *f*, *g* and *h*. Note that this indifference curve falls below the curve *IC* given in Fig. 3.1. Similarly, he may be faced with many other combinations with *more* of one or both the goods—each combination yielding the same satisfaction—but greater than the satisfaction indicated by the lower indifference curves. Thus, another indifference curve can be drawn above the *IC* given in Fig. 3.1, say, through points *j*, *k*, and *l.* This exercise may be repeated as many times as one wants, each time generating a new indifference curve.
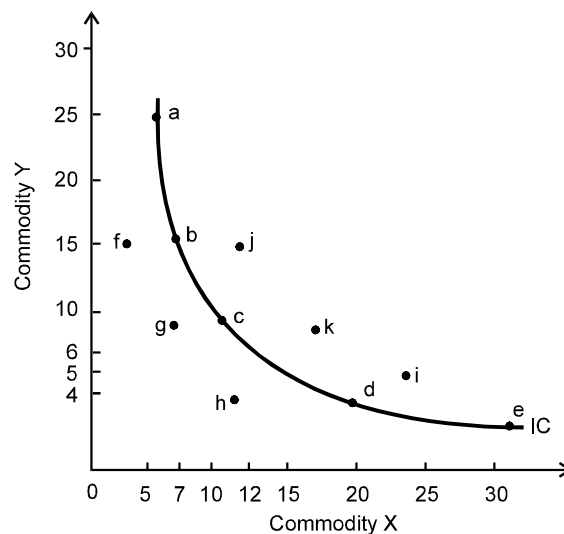
Fig. 3.1 *Indifference Curve*
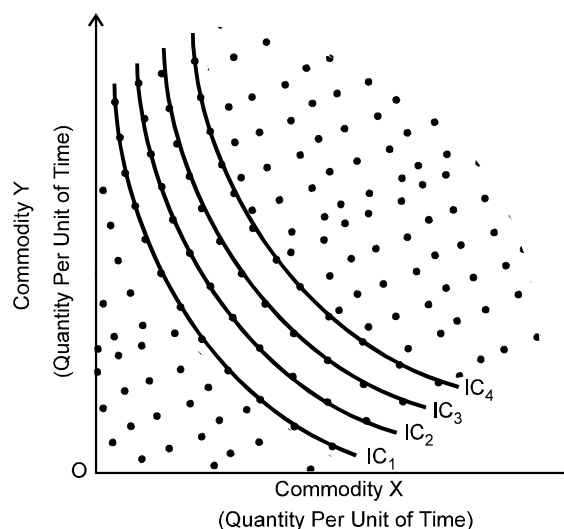
Fig. 3.2 *The Indifference Map*

In fact the area between *X* and *Y* axes is known as *indifference plane* or *commodity space*. This plane is full of finite points and each point on the place indicates a different combination of goods *X* and *Y*. Intuitively, it is always possible to locate any two or more points indicating different combinations of goods *X* and *Y* yielding the same satisfaction. It is thus possible to draw a number of indifference curves without intersecting or touching the other, as shown in Fig. 3.2. The set of indifference curves, IC_1, IC_2, IC_3 and IC_4 drawn in this manner make the *indifference map.* In fact, an indifference map may contain any number of indifference curves, ranked in the order of consumer's preferences.

3.3.3 The Diminishing Marginal Rate of Substitution

When a consumer makes different combination of two goods, yielding the same level of satisfaction, he substitutes one good for another. The rate at which he substitutes one good for the other is called the 'Marginal Rate of Substitution (*MRS*)'. One of the basic postulates of *indifference curve analysis* is that (*MRS*) diminishes. The axiomatic assumption of ordinal utility theory is analogous to the assumption of 'Diminishing Marginal Utility' in *cardinal utility theory*. The *postulate of diminishing marginal rate of substitution states* an observed behavioural rule that when a consumer substitutes one commodity (say *X*) for another (say *Y*), the 'Marginal Rate of Substitution' (*MRS*) decreases as the stock of *X* increases and that of *Y* decreases.

Measuring Marginal Rate of Substitution (MRS)

Conceptually, the *MRS* is the rate at which one commodity can be substituted for another, the level of satisfaction remaining the same. The *MRS* between two commodities, *X* and *Y*, can also be defined as the number of units of *X* which are required to replace one unit of *Y* (or number of units of *Y* that are required to

replace one unit of *X*), in the combination of the two goods so that the total utility remains the same. It implies that the utility of units of *X* (or *Y*) given up is equal to the utility of additional units of *Y* (or *X*) added to the basket.

To explain symbolically the concept of *MRS*, let us suppose that the utility function of a consumer is given as

$$U = f(X, Y)$$

Let us now suppose that the consumer substitutes *X* for *Y*. When the consumer foregoes some units of *Y*, his stock of *Y* decreases by $-\Delta Y$. His loss of utility may be expressed as

$$-\Delta Y \cdot MU_y$$

On the other hand, as a result of substitution, his stock of *X* increases by ΔX. His utility from ΔX equals

$$+\Delta X \cdot MU_x$$

For the total utility *U* to remain the same, $-\Delta Y. MU_y$ must be equal to $\Delta X.MU_x$. That is,

$$-\Delta Y.MU_y + \Delta X.MU_x = 0 \qquad \qquad ...(3.1)$$

Rearranging the terms in Eq. (3.2), we get *MRS* of *X* for *Y* as

$$-\frac{\Delta Y}{\Delta X} = \frac{MU_x}{MU_y} \qquad \qquad ...(3.2)$$

Here, $\Delta Y/\Delta X$ is simply the *slope* of the indifference curve, which gives the $MRS_{x,y}$ when *X* is substituted for *Y*. Similarly, $\Delta X/\Delta Y$ gives $MRS_{y,x}$ when *Y* is substituted for *X*.

and
$$\left[\begin{array}{l} MRS_{x,y} = -\dfrac{\Delta Y}{\Delta X} = \dfrac{MU_x}{MU_y'} \\[2mm] MRS_{y,x} = -\dfrac{\Delta X}{\Delta Y} = \dfrac{MU_y}{MU_x} \end{array}\right] = \text{Slope of indifference curve}$$

The Diminishing MRS

As mentioned basic postulate of ordinal utility theory is that the $MRS_{x,y}$ (or $MRS_{y,x}$) decreases. That is, the number of units of a commodity that a consumer is willing to sacrifice for an additional unit of another goes on decreasing when he goes on substituting one commodity for another. The diminishing $MRS_{x,y}$ which can be obtained from Table 3.1, are presented in Table 3.2.

Table 3.2 *The Diminishing* MRS *between Commodities* X *and* Y

Movements on IC	Change in Y ($-\Delta Y$)	Change in X (ΔX)	$MRS_{y,x}$ ($\Delta Y/\Delta X$)
From point *a* to *b*	– 10	2	– 5.0
From point *b* to *c*	– 5	5	– 1.0
From point *c* to *d*	– 4	8	– 0.5
From point *d* to *e*	– 2	10	– 0.2

As Table 3.1 shows, when the consumer moves from point *a* to *b* on the indifference curve (Fig. 3.1) he gives up 10 units of commodity *Y* and gets only 2 units of commodity *X*, so that

$$MRS_{y,x} = \frac{-\Delta Y}{\Delta X} = \frac{-10}{2} = -5$$

As he moves down from points *b* to *c,* he losses 5 units of *Y* and gains 5 units of *X*, giving

$$MRS_{x,y} = \frac{-\Delta Y}{\Delta X} = \frac{5}{5} = -1$$

The $MRS_{x,y}$ goes on decreasing as the consumer moves further down along the indifference curve. The diminishing marginal rate of substitution causes the indifference curves to be convex to the origin.

The diminishing marginal rate of substitution can also be illustrated graphically, as shown in Fig. 3.3.

As the consumer moves from point *a* to *b,* to *c*, and to *d,* he gives up a constant quantity of *Y*, (i.e., $\Delta Y_1 = \Delta Y_2 = \Delta X_3$). To substitute a constant quantity of $\Delta Y,$ he requires an increasing quantity of X (i.e., $\Delta X_1 < \Delta X_2 < \Delta X_3$). Since *MRS* is given by the slope of the indifference curve, (i.e., $\Delta Y/\Delta X$), arranging the slopes between points *a* and *b, b* and *c*, and *c* and *d,* in descending order, we get

$$\frac{\Delta Y_1}{\Delta X_1} > \frac{\Delta Y_2}{\Delta X_2} > \frac{\Delta Y_3}{\Delta X_3}$$

These inequalities show that *MRS* ($=\Delta Y/\Delta X$) goes on decreasing as the consumer moves from point *a* towards point *d*.

The diminishing *MRS* is geometrically illustrated in Fig. 3.3(b). The lines tangent to the indifference curve at points *a*, *b* and *c* measure the slope of the curve at these points. It can be seen from the figure, that as the consumer moves from point *a* towards *d*, the tangential lines become flatter indicating decrease in the slope of the indifference curve. This also proves the decrease in *MRS* all along the indifference curve.
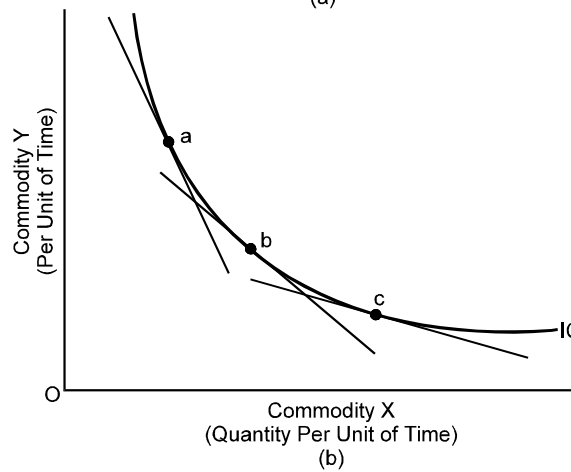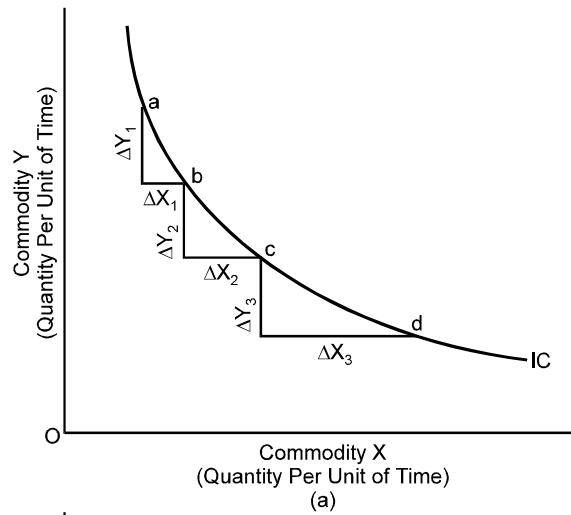
Fig. 3.3 *Diminishing Marginal Rate of Substitution*

Why Does MRS Diminish?

The negative slope of the indifference curve implies that two commodities are not perfect substitutes for each other. In case they are perfect substitutes, the indifference curve will be a straight line with a negative slope. Since goods are not perfect substitutes for each other, the *subjective value attached to the additional quantity (i.e., MU) of a commodity decreases fast in relation to the other commodity whose total quantity is decreasing.* Therefore, when the quantity of one commodity (say, X) increases and that of other (say, Y) decreases, it becomes increasingly difficult for the consumer to sacrifice more and more units of commodity Y for one unit of X. But, if he is required to sacrifice additional units of Y, he will demand increasing units of X to maintain the level of his satisfaction. As a result, the *MRS* decreases.

Furthermore, when the combination of two goods at a point of indifference curve is such that it includes a large quantity of one commodity, (say, Y) and a small quantity of the other (commodity X), then consumer's capacity to sacrifice Y

is greater than to sacrifice *X*. Therefore, he can sacrifice a large quantity of *Y* in favour of a smaller quantity of *X*. This is an observed behavioural rule that the consumer's willingness and capacity to sacrifice a commodity is greater when its stock is greater and it is lower when the stock of a commodity is smaller. These are the reasons why *MRS* decreases all along on the indifference curve.

3.4 REVEALED PREFERENCE THEORY

The cardinal and ordinal utility approaches to demand analysis, discussed in the preceding sections, are based on two different concepts of utility–cardinal or ordinal. The cardinal approach assumes absolute or cardinal measurability of utility and ordinal approach assumes relative or ordinal or introspective measurability of utility. While cardinal measurement of utility is not practicable, the introspective ordinal utility is non-observable. Thus, both these approaches involve problems of measurability of utility. In an attempt to overcome this problem, Samuelson proposed in 1947 another theory of demand called 'Revealed Preference Theory' of consumer behaviour.

The main merit of the revealed preference theory is that the 'law of demand' can be directly derived from the revealed preference axioms without using indifference curves and most of its restrictive assumptions. What is needed is simply to record the observed behaviour of the consumer in the market. The consumer reveals his behaviour by the basket of goods he buys at different prices.

Besides, the revealed preference theory is also capable of establishing the existence of indifference curves and their convexity. For its merits, revealed preference theory is treated as the 'third root of the logical theory of demand'.

3.4.1 Assumptions

Samuelson's revealed preference theory is based on the following straight forward assumptions:

1. **Rationality.** The consumer is assumed to be a rational being. In his order of preferences, he prefers a larger basket of goods to the smaller ones.

2. **Transitivity.** Consumer's preferences are assumed to be *transitive*. That is, given alternative baskets of goods, *A*, *B* and *C*, if he treats *A* > *B* and *B* > *C*, then he treats *A* > *C*.

3. **Consistency.** It is also assumed that during the analysis, consumer's taste remains constant and consistent. *Consistency* implies that if a consumer, given his circumstances, prefers *A* to *B* he will not prefer *B* to *A* under the same conditions.

4. **Effective price inducement.** Given the collection of goods, the consumer can be induced to buy a particular collection of goods by providing him sufficient price incentives. That is, for each collection, there exists a price line which makes it attractive for the consumer.

3.4.2 Revealed Preference Axiom

The revealed preference axiom can be stated as follows. Given the budgetary constraint and alternative baskets of goods having the same price, if a consumer chooses a particular basket, he reveals his preference for the basket. For example, suppose there are two alternative baskets *A* and *B* of two goods *X* and *Y*. Both the baskets being equally expensive, if a consumer chooses basket *A* rather than basket *B*, he reveals his preference for basket *A*.

If a consumer chooses a particular basket, he does so either because he likes it more or it is less expensive than the other. In our example, if the consumer chooses basket *A* rather than *B* because *A* is cheaper, then the preference for *A* is not revealed because the consumer might regret not having been able to buy basket *B*. But, if both the baskets are equally expensive, then there is only one plausible explanation, i.e., the consumer likes basket *A* more than basket *B*. In this case, the consumer reveals his preference for basket *A*.
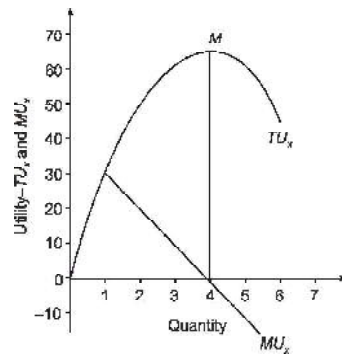
Fig. 3.4 *Revealed Preference*

The revealed preference axiom has been illustrated in Fig. 3.4. The consumer's budgetary constraint has been shown by his budget line *MN*. If he chooses a particular bundle of *X* and *Y* represented by point *A* on the budget line, it implies that he prefers point *A* to any other point on the budget line, say point *B*. Since point *B* is on the budget line, it is as much expensive as *A*. If the consumer chooses point *A*, it means that *A* is revealed to be preferable to *B* and *B* is revealed inferior to *A*. Any point below the budget line, like point *C*, represents a smaller and cheaper basket of *X* and *Y* and hence it is revealed inferior to *A*. Therefore, any point above the budget line, like point *D*, represents a larger and more expensive basket of goods than indicated by point *A*. Hence point *D* is preferable to point *A*.

3.5   THEORY OF CONSUMER'S SURPLUS

In the preceding analyses of consumer behaviour, it was assumed that consumers are aware of prices of goods and services they consume. In real life, however, the consumers may not necessarily be aware of all the prices. They find often that the price which they are willing to pay is different from what they are actually required

to pay. If the price a consumer is willing to pay is higher than the price which he actually pays, then the consumer is said to have a *surplus*. This **surplus** is called **consumer surplus.**

The conept of *consumer's surplus* is believed to have been originated by a French engineer, Arsene Julis Dupuit, in 1844, in his effort to measure social benefit of such collective goods as roads, canals and bridges. In his opinion, the value of the benefit of such collective goods was greater than the price actually charged because most people would be willing to pay a higher price than they actually paid. The concept was later refined by Marshall who provided also a measure of consumer's surplus. His premise of measuring consumer's surplus was, however, rejected by the *ordinalists* who attempted to provide a different method of measuring consumer's surplus through their indifference curve technique.

In this section, we have discussed the Marshallian and Hicksian methods of measuring consumer's surplus and their merits and demerits. Let us first look at Marshallian concept and measure of consumer surplus and its drawbacks.

3.5.1 Marshallian Concept of Consumer Surplus

Although the concept of consumer surplus was originated by Dupuit as early as 1844, it remained an immeasurable concept until Marshall suggested, as late as 1920, a method of measuring consumer's surplus in money terms. Marshall defined consumer's surplus as "the excess of the price which [a consumer] would be willing to pay rather than go without the thing, over that which he actually does pay." According to Marshallian theory of demand, what a consumer is willing to pay for one unit of a commodity measures the money value of his expected utility and what he actually pays measures the monetary cost of the expected utility. The difference between the two values is called *'consumer surplus'*. For example, suppose you are prepared to pay ₹ 500 for a ticket to watch a cricket match but you pay only ₹ 200, the actual price of the ticket match, you have a consumer surplus of ₹ 300.

The Marshallian concept of consumer's surplus and its measurement are graphically illustrated in Fig. 3.5. The consumer's willingness to pay is shown by his straight line demand curve *MN*. The curve *MN* also indicates the utility derived from each successive unit of a commodity. The market price, i.e., the price which a consumer actually pays, is given by *OP*. At price *OP,* the consumer buys *OQ* units. The total utility derived by the consumer from *OQ* units is shown by the area *OMBQ,* for which the consumer pays *OPBQ* = *OQ* × *OP*. Thus, in Marshallian sense, total consumer surplus equals *PMB* = *OMBQ* – *OPBQ*. That is, the shaded area *PMB* represents the consumer's surplus in Marshallian sense when consumer buys *OQ* units of a commodity *X*.
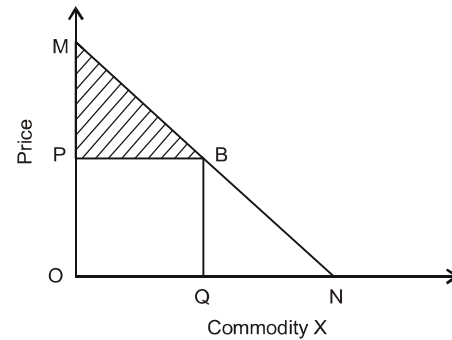
Fig. 3.5 *Consumer's Surplus*

Assumptions: The above analysis of consumer's surplus is based on the following assumptions.

First, it is assumed that the market price is given so that neither the sellers nor the buyers can affect the price and the consumer is not aware of the price. The consumer's surplus will not exist if there is a monopolist and he adopts first degree price discrimination in his pricing policy.

Secondly, the utility is cardinally measurable and *MU* of consumer's money income remains constant throughout.

Thirdly, the utility of each commodity is absolute and is independent of other goods and services consumed by the consumer.

Fourthly, there is no close substitute for the commodity in question. For, if close substitutes are available, there may not be any difference between 'what the consumer would be willing to pay' and 'what he actually pays' for the commodity in question.

Critical Appraisal: The Marshallian concept and measurement of consumer's surplus have been criticised on the following grounds:

First, economists have pointed out difficulties in measuring the consumer's surplus as defined by Marshall and represented by 'a triangle', as shown in Fig. 3.5. In the words of Mark Blaug, "It is sometimes objected that demand curves are usually asymptotic to the price axis. If the individual's offer for the first unit is not defined so that the demand curve does not touch the Y-axis, the integral under the demand curve is infinite. But this objection is easily overcome by measuring consumer's surplus from some selected value of $q_x > 0$[2] (where q_x is the quantity of commodity *X*).

Second, a "more fatal objection" to Marshall's method of measuring consumer's surplus as 'the triangle' under the demand curve is that real income does not remain constant along the demand curve even for 'unimportant' commodities. As price falls along the demand curve (as shown in Fig. 3.5), real income makes the estimate of consumer's surplus as ambiguous one.

Third, it is generally alleged that the assumptions on which Marshallian concept of consumer's surplus is based are unrealistic: *MU* of money does not remain constant; cardinal measurement of utility is not possible; utilities of various goods consumed by a consumer are not independent of each other; most goods have their substitutes—close or remote, and so on. Therefore, it is alleged that the Marshallian consumer's surplus is imaginary and hypothetical.

Fourth, in the ultimate analysis of the consumer's purchases of various goods and services, consumer's surplus is reduced to zero. For, a consumer's willingness to pay (i.e., 'potential price') cannot exceed his income, i.e., what he actually pays out. It means that, when all purchases have been made, the consumers willingness to pay (which equals his income) equals what he actually pays (i.e., his income).

Fifth, the concept of consumer's surplus cannot be convincingly applied to 'essential' and prestigious goods. For example, a starving affluent person may be willing to pay a million rupees for a piece of bread while he may be required to pay only ten rupees. As such, his consumer's surplus will be equal to ₹ 9,99,990 which seems ridiculous. In case of prestigious goods, e.g., rare paintings, diamonds, jewellery, etc., what a buyer is willing to pay, generally, equals what he actually pays. It means there is no consumer's surplus. Thus, Marshallian concept of consumer surplus becomes illusory.

Due to these weaknesses of Marshallian concept of consumer surplus, Samuelson considers this concept as of only 'historical and doctrinal interest' and suggests that 'the economists had best dispense with it'. Hicks has however tried to rehabilitate the consumer's surplus as this concept is of great importance in the *economics of welfare.*

3.5.2 Hicksian Method of Measuring Consumer Surplus

According to Hicks, Marshallian consumer surplus can be measured by using indifference curve analysis also under Marshallian assumption of constant *MU* of money and variable *MU* of money.

Consumer Surplus Under Constant MU of Money: We discuss and illustrate here the measurement of consumer's surplus by using *IC* technique under Marshall's assumption of constant *MU* of money income.

In Fig. 3.6, Y-axis measures consumer's money income, and X-axis measures the quantity of commodity *X*. Assuming money income to be given at *OM,* the line *MN* is consumer's budget line. It shows various combinations of money and commodity *X*, which a consumer can choose from, given the price of *X* at $OM/ON = P_x$. The consumer is shown to be in equilibrium at point *E* on indifference curve IC_2. At point *E*, he buys *OQ* units of *X* for which he pays MP_2 of his income. The amount MP_2 is what the consumer actually pays for $O\widehat{Q}$ ($=P_2E$) units of *X*.
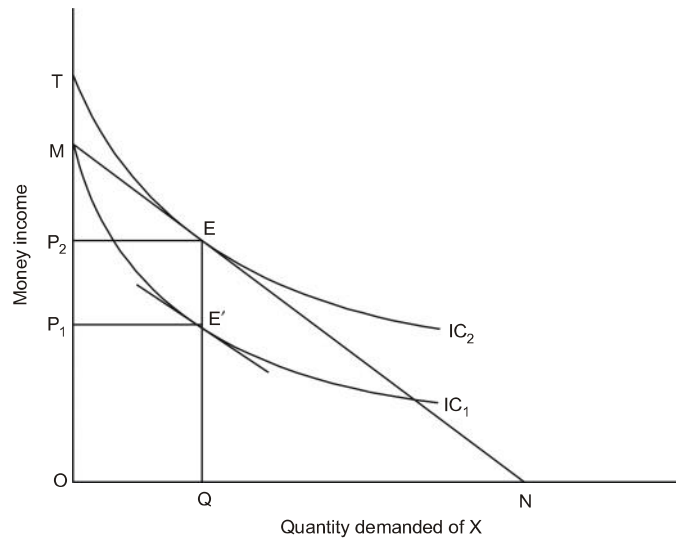
Fig. 3.6 Marshallian Consumer's Surplus through Indifference Curve Analysis

Let us now find what the consumer would be willing to pay for OQ units of X, rather than go witout it. As Fig. 3.6 shows, he would be willing to pay TP_2 for OQ as indicated by the slope of indifference curve IC_2. It implies that he needs an extra income of TM becaus his income is only OM.

Hicks has devised a method to find an equivalent of TP_2 within his budgetary constraints. This he has done by drawing a lower indifference curve having two qualifications (*i*) that the lower indifference curve must pass through point M; and (*ii*) that the lower indifference curve must be vertically parallel to the upper indifference curve. The second condition is necessary to comply with the Marshallian assumption that MU of money remains constant. The indifference curve having these qualifications is shown by indifference curve IC_1. Since the two indifference curves, IC_1 and IC_2, are vertically parallel, they have the same slope at a given quantity. For example, point E on IC_2 and E' on IC_1 refer to the same quantity OQ and have the same slope. It implies that point E' satisfies the equilibrium condition. Note also that IC_1 is simple reproduction of IC_2 at a lower level of utility.

The further analysis can be carried out as follows. Since point M and $E¢$ are on the same indifference curve, IC_1, it means that the consumer would be equally well off at these points. That is, his total satisfaction from OM money income and zero units of X will be the same as from OP_1 of money income and OQ units of X. It means that he would be willing to pay $OM - OP_1 = MP_1$ of his income for OQ units of X. Thus, what the consumer is willing to pay for OQ is MP_1 and what he actually pays (given the P_x) is MP_2. Therefore,

$$\text{Consumer surplus} = MP_1 - MP_2 = P_1P_2 = E'E$$

Since IC_1 and IC_2 are vertically parallel, $E'E = TM$.

The discussion on the consumer surplus takes us to the end of our disucssion on the theory of individual demand for a product.

Check Your Progress

1. List the three attributes of a want.

2. Who invented the indifference technique?

3. What is the main advantage of using the revealed preference theory?

3.6 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. A want has three attributes – desire to buy, willingness to pay and ability to pay.

2. The indifference technique was invented and used by Francis Y. Edgeworth (1881) to show the possibility of exchange of commodities between two individuals.

3. The main merit of the revealed preference theory is that the 'law of demand' can be directly derived from the revealed preference axioms without using indifference curves and most of its restrictive assumptions.

3.7 SUMMARY

- Conceptually, the term 'demand' implies a 'desire' or 'want' for a commodity backed by the ability and willingness to pay for it.

- The desires without adequate purchasing power and willingness to pay do not affect the market, nor do they generate production activity. A want with *three attributes – desire to buy, willingness to pay and ability to pay –* becomes effective demand.

- The indifference technique was invented and used by Francis Y. Edgeworth(1881) to show the possibility of exchange of commodities between two individuals.

- The fundamental departure that indifference curve analysis makes from the Marshallian marginal utility analysis is the hypothesis that utility can be measured only ordinally, not cardinally.

- An indifference curve may be defined as *the locus of points, each representing a different combination of two goods but yielding the same level of utility or satisfaction.*

- When a consumer makes different combination of two goods, yielding the same level of satisfaction, he substitutes one good for another. The rate at which he substitutes one good for the other is called the 'Marginal Rate of Substitution (*MRS*)'.

- The main merit of the revealed preference theory is that the 'law of demand' can be directly derived from the revealed preference axioms without using indifference curves and most of its restrictive assumptions.

- The concept of *consumer's surplus* is believed to have been originated by a French engineer, Arsene Julis Dupuit, in 1844, in his effort to measure social benefit of such collective goods as roads, canals and bridges.

- According to Hicks, Marshallian consumer surplus can be measured by using indifference curve analysis also under Marshallian assumption of constant *MU* of money and variable *MU* of money.

3.8   KEY WORDS

- **Indifference curve:** It may be defined as the locus of points, each representing a different combination of two goods but yielding the same level of utility or satisfaction.

- **Demand:** It is an economic principle referring to a consumer's desire and willingness to pay a price for a specific good or service.

3.9   SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. Mention the assumptions of the ordinal utility theory.
2. How is demand analysed?
3. Define consumer surplus in your own words.

Long Answer Questions

1. Describe marginal rate of substitution.
2. Explain the revealed preference axiom with the help of a diagram.
3. Discuss the Marshallian concept of consumer surplus.

3.10 FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.

UNIT 4 THEORIES OF SEARCH

Structure

4.0 INTRODUCTION

Asymmetric information deals with the study of decisions in transactions where one party has more or better information than the other. This asymmetry creates imbalance of power in transactions which can sometimes cause to transaction to go awry a kind of market failure in the worst case. This unit will introduce you to the theory of search, asymmetric information, lemon and market signalling.

4.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the theory of search
- Explain the theory of asymmetric information
- Analyse the lemon problem
- Define market signalling

4.2 THEORY OF SEARCH

Search theory is a study about transactional friction between two parties that prevent them from a match in an efficient time frame. Search theory has mainly been used to explain inefficiencies in the market for employment, but it also has broad applicability to any form of "buyer" and "seller" whether for a product,

house or even a spouse/partner. Under classical competitive equilibrium models, buyers and seller can transact in a frictionless world with complete and open information; clearing prices are immediately met because supply and demand forces react freely. However, in the real world, this does not happen.

Two parties that wish to transact in business - an employer and job seeker, a seller of a good and a buyer - encounter frictions in their search for each other. These frictions can take the form of mismatched geographies, price expectations and specification requirements, as well as slow response and negotiation times by one of the parties involved. Government or corporate policy may even interfere with an efficient search process.

4.3   ASYMMETRIC INFORMATION

The theory of asymmetric information was developed in the 1970s and 1980s by George Akerlof, Michael Spence and Joseph Stiglitz. The concept of information asymmetry was able to explain many common phenomena that could not be otherwise explained when it was first introduced in the early 1970s. Since then it has become a valuable tool in the field of economics and it is used to explain a diverse set of phenomena. Its significance was established well before the year 2001 when the original authors of the theory, George Akerlof, Michael Spence and Joseph Stiglitz received the Nobel Prize in Economic Sciences. In the paper 'The Market for Lemons: Quality Uncertainty and the Market Mechanism', Akerlof develops asymmetric information with the example case of automobile market.

Asymmetry information deals with the study of decisions in transactions where one party has more or better information than the other. This asymmetry creates an imbalance of power in transactions, which can sometimes cause the transactions to go awry, a kind of market failure in the worst case. Examples of this problem are adverse selection, moral hazard, and monopolies of knowledge. In simple terms, the theory proposes that an imbalance of information between buyers and sellers can lead to inefficient outcomes in certain markets. Asymmetric information is a situation in which either the buyer or the seller know more than the other about the product to exchanged.

Example the Buyer Know More: Let take an example where the buyer knows more. In the insurance business, the two transactors are the insurance company (the seller) and the one seeking insurance (the buyer). In this business, the buyer has more information about the life of the product to be insured. In the market for 'credit' the credit institution, like banks and other financial institutions (the sellers) have less information about the creditworthiness of those seeking credit (the buyers).

Example the Sellers Know More: Let us now take examples where the sellers know more. In the market for used cars the product sellers know much more about the condition of the car than the buyer. Similarly in the labour market, the

worker (the seller) has more information about the quality of labour he can perform, than the employer (the buyer). The same is true about the market for all goods.

Implication: Asymmetric information leads to failure of market. For example, when buyers of product are not sufficiently informed, they are not able to determine the true quality of product at the time of purchase. As a result, most buyers start believing that most of the units of the good that enter in the market are of low quality. This can happen in the case of market for used cars.

Similar problem may arise if the seller knows less than the buyer. When the sellers are not significantly informed, they are not able to differentiate between the genuine and the non-genuine buyers. To avoid the risk of loss they may refuse to sell their product to a particular category of buyers. For example, the insurance companies may refuse to give insurance to the senior citizens.

Market failure due to the buyers being less informed, takes the form of **Adverse Selection** market failure due to sellers being less informed takes the form of **Moral Hazard.** Let us explain each type of market failure and the ways to avoid the failure.

4.3.1 The Adverse Selection

The Lemon Problem: This type of market failure can be best explained with the help of an example. In the market for the used cars, the seller knows much more about the condition of the car than the prospective buyer. The buyer may try many things to judge the quality of the used car, but he has no fool proof method for doing so. There is a psychological aspect of the problem also. The buyer may think that the very fact that the car is for sale, denotes that the car must be of bad quality and the seller want to get rid of it. As a result, the buyer is always suspicious of the quality. The slang used for bad quality is *Lemon.* The buyer thinks that car is a lemon the current of such thinking in the mind of the buyer is called *lemon problem.*

The suspicion of quality in the minds of buyers leads them to believe that most of the used cars sold are of low quality. This brings down the market price of the used cars. As a result, the sellers of high quality used cars finding the price too low withdraw from the market. Only low quality cars are offered for sale. *The low quality goods drive away high quality goods from the market*. The buyer offers the same price for both low and the high quality used cars because they are not able to determine the quality.

Graphically representation: We can explain the adverse selection with the following example:

Figure 4.1 (a) shows the market for the high quality cars and figure 4.1 (b) for the low quality cars. Let us first assume that there is no asymmetric information. In this case, the higher quality cars are sold at higher price and the low quality cars are sold at a lower price. The equilibrium price for the high quality cars is OP_H and for the low quality cars is OP_L.
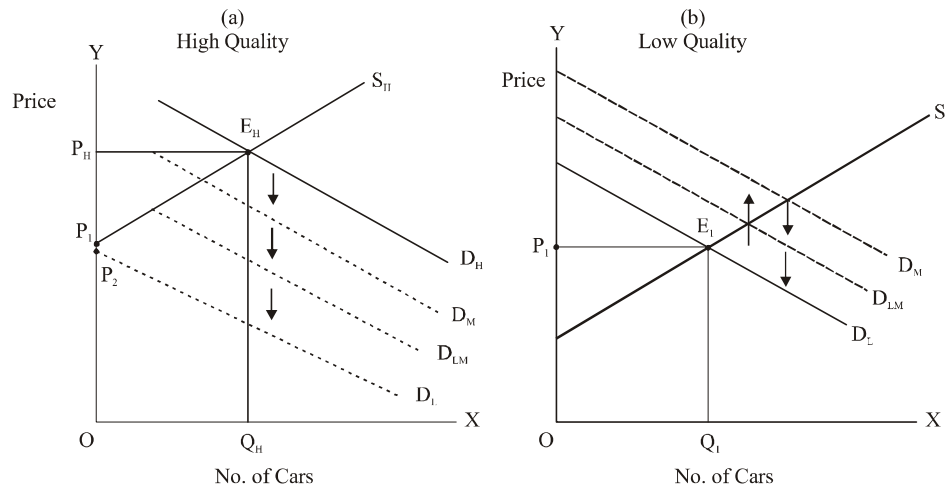
Fig. 4.1(a) and (b) *Graphical Representation of Asymmetric Information*

Now assume that there is asymmetric information. The buyers do not know about the quality of the used cars. They only guess. Suppose the initial guess is that all cars are of medium quality. This shift demand curve of higher quality cars D_H down wards to D_M and that of lower quality cars from D_L upward to D_M. M signifies medium quality.

This thought of the buyers increases the sale of low quality cars. This makes the buyers realise that most cars sold are of low quality. This thinking further shifts D_M curve of high quality cars downwards to D_{LM} curve, and also in case of low quality cars downwards to D_{LM}. LM signifies lower medium quality. This further increases the sale of low quality cars.

The shifting demand curves continue until only the low quality cars are sold. In both cases, it settles at D_L. See what happens in case of high quality cars. The minimum price at which the sellers are willing to start selling high quality cars in OP_1. The buyers on the other hand, on the assumption that all the cars of low quality, are not prepared to pay more than OP_2 for any car. Since OP_1 is higher than OP_2, no high quality car will be offered for sale. Eventually, only low quality cars will be sold. In this way, low quality cars drive away high quality cars out of market. This is *lemon problem*.

The seller having less information

In the above case, the buyer had less information. Let us now take an example where the seller has less information. In the market for life insurance, the buyers know much more about their general health than any insurance company hopes to know. Even a medical examination of the buyer may not reveal true health of the buyer. As a result, unhealthy people are more likely to purchase insurance. This results in rise of the proportion of the unhealthy people seeking insurance.

This forces the price of insurance to rise. As a result of high price, healthy person choose not to be insured. This further raises the proportion of unhealthy

people and a further rise in the price of insurance. The process continues until nearly all who want to buy insurance are unhealthy. This makes insurance business unprofitable. The insurance companies start refusing insurance to the people above a particular age assuming every person in that age group to be unhealthy. For example, insurance companies refuse to cover senior citizens.

There are many other examples of asymmetric information. In the market for credit, the lender has incomplete information about the borrower. In case of restaurants, the customers have incomplete knowledge about the food quality served. In case of retails stores, a buyer has incomplete knowledge about the after-sales service.

What is the solution? When the buyers have incomplete information and if the sellers want to save themselves from the problem to stay in the market, the sellers must adopt some method of convincing the buyers about the quality of the products they offer.

4.3.2 Ways to deal with the Lemon Problem

Some of the ways of dealing with problem are:

(1) **Create reputation about the quality**: This is to convince the buyers about the quality of the product. In the absence of complete information, buyers do go by the reputation of the sellers.

Developing reputation is not an easy task. Sometimes no matter what efforts a seller makes, he may not be in a position to develop reputation. His nature of business may be such that he does not have regular customers stores as the bus stands, railway stations, highways etc. are of such nature. A business must have regular customers to develop a reputation.

(2) **Standardisation of the products:** When it is not possible to develop a reputation, the alternative is standardisation. Customers when visit unfamiliar stores ask for only the standardised, and well-known for quality products. For example, the visitors to bus stands, railway station, exhibitions and others look for standardised products, the known brands in comparison to the unknown brands of the cold drinks, biscuits, packed foods etc. the visitors are prepared to pay more for the standardised products.

(3) **Market signalling:** Market signalling is a process by which the sellers send signals to the buyers conveying information about the product quality.

Take an example of labour markets. The market suffers from asymmetric information. The employers (buyers) know much less about the productivity potential of the prospective employees (sellers). This create adverse selection problem. The employer firms would be much better off if they knew how productive potentials of the employees were before they hired them. How can a firm do this? The firm must look or a signal to separate low productive workers from high productivity workers. For example, education of the employees is a strong signal of this type.

The firm can evolve criteria to measure the education level of a prospective employee. The criteria may be the number of years of schooling, degree, diploma, reputation of the school, college, university, the marks obtained, the project reports, etc. The firm can use education as a criterion in the selection of workers assuming that more productive workers attain higher level of education.

The workers, on the other hand, can use their education level as a market signal to obtain better paying jobs. In India, the firms do offer highly paid jobs to the MBAs, MCAs, BEs, and CAs from reputed institutes. This is market signalling**.**

Education as a market signal: A model

How do workers decide as to how much to spend on education, to use education as market signal. Workers compare cost and benefit of obtaining higher education. A worker goes in for a higher education level only when the benefit is either greater, or at least equal to, the cost.

The cost is on tuition fees, books, opportunity cost of foregone wages, the psychic cost of having to work hard, etc. The cost is greater for the low productivity group than the high productivity group because of the two reasons. First, the low productivity group may be less studious. Second, the low productivity group progresses slowly and takes a longer time to obtain a particular level of education.

The benefit is the increase in wage associated with each level of education. If we assume that firm do pay higher wages to the higher educated people, and since education involves cost, the workers follow the following criterion:

Benefit ≥ Cost

They decide to go in for higher education if the benefits are greater, or at least equal to, the cost. If the benefit is lower than the cost, they decide not to go to in for higher education.

Illustration 4.1: Suppose the firm, suffering from the adverse selection, decide to use education as a signal to separate the 'Low Productivity' (LP) group from the 'High Productivity' (HP) group. Further suppose that the firms decide to give ₹ 60,000 per annum to the LP group and ₹ 1,20,000 to the HP group. The firms expect an average of 10 years of work from each group employee. It means that the firms would pay ₹ 6 lakh to an employee from LP group and ₹ 12,00,000 to an employee from the HP group over a period of 10 years.

Let the education levels be measured in terms of level 0,1,2,3 ... etc. Suppose the level below 4 is the LP group and above level 4 is the HP group. Over 10 years period, a LP worker earns ₹ 6 lakh and a HP workers ₹ 12 lakhs.

The education level range for the group is 0-3. It means that even at the 0 level the worker earns the same as at level 3. So, in this group no education is necessary. However, to reach the HP group, level 4 and above is a must.

The worker will try to attain only level 4 because he has no extra benefit at level 5 or above, but he has to incur extra cost. The choice of the worker is thus limited to attain level 4 or remain at 0. He will try to attain level 4 only if the benefit of salary at level 4 is greater, or at least equal to, the cost of reaching that level. If the cost is greater than the benefit, the worker will drop the idea of attaining level 4.

Suppose, the LP work has to incur ₹ 2.5 lakh per level of education, and the HP worker ₹ 1.25 lakh per level. To reach level 4 the LP worker must spend ₹ 10 lakh, while the benefit to him only ₹ 6 lakh. He gets 67 lakh even if his education level is 0. Since the cost is greater than benefit, the LP worker decides not to go for attaining level 4.

The HP workers has to spend only ₹ 5 lakh to reach level 4. The benefit to him is ₹ 6 lakh. Since the benefit is greater than the cost, the HP workers decides to go in for higher education.

The overall result is that only the HP workers will go in for higher education. It makes the higher education serve as a signal of higher productivity to the employer firm and solve the problem of adverse selection. The firm is able to calearly identify the LP and the HP groups. Graphically:- let the education level be termed L. We are given the cost function of the two groups as:

C (LP)=2.5L

C(HP)1.25L

Where C is Total Cost of education.

Refer to the figure (a) and (b). In (a) , the LP group graph, the total cost curve is higher than the benefit curve throughout. Therefore the LP group will not go in for higher education. In (b) the HP group graph, the benefit cureve is higher than the cost curve at level 4 . The HP group will go in for higher education.

(a)
Low Productivity Group

(a)
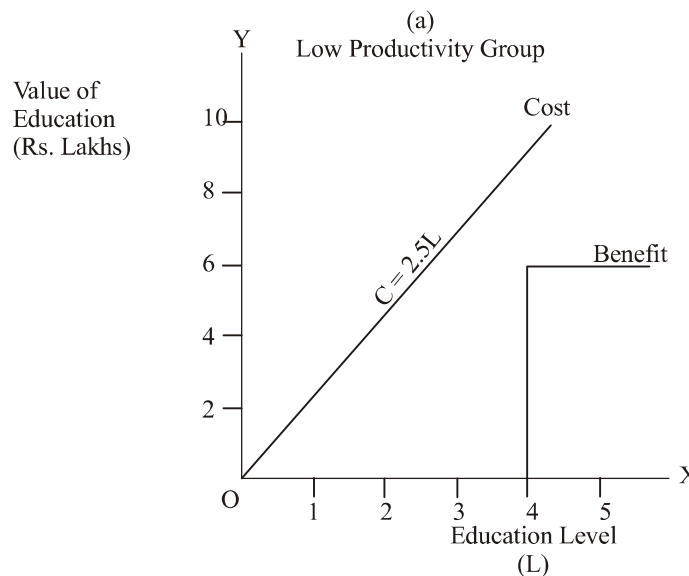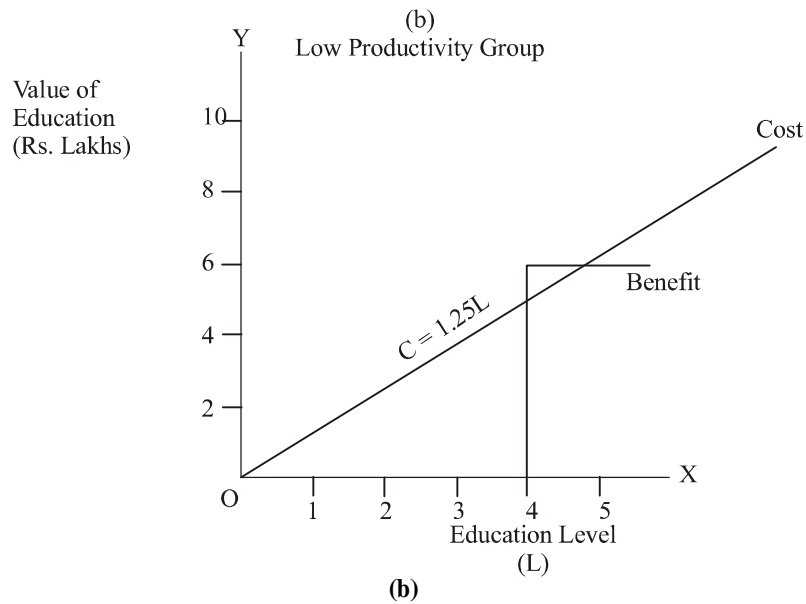
Fig. 4.2 *Graphical Representation of Market Signal*

(4) Give Guarantees and Warranties: This is another way before the firms to signal that their product is of higher quality. The general assumption is that only the firms producing the high quality products are in a position to give guarantees and warranties.

If a defect is discovered in the product of the firm, the firm has to either replace the product with new one or service the defect. Doing so involves high cost. The firms producing low quality products will not offer the guarantees and warranties because it is a costly affair. The firms producing high quality products will gladly offer the same knowing that the chances of defect arising in their products are low. Even if defect occurs they would gladly service it because they are already charging a high price. It then becomes a signal to the buyer that the firm offering guarantees and warranties product high quality products. This solves the firm's problem of adverse selection.

4.4 THE MORAL HAZARD PROBLEM

Meaning: Moral hazard arises when the cost of the behaviour of the buyers is borne by the seller. This is a symmetric information problem. The problem is best explained by taking an example.

Example: Suppose a person's home is fully insured against theft. When the house was not insured the owner took extra care to see that doors are properly locked. When the house is insured the owner may become a little careless about locking the door and other security measures. He knows that he will be fully compensated by the insurance company in case of their theft. The insurance changes the behaviour of the insured. This increases the possibilities of thefts, and in turn, increases the

cost of insurance, for no fault of the insurance company. This is the problem of moral hazard.

Implication: The increase in the cost of insurance forces the insurance company to increase premiums for everyone. The company may also refuse to sell insurance at all.

The problem arises because the insurance company (the seller) cannot observe the actions of the insured (the buyer). As a result, the insurance company has to part with higher payments as claims.

The problem can arise in many other cases. It can arise to the employers who cannot effectively monitor the behaviour of the workers. The workers take advantage and perform below their capabilities. The problem can also arise to the companies offering guarantees and warranties who cannot effectively monitor the behaviour of their customers who become careless in handling the products sold to them.

By increasing cost of the producer moral hazard leads to inefficient allocation of resources.

4.4.1 The Principal-Agent Problem

The problem is that the agents pursue their own goals. Even at cost of obtaining lower profits for the principal. Let us take an example. In most firms, the owners cannot monitor everything that the employee do. As a result, the employees put in lesser efforts than they are expected to do. This creates the principal-agent-problem. In this example, owners are the principals and the employees are the agents.

Let us take another example. In the most of the private corporate firms, individual shareholders have only a small percentage of the firm's total equity. This makes it difficult to monitor the performance of the managers. Realising the helplessness of the shareholders (principal), the managers (agents) pursue their own goals rather than the goal of the shareholders. Managers are more interested in their perks, power, fringe benefits and long job tenure. They can achieve these objectives more by maximizing total revenue rather than total profits.

There may be checks but these checks do not produce the desired results because it is not possible to apply these checks efficiently.

- One check may be that the shareholder can complain.
- Another check is that if the firm is badly managed some other firm may take over and appoint new managers.
- Third, the managers who pursue owners' goal may be in great demand and earn higher wages.

The check exists but it may not be possible to apply them efficiently. For example, the corporation takeover may be motivated not by efficiency but for personal and economic power. The market for managers may be imperfect.

Therefore, the solution lies somewhere else. The solution may lie in certain incentives provided to the managers so that they pursue the goals of the owners.

The same problem may be faced in public enterprises. The managers may be interested more in perks and power which they can obtain by expanding the firm beyond the efficiency level. There are legislative checks but they may be less effective because the managers have more information than legislature.

The Solution: The principal-agent problem is that the principal (owner) cannot effectively monitor the activities of agents(managers). Taking advantage of this factor, managers pursue their own goal and ignore the profit maximisation goal of the owners.

The solution must be such that the agents come as close as possible in meeting the owner's goal. Since the monitoring is ineffective, the solution lies in giving incentives to the agents to induce them to follow the principal's goals.

How does the incentive work to the advantage of both the owners and the managers is explained with the help of an example.

The Example of Solution: Let the degree of effort put in by the manager be designated 'a'. Let the low effort have a=0 and the high effort a=1. Putting in high effort by managers involves extra cost in terms of the lost leisure and unpleasant work time. Let this extra cost be designated as 'C'. Suppose the extra cost is ₹ 10,000 per degree of effort. This make the extra cost function C=10,000a.

Suppose when the manager put in low effort (a=0) the total revenue (R) of the firm is expected to be ₹ 10,000 or ₹ 20,000 with equal profitability. It means that the average expected R is ₹ 15,000 =(10,000+20,000/2). Suppose further that when the manager puts in extra effort (a=1) the revenue (R) of the firm is expected to be ₹ 20,000 or ₹ 40,000 with equal probability. This gives an average R of ₹ 30,000.

Given the R outcome, the incentive scheme to be adopted will depend upon the nature of production, the degree of uncertainty and the objectives of the owners and the managers. The best scheme for the managers is one which gives them a share in the earnings of the firm. It means that in addition to the regular wage they get incentive wage (W_1). Suppose the owners offer the following incentives scheme:

If the R =10,000 or 20,000 $W_1 = 0$

If the R= 20,000 or 40,000 $W_1 = 24,000$

It means that if the managers put in low effort (a=0) the W_1 =0. If they put in high effort (a=1) the $W_1 = 24000$. To the owners the average expected W1 =12000 (=0+24000/2) . if they put higher effort.

Suppose to put in high effort (a=1) the manager incurs a cost of ₹ 10,000. Since the W_1 is ₹ 12,000 and the cost is 10,000 the net W_1 is ₹ 2,000(=12,000-10,000). To the owners average expected R =30,000 and the cost (=W1) is ₹ 12000. The expected profit for the owners is thus ₹18,000(30,000-12,000). The scheme makes both the managers and the owners better off.

An Alternative: An alternative to the bonus can be the revenue sharing scheme. Suppose the scheme is : Incentive wage = 18000

It means that if R= 18,000   W1=0 i.e. R sharing is Zero. If the manager makes high effort the expected R= 30,000. Then W1=30,000 - 18000= 12,000. It means that R sharing is 12,000. Subtracting the cost of extra effort ₹ 10,000, the net W1 = 2000.

In this example, the outcome of both the alternative is the same, but may not necessarily be the same in all cases.

Therefore, when it is not possible to monitor the efforts put in by the agents, an incentive scheme that rewards the outcome of high levels of efforts can induce the agents to pursue the goals the owners set.

4.4.2 Shirking in the Labour Market: A Moral Hazard

The Problem: It is another principle-agent problem. The model assumes perfectly competitive markets in which all the workers are equally productive and earn the same wage. This is according to the Marginal Productivity Theory of Wages.

The model says that if all the workers get the same wage and if all the workers are employed, they have an incentive to shirk. The argument is like this, if a worker is caught shirking and is removed, he can easily get hired by some other firm and at the same wage because of perfectly competitive condition in the market. The threat of removal does not work as a disincentive for shirking or as an incentive to be productive.

The solution: Efficient wage

The model says that an incentive for not shirking the firm must offer workers a higher wage. If even at a higher wage a worker shirks, get caught and removed he can find employment elsewhere but only at lower rate. This works as a disincentive for shirking and the firm has no problem. The wage at which no shirking occurs is called *efficiency wage*. This works as a shirk disincentive wage. The shirking model is, therefore also called efficiency wage theory.

4.4.3 Advantages and Disadvantages of Asymmetric Information

Advantages of Asymmetric Information

- Growing asymmetrical information is a desired outcome of a market economy. As workers specialize and become more productive in their fields, they can provide greater value to workers in other fields. For example, a stockbroker's services are more valuable to customers who do not know enough to buy or sell their own stocks with confidence.

- One alternative to ever-expanding asymmetric information is for workers to study in all fields, rather than specialize in fields where they can provide the most value. Associated with this alternative are large opportunity costs

and possibly a lower aggregate output, which would lower standards of living.

- Another alternative is to make information abundantly available and inexpensive, such as through the internet. It is important to note that this does not replace asymmetric information. It only has the effect of moving information asymmetries away from simpler areas and into more complex areas.

Disadvantages of Asymmetric Information

- In certain circumstances, asymmetric information may lead to adverse selection or moral hazard. These are situations where individual economic decisions are hypothetically worse than they would have been had all parties possessed more symmetrical information. Most of the time, the solutions to adverse selection and moral hazard are not complicated.

- Consider adverse selection in life insurance or fire insurance. High-risk customers, such as smokers, the elderly, or those living in dry environments, may be more likely to purchase insurance. This could raise insurance premiums for all customers, forcing the healthy to withdraw. The solution is to perform actuarial work and insurance screening and then charge different premiums to customers based on their associated potential risks.

Check Your Progress

1. Who developed the theory of asymmetric information and when?
2. State one disadvantage of asymmetric information.

4.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The theory of asymmetric information was developed in the 1970s and 1980s by George Akerlof, Michael Spence and Joseph Stiglitz.

2. In certain circumstances, asymmetric information may lead to adverse selection or moral hazard. These are situations where individual economic decisions are hypothetically worse than they would have been had all parties possessed more symmetrical information. Most of the time, the solutions to adverse selection and moral hazard are not complicated.

4.6 SUMMARY

- Search theory is a study about transactional friction between two parties that prevent them from a match in an efficient time frame.

- Two parties that wish to transact in business - an employer and job seeker, a seller of a good and a buyer - encounter frictions in their search for each other.

- The theory of asymmetric information was developed in the 1970s and 1980s by George Akerlof, Michael Spence and Joseph Stiglitz.

- Asymmetry information deals with the study of decisions in transactions where one party has more or better information than the other.

- The low quality goods drive away high quality goods from the market. When it is not possible to develop a reputation, the alternative is standardisation.

- Market signalling is a process by which the sellers send signals to the buyers conveying information about the product quality.

- Moral hazard arises when the cost of the behaviour of the buyers is borne by the seller. This is a symmetric information problem.

- Growing asymmetrical information is a desired outcome of a market economy. As workers specialize and become more productive in their fields, they can provide greater value to workers in other fields.

4.7 KEY WORDS

- **Market signalling:** It is a process by which the sellers send signals to the buyers conveying information about the product quality.

- **Perfect competition:** It is a market structure where a large number of buyers and sellers are present, and all are engaged in the buying and selling of the homogeneous products at a single price prevailing in the market.

- **Shirking:** It is defined as the tendency to do less work when the return is smaller.

4.8 SELF ASSESSMENT QUESTIONS

Short Answer Questions

1. What is asymmetric information?
2. What is lemon problem?
3. What is moral hazard problem?
4. What is efficiency wage?

Long Answer Questions

1. Explain the meaning and implication of asymmetric information with the help of an example.

2. Discuss the problem of adverse selection and represent it graphically.

3. Describe the ways of dealing with lemon problem.

4. Explain the meaning and implication of moral hazard.

5. What is the problem of shirking in the labour market? What is the solution and what are the implications of the solution?

4.9 FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.

UNIT 5 THE EFFICIENT MARKET HYPOTHESIS

Structure

5.0 INTRODUCTION

In the previous unit, you studied about the theory of search, asymmetric information, market signalling and the moral hazard problems.

Efficient market theory states that share price fluctuations are random and do not follow any regular pattern. Meanwhile, technical analysts see meaningful patterns in their charts. This raises these questions: Does the intrinsic value of stock have any meaning? Is it related to the stock price?

5.1 OBJECTIVES

After going through this unit, you will be able to:

- State the meaning of efficient market hypothesis

- List the types of efficient market hypothesis

- Identify the limitations of efficient market hypothesis

5.2 THE EFFICIENT MARKET HYPOTHESIS

5.2.1 Meaning

One should know the meaning of phrases such as market efficiency, liquidity traders and information traders, before trying to understand the random walk theory.

Market Efficiency

The expectations of investors regarding future cash flows are translated into or reflected in share prices. Market efficiency is the accuracy and speed with which the market translates the expectation into prices. There are two types of market efficiencies:

- Operational efficiency
- Informational efficiency

Operational efficiency

Factors like the time taken to execute the order and the number of bad deliveries measure the operational efficiency of a stock exchange. Operational efficiency of the market is a matter of concern for the investors. The efficient market hypothesis (EMH) does not take into account this efficiency.

Informational efficiency

It is a measure of swiftness with which the market reacts to new information. The market frequently receives new information in the form of economic reports, company analysis, political statements and notification of new industrial policy. How does it react to this? Security prices adjust themselves rapidly and accurately. They never take a long time to adjust to new information. For instance, a company's announcement of the issue of bonus shares can lead to a rise in the price of the stock. Likewise, stock index movements reflect significant changes in the policy decisions of the government.

Liquidity Traders

These traders' investments and resale of shares depend upon their personal fortune. Liquidity traders may sell their shares to pay their bills. They do not analyse before they invest.

Information traders

Information traders base their buy or sell strategy on proper analyses. They estimate the intrinsic value of shares. They enter the market on the basis of the deviation of the market value of shares from the intrinsic value. They sell if the market value is higher than the intrinsic value and vice versa. The buying and selling of shares triggered by demand and supply forces brings the market price back to its intrinsic value.

Arbitrage

The concept of arbitrage provides the basis for the efficient market hypothesis. The arbitrageurs absorb the potential excess return opportunities in the market. Let us take one example. Arnica stock is expected to yield 15 per cent return per annum. If the current price is lower than the expected price, making the return

high, there is an unexpected opportunity for exceptional gain. The arbitrageurs buy heavily, which leads to a rise in Arnica's stock price. Thus, the equilibrium return and the expected return become equal. This eliminates the possibility of earning an abnormal return.

Likewise, if the optimal forecast of return is negative, say minus 2 per cent, and the equilibrium return is 15 per cent, this investment is not worth it. The arbitrageurs may sell the stock leading to a drop in price. This makes the optimal forecast price equal to the equilibrium price.

The Random-walk Theory

In 1900, French mathematician Louis Bachelier wrote an article suggesting that security price fluctuations were random. In 1953, British statistician Maurice Kendall reported that the stock price series is a wandering one. Each change is independent of the previous one. In 1970, American economist Eugene Fama stated that efficient markets fully reflect the available information. If markets are efficient, security prices reflect normal returns for their level of risk. Fama suggested that the efficient market hypothesis can be divided into three categories. They are the 'weak form', the 'semi-strong form' and the 'strong form'. The level of information considered in the market is the basis for this segregation. Figure 5.1 illustrates market efficiency.
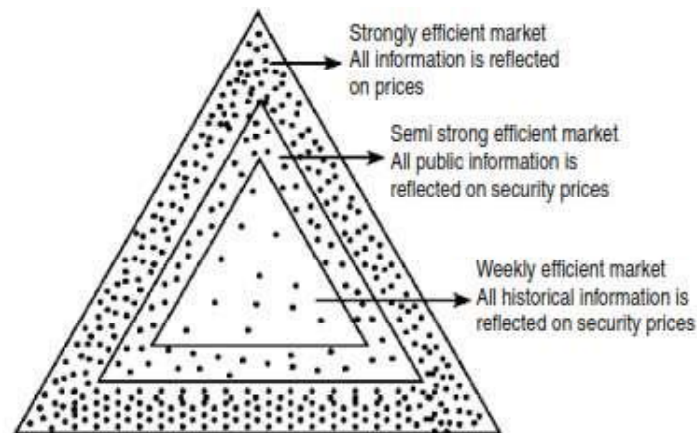
Fig. 5.1 *Levels of information and the markets*

Weak form of Efficient Market Hypothesis

Historical prices form the basis of information used in the weak form of efficient market hypothesis (EMH). According to it, current prices reflect all the information found in past prices and traded volumes. Future prices cannot be predicted by analysing past prices. Everyone has access to past prices, even though some people can get these more easily than others. Liquidity traders may sell their stocks without considering the intrinsic value of the shares and cause price fluctuations. Buying and selling activities of the information traders lead the market price to align itself with the intrinsic value. Figure 5.2 shows a weakly efficient market.
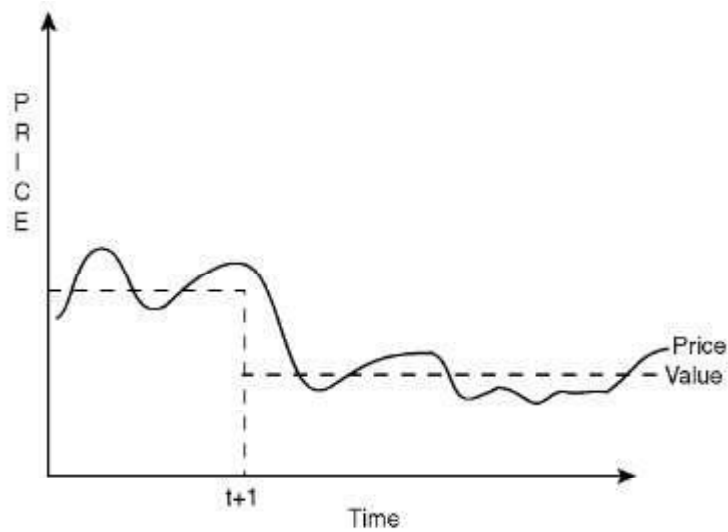
Fig. 5.2 *A Weakly Efficient Market*

The dotted line in Figure 5.2 represents the intrinsic value. The intrinsic value changes at times t and t + 1. In the weak form of market, the price of the stock and its intrinsic value diverge significantly. The supply and demand for the stock or any other asset has a tendency to move towards the equilibrium return. Supply and demand match each other in the equilibrium return. Thus,

$E(R) = Eq (R)$

Where, $E(R)$ is the expected return and $Eq (R)$ is the equilibrium return.

The expected return is equal to the optimal forecast of return, which is $F(R)$.

Then,

$F(R) = Eq (R)$

This indicates that the current price will be equal to the optimal forecast of a stock's return, which in turn is equal to the equilibrium return. The optimal forecast uses all the available information in the market. Hence, EMH states that a security's price reflects all the available information.

In the weak form of efficient market, short-term traders may earn a positive return. On average, short-term traders will not outperform the blindfolded investor picking the stock with a dart. That is, traders may earn by the naive buy-and-hold strategy and while some may incur a loss, the average buy-and-hold strategist cannot be beaten by the chartist. Many studies by market analysts have proven the weak form of EMH. Empirical tests of the weak form are presented here.

Filter rule: Investors use technical trading strategies based on historical prices to earn returns. Filter rule is one such strategy. According to it, if the price of a security rises by X per cent, an investor should buy and hold the stock until its

price declines by at least X per cent from a subsequent high. Short sellers can use the filter to earn profits by liquidating their holdings when the price decreases from a peak level by X per cent. They can take up short position as the price declines till the price reaches a new low and then increases by X per cent. Different traders use different filter rules. It ranges from as low as 0.5 per cent to as high as 50 per cent.

The filter rule can be explained with the help of an example. Take a hypothetical company XY and assume the filter to be 10 per cent. The price fluctuates between ₹ 20 and 30. Assume the starting point is ₹ 20. When there is an increase in the price of the share to ₹ 22, i.e., a 10 per cent rise, one buys it. The rally may continue up to ₹ 30 and decline. A fall in the price gives a sell signal at ₹ 27, i.e., 10 per cent of ₹ 30, and the trader can take up a short position till it reaches its low level. When there is an increase in price, the same exercise is followed.

Several studies have found that after commissions the average gains produced by the filter rules are far below the gains of a naive buy-and-hold strategy adopted by the investor.

Runs test: A runs test finds out whether the series of price movements occurs by chance. A run is an uninterrupted sequence of the same observation. Tossing a coin gives the following sequence of occurrence.

H H T T T H H H T H H

Here, occurrence of H H is a run and T T is another run. When the sequence of observations changes, it is counted as a run.

$$Runs\ test\ Z = \frac{R - \bar{X}}{\sigma}$$

$$R = \text{number of runs}$$

$$\bar{X} = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

$$\sigma^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

R = number of runs

$n_1 + n_2$ = number of observations in each category

σ = standard deviation

Z = standard normal variate

The following example explains the calculation of runs test. Table 5.1 gives the Real company stock prices along with their runs.

Table 5.1 *Real Company Stock Prices*

Date	Price	Runs	Date	Price	Runs
Sept. 20	43.05		Oct. 19	54.70	
21	43.40	+1	20	58.95	+11
22	41.75	−2	21	60.30	
23	42.65	+3	25	59.65	
24	43.60		26	58.65	
28	43.05	−4	27	56.80	
29	43.40	+5	28	53.50	−12
Sept. 1	46.80		29	51.50	
4	46.60	−6	Nov. 1	48.40	
5	47.50	+7	2	52.30	+13
6	47.40	−8	3	56.05	
7	52.15		4	55.15	−14
8	52.50	+9	5	56.40	
11	53.45		7	57.15	
12	57.55		9	57.25	+15
13	57.45		10	57.55	
14	55.90	−10	11	56.75	−16
15	54.15				

$$Runs\ test\ Z = \frac{R - \overline{X}}{\sigma}$$

$$R = \text{number of runs}$$

$$\overline{X} = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

$$\overline{X} = \frac{2 \times 19 \times 15}{34} + 1 = 17.76$$

$$\sigma^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

$$\sigma^2 = \frac{2 \times 19 \times 15 [(2 \times 19 \times 15) - 19 - 15]}{(19 + 15)^2 (19 + 15 - 1)} = 8.01$$

$$\sigma = 2.83$$

$$= \frac{16 - 17.6}{2.83} = -0.565$$

The consecutive rise in prices would be counted as a positive run and the decline would be counted as a negative run.

According to the probability theory, 95 per cent of the area under the normal curve lies within ± 1.96 standard deviation of the mean. Since the calculated value of minus 0.565 is less than minus 1.96, the runs have occurred by chance.

Published results of the studies using runs test suggest that runs in the price series of stocks are not significantly different from the runs in the series of random numbers.

Serial correlation The serial correlation technique tests the independence of successive price changes. Serial correlation or autocorrelation measures the correlation co-efficient in a series of numbers with the lagging values of the same series. Price changes in period $t + 1$ (or $t +$ any number) are correlated with the price changes of the preceding period. Scatter diagrams can be used to find out the correlation. If there is a correlation between the price of t and $t + 1$ period, the points in the graph will form a straight line. If the price rise (or fall) in period t is followed by price increase (or fall) in period $t + 1$, then the correlation co-efficient will be $+1$. Many studies conducted on the security price changes have failed to show any significant correlations. Fama computed serial correlations for 30 stocks for the period 1958–62 with varying t periods from $t + 1$ to $t + 10$. The values of the auto-correlations are usually insignificant with multiple values falling within the range of 0.10 to minus 0.10. If there is little correlation between stock prices over time, chart analyses cannot be of much use in predicting the future.

5.2.2 Types

The National Stock Exchange has carried out research on the functioning of the market. The research deals with the serial correlation factor in index returns.

This study looks at whether the Indian stock market is predictable by using returns of S&P CNX Nifty index. Serial correlation with lag one is calculated. If markets are informational efficient, the serial correlation will not be statistically significant.

Table 5.2 shows the monthly serial correlation for three years 1995–96, 1996–97, and 1997–98. Logarithmic returns are used to check serial correlation at lag one. Table 5.2 also gives the t-statistics, which help to determine if the calculated correlation parameter is different from zero in the statistical sense.

Table 5.2 *Predictability of Stock Returns*

Month	1995-96		1996-97		1997-98	
	Correlation	T Statistic	Correlation	T Statistic	Correlation	T Statistic
Apr	0.68	2.65	–0.12	–0.52	–0.31	–1.31
May	0.30	1.28	0.05	0.22	0.38	1.75
June	0.33	1.53	0.21	0.94	0.23	1.04
July	0.49*	2.18	0.02	0.11	0.17	0.81
Aug	–0.24	–0.99	0.31	1.41	0.26	1.15
Sep	0.29	1.31	–0.03	–0.16	0.42*	1.99
Oct	0.15	0.64	0.21	0.98	–0.35	–1.46
Nov	0.09	0.42	0.05	0.23	–0.21	–0.92
Dec	–0.04	–0.17	–0.33	–1.45	0.11	0.51
Jan	0.41	1.94	0.04	0.20	0.34	1.54
Feb	0.23	1.02	0.38	1.75	0.25	1.10
Mar	0.13	0.57	0.14	0.60	–0.17*	–3.17

On the whole, out of the 36 months analysed, the correlation coefficients were statistically different from zero in only four months. The results thus indicate that the S&P CNX Nifty is an efficient index, and that the information assimilation and price adjustment processes are quick.

Semi-strong Form

The semi-strong form of the EMH states that the security price adjusts quickly to all publicly available information. In semi-strong efficient markets, security prices fully reflect all publicly available information. The prices not only reflect past prices, but also the available information regarding the earnings of the corporation, dividend, bonus issue, right issue, mergers, acquisitions, and so on. In the semi-strong efficient market, a few insiders can earn a profit on short-run price changes rather than investors who adopt the naive buy-and-hold policy.

In the case of a competitive market, supply and demand forces determine the price. The price at the equilibrium level of supply and demand represents the consensus opinion of the market. The intrinsic value of the stock and the equilibrium price are the same. Whenever new information arrives on the market, supply and the demand factors react to it. If the market processes the new information quickly, a new price will come out of it. To be semi-strongly efficient, a market needs timely and accurate dissemination of information and assimilation of news. Only then can the market reflect all relevant information quickly. Studies show that stock markets in the US strongly support the semi-strong hypothesis because the prices adjust rapidly to new information.

Empirical evidence Fama, Fisher, Jensen and Roll (1969) were the forerunners in assessing the semi-strong form of EMH. They analysed the impact of stock split on share prices. Their study was noteworthy because of the following reasons:

- It provided evidence of a semi-strong form of the market.
- It analysed if stock splits increase the wealth of the shareholders.
- It helped researchers develop a design to test market efficiency.

The authors developed a method to compute abnormal returns by using the simple regression technique. Regressing the security return against the return of the stock market index gives the normal return. The following equation shows it.

$$r_{it} = \alpha_1 + \beta_1 r_{mt} + e_{it}$$

Where,

r_{it} = realised return for the i the stock in the period t
r_{mt} = realised return for index in period t
α_1, β_1 = regression coefficients
e_{it} = error term, or residual for the period t

The normal return for any period is as follows:

Normal return $= a_1 + \beta_1 \, r_{mt}$

Here, e_{it} indicates the abnormal return.

For any period, AR_{it}

$e_{it} = r_{it} - (a_1 + \beta_1 \, r_{mt})$

$e_{it} = AR_{it}$

This method of estimating the abnormal return is called the residual analysis. The regression equation represents normal return and e_{it} represents abnormal return.

The average abnormal return (AAR) can be obtained by adding the abnormal returns over time and dividing it by *n*. The AAR can be measured around a date of event or the announcement date of stock split or bonus issue

$$AAR_t = \frac{1}{n} \sum_{i=1}^{N} AR_{it}$$

CAAR = SAAR. Adding the AAR_{it} for each period gives the cumulative average abnormal return (CAAR). Period of study begins several weeks before the event takes place and ends several weeks after the event. The CAAR provides a snapshot of average price behaviour of securities over time. If the markets are efficient, the CAAR should be close to zero.

The authors reviewed hundreds of cases of efficient corporations and heterogeneous sample periods to study the effect of stock splits. They examined 940 stock splits from 1927 to 1959 in the New York Stock Exchange. The price behaviour was analysed 29 months before and after the date of the stock split. They found the CAAR for all 940 observations. They found the level of CAAR to be insignificant and that it essentially fell from the date of announcement of the split. According to them, the simple strategy of buying shares after a stock split did not appear to generate abnormal returns. The study results provide evidence for the semi-strong form of EMH.

Ball and Brown analysed the market's ability to integrate the reported annual earnings per share. Their study showed that the actual good price earnings were higher than the predicted good price earnings and at the same time, the low price earnings were lower than the predicted low price earnings. Ball and Brown found that even before the announcement of a good report, the respective shares experienced an increase in price. Likewise, even before the announcement of the negative earnings report, the share prices decreased. Both the groups generated only normal returns after the announcement providing support for the semi-strong form.

Scholes in his study found that the market is also efficient in identifying the seller. He analysed the price effects of large secondary offerings. Usually, price

tends to fall before the secondary offering. This is mainly due to the information effect and not due to selling pressures. If heavy selling is associated with the corporation's members and officers, the prices tend to fall at a faster rate. If it is sold by groups other than these, the decrease in price is small.

5.2.3 Strong Form

The strong form EMH states that security prices fully reflect all information. It represents an extreme hypothesis that most observers do not believe to be literally true. This hypothesis maintains that not only is the publicly available information useless to the investor or analyst, but that all information is useless. Information, whether public or inside, cannot be used consistently to earn superior investor returns in the strong form. This implies that security analysts and portfolio managers, who have access to more information than ordinary investors, are not able to use it to earn more profits.

Empirical evidence Many of the tests of the strong form of EMH deal with mutual fund performances. Financial analysts have studied the risk-adjusted rates of return from hundreds of mutual funds and found that the professionally managed funds are not able to outperform the naive -buy-hold strategists. Jensen studied 115 funds over a decade. He concluded that even though the analysts have a wide range of contacts and associations in both the business and financial committees, they are unable to predict returns accurately enough to recover the research and transaction costs. He holds this as a striking piece of evidence for the strong form of the efficient market hypothesis.

5.2.4 Limitations

According to the theory, successive price changes or changes in return are independent, and these successive price changes have a random distribution. The random walk model argues that stock prices fully reflect all publicly available information and further that the stock prices instantaneously adjust to the available new information. The theory mainly deals with successive changes rather than the price or return levels.

According to the theory, the market may have imperfections like transaction costs and delays in disseminating relevant information to all market investors, but these sources of inefficiency may not result in excess returns above the normal or equilibrium returns. The equilibrium return is the return earned by the naive buy-and-hold strategy.

Investors should note that the random theory says nothing about the relative price changes that are occurring across the securities. Some securities may outperform others. Again, it does not say anything about the decomposing of price into market, industry or firm factors. All these factors consider relative prices but not with the absolute price changes. The random walk hypothesis deals with the absolute price changes and not with the relative prices.

The prices may change at random, but this does not imply that there would not be any upward or downward trend in the price. The random walk hypothesis is entirely consistent with the upward and downward movements of stock prices.

Market Inefficiencies

Many studies have proven the prevalence of market efficiency. At the same time, several studies also contradict the concept of market efficiency.

Overreactions of the market Recent studies have shown that the market overreacts to corporate news. If the corporation announces a reduction in the earnings or closure of a unit, the market overreacts. The price of the stock may decline. After the initial decline, it may take several weeks for the stock price to reach the normal level. During this period, the investor may buy the stock when the price is low and sell it once the normal level is reached. This strategy helps him to earn an abnormal return, which goes against the efficient market hypothesis.

For example, take the studies conducted Joy, Litzenberger and McEnally (JLM) over the period 1963–68, which gave different results. The authors examined the quarterly earnings of the stock prices. The earnings of one quarter were compared with the earnings of the same quarter of the previous year. If the current year's earnings were 40 per cent or higher than the earnings for the same quarter in the previous year, the earnings were classified as better than anticipated. If the current quarter's earnings were below 40 per cent of the previous year's earnings, they were classified as worse than expected.

Then the abnormal returns were calculated from 13 weeks before the announcement of earnings to 26 weeks after the announcement. The stocks whose earnings were significantly greater than anticipated gave positive abnormal returns. The stocks whose earnings were below the anticipated earnings generated negative abnormal returns.

The authors' main claim is that after the announcement of the earnings, stocks that reported earnings substantially higher than that of the previous year continued to generate positive abnormal returns. According to their study, investors could have earned positive abnormal returns of around 6.5 per cent over the next 26 weeks simply by buying stocks that reported earnings 40 per cent above the last quarterly earnings. Meanwhile, for those stocks with earnings substantially below that of the previous year, the cumulative average abnormal return remained relatively stable. This shows evidence against the semi-strong market hypothesis because it states that when the information is made public, the analyst could not earn abnormal profits. A study by C P Jones and R S Rendleman for the period 1971–80 also gave results similar to those of JLM.

Reversal to mean return Some studies have found that stock returns have a tendency to return to their average level. Stocks that currently yield low returns tend to yield high returns in future. Likewise, the stocks that perform well at present may not yield high returns in future. The returns may go back to the average level.

This gives an opportunity to predict the future price, which is contrary to the random walk theory.

Delayed absorption of new information Usually, stock prices react quickly to information. Research has proved that stock prices tend to increase continuously for some time after the announcement of good profits. Likewise, they continue to decline for some time after the reporting of low profits.

Low P/E effect Many studies have provided evidence that stocks with low price-earnings ratios (P/Es) yield higher returns than stocks with higher P/Es. This is known as the low PE effect. A study by Basu in 1977 looked at risk adjusted return and even after the adjustment there was excess return in the low price-earnings stocks. If historical information of P/E ratios can help investor to obtain superior stock returns, it questions the validity of the semi-strong form of market hypothesis. Basu stated that low P/E portfolio experienced superior returns relative to the market, and high P/E portfolio performed in an inferior manner relative to the overall market. Since his result directly contradicts the semi-strong form of efficient market hypothesis, it is noteworthy.

Small firm effect The theory of the small firm effect maintains that investing in small firms (those with low capitalization) provides superior risk adjusted returns. Banz found the size of the firm to be highly correlated with returns. Banz examined historical monthly returns of NYSE common stocks for the period 1931–75. He formed portfolios consisting of the 10 smallest firms and the 10 largest firms and computed the average return for these portfolios. The small firm portfolio outperformed the large firm portfolio.

Several other studies have confirmed the existence of a small firm effect. The size effect raised doubts about the risk associated with small firms. The risks associated with them are underestimated, and they do not trade as frequently as the large firms. The correct measurement of risk and return of small portfolios tends to eliminate at least 50 per cent of the small firm effect.

The weekend effect French examined the returns generated by the S&P index for each day of the week. Stock prices tend to rise all week, reaching a peak on Fridays. Usually, stocks trade on Monday at low prices, before they begin the week's price increase. Buying on Monday and selling on Friday from 1953 to 1977 would have generated average annual return of 13.4 per cent while a simple buy-and-hold would have yielded 5.5 per cent annual return. If the transaction costs are taken into account, the naive buy-and-hold strategy would have provided higher returns. The weekend effect knowledge is of value to investors. Purchases planned on Thursday or Friday can be delayed until Monday, while sales planned for Monday can be delayed until the end of the week. The weekend effect is a small but significant deviation from perfectly random price movements and violates the weekly efficient market hypothesis.

B Venkatesh of the BL Research Bureau stated that the Bombay Stock Exchange reveals a clear pattern. Usually, Monday sees trading blues, and Friday, frenzied activity. The Friday rush is related to speculators covering their open position. If the short sellers fail to cover their position within this period, their open positions results in the auction where prices are dear.

Efficient Markets and the Investor

Every theory is aimed at the benefit of the investor. A technical analysis helps the investor to formulate his entry and exit strategy. The efficient market hypothesis also has its message. But all the investment tips, investment advice or technical analysis cannot guarantee an investor will earn abnormal profits. The publicly available information too cannot help the investor to outperform the market. All this leads to the conclusion that predictions about the market are inadequate. The constant buying and selling of stocks will result in payment of commission to brokers. Instead, the investor should pursue a buy-and-hold strategy with a careful selection of stocks. Yet, the anomalies of an efficient market hypothesis suggest that a prudent investor can earn higher with predictions than by following a buy-and-hold strategy.

Behavioural Finance

The anomalies of the efficient market hypothesis led to the evolution of behavioural finance. According to Andre Sheifer, it borrows concepts from the social sciences such anthropology sociology, and psychology to explain the behaviours of security price.

Psychologists opine that investors are unhappy when loss occurs and happy with the profit. This leads to the conclusion that investors are risk averse. The arbitrageurs who make smart money reduce the profit opportunities. They borrow stocks from the brokers and sell them. They hope to obtain the sold stock at a low price, and give it back to the brokers. The drawback of this process is that stock prices may go up. When stock prices rise, investors incur a loss. From the point of view of psychologists, the investors are risk averse. Hence, this psychology limits the short selling activity in the stock market. This leads to an overvaluation of stocks.

Psychologists also observe that investors are overconfident in their own judgments. They act according to their beliefs. While one investor may feel that 'X' stock price is going to fall and it is better to get rid of it, another may judge it to be a good investment at the prevailing price. Hence, the trading volume in the stock market may be high. The efficient market hypothesis has no explanation for it.

Behaviour finance explains that the stock market bubbles are due to the social contagion effect and overconfidence of the investor. The media and word-of-mouth enthusiasm create stock market bubbles. Once the stock prices go up,

the investor feels it will rally in future. This leads to a positive feedback loop, and stock prices continue to move up. This creates a speculative bubble without the backing of strong fundamentals. This results in a stock market crash.

Check Your Progress

1. Name the two kinds of market efficiencies.
2. Who are liquidity traders?

5.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The two kinds of market efficiencies are the following:
   - Operational efficiency
   - Informational efficiency

2. These traders' investments and resale of shares depend upon their personal fortune. Liquidity traders may sell their shares to pay their bills. They do not analyse before they invest.

5.4 SUMMARY

- The expectations of investors regarding future cash flows are translated into or reflected in share prices. Market efficiency is the accuracy and speed with which the market translates the expectation into prices.

- Information traders base their buy or sell strategy on proper analyses. They estimate the intrinsic value of shares. They enter the market on the basis of the deviation of the market value of shares from the intrinsic value.

- In 1900, French mathematician Louis Bachelier wrote an article suggesting that security price fluctuations were random. In 1953, British statistician Maurice Kendall reported that the stock price series is a wandering one. Each change is independent of the previous one.

- A runs test finds out whether the series of price movements occurs by chance. A run is an uninterrupted sequence of the same observation.

- The serial correlation technique tests the independence of successive price changes. Serial correlation or autocorrelation measures the correlation co-efficient in a series of numbers with the lagging values of the same series.

- The semi-strong form of the EMH states that the security price adjusts quickly to all publicly available information. In semi-strong efficient markets, security prices fully reflect all publicly available information.

- The strong form EMH states that security prices fully reflect all information. It represents an extreme hypothesis that most observers do not believe to be literally true.

- The random walk model argues that stock prices fully reflect all publicly available information and further that the stock prices instantaneously adjust to the available new information.

5.5 KEY WORDS

- **Market efficiency:** It is the accuracy and speed with which the market translates the expectations into prices.

- **Arbitrage:** It is the simultaneous purchase and sale of an asset to profit from an imbalance in the price.

5.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. Write a short note on the Random-walk theory.

2. What is the significance of the study of Fama, Fisher, Jensen and Roll conducted in the field of EMH?

3. Why is it necessary for the investor to pursue a buy-and-hold strategy with careful selection of stocks?

Long Answer Questions

1. 'The strong form EMH states that security prices fully reflect all information.' Explain the statement.

2. Discuss the evolution of behavioural finance.

5.7 FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.

BLOCK - II

THEORY OF PRODUCTION

UNIT 6 THEORY OF PRODUCTION

Structure

6.0   INTRODUCTION

Be it profit maximization or any other objective of business firms, achieving optimum efficiency in production and minimising cost for a given production is one of the prime concerns of the business managers. In fact, the very survival of a firm in a competitive market depends on their ability to produce at a competitive cost. Therefore, managers of business firms endeavour to minimize the production cost or, what is the same thing, maximize output from a given quantity of inputs. In their effort to minimize the cost of production, the fundamental questions which managers are faced with are:

   (*i*) How can the use of production inputs be optimized or cost minimized?

  (*ii*) How does output behave when quantity of inputs is increased?

 (*iii*) How does technology matter in reducing the cost of production?

 (*iv*) How can the least-cost combination of inputs be achieved?

  (*v*) Given the technology, what happens to the rate of return when production scale is increased or more plants are added to the firm?

The theory of production provides a theoretical answer to these questions through abstract models built under hypothetical conditions. The production theory may therefore not provide solution to the real life problems. But it does provide tools

and techniques to analyse the production conditions and to find solution to the practical business problems.

This unit is devoted to the discussion of the theory of production. Production theory deals with quantitative relationships between output and inputs given the production technology. Let us first discuss some basic concepts used in production analysis.

6.1 OBJECTIVES

After going through this unit, you will be able to:

- Describe the theory of production
- Explain the law of variable proportions
- Discuss the law of returns to scale

6.2 THEORY OF PRODUCTION: AN INTRODUCTION

Let us begin our discussion on the theory of production by first understanding the basic concepts of productions. In economics, the term 'production' means a process by which resources (men, material, time, etc.) are transformed into a *different* and more useful commodity or service. In general, production means transforming inputs (labour, machines, raw materials, time, etc.) into an output with value added. This concept of production is however limited to only 'manufacturing'.

In economic sense, *production process* may take a variety of forms other than manufacturing. For example, transporting a commodity from one place to another where it can be consumed or used in the process of production is *production*. For example, a sand dealer collects and transfers the sand from the river bank to the construction site; a coal miner does virtually nothing more than transporting coal from coal mines to the market place. Similarly, a fisherman only transports fish to the market place. Their activities too are 'production'. Transporting men and materials from one place to another is a productive activity: it produces *service*. Storing a commodity for future sale or consumption is also 'production'. Wholesaling, retailing, packaging, assembling are all productive activities. These activities are just as good examples of production as manufacturing. Cultivation or farming is the earliest form of productive activity.

Besides, production process does not necessarily involve physical conversion of raw materials into tangible goods. Some kinds of production involve an intangible input to produce an intangible output. For example, in the production of legal, medical, social and consultancy services both input and output are intangible; lawyers, doctors, social workers, consultants, hair-dressers, musicians, orchestra players are all engaged in producing intangible goods.

Input and Output

An *input* is a good or service that is used into the process of production. In the words of Baumol, "An input is simply anything which the firm buys for use in its production or other proceses." An *output* is any good or service that comes out of production process.

The term 'inputs' needs some more explanations. Production process requires a wide variety of inputs, depending on the nature of product. But, economists have classified inputs as (*i*) labour, (*ii*) capital, (*iii*) land, (*iv*) raw materials, (*v*) time and (*vi*) technology. While capital and land are treated as 'stock' variable, all other variables are 'flow' variables, since they are measured per unit of time.

Fixed and Variable Inputs

For the purpose of theoretical analysis, inputs are classified under two categories:

   (*i*) *fixed inputs* or *fixed factors*, and

  (*ii*) *variable inputs* or *variable factors*.

Fixed and variable inputs are defined in economic sense and in technical sense. In **economic sense**, a fixed input is one whose supply is inelastic in the short run. Therefore, all of its users together cannot buy more of it in the short-run. In **technical sense,** a fixed factor is one that remains fixed (or constant) for a certain level of output.

A *variable input* is defined as one whose supply in the short-run in elastic, e.g., labour and raw material, etc. All the users of such factors can employ a larger quantity in the short-run. Technically, a variable input is one that changes with the change in output. In the long run, all inputs are treated to be variable.

Short-Run and Long-Run

The reference to *time* period involved in production process is another important concept used in production analysis. The two reference periods are *short run* and *long run*. The *short run* refers to a period of time in which the supply of certain inputs (e.g., plant, building, machinery, etc.) is fixed or is inelastic. In short run, therefore, production of a commodity can be increased by increasing the use of only variable inputs like labour and raw materials.

It is important to note that 'short run' and 'long run' are economists' jargon. They do not refer to any fixed time period. While in some industries short run may be a matter of few weeks or few months, in some others (e.g., automobile, ship, plane and power industries), it may mean two, three or more years.

The *long run* refers to a period of time in which the supply of all the inputs is elastic, but the period is not long enough to permit a change in technology. That is, in the long run, all the inputs are variable. Therefore, in the long-run, production of a commodity can be increased by employing more of both variable and fixed inputs.

The economists use another term, i.e., *very long run* which refers to a period in which the technology of production is supposed to change. In the very long run, the production function also changes. The technological advances result in a larger output from a given quantity of inputs.

6.3 THE LAW OF VARIABLE PROPORTIONS

Before we proceed to discuss the laws of production, let us have a look at the meaning and the kinds of laws of production with reference to this frame.

The laws of production state the nature of relationship between output and input. The traditional theory of production studies the marginal input-output relationships under (*i*) short run, and (*ii*) long run conditions. In the short run, input-output relations are studied with one variable input, other inputs held constant. This changes the proportion of inputs. Therefore, the laws of production under these conditions are called 'The Laws of Variable Proportions' or the 'Laws of Returns to a Variable Input'. In the long-run, input-output relations are studied assuming all the input to be variable. The long-run input-output relations are studied under the 'Laws of Returns to Scale'. In subsequent sections, we explain the 'laws of return to a variables input' and the laws of 'returns to scale'.

6.3.1 Short-Run Laws of Production

Production with one Variable Input

Some factors of production have elastic supply even during the short period. Such factors are called **variable factors.** In the short-run, therefore, the firms can employ a large quantity of the variable factor. In other words, firms can employ in the short run, varying quantities of variable inputs against a given quantity of fixed factors. This kind of change in input combination leads to *variation in factor proportions.* The laws which bring out the relationship between varying factor proportions and output are therefore known also as the *Laws of Returns to a Variable proportions.* This law is more popularly known as the *Law of Diminishing Returns.* In this section, we explain the laws of returns to variable input.

The Law of Returns to a Variable Input : The Law of Diminishing Returns

The law of diminishing returns states that *when more and more units of a variable input are applied to a given quantity of fixed inputs, the total output may initially increase at an increasing rate and then at a constant rate but it will eventually increase at diminishing rates.* In other words, when a firm using two inputs—labour and capital—increases the number of labour, capital remaining constant, the marginal productivity of labour may initially increase, but it does decrease eventually. This is called the law of diminishing returns to the variable input.

Assumptions. The law of diminishing returns is based on the following assumptions: (*i*) the state of technology is given, (*ii*) labour is homogeneous, and (*iii*) input prices—wages and interest—are given.

To illustrate the law of diminishing returns, let us assume (*i*) that the coal-mining firm (in our earlier example) has a set of mining machinery as its capital (*K*), fixed in the short run, and (*ii*) that it can employ more of mine-workers to increase its coal production. Thus, the short run production function for the firm will take the following form.

$$Q_c = f(L)$$

Let us assume that the labour-output relationship in coal production based on actual data is given by a hypothetical production function of the following form.

$$Q_c = -L^3 + 15L^2 + 10L \qquad \ldots(6.1)$$

Given the production function (7.4), we may substitute different numerical values for *L* in the function and work out a series of Q_c, i.e., the quantity of coal (say, thousand tonnes) that can be produced with different number of workers. For example, if *L* = 5, then by substitution,

$$Q_c = -5^3 + 15 \times 5^2 + 10 \times 5$$

$$= -125 + 375 + 50$$

$$= 300 \text{ (thousand tonnes)}$$

A tabular array of output levels associated with different number of workers from 1 to 12, in our hypothetical coal-production example, is given in Table 6.1 (Cols. 1 and 2).

What we need now is to work out *marginal productivity of labour* (MP_L) to find the trend in the contribution of the marginal labour and *average productivity of labour* (AP_L) to find the average contribution of labour.

Tables 6.1 *Three Stages of Production*

No. of workers (N)	Total product (TP_L) (000 tonnes)	Marginal Product* (MP_L)	Average Product (AP_L)	Stages of Production
(1)	(2)	(3)	(4)	(5)
1	24	24	24	I
2	72	48	36	Increasing
3	138	66	46	returns
4	216	78	54	
5	300	84	60	
6	384	84	64	
7	462	78	66	II
8	528	66	66	Diminishing
9	576	48	64	returns
10	600	24	60	
11	594	– 6	54	III
12	552	– 42	46	Negative returns

$*MP_L = TP_n - TP_{n-1}$. MP_L calculated by differential method will be different from that given in Col. 3.

Marginal Productivity of Labour (MP_L) can be obtained by differentiating the production function (6.1). Thus,

$$MP_L = \frac{\partial Q}{\partial L} = -3L^2 + 30L + 10 \qquad \ldots(6.2)$$

By substituting numerical value for labour (L) in Eq. (6.2), MP_L can be obtained at different levels of labour employment. However, this method can be used only where labour is perfectly divisible and δL 0. Since, in our example, each unit of $L = 1$, calculus method cannot be used.

Alternatively, where labour can be increased at least by one unit, i.e., $\Delta L = 1$, MP_L can be obtained as

$$MP_L = TP_L - TP_{L-1}$$

The MP_L worked out by this method is presented in col. 3 of Table 6.1.

Average Productivity of labour (AP_L) can be obtained by dividing the production function by L. Thus,

$$AP_L = \frac{-L^3 + 15L^2 + 10L}{L} = -L^2 + 15L + 10 \qquad \ldots(6.3)$$

Now AP_L can be obtained by substituting the numerical value for L in Eq. (6.3). AP_L obtained by this method is given in col. 4 of Table 7.1.

The information contained in Table 6.1 is presented graphically in panels (*a*) and (*b*) of Fig. 6.1. Panel (*a*) of Fig. 6.1 presents the total product curve (TP_L) and panel (*b*) presents marginal product (MP_L) and average product (AP_L) curves. The TP_L schedule demonstrates the law of diminishing returns. As the curve TP_L shows, the total output increases at an increasing rate till the employment of the 5th worker, as indicated by the increasing slope of the TP_L curve. (See also col. 3 of the table.) Beyond the 6th worker, TP_L increases (until the 10th worker) but the rate of increase in TP_L (i.e., marginal addition to TP_L) begins to fall and turns negative 11th worker onwards. This shows the operation of the law of diminishing returns, i.e., as units of labour increase, capital remaining constant, the total output increases initially at an increasing rate but finally at a decreasing rate.

Three Stages in Production

Table 6.1 and Fig. 6.1 present the three usual stages in the application of the laws of diminishing returns. In Stage I, TP_L increases at increasing rate. This is indicated by the rising MP_L till the employment of the 5th worker. Given the production function (Eq. 6.1), the 6th worker produces as much as the 5th worker. The output from the 5th and the 6th workers represents an intermediate stage of constant returns to the variable factor, labour.

In Stage II, TP_L continues to increase but at diminishing rates, *i.e.*, MP_L begins to decline. This stage in production shows the law of diminishing returns to the variable factor. Total output reaches its maximum level at the employment of

the 10th worker. Beyond this level of labour employment, TP_L begins to decline. This marks the beginning of Stage III in production.

To conclude, given the employment of fixed factor (capital), when more and more workers are employed, the return from the additional worker may initially increase but will eventually decrease.

Factors Behind the Laws of Returns

As shown in Fig. 6.1, the marginal productivity of workers (MP_L) increases in Stage I and it decreases in Stage II. Stage I shows the Law of Increasing Returns and Stage II shows the Law of Diminishing Returns.
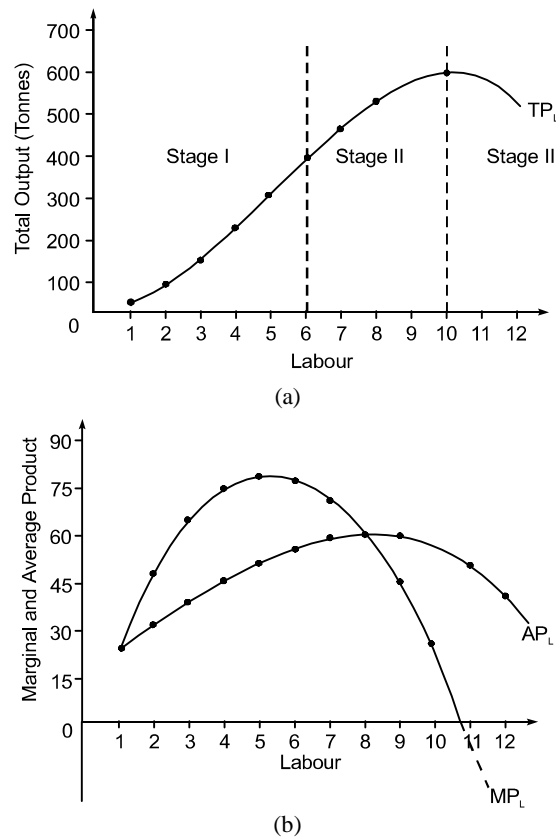
Fig. 6.1 *Total, Average and Marginal Products*

The reasons which underly the application of the laws of returns in Stages I and II may be described as follows. One of the important factors causing increasing returns to a variable factor is the **indivisibility of fixed factor** (capital). It results in under-utilisation of capital if labour is less than its optimum number. Let us suppose that optimum capital-labour combination is 1:6. If capital is indivisible and less than 6 workers are employed, then capital would remain underutilised. When more and more workers are added, utilization of capital increases and also the productivity of additional worker. Another reason for increase in labour productivity is that employment of additional workers leads to advantages of **division of labour**, until optimum capital-labour combination is reached.

Once the optimum capital-labour ratio is reached, employment of additional workers amounts to substitution of capital with labour. But, technically, one factor can substitute another only upto a limited extent. Therefore, with increase in labour, capital per unit of labour decreases. This causes decrease in the productivity of the marginal labour. That is, employment of more and more labour against a given capital, causes decrease in MP_L.

Empirical Validity of the Law of Diminishing Returns. The law of diminishing returns is *an empirical law,* frequently observed in various production activities. This law, however, may not apply universally to all kinds of productive activities since the law is not as true as the law of gravitation. In some productive activities, it may operate quickly, in some its operation may be delayed; and in some others, it may not appear at all. This law has been found to operate in agricultural production more regularly than in industrial production. The reason is, in agriculture, natural factors play a predominant role whereas man-made factors play the major role in industrial production. Despite these variations and limitations of the law, if increasing units of an input are applied to the fixed factors, the marginal returns to the variable input decrease eventually.

6.3.2 The Law of Diminishing Returns and Business Decisions

We have discussed above the law of diminishing returns in a theoretical framework. Let us now look at the applicability of this law to business decision-making.

The law of diminishing returns as presented graphically has a relevance to the business decisions. The graph can help in identifying the rational and irrational stages of operations. It can also provide answer to such questions as (*i*) how much labour to employ to maximise the output; and (*ii*) what number of workers to apply to a given fixed input so that per unit cost in minimized when output is maximumized. Fig 6.1 exhibits the three stages of production. In Stage III, has a very high labour-capital ratio. As a result, employment of additional workers proves not only unproductive but also causes a decline in the *TP*. Similarly, in Stage I, capital is presumably underutilized. So a firm operating in Stage I is required to increase labour, and a firm operating in Stage III is required to reduce labour, with a view to maximising its total production. From the firm's point of view, setting an output target in Stages I and III is irrational. The only meaningful and rational stage from the firm's point of view is Stage II in which the firm can find answer to the questions 'how many workers to employ'.

Figure 6.1 shows also that the firm should employ a minimum of 6 workers and a maximum of 10 workers even if labour is available free of cost. This means that the firm has a limited choice ranging from 6 to 10 workers. How many workers to employ against the fixed capital and how much to produce can be answered, only when the price of labour, i.e., wage rate, and that of the product are known. This question is answered below.

Optimum Employment of Labour

It may be recalled from Fig. 6.1 that an output maximizing coal-mining firm would like to employ 10 workers—since at this level of employment, the output is maximum and $MP_L = 0$. The firm can, however, employ 10 workers only if workers are available free of cost. But labour is not available free of cost—the firm is required to pay wages to the workers. Therefore, the question arises 'how many workers should the firm employ—10 or less or more than 10—to maximise its profit. A simple answer to this question is that the number of workers to be employed depends on the output that maximizes firm's profit, given the product price and the wage rate. This point can be proved as follows.

As a rule, total profit is maximum at the level of output at which

$$MC = MR$$

In our example here, since labour is the only variable input, marginal cost (MC) equals marginal wages (MW), i.e., $MC = MW$.

As regards MR, in case of factor employment, the concept of **Marginal Revenue Productivity** is used. The marginal revenue productivity of labour is the money value of marginal product of labour (MP_L). In specific terms, marginal revenue productivity of labour (MRP_L) equals marginal physical productivity (MP_L) of labour multiplied by the price (P) of the product, i.e.,

$$MRP_L = MP_L \times P$$

For example, suppose that the price (P) of coal is given at ₹ 10 per quintal. Now, MRP_L can be known by multiplying its MP_L (as given in Table 6.1) by ₹ 10. For example, MRP_L of the 3rd worker (see Table 6.1) equals $66 \times 10 = ₹ 660$ and of the 4th worker, $78 \times 10 = 780$. Likewise, if whole column (MP_L) is multiplied by ₹ 10, it gives us a table showing the number of workers and the marginal revenue productivity of workers. Let us suppose that wage rate (per time unit) is given at ₹ 660. Given the wage rate, the profit maximising firm will employ only 8 workers because at this employment, MRP_L = wage rate = MRP_L of 8th worker $= 66 \times 10 = ₹ 660$. If the firm employs the 9th worker, $MRP_L = 48 \times 10 = ₹ 480$ $< ₹ 660$. Clearly, the firm loses ₹ 180 on the 9th worker. And, if the firm employs less than 8 workers, it will not maximize its profit.

To generalize, if relevant series of MRP_L is graphed, it gives the MRP_L curve as shown in Fig. 6.2. Similarly, the MRP_L curve for any input may be drawn and compared with MC (or MW) curve. Labour being the only variable input, in our example, let us suppose that wage rate in the labour market is given at OW (Fig. 6.2). When wage rate remains constant, average wage (AW) and marginal wage (MW) are equal, i.e., $AW = MW$, for the whole range of employment in the short run. When $AW = MW$, a large number of labour can be employed at wage rate OW as shown by a straight horizontal line marked $AW = MW$.

With the introduction of MRP curve and $AW = MW$ line (Fig. 6.2), a profit maximising firm can easily find the maximum number of workers which can be

optimally employed against a fixed quantity of capital. Once the maximum number of workers is determined, the optimum quantity of the product is automatically determined.
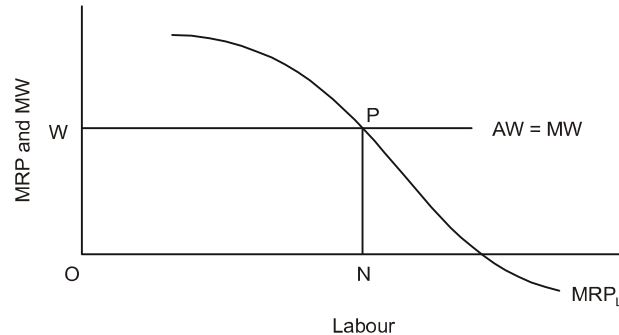
Fig. 6.2 *Determination of Labour Employment in the Short-Run*

The marginality principle of profit maximization tells that profit is maximum where $MR = MC$. This is a necessary condition of profit maximisation. Fig. 6.2 shows that $MRP_L = MW (= MC)$ are equal at point P, the point of intersection between MRP_L and $AW = MW$. The number of workers corresponding to this point is ON. A profit maximising firm should therefore employ only ON workers. Given the number of workers, the total output can be known by multiplying ON with average labour productivity (AP).

6.3.3 Long-Term Laws of Production–I: Tools of Analysis

In the preceding section, we have discussed the technological relationship between inputs and output assuming labour to be the only variable input, capital remaining constant. This is a short-run phenomenon. In this section, we will discuss the relationship between inputs and output under the condition that both the inputs, capital and labour, are variable factors. This is a long-run phenomenon. In the long-run, supply of both the inputs is supposed to be elastic and firms can hire larger quantities of both labour and capital. With larger and larger employment of capital and labour, the scale of production increases. The technological relationship between changing scale of inputs and output is explained under the **laws of returns to scales**. The laws of returns to scale can be explained through the *production function* and *isoquant curve* technique. The most common and simple tool of analysis is isoquant curve technique. We have, therefore, first introduced and elaborated on this tool of analysis—the isoquant curve. The laws of return to scale have then been explained through isoquant curve technique and through production function in the subsequent sections.

6.3.4 Production Isoquant: A Tool of Production Analysis

The Isoquant Curve *An isoquant curve is locus of points representing various combinations of two inputs—capital and labour—yielding the same output.* An 'isoquant curve' is analogous to an 'indifference curve', with two

points of distinction: (*a*) an indifference curve is made of two consumer goods while an isoquant curve is constructed of two producer goods (labour and capital), and (*b*) while an indifference curve measures 'utility', an isoquant measures 'output'.

Isoquant curves are drawn on the basis of the following assumptions:

- (*i*) there are only two inputs, viz., labour (*L*) and capital (*K*), to produce a commodity *X*;

- (*ii*) the two inputs—*L* and *K*—can be substituted one for another but at a diminishing rate; and

- (*iii*) the technology of production is given and labour and capital can be substituted only to a certain extent.

Given these assumptions, it is always possible to produce a given quantity of commodity *X* with various combinations of capital and labour. The input combinations are so formed that the substitution of one factor for the other leaves the output unaffected. This technological fact is presented through an Isoquant Curve ($IQ_1 = 100$) in Fig. 6.3. The curve IQ_1 all along its length represents a fixed quantity, 100 units of product *X*. This quantity of output can be produced with a number of labour-capital combinations. For example, points *A*, *B*, *C*, and *D* on the isoquant IQ_1 show four different combinations of inputs, *K* and *L*, as given in Table 6.2, all yielding the same output—100 units. Note that movement from *A* to *D* indicates decreasing quantity of *K* and increasing number of *L*. This implies substitution of labour for capital such that all the input combinations yield the same quantity of commodity *X*, i.e., $IQ_1 = 100$.
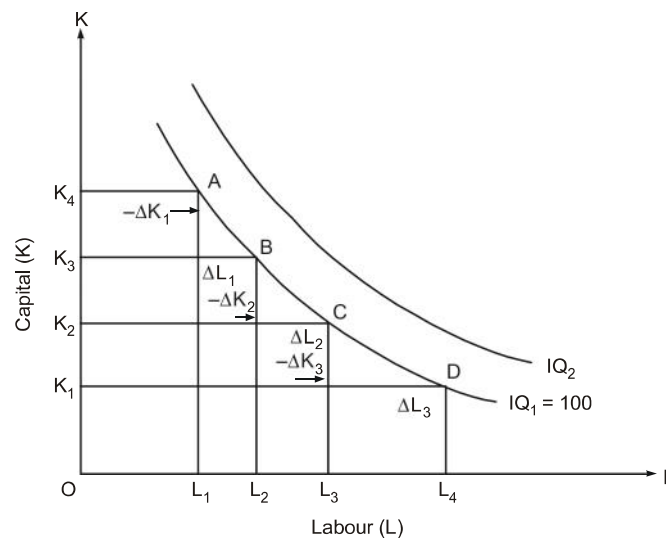
Fig. 6.3 *Isoquant Curves*

Table 6.2 *Capital-Labour Combinations and Output*

Points	Input Combinations			Output
	K	+	L	
A	OK_4	+	OL_1	= 100
B	OK_3	+	OL_2	= 100
C	OK_2	+	OL_3	= 100
D	OK_1	+	OL_4	= 100

Properties of Isoquants

Isoquants have the same *properties as indifference curves.* Properties of isoquants are explained below in terms of inputs and outputs.

(a) Isoquants have a negative slope. An isoquant has a negative slope in the *economic region* or in the range of labour-capital substitutability. The economic region is the region on the isoquant plane in which substitution between inputs is technically feasible. It is also known as the product maximizing region. *The negative slope of the isoquant implies substitutability between the inputs*. It means that if one of the inputs is reduced, the other input has to be so increased that the total output remains unaffected. For example, movement from point A to B on IQ_1 (Fig. 6.3) means that K_4K_3 units of capital are substituted with L_1L_2 units of labour to maintain the same level of output.

(b) Isoquants are convex to the origin. Convexity of isoquants implies not only the substitution between the inputs but also diminishing *marginal rate of technical substitution* (*MRTS*) between the inputs. The *MRTS* is defined as

$$MRTS = \frac{-\Delta K}{\Delta L} = \text{slope of the isoquant}$$

In plain words, *MRTS* is the rate at which one input can substitute the other, output remaining the same. Movement along the isoquant gives the rate at which a marginal unit of *K* can be substituted with labour without affecting the total output or otherway round. This rate is indicated by the slope of the isoquant. The *MRTS* decreases for two reasons: (*i*) no factor is a perfect substitute for another, and (*ii*) inputs are subject to diminishing marginal return. Therefore, more and more units of an input are needed to replace each successive unit of the other input. For example, suppose units of ΔK (minus sign ignored) in Fig. 6.3 are equal, i.e.,

$$\Delta K_1 = \Delta K_2 = \Delta K_3$$

the subsequent units of *L* substituting ΔK go on increasing, i.e.,

$$\Delta L_1 < \Delta L_2 < \Delta L_3$$

As a result, *MRTS* goes on decreasing, i.e.,

$$\frac{\Delta K_1}{\Delta L_1} > \frac{\Delta K_2}{\Delta L_2} > \frac{\Delta K_3}{\Delta L_3}$$

(c) Isoquants do not intersect nor are tangent to each other. The intersection or tangency between any two isoquants implies that a given quantity of a commodity can be produced with a smaller as well as a larger input-combination. This is untenable so long as marginal productivity of inputs is greater than zero. In Fig. 6.4, two isoquants intersect each other at point *M*. Consider two other points—point *J* on isoquant marked 100 and point *K* on isoquant marked 200. One can easily infer that a quantity that can be produced with the combination of *K* and *L* at point *M* can be produced also with factor combinations at points *J* and *K*. On the isoquant 100, factor combinations at points *M* and *J* are equal in terms of their output. On the isoquant marked 200, factor combinations at *M* and *K* are equal in terms of their output. Since point *M* is common to both the isoquants, it follows that input combinations at *J* and *K* are equal in terms of output. This implies that in terms of productivity,
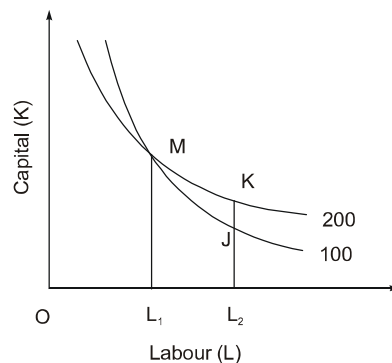
Fig. 6.4 *Intersecting Isoquants*

$$OL_2(L) + JL_2(K) = OL_2(L) + KL_2(K)$$

Since OL_2 is common to both the sides, it means that in terms of output,

$$JL_2 = KL_2$$

But, as can be seen in Fig. 7.4,

$$JL_2 < KL_2$$

But the intersection of the isoquants means that JL_2 and KL_2 are equal in terms of productivity. This cannot be possible because output generated by $KL_2(K)$ is greater than that by $JL_2(K)$. That is why isoquant will not intersect or be tangent to each other. If they do, it violates the law of production.

(d) Upper isoquants represent higher level of output. Between any two isoquants, the upper one represents a higher level of output than the lower one. The reason is, an upper isoquant implies a larger input combination throughout. A larger input combination produces a larger output. Therefore, upper isoquants indicate a higher level of output.

For instance, the upper isoquant, IQ_2 in Fig. 6.5 will always mean a higher level of output than IQ_1. For, any point at IQ_2 consists of more of either capital or labour or both. For example, consider point *a* on IQ_1 and compare it with any point at IQ_2. For example, the point *b* on IQ_2 indicates more of capital (*ab*), point

d more of labour (*ad*) and point *c* more of both. Therefore, IQ_2 represents a higher level of output (200 units) than IQ_1 indicating 100 units.
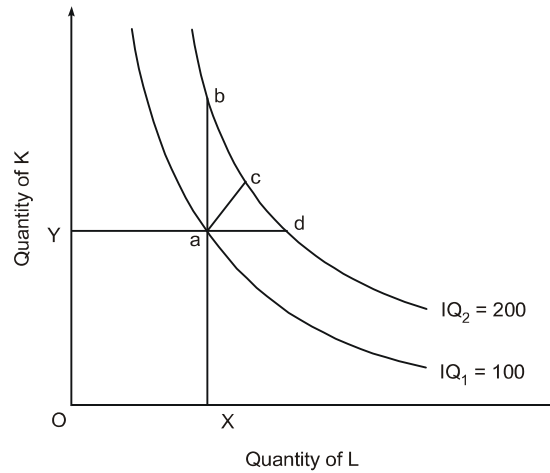
Fig. 6.5 *Comparison of Output at Two Isoquants*

Isoquant Map, Ridge Line and Economic Region

Isoquant map. One way to present a production function on a two-dimensional plane is to use its *isoquant map.* An *isoquant map* is a set of isoquants presented on a two-dimensional plane as shown by isoquants Q_1, Q_2, Q_3 and Q_4 in Fig. 6.6. Each isoquant shows various combinations of two inputs *L* and *K* that can be used to produce a given level of output. An upper isoquant is formed by a greater quantity of one or both the inputs than the input combination indicated by the lower isoquants. For example, isoquant Q_2 indicates a greater input-combination than that shown by isoquant Q_1 and so on.

Since upper isoquants indicate a larger input-combination than the lower ones, each successive upper isoquant indicates a higher level of output than the lower ones. For example, if isoquant Q_1 represents an output equal to 100 units, isoquant Q_2 represents an output greater than 100 units. As one of the properties of isoquants, no two isoquants can intersect or be tangent to one another.

Economic region. The convexity of isoquant might imply that labour and capital can be substituted one for another to any extent. But, technically it is not possible. There is a limit to which one input can be substituted for another. We know that *MRTS* decreases along the convex isoquant. The limit to which the *MRTS* can decrease is zero. *A zero MRTS implies that there is a limit to which one input can substitute another*. It also determines the minimum quantity of an input which must be used to produce a given output. Beyond this point, an additional employment of one input will necessitate employing additional units of the other input. Such a point on an isoquant may be obtained by drawing a tangent to the isoquant and parallel to the vertical and horizontal axes, as shown by dashed lines in Fig. 6.6. By joining the resulting points *a, b, c* and *d*, we get a line called the *upper ridge line*, *Od*. Similarly, by joining the points *e, f, g* and *h*, we

get the *lower ridge line, Oh*. The ridge lines are locus of points on the isoquants where the marginal products (*MP*) of the inputs are equal to zero. The upper ridge line implies that *MP* of capital is zero along the line, *Od*. The *lower ridge line* implies that *MP* of labour is zero along the line, *Oh*.
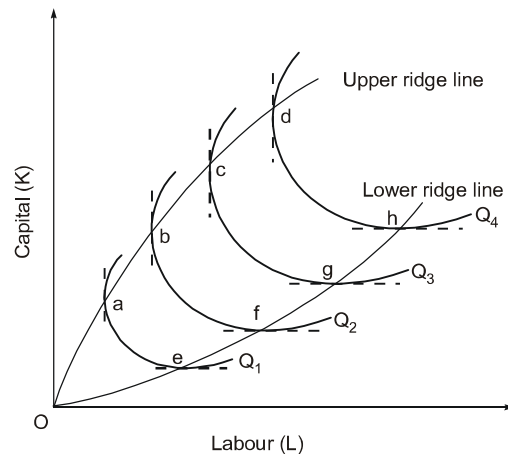
Fig. 6.6 *Isoquant Map*

The area between the two ridge lines, *Od* and *Oh*, is called '**Economic Region**' or 'technically efficient region' of production. Any production technique, i.e., capital-labour combination, within the economic region is technically efficient to produce a given output. And, any production technique outside this region is technically inefficient since it requires more of both inputs to produce the same quantity.

6.4 THE LAW OF RETURNS TO SCALE

Having introduced the isoquants—the basic tool of analysis—we now return to the laws of returns to scale. The laws of returns to scale state the behaviour of output in response to a proportional and simultaneous change in inputs. Increasing inputs proportionately and simultaneously is, in fact, an expansion of the scale of production.

When a firm expands its scale, i.e., it increases both the inputs proportionately, then there are three technical possibilities:

(*i*) total output may increase more than proportionately;

(*ii*) total output may increase proportionately; and

(*iii*) total output may increase less than proportionately.

Accordingly, there are three kinds of returns to scale:

(*i*) Increasing returns to scale;

(*ii*) Constant returns to scale, and

(*iii*) Diminishing returns to scale.

So far as the sequence of the laws of 'returns to scale' is concerned, the law of increasing returns to scale is followed by the law of constant and then by the law of diminishing returns to scale. This is the most common sequence of the laws of returns to scale.

1. Increasing Returns to Scale

When a certain proportionate increase in both the inputs, *K* and *L*, leads to a more than proportionate increase in output, it exhibits *increasing returns to scale*. For example, if both the inputs, *K* and *L*, are successively doubled and the corresponding output is more than doubled, the returns to scale is said to be increasing. The increasing returns to scale is illustrated in Fig. 6.7. The movement from point *a* to *b* on the line *OB* means doubling the inputs. It can be seen in Fig. 6.7 that the combination of inputs *L* and *K*, increases from $1K + 1L$ to $2K + 2L$. As a result of doubling the inputs, output is more than doubled: it increases from 10 to 25 units, *i.e.*, a 100 per cent increase in inputs results in 120 per cent increase in output. Similarly, the movement from point *b* to point *c* indicates a 50% increase in inputs as a result of which the output increases from 25 units to 50 units, i.e., by 200%. This kind of relationship between the inputs and output shows *increasing returns to scale*.
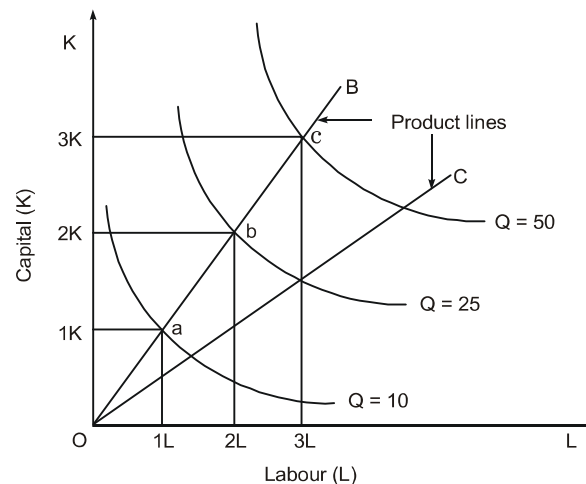
Fig. 6.7 *Increasing Returns to Scale*

Factors Leading to Increasing Returns to Scale   There are at least three plausible reasons for increasing returns to scale, called *economics of seale*.

(i) Indivisibility of Machinery and Managerial Manpower.   Certain inputs, particularly mechanical equipments and managerial manpower, used in the process of production are available in a given size. Such inputs cannot be divided into parts to suit small scale of production. For example, half a turbine cannot be used to produce electricity and one-third of a composite harvester and earth-movers cannot be used productively. Similarly, half of a production manager cannot be employed, if part-time employment is not acceptable to the manager. Because of indivisibility of machinery and managers, given the state of technology, they have

to be employed in a minimum quantity even if scale of production is much less than the capacity output. Therefore, when scale of production is expanded by increasing all the inputs, the productivity of indivisible factors increases exponentially because of technological advantage. This results in increasing returns to scale.

(ii) Higher degree of specialization. Another factor causing increasing returns to scale is higher degree of specialization of labour, manager and machinery, which becomes possible with increase in scale of production. The use of specialized labour suitable to job needs and composite machinery increases productivity per unit of inputs. Their cumulative effects contribute to the increasing returns to scale. Besides, employment of specialized managerial personnel, e.g., administrative manager, production managers, sales manager and personnel manager, contributes a great deal in increasing production.

(iii) Dimensional relations. Increasing returns to scale is also a matter of dimensional relations. For example, when the length and breadth of a room ($15' \times 10' = 150$ sq. ft.) are doubled, then the size of the room is more than doubled. It increases to $30' \times 20' = 600$ sq. ft. which is more than double the room size. Similarly, when diameter of a pipe is doubled, the flow of water is more than doubled. In accordance with this dimensional relationship, when the labour and capital are doubled, the output is more than doubled and so on.

2. Constant Returns to Scale

When the increase in output is proportional to the increase in inputs, it exhibits *constant returns to scale.* For example, if both the inputs, *K* and *L*, are doubled subsequently and output is also doubled, subseqently then the returns to scale are said to be constant. Constant returns to scale are illustrated in Fig. 6.8. The lines *OA* and *OB* are 'product lines' indicating two hypothetical techni-ques of production. The isoquants marked $Q = 10$, $Q = 20$ and $Q = 30$ indicate the three different levels of output. In the figure, the movement from points *a* to *b* indicates doubling both the inputs. When inputs are doubled, output is also doubled, i.e., the output increases from 10 to 20, i.e., a 50% increase in output.
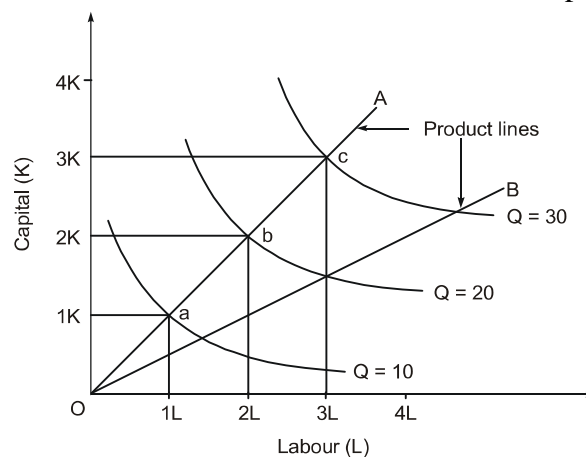
Fig. 6.8 *Constant Returns to Scale*

Similarly, movement from point *b* to *c* indicates a 50 per cent increase in labour as well as capital. This increase in inputs results in an increase of output from 20 to 30 units, i.e., a 50 per cent increase in output. In simple words, a 50 per cent increase in inputs leads a 50 per cent increase in output. This relationship between the proportionate change in inputs and proportional change in output may be summed up as follows:

$$1K + 1L = 10$$
$$2K + 2L = 20$$
$$3K + 3L = 30$$

This relationship between inputs and output exhibits *constant returns* to *scale*.

The constant returns to scale are attributed to the limits of the economies of scale. With expansion in the scale of production, economies arise from such factors as indivisibility of fixed factors, greater possibility of specialization of capital and labour, use of labour-saving techniques of production, etc. But there is a limit to the economies of scale. When economies of scale reach their limits and diseconomies are yet to begin, returns to scale become constant. The constant returns to scale also take place where factors of production are perfectly divisible and where technology is such that capital-labour ratio is fixed. When the factors of production are perfectly divisible, the production function is homogeneous of degree 1 showing constant returns to scale.

3. Decreasing Returns to Scale

The firms are faced with *decreasing returns to scale* when a certain proportionate increase in inputs, *K* and *L*, leads to a less than proportional increase in output. For example, when inputs are doubled and output is less than doubled, then decreasing returns to scale is in operation. The decreasing returns to scale is illustrated in Fig. 6.9. As the figure shows, when the inputs *K* and *L* are doubled, i.e., where capital-labour combination is increased from $1K + 1L$ to $2K + 2L$, the output increases from 10 to 18 units, which is less that the proportionate increase. The movement from point *b* to *c* indicates a 50 per cent increase in the inputs. But, the output increases by only 33.3 per cent. This exhibits *decreasing returns to scale*.
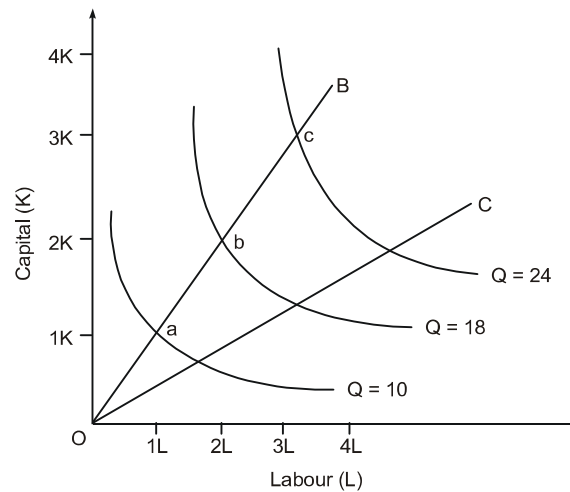
Fig. 6.9 *Decreasing Returns to Scale*

Causes of Diminishing Returns to Scale. The decreasing returns to scale are attributed to the diseconomies of scale. The most important factor causing diminishing returns to scale is 'the diminishing return to management', i.e., managerial diseconomies. As the size of the firms expands, managerial efficiency decreases. Another factor responsible for diminishing returns to scale is the limitedness or exhaustibility of the natural resources. For example, doubling of coalmining plant may not double the coal output because of limitedness of coal deposits or difficult accessibility to coal deposits. Similarly, doubling the fishing fleet may not double the fish output because availability of fish may decrease in the ocean when fishing is carried out on an increased scale.

Check Your Progress

1. Mention the classification of inputs.
2. State the assumptions which are used for drawing isoquant curves.

6.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Economists have classified inputs as (*i*) labour, (*ii*) capital, (*iii*) land, (*iv*) raw materials, (*v*) time and (*vi*) technology.

2. Isoquant curves are drawn on the basis of the following assumptions:

    (*i*) there are only two inputs, viz., labour (*L*) and capital (*K*), to produce a commodity *X*;

    (*ii*) the two inputs—*L* and *K*—can be substituted one for another but at a diminishing rate; and

    (*iii*) the technology of production is given and labour and capital can be substituted only to a certain extent.

6.6 SUMMARY

- Ñ Be it profit maximization or any other objective of business firms, achieving optimum efficiency in production and minimising cost for a given production is one of the prime concerns of the business managers.

- Ñ In economics, the term 'production' means a process by which resources (men, material, time, etc.) are transformed into a *different* and more useful commodity or service. In general, production means transforming inputs (labour, machines, raw materials, time, etc.) into an output with value added. This concept of production is however limited to only 'manufacturing'.

- Ñ An *input* is a good or service that is used into the process of production. In the words of Baumol, "An input is simply anything which the firm buys for use in its production or other proceses." An *output* is any good or service that comes out of production process.

- Ñ The laws of production state the nature of relationship between output and input. The traditional theory of production studies the marginal input-output relationships under (*i*) short run, and (*ii*) long run conditions.

- Ñ *An isoquant curve is locus of points representing various combinations of two inputs—capital and labour—yielding the same output.*

- Ñ The laws of returns to scale state the behaviour of output in response to a proportional and simultaneous change in inputs. Increasing inputs proportionately and simultaneously is, in fact, an expansion of the scale of production.

- Ñ When the increase in output is proportional to the increase in inputs, it exhibits *constant returns to scale*.

6.7 KEY WORDS

- Ñ **Production:** It means a process by which resources (men, material, time, etc.) are transformed into a different and more useful commodity or service.

- Ñ **Isoquant:** It is a curve that shows all the combinations of inputs that yield the same level of output.

6.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. Identify the variable factors in production.

2. Define the law of variable returns with reference to a variable input.

3. State the differences between isoquant curve and indifference curve.

Long Answer Questions

1. Explain the factors behind the law of returns.

2. Critically analyse the application of the law of diminishing returns to business decisions.

3. Describe the kinds of returns to scale with the help of diagrams.

6.9   FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.

UNIT 7 PRODUCTION FUNCTION

Structure

7.0 INTRODUCTION

Production function is a tool of analysis used to explain the input-output relationship. A production function describes the technological relationship between inputs and output in physical terms. In its general form, it tells that production of a commodity depends on certain specific inputs. In its specific form, it presents the quantitative relationships between inputs and output. Besides, the production function represents the technology of a firm, of an industry or of the economy as a whole. A production function may take the form of a schedule or table, a graphed line or curve, an algebraic equation or a mathematical model. But each of these forms of a production function can be converted into its other forms.

7.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the production function
- Define Cobb Douglas production function
- Analyse the CES production function
- Discuss technical progress and production function

7.2 PRODUCTION FUNCTION

Before we illustrate the various forms of a production function, let us note how a complex production function is simplified and the number of inputs in the production

function (used as independent variables) is reduced to a manageable number, especially in theoretical analysis or model building.

An empirical production function is generally very complex. It includes a wide range of inputs, viz., (*i*) land; (*ii*) labour, (*iii*) capital, (*iv*) raw material, (*v*) time, and (*vi*) technology. All these variables enter the actual production function of a firm. The long-run production function is generally expressed as

$$Q = f(L_d, L, K, M, T, t)$$

where L_d = land and building, L = labour, K = capital, M = materials, T = technology, and t = time.

The economists have however reduced the number of variables used in a production function to only two, viz., *capital* (K) and *labour* (L), for the sake of convenience and simplicity in the analysis of input-output relations and production function is expressed as

$$Q = f(L, K)$$

The arguments for eliminating other inputs from the production function are given below.

Land and *building* (L_d), as *inputs,* are constant for the economy as a whole, and hence it does not enter into the aggregate production function. However, land and building are not a constant variable for an individual firm or industry. In the case of individual firms, land and building are lumped with 'capital'.

In case of 'raw materials' it has been observed that this input 'bears a constant relation to output at all levels of production'. For example, cloth bears a constant relation to the number of garments. Similarly, for a given size of a house, the quantity of bricks, cement, steel, etc. remains constant, irrespective of number of houses constructed. To consider another example, in car manufacturing of a particular brand or size, the quantity of steel, number of the engine, and number of tyres and tubes are fixed per car. This constancy of input-output relations leaves the methods of production unaffected.

So is the case, generally, with time and space. That is why, in most production functions only two inputs—labour and capital—are included.

We will illustrate the tabular and graphic forms of a production function when we move on to explain the laws of production. Here, let us illustrate the algebraic or mathematical form of a production function. It is this form of production function which is most commonly used in production analysis.

To illustrate the algebraic form of production function, let us suppose that a coal-mining firm employs only two inputs—capital (K) and labour (L)—in its coal production activity. In this case, the general form of its production function would be expressed as

$$Q = f(K, L) \qquad \qquad \ldots(7.1)$$

where Q = the quantity of coal produced per time unit, K = capital, and L = labour.

The production function (Eq. 7.1) implies that quantity of coal produced depends on the quantity of capital, *K*, and labour, *L*, employed to produce coal. Increasing coal production will require increasing *K* and *L*. Whether the firm can increase both *K* and *L* or only *L* depends on the time period it takes into account for increasing production, i.e., whether the firm considers a *short-run* or a *long-run*.

By definition, supply of capital is *inelastic* in the short run and *elastic* in the long run. In the short run, therefore, the firm can increase coal production by increasing labour only since the supply of capital in the short run is fixed. In the long run, however, the firm can employ more of both capital and labour. Accordingly, there are two kinds of production functions:

> (*i*) *Short-run production function;* and

> (*ii*) *Long-run production function.*

The short-run production function or what may also be termed as *'single variable production function',* can be expressed as

$$Q = f(L) \text{ (with constant K)} \qquad \ldots(7.2)$$

In the long-term production function, both *K* and *L* are included and the function takes the form

$$Q = f(K, L) \qquad \ldots(7.3)$$

Assumptions: A production function is based on the following assumptions.

(*i*) Perfect divisibility of both inputs and output;

(*ii*) There are only two factors of production – labour (*L*) and capital (*K*);

(*iii*) Limited substitution of one factor or the other;

(*iv*) A given technology; and

(*v*) Inelastic supply of fixed factors in the short-run.

A change in these assumptions would necessitate a change in the production function.

The two most important long-run production functions used in economic literature to analyse input-output relationships are Cobb-Douglas and 'Constant elasticity of Substitution' (*CES*) production functions. These production functions are discussed in detail in a subsequent section.

7.3 COBB DOUGLAS

In the preceding section, the laws of returns to scale have been illustrated through a normal and a power function. In this section, we show the application of the most widely used production functions – the multiplicative *power function.* The most popular production function of this category is *'Cobb-Douglas Production Function* of the form

$$Q = AK^a\ L^b \qquad \qquad \ldots(7.4)$$

where *A* is a positive constant; *a* and *b* are positive fractions; and $b = 1 - a$.

The Cobb-Douglas production function is often used in its following form.

$$Q = AK^a\ L^{1-a} \qquad \qquad \ldots(7.5)$$

Properties of Cobb-Douglas Production Function. A power function of this kind has several important properties.

First, the multiplicative form of the power function (7.6) can be changed into its log-linear form as

$$\log Q = \log A + a \log K + b \log L \qquad \qquad \ldots(7.6)$$

In its logarithmic form, the function becomes simple to handle and can be empirically estimated using linear regression analysis.

Secondly, power functions are homogeneous and the degree of homogeneity is given by the sum of the exponents *a* and *b*. If $a + b = 1$, then the production function is homogeneous of degree 1 and implies **constant returns to scale**.

Thirdly, parameters *a* and *b* represent the elasticity co-efficient of output for inputs *K* and *L*, respectively. The output elasticity co-efficient $(\in)$ in respect of capital may be defined as proportional change in output as a result of a given change in *K*, keeping *L* constant. Thus,

$$\in_k = \frac{\partial Q/Q}{\partial K/K} = \frac{\partial Q}{\partial K} \cdot \frac{K}{Q} \qquad \qquad \ldots(7.7)$$

By differentiating the production function $Q = AK^a L^b$ with respect to *K* and substituting the result in Eq. (7.7), we can find the elasticity co-efficient. We know that

$$\frac{\partial Q}{\partial K} = a\ AK^{a-1}\ L^b$$

By substituting the values for *Q* and $\partial Q/\partial K$, we get

$$\in_k = a\ AK^{a-1}\ L^b \left(\frac{K}{AK^a L^b} \right) = a \qquad \qquad \ldots(7.8)$$

Thus, output-elasticity coefficient for *K* is '*a*'. The same procedure may be adopted to show that *b* is the elasticity co-efficient of output for *L*.

Fourthly, constants *a* and *b* represent the relative share of inputs, *K* and *L*, in total output *Q*. The share of *K* in *Q* is given by

$$\frac{\partial Q}{\partial K} \cdot K$$

Similarly, the share of *L* in *Q* is given by

$$\frac{\partial Q}{\partial L} \cdot L$$

The relative share of K in Q an e obtained as

$$\frac{\partial Q}{\partial L} K \cdot \frac{1}{Q} = \frac{a\,AK^{a-1}L^b \cdot K}{AK^a\,L^b} = a$$

Similarly, it can be shown that b gives the relative share of L in Q.

Finally, Cobb-Douglas production function in its general form, $Q = K^a\,L^{1-a}$ implies that at zero cost, there will be zero production.

Some Input-Output Relationships

Some of the concepts used in production analysis can be easily derived from the Cobb-Douglas production function as shown below.

(*i*) Average Product (*AP*) of L and K:

$$AP_L = A\,(K/L)^{1-a}$$
$$AP_K = A\,(L/K)^1$$

(*ii*) Marginal Product of L and K:

$$MP_L = a.A\,(K/L)\,a = a\,(Q/L)$$
$$MP_K = (a-1)\,A\,(L/K)\,a = (1-a)\,Q/K$$

(*iii*) Marginal Rate of Technical Substitution:

$$MRTS_{L,\,k} = \frac{MP_L}{MP_K} = \left[\frac{a}{1-a}\cdot\frac{K}{L}\right]$$

7.4 CES PRODUCTION FUNCTION

In addition to the Cobb-Douglas production function, there are several other forms of production function, viz., 'constant elasticity substitution' (*CES*), 'variable elasticity of substitution' (*VES*), Leontief-type, and linear-type. Of these, the *constant elasticity substitution* (*CES*) production function is more widely used, apart from Cobb-Douglas production function. We will, therefore, discuss the *CES* production function briefly.

The *CES* production function is expressed as

$$Q = A[aK^{-\beta} + (1-a)L^{-\beta}]^{-1/\beta} \qquad \qquad \text{... (7.9)}$$

or $Q = A[aL^{-\beta} + (1-a)K^{-\beta}]^{-1/\beta}$

$(A > 0,\ 0 < \alpha < 1,\ \text{and } \beta > -1)$

where L = labour, K = capital, and A, α and β are the three parameters.

An important property of the *CES* production function is that it is homogeneous of degree 1. This can be proved by increasing both the inputs, K and L, by a constant factor and finding the final outcome. Let us suppose that inputs K and L are increased by a constant factor m. Then the production function given in Eq. (7.9) can be written as follows.

$$Q' = A[\alpha(mK)^{-\beta} + (1 - \alpha)\ (mL)^{-\beta}]^{-1/\beta}$$
$$= A[m^{-\beta}\ \{aK^{-\beta} + (1 - \alpha)L^{-\beta}\}]^{-1/\beta}$$
$$= (m^{-\beta})^{-1/\beta} \times A[\alpha K^{-\beta} + (1 - \alpha)\ L^{-\beta}]^{-1/\beta} \quad ...(10.30)$$

As given in Eq. (7.9) the term $A[\alpha K^{-\beta} + (1 - a)\ L^{-b}]^{-1/\beta} = Q$, By substitution, therefore, we get

$$Q' = mQ$$

Thus, the *CES* production function is homogeneous of degree 1.

Given the production function (7.9), the marginal product of capital (K) can be obtained as

$$\frac{\delta Q}{\delta K} = \frac{1 - \alpha}{A^{\beta}} \cdot \left[\frac{Q}{L}\right]^{\beta+1}$$

and of labour (L) as

$$RTS = \frac{\alpha}{1 - \alpha}\left[\frac{L}{K}\right]^{\beta+1}$$

The rate of technical substittion (*RTS*) can be obtained as

Merits of CES Production Function. *CES* production function has certain advantages over the other functions:

  (*i*) it is a more general form of production function;

 (*ii*) it can be used to analyze all types of returns to scale, and

(*iii*) it removes many of the problems involved in the Cobb-Douglas production function.

Limitations. The *CES* production function has, however, its own *limitations*. Some economists claim that it is not a general form of production function as it does not stand the empirical test. In other words, it is difficult to fit this function to empirical data. Also, Uzawa finds that it is difficult to generalize this function to *n*-number of factors. Besides, in this production function, parameter b combines the effects of two factors, *K* and *L*. When there is technological change, given the scale of production, homogeneity parameter b may be affected by both the inputs. This production function does not provide a measure to separate the effects on the productivity of inputs.

7.5 TECHNICAL PROGRESS AND PRODUCTION FUNCTION

Briefly, technical progress is the adoption of those techniques of production by which resource cost per unit of output comes down. These techniques are adopted only if it is more profitable to do so. A production function relates physical output of a production process to physical inputs or factors of production.

Throughout history, technical progress has played a crucial role in stimulating economic growth. It adds to the productive capacity of an economy. Its production possibility curve (PPC) moves outwards, and widens the scope for its international trade. In an economy, technical progress is said to take place when producers switch over to 'superior' or 'more productive' techniques of production. For this to happen, these superior techniques have to be more profitable as well. Here, the concept of technical superiority means a saving in resource-cost per unit of output (which is the same thing as more output per unit of resource-cost). The term 'resource cost' itself refers to the physical quantities of factors of production and not their money costs. The source of saving in resource costs lies in an increase in the marginal productivity of one or more inputs. For this reason, a technically more productive method may be accompanied by a change in the proportion in which factors of production are used.

7.6 CLASSIFICATION OF TECHNICAL PROGRESS

Technical progress (or technological progress) is an economic measure of innovation. Technical change is thought to be 'neutral' when it raises the productivity of all factors inputs in production (e.g., capital and different types of labour) equally or in the same proportion. Technical change becomes non-neutral when it increases the productivity of some factors more than others. An example on non-neutral technical change is so-called 'skill-biased technical change,' the idea that computerization has raised the productivity of skilled workers more than less-skilled workers.

Technical progress can be classified into two parts:

- **Embodied technical progress:** It refers to improved technology that is exploited by investing in new equipment. New technical changes made are embodied in the equipment.

- **Disembodied technical progress:** improved technology which allows increase in the output produced from given inputs without investing in new equipment.

In the real world, many innovations do not require replacing the entire or some part of the equipment. It can be improved for better use depending upon the change required. Hence, technological progress, whether it is embodied or disembodied, is matter of degree.

Kaldor's technical progress function

The technical progress function developed by Nicholas Kaldor measures technical progress as the rate of growth of labour productivity. It is described by the following statements:

(i) The larger the rate of growth of capital/input per worker, the larger the rate of growth of output per worker, of labour productivity. The rate of growth

of labour productivity is thus explained by the rate of growth of capital intensity.

(ii) In equilibrium capital/input per worker and output per worker grow at the same rate, the equilibrium rate of growth.

(iii) At growth rates below the equilibrium rate of growth, the growth rate of output per worker is larger than the growth rate of capital/input per worker.

(iv) At growth rates above the equilibrium rate of growth it is the other way round, the rate of growth of output per worker is less than the rate of growth of capital/input per worker.

Different technical progress models

- Edward Denison's empirical data showed how technology was a major contributor to economic growth.

- Simon Kuznets in his model proved the importance of technology innovation in growth of an economy.

- Solow identified technology in his aggregate production function. He also suggested a neutral technical progress and economic growth model.

- Harrod also proposed a neutral technical progress and economic growth model.

- The most important neutral technical progress model was discussed by J.R. Hicks. Hicks-neutral technical change is change in the production function of a business or industry which satisfies certain economic neutrality conditions. The concept of Hicks neutrality was first put forth in 1932 by him in the book *The Theory of Wages*. A change is considered to be Hicks neutral if the change does not affect the balance of labour and capital in the products' production function.

Hicks and Harrod's model will be discussed in more detail in the subsequent sections.

7.6.1 Hicks' Model of Neutral Technical Progress

J. R. Hicks classifies technical progress into three varieties (with each variety having its own type of impact upon the PPC of the country concerned), viz., (i) neutral, (ii) labour-saving, and (iii) capital-saving. Thus, for a given level of output,

- a proportionate saving in both L and K (no change in the ratio K/L) denotes neutral technical progress;

- in labour-saving (or labour-stretching) technical progress, the proportionate saving in L is more than the proportionate saving in K—(an increase in the ratio K/L); and

- in capital-saving (or capital-stretching) technical progress, the proportionate saving in K is more than the proportionate saving in L—(a) decline in the ratio K/L).

The three types of technical progress are illustrated in Figure 7.1. For the sake of simplicity it is assumed that technical progress takes place only in one commodity X and in only one country (Cl). In all the three Panels of Figure 7.1, factors of production L and k are measured along horizontal and vertical axes, respectively. We draw two isoquants for a given quantity of commodity X, one representing the position before technical progress (Q) and the second after that (Q'). Every technical innovation shifts the isoquant closer to the point of origin with production equilibrium moving from point P to P'. Correspondingly, the slope (K/L) of the ray from the origin to the point of production equilibrium also changes; and this change in slope, in turn, tells us the variety of the technical progress. In case of a neutral technical progress, the slope of the ray remains the same, while it increases in labour-saving variety and declines in a capital-saving variety.

It is also noteworthy that the shift of isoquant towards the point of origin is on account of an increase in the marginal productivity of one or both factors of production. It also means that, with technical progress, a given quantity of commodity X can be produced with a smaller resource cost. This causes a leftward shift in the isocost line also. The new isocost line (C_1C_1') is parallel to original one (CC') because there is no change in the prices of factors of production.

Panel A of Figure 7.1 illustrates a neutral technical progress; in Panel B it is labour-saving and in Panel C, it is capital-saving.

Technical progress and trade

As in the case of factor endowment, technical progress also shifts PPC of the country outwards. The extent of outward shift and change in slope of PPC depend upon (he nature of technical progress and the extent to which there is a saving in factors of production.

The inferences regarding the impact of technical progress on trade are similar to the ones discussed in the case of factor endowment, and need not be repeated here.

Growth, consumer preferences and trade

Economic growth of a country adds to its capacity to produce, consume, export and import. But economic growth may or may not actually lead to an increase in its international trade. That would depend upon the interaction of a number of additional factors including the following:

- policies pursued by the authorities of the country experiencing economic growth;

- changes, if any, in the import needs of its industries and changes in the tastes and preferences of its consumers,

- policies pursued by its trading partners, and their demand and supply functions;
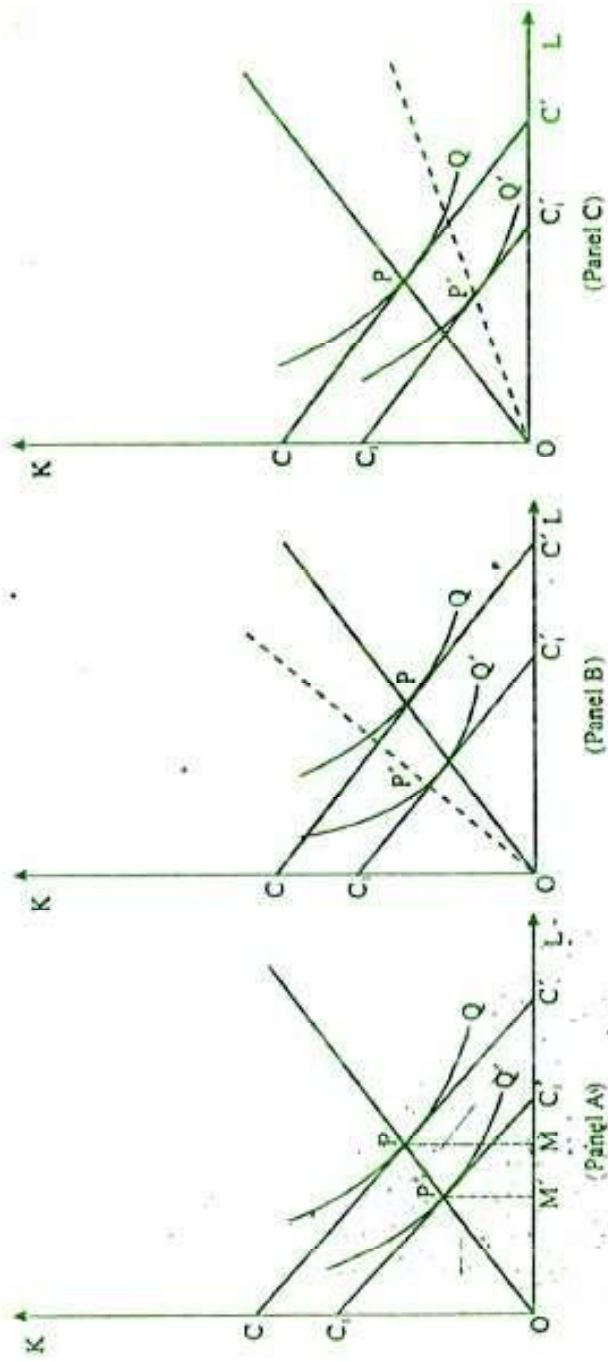
Fig. 7.1 *Varieties of Technical Progress. Technical progress shifts a given isoquant of good X from its old position Q to a new position Q' which is nearer the origin. O. It means that, same quantity of X can be produced with smaller quantities of L and K. The slope of the isocost line CC' or C_1C_1' remains unchanged because of constant prices of inputs L and K. Production equilibrium of C1 shifts from P to P'. Technical progress is classified by comparing the slopes of the rays drawn from the origin O to the equilibrium points P and P'. In panel A both P and P' are on the same ray and technical progress is 'neutral'. In panel B, the slope (K/L) of the ray passing through P' is greater than the slope of ray passing through P and this technical progress is therefore, labour-saving or labour-stretechnig or capital-intensive. In panel C, by the same criterion, technical progress is capital-saving or capital-stretching or labour-intensive.*

- types of the goods produced, such as those which are exported and have a demand abroad, or those which are substitutes for imports;

- shifts in terms of trade, and soon.

Further, the extent of impact on trade depends upon the intensity and interaction of these forces. The net result may be any of the following. It is termed

- protrade when there is a more than proportionate increase in its international trade,

- neutral when there is an equi-proportionate increase in its international trade, and

- antitrade when there is less than proportionate increase, or no increase, or a reduction in its international trade.

The impact on trade may obviously be different if some additional factor comes into play, such as a change in terms of trade, a non-proportionate increase in the productive capacity of the country, a change in consumer preferences, and the like. Similarly, the policy pursued by the authorities has also its impact on the volume and composition of trade.

Concluding Remarks: Growth adds to the production potential of a country. This, in turn, is deeply influenced by the (i) nature of its factor endowment, and/or (ii) the type of technical progress which caused the growth. In addition, the path of growth and its impact on trade are deeply influenced by the

- input needs of the production activities,

- tastes and preferences of consumers (who themselves are partially influenced by the effect of growth on income distribution) and,

- economic and trade policies pursued by the authorities.

Depending upon the complex interplay of these forces, economic growth of a country may lead to proportionate or non-proportionate increase in trade. Thus, for example, if economic growth adds to the output of those goods which the country does not want to export, or for which enough price-elastic demand does not exist abroad, it will not add to its exports. Similarly, unless the growing country can compete in the international markets on the basis of (i) quality and price of its exports, and (ii) acceptance of buyers abroad, its economic growth may not add to its international trade.

Growth adds to production potential of an economy. But it may or may not lead to an increase in its international trade. Trade may even decrease. The net effect on trade would depend upon the interaction of a number of additional factors including the nature of production activities and consumer preferences, etc.

7.6.2 Harrod's Growth Model

Before describing Harrod's theory of economic growth, it is worthwhile to make a few remarks concerning the ideas and goals which Harrod had in mind.

1. Harrod has discussed the following three different concepts of the rate of economic growth.
   (*a*) The actual rate of growth which gives the increase in total output attained during any given time period;
   (*b*) The natural rate of growth which is determined by the growth in the economy's labour force and technological improvements and which may be called the full employment rate of growth; and
   (*c*) The warranted rate of growth which is that rate of growth which the entrepreneurs. counted on and which, if realised, will be repeated by them.
2. Harrod attempts to show how the steady–equilibrium–growth may take place in the system.
3. He also attempts to show that once this steady rate of growth is interrupted, cumulative factors tend to perpetuate the divergence. As a consequence, the economy will experience either a secular stagnation or a secular exhilaration.
4. Harrod's theory is based on the naive acceleration principle as an explanation of the level of investment and the additional investment needed to produce additional output (the capital-output ratio makes it so).
5. Included in Harrod's warranted rate of growth is the equilibrium rate of growth between the (*i*) saving and investment; and (*ii*) total supply and total demand. Both these equilibrium rates are crucial to his theory. When the realised investment exceeds or is less than the planned investment, the warranted rate of growth is interrupted.
6. Under the warranted rate of growth, the net investment equals the amount necessary to produce the increased output of the period.
7. Under the warranted rate of growth, the production decisions are made first and thereafter investment is undertaken to satisfy these production decisions; in a sense, production is assumed to create its own demand.

In short, in Harrod's dynamic system there are two points of crucial importance. *Firstly*, to avoid too much or too little fluctuations in production, income must rise at an ever increasing rate. All the burden is placed on investment because, according to Harrod, saving intentions are always realised. *Secondly,* in Harrod's model even a slight deviation from the warranted rate of growth path tends to be "self-sustaining and possibly self-aggravating." In other words, Harrod's warranted rate of growth is a path which, if once lost, is difficult to regain. It is on account of its precarious balance and volatile behaviour—when there is even a slight deviation from the path of warranted rate of growth—that this model has been labelled as a "razor-edge" model.

Actual Rate of Growth: Harrod explains the actual rate of growth by using his 'fundamental equation' as

$$GC = s \qquad \qquad ...(7.10)$$

where *G* represents the rate of growth of income ($\Delta Y/Y$); *C* represents the ratio of net investment to the change in income ($I_n/\Delta Y$ or $\Delta K/\Delta Y$) and *s* denotes the fraction of income saved (*S/Y*). Equation (1.1) can, therefore, be rewritten as

$$\frac{\Delta Y}{Y} \times \frac{I_n}{\Delta Y} = \frac{S}{Y}$$

Cancelling out ΔY, we obtain

$$\frac{I_n}{Y} = \frac{S}{Y}$$

This means, as should be expected, that *ex post* saving (*S*) equals *ex post* investment (I_n) in any period. With a constant *C*, a higher growth rate means that the growth of capital must be higher (to produce more goods) and that saving must have been higher since income is higher and/or since full employment is assumed. Also, with the given rate of growth, higher is the value of *C*, greater must be the growth of capital (and saving). This occurs since more capital is required per unit of output.

Harrod's fundamental equation had been criticised for giving undue importance to the acceleration principle and for neglecting the autonomous investment. To counter this criticism, Harrod added a *k* term to his fundamental equation modifying it thus:

$$GC = s - k$$

where *k* is the proportion between autonomous investment and net income (I_d/Y).

Warranted Rate of Growth: The concept of the warranted rate of growth is primarily related to the behaviour of entrepreneurs. It is that rate of growth "at which producers will be content with what they are doing. It is the "entrepreneurial equilibrium; it is the line of advance which, if achieved, will satisfy profit takers that they have done the right thing." An entrepreneur will continue producing (and buying inventories) at the same rate in this period as in the immediately preceding period only if he has succeeded in selling his goods. Harrod has discussed this in terms of the aggregate business behaviour. The warranted rate of growth is that rate at which the demand is high enough to enable the business community to sell the goods it has produced (and inventories it has bought). When this happens, the businessmen are happy and they will repeat the performance in the next time period, *i.e.,* businessmen will continue to produce at the same percentage rate of growth. Businessmen then decide to produce a certain quantity of goods and they make the necessary investment to enable them to produce these goods and the decided quantiry.

Thus, there develops a sequence of production decisions, investment and income creation for the consumers. If actual sales and investment correspond those that have been planned, their warranted rate will have been achieved and the same percentage rate of growth will take place in the next time period. If the warranted rate of growth is realised then *ex ante* and *ex post* investment and *ex ante* saving are in equilibrium. If output grows fast enough so that

investment equals savings, the warranted rate of growth will prevail. To express this in Harrod's words, "if the advance is to be maintained ... the quantity of the addition to capital actually accruing must be what is needed."

Harrod's equation for the warranted rate of growth is

$$G_w C_r = s \qquad \qquad ...(1.2)$$

The warranted rate of growth can be represented by $\Delta Y/Y$. The term C_r in the equation is analogous to C in equation (1.2) except that there C is a *ex post* term while C_r in equation (1.1) is an equilibrium term which gives the requirement for new capital, or the required capital coefficient. It is the ratio of addition to capital stock which is required to produce the additional output, *i.e.,*

$$\Delta K/\Delta Y = I_n \Delta Y$$

It is, therefore, a marginal concept and its value may be different from the value of the capital-output ratio for the economy as a whole. Harrod, however, makes the assumption that the two ratios are the same and that the capital-output ratio is, therefore, constant. For achieving an even ratio of growth the actual and warranted rates of growth should be equal, *i.e., $GC = C_w C_r = s$.*

Here it is necessary to indicate certain characteristics of the two growth equations. In the actual rate of growth equation $GC = s$, with a given value of *s,* if G is higher then C must be lower. If both the equations are equal to the same s, and if the actual rate of growth G is greater than the warranted rate of growth G_w, then the ratio of increased capital to increased output C (and, therefore, net investment) must be less than C_r which is the required increase in capital necessary to produce the additional output and *vice versa.* We can elaborate this point by means of the following numerical example.

$$s = 4;\ G = 6;\ C = 8;\ G_w = 4;\ \text{and}\ C_r = 12$$
$$G > G_w$$

Therefore, $C_r > C$

$$G \times C = s$$
$$\downarrow \qquad\quad \downarrow$$
$$6 \times 8 = 48$$
$$G_w \times C_w = s$$
$$\downarrow \qquad\quad \downarrow$$
$$4 \times 12 = 48$$

Again, the actual rate of growth ($G = 6$) exceeds the warranted rate of growth *(G_w=4)*. In so far as the warranted rate of growth is a crucial factor in determining the amount of the capital and inventories which are produced, it is evident that if $G > G_w$ shortage must result. "There will be insufficient goods in the pipeline and/or insufficient equipment…" Such a situation shows that the planned investment is greater than the realised investment and the inventories have been exhausted. Consequently, the orders increase. This divergence between the actual and warranted rates of growth creates an extremely unstable condition in a growing economy and the forces of disequilibrium are at work causing "the system to depart farther and farther from the required line of advance."

The fact that *ex ante* investment exceeds the *ex post* investment means that the aggregate demand exceeds the aggregate supply. There is not enough capital to supply this. Entrepreneurs try to meet this shortfall by raising investment. This increase in investment causes further increase in income and demand leading to further increase in investment to meet the increased demand. Consequently, demand and income increase further and investment increases again resulting in a cumulative departure away from the equilibrium rate of growth. The economy shoots off into hyperinflation.

The only way in which this process can be short-circuited is by increasing the value of *s* (the average propensity to save). According to Harrod, it is not possible to vary this saving-income ratio significantly. Finally, since many people are involved in making the decisions which determine G, the value of G will not automatically become equal to the value of G_w. The opposite cumulative process of the shrinking output, employment and investment occurs when $G < G_w$. In other words, when $G < G_w$ the economy slides into deep depression.

When $G < G_w$, *i.e.,* when the actual rate of growth is less than the warranted rate of growth, if follows from the equality of the actual and warranted growth rate equations that $C < G_r$, *i.e.,* the ratio of the actual net investment to the increment in income is greater than the ratio of the net investment to output necessary to satisfy the equilibrium or warranted rate of growth. We can once again elaborate this point by means of a numerical example setting $s = 48$; $G=4$; $G_w = 6$; $C = 12$; and $C_r = 8$.

$$G_w > G \text{ or } G < G_w$$

Therefore,
$$C > C_r \text{ or } C_r < C;$$

$$G \times C = s$$
$$\downarrow \qquad \downarrow$$
$$4 \times 12 = 48$$
$$G_w \times C_r = s$$
$$\downarrow \qquad \downarrow$$
$$6 \times 8 = 48$$

This means that net investment was too high or that the *ex post* investment exceeds the *ex ante* investment or that the total output produced was not sold. This implies the presence of excess capacity in the economy causing a fall in investment and income in the next time period. Again, once the economy leaves the path of warranted rate of growth as a result of too much realised investment, it moves away and away from this path in a cumulative manner. Thus, in Harrod's growth model, the equilibrium or warranted growth rate path is like a razor's edge.

Natural Rate of Growth: According to Harrod, the natural rate of growth is "the rate of advance which the increase in population and technical improvements allow. The equation for the natural rate of growth is

$$G_n C_r = \text{ or } \neq s$$

where G_n represents the long-run maximum average rate of growth. Although G might exceed G_n during the recovery from a recession, this will be temporary due to the limit set by the population growth and technological progress.

The relationship between G_n and G_w is of great significance in determining whether the economy will move toward secular exhilaration and *vice versa*. According to Harrod, if $G_w > G_n$, the economy will tend towards secular stagnation. This seems somewhat paradoxical because in such a situation it is natural to think that entrepreneurs would like to move forward at a higher rate than could be allowed by labour and productivity growth. When $G_w > G_n$, then G must on an average and over a period of time be less than G_w because the ceiling for G is set by G_n. When G_w is higher than G_n, the investment goods which have been produced remain unused due to the shortage of labour. The shortage of labour sets a limit to the rate at which the output can increase. The rate of output increase is then not equal to G_w; plants, therefore, remain idle and excess capacity results. Such a situation dampens the investment incentives. The decline in investment, income and employment soon becomes cumulative leading to secular stagnation in the same manner in which it occurs when G is below G_w. If G_n exceeds G_w, the economy tends towards secular exhilaration.

Although it is not necessary that $G_w = G_n$, it would, however, be advantageous for the society if $G_w = G_n$ (in which case $G_n = G_w = G$). If, however $G_w > G_n$, steps can be taken to reduce the rate of saving, to raise the level of expenditure and reduce the level of investment (which had become too high and redundant). If $G_n > G_w$, steps can be taken to raise the rate of saving. An increase in saving will allow the warranted rate of growth to increase and release resources for use in the production of capital goods.

Criticisms: Harrod's growth theory has been criticised on several grounds. The assumption of fixed output-capital ratio and an undue reliance on the acceleration principle have severely limited the practical utility of the theory. Further, production plans are less likely to determine investment (and aggregate demand) than the increase in the aggregate demand. It is also doubtful that simply because the warranted rate of economic growth has been realised the entrepreneurs will automatically increase the output by exactly the same percentage rate in the next time period. However, despite its weaknesses, the theory nevertheless focuses our attention on the important fact that over-investment will cause capital scarcity while under-investment will lead to an excess capacity in the economy.

Capital Deepening and Capital Broadening Technical Progress

Economists have, for very good reasons, long emphasized the lack of investment in capital resources as an obstacle to economic development. Saving and investment from current consumption is one central determinant of the rate of economic growth. Today, the more rapidly growing nations are those that are achieving high rates of capital formation. The output produced by ₹ 1 of capital investment, as reflected by the productivity of capital (the output/capital ratio) is dependent upon a host of

qualitative and vaguely understood growth factors, all of which are mutually interdependent. Poor nations do not have a large stock of capital resources relative to labour, however. But technological changes cannot take place in the absence of capital resources, thus impeding progress further.

Paul Romer in his seminal 1986 article on technological progress stated that the concept of capital is broadened to encompass human capital (the knowledge and skills embodied in the workforce), the law of diminishing returns may not apply. The law of diminishing returns basically states that if one input in the production of a commodity is increased while all other inputs are held fixed, a point will eventually be reached at which additions of the input yield progressively smaller, or diminishing, increases in output. Romer argued that what was required was an equilibrium model of endogenous technical change where long-term growth is mainly determined by the amassing of knowledge by forward looking, profit maximizing agents. Romer's initial models suggest either constant or increasing returns to capital.

Deepening of Capital

Capital deepening is a situation where the capital per worker increases in the economy. This is also referred to as increase in the capital intensity. Capital deepening is often measured by the rate of change in capital stock per labour hour.

When capital resources (K) grow more rapidly than the labour supply (L):

- There is an increase in the ratio between K and L.
- Capital deepening is said to occur. In other words, each unit of labour works with more and more capital resources.

But what are the consequences of the deepening of capital? After all, capital deepening (a rising ratio of K/L) is akin to Ricardo's observation of diminishing returns to the use of increased amounts of labour applied to relatively fixed supplies of land. Does not capital deepening lead to diminishing returns to capital resources? The answer is a yes and no. Capital deepening does lead to diminishing returns, but diminishing returns to capital can be (and are) offset.

Under capital-deepening growth conditions, a relatively fixed labour factor cooperates in production with larger and larger amounts of capital, which are subject to diminishing returns. Also, we must remember that the return to investment in capital resources will tend to fall if additional capital investments are successively less productive as diminishing returns occur. Capital deepening accompanied by diminishing returns means that each additional dollar of capital produces less and less additional output. Investment would then decline, and ultimately, dry up because it would be increasingly unprofitable to invest. If population continues to expand, deepening of capital would lead to real output per person ultimately falling to subsistence levels, while profits and investment in physical capital would disappear just as Ricardo predicted.

Technical Progress and Relative Factor Share

The period from 1930s to 1970s saw a lot of debates on the classification of technical progress into neutral, labour-intensive inventions and capital-intensive inventions. In his book, *Economics of Welfare*, Pigou distinguished between technical progress increasing, decreasing or leaving unaffected the ratio of capital to labour. In 1932, as stated earlier, Hicks suggested that if the inventions increased the marginal products of capital and labour, they can be said to be neutral inventions. In 1948, Harrod claimed that the inventions that did not affect the capital-output ratio in any way were neutral inventions.

Neutral technological change refers to the behaviour of technological change in models. A Hicks-neutral technical change is a change in the production function of an enterprise or business that meets certain economic neutrality conditions. This concept by Hicks was expounded by him in his book *The Theory of Wages*. You can say that a change is Hicks-neutral if the change does not affect the balance of labour and capital in the product's production function.

In other words, a technological innovation is said to be Hicks neutral if a change in technology does not alter the ratio of capital's marginal product to labour's marginal product for a given capital to labour ratio. A technological innovation is said to be Harrod neutral if the technology is labour-augmenting (i.e., it helps labour). A technological innovation is said to be Solow neutral if the technology is capital-augmenting (i.e., it helps capital).

Hicks asserted that relative factor prices between capital and labour gives firms an incentive to innovate towards saving the relatively expensive production factor. To explain this, he introduced the concept of biased technical change in a one-sector economy. In a multi-sector economy, skill biases at the aggregate level occur due to biases at the sectoral level as well as due to shifts in demand or differences in unbiased technical change between sectors.

Hicks-neutral technical change may be defined as an increase in productivity which leaves relative factor prices (wages) steady for a given factor employment ratio. Such increase also leaves the factor employment ratio stable for a given wage ratio. Thus, technical change is Hicks neutral so far as parameters that capture factor saving technical progress change proportionately. In a one-sector economy technologically induced changes in relative wages always indicate factor biases in technical change. This is no longer true in case of a multi-sector economy, in which biases in demand at the aggregate level can result not only from non-proportionate changes in the parameters but also from sector biases.

Hicks also introduced a classical definition of input-biased and input-neutral technical change. According to him, the input expansion path is not affected by technical change. Later, Hicks also introduced the taxonomy of productivity growth that distinguishes between import-biased growth, export-biased growth and neutral productivity growth. This taxonomy was based on how technological change affected the shape of the production facilities frontier.

Harrod's concept of the neutral

In his book *The Trade Cycle*, Harrod proposed the concept of the trade cycle. Following the principles of the Keynesian multiplier, Harrod suggests that an increase in income will lead to an increase in investment and cumulative demand, but this trade cycle will eventually lead to a depression because of the continually decreasing rate of growth. As people are able to generate more income, so their savings are increased; this implies a decrease in demand and eventually a reduction in the rate of growth of income. But technological growth is one feature that can break this cycle and unstable equilibrium.

Technological advancement may lead to a change in the amount of additional capital needed to meet the prospected increase in consumption. Thus, neutrality may be regarded as constancy in the capital-output ratio at a given rate of interest. But if this were true, technological progress would not affect economic growth in any way. Harrod, however, was more interested in the economic growth as triggered by an increase in consumption and technological progress was shown to act merely as a catalyst in this equation. It was in his book *Towards a Dynamic Economics* (1948) that Harrod actually expounded his concept of neutral invention.

On the other hand, Hicks defined a neutral invention as one that leaves the ratio of marginal productivity of capital to the marginal productivity of labour unaffected when the relative amounts of the factors are unchanged. Harrod, on the other hand, tried to classify inventions on how they affected the length of the productive process. This classification would help in ascertaining the effect of each kind of invention on the distribution of income. Robinson raised some doubts on Harrod's classification as Harrod's approach did not take production function into account while determining the distribution of income. Robinson and Harrod never agreed with each other. Harrod's approach was more focussed on the amount of capital required to meet the prospected increase in consumption demands. He, however, did take the effects of changing technology into account and the effect of other variants as well. Harrod's concept worked in a dynamic context where a moving equilibrium always persisted and where the stress was on the variables expressing a proportion and not a value.

Harrod maintained that an increase in the level of wages is not capable of substituting capital for labour unless there is a reduction in the rate of interest. For him, any change in the method of production takes place as a direct result of the variations of interest. Harrod asserted that if there is no change in the rate of interest and the length of the productive process, then 'the proportion of the value of the final product that is to be assigned to interest is unchanged; i.e., the share of capital is unchanged'. Harrod rejected the use of production function.

> In favour of the 'period of production' the following points may be made. First, at a given rate of interest industries can be ranked as of greater or less capital intensity by the proportion that interest bears to the value of the product (which simply reflects the length of the production period). Secondly, at a given rate of

interest, improvements can be ranked as capital-saving or capital-requiring by whether they lower or raise the total interest payment as a proportion of the value of the product

<div align="center">(Harrod 1961b, pp. 786-87; see also 1961a, p. 300).</div>

The object of Hick's study was to study the distribution of income and so he defined 'neutral' as those inventions that did not affect the factors affecting the division of income between labour and capital- i.e., the ratio of their marginal productivities. Harrod's classification considers the existing capital-output ratio and falls into the danger of oversimplifying things. His notions work only when neutrality functions in accordance with what he expounds. Harrod explains that:

[t]his definition is based on the idea that existing output can be sustained by existing capital and that additional capital is only required to sustain additional output. This follows from the assumption that the capital/income ratio is constant, i.e. that the length of the production process is unchanged and this follows from the two assumptions on which we are presently working, namely, (1) that inventions are neutral and (2) that the rate of interest is constant

<div align="center">(Harrod 1948, pp. 82-83)</div>

Check Your Progress

1. List the assumptions on which the production function is based.
2. Name the various forms of production function.

7.7 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. A production function is based on the following assumptions.
   (*i*) Perfect divisibility of both inputs and output;
   (*ii*) There are only two factors of production – labour (*L*) and capital (*K*);
   (*iii*) Limited substitution of one factor or the other;
   (*iv*) A given technology; and
   (*v*) Inelastic supply of fixed factors in the short-run.
2. The various forms of production function are Constant Elasticity Substitution (CES), variable Elasticity Substitution (VES), Leontief-type and linear-type.

7.8 SUMMARY

- *Production function* is a tool of analysis used to explain the input-output relationship. A production function describes the technological relationship between inputs and output in physical terms.

- In addition to the Cobb-Douglas production function, there are several other forms of production function, viz., 'constant elasticity substitution' (*CES*), 'variable elasticity of substitution' (*VES*), Leontief-type, and linear-type. Of these, the *constant elasticity substitution* (*CES*) production function is more widely used, apart from Cobb-Douglas production function.

- Technical progress (or technological progress) is an economic measure of innovation. Technical change is thought to be 'neutral' when it raises the productivity of all factors inputs in production (e.g., capital and different types of labour) equally or in the same proportion.

- Growth adds to production potential of an economy. But it may or may not lead to an increase in its international trade. Trade may even decrease. The net effect on trade would depend upon the interaction of a number of additional factors including the nature of production activities and consumer preferences, etc.

- Economists have, for very good reasons, long emphasized the lack of investment in capital resources as an obstacle to economic development. Saving and investment from current consumption is one central determinant of the rate of economic growth.

- The period from 1930s to 1970s saw a lot of debates on the classification of technical progress into neutral, labour-intensive inventions and capital-intensive inventions.

7.9   KEY WORDS

- **Technical progress:** It is defined as an economic measure of innovation.
- **Capital deepening:** It is a situation where the capital per worker increases in the economy.

7.10 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. What are the limitations of the CES production function?
2. Mention the classification of technical progress.
3. What are the criticisms raised against Harrod's model of technical progress?

Long Answer Questions

1. Discuss the properties of the Cobb Douglas production function.

2. Explain Hicks' model of neutral technical progress.

3. 'Neutral technological change refers to the behaviour of technological change in models.' Elucidate the statement.

7.11 FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.

BLOCK - III

THEORY OF COSTS

UNIT 8 THEORY OF COST

Structure

8.0 INTRODUCTION

In the previous unit, we have discussed the production function in detail. However, business decisions are generally taken on the basis of money value of the inputs and outputs. The total money spent on acquiring the inputs—labour, capital, materials and technology—gives the measure of the total *cost of production*. Money received from the sale of output gives the total revenue. In this unit, we are concerned with the theory of cost, i.e., the laws pertaining to cost-output relationship.

The cost of production is an important factor in almost all business analysis and decisions, specially those which pertain to the following aspects.

(*a*) locating the weak points in production management;

(*b*) minimizing the cost as far as possible, for a given output;

(*c*) finding the optimum level of output;

(*d*) determination of price and dealers margin; and

(*e*) estimating or projecting the cost of business operation.

Besides, cost analysis assumes a great significance in all major business decisions because the term 'cost' has different meaning under different settings and is subject to varying interpretations. It is, therefore, essential that only the relevant concept of costs is used in the business decisions. This unit deals with these aspects of cost analysis. We begin with cost concepts used in business decisions. The traditional theory of cost and the modern theory of cost have been discussed in detail in the subsequent units.

8.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the various cost concepts
- Analyse the importance of the theory of cost

8.2 THEORY OF COST: INTRODUCTION

The cost concepts which are relevant to business analysis and decision-making can be grouped, on the basis of their nature and purpose, under two overlapping categories: (*i*) concepts used for accounting purposes, and (*ii*) analytical cost concepts used in economic analysis of business activities. We will discuss here some important concepts of the two categories. It is important to note here that this classification of cost concepts is only a matter of analytical convenience.

Accounting Cost Concepts

1. **Opportunity Cost and Actual Cost.** Resources available to any person, firm or society are scarce but have alternative uses with different returns. Income maximizing resource owners put their scarce resources to their most productive use and thus, they forego the income expected from all other uses of the resources. The income foregone is called *opportunity cost*. While measuring the opportunity cost, the return from the second best use only is taken into account. The *opportunity cost* may be defined as the expected returns form the second best use of the resources were foregone due to the scarcity of resources. The opportunity cost is also called *alternative cost.* If resources available to a person, a firm or a society were unlimited there would be no opportunity cost.

   For example, suppose that a firm has a sum of ₹ 100,000 for which it has only two alternative uses. It can buy either a printing machine or alternatively a lathe machine both having productive life of 10 years. From the printing machine, the firm expects an annual income of ₹ 20,000 and from the lathe, ₹ 15,000. A profit maximizing firm would invest its money in the printing machine and forego the expected income from the lathe. The opportunity cost of the income from printing machine is the expected income from the lathe, i.e., ₹ 15,000.

   Associated with the concept of opportunity cost is the concept of **economic rent** or **economic profit.** In our example of expected earnings firm printing machine and economic rent of the printing machine is the excess of its earning over the income expected from the lathe. That is, economic rent equals ₹ 20,000 – ₹ 15,000 = ₹ 5,000. The implication of this concept for a business man is that investing in the printing machine is preferable so long as its

economic rent is greater than zero. Also, if firms know the economic rent of the various alternative uses of their resources, it will be helpful in the choice of the best investment avenue.

In contrast to the concept of *opportunity cost,* **actual costs** are those which are actually incurred by the firm in payment for labour, material, plant, building, machinery, equipment, travelling and transport, advertisement, etc. The total money expenses, recorded in the books of accounts are for all practical purposes, the **actual costs.** In our example, the cost of printing machine, i.e., ₹ 100,000 is the actual cost. Actual cost comes under the accounting cost concept.

2. **Business Costs and Full Costs.** *Business costs* include all the expenses which are incurred to carry out a business. The concept of business costs is similar to the actual or real costs. Business costs "include all the payments and contractual obligations made by the firm together with the book cost of depreciation on plant and equipment." These cost concepts are used for calculating business profits and losses and for filing returns for income-tax and also for other legal purposes.

The concept of **full cost***,* includes business costs, opportunity cost and normal profit. The opportunity cost includes the expected earning from the second best use of the resources, or the market rate of interest on the total financial capital and also the value of an entrepreneur's own services which are not charged for in the current business. Normal profit is a necessary minimum earning in addition to the opportunity cost, which a firm must receive to remain in its present occupation.

3. **Explicit and Implicit or Imputed Costs:** *Explicit costs* are those which fall under actual or business costs entered in the books of accounts. The payments on account of wages and salaries, materials, license fee, insurance premium, depreciation charges are the examples of explicit costs. These costs involve cash payment and are recorded in normal accounting practices.

In contrast to explicit costs, there are certain other costs which do not take the form of cash outlays, nor do they appear in the accounting system. Such costs are known as **implicit** or **imputed costs***.* Opportunity cost is an important example of implicit cost. For example, suppose an entrepreneur does not utilize his services in his own business and works as a manager in some other firm on a salary basis. If he sets up his own business, he foregoes his salary as manager. This loss of salary is the opportunity cost of income from his own business. This is an implicit cost of his own business. Thus, implicit wages, rent, and implicit interest are the wages, rents and interest which an owner's labour, building and capital, respectively, can earn from their second best use.

Implicit costs are not taken into account while calculating the loss or gains of the business, but they do appear as an important consideration in whether

or not to retain a factor in its present use. The explicit and implicit costs together make the **economic cost**.

4. **Out-of-Pocket and Book Costs:** The items of expenditure which involve cash payments including, both recurring and non-recurring expenses, are known as *out-of-pocket costs*. All the explicit costs (e.g., wages, rent, interest, cost of materials and maintenance, transport expenditure, etc.) fall in this category. On the contrary, there are certain actual business costs which do not involve cash payments, but a provision is therefore made in the books of account and they are taken into account while finalising the profit and loss accounts. For example, payments made by a firm to itself, depreciation allowances and unpaid interest on the owner's own fund are the example of *book costs.*

Analytical Cost Concepts

1. **Fixed and Variable Costs.** *Fixed costs* are those which remain fixed in volume over a certain level of output. Fixed cost does not vary with variation in the output between zero and a certain level of output. In other words, costs that do not vary for a certain level of output are known as *fixed costs*. The fixed costs include (*i*) costs of managerial and administrative staff, (*ii*) depreciation of machinery, building and other fixed assets, (*iii*) maintenance of land, etc. The concept of fixed cost is associated with the short-run. In the long run, not cost is fixed.

   Variable costs are those which vary with the variation in the total output. Variable costs include cost of raw material, running cost of fixed capital, such as fuel, repairs, routine maintenance expenditure, direct labour charges associated with the level of output, and the costs of all other inputs that vary with output.

2. **Total, Average and Marginal Costs.** *Total Cost* (*TC*) is the total expenditure incurred on the production of goods and service. It refers to the total outlays of money expenditure, both explicit and implicit, on the inputs used to produce a given level of output. It includes both fixed and variable costs. That is,

$$TC = TFC + TVC$$

   Average Cost (*AC*) is of statistical nature—it is not actual cost. It is obtained by dividing the total cost (*TC*) by the total output (*Q*), i.e.,

$$AC = \frac{TC}{Q} \qquad \qquad \text{...(8.1a)}$$

   Marginal Cost (*MC*) is the addition to the total cost on account of producing one additional unit of the product. Or, marginal cost is the cost of the marginal unit produced. Marginal cost is calculated as $TC_n - TC_{n-1}$ where *n* is the number of units produced. Alternatively, given the cost function, *MC* can be defined as

$$MC = \frac{\partial TC}{\partial Q} \qquad \qquad \dots(8.1b)$$

These cost concepts are discussed in further detail in the following section. The total, average and marginal cost concepts are used in the economic analysis of firm's production activities.

3. **Short-Run and Long-Run Costs.** Short-run and long-run cost concepts are related to variable and fixed costs, respectively, and often figure in economic analysis interchangeably.

Short-run costs are the costs which have short-run implication in production process and vary with the variation in output, the size of the firm remaining the same. *Short-run* costs are the same as *variable costs*. *Long-run* costs, on the other hand, are the costs which are incurred on the fixed assets like plant, building, machinery, etc. It is important to note that the running cost and depreciation of the capital assets are included in the short-run or variable costs.

Long-run costs are by implication the same as *fixed costs*. In the long-run, however, even the fixed costs become variable costs as the size of the firm or scale of production increases. Broadly speaking, 'the short-run costs are those associated with variables in the utilization of fixed plant or other facilities whereas long-run costs are associated with the changes in the size and kind of plant.'

4. **Incremental Costs and Sunk Costs.** Conceptually, *incremental costs* are closely related to the concept of marginal cost but with a relatively wider connotation. While marginal cost refers to the cost of the marginal unit of output, incremental cost refers to the total additional cost associated with the decisions to expand the output or to add a new variety of product, etc. The concept of incremental cost is based on the fact that in the real world, it is not practicable (for lack of perfect divisibility of inputs) to employ factors for each unit of output separately. Besides, in the long run, when firms expand their production, they hire more of men, materials, machinery and equipments. The expenditures of this nature are incremental costs and not the marginal cost (as defined earlier). Incremental costs arise also owing to the change in product lines, addition or introduction of a new product, replacement of worn out plant and machinery, replacement of old technique of production with a new one, etc.

The *sunk costs* are those which are incurred once for all. Such costs cannot be altered, increased or decreased, by varying the rate of output. For example, once it is decided to make incremental investment expenditure and the funds are allocated and spent, all the preceding costs are considered to be the sunk costs since they accord to the prior commitment and cannot be revised or reversed or recovered when there is a change in market conditions or change in business decisions.

5. **Historical and Replacement Costs.** *Historical cost* refers to the cost of an asset acquired in the past whereas *replacement cost* refers to the expenditure made for replacing an old asset. These concepts owe their significance to the unstable nature of price behaviour. Stable prices over time, other things given, keep historical and replacement costs on par with each other. Instability in asset prices makes the two costs differ from each other.

Historical *cost* of assets is used for accounting purposes, in the assessment of the net worth of the firm. The replacement cost figures in business decisions regarding the renovation of the plant.

6. **Private and Social Costs.** We have so far discussed the cost concepts that are related to the working of the firm and that are used in the cost-benefit analysis of business decisions. There are, however, certain other costs which arise due to the functioning of the firm but do not normally figure in the business decisions nor are such costs explicitly borne by the firms. The costs on this category are borne by the society. Thus, the total cost generated by a firm's working may be divided into two categories: (*i*) costs paid out or provided for by the firms, and (*ii*) costs not paid or borne by the firms including the use of resources freely available plus the disutility created in the process of production. The costs of the former category are known as *private costs* and costs of the latter category are known as *external* or *social costs*. To mention a few examples of social cost: Mathura Oil Refinery discharging its wastage in the Yamuna river causes water pollution. Mills and factories located in a city cause air pollution, and so on. Let us now look at these concepts of cost in some detail.

Private costs are those which are actually incurred or provided for by an individual or a firm on the purchase of goods and services from the market. For a firm, all the actual costs both explicit and implicit are private costs. Private costs are internalized costs that are incorporated in the firm's total cost of production.

Social costs, on the other hand, refer to the total cost borne by the society due to production of a commodity. Social cost includes both private cost and the external cost. Social cost includes (*a*) the cost of resources for which the firm is not required to pay a price, i.e., atmosphere, rivers, lakes, and also for the use of public utility services like roadways, drainage system, etc., and (*b*) the cost in the form of 'disutility' created through air, water and noise pollution, etc. The costs of category (*c*) are generally assumed to equal the total private and public expenditure incurred to safeguard the individual and public interest against the various kinds of health hazards created by the production system. The private and public expenditure, however, serve only as an indicator of 'public disutility'—they do not give the exact measure of the public disutility or the social costs.

8.2.1 Importance

Costs form a vital component of the field of economics as economics is concerned with the study of choices. The alternatives that an individual selects are based by weighing his unlimited wants with the limited available resources to accomplish what he want. Hence, it is not possible for the consumer to make his decision without taking the cost aspect into consideration.

Cost-benefit analysis occupy a significant place in the day-to-day life as well as in the management of businesses and large corporations. Take for example. Whenever, we go in the market to make a purchase, we do a cost-benefit analysis. If the cost offered is too high, the purchase is not made. However, if the benefit is high then you might pay more as compared to purchasing an object which offers lesser benefits.

Check Your Progress

1. What are explicit costs?
2. Give examples of book cost.

8.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Explicit costs are those which fall under actual or business costs entered in the books of accounts. The payments on account of wages and salaries, materials, license fee, insurance premium, depreciation charges are the examples of explicit costs.

2. For example, payments made by a firm to itself, depreciation allowances and unpaid interest on the owner's own fund are the example of *book costs.*

8.4 SUMMARY

- The cost concepts which are relevant to business analysis and decision-making can be grouped, on the basis of their nature and purpose, under two overlapping categories: (*i*) concepts used for accounting purposes, and (*ii*) analytical cost concepts used in economic analysis of business activities.

- The concept of full cost, includes business costs, opportunity cost and normal profit.

- In contrast to explicit costs, there are certain other costs which do not take the form of cash outlays, nor do they appear in the accounting system. Such costs are known as implicit or imputed costs.

- The items of expenditure which involve cash payments including, both recurring and non-recurring expenses, are known as *out-of-pocket costs*.

- Fixed costs are those which remain fixed in volume over a certain level of output. Fixed cost does not vary with variation in the output between zero and a certain level of output.
- Total Cost (TC) is the total expenditure incurred on the production of goods and service.
- Marginal Cost (MC) is the addition to the total cost on account of producing one additional unit of the product.
- Historical cost refers to the cost of an asset acquired in the past whereas *replacement cost* refers to the expenditure made for replacing an old asset.

8.5 KEY WORDS

- **Fixed cost:** These are those which remain fixed in volume over a certain level of output. Fixed cost does not vary with variation in the output between zero and a certain level of output.
- **Opportunity cost:** It may be defined as the expected returns form the second best use of the resources were foregone due to the scarcity of resources.

8.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. State the differences between business cost and full cost.
2. Write a short note on short-run and long-run cost.
3. Why is opportunity cost also known as alternative cost?

Long Answer Questions

1. 'Social cost includes both private cost and the external cost.' Explain the statement.
2. Analyse the importance of the theory of cost.

8.7 FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.

UNIT 9 THE TRADITIONAL THEORY OF COSTS

Structure

9.0 INTRODUCTION

In the previous unit, you were introduced to the theory of cost wherein you studied about to the cost concepts and the significance of cost analysis. This unit will further move ahead and introduce you to the concepts covered under the traditional theory of cost namely, short-run cost function and cost curves and the concept of long-run marginal cost curves.

9.1 OBJECTIVES

After going through this unit, you will be able to:

- Define cost function
- Explain short-run cost functions and cost curves
- Discuss the concept of long-run marginal cost curve

9.2 TRADITIONAL THEORY OF COSTS: AN OVERVIEW

Cost function is a symbolic statement of the technological relationship between cost and output. In its general form, it is expressed by an equation. Cost function can be expressed also in the form of a schedule and a graph. In fact, tabular, graphical, and algebraic equation forms of cost function can be converted in the form of each other. Going by its general form, total cost (*TC*) function is expressed as follows.

$$TC = f(Q)$$

using cost function. The long-run cost-output relations are discussed in the following section.

9.2.1 Cost Concepts used in Cost Analysis

Before we discuss the cost-output relations, let us first look at the cost concepts and the components used to analyse the short-run cost-output relations.

The basic analytical cost concepts used in the analysis of cost behaviour are Total, Average and Marginal costs. The total cost *(TC)* is defined as the actual cost that must be incurred to produce a given quantity of output. The short-run *TC* is composed of two major elements: (*i*) *total fixed cost (TFC),* and (*ii*) *total variable cost (TVC)*. That is, in the short-run,

$$TC = TFC + TVC \qquad\qquad ...(9.1)$$

As mentioned earlier, *TFC* (i.e., the cost of plant, machinery building, etc.) remains fixed in the short-run, whereas *TVC* varies with the variation in the output.

For a given quantity of output *(Q)*, the average total cost *(AC)*, average fixed cost *(AFC)* and average variable cost *(AVC)* can be defined as follows:

$$AC = \frac{TC}{Q} = \frac{TFC + TVC}{Q}$$

$$AFC = \frac{TFC}{Q}$$

$$AVC = \frac{TVC}{Q}$$

and $\qquad\qquad AC = AFC + AVC \qquad\qquad ...(9.2)$

Marginal cost *(MC)* is defined as the change in the total cost divided by the change in the total output, i.e.,

$$MC = \frac{\Delta TC}{\Delta Q} \qquad\qquad ...(9.3)$$

or as the first derivative of cost function, i.e., $\frac{\partial TC}{\partial Q}$.

Note that since $\Delta TC = \Delta TFC + \Delta TVC$ and, in the short-run, $\Delta TFC = 0$, therefore, $\Delta TC = \Delta TVC$. Furthermore, under the marginality concept, where $\Delta Q = 1$, $MC = \Delta TVC$. Now we turn to cost function and derivation of cost curves.

9.2.2 Short-Run Cost Functions and Cost Curves

The cost-output relations are determined by the cost function and are exhibited through cost curves. The shape of the cost curves depends on the nature of the cost function. Cost functions are derived from actual cost data of the firms. Given the cost data, estimated cost functions may take a variety of forms, yielding different kinds of cost curves. The cost curves produced by *linear, quadratic* and *cubic cost functions* are illustrated below.

1. Linear Cost Function. A linear cost function takes the following form.

$$TC = a + bQ \qquad\qquad …(9.4)$$

(where TC = total cost, Q = quantity produced, $a = TFC$, and $b = \partial TC/\partial Q$).

Given the cost function (Eq. 9.4), AC and MC can be obtained as follows.

$$AC = \frac{TC}{Q} = \frac{a + bQ}{Q} = \frac{a}{Q} + b$$

and $MC = \dfrac{\partial TC}{\partial Q} = b$

Note that since '*b*' is a constant factor, MC remains constant throughout in case of a linear cost function.

Assuming an actual cost function given as

$$TC = 60 + 10Q \qquad\qquad …(9.5)$$

the cost curves (TC, TVC and TFC) are graphed in Fig. 9.1.
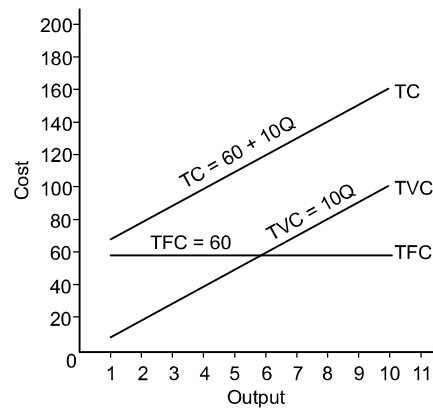
Fig. 9.1 *Linear Cost Functions*

Given the cost function (Eq. 9.5),

$$AC = \frac{60}{Q} + 10$$

and $\qquad\qquad MC = 10$

Fig. 9.2 shows the behaviour of TC, TVC and TFC. The straight horizontal line shows TFC and the line marked $TVC = 10Q$ shows the movement in TVC. The total cost function is shown by $TC = 60 + 10Q$.
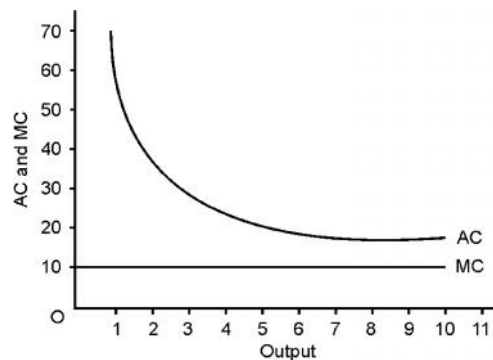
Fig. 9.2 *AC and MC Curves Derived from Linear Cost Function*

More important is to notice the behaviour of *AC* and *MC* curves in Fig. 9.2. Note that in case of a linear cost function *MC* remains constant, while *AC* continues to decline with the increase in output. This is so simply because of the logic of the linear cost function.

2. Quadratic Cost Function. A quadratic cost function is of the form

$$TC = a + bQ + Q^2 \qquad\qquad ....(9.6)$$

where *a* and *b* are constants.

Given the cost function (Eq. 9.6), *AC* and *MC* can be obtained as follows.

$$AC = \frac{TC}{Q} = \frac{a + bQ + Q^2}{Q} = \frac{a}{Q} + b + Q \qquad\qquad ...(9.7)$$

$$MC = \frac{\partial TC}{\partial Q} = b + 2Q \qquad\qquad ...(9.8)$$

Let us assume that the actual (or estimated) cost function is given as

$$TC = 50 + 5Q + Q^2 \qquad\qquad ...(9.9)$$

Given the cost function (Eq. 8.10),

$$AC = \frac{50}{Q} + Q + 5 \qquad \text{and} \qquad MC = \frac{\partial C}{\partial Q} = 5 + 2Q$$

The cost curves that emerge from the cost function (9.9) are graphed in Fig. 9.3 (*a*) and (*b*). As shown in panel (*a*), while fixed cost remains constant at 50, *TVC* is increasing at an increasing rate. The rising *TVC* sets the trend in the total cost *(TC)*. Panel (*b*) shows the behaviour of *AC*, *MC* and *AVC* in a quadratic cost function. Note that *MC* and *AVC* are rising at a constant rate whereas *AC* first declines and then increases.
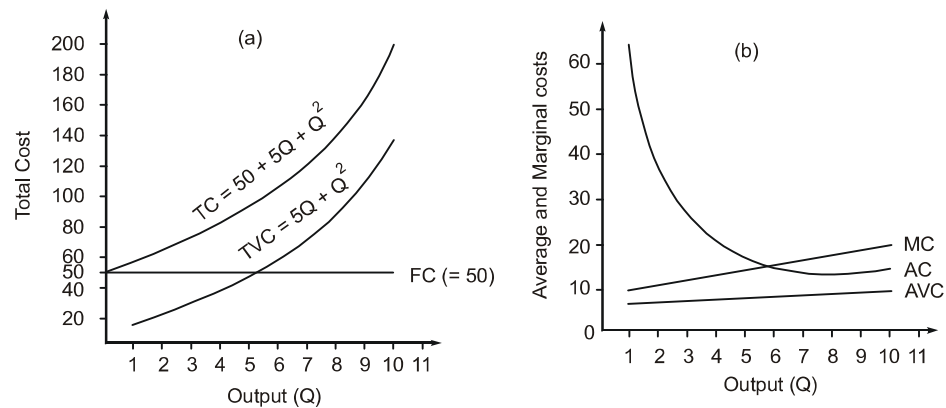
Fig. 9.3 *Cost Curves Derived from a Quadratic Cost Function*

3. Cubic Cost Function. A cubic cost function is of the form

$$TC = a + bQ - cQ^2 + Q^3 \qquad\qquad ...(9.10)$$

where *a*, *b* and *c* are the parametric constants.

From the cost function (9.10), *AC* and *MC* can be derived as follows.

$$AC = \frac{TC}{Q} = \frac{a + bQ - cQ^2 + Q^3}{Q} = \frac{a}{Q} + b - cQ + Q^2$$

and $MC = \dfrac{\partial TC}{\partial Q} = b - 2cQ + 3Q^2$

Let us suppose that the cost function is empirically estimated as

$TC = 10 + 6Q - 0.9Q^2 + 0.05Q^3$ …(9.11)

Given the cost function (9.11), the *TVC* function can be derived as

$TVC = 6Q - 0.9Q^2 + 0.05Q^3$ …(9.12)

The *TC* and *TVC*, based on Eqs. (9.11) and (9.12), respectively, have been calculated for *Q* = 1 to 16 and presented in Table 9.1. The *TFC*, *TVC* and *TC* have been graphically presented in Fig. 9.4. As the figure shows, *TFC* remains fixed for the whole range of output, and hence, takes the form of a horizontal line—*TFC*. The *TVC* curve shows that the total variable cost first increases at a decreasing rate and then at an increasing rate with the increase in the output. The rate of increase can be obtained from the slope of the *TVC* curve. The pattern of change in the *TVC* stems directly from the law of increasing and diminishing returns to the variable inputs. As output increases, larger quantities of variable inputs are required to produce the same quantity of output due to diminishing returns. This causes a subsequent increase in the variable cost for producing the same output.
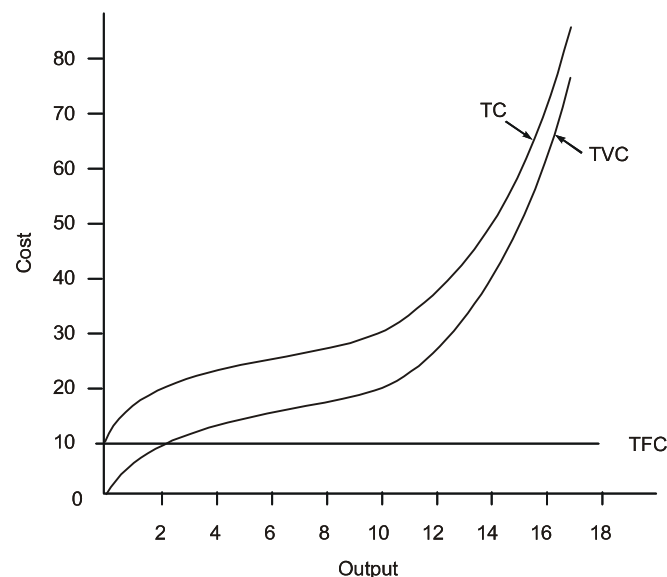
Fig. 9.4 *TC, TFC and TVC Curves*

Table 9.1 *Cost-Output Relations*

Q	FC	TVC	TC	AFC	AVC	AC	MC
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0	10	0.0	10.00	—	—	—	—
1	10	5.15	15.15	10.00	5.15	15.15	5.15
2	10	8.80	18.80	5.00	4.40	9.40	3.65
3	10	11.25	21.25	3.33	3.75	7.08	2.45
4	10	12.80	22.80	2.50	3.20	5.70	1.55
5	10	13.75	23.75	2.00	2.75	4.75	0.95
6	10	14.40	24.40	1.67	2.40	4.07	0.65
7	10	15.05	25.05	1.43	2.15	3.58	0.65
8	10	16.00	26.00	1.25	2.00	3.25	0.95
9	10	17.55	27.55	1.11	1.95	3.06	1.55
10	10	20.00	30.00	1.00	2.00	3.00	2.45
11	10	23.65	33.65	0.90	2.15	3.05	3.65
12	10	28.80	38.80	0.83	2.40	3.23	5.15
13	10	35.75	45.75	0.77	2.75	3.52	6.95
14	10	44.80	54.80	0.71	3.20	3.91	9.05
15	10	56.25	66.25	0.67	3.75	4.42	11.45
16	10	70.40	80.40	0.62	4.40	5.02	14.15

From Eqs. (9.11) and (9.12), we may derive the behavioural equations for *AFC*, *AVC* and *AC*. Let us first consider *AFC*.

Average Fixed Cost (AFC). As already mentioned, the costs that remain fixed for a certain level of output make the total fixed cost in the short-run. The fixed cost is represented by the constant term '*a*' in Eq. (9.10) and $a = 10$ as given in Eq. (9.11). We know that

$$AFC = \frac{TFC}{Q} \qquad ....(9.13)$$

Substituting 10 for *TFC* in Eq. 9.8, we get

$$AFC = \frac{10}{Q} \qquad ....(9.14)$$

Equation (9.14) expresses the behaviour of *AFC* in relation to change in *Q*. The behaviour of *AFC* for *Q* from 1 to 16 is given in Table 9.1 (col. 5) and presented graphically by the *AFC* curve in Fig. 9.5. The *AFC* curve is a rectangular hyperbola.

Average Variable Cost (AVC). As defined above, $AVC = TVC/Q$. Given the *TVC* function (Eq. 9.12), we may express *AVC* as follows.

$$AVC = \frac{6Q - 0.9Q^2 + 0.05Q^3}{Q}$$

$$= 6 - 0.9Q + 0.05Q^2 \qquad ...(9.15)$$

Having derived the *AVC* function in Eq. (9.15), we may easily obtain the behaviour of *AVC* in response to change in *Q*. The behaviour of *AVC* for $Q = 1$ to 16 is given in Table 9.1 (col. 6), and graphically presented in Fig. 9.2 by the *AVC* curve.
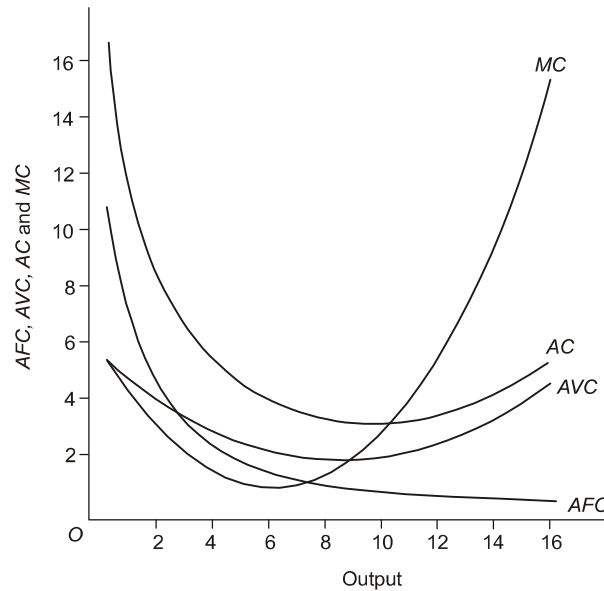
Fig. 9.5 *Short-run AFC, AVC, AC and MC Curves*

Critical Value of AVC. From Eq. (9.9), we may compute the critical value of *Q* in respect of *AVC*. *The critical value of Q (in respect of AVC) is one that minimizes AVC.* The *AVC* will be minimum when its rate of decrease equals zero. This can be accomplished by differentiating Eq. (8.15) and setting it equal to zero. Thus, critical value of *Q* can be obtained as

Critical value of $\qquad Q = \dfrac{\partial AVC}{\partial Q} = -\,0.9 + 0.10\,Q = 0$

$$0.10\,Q = 0.9$$
$$Q = 9$$

In our example, the critical value of $Q = 9$. This can be verified from Table 9.1. The *AVC* is minimum (1.95) at output 9.

Average Cost (AC). The average cost (*AC*) is defined as $AC = \dfrac{TC}{Q}$.

Substituting Eq. (9.11) for *TC* in the above equation, we get

$$AC = \frac{10 + 6Q - 0.9Q^2 + 0.05Q^3}{Q}$$

$$= \frac{10}{Q} + 6 - 0.9Q + 0.05Q^2 \qquad\qquad ...(9.16)$$

The Eq. (9.16) gives the behaviour of *AC* in response to change in *Q*. The behaviour of *AC* for *Q* = 1 to 16 is given in Col. 7 of Table 9.1 and graphically presented in Fig. 9.5 by the *AC* curve. Note that *AC* curve is *U*-shaped.

Minimization of AC. One objective of business firms is to minimize *AC* of their product or, which is the same as, to optimize the output. The level of output that minimizes *AC* can be obtained by differentiating Eq. (9.16) and setting it equal to zero. Thus, the optimum value of *Q* can be obtained as follows.

$$\frac{\partial AC}{\partial Q} = \frac{10}{Q^2} - 0.9 + 0.1Q = 0$$

When simplified (multiplied by Q^2) this equation takes the quadratic form as

$$- 10 - 0.9Q^2 + 0.1Q^3 = 0$$

or $$Q^3 - 9Q^2 - 100 = 0 \qquad ...(9.17)$$

By solving equation (9.17) we get *Q* = 10.

Thus, the critical value of output in respect of *AC* is 10. That is, *AC* reaches its minimum at *Q* = 10. This can be verified from Table 9.1.

Marginal Cost (MC). The concept of marginal cost (*MC*) is useful particularly in economic analysis. *MC* is technically the first derivative of the *TC* function. Given the *TC* function in Eq. (9.11), the *MC* function can be obtained as

$$MC = \frac{\partial TC}{\partial Q} = 6 - 1.8Q + 0.15Q^2 \qquad ...(9.18)$$

Eq. (9.18) represents the behaviour of *MC*. The behaviour of *MC* for *Q* = 1 to 16 computed as $MC = TC_n - TC_{n-1}$ is given in Table 9.1 (col. 8) and graphically presented by the *MC* curve in Fig. 9.5. The critical value of *Q* with respect to *MC* is 6 or 7. This can be seen from Table 9.1.

9.2.3 Cost Curves and the Law of Diminishing Returns

Now we return to the law of variable proportions and explain it through the cost curves. Figures 9.4 and 9.5 represent the cost curves conforming to the short-term law of production, i.e., the law of diminishing returns. Let us recall the law: it states that when more and more units of a variable input are applied, other inputs held constant, the returns from the marginal units of the variable input may initially increase but it decreases eventually. The same law can also be interpreted in terms of decreasing and increasing costs. The law can then be stated as, if more and more units of a variable input are applied to a given amount of a fixed input, the marginal cost initially decreases, but eventually increases. Both interpretations of the law yield the same information—one in terms of **marginal productivity** of the variable input, and the other in terms of the **marginal cost**. The former is expressed through a production function and the latter through a cost function.

Figure 9.5 presents the short-run laws of return in terms of cost of production. As the figure shows, in the initial stage of production, both *AFC* and *AVC* are declining because of some internal economies. Since *AC* = *AFC* + *AVC*, *AC* is

also declining. This shows the operation of the law of increasing returns to the variable input. But beyond a certain level of output (i.e., 9 units in our example), while *AFC* continues to fall, *AVC* starts increasing because of a faster increase in the *TVC*. Consequently, the rate of fall in *AC* decreases. The *AC* reaches its minimum when output increases to 10 units. Beyond this level of output, *AC* starts increasing which shows that the law of diminishing returns comes into operation. The *MC* curve represents the change in both the *TVC* and *TC* curves due to change in output. A downward trend in the *MC* shows increasing marginal productivity of the variable input due mainly to internal economy resulting from increase in production. Similarly, an upward trend in the *MC* shows increase in *TVC*, on the one hand, and decreasing marginal productivity of the variable input, on the other.

Some important relationships between costs used in analysing the short-run cost-behaviour may now be summed up as follows:

(*a*) Over the range of output both *AFC* and *AVC* fall, *AC* also falls because *AC* = *AFC* + *AVC*.

(*b*) When *AFC* falls but *AVC* increases, change in *AC* depends on the rate of change in *AFC* and *AVC*.

    (*i*) if decrease in *AFC* > increase in *AVC*, then *AC* falls,

    (*ii*) if decrease in *AFC* = increase in *AVC*, *AC* remains constant, and

    (*iii*) if decrease in *AFC* < increase in *AVC*, then *AC* increase.

(*c*) The relationship between *AC* and *MC* is of a varied nature. It may be described as follows:

(*i*) When *MC* falls, *AC* follows, over a certain range of initial output. When *MC* is falling, the rate of fall in *MC* is greater than that of *AC*, because in the case of *MC* the decreasing marginal cost is attributed to a single marginal unit while, in case of *AC*, the decreasing marginal cost is distributed over the entire output. Therefore, *AC* decreases at a lower rate than *MC*.

(*ii*) Similarly, when *MC increases, AC also increases but at a lower rate* for the reason given in (*i*). There is, however, a range of output over which the relationship does not exist. Compare the behaviour of *MC* and *AC* over the range of output from 6 to 10 units (Fig. 9.5). Over this range of output, *MC* begins to increase while *AC* continues to decrease. The reason for this can be seen in Table 9.1: when *MC* starts increasing, it increases at a relatively lower rate which is sufficient only to reduce the rate of decrease in *AC*—not sufficient to push the *AC* up. That is why *AC* continues to fall over some range of output even if *MC* increases.

(*iii*) The *MC curve intersects the AC at its minimum point*. This is simply a mathematical relationship between *MC* and *AC* curves when both of them are obtained from the same *TC* function. In simple words, when *AC* is at its minimum, it is neither increasing nor decreasing: it is constant. When *AC* is constant, *AC* = *MC*. That is the point of intersection.

Output Optimization in the Short-Run

Optimization of output in the short-run has been illustrated graphically in Fig. 9.5. Optimization technique is repeated here for the sake of completeness.

Let us suppose that a short-run cost function is given as

$$TC = 200 + 5Q + 2Q^2 \qquad \qquad \text{…(9.19)}$$

We have noted above that an optimum level of output is one that equalizes *AC* and *MC*. In other words, at optimum level of output, *AC* = *MC*. Given the cost function in Eq. (9.19),

$$AC = \frac{200 + 5Q + 2Q^2}{Q} = \frac{200}{Q} + 5 + 2Q \qquad \text{... (9.20)}$$

and

$$MC = \frac{\partial TC}{\partial Q} = 5 + 4Q \qquad \qquad \text{... (9.21)}$$

By equating *AC* and *MC* equations, i.e., Eqs. (9.20) and (9.21), respectively, and solving them for *Q*, we get the optimum level of output. Thus,

$$\frac{200}{Q} + 5 + 2Q = 5 + 4Q,$$

$$\frac{200}{Q} = 2Q$$

$$2Q^2 = 200 \quad \text{or} \quad Q = 10$$

Thus, given the cost function (9.19), the optimum output is 10.

Long-Run Cost-Output Relations

By definition, long-run is a period in which all the inputs–specifically, labour and capital–become variable. The variability of inputs is based on the assumption that in the long-run supply of all the inputs, including those held constant in the short-run, becomes elastic. The firms are, therefore, in a position to expand the scale of their production by hiring a larger quantity of both labour and capital. The long-run-cost-output relations, therefore, imply the relationship between the changing scale of the firm and the total output, whereas in the short-run this relationship is essentially one between the total output and the variable cost (labour).

Derivation of Long-run Cost Curves

To understand the long-run-cost-output relations and to derive long-run cost curves it will be helpful to imagine that a long-run is composed of a series of short-run production decisions. As a corollary of this, long-run cost curve is composed of a series of short-run cost curves. Based on this principle, we may now derive the long-run cost curves and study their relationship with output.

Long-run Total Cost Curve (LTC). In order to draw the long-run total cost curve, let us begin with a short-run situation. Suppose that a firm having only one plant has its short-run total cost curve as given by STC_1, in panel (*a*) of Fig. 9.6. Let us now suppose that the firm decides to add two more plants to its size over time, one after the other. As a result, two more short-run total cost curves are

added to STC_1, in the manner shown by STC_2 and STC_3 in Fig. 9.6 (*a*). The *LTC* can now be drawn through the minimum points of STC_1, STC_2 and STC_3 as shown by the *LTC* curve corresponding to each *STC*.
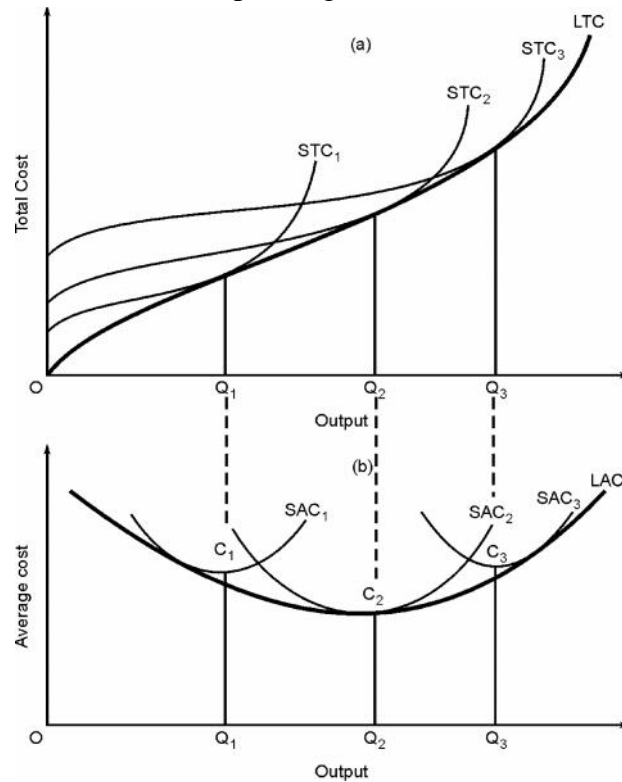
Fig. 9.6 *Long-run Total and Average Cost Curves*

Long-run Average Cost Curve (LAC). The long-run average cost curve (*LAC*) is derived by combining the short-run average cost curves (SAC_s). Note that there is one *SAC* associated with each *STC*. Given the STC_1, STC_2, STC_3 curves in panel (*a*) of Fig. 9.6 there are three corresponding *SAC* curves as given by SAC_1, SAC_2, and SAC_3 curves in panel (*b*) of Fig. 9.6. Thus, the firm has a series of *SAC* curves, each having a bottom point showing the minimum *SAC*. For instance, C_1Q_1 is minimum *AC* when the firm has only one plant. The *AC* decreases to C_2Q_2 when the second plant is added and then rises to C_3Q_3 after the addition of the third plant. The *LAC* curve can be drawn through the bottom points of the SAC_1, SAC_2 and SAC_3 as shown in Fig. 9.6(*b*) The *LAC* curve is also known as the **'Envelope Curve'** or **'Planning Curve'** as it serves as a guide to the entrepreneur in his plans to expand production.

Alternatively, the *SAC* curves can be derived from the data given in the *STC* schedule, from *STC* function or straightaway from the *STC* curve. Similarly, *LAC* and can be derived from *LTC*-schedule, *LTC* function or form *LTC*-curve.

The relationship between *LTC* and output, and between *LAC* and output can now be easily derived. It is obvious from the *LTC* that the long-run cost-output relationship is similar to the short-run cost-output relation. With the subsequent increases in the output, *LTC* first increases at a decreasing rate, and then at an

increasing rate. As a result, *LAC* initially decreases until the optimum utilization of the second plant and then it begins to increase. These cost-output relations follow the 'laws of returns to scale'. When the scale of the firm expands, unit cost of production initially decreases, but ultimately increases as shown in Fig. 9.6(*b*). The decrease in unit cost is attributed to the internal and external economies of scale and the eventual increase in cost is linked to the internal and external diseconomies. The economies and diseconomies of scale are discussed in the following section.

Long-run Marginal Cost Curve (LMC). The *long-run marginal cost curve* (*LMC*) is derived from the short-run marginal cost curves (*SMCs*). The derivation of *LMC* is illustrated in Fig. 9.7 in which *SACs* and *LAC* are the same as in Fig. 9.6 (*b*) To derive the *LMC*, consider the points of tangency between *SACs* and the *LAC*, i.e., points *A*, *B* and *C*. In the long-run production planning, these points determine the output levels at the different levels of production. For example, if we draw perpendiculars from points *A*, *B* and *C* to the *X*-axis, the corresponding output levels will be OQ_1, OQ_2 and OQ_3. The perpendicular AQ_1 intersects the SMC_1 at point *M*. It means that at output OQ_1, *LMC* is MQ_1. If output increases to OQ_2, *LMC* rises to BQ_2. Similarly, CQ_3 measures the *LMC* at output OQ_3. A curve drawn through points *M*, *B* and *N*, as shown by the *LMC*, represents the behaviour of the marginal cost in the long-run. This curve is known as the long-run marginal cost curve, *LMC*. It shows the trends in the marginal cost in response to the changes in the scale of production.

Some important inferences may be drawn from Fig. 9.7. The *LMC* must be equal to *SMC* for the output at which the corresponding *SAC* is tangent to the *LAC*. At the point of tangency, *LAC* = *SAC*. Another important point to notice is that *LMC* intersects *LAC* when the latter is at its minimum, i.e., point *B*. There is one and only one short-run plant size whose minimum *SAC* coincides with the minimum *LAC*. This point is *B* where

$$SAC_2 = SMC_2 = LAC = LMC$$
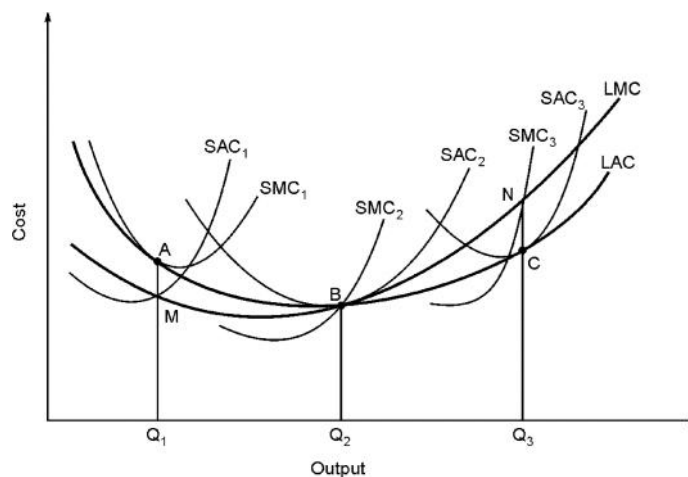
Fig. 9.7 Derivation of LMC

Conceptually, the optimum size of a firm is one which ensures the most efficient utilization of resources. Practically, the optimum size of the firm is one which minimizes the *LAC*. Given the state of technology over time, there is technically a unique size of the firm and level of output associated with the least-cost concept. In Fig. 9.7, the optimum size consists of two plants which produce OQ_2 units of a product at minimum long-run average cost (*LAC*) of BQ_2. The downtrend in the *LAC* indicates that until output reaches the level of OQ_2, the firm is of less than optimal size. Similarly, expansion of the firm beyond production capacity OQ_2, causes a rise in *SMC* and, therefore, in *LAC*. It follows that given the technology, a firm aiming to minimize its average cost over time must choose a plant which gives minimum *LAC* where $SAC = SMC = LAC = LMC$. This size of plant assures the most efficient utilization of the resource. Any change in output level—increase or decrease—will make the firm enter the area of inoptimality.

Check Your Progress

1. State the kinds of cost functions.
2. Define total cost.

9.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. There are two kinds of cost functions: (i) short-run cost functions (ii) long-run cost functions.
2. Total cost is defined as the actual cost that must be incurred to produce a given quantity of output.

9.4 SUMMARY

- Cost function is a symbolic statement of the technological relationship between cost and output.
- The theory of cost deals with the behaviour of cost in relation to a change in output. In other words, the cost theory deals with cost-output relations.
- Depending on whether cost analysis pertains to short-run or to long run, there are two kinds of cost functions: (*i*) short-run cost functions, and (*ii*) long-run cost functions,
- The basic analytical cost concepts used in the analysis of cost behaviour are Total, Average and Marginal costs.
- The cost-output relations are determined by the cost function and are exhibited through cost curves. The shape of the cost curves depends on the nature of the cost function.

- The concept of marginal cost (*MC*) is useful particularly in economic analysis.

- By definition, long-run is a period in which all the inputs–specifically, labour and capital–become variable. The variability of inputs is based on the assumption that in the long-run supply of all the inputs, including those held constant in the short run, becomes elastic.

9.5 KEY WORDS

- **Marginal cost:** It is the additional cost incurred for the production of an additional unit of output.

- **Average cost:** It is a cost accounting term that is sometimes referred to as unit cost or weighted average cost.

9.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. State the formula for cubic cost function.
2. Write a short note on output optimization in the short-run.

Long Answer Questions

1. Explain linear cost function with the help of diagrams.
2. Discuss the relationship between cost curves and the law of diminishing returns.

9.7 FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.

UNIT 10 THE MODERN THEORIES OF COSTS

Structure

10.0 INTRODUCTION

In the previous unit, you studied in detail the traditional theory of cost. This unit will introduce you to the modern theory of cost. Modern economists are of the opinion that whenever a firm installs a plant this plant has some more capacity or reserve capacity to produce in addition to a specific level of output. Thus, the modern cost curves are based upon the idea of reserve capacity.

10.1 OBJECTIVES

After going through this unit, you will be able to:

- State the differences between the traditional and modern theory of cost

- Explain the modern approach to short-run cost behaviour

- Discuss the modern approach to long-run cost behaviour

10.2 MODERN THEORY OF COST: AN OVERVIEW

Some economists, especially George Stigler, have questioned theoretically as well as empirically the U-shaped cost curves of 'the traditional theory of cost' and have attempted to establish that the shape of the cost curves, at least in the long run, is L-shaped. However, this point of view does not appear to have received a general recognition by the economists or as much attention as the traditional theory of cost, at least in the context of pricing theory. One possible reason is that

the traditional theory of cost has a greater application to the theory of price determination and has a greater predicting power than the 'modern theory'. However, this section provides a brief description of the 'modern approach' to the theory of cost. Incidentally, like traditional theory of cost, modern theory too analyses cost-output relationships in the short-run and long-run framework.

10.2.1 Modern Approach to Short-run Cost Behaviour

Like traditional theory of cost, modem theory recognizes that in the short-run,

$$TC = TFC + TVC$$

and $$AC = AFC + AVC$$

In traditional as well as in modem theory of cost, *TFC* includes the following elements of costs:

(*i*) the salaries of administrative staff and related expenses;

(*ii*) the salaries of direct production labour paid on fixed-term basis;

(*iii*) standard depreciation allowance; and

(*iv*) maintenance cost of land and building.

This point onwards, the modem theory deviates from the traditional theory. *Traditional theory* assumes optimum capacity of a plant to be technically given (where *SAC* in minimum) and a cost-minimizing firm has no choice but to utilize the plant to its optimum capacity. On the other hand, *modern theory of cost* emphasizes that firms, in their production planning, choose a plant with flexible capacity, i.e., a plant with **built-in reserve capacity.** According to the modern theory, firms want to have some *reserve capacity,* as a matter of *planning* for the following reasons:

(*i*) to meet the 'seasonal' and eventual increase in demand;

(*ii*) to avoid loss of production due to break-down and repair works;

(*iii*) to have provision for meeting anticipated growth in demand;

(*iv*) to take the advantage of technology providing built-in reserve capacity;

(*v*) to build excess capacity in land and building for expansion, if required; and

(*vi*) to make full utilization of excess 'organizational and administrative' capacity.

Under these conditions, a firm does not necessarily choose a plant that gives the lowest cost of production. Instead, it chooses a plant (a set of machinery) that gives 'maximum flexibility' in production with minor adjustment in technique. For example, let us suppose that the firm has the option of setting up a plant which has an *absolute limit* to produce a commodity at the minimum cost. This absolute limit is shown by the quantity OQ_2 in Fig. 10.1. Note that if the firm chooses this plant, it can produce a maximum quantity of OQ_2 at the minimum *AFC* (average

fixed cost) as shown in the figure by the boundary line BQ_2. Since there is no excess capacity, the firm cannot produce any quantity beyond OQ_2 even if demand increases and hence the firm will not be able to take the advantage of rising demand for its product. Therefore, the firm chooses a flexible plant capable of producing more than OQ_2 with minor adjustment or alternation in the production technique. For example, let us suppose that the firm chooses a flexible plant with absolute limit of output OQ_1 as shown by the boundary line AQ_1. Now let the firm anticipate a rise in demand for its product and add a small-unit machinery to its flexible plant at the output level OQ_1. As Fig 10.1 shows, with the addition of a small-unit machinery, firm' AFC increases from point c to point a on the boundary line AQ_1. But what is important from the firm's point of view is that the firm can increase its production beyond OQ_2 to meet the anticipated increase in demand. Though its AFC increases initially, it declines as production increases, as shown by the curve ab and goes below the limit set by the inflexible plant and the firm is a gainer.
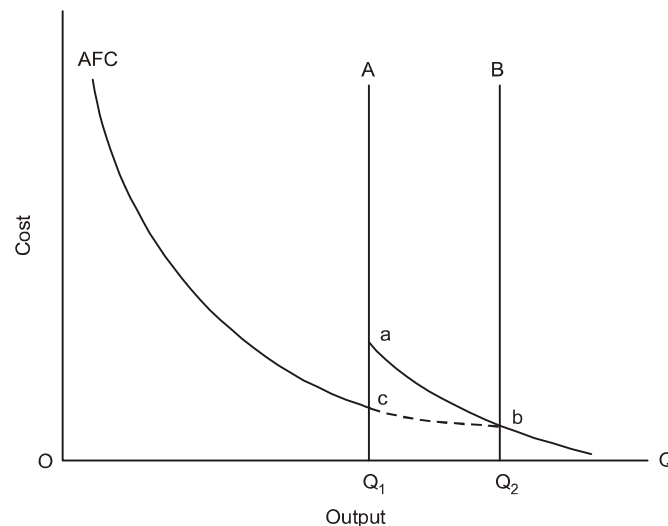
Fig. 10.1 The Built-in Reserve Capacity and AFC Curve

What Happens to the Average Variable Cost? The *average variable cost,* as in traditional theory, includes *average cost* of (*a*) direct labour, (*b*) raw materials, and (*c*) running cost of machinery. There is however a difference between the *short-run average variable cost* (*SAVC*) curves of the traditional and modem cost theories. While in traditional theory, the *SAVC* curve is U-shaped, in modem theory, it is saucer-shaped or bowl-shaped. Part (*a*) of Fig. 10.2 shows *SAVC* curve of the traditional theory, and part (*b*) shows *SAVC* curve of the modern theory.
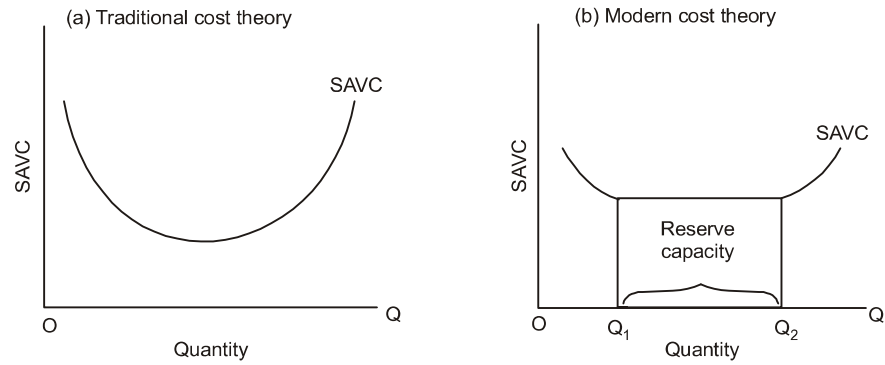
Fig. 10.2 The Traditional and Modern SAVC Curves

As part (*b*) of Fig. 10.2 shows, according to the modem theory cost, the *SAVC* remains constant over a long stretch of output between OQ_1 and OQ_2. The constancy of *SAVC* in the modem theory is attributed to the *built-in reserve capacity* of the flexible plant. The utilization of the built-in reserve pacity keeps the *SAVC* constant. This is an 'innovative' aspect of the modern theory of cost. In the traditional theory, there is no such built-in reserve capacity and therefore *SAVC* begins to rise once the technically efficient level of output is reached.

SAVC and SMC curves: A more important aspect of the modem theory of cost is the nature of and relationship between the *SAVC* and the *SMC* curves. The derivation of *SAVC* and the *short-run marginal cost* (*SMC*) curves is illustrated in Fig. 10.3. The *SAVC* curve is the same as in Fig. 10.2. The *SMC* curve follows the pattern of the traditional theory. The *SMC* decreases with increase in output upto a certain level.
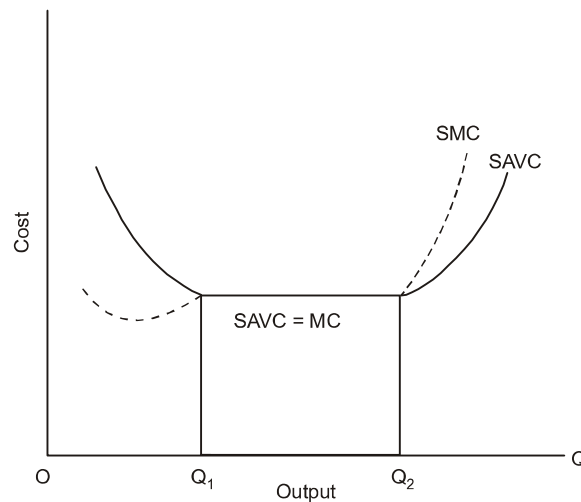
Fig. 10.3 Modern SAVC and SMC Curves

This behaviour of *SMC* curve is shown in Fig. 10.3 till the output OQ_1. However, in the range of output, between OQ_1 and OQ_2, the *SAVC* is constant. It is therefore equal to *SMC*. We know that when *SMC* begins to rise, it rises

faster than *SAVC.* This behaviour of *SMC* is shown begin at output OQ_2 and continue beyond. Beyond output OQ_2, the *SMC* begins to rise and it rises faster than the *SAVC* as is the case in the traditional-theory.

Short-run Average Cost (SAC) Curves: As in traditional theory, in modem theory of cost, $SAC = AFC + SAVC$. The *AFC* includes *normal profit.* Derivation of the *SAC* curve in the modem theory is illustrated in Fig. 10.4. The *SAVC* curve (and also the *SMC* curve) is similar to one given in Fig. 10.3. For the derivation of the *SAC* curve, the *AFC* curve is added to Fig. 10.4. The *SAC* curve is the vertical summation of the *SAVC* and *AFC* curves.

As Fig. 10.4 shows, *AFC* falls continuously whereas *SAVC* decreases till output OQ_1 and remains constant between output OQ_1 and OQ_2. Therefore, a vertical summation of *AFC* and *SAVC* curves gives the *SAC* curve which declines continuously till output OQ_2. Thus, in modem theory of cost, *SAC* decreases until the *built-in reserve capacity* is fully exhausted. The *reserve capacity* is exhausted at output OQ_2.
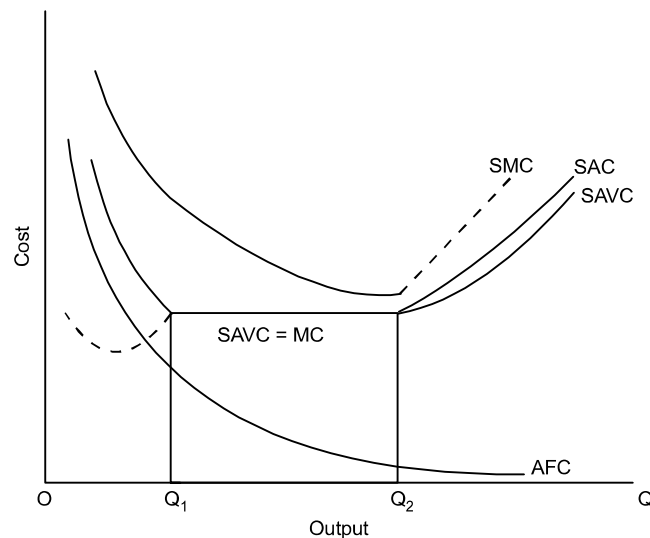
Fig. 10.4 Derivation of the Modern SAC Curve

Beyond output OQ_2, therefore, *SAC* begins to increase and goes on increasing following the increase in *SAVC* while decreasing *AFC* loses its significance.

10.2.2 Modern Approach to Long-run Cost Behaviour: The L-shaped Scale Curve

In respect of long-run cost behaviour, the modern theory of cost distinguishes between *production costs* and *managerial costs.* Both these costs are variable in the long run. The behaviour of these costs determines the shape of the long-run average cost curve (*LAC*). According to the modern theory, the long-run *LAC* is broadly L-shaped. Let us now look at the behaviour of the *production* and *managerial costs* in the long run and how they determine the shape of the *LAC* curve.

Production Cost Behaviour. Production cost decreases steeply in the beginning with the increase in the scale of production but the rate of decrease slows down as the scale increases beyond a certain level of production. The decrease in the production costs is caused by the technical economies which taper off when the scale of production reaches its *technical optimum scale.* Nevertheless, some economies of scale are always available to the expanding firms due to (*i*) 'decentralization and improvement in skills'; and (*ii*) decreasing cost of repairs per unit of output. In addition, in case of multi-product firms producing some of their raw materials and equipments have economies in material cost compared to purchases made from outside.

Managerial Cost Behaviour. The modern theory of cost assumes that, in modern management technology, there is a fixed managerial or administrative set up with a certain scale of production. When the scale of production increases, management set up has to be accordingly expanded. It implies that there is a link between the scale of production and the cost of management. According to the modern theory, the managerial cost first decreases but begins to increase as the scale of production is expanded beyond a certain level.

What Makes LAC L-shaped? *The net effect of decreasing production cost and increasing managerial cost determines the shape of the long-run average cost (LAC).* Recall that production cost continues to decrease, though slowly beyond a certain scale of production and managerial cost too decreases initially but rises later. In the initial stage of production, therefore, the *LAC* decreases very steeply. Beyond a certain scale of production, however, while production cost continues to decline, management cost begins to rise. According to the modern theory of cost, the rise in managerial cost is more than offset by the decrease in the production costs. Therefore, the *LAC* continues to fall but very slowly. *In case the decrease in production cost* is *just sufficient to offset the rise in the managerial cost, the LAC becomes constant.* This makes *LAC* an L-shaped curves.

Derivation of the LAC Curve: The derivation of the *LAC* curves is illustrated in Figs. 10.5 and 10.6. Fig. 10.5 shows the decreasing *LAC* curve. Let us suppose that, given the technology, the optimum scale of production consists of four plants and *SAC* curves from SAC_I to SAC_4 in Fig. 10.5 represent the addition of four plants to the production scale in each period of time. Clower and Due have found that firms use 'normally' only 2/3 to 3/4 of the plant size. This is called 'load factor'. The *load factor* is the "ratio of average actual rate of use to the capacity or best rate of use, and this load factor will generally be smaller than one". The points *A, B, C* and *D* on the *SACs* mark the 'load factor' in case of each plant, respectively—it may be any value between 2/3 and 3/4 of the plant size. By drawing a curve through the 'load-factor' points, we get the *LAC* curve. If there is a larger number of plants, we will get much larger number of 'load factor' points and draw a smooth *LAC* curve as shown in Fig. 10.5.
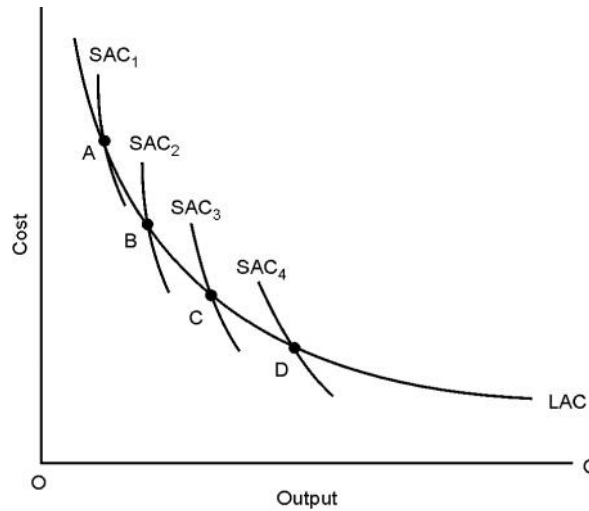
Fig. 10.5 Derivation of LAC Curve in Modern Theory

To compare the *LAC* of the modern and traditional theories of cost, two points need to be noted: (*i*) the *LAC* curve of the modern theory does not show the tendency to turn up even at a very large scale of production whereas the traditional *LAC* curve does turn up; (*ii*) unlike traditional *LAC* forming an *envelope curve,* modern *LAC* intersects the *SACs.*

If scale of production has a *minimum optimal scale of plant,* as shown by output level *OQ* in part (*b*) of Fig. 10.6, all economies of scale are achieved at output *OQ* and the *LAC* becomes constant even if scale of production is expanded. In this case, the *LMC* lies below the *LAC* till the *minimum optimal scale of plant* is reached, as shown in part (*a*) of Fig. 10.6. When the firm operates in the range of no-seale-economies, i.e., beyond output *OQ* in part (*b*) of Fig. 10.6, the *LAC* becomes constant and the *LAC* curve takes the shape of a horizontal line. Both the parts (declining and constant) of the *LAC* curve put together make it roughly L-shaped.
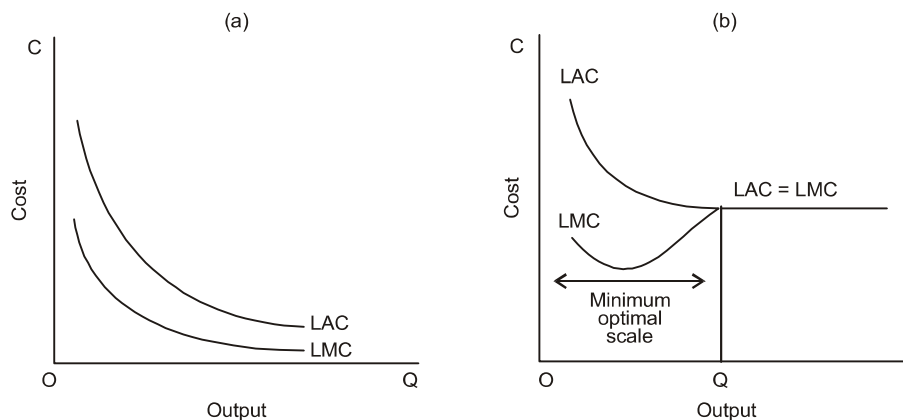
Fig. 10.6 Derivation of the L-shaped LAC Curve

From practical point of view, the modern *LAC* curve is regarded to be more realistic. But from analytical and prediction point of view, the traditional cost curves still hold the ground firmly. In fact, the so called 'modern theory of cost' is a modification of the traditional theory on the basis of empirical data in some manufacturing industries of some countries.

10.2.3 Learning Curve

In many industries, some firm specialize in one particular good or service and continue to produce it over time. As a result, they gain additional knowledge, expertise and efficiency and are able to make technological advances in the field of production. The economists have found that these factors result in continuous decline in their average cost with increase in output. A curve drawn to show the continuous decline in the average cost is called *learning curve.* The *learning curve* is, in fact, the long-run average cost curve showing a continuous decline in average cost with increase in output over time, as shown in Fig. 10.7(a) and (b). The shape of the learning curve depends on the nature of the product. As shown by Mansfield *et. al.* Fig. 10.7(a) shows the learning curve for the production of portable turbine and Fig. 10.7(b) shows the learning curve for production of optical equipment. The continuous decrease in long-run average cost, as shown in Fig. 10.7, results from the continuous technological improvement, on the one hand, and knowledge, expertise and efficiency gained by the firms by producing the same commodity over time, on the other.

It is **important** to note here that average cost is estimated by dividing **cumulative total cost** by **cumulative total output** increasing at constant rate. This implies that when firm continues to produce the same commodity over time, its cumulative cost increases less than proportionate to a constant increase in output.
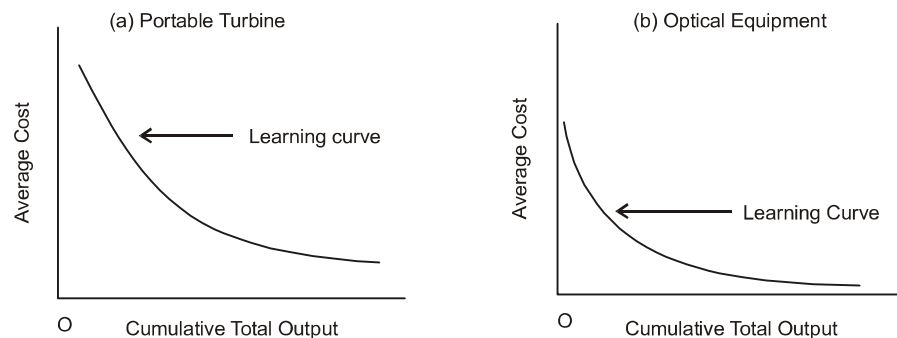
Fig. 10.7 The Learning Curve

The learning curve is generally expressed by the following form of *AC* function.

$$AC = BQ^a$$

where *AC* average cost; *Q* is output; *B* is the cost of the first unit of output; and power *a* is a *negative* constant.

In its double logarithmic form, the *AC* function for learning curve can be expressed as

$$\text{Log } AC = \log B + a \log Q$$

In its logarithmic form of the learning cost function, *a* gives the slope of the learning curve. The numerical values of parameters *B* and *a* are estimated by regression techniques by using historical data of cumulative output and cost.

Many firms in both manufacturing and service sectors have adopted pricing strategy based on learning curve, mainly by the firms manufacturing airplanes, ships, semi-conductor chips, domestic appliances, refined petroleum products and power plants. The learning curve is used to assess and forecast the future requirements for inputs including manpower, raw materials, and machinery.

Check Your Progress

1. State one difference between the traditional and modern theory of cost.

2. What does the average variable cost under the traditional theory of cost consist of?

10.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Traditional theory assumes optimum capacity of a plant to be technically given (where SAC in minimum) and a cost-minimizing firm has no choice but to utilize the plant to its optimum capacity. On the other hand, *modern theory of cost* emphasizes that firms, in their production planning, choose a plant with flexible capacity, i.e., a plant with built-in reserve capacity.

2. The average variable cost, as in traditional theory, includes average cost of (a) direct labour, (b) raw materials, and (c) running cost of machinery.

10.4 SUMMARY

- Some economists, especially George Stigler, have questioned theoretically as well as empirically the U-shaped cost curves of 'the traditional theory of cost' and have attempted to establish that the shape of the cost curves, at least in the long run, is L-shaped.

- The *average variable cost,* as in traditional theory, includes *average cost* of (*a*) direct labour, (*b*) raw materials, and (*c*) running cost of machinery.

- A more important aspect of the modem theory of cost is the nature of and relationship between the *SAVC* and the *SMC* curves.

- In respect of long-run cost behaviour, the modern theory of cost distinguishes between *production costs* and *managerial costs*.

- Production cost decreases steeply in the beginning with the increase in the scale of production but the rate of decrease slows down as the scale increases beyond a certain level of production.
- The net effect of decreasing production cost and increasing managerial cost determines the shape of the long-run average cost (LAC).
- In many industries, some firm specialize in one particular good or service and continue to produce it over time. As a result, they gain additional knowledge, expertise and efficiency and are able to make technological advances in the field of production.
- The learning curve is used to assess and forecast the future requirements for inputs including manpower, raw materials, and machinery.

10.5 KEY WORDS

- **Learning curve:** It is the long-run average cost curve showing a continuous decline in average cost with increase in output over time.
- **Decentralization:** It is the process by which the activities of an organization, particularly those regarding planning and decision-making, are distributed or delegated away from a central, authoritative location or group.

10.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. State the difference between short-run average variable cost curves of the traditional and modern cost theories.
2. Mention the uses of the learning curve in economics.

Long Answer Questions

1. Discuss the 'innovative' aspect of the modern theory of cost.
2. Explain the modern approach to long-run cost behaviour.

10.7 FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.

UNIT 11 ECONOMIES OF SCALE

Structure

11.0 INTRODUCTION

Economies of scale refers to the cost advantages which arises when average cost start decreasing as the output increases. Economies of scale is a concept which applies to a variety of organisational and business situations and at various levels such as business or manufacturing unit, plant or an entire enterprise. The shape of LAC or the Long-run Average Cost function is determined by the economies and diseconomies of scale. Economies of scale give rise to lower per-unit cost for several reasons including specialisation, division of labour, integrated technology, larger advertising boost production volume. In this unit, we will discuss the concept of economies of scale in detail.

11.1 OBJECTIVES

After going through this unit, you should be able to:

- Discuss the concept of economies of scale

- Explain the diseconomies of scale

- Examine the concept of elasticity of cost

11.2 OVERVIEW OF ECONOMIES OF SCALE

Economies of scale refers to the cost advantage experienced by a firm when it increases its level of output. The advantage arises due to the inverse relationship between per-unit fixed cost and the quantity produced by the firm.

While optimization of output in the long run is an important concern of business firms, *cost minimization* is an equally import decision area. Cost of production depends not only on *internal factor* – the productivity of inputs – but also on

many *external factors* – the factors that arise out of the firm. Since we are concerned in this chapter with the theory of cost, in this section we give a detailed analysis of internal and external economies and diseconomies of scale and how they determine the trend in cost of production. To begin with, let us have look at the trend of long-run average cost curve (*LAC*). It is to be noted that *LAC* decreases with the expansion of production scale up to *OQ*2 and then it begins to rise. Decrease in *LAC* is caused by the economies of scale and increase in *LAC* is caused by diseconomies of scale. Economies of scale result in cost saving and diseconomies lead to rise in cost. *Economies and diseconomies of scale* determine also the returns to scale. Increasing returns to scale operate till economies of scale are greater than the diseconomies of scale, and returns to scale decrease when diseconomies are greater than the economies of scale. When economies and diseconomies are in balance, returns to scale are constant. In this section, we briefly discuss the various kinds of economies and diseconomies of scale and their effect on cost of production.
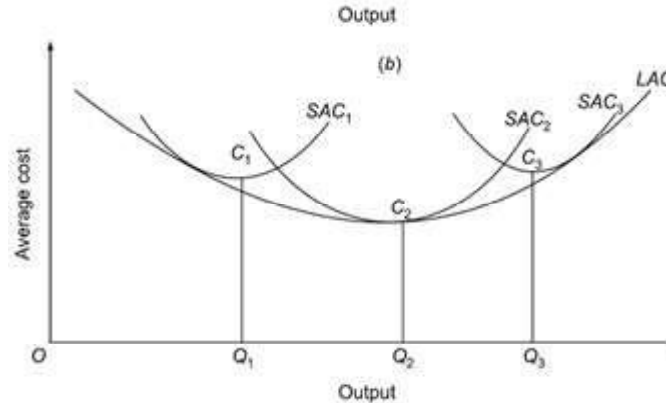
Fig. 11.1 *Long-run Total and Average Cost Curves*

Economies of Scale

The economies of scale are classified as (*a*) Internal or Real Economies, and (*b*) External or Pecuniary Economies.

A. Internal Economies: Internal economies, also called 'real economies', are those that arise within the firm with addition of new production plants. This means that internal economies are available exclusively to the expanding firm. Internal economies may be classified under the following categories.

   (*i*) Economies in production;

  (*ii*) Economies in marketing;

 (*iii*) Managerial economies, and

 (*iv*) Economies in transport and storage.

  (i) Economies in Production: Economies in production arise from two sources: (*a*) technological advantages, and (*b*) advantages of division of labour based on specialization and skill of labour.

Technological advantages: Large-scale production provides an opportunity to the expanding firms to avail the advantages of technological advances. Modern technology is highly specialized. The advanced technology makes it possible to conceive the whole process of production of a commodity in one composite unit of production. For example, production of cloth in a textile mill may comprise such plants as (*i*) spinning; (*ii*) weaving; (*iii*) printing and pressing; and (*iv*) packing, etc. Likewise, a composite dairy scheme may consist of plants like (*i*) chilling; (*ii*) milk processing; and (*iii*) bottling. Under small-scale production, the firm may not find it economical to have all the plants under one roof. It would, therefore, not be in a position to take the full advantage of a composite technology. But, when scale of production expands and firms hire more capital and labour, their total output increases more than proportionately till the optimum size of the firm is reached. It results in lower cost of production.

Advantages of division of labour and specialization: When a firm's scale of production expands, more and more workers of varying skills and qualifications are employed. With the employment of larger number of workers, it becomes increasingly possible to divide the labour according to their qualifications, knowledge, experience, expertise and skills and to assign them the function to which they are best suited. This is known as division of labour. *Division of labour* leads to a greater specialization of manpower. It increases productivity of labour and, thereby, reduces cost of production. Besides, specialized workers develop more efficient tools and techniques and gain speed of work. These advantages of division of labour improve productivity of labour per unit of labour cost and time. Increase in labour productivity decreases to per unit cost of production.

(ii) Economies in Purchase of Inputs: Economies in input purchases arise from the large-scale purchase of raw materials and other material inputs and large-scale selling of the firm's own products. As to economies in the purchase of inputs, the large-size firms normally make bulk purchases of their inputs. The large scale purchase entitles the firm for certain discounts in input prices and other concessions that are not available on small purchases. As such, the growing firms gain economies on the cost of their material inputs. The internal economies arise also in marketing the firm's own product as (*a*) economies in advertisement cost; (*b*) economies in large-scale distribution through wholesalers, etc.; and (*c*) other large-sale economies. With the expansion of the firm, the total production increases. But the expenditure on advertising the product does not increase proportionately. Similarly, selling through the wholesale dealers reduces the cost of distribution of the firm's production. The firm also gains on large scale distribution through better utilization of 'sales force, distribution of sample, etc.'

(iii) **Managerial Economies:** Managerial economies arise from (*a*) specialization in managerial activities, i.e., the use of specialized managerial personnel, and (*b*) systemization of managerial functions. For a large-size firm, it becomes possible to divide its management into specialized departments under specialized personnel, such as production manager, sales manager, HR manager, financial manager, etc. The management of different departments by speciallized managers increases the efficiency of management at all the levels of management because of the decentralization of decision-making. It increases production, given the cost. Large-scale firms have the opportunity to use advanced techniques of communication, telephones and telex machines, computers, and their own means of transport. All these lead to quick decision-making, help in saving valuable time of the management and, thereby, improve the managerial efficiency. For these reasons, managerial cost increases less than proportionately with the increase in production scale upto a certain level, of course.

(iv) **Economies in Transport and Storage:** Economies in transportation and storage costs arise from fuller utilization of transport and storage facilities. Transportation costs are incurred both on production and sales sides. Similarly, storage costs are incurred on both raw materials and finished products. The large-size firms may acquire their own means of transport and they can, thereby, reduce the unit cost of transportation, at least to the extent of profit margin of the transport companies. Besides, own transport facility prevents delays in transporting goods. Some large-scale firms have their own railway tracks from the nearest railway point to the factory, and thus they reduce the cost of transporting goods in and out. For example, Bombay Port Trust has its own railway tracks, oil companies have their own fleet of tankers. Similarly, large-scale firms can create their own godowns in various centres of product distribution and can save on cost of storage.

B. External or Pecuniary Economies of Scale

External economies are those that arise outside the firm and accrue to the expanding firms. External economies appear in the form of money saving on inputs, called *pecuniary economies*. Pecuniary economies accrue to the large-size firms in the form of discounts and concessions on (*i*) large scale purchase of raw material, (*ii*) large scale acquisition of external finance, particularly from the commercial banks; (*iii*) massive advertisement campaigns; (*iv*) large scale hiring of means of transport and warehouses, etc. These benefits are available to all the firms of an industry but large scale firms benefit more than small firms. Besides, expansion of an industry encourages the growth of ancillary industries that supply inputs. In the initial stages, such industries also enjoy the increasing returns to scale. In a competitive market, therefore, input prices go down. The benefit of decreasing input prices accrues to the expanding firms in addition to discounts and concessions. For example, growth of the automobile industry helps the development of tyre industry and other motor

parts manufacturing units. The economies of scale reaped by such industries flow also to automobile industry. If Maruti Udyog Limited starts producing tyres for its own cars and ancillaries, cost of Maruti cars may go up. Consider another example: growth of computer industry encourages growth of firms that manufacture and supply computer chips and other software. Competition between such firms and law of increasing returns reduces the cost of inputs. Reduction in input costs is an important aspect of external economies.

11.2.1 Diseconomies of Scale

The economics of scale have their own limits, i.e., scale economies exist only up to a certain level of production scale. The expansion of scale of production beyond that limit creates condition for *diseconomies of scale.* Diseconomies of scale are disadvantages that arise due to the expansion of production scale beyond its optimum level and lead to rise in the cost of production. Like economies, diseconomies may be internal and external. Let us describe the nature of **internal** and **external diseconomies** in some detail.

1. **Internal Diseconomies:** Internal diseconomies are those that are exclusive and internal to a firm as they arise within the firm. Like everything else, economies of scale have a limit too. This limit is reached when the advantages of division of labour and managerial staff have been fully exploited; excess capacity of plant, warehouses, transport and communication systems, etc., is fully used; and economy in advertisement cost tapers off. Although some economies may still exist, diseconomies begin to outweigh the economies and the costs begin to rise.

   Managerial Inefficiency: Diseconomies begin to appear first at the management level. Managerial inefficiencies arise, among other things, from the expansion of scale itself. With fast expansion of the production scale, personal contacts and communications between (*i*) owners and managers, (*ii*) managers and labour, and (*iii*) between the managers of different departments or sections get rapidly reduced. The lack of fast or quick communication causes delays in decision-making affecting production adversely. *Secondly*, close control and supervision is replaced by remote control management. With the increase in managerial personnel, decision-making becomes complex and delays in decision-making become inevitable. *Thirdly*, implementation of whatever decisions are taken is delayed due to coordination problem in large scale organisations. *Finally*, with the expansion of the scale of production, management is professionalized beyond a point. As a result, the owner's objective function of profit maximization is gradually replaced by managers' utility function, like job security and high salary, standard or reasonable profit target, satisfying functions. All these lead to laxity in management and, hence to a rise in the cost of production.

   Labour Inefficiency: Increasing number of labour leads to a loss of control over labour management. This affects labour productivity adversely. Besides, increase in the number of workers encourages labour union activities that cause loss of output per unit of time and hence, rise in the cost of production.

2. External Diseconomies: External diseconomies are the disadvantages that arise outside the firm, especially in the input markets, due to natural constraints, especially in agriculture and extractive industries. With the expansion of the firm, particularly when all the firms of the industry are expanding, the discounts and concessions that are available on bulk purchases of inputs and concessional finance come to an end. More than that, increasing demand for inputs puts pressure on the input markets and input prices begin to rise causing a rise in the cost of production. These are *pecuniary diseconomies*. On the production side, the law of diminishing returns to scale come into force due to excessive use of fixed factors, more so in agriculture and extractive industries. For example, excessive use of cultivable land turns it into barren land; pumping out water on a large scale for irrigation causes the water table to go down resulting in rise in cost of irrigation; extraction of minerals on a large scale exhausts the mineral deposits on upper levels and mining further deep causes rise in cost of production; extensive fishing reduces the availability of fish and the catch, even when fishing boats and nets are increased. These kinds of diseconomies make the *LAC* move upward.

Check Your Progress

1. Mention the classification of economies of scale.

2. What are internal economies?

11.3 PRICE ELASTICITY OF COST

In unit 8 of this book, we have discussed the basic cost concepts related to elasticity of cost.

Elasticity of Cost

If output (Q) is produced at a total cost (T), the cost function is T=f (0.) The elasticity of total cost is the ratio of the proportional change in total cost to the proportional change in output. It may be explained as-
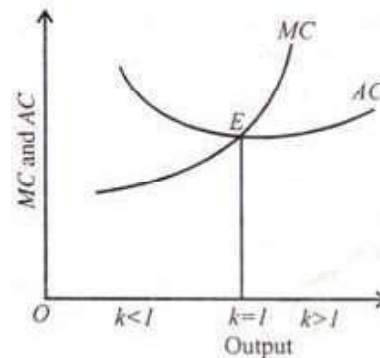
Fig. 11.1 *Elasticity of Cost*

Thus, cost elasticity (κ) is equal to the ratio of marginal cost (dT/dQ) to average cost (T/Q). It follows from this that if $MC_{\leq}^{\geq} AC$, $\kappa_{\leq}^{\geq} 1$. It means that when MOAC, $\kappa > 1$. Diagrammatically, when the MC curve is rising and is above the AC curve, $\kappa > 1$, as shown by the area right to point E in the Figure 11.1.

It is the case of decreasing returns. When $MC = AC$, $\kappa = 1$, it is the point E where the MC curve cuts the AC curve from below in the figure. It is the case of constant returns. When $MC < AC$, $\kappa < 1$, shown as the area to the left of point E in the figure, where the MC curve is falling and is below the AC curve. It is the case of increasing returns.

Since the average cost and the marginal cost are derived from the total cost in relation to the output, the shapes of the AC curve and the MC curve can also be checked from the shape of the total cost curve. If P is the point on the total cost curve at a given output Q, then the average cost is to be read off as the gradient of OP and the marginal cost as the tangent at P. This is shown in Figure 11.2.
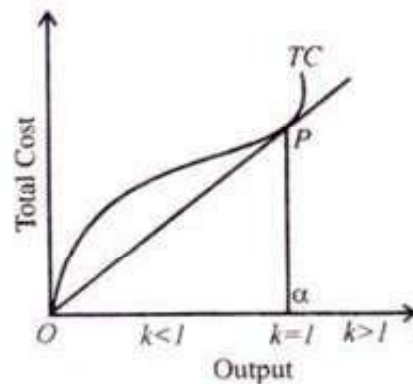
Fig. 11.2 *Average Cost Curve and Marginal Cost Curve from Total Cost cruve*

The figure, further, reveals that the elasticity of total cost increases continuously with increases in output from less than unity to greater than unity. At first, cost elasticity is less than unity for small outputs, and finally, it is greater than unity for large outputs. In other words, if we take $\kappa = 1$ at some definite level of output, $Q = \alpha$, then $\kappa < 1$ for outputs $Q < \alpha$, and $\kappa > 1$ for outputs $Q > \alpha$. This is illustrated in Figure 11.2.

Elasticity of Average Cost

The elasticity of total cost is given by

$$k \ or \ E \ (T) = \frac{dT}{dQ} \cdot \frac{Q}{T} \ and \ average \ cost \ is \ {}^{T}/_{Q}.$$

Therefore replacing T by ${}^{T}/_{Q}$

$$= E({}^{T}/_{Q}) = \frac{d({}^{T}/_{Q})}{dQ} \cdot \frac{Q}{{}^{T}/_{Q}}$$

$$= \frac{d}{dQ} \cdot \left({}^{T}/_{Q}\right) \cdot \frac{Q^2}{T}$$

$$\frac{Q^2}{T} \cdot \left[Q\,\dfrac{\dfrac{dT - T}{dQ}}{- - - -} \right]$$

$$\frac{Q^2}{T} \cdot \frac{1}{Q^2} \left[Q\,\frac{dT - T}{dQ} \right]$$

$$\frac{Q}{T} \cdot \frac{dT}{dQ} - 1 = k - 1$$

The following results follow from this:

(1) If A: (the elasticity of total cost) is greater than, equal to or less than unity, the elasticity of average cost is greater than, equal to or less than zero-, and

(2) The elasticity of total cost exceeds the elasticity of average cost by unity, i.e., E (T/Q) = κ-1 or κ-E (T/Q) =1.

Elasticity of Marginal Cost

As we know, the elasticity of total cost is given by E (T) = dT/dQ. Q/T. Therefore, the marginal cost is (dT/dQ). Replacing T by dT/dQ.

$$E\left|\frac{dT}{dQ}\right| = d\,\frac{dT/dQ}{dQ} \cdot \frac{Q}{dT/dQ}$$

$$= \frac{d}{dQ}\left(dT/dQ\right) \cdot \frac{Q}{dT/dQ} \qquad \qquad ...(11.1)$$

Since k is given by,

$$\mathbf{K} = \square_{\square} \cdot {}^{\square} \; \square\square \; \llcorner \; {}^{\square}\square_{\square\square}{}^{\square}\square_{\square} \qquad \qquad ...(11.2)$$

Substituting the value of (11.2) in (11.1), we get

$$E\left|\frac{dT}{dQ}\right| = \frac{d}{dQ}\left|\frac{dT}{dQ}\right|\frac{Q^2}{T}$$

Elasticity of Productivity

The output of a firm is determined by the various inputs used by it. Assuming all inputs (λ) are used in fixed proportions to produce output Q, then Q =f (λ). The elasticity of productivity is defined as the ratio of proportional change in output to the proportional change in inputs. It can be written as

$$\in = \frac{dQ}{Q} \div \frac{d\lambda}{\lambda} = \frac{dQ}{d\lambda} \cdot \frac{\lambda}{Q}$$

Elasticity of productivity helps in understanding the nature of production function in economic theory. If $\in > 1$, it is the case of increasing returns because a small proportionate increase in inputs leads to more than proportionate increase in output.

If $\in \, < 1$, it is the case of decreasing returns because a small proportionate increase in inputs leads to a less than proportionate increase in output. If $\in \, = 1$, there are constant returns. Since the concept of elasticity of productivity is based on the assumption of fixed proportions of inputs used in production, total costs become proportional to the inputs employed.

Therefore, e is the inverse of the elasticity of total cost. But the basic difference between the elasticity of productivity and the elasticity of total cost arises from the fact that the inputs are used in fixed proportions only in the case of ε and not for k.

Check Your Progress

3. What is the elasticity of average cost when the elasticity of total cost is greater than, equal to or less than unity?

11.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The economies of scale are classified as (*a*) Internal or Real Economies, and (*b*) External or Pecuniary Economies.
2. Internal economies, also called 'real economies', are those that arise within the firm with addition of new production plants. This means that internal economies are available exclusively to the expanding firm.
3. If the elasticity of total cost is greater than, equal to or less than unity, the elasticity of average cost is greater than, equal to or less than zero.

11.5 SUMMARY

- Economies of scale refers to the cost advantage experienced by a firm when it increases its level of output. The advantage arises due to the inverse relationship between per-unit fixed cost and the quantity produced by the firm.
- Cost of production depends not only on *internal factor* – the productivity of inputs – but also on many *external factors* – the factors that arise out of the firm.
- Diseconomies of scale are disadvantages that arise due to expansion of production scale and lead to rise in the cost of production.
- The elasticity of total cost is the ratio of the proportional change in total cost to the proportional change in output.
- Since the average cost and the marginal cost are derived from the total cost in relation to the output, the shapes of the AC curve and the MC curve can also be checked from the shape of the total cost curve.

- The output of a firm is determined by the various inputs used by it. Assuming all inputs (λ) are used in fixed proportions to produce output Q, then Q =f (λ). The elasticity of productivity is defined as the ratio of proportional change in output to the proportional change in inputs.

11.6 KEY WORDS

- **Economies of scale:** It refers to the cost advantage experienced by a firm when it increases its level of output.

- **Diseconomies of scale:** It refers to the disadvantages that arise due to expansion of production scale and lead to rise in the cost of production.

- **Elasticity of total cost:** It is the ratio of the proportional change in total cost to the proportional change in output.

- **Elasticity of productivity:** It is defined as the ratio of proportional change in output to the proportional change in inputs.

11.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. Write a brief note on internal economies.
2. What are external economies?

Long Answer Questions

1. Compare and contrast internal economies and external economies.
2. Explain the concept of elasticity of cost.

11.8 FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.

BLOCK - IV
PRICE AND OUTPUT DETERMINATION UNDER
DIFFERENT MARKET STRUCTURES

UNIT 12 PRICE AND OUTPUT DETERMINATION UNDER PERFECT COMPETITION

Structure

12.0 INTRODUCTION

In the previous unit, you studied about economies of scale, its importance and elasticity of costs. In this unit, you will study about theory of price and output determination under perfect competition. For this purpose, two basic points need to be noted at the outset. One, the main consideration behind the determination of price and output is to achieve the objective of the firm. Two, although there can be various business objectives, traditional theory of price and output determination is based on the assumption that all firms have only one and the same objective to achieve, i.e., profit maximization. These two basic conditions need to be borne in mind throughout this and the next subsequent units.

12.1 OBJECTIVES

After going this unit, you will be able to:

- List the characteristics of perfect competition

• Explain the derivation of firm's supply curve with diagrams

• Define constant cost industry

12.2 PERFECT COMPETITION: AN OVERVIEW

Perfect competition refers to a market condition in which a very large number of buyers and sellers enjoy full freedom to buy and to sell a homogenous good and service and they have perfect knowledge about the market conditions, and factors of production have full freedom of mobility. Although this kind of market situation is a rare phenomenon, it can be located in local vegetable and fruit markets. Another area which was often considered to be perfectly competitive is the stock market. However, stock market are controlled and regulated in India and a few big market players influence the market conditions in a serious and dangerous way. Therefore, stock market in India is not perfectly competitive. In the following section, the main features of perfect competition are discussed in detail.

12.2.1 Characteristics of Perfect Competition

The following are the main features or characteristics of a perfectly competitive market.

(i) Large number of buyers and sellers. Under perfect competition, the number of sellers is assumed to be so large that the share of each seller in the total supply of a product is very small or insignificant. Therefore, no single seller can influence the market price by changing his supply or can charge a higher price. Therefore, *firms are price-takers, not price-makers.* Similarly, the number of buyers is so large that the share of each buyer in the total demand is very small and that no single buyer or a group of buyers can influence the market price by changing their individual or group demand for a product.

(ii) Homogeneous product. The goods and services supplied by all the firms of an industry are assumed to be homogeneous or almost identical. Homogeneity of the product implies that buyers do not distinguish between product supplied by the various firms of an industry. Product of each firm is regarded as a perfect substitute for the products of other firms. Therefore, no firm can gain any competitive advantage over the other firms. This assumption eliminates the power of all the firms to charge a price higher than the market price.

(iii) Perfect mobility of factors of production. Another important characteristic of perfect competition is that the factors of production are freely mobile between the firms. Labour can freely move from one firm to another or from one occupation to another, as there is no barrier to labour mobility—legal, language, climate, skill, distance or otherwise. There is no trade union. Similarly, capital can also move freely from one firm to another. No firm has any kind of monopoly over any industrial input. This assumption guarantees that factors of production—land, labour, capital, and entrepreneurship—can enter or quit a firm or the industry at will.

(iv) Free entry and free exist. There is no legal or market barrier on the entry of new firms to the industry. Nor is there any restriction on the exit of the firms from the industry. A firm may enter the industry or quit it at its will. Therefore, when firms in the industry make supernormal profit for some reason, new firms enter the industry. Similarly, when firms begin to make losses or more profitable opportunities are available elsewhere, firms are free to leave the industry.

(v) Perfect knowledge. Both buyers and sellers have perfect knowledge about the market conditions. It means that all the buyers and sellers have *full information* regarding the prevailing and future prices and availability of the commodity. As Marshall put it, " ... though everyone acts for himself, his knowledge of what others are doing is supposed to be generally sufficient to prevent him from taking a lower or paying a higher price than others are doing." Information regarding market conditions is available free of cost. There is *no uncertainty in the market.*

(vi) No government interference. Government does not interfere in any way with the functioning of the market. There are no discriminatory taxes or subsidies; no licencing system, no allocation of inputs by the government, or any other kind of direct or indirect control. That is, the government follows the free enterprise policy. Where there is intervention by the government, it is intended to correct the market imperfections if there are any.

(vii) Absence of collusion and independent decision-making by firms. Perfect competition assumes that there is no collusion between the firms, i.e., they are not in league with one another in the form of guild or cartel. Nor are the buyers in any kind of collusion between themselves. There are no consumers' associations, etc. This condition implies that buyers and sellers take their decisions independently and they act independently.

Perfect vs. Pure Competition: Sometimes, a distinction is made between *perfect competition* and *pure competition.* The differences between the two is a matter of degree. While 'Perfect Competition' has all the features mentioned above, under 'Pure Competition', there is *no perfect mobility of factors* and *perfect knowledge* about market-conditions. That is, *perfect competition less 'perfect mobility' and 'perfect knowledge' is pure competition.* 'Pure competition' is 'pure' in the sense that it has absolutely no element of monopoly.

The perfect competition, with characteristics mentioned above is considered as a rare phenomenon in the real business world. The actual markets that approximate to the conditions of a perfectly competitive market include markets for stocks and bonds, and agricultural market *(mandis).* Despite its limited scope, perfect competition model has been widely used in economic theories due to its analytical value.

12.2.2 Price Determination Under Perfect Competition

Under perfect competition, an individual firm *does not* determine the price of its product. Price for its product is determined by the market demand and market supply. The determination of market price has been reproduced in Fig. 12.1 (a).

The demand curve, *DD',* represents the market demand for the commodity of an industry as a whole. Likewise, the supply curve, *SS',* represents the total supply created by all the firms of the industry (derivation of industry's supply curve has been shown in a following section). As Fig. 12.1 (a) shows, market price for the industry as a whole is determined at *OP.* This price is given for all the firms of the industry. No firm has power to change this price. At this price, a firm can sell any quantity. It implies that the demand curve for an individual firm is a straight horizontal line, as shown by the line *dd'* in Fig. 12.1 (b), with infinite elasticity.
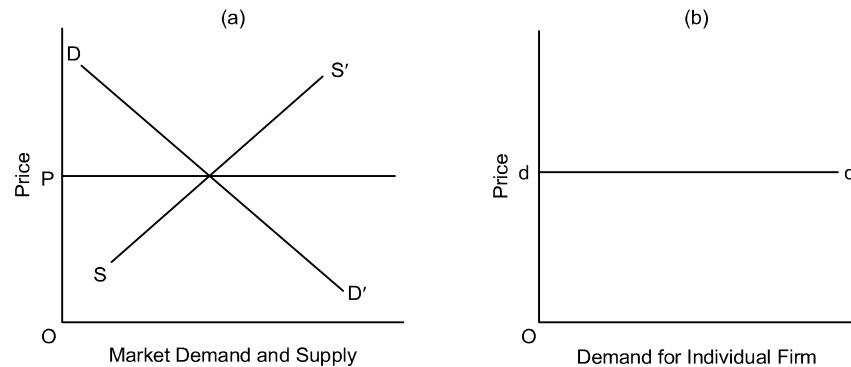
Fig. 12.1 *Determination of Market Price and Demand for Individual Firms*

No control over cost. Because of its small purchase of inputs (labour and capital), a firm has no control over input prices. Nor can it influence the technology. Therefore, cost function for an individual firm is given. This point is, however, not specific to firms in a perfectly competitive market. This condition applies to all kinds of market except in case of bilateral monopoly.

What Are the Firm's Options? The firm's option and role in a perfectly competitive market are very limited. The firm has no option with respect to price and cost. It has to accept the market price and produce with a given cost function. The only option that a firm has under perfect competition is to produce a quantity that maximises its profits given the price and cost. Under profit maximising assumption, a firm has to produce a quantity which maximises its profit and attains its equilibrium.

Equilibrium of the Firm in Short Run

A profit maximising firm is in equilibrium at the level of output which equates its $MC = MR$. However, the level of output which meets the equilibrium condition for a firm varies depending on cost and revenue functions. The nature of cost and revenue functions depends on whether one is considering *short-run* or *long-run.* While the revenue function is generally assumed to be given in both short and long runs, the short-run cost function is not the same in the short and long-runs. The short-run cost function is different from the long-run cost function because in the short run, some inputs (e.g., capital) are held constant while all factors are variable in the long-run. In this section, we will discuss *firm's short-run equilibrium.* Long-run equilibrium of the firm will be discussed in the forthcoming section.

Assumptions: The short-run equilibrium of a firm is analyses under the following assumptions.

(*a*) capital is fixed but labour is variable;

(*b*) prices of inputs are given;

(*c*) price of the commodity is fixed; and

(*d*) the firm is faced with short-run U-shaped cost curves.

The firm's equilibrium in the short-run is illustrated in Fig. 12.2. Price of a commodity is determined by the market forces—demand and supply—in a perfectly competitive market at *OP.* The firms, therefore, face a straight-line, horizontal demand curve, as shown by the line *P = MR.* The straight horizontal demand line implies that price equals marginal revenue, i.e., *AR = MR.* The short-run average and marginal cost curves are shown by *SAC* and *SMC,* respectively.
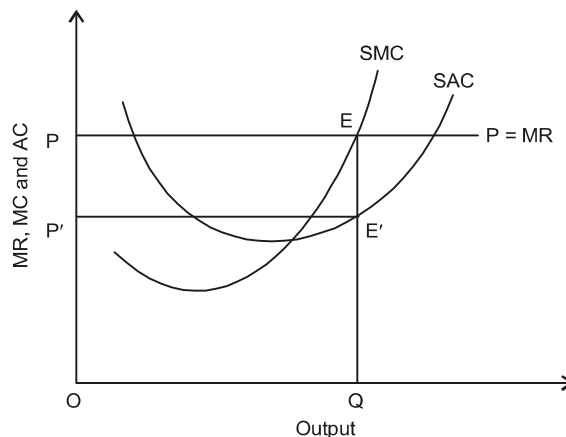
Fig. 12.2 *Short-run Equilibrium of the Firm*

It can be seen in Fig. 12.2 that *SMC* curve intersects the *P = MR* line at point *E*, from below. At point *E*, therefore, *SMC = MR.* A perpendicular drawn from point *E* to the output axis determines the equilibrium output at *OQ.* It can be seen in the figure that output *OQ* meets both the *first* and the *second order conditions* of profit maximisation. At output *OQ,* therefore, profit is maximum. The output *OQ* is, thus, the equilibrium output. At this output, the firm is in equilibrium and is making maximum profit. Firm's total pure profit is shown by the area *PEE′P′* which equals *PP′ × OQ* where *PP′* is the per unit super normal profit at output *OQ.*

Does a Firm Always Make Profit in the Short Run?: Figure 12.2 shows that a firm makes supernormal profit in the shortrun. A question arises here: Does a firm make always a supernormal profit in the short run? The answer is 'not necessarily'. As a matter of fact, in the shortrun, a firm may make a supernormal profit or a normal profit or even make losses. Whether a firm makes abnormal profits, normal profits or makes losses depends on its cost and revenue conditions. If its short-run average cost (*SAC*) is below the price (*P = MR*) at equilibrium

(Fig. 12.2), the firm makes abnormal or pure profits. If its *SAC* is tangent to *P = MR*, as shown in Fig. 12.3 (a), the firm makes only normal profit as it covers only its *SAC* which includes *normal profit.* But, if its *SAC* falls above the price (*P = MR*), the firm makes losses. As shown in Fig. 12.3 (b), the total loss equals the area *PP′E′E* (= *PP′* × *OQ*), while per unit loss is *PP′ = EE′*.
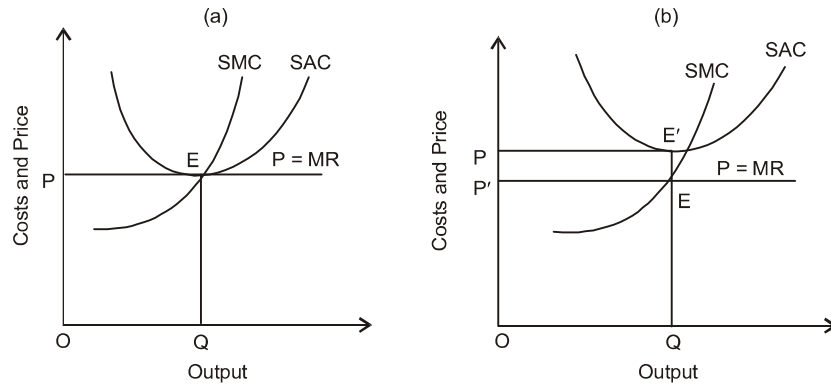
Fig. 12.3 *Short-run Equilibrium of Firm with Normal and Losses*

Shut-down or Close-down Point: In case a firm is making loss in the shortrun, it must minimise its losses. In order to minimise its losses, it must cover its *short-run average variable cost (SAVC).* The behaviour of *short-run average variable cost* is shown by the curve *SAVC* in Fig. 12.4. A firm unable to recover its minimum *SAVC* will have to close down. The firm's *SAVC* is minimum at point *E* where it equals the *MC*. Note that *SMC* intersects *SAVC* at its minimum level as shown in Fig. 12.4.
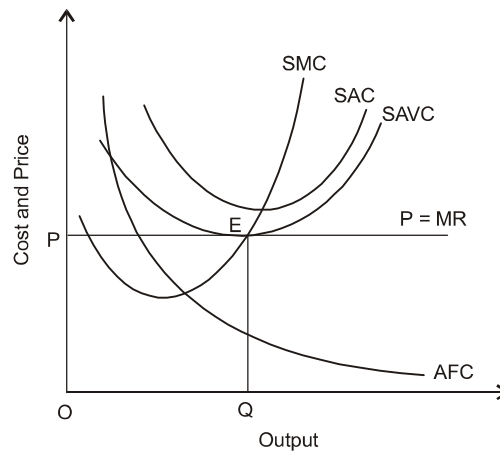
Fig. 12.4 *Shut-down Point*

Another condition which must be fulfilled is *P = MR = SMC*. That is, for loss to be minimum, *P = MR = SMC = SAVC*. This condition is fulfilled at point *E* in Fig. 12.4. At point *E*, the firm covers only its fixed cost and variable cost. It does not make any profit—rather it makes losses. The firm may survive for a short period but not for long. Therefore, point *E* denotes the 'shut-down point' or

'break-down point', because at any price below *OP,* it pays the firm to close down as it minimises its loss.

Derivation of Firm's Supply Curve

The supply curve of an individual firm is derived on the basis of its equilibrium output. The equilibrium output, determined by the intersection of *MR* and *MC* curves, is the optimum supply by a profit maximising (or cost minimising) firm. Under the condition of increasing *MC,* a firm will increase supply only when price increases. This forms the basis of a firm's supply curve. The derivation of supply curve of a firm is illustrated in Fig. 12.5 (a) and (b). As the figure shows, the firm's *SMC* passes through point *M* on its *SAVC.* The point *M* marks the minimum of firm's *SAVC* which equals MQ_1. The firm must recover its $SAVC = MQ_1$ to remain in business in the short run. Point *M* is the shut-down point in the sense that if price falls below OP_1, it is advisable for the firm to close down. However, if price increases to OP_2, the equilibrium point shifts to *R* and output increases to OQ_2. Note that at output OQ_2, the firm covers its *SAC* and makes normal profit. Let the price increase further to OP_3 so that equilibrium output rises to OQ_3. When price rises to OP_4, the equilibrium output rises to OQ_4 and the firm makes abnormal profit. By plotting this information, we get a *supply curve (SS')* as shown in Fig. 12.5 (b).
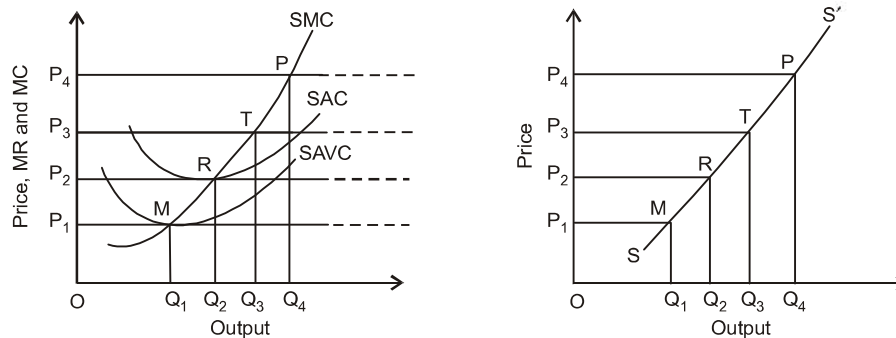
Fig. 12.5 *Derivation of a Firm's Supply Curve*

Derivation of Industry's Supply Curve

The industry supply curve, *or what is also called* **market supply curve, is the horizontal summation of the supply curve of the individual firms.** If cost curves of the individual firms of an industry are identical, their individual supply curves are also identical. In that case, industry supply curve can be obtained by multiplying the individual supply at various prices by the number of firms. In the shortrun, however, the individual supply curves may not be identical. If so, the market supply curve can be obtained by summing horizontally the individual supply curves. Let us consider only two firms having their individual supply curves as S_1 and S_2 as shown in Fig. 12.6 (a). At price OP_1, the market supply equals $P_1A + P_1B$. Suppose $P_1A + P_1B$ equals P_1M as shown in Fig. 12.6

(b). [Note that output scale in part (b) is different from that in part (a).] Similarly, at price OP_2, the industry supply equals $P_2C + P_2C$ or $2(P_2C) = P_2N$ as shown in Fig. 12.6 (b). In the same way, point T is located. By joining the points M, N and T, we get the market or industry supply curve, SS'.
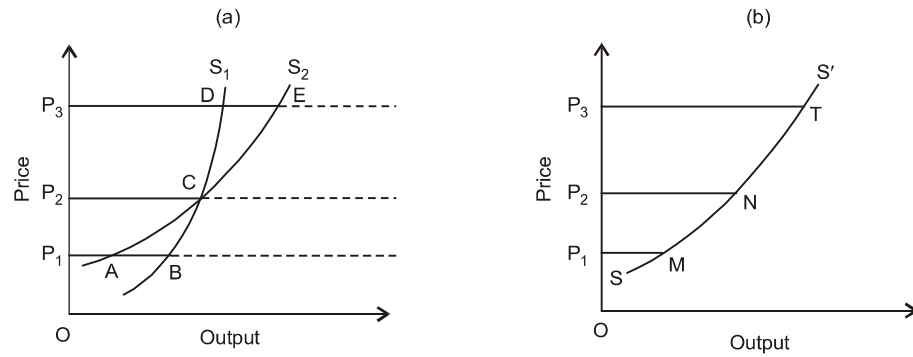
Fig.12.6 *Derivation of the Industry Supply Curve*

Equilibrium of Industry and Firm in Short-run

We have discussed above the equilibrium of the firm in the short-run in a perfectly competitive market. To complete the discussion on short-run price and output determination, now we discuss the short-run equilibrium of the industry.
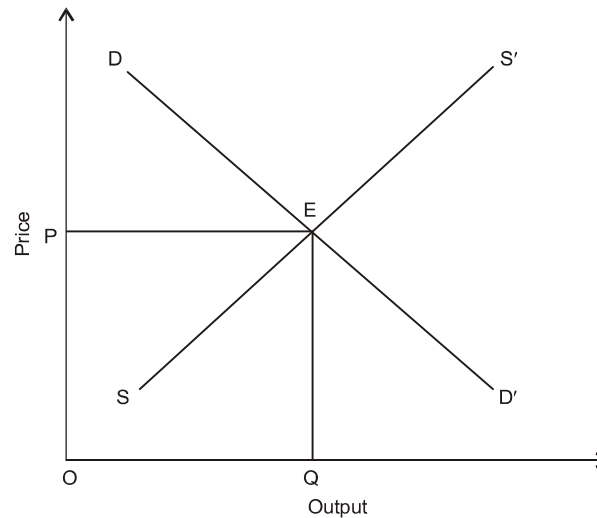
Fig. 12.7 *Equilibrium of the Industry*

An industry is in equilibrium in the shortrun when market is cleared at a given price, i.e., when the total supply of the industry equals the total demand for its product. The price at which market is cleared is called equilibrium price. When an industry reaches its equilibrium, there is no tendency to expand or to contract the output. The equilibrium of industry is shown at point E in Fig. 12.7. The industry demand curve DD' and supply curve SS' intersect at point E, determining equilibrium price OP. At this price, $D = S$. The industry is supplying as much as

consumers demand. In the short-run equilibrium of the industry, some individual firms may make pure profits, some normal profits and some may make even losses, depending on their cost and revenue conditions. As we have explained below, this situation will, however, not continue in the long-run.

Link between Short-run Equilibrium of the Industry and of the Firm

The short-run equilibrium of the firm and industry have been analysed separately in the previous sections. There exists, however, a link between the equilibrium of a firm and that of the firm industry. In a perfectly competitive market, change in the equilibrium of an individual firm does not affect the industry's equilibrium, simply because the total output of a single firm constitutes a small fraction of the industry's output. But, a change in the industry's equilibrium does alter the equilibrium of an individual firm. In this section, we show how individual firms move from one equilibrium position to another, when industry's equilibrium changes. For the sake of simplicity, we assume that all the firms of an industry have *identical cost conditions.*

The link between industry's and firm's equilibrium is illustrated in Fig. 12.8. Suppose industry's initial demand and supply curves are given as *DD* and *SS*, respectively, in Fig. 12.8 (a). As shown in panel (a) of the figure, industry's demand and supply curves intersect at point *P*, determining the market price at $PQ = OP_1$ and industry's equilibrium output is *OQ*. Thus, the price *PQ* is given to all the firms of the industry. Given the price *PQ* and firm's cost curves, an individual firm finds its equilibrium at point *E* in panel (b), where its *MC = MR*. Firm's equilibrium output is *OM* [see Fig. 12.8 (b)]. At price *EM = PQ,* the firm is making an abnormal profit in the short run to the extent of EN per unit of output. The firm's total pure profit is shown by the shaded rectangle.
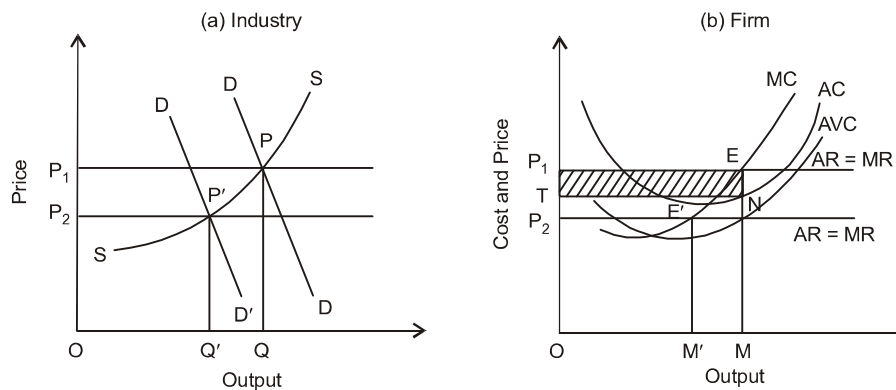
Fig. 12.8 *Industry's Vs Firm's Equilibrium in the Short Run*

Now let the industry demand curve *DD* shift downward for some reason to *DD′*, supply curve remaining unchanged. As a result, market price falls to $OP_2 = P′Q′$ and industry's equilibrium output falls to *QQ′*. With the fall in price, firm's equilibrium shifts from point *E* to *E′* where its *MC = MR* at *F′M′*. Thus, change in industry's equilibrium changes firm's equilibrium. Although firms are making loss

because their *AR* is less than their *AC*, this is only a short-run situation. In the long run, loss-making firms will quit the industry. As a result, supply curve *SS* will shift leftward causing price to go up. Losses will therefore disappear in the long run. This point is discussed below further.

Equilibrium of the Firm and Industry in Long-run

The long-run refers to a time period in which short-run conditions do not exist. Recall that short-run is, by definition, a period in which (*i*) firm's cost and revenue curves are given, (*ii*) firms cannot change their size—their capital is fixed, (*iii*) existing firms do not have the opportunity to leave the industry, and (*iv*) new firms do not have the opportunity to enter the industry. In contrast, long-run is a period in which these constraints disappear. Long-run permits change in technology and employment of both, labour and capital, i.e., firms can change their size. Some of the existing firms may leave and new firms may enter the industry. In the long-run, supply curve not only shifts rightward but also becomes more elastic.

In this section, we will analyse the equilibrium of the firm and industry in the long-run. It should be noted that the process by which firms and industry reach their respective long-run equilibrium, is a continuous process of adjustment and readjustment of price and output with the changing conditions in the long-run. The process of equilibrium adjustment by the firm and industry is discussed below.

Equilibrium of the Firm

The long-run equilibrium of the firm is illustrated in Fig. 12.9. To explain it furhter, let us begin with a short-run situation. Suppose (*i*) that short-run price is given at OP_1, in panel (a) of Fig. 12.9, and (*ii*) that firms' short-run cost curves are identical and are given by SAC_1 and SMC_1, as shown in panel (b). Given the price OP_1, firms are in equilibrium at point E_1. It can be seen in part (b) of Fig. 12.9 that the firms are making an abnormal profit to the extent of $AR - SAC$, i.e., $E_1M = E_1Q_1 - MQ_1$ per unit of output. The abnormal profit brings two major changes in the industry.

One, existing firms get incentive to increase the scale of their production. Their average and marginal costs go down caused by the economies of scale. This phenomenon is shown by SAC_2 and SMC_2. When we draw the *LAC* and *LMC* curves, these curves show decreasing costs in the long-run.

Two, attracted by the abnormal profit, new firms enter the industry increasing the total supply causing a rightward shift in the industry supply curve.
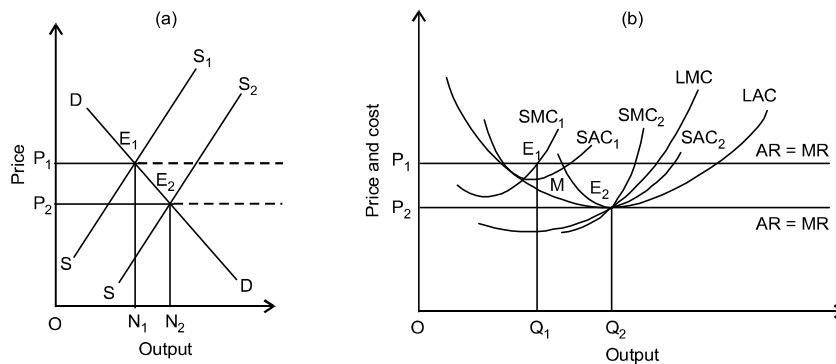
Fig. 12.9 *Long-run Equilibrium of the Firm and Industry*

For these reasons, the industry supply curve, SS_1, shifts rightward to SS_2 [Fig. 12.9(a)]. The shift in supply curve (demand curve, *DD*, remaining the same) brings down the market price to OP_2 which is the long-run equilibrium price. Thus, equilibrium price is once again determined for the industry a OP_2.

Given the new market price, OP_2, firms attain their equilibrium in the long-run where $AR = MR = LMC = LAC = SMC = SAC$. This point is shown at point E_2 in Fig. 12.9 (b). As the figure shows, the firms of an industry reach their equilibrium in the long-run where both short-run and long-run equilibrium conditions coincide. In a perfectly competitive market, the cost and revenue conditions are given for the firms. Therefore, when price goes down to OP_2, what firms are required to do is to adjust their output to the given revenue and cost conditions in order to maximise their profit. Let us now illustrate the process of adjustment of output so as to reach the equilibrium in the long-run.

So long as economies of scale are available to the firms, the *LAC* tends to decrease and it pays firms to expand their plant-size. This process of output adjustment continues until industry reaches its equilibrium. The industry reaches its equilibrium where *LAC* is tangent to $P = AR = $ MR for each firm in the industry. This position is shown at point E_2, in Fig. 12.9 (b). At point E_2, the point of equilibrium, $P = MR = LMC = LAC = SMC = SAC$. Since $P = LAC$, the firms make only normal profits in the long-run. If firms deviate from point E_2, due to some short-run disturbances, the market forces of demand and supply come into operation and restore the equilibrium.

12.2.3 Equilibrium of the Industry

An industry is in equilibrium at a price and output at which market is ceared, i.e., where market demand equals market supply. The equilibrium of the industry is illustrated in Fig. 12.9 (a). When an industry is in equilibrium, all its firms are supposed to be in equilibrium [as shown in Fig. 12.9 (b)], and earn only normal profits. This is so because under the conditions of perfect competition, all the firms are assumed to achieve the same level of efficiency in the long-run. Since industry yields only normal profits, there is no incentive for new firms to enter the industry. These conditions are fulfilled at price OP_2 in Fig. 12.9 (a) and (b). At price OP_2,

all the firms are in equilibrium, where each firm has its $LMC = LMR = SMC = SAC = P = LAC$.

Since $P = LAC$, all the firms are earning only normal profit. At industry's equilibrium output ON_2, market demand equals market supply [Fig. 12.9 (a)]. At price OP_2, therefore, market is cleared. The output ON_2 may remain stable in the long-run. For, there is no incentive for new firms to enter the industry and no reason for the existing ones to leave the industry. The industry is, therefore, in equilibrium.

12.3 LONG-RUN SUPPLY CURVE OF THE INDUSTRY

We have earlier derived the short-run supply curve of the industry by summing up horizontally the supply curves of the individual firms (see Fig. 12.6). The long-run supply curve of a competitive industry, however, has nothing to do with the LMC curves. The shape of the long-run supply curve of an industry, under perfect competition, depends on whether factor prices remain constant or increase, when demand for inputs increases in the long-run as a result of expansion of the output of the industry. Depending on whether industry's cost is constant, increasing or decreasing, industries are classified as *constant cost, increasing cost* or *decreasing cost industry*. Let us, now, derive the long-run supply curve of industries conforming to their cost behaviour.

12.3.1 Constant Cost Industry

An industry for which factor prices remain constant is referred to as constant cost industry. In other words, **when the expansion of output in an industry does not entail a change in factor prices, the industry is said to be a constant cost industry**. The shape of the supply curve of such an industry is illustrated by the line *LRS* in Fig. 12.10 (a) and (b). To explain the horizontal shape of the supply curve, let us suppose that the industry is in equilibrium at point *P* where demand curve DD_1 and supply curve SS_1 intersect. The industry is in equilibrium at price OP_1 and output OQ_1. At price OP_1, all firms are in equilibrium as their $LMC = P = MR = SMC = SAC$.
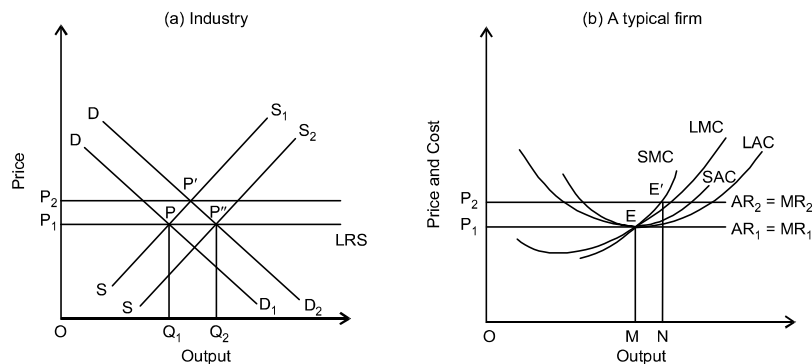
Fig. 12.10 *Long-run Supply Curve of the Constant Cost Industry*

Now, let the demand curve DD_1 shift to DD_2 due to, say, increase in consumers' income or increase in population or due to both, supply curve remaining the same. As a result, market price increases to OP_2. In the short-run, this increase in price can induce an increase in supply by the firms only by *MN* as shown in Fig. 12.10 (b), as determined by the point of intersection between firms' *SMC* and new price line through P_2. The firms enjoy abnormal profit to the extent of AR_2 – *LAC*.

The abnormal profits attract new firms into the industry. The entry of new firms leads to increase in demand for inputs. However, the industry being a constant cost industry, factor prices do not increase—may be due to simultaneous increase in input supply. Cost of production for all the firms remains constant at the previous level. But, due to the entry of new firms, market supply increases and market supply curve shifts from SS_1 to SS_2 [Fig. 12.10 (a)]. Consequently, in the long-run, market price falls to its previous level, OP_1, and firms return to their previous equilibrium point *E*. But the industry output increases from OQ_1 to OQ_2 and industry moves from equilibrium point *P* to *P″*. By joining the two points of industry equilibrium *P* and *P″*, and extending it further, we get long-run supply curve (*LRS*) of the constant cost industry. Obviously, the long-run supply curve (*LRS*) of a constant cost industry is a horizontal straight line, as given by the line *LRS*.

Increasing Cost Industry

An increasing cost industry is one which faces increasing input prices. The increase in input prices may be caused by increase in demand for inputs but their supply does not keep pace with rising demand. The long-run supply curve of an **increasing cost industry** has a positive slope as illustrated in Fig. 12.11 (a).

The derivation of long-run market supply curve under increasing cost condition is demonstrated in Fig. 12.11. Let the original demand and supply curves of the industry be represented, respectively, by DD_1 and SS_1, and industry be in equilibrium of point *A*. Let us also suppose that for some reason demand curve DD_1 shifts to DD_2, supply curve remaining the same. As a result, short-run market price increases from OP_1 to OP_3 given the supply curve S_1. With this increase in price, the demand curve for the firms shifts upward to $AR_3 = MR_3$ [Fig. 12.11 (b)]. The firms, therefore, enjoy a super normal or economic profit to the extent of P_1P_3. This supernormal profit attracts new firms to the industry and demand for inputs increases. Since by assumption, the supply of inputs is less than infinitely elastic, the entry of new firms causes an increase in demand for inputs and, therefore, an increase in the input prices. Consequently, cost curves, both short-run and long-run, shift upward tram LAC_1 to LAC_2. In this process of adjustments, however, industry-supply increases and market supply curve SS_1 shifts rightwards to SS_2, With this shift in the supply curve, the industry reaches another equilibrium position at point *C* where new demand and supply curves intersect. A new market price is determined at OP_2. At price OP_2 [Fig. 12.11 (b)], the long-run and

short-run cost curves are tangent to the price line $OP_2 = AR_2 = MR_2$. The individual firms shift to a new long-run equilibrium point E_2, their individual output remaining the same. Whether equilibrium output of the firms remains constant, increases or decreases, depends, respectively, on whether cost curves shifts upward vertically, upward to the right or upward to the left.
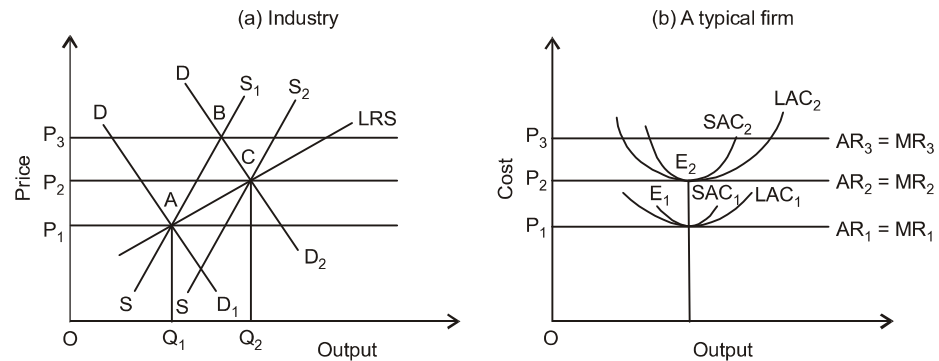
Fig. 12.11 *Long-run supply curve of an Increasing Cost Industry*

Note that at price OP_2, both industry and individual firms are in equilibrium. In the absence of any further disturbance, the equilibrium of both firms and industry will remain stable. Thus, at the new equilibrium price OP_2, the industry-output increases from OQ_1 to OQ_2 and corresponding equilibrium points are A and C, respectively. By joining the long-run equilibrium points A and C, we get the long-run supply curve for the industry, as shown by the curve *LRS*. Obviously, the *LRS* has a positive slope in an increasing cost industry.

Decreasing Cost Industry

If expansion of output of an industry is associated with decrease in the input prices, the industry is referred to as a *decreasing cost industry.* A decreasing cost industry has a long-run supply curve with a negative slope.

The derivation of long-run industry supply curve (*LRS*) under decreasing cost condition is illustrated in Fig. 12.12(a) and (b). Let the industry be initially in equilibrium at point A [Fig. 12.12 (a)] and firms at E_2 [Fig. 12.12 (b)]. Now suppose that demand curve shifts from DD_1 to DD_2 and, consequently, price rises from OP_2 to OP_3. The short-run equilibrium of firms at price OP_1 [Fig. 12.12 (b)] moves upward on the SMC_2 where the firms make abnormal profits. The abnormal profits attract new firms to the industry causing increase in demand for inputs. In case industries are enjoying increasing returns to scale due to economies of scale, the increase in demand for inputs would encourage increased supply of inputs. Increase in the supply of inputs causes input prices to fall. The industry, therefore, enjoys the external economies to scale. As a result, their long-run and short-run cost curves shift downward, from LAC_2 to LAC_1 [Fig. 12.12 (b)].
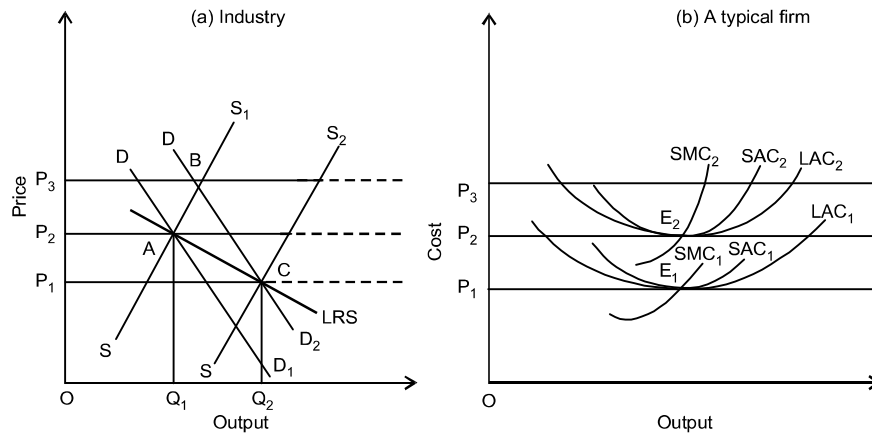
Fig. 12.12 *Long-run Supply Curve of a Decreasing Cost Industry*

From the industry's point of view, industry supply increases due to the entry of new firms, even if the existing firms maintain their old level of output. Therefore, the industry supply curve shifts from SS_1 to SS_2 which intersects the new demand curve DD_2 at point C. Thus, the equilibrium of the industry shifts from A to C. Industry output increases from OQ_1 to OQ_2. In the absence of any external disturbance, the industry equilibrium point C, would tend to stabilise. By joining the two equilibrium points A and C and extending it further, we get the long-run supply curve (*LRS*) of the decreasing cost industry. The *LRS* has a negative slope.

Whether Decreasing Cost: Some authors argue that the 'phenomenon of decreasing cost... is not consistent with all the requirements of perfect competition. However, the possibility of a decreasing cost industry cannot be ruled out in a very long period. One reason for this is the likelihood of the existence of large external economies of scale, particularly in case of young industries in the undeveloped areas. An increase in the number of industries and the consequent growth of transportation, marketing facilities and financial institutions may reduce the industry's cost of production. Nevertheless, it depends on how substantial are the external economies of scale. R.G. Lipsey has cited the car industry of England as an example of decreasing cost industry. In his own word, "As the output of cars increased, the industry's demand for tyres grew greatly. This ... would have increased the demand for rubber and tended to raise its price, but it also provided the opportunity for tyre manufacturers to build large modern plants and reap the benefits of increasing returns in tyre production. At first, these economies were large enough to offset any factor price increases and tyre price charged to car manufacturers all. Thus car costs fell because of lower prices of an important input."

Conclusion: To conclude, whether costs of an industry remain constant or decrease due to increase in the price of some of its inputs, depends also on what proportion of the total input supply is consumed by the industry. For example, output of pencil industry can be increased without substantially affecting the lumber prices as pencil industry uses a small proportion of lumber output.

But a large increase in the output of furniture industry will not leave lumber prices unaffected. Similarly, output of a pin industry can be substantially increased without affecting the steel price. But a substantial increase in car output cannot leave steel prices unaffected.

Another factor which may cause a rise in input prices is whether or not input industries enjoy economies of scale.

Moreover, the most common cases are of the constant and increasing cost industries. Decreasing , cost industries are most unlikely to exist for a long time. The constant and decreasing cost industries, tend over time to become increasing cost industries because external economies have a limit.

Check Your Progress

1. List any two characteristics of perfect competition.

2. How is the supply curve of an individual firm derived?

12.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Two characteristics of perfect competition are the following:

   - Large number of buyers and sellers

   - Perfect mobility of factors of production

2. The supply curve of an individual firm is derived on the basis of its equilibrium output.

12.5 SUMMARY

- Perfect competition refers to a market condition in which a very large number of buyers and sellers enjoy full freedom to buy and to sell a homogenous good and service and they have perfect knowledge about the market conditions, and factors of production have full freedom of mobility.

- Under perfect competition, the number of sellers is assumed to be so large that the share of each seller in the total supply of a product is very small or insignificant.

- Perfect competition assumes that there is no collusion between the firms, i.e., they are not in league with one another in the form of guild or cartel.

- While 'Perfect Competition' has all the features mentioned above, under 'Pure Competition', there is no perfect mobility of factors and perfect knowledge about market-conditions. That is, perfect competition less 'perfect mobility' and 'perfect knowledge' is pure competition.

- The supply curve of an individual firm is derived on the basis of its equilibrium output. The industry supply curve, or what is also called market supply curve, is the horizontal summation of the supply curve of the individual firms.

- An industry is in equilibrium at a price and output at which market is ceared, i.e., where market demand equals market supply.

- An industry for which factor prices remain constant is referred to as constant cost industry. In other words, when the expansion of output in an industry does not entail a change in factor prices, the industry is said to be a constant cost industry.

- If expansion of output of an industry is associated with decrease in the input prices, the industry is referred to as a *decreasing cost industry.* A decreasing cost industry has a long-run supply curve with a negative slope.

12.6 KEY WORDS

- **Perfect competition:** It refers to a market condition in which a very large number of buyers and sellers enjoy full freedom to buy and to sell a homogenous good and service and they have perfect knowledge about the market conditions, and factors of production have full freedom of mobility.

- **Constant cost industry:** When the expansion of output in an industry does not entail a change in factor prices, the industry is said to be a constant cost industry.

12.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. What are the features of perfect competition? Distinguish between perfect and pure competition.

2. What is the relative position of a firm in a perfectly competitive industry? How does it choose its price and output?

3. Under perfect competition average revenue equals average cost in the long run equilibrium. Yet why do firms produce under such a condition?

4. Bring out the essential difference in the nature of equilibrium of a firm under perfect competition in the short-run and in the long-run.

5. Write a short note on the relationship between firm's short-run cost curves and supply curve.

6. If all the firms in a perfectly competitive industry have U-shaped cost curves, can then supply curve of the industry be downward sloping?

Long Answer Questions

1. Analyse the equilibrium of a firm under the conditions of perfect competition in the short-run? Discuss in this regard the importance of *AR, AC, MR* and *MC* under perfect competition.

2. Explain the short-run equilibrium of a competitive firm. When would a competitive firm close down its business in the short-run?

3. Do you agree that perfect competition leads to optimum size of the firm? Give reasons for your answer.

4. Show graphically long-run supply curves of an industry is drawn under perfect competition? Also illustrate graphically the derivation of the long run supply curve of a firm under perfect competition.

5. Suppose a competitive firm is in long-run equilibrium. What will happen to price in the long-run if there is a rise in demand for the product of the industry?

12.8 FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.

UNIT 13   PRICE AND OUTPUT DETERMINATION UNDER MONOPOLY

Structure

13.0   INTRODUCTION

In the preceding unit, we have discussed the theory of price and output determination under perfect competition. In this unit, we discuss the theory of price and output determination under monopoly—another extreme case of market. Recall that under perfect competition, a firm is a **price-taker**. In contrast, under monopoly, a firm is a **price-maker**. The following aspects of monopoly theory are discussed here.

   (*i*) Meaning and sources of monopoly power,

  (*ii*) Price and output determination in the short and long runs,

 (*iii*) Price discrimination by a monopoly firm,

 (*iv*) Pricing under multiplant monopoly,

  (*v*) Comparison of price and output under monopoly and perfect competition,

 (*vi*) Measures of monopoly power, and

(*vii*) Government control of monopolies.

13.1 OBJECTIVES

After going through this unit, you will be able to:

- Define monopoly

- State the sources and kinds of monopolies

- Analyse revenue curve under monopoly

- Discuss the rule of thumb pricing under monopoly

- Explain price discrimination under monopoly

- List the necessary conditions for price discrimination

- Compute the dead-weight loss under monopoly

- Discuss the measures of monopoly power

- Identify the measures adopted by the government to regulate monopoly prices

13.2 MONOPOLY: AN OVERVIEW

The word monopoly has been derived from Greek word *monos,* meaning 'alone' and *polein* meaning 'seller'. By definition, *monopoly is a market situation in which there is a single seller of a commodity of* '*lasting destination*' *without close substitutes.* A monopoly firm enjoys an absolute power to produce and sell a commodity. This, however, does not mean that a monopoly firm is absolutely free from any kind of competition. Monopoly firms too, have to face *indirect competition,* especially in regard to setting the price of the product. There are at least *two potential sources of indirect competition.*

First potential source of indirect competition is the rivalry between monopoly good and other goods produced by other monopolies and competitive firms. Therefore, a monopolist cannot charge any price for its product. For example, North Delhi Power Limited, a public sector electricity producing and supplying company is at present a monopolist in Delhi. Pricing of electricity is under the control and regulation of the government. Therefore, while fixing the electricity price, it will have to take into account in government's pricing policy. Besides, it will have to take into account what people can afford after meeting such essential needs as food, clothing, shelter, education and medicine.

The **second** source of potential indirect competition comes from the availability and price of *inferior substitutes.* For example, the Mahanager Telephone Nigam Limited (Delhi) still a monopolist in communications, is facing competition from the cellphone companies, Tata Telephones, Airtel and BSNL. Therefore, its monopoly power is considerably eroded. So is the case with Delhi Transport Corporation with increasing number of private operators of charted buses.

Given these problems confronting the monopolies, one can hardly find many cases of a **pure** or **absolute monopolies**. However, the theoretical discussion on price and output determination under monopoly is based on the case of a *pure monopoly,* i.e., a monopoly firm enjoying absolute power in determining the price and output of its product.

Finally, an **important feature** of a pure monopoly is that a monopolized industry is a single-firm industry, i.e., there is no distinction between the firm and the industry. Therefore, there is no distinction between market demand curve and monopoly firm's own demand curve, i.e., the demand curve for the monopoly firm's product is same as the market demand curve.

13.2.1 Sources and Kinds of Monopolies

The emergence and survival of monopoly is attributed to the factors which prevent the entry of other firms into the industry. The barriers to entry are, therefore, the sources of monopoly power. The major sources of barriers to entry are: (*i*) legal restrictions, (*ii*) sole control over the supply of certain scarce and key raw materials; (*iii*) efficiency; (*iv*) economies of scale; and (*v*) patent rights. How these factors prevent the entry of new firms is discussed subsequently in this unit.

(i) Legal restrictions. Some monopolies are created by law in public interest. Such monopolies may be created in both public and private sectors. Most of the state monopolies in the public utility sector, including postal, telegraph, and telephone services, generation and distribution of electricity, railways, airlines and state roadways, etc., are public monopolies. Such monopolies, wherever they exist, are created by the public law. The state may create monopolies in the private sector also through licence or patent. Such monopolies are intended to reduce cost of production by the economies of scale and investment in technical innovations. Such monopolies are also known as **franchise monopolies**.

(ii) Control over key raw materials. Some firms acquire monopoly power from their overtime control over certain scarce and key raw materials that are essential for the production of certain other goods, e.g., bauxite, graphite, diamond, etc. For instance, Aluminium Company of America had monopolised the aluminium industry before the World War II because it had acquired control over almost all sources of bauxite supply. Such monopolies are often called 'raw material monopolies'. The monopalies of this kind emerge also because of monopoly over certain specific technical knowledge or techniques of production.

(iii) Efficiency. A primary and technical reason for growth of monopolies is the economies of scale. In some industries, long-run minimum cost of production, i.e., the most efficient scale of production, coincides almost with the size of the market. In such industries or prodcuts, a large-size firm finds it profitable, in the long-run, to eliminate competition by cutting down its price for a short period. Once a monopoly is established, it becomes almost impossible for the new firms to enter the industry and survive. Monopolies born out of efficiency are known as **natural monopolies**. A natural monopoly may emerge out of the technical conditions of efficiency or may be created by the law on efficiency grounds in public interest.

(iv) Economies of scale. Under free economic system, a large firm enjoying large economies of scale is capable of cutting down the price low enough to eliminate the competitors. This method is adopted in the absence of Monopoly Control Act in the country. Once competing firms are elimited, the large firm enjoys the powers of a monopoly firm.

(v) Patent rights. Another source of monopoly is the patent rights of the firm for a product or for a production process. Patent rights are granted by the government to a firm to produce a commodity of specified quality and character or to use a specified technique of production. Patent rights give a firm exclusive rights to produce the specified commodity or to use the specified technique of production. Such monopolies are called **patent monopolies**.

13.2.2 Revenue Curves Under Monopoly

The tools that are used to analyze price and output determination include cost curves and revenue curves. It is therefore essential that we acquaint ourselves with cost and revenue curves faced by monopoly firms. Let it be noted at the first instance that cost curves—the *AC* and *MC* curves—faced by the monopoly firm are U-shaped, just as those faced by the firms under perfect competition. But a monopoly firm has option to charge a high or a low price at its discretion. Therefore, monopoly firm has a downward sloping demand curve. As a result, the *AR* and *MR* curves that a monopoly firm faces are different from those faced by the firms under perfect competition. Therefore, before we discuss price and output determination and firm's equilibrium under monopoly, it will be useful to look into the nature of revenue curves faced by a monopoly firm.

As we have noted earlier, in a perfectly competitive market, there is dichotomy between the firm and the industry—firms face a horizontal, straight-line demand curve and industry faces a downward sloping demand curve. Since under monopoly, there is a single firm in the industry, *the industry is a single-firm-industry and industry demand curve has a negative slope.* A monopoly firm faces, therefore, a downward sloping demand curve—it may be a linear or a non-linear demand curve. It is **important** to note here that, given the demand curve, a monopoly firm has the option to choose between the prices to be charged and output to be sold. Once it chooses a price, the demand for its output is fixed. Similarly, if the firm decides to sell a certain quantity of output, then its price is fixed. It cannot charge any other price inconsistent with the demand curve.

For example, suppose that the demand curve for a monopolised industry is given as *DM* in Fig. 13.1. Demand curve, *DM,* shows the quantities that can be sold at different prices. For instance, if monopoly firm chooses price *OP,* the quantity that it can sell at this price is fixed at *OQ* – no other quantity can be sold at this price. Similarly, if it decides to sell quantity *ON,* its price has to be fixed at *OP'* – it cannot sell *ON* output at a higher price. This means that if demand curve is given, the options of monopoly firm becomes limited – it can choose either a price or a quantity at a time. The firm cannot choose a price and a quantity

inconsistent with the demand curve. This implies that a monopoly firm has the option of charging a high price and selling a lower quantity or charging a lower price and selling a larger quantity.
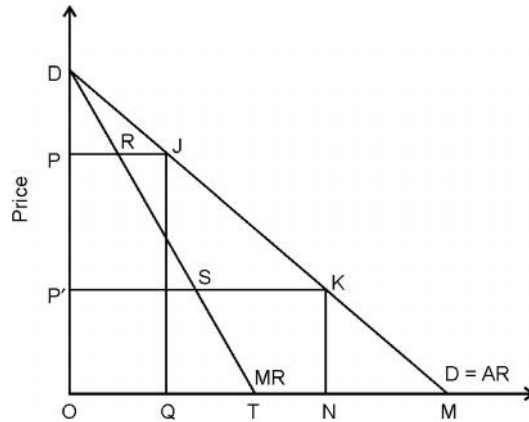
Fig. 13.1 *AR and MR Curves for Monopoly*

The Relationship between AR and MR Curves: Another point that needs to be noted is the relationship between monopoly's *AR* and *MR* curves. As noted earlier, the *AR* curve for a firm is the same as its demand curve. So is the case with a monopoly firm. Since a monopoly firm faces a downward sloping demand curve, its *AR* also slopes downward to the right. For example, the demand curve *DM* in Fig. 13.1 is the same as the firm's *AR* curve.

What is much more *important* in the analysis of equilibrium of a monopoly firm is the *relationship between the AR and MR* curves. when price is fixed, as in case of perfect competition, firm's demand curve takes the form of a horizontal line. In that case, *AR* = *MR* and *MR* is a straight line too. But, in case of a monopoly firm, demand curve has a negative slope. Therefore, its *MR* curve too has a negative slope. There is, however, a *specific relationship* between *AR* and *MR,* i.e., *the slope of MR curve* is *twice that of that AR curve.* That is, given the linear demand function, marginal revenue curve is twice as steep as the average revenue curve. This relationship can be proved as follows. Let us assume that a monopoly firm is faced with a price function given as

$$P = a - bQ \qquad \qquad \dots(13.1)$$

We know that $TR = Q \cdot P$

By substituting Eq. 13.1 for *P,* we get

$$TR = Q(a - bQ)$$
$$= aQ - bQ^2 \qquad \qquad \dots(13.2)$$

Since *MR* equals the first derivative of the *TR* function,

$$MR = \frac{\partial TR}{\partial Q} = \frac{\partial(aQ - bQ^2)}{\partial Q}$$

$$= a - 2bQ \qquad \qquad \dots(13.3)$$

Note that the slope of the price function (13.1) equals b whereas the slope of the MR-function (13.3) equals $2b$, it means that the slope of the MR-function is *twice* that of the AR-function. It implies that MR curve is always to the left of AR curve and MR bisects the demand at all levels of price. For example, in Fig. 13.1, if price is OP, demand is PJ, and MR passes through point R which divides PJ in two equal parts. Geometrically, $PR = RJ$. Similarly, at price OM, demand equals $P'K = P'S + SK$ where $P'S = SK$, i.e., $P'S = P'K/2$.

Price and Output Determination in Short-run

Having defined monopoly and discussed the nature of revenue curves it faces, let us now discuss the price and output determination under monopoly. According to the traditional theory of firm, a monopoly firm (or otherwise) is said to be in equilibrium where it maximises its profit. Maximisation of total profit is a matter of time required to adjust output to the price. Therefore, as in case of perfect competition, equilibrium of a monopoly is studied under both *short-run* and *long-run* conditions. In this section, we explain price and output determination under monopoly in the *short-run*. The equilibrium of monopoly in the *long-run* will be discussed in the next section.

The short-run equilibrium of the monopoly firm is illustrated in Fig. 13.2. The short-run revenue curves of the monopoly firm are shown by the AR and MR curves and its short-run cost curves are shown by the SAC and SMC curves. Given the revenue and cost conditions and the profit maximisation rule, the equilibrium of the monopoly firm can easily be traced. Recall once again that profit is maximum where $MR = MC$. It can be seen in the figure that MR and SMC curves intersect at point N. Note that point N satisfies both the conditions of profit maximisation: (*i*) $MR = MC$, and (*ii*) SMC curve intersects MR curve from below. Point N, therefore, determines the equilibrium output and price. An ordinate drawn from point N to X-axis determines the profit maximising output at OQ. The ordinate NQ extended upward to the AR curve gives the price PQ at which output OQ can be sold, given the demand function. Thus, at monopoly equilibrium determines, both equilibrium output and price are determined simultaneously.
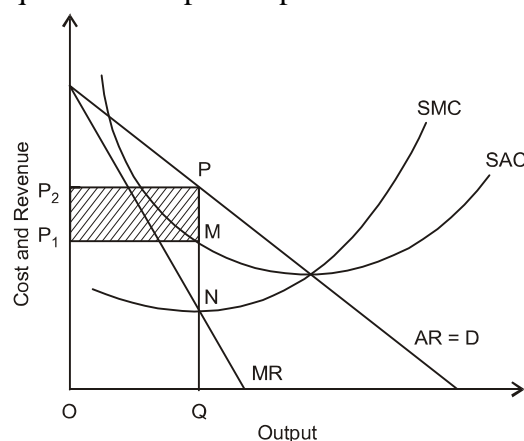
Fig. 13.2 *Monopoly Equilibrium: MR-MC Approach*

Once equilibrium price and output are determined, given the revenue and cost curve, the maximum monopoly profit can be easily determined as follows.

Per unit monopoly profit $= AR - SAC = PQ - MQ = PM$

Given the equilibrium output *OQ,* total monopoly profit can be obtained by multiplying per unit profit *PM* by the equilibrium output, *OQ.* That is, total monopoly profit $= OQ . PM$.

Since $OQ = P_1M,$ total monopoly profit at equilibrium can be written as

$$P_1M \cdot PM = P_1P_2PM$$

The total monopoly profit is shown by the shaded area in the Fig. 13.2. Since cost and revenue conditions of the monopoly firm are supposed to be given, the monopoly equilibrium is supposed to be stable.

Two Common Misconceptions about monopoly

In general usage of the term, there are two common misconceptions about monopolies: (*a*) that monopolies always make profits, and (*b*) that they can charge any price arbitratily. These misconceptions must be clarified before we proceed to discuss other related issues.

(a) Does a monopoly firm necessarily make profit in the short-run? There is no certainty that a monopoly firm will always make profits in the short-run. In fact, whether a monopoly makes profit or loss in the short-run depends on its revenue and cost conditions. It is quite likely that its *SAC* curve lies above its *AR* curve as shown in Fig. 13.3. At profit maximising output (*OQ*), *SAC* exceed *AR* by *PM*. The monopoly firm therefore, makes losses to the extent of $PM \times OQ$ $= P_1MPP_2$ in the short-run. The firm may yet survive in the short-run as it is sure of making profits in the long-run when economies of scale become available to the firm. The monopoly firm will however stick to the profit maximisation rule (i.e. $MR = MC$) in order to minimise its losses in the short-run. Furthermore, if a monopoly firm operates in the short-run at a level of output where its $MR = SMC$ and $AR = SAC,$ it makes just normal profits like a competitive firm in the long-run.
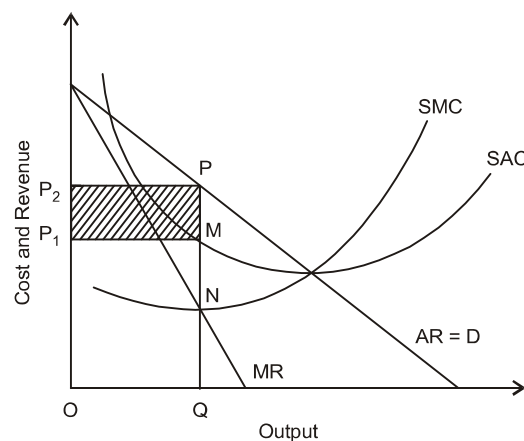
Fig. 13.3 *Monopoly Equilibrium in the Short-Run: Losses*

(b) Can monopolies charge any arbitrary price? Another common misconception about monopoly is that a monopoly firm, by virtue of being a single seller of a commodity, can charge *any* price or an exhorbitantly high price for its product. In fact, the demand curve faced by a monopolist, is also the industry's demand curve. And, most market demand curves are negatively sloped, being highly elastic in the region of upper half and highly inelastic in the lower half. As shown in Fig. 13.4, point *M* marks the mid-point of the demand curve *PQ*. As proved earlier, at mid-point of the demand curve, i.e., at point *M*, *e* = 1. Over the upper half of the demand curve, *e* > 1; and over its lower half, *e* < 1. Therefore, a profit maximizing or a loss minimizing monopoly firm cannot charge any price of its own choice.

The Rule-of-Thumb Pricing under Monopoly

If the monopoly firm is aware of its *MC* and price elasticity (*e*) of demand for its product, it can easily find out its profit-maximising price using the relationship between *AR* and *MR*. Suppose firm's *MC* = 4 and elasticity of its demand curve *e* = – 2. Given these variables, the profit maximising price can be obtained as follows.

Profit is maximum where

$$MC = MR$$

Since *MC* = 4, at equilibrium, *MR* must be equal to 4, i.e.,

$$MC = 4 = MR$$

We know that $MR = P\left(1 - \dfrac{1}{e}\right)$

Since *MC* = 4 = *MR,* and *e* = – 2, by substitution, we get

$$4 = P\left(1 - \frac{1}{2}\right)$$

or

$$P = \frac{4}{P\left(1 - \dfrac{1}{2}\right)}$$

$$P = 8$$

Thus, *under the given conditions,* profit-maximising price = 8. Note that if *MR* is known, *MC* can be found out, and if *P* and *e* are known, *MC* and *MR* can be worked out for the equilibrium output.

There is No Monopoly Supply Curve in Short-run

In this section, we will answer the question whether there exists a supply curve under monopoly. As mentioned earlier, economists believe that, **there is no unique or precise supply curve under monopoly.** Let us examine this issue.

A supply curve exhibits a unique relationship between price and quantity supplied. This unique relationship between market price and quantity supplied

does not exist under monopoly, at least at the theoretical level. The argument runs as follows. A profit maximising monopoly firm determines its equilibrium output where $MR = MC$ and $P > MC$. Therefore, a *unique* relationship between price ($AR = P$) and quantity supplied cannot be established. The reason is that there are two different possibilities: (*i*) that given the *MC*, the same output is supplied at different prices, and (*ii*) that at a given price, different quantities may be supplied if the two downward sloping demand curves have different elasticities. The two cases are illustrated in Figs. 13.4 and 13.5, respectively.
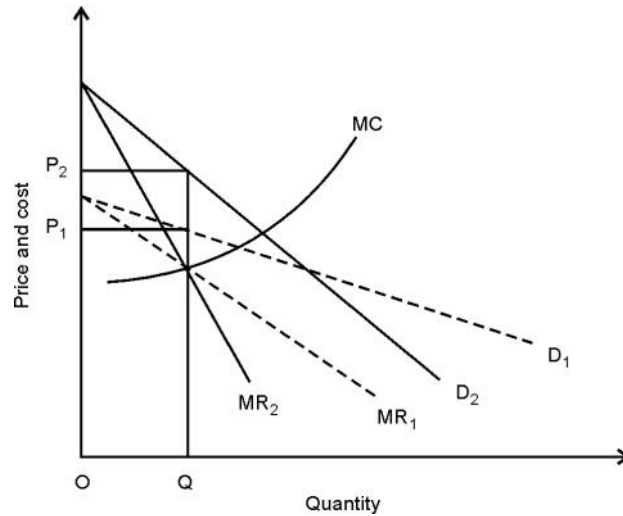
Fig. 13.4 *Same Quantity Supplied at Two Different Prices: Case I*
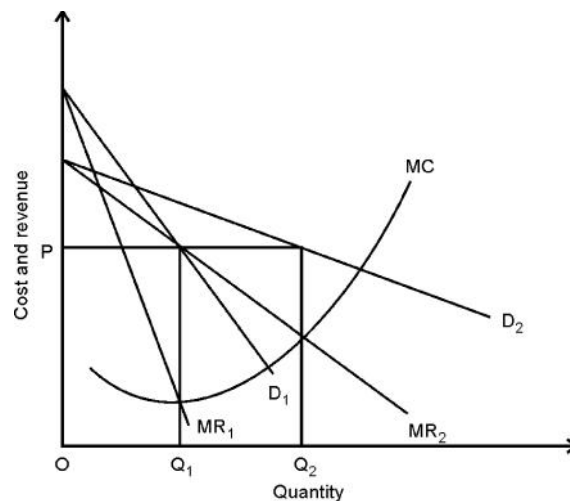
Fig. 13.5 *Different Quantities Supplied at the Same Price: Case II*

Figure 13.4 shows that, given the *MC* curve, if there are two demand curves with different slopes—as is the general case—the same quantity (OQ) can be supplied at two different prices—OP_1 when demand curve is D_1 and OP_2 when demand curve is D_2. Obviously, there is no unique relation-ship between price and quantity supplied.

Figure 13.5 presents the case of two different quantities supplied at the same price, *OP.* Given the *MC* curve, quantity OQ_1 is supplied when demand curve is D_1 and quantity OQ_2 is supplied when demand curve is D_2 at the same price, *OP.* In this case too, there is no unique relationship between price and quantity supplied. These points lead one to the conclusion that *there is no unique supply curve under monopoly.*

Two points are important to remember here. *One*, it is only the monopoly firm which faces two demand curves. *Two*, the absence of supply curve under monopoly is a only a short-run affair.

Monopoly Equilibrium in the Long-run

The long-run equilibrium conditions of a monopolist differ from those faced by the competitive firms in another important respect, i.e., the entry of new firms into the industry. While in a competitive market, there is free entry to the industry, a monopoly firm is protected by the *barriers to entry.* The barriers to entry may be in the form of patent rights, legal protection, economies of scale and the well established long standing of the monopolist.

In the long run, a monopolist gets an opportunity to expand the size of its plant with a view to maximising its long-run profits provided the size of the market is fairly large. The expansion of the plant size may, however, be subject to such conditions as: (*a*) size of the markets; (*b*) expected economic profits; and (*c*) risk of inviting legal restrictions, like *MRTP*.

A general case of monopoly equilibrium in the long-run is presented in Fig. 13.6, assuming none of the above conditions limits the expansion of monopoly firm. The *AR* and *MR* curves show the market demand and marginal revenue conditions, respectively, faced by the monopoly. Let us begin the analysis with short-run case. The short-run average and marginal cost conditions are shown by *SAC* and *SMC* curves. The *SMC* and *MR* curves intersect at point *A* determining the short-run, equilibrium output at OQ_1 and price at P_1Q_1. The firm is making super-normal profit. Given the demand curve, $AR = D$, the firm has a wide scope for expansion of production. Therefore, the firm adds new plants which results in decrease in long-run production cost. The *LAC* and *LMC* curves show the long-term cost conditions. Given the revenue and cost conditions as shown in the figure, the point of intersection between *LMC* and *MR* curves determine the equilibrium output at OQ_2. Given the *AR* curve, price is determined at P_2Q_2. Thus the long-run equilibrium output is OQ_2 and equilibrium price is P_2Q_2. This price-output combination maximises the monopolist's long-run profits. The total long-run profit has been shown by the area $LMSP_2$. Here the monopoly firm is in the long-run equilibrium.
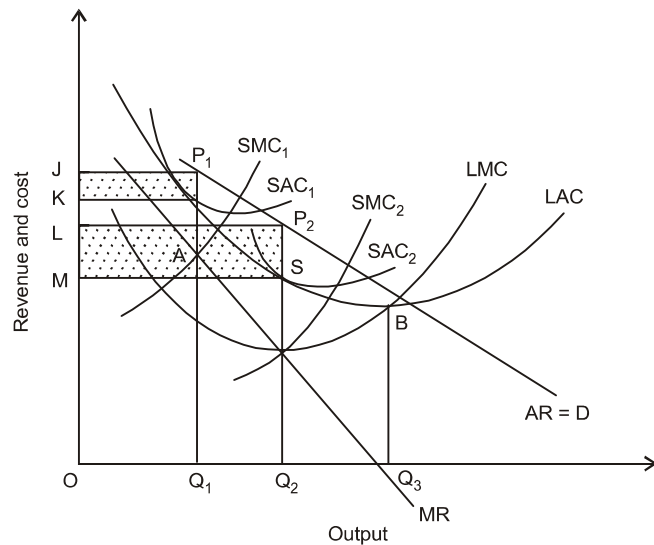
Fig. 13.6 *Long-run Equilibrium of the Monoply Firm*

Capacity Utilisation Under Monopoly

Underutilization of capacity by the monopolies, in general, is an important issue pertaining to the social cost of monopolies. Underutilisation of capacity is measured by the difference between the optimum production capacity of the firm and the capacity actually used. The optimum capacity of a firm is given by its minimum *LAC*. A firm producing less than its optimum capacity is underutilizing its capacity and a firm producing more than its optimum capacity is overutilizing its capacity.

The question that we will answer in this section is: **Does a monopoly firm always produce at less than its optimum capacity?** The answer is: not necessarily—it depends on the cost and revenue conditions of the firm and the market size.

(a) Underutilization of Capacity: The case of capacity underutilization by the monopoly firm can be seen in Fig. 13.6. The optimum size of the monopoly firm is given by point B where *LAC* is at its minimum and *LAC* = *LMC*. At this point, the optimum output is OQ_3. But, a profit maximising monopoly firm uses capacity only upto OQ_2 which is less than its optimum capacity. There is, thus, underutilization of capacity by the monopoly firm to the extent of Q_2Q_3.

(b) Overutilisation of Capacity: If size of the market and cost conditions permit, the monopoly firm may not only use its capacity to its optimum level, but may even exceed the optimum size of the plant and overutilise its long-run capacity. Figure 13.7 illustrates the case of more-than-optimal utilisation of the firm's production capacity. This is a case of over-utilisation of the capacity. The optimum size of the firm's production is given at point *B*, the point of intersection between *LAC* and *LMC,* whereas the monopoly firm chooses output at *M* where its profit is maximum.
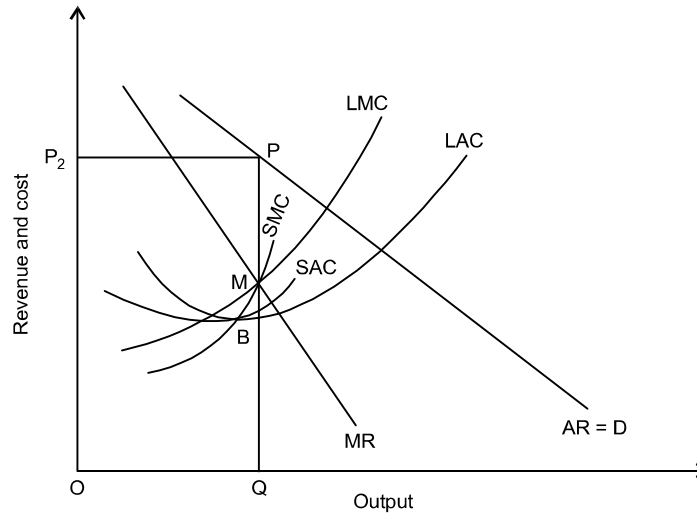
Fig. 13.7 *Monopoly Equilibrium: Overutilisation of Capacity*
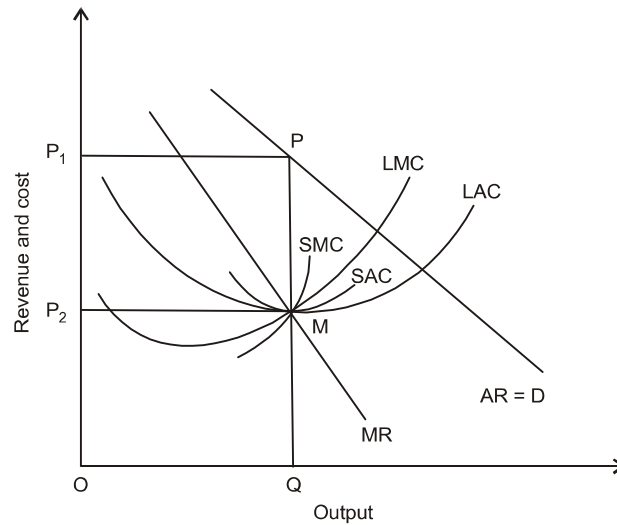
Fig. 13.8 *Monopoly Equilibrium at Optimal Size of the Plant*

Alternatively, the monopoly firm may find its equilibrium just at its optimum size. This is possible only when the market-size is just large enough to permit optimisation or optimum utilisation of the firm's capacity. This possibility has been illustrated in Fig. 13.8. This may also happen as a matter of coincidence and when production is accurately planned, given the revenue curves.

Equilibrium of Multiplant Monopoly

We have so far analysed the equilibrium of a single-plant monopoly firm. In this section, we explain the case of a monopoly firm producing a homogeneous product in more than one plant. For the sake of simplicity, we assume (*i*) that a monopoly firm has two plants *A* and *B;* (*ii*) that its cost conditions differ from one plant to another; and (*iii*) that the firm is aware of its *AR* and *MR* curves. Let us now see, how a profit maximising monopoly determines its total output and

how it allocates the total output between the two plants so that output of each unit is optimum.
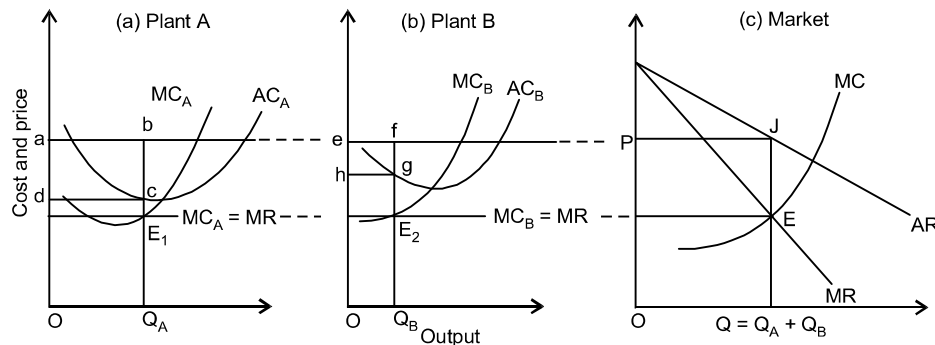
Fig. 13.9 *Multiplant Monopoly: Output Allocation*

Let the cost conditions in the two plants be given as shown in Fig. 13.9 (a) and (b). The total marginal cost, *MC,* can be obtained by horizontally summing the MC_A and MC_B curves so that $MC = MC_A + MC_B$. The *MC* curve is shown in panel (*c*) of Fig. 13.9. When *MR* and *MC* are known, the monopolist can easily find the profit maximising output and price. The profit maximising output would be *OQ* as shown in panel (c) of Fig. 13.9. Now the problem is how to allocate *OQ* between the two plants so that output of each plant is optimum. To optimise the output in each plant, the following marginal rule must be satisfied.

$$MR = MC_A = MC_B$$

The profit maximising output of each plant that satisfies this rule can be obtained by drawing a horizontal line $MR = MC$ (parallel to X-axis) from point *E* through MC_B and MC_A. As shown in Fig. 13.9, the points of intersection E_1 and E_2 determine the optimum output for plants *A* and *B*, respectively. Note that at point E_2, $MC_B = MR$ and at point E_1, $MC_A = MR$. Thus, the optimum output is determined at OQ_A for plant *A* and OQ_B for plant *B*.

Turning to the question of price, the profit maximisation price is determined at $OP (= JQ)$. At price *OP,* total profit from plant *A* is *abcd* $(= ad \times OQ_A)$ and from plant *B*, it is *efgh* $(= eh \times OQ_B)$. Although this analysis is based on only two plants, it can be extended to a larger number of plants.

The Long-Run Adjustments: In the long run, a multiplant monopoly firm adjusts the size and number of plants so as to maximise its long-run profits. The monopolist makes long-run adjustments when the existing size of the minimum-cost-plant is smaller compared to the size of the market, and there exist economies of scale. If such conditions do exist, the monopolist would adjust the size of each plant in the long-run so that the minimum of *SAC* coincides with the minimum of *LAC,* as shown in Fig. 13.10 (a). If the monopolist does not increase the number of plants or if addition of new plants does not affect the factor prices, it will have a constant minimum cost at *PQ*¼ In that case, its *LAC = LMC.* The long-run cost conditions for a multiplant monopoly firm can then be shown by the horizontal *LAC = LMC* line. The long-run equilibrium of the multiplant monopoly firm is determined by the

intersection of *LMC* and *MR* curves. The *LMC* and *MR* curves intersect at point *P* in Fig. 13.10 (b). Thus, the firm will be in the long-run equilibrium at output $OQ = nQ'$ (where n = number of plants).
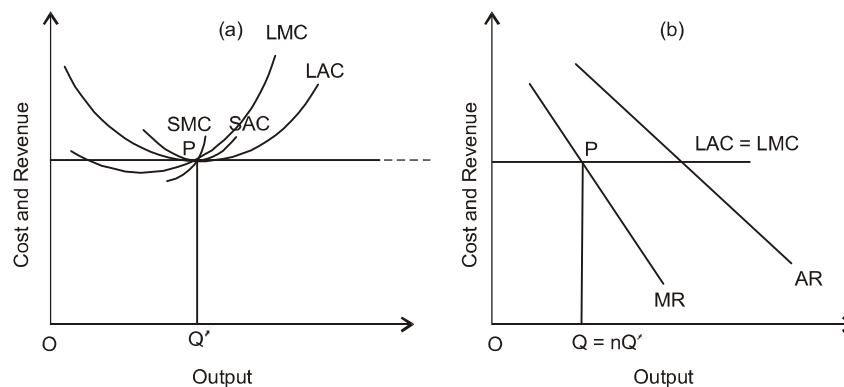
Fig. 13.10 *Long-run Equilibrium of Multiplant Monopoly*

However, if addition of new plants causes an increase in the factor prices, the cost curves for each plant would shift upward, as shown in Fig. 13.11. The monopoly firm would then be able to produce more by adding new plants only at an increasing cost. Under this condition, its *LMC* would be positively sloped as shown by *LMC* curve in output of the multiplant monopolist will be at output $OQ = nQ'$ and its equilibrium price will be *QP*.
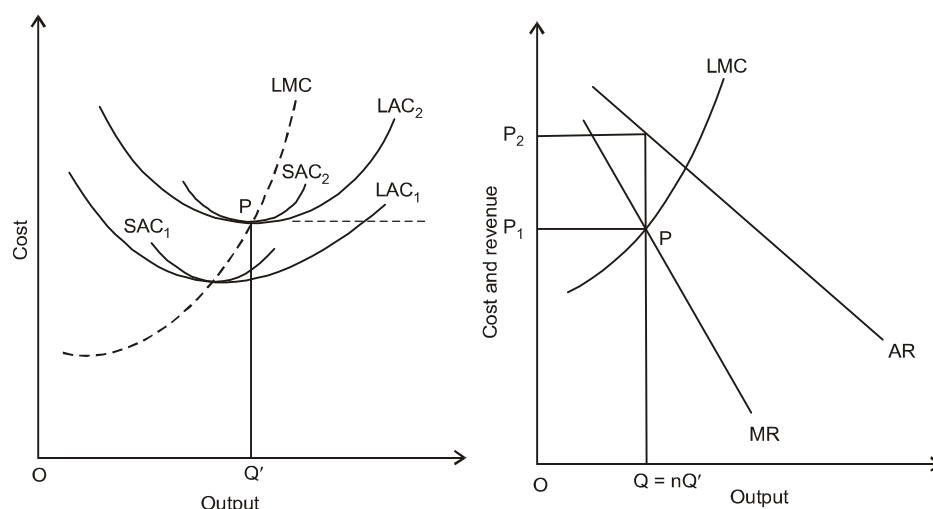
Fig. 13.11 *Long-run Equilibrium of Multiplant Firm: Increasing Costs*

13.3 PRICE DISCRIMINATION UNDER MONOPOLY

The theory of pricing under monopoly, as discussed above, gives the impression that once a monopolist fixes the price of its product, the same price will be charged from all the consumers. This is, however, not the case generally.

A monopolist, simply by virtue of its monopoly position, is capable of charging different prices from different consumers or groups of consumers. When the same (or somewhat differentiated) product is sold at different prices to different set of consumers, it is called **price discrimination**. When a monopolist sells an identical product at different prices to different buyers, it is called a **discriminatory monopoly**.

Consumers are discriminated in respect of price on the basis of their income or purchasing power, geographical location, age, sex, quantity they purchase, their association with the seller, frequency of purchase, purpose of the use of the commodity or service, and on several other grounds which the seller may find suitable.

A common example of consumers being discriminated on the basis of their incomes is found in medical professions. Consulting physicians charge different fees from different clients on the basis of their paying capacity even if quantity and quality of service rendered is the same. Price discrimination on the basis of age is found in railways, roadways and airways: children between 3 and 12 years are charged only half the adult rates. Price discrimination on the basis of quantity purchased is very cornmon. It is generally found that private businessmen charge lower price (or give discount) when bulk purchase is made. In case of public utility services, however, lower rates are charged when commodity or service is consumed in smaller quantity. For example, North Delhi Power Limited (NDPL) charges lower tariff rates on lower slabs of electricity consumption. The most common practice of price discrimination is found in entertainment business, e.g., cinema shows, musical concerts, game-shows, etc.

The product or service in question may be identical or slightly differentiated. For example, services of consulting physician and lawyer are identical. The services of railways, roadways and entertainment shows may be slightly differentiated by providing more comfortable seats, sleepers, security and airconditioning, etc. for the purpose of price discrimination. The modification in service may involve some additional cost. But price differentials are much more than is justified by cost differentials.

Although price discrimination is a common practice under monopoly, it should not mean that this practice exists only under monopoly. Price discrimination is quite common also in other kinds of market structures, particularly when market imperfection exists.

13.3.1 Necessary Conditions for Price Discrimination

As noted above, monopoly firms discriminate between consumers in regard to price charged from them. However, there are certain necessary conditions for price discrimination. The conditions are described here briefly.

First, *markets are so separated that resale is not profitable.* The market for different classes of consumers are so separated that buyers of low-price market do not find it profitable to resell the commodity in the high-price market

because of (*i*) high cost of transportation, e.g., domestic versus foreign markets; (*ii*) exclusive use of the commodity, e.g., doctor's services, entertainment shows, etc., and (*iii*) lack of distribution channels, e.g., resale of electricity.

Second, *price-elasticity of demand is different in different markets.* If market is divided into submarkets, the elasticity of demand must be different in each sub-market. It is the difference in price elasticities that provides opportunity for price discrimination. If price elasticities of demand in different markets are the same, price discrimination would not be gainful.

Third, there must be *imperfect competition in the market.* The seller must possess some monopoly power *over* the supply of the product to be able to distinguish between different classes of consumers and to charge different prices.

13.3.2 Degrees of Price Discrimination

The degree of price discrimination refers to the extent to which a monopolist can divide the market and can take advantage of market division in extracting the consumer's surplus. According to Pigou, there are three degrees of price discrimination practiced by the monopolists: (*i*) first degree price discrmination; (*ii*) second degree price discrimination; and (*iii*) third degree price discrimination.

(i) First degreeprice discrimination. The discriminatory pricing that aims at taking away the entire consumer surplus is called **first degree price discrimination**. The first degree discrimination is possible only when a seller is in a position to know the price each buyer is willing to pay. That is, the monopolist knows buyer's demand curve for the product. In that case the seller first sets a price at the highest possible level at which all those who are willing to buy, purchase at least one unit each of the commodity. When the consumer surplus of this section of consumers is exhausted, he gradually lowers down the price so that the consumer surplus of the users of the subsequent units can be extracted. This procedure is continued until the whole consumer surplus available at the price where $MR = MC$ is extracted. Also, consider the case of services of exclusive use, e.g., medical services. A doctor who knows or can guess the paying capacity of his patients, can charge the highest possible fee from visibly the rich patients and the lowest fee from the poorest one. The first degree of price discrimination is the limit of discriminatory pricing.

(ii) Second degree price discrimination. Under the second degree of discriminatory pricing, consumers are classified under different categories and a different price is charged from the different categories of consumer, e.g., consumers belonging to high, middle and low income categories. The second degree price discri-mination is also called 'block pricing system'. A monopolist adopting the **second degree price discrimination** intends to siphon off only the major part of the consumer surplus, rather than the entire of it.

The second degree price discrimination is feasible where:

   (*i*) the number of consumers is large and price rationing can be effective, as in case of utilities like electricity and natural gas and also consumer durabilities;

(*ii*) demand curves of all the consumers are identical; and

(*iii*) a single rate is applicable for a large number of buyers.
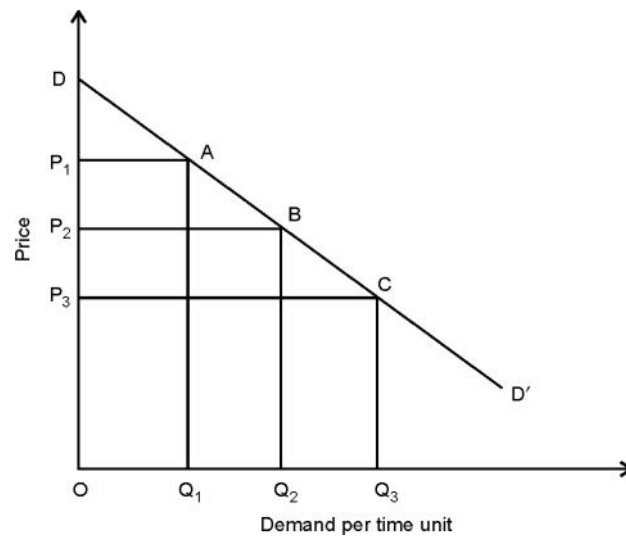
Fig. 13.12 *Second Degree Price Discrimination*

The second degree price discrimination is illustrated in Fig. 13.12. The monopolist fixes the price of its product first at OP_1 for the high income class of consumers and sells OQ_1 quantity. Once this market segment is exploited, the monopolist reduces the price to OP_2 and sells Q_1Q_2 to the second category of consumers, say, the middle income class. Finally, it cuts down the price to OP_3 and ex-ploits the third category of consumers. This process is used over a period of time. Thus, by adopting a **block-pricing system,** the monopolist maximises his total revenues (*TR*) as

$$TR = (OQ_1 \cdot AQ_1) + (Q_1Q_2 \cdot BQ_2) + (Q_2Q_3 \cdot CQ_3)$$

If a monopolist is restrained from price discrimination and is forced to choose anyone of the three prices, OP_1, OP_2 or OP_3 his total revenue will be much less.

(iii) Third degree price discrimination. When a profit maximising monopoly firm sets different prices in different markets having demand curves with different elasticities, it is using third degree price discrimination. A monopolist is often faced with two or more markets, completely separated from each other—each having a demand curve with different elasticity. Therefore, a uniform price cannot be set for all the markets without loosing the possible profits. The monopolist is, therefore, required to allocate total output between the different markets so that profit can be maximised in each market. Profit in each market would be maxmum only when $MR = MC$ in each market. The monopolist, therefore, allocates its total output between the markets in such proportions that in all the market his $MR = MC$.

The process of output allocation and determination of price for different markets is illustrated in Fig. 13.13. Suppose that a monopolist has to sell goods in only two markets, *A* and B. The two markets are so separatd that resale of

commodity is not feasible. The demand curve (D) and marginal revenue curve (MR) given in Fig. 13.13 (a) represent the AR and MR curves in market A and the curves D_b and MR_b in Fig. 13.13 (b) represent the AR and MR curves, respectively, in market B. The horizontal summation of demand curves D_a and D_b gives the total demand curve for the two markets, as shown by the curve $AR = D$, and the horizontal summation of MR_a and MR_b is given by the curve MR in Fig. 13.13 (c). The firm's marginal cost is shown by MC which intersects the aggregated MR at point T. Thus, optimum level of output for the firm is determined at OQ. The whole of OQ cannot be profitably sold in anyone market because of their limited size.
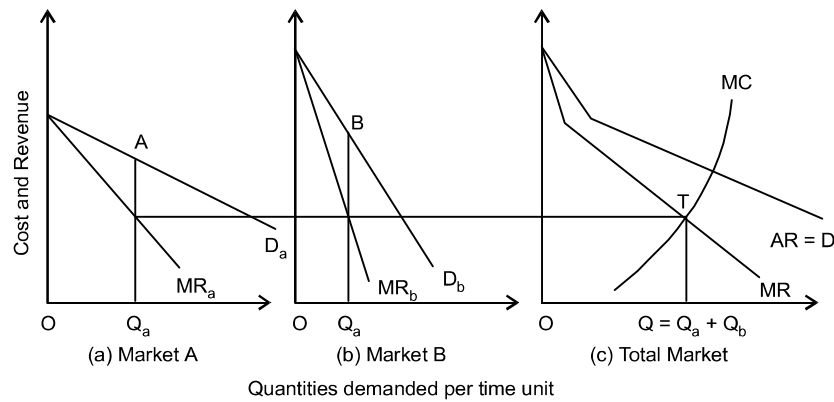
Fig. 13.13 *Third Degree Price Discrimination*

Therefore, the monopolist would allocate output OQ between the two markets in such proportions that the necessary condition of profit maximisation is satisfied in both the markets (i.e., $MC = MR$). The profit maximising output for each market can be obtained by drawing a line from point T, parallel to X-axis, through MR_b and MR_a. The points of intersection on curved MR_a and MR_b determine the optimum share for each market. As shown in Fig. 13.13, the monopolist maximises profit in market A by selling OQ_a units at price AQ_a and by selling OQ_b units in market B at price BQ_b. The firm's total equilibrium output is $OQ = OQ_a + OQ_b$.

Since at OQ_a, $MR_a = MC$ in market A, and at OQ_b, $MR_b = MC$ in market B,

$$MC = TQ = MR_a = MR_b$$

Thus, the profit maximizing condition is satisfied in both the sub-markets and the monopoly firm adopting the *third degree* method of price discrimination maximises its profits.

The third degree method of price discrimination is most suitable where the total market is divided between the home and foreign markets. This practice is called **dumping**—not necessarily by monopoly firms. However, it may be suitably practised between any two or more markets separated from each other by any two or more of such factors as geographical distance, transport barriers, cost of transportation, legal restrictions on the inter-regional or interstate transfer of commodities by individuals, etc.

13.3.3 Social Desirability of Price Discrimination

Price discrimination has had a bad reputation and condemmed as illegal and immoral. The objection is: Why charge a higher price from some consumer's and a lower price from others while there is no extra advantage to those who pay a higher price, or why benefit some at the cost of the some others? In the United Kingdom and the United States, railways were prohibited to charge discriminatory rates. Discriminatory pricing has also been criticised as a destructive tool in the hands of monopoly. For, in the past, large corporations had sought to use price discrimination to prevent the growth of competition. Besides, price discrimination may cause mal-allocation of resources and, hence, may be a deterrent to social welfare.

This is, however, not the case always. In some cases price discrimination is socially advantageous. In fact, as Lipsey has observed. "Whether an individual judges price discrimination to be good or bad is likely to depend upon the details of the case as well as upon his own personal value judgements." He adds, "Certainly there is nothing in economic theory to suggest that price discrimination is always in some sense worse than non-discrimination under conditions of monopoly or oligopoly."

Price discrimination is however considered to be *desirable* when it adds to total social welfare. Price discrimination is generally justified on the following grounds.

First, price discrimination is socially desirable in case of goods and services which are essential for the society as a whole but their production is unprofitable because long-run average cost curve (*LAC*) lies much above the aggregated market demand curve. In that case, such goods and services cannot be produced. But, production of such goods and services can be possible provided price discrimination is permitted. Price discrimination, thus, becomes essential for the survival of the industry, on the one hand, and availability of such goods and services, on the other.
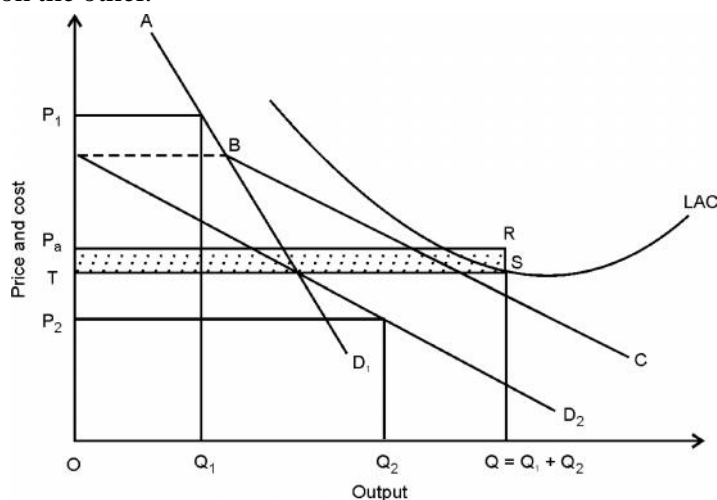
Fig. 13.14 *Price Discrimination for Industry's Survival*

The desirability of price discrimination is illustrated in Fig. 13.14. Suppose (*i*) there are two markets, I and II with their individual demand curves given as D_1 and D_2, respectively, (*ii*) market demand curve is given by *ABC,* and (*iii*) the long-run average cost curve is given by curve *LAC*. Note that *LAC* lies throughout above the total demand curve *ABC.* Therefore, a profitable production is not possible if a price along the market demand curve *ABC* is to be charged from all its consumers. However, if price discrimination is adopted and prices are so charged in the two markets that the total revenue exceeds *LAC* at some level of output, then monopoly may profitably survive to the advantage of the society. Let us suppose that the monopoly firm sets price OP_1 in the market I in which demand is less elastic and price OP_2 in market II in which demand is highly elastic. It would sell OQ_1 units at price OP_1 in market I and OQ_2 at price OP_2 in market II. The total output would then be $OQ = OQ_1 + OQ_2$. His total revenue (*TR*) would be

$$TR = (OP_1 \times OQ_1) + (OP_2 \times OQ_2)$$

and
$$AR = \frac{(OP_1 \ Î \ OQ_1 < OP_2 \ Î \ OP_2)}{OQ}$$

Suppose *AR* is estimated to be OP_a as shown in Fig. 13.14.

At output *OQ,* the *LAC* is $OT = QS$. Thus, the total cost,
$$TC = OQ \times OT = OQST$$

and its total revenue,
$$TR = OQ \times OP_a = OQRP_a$$

Since $OQRP_a > OQST$, the monopoly firm not only covers its cost but also makes a pure profit. Its total profit (a) can be expressed as
$$a = OQRP_a - OQST = P_aRST$$

This kind of situation arises mostly in public utility services like railways, roadways, post and telegraph services, etc., in which high-paying sections of the market subsidise the low-paying sections. In other words, rich subsidise the consumption of the poor.

Secondly, discriminatory pricing can be adopted with justification where a uniform, single profitable price is likely to restrict the output and deprive many (particularly the people of lower income groups) of the essential good or service. For example, if specialist doctors in private practice, who charge a discriminatory price for their services, are asked to charge a uniform fee from all the patients, they would charge a fee high enough to maintain the level of their income. But the high fee deprives the poor of the doctor's services and may force them to opt for an inferior or inadequate treatment. This leads to a socially undesirable situation because, on the one hand, rich patients who could pay a higher fee pay a price lower than what they could afford, and could, on the other hand, poor patients are deprived of proper medical care. What is worse, doctor's services remain underutilized.

Thirdly, there is a section of consumers which gains more than the people of other sections from the use of a product. For example, factory-owners gain more from the use of electricity than the households; irrigation projects benefit big farmers more than the marginal farmers; motorists benefit from roads more than pedestrians, and so on. In such cases, uniform price would be unjustified from a normative point of view, provided the objective is not to restrain the household consumption of certain goods like electricity and spare it for productive purposes. There is, on the other hand, full justification for discriminatory pricing of electricity.

13.3.4 Monopoly Vs. Perfect Competition

This section makes a comparative analysis of monopoly and perfect competition with respect to (*i*) price and output, and (*ii*) loss and gain to the society with the purpose of examining the desirability of monopoly.

Long-Run Price and Output

This section answers the question whether perfect competition or monopoly is more desirable from society's point of view. This question can be answered by comparing the price and output in the two kinds of markets. Comparison of long-run price and output is more justified than those of short-run because long-run equilibrium is settled and shows the final status of the firms, whereas short-run equilibrium subject to change making comparison uncertain. Therefore, we compare here long-run price and output under monopoly and under perfect competition.

Figure 13.15 presents a comparative analysis of equilibrium price and output under perfect competition and monopoly in the long-run. Let us assume that LMC and LAC curves are identical for both a competitive industry and a monopoly.
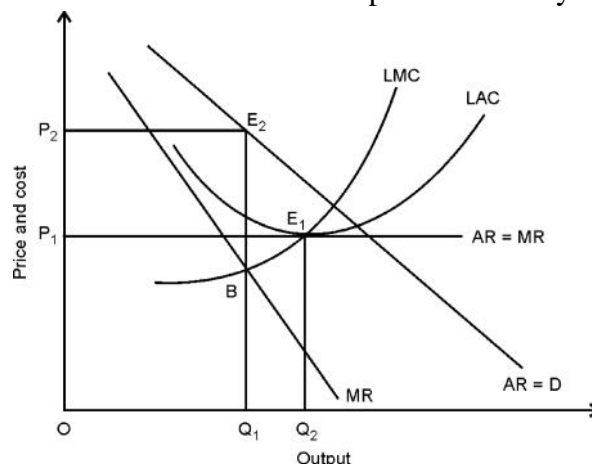
Fig. 13.15 Comparision of Price and Output: Monopoly Vs. Competitive Industry

The long-run equilibrium of a competitive industry requires that all its firms are in equilibrium. That is, all the firms have their $AR = MR = LAC = LMC$. This condition is satisfied at point E_1 in Fig. 13.15. Thus, in a competitive industry,

equilibrium price is OP_1 and equilibrium output OQ_2. Now if this industry were to be monopolised, the revenue conditions (AR and MR) will change from horizontal one to be downward sloping ones as shown by $AR = D$ and MR curves. The monopoly firm will reach its equilibrium and maximise its total profits at the level of output where $MR = MC$. This level of output is indicated by point B on the downward sloping MR curve. Thus, the equilibrium output under monopoly will be OQ_1 and the equilibrium price will be OP_2. The following table presents a comparison of the equlibrium price and output under perfect competition and monopoly.

Price/Output	Perfect Competition	Monopoly	Comparison
Price	OP_1	OP_2	$OP_1 < OP_2$
Output	OQ_2	OQ_1	$OQ_2 > OQ_1$

Conclusion. If both monopoly and competitive industries are faced with identical cost conditions, the output under perfectly competitive conditions is higher than under monopoly since $OQ_2 > OQ_1$, and price in the competitive industry is lower than in monopoly as $OP_1 < OP_2$. Perfect competition is, therefore, more desirable from social welfare angle, under the given conditions.

The Dead-Weight Loss Under Monopoly

On the basis of the fore-going conclusion, it can be said that monopoly firms are less efficient than perfectly competitive firms. Another and a very serious disadvantage of monopoly is that it causes loss of social welfare and distortions in resource allocation. The suboptimal allocation of resources and loss of social welfare are illustrated in Fig. 13.16, assuming a con-stant-cost industry whose cost conditions are shown by the line $LAC = LMC$. The revenue con-ditions of the industry are shown by the AR and MR curves. Given the cost and revenue conditions, a perfectly competitive industry will produce OQ_2 at which is $LAC = LMC = AR$. Its price will be OP_1.
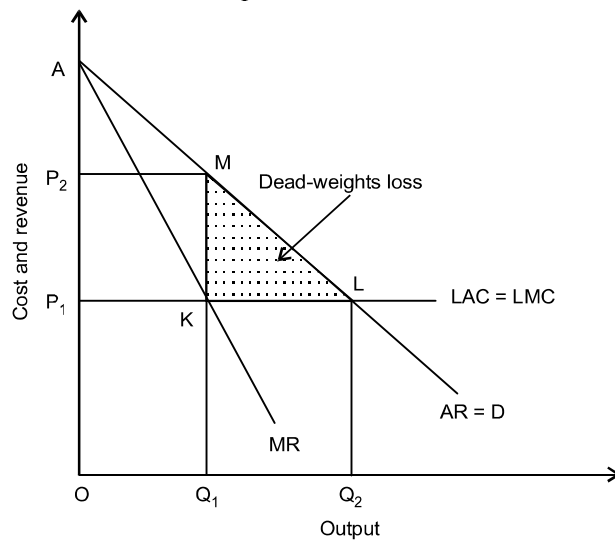
Fig. 13.16 *Price and Output under Monopoly and Perfect Competition*

On the other hand, the monopoly firm produces an output that equalises its *LMC* and *MR.* Thus, monopoly firm produces OQ_1 and charges price OP_2. The comparison of prices and outputs under monopoly and perfect competition stands as follows.

Price and Output Under Pefect Competition and Monopoly

Variable	Monopoly	Competitive Industry	Comparison
Output	OQ_1	OQ_2	$OQ_1 < OQ_2$
Price	OP_2	OP_1	$OP_2 > OP_1$

The Dead-Weight Loss: The dead-weight loss is measured in terms of loss of consumer surplus at no body's gain. The total consumer surplus equals the difference between the total utility which a society derives from the consumption of a commodity and the total price that it pays for that commodity. If an industry is perfectly competitive, the total output available to the society will be OQ_2 at price OP_1 (see Fig. 13.16). The total price which the society pays for OQ_2 is given by the area $OP_1LO_2 = OP_1 \times OQ_2$. The total utility which it gains from the output OQ_2 is given by the area $OALQ_2$ which, in Marshallian terminology, is the value which society would be willing to pay for output OQ_2. Thus,

Consumer's surplus $= OALQ_2 - OP_1LQ_2 = ALP_1$

If the industry is monopolised, the equilibrium output is set at OQ_1 and price at OP_2. This leads to a loss of a part of consumer surplus.

Loss of consumer surplus under monopoly $= ALP_1 - AMP_2 = P_2MLP_1$.

Of this total loss of consumer surplus (P_2MLP_1), P_2MKP_1 goes to the monopolist as monopoly profit or pure profit. The remainder $MKL = P_2MLP_1 - P_2MKP_1$ goes to none. Therefore, it is termed as **dead-weight loss** to the society. The dead-weight loss is taken to be the loss of social welfare due to monopoly.

Empirical Evidence of Dead-Weight Loss: Theoretically, there is a strong possibility of a dead-weight loss under monopoly in comparison to perfect competition. However, empirical evidence of the dead-weight loss attributed to monopoly is inconclusive. Harberger was the first to estimate the loss of welfare— in fact, loss of efficiency—due to monopoly in the United States. To estimate the loss of welfare, he assumed constant cost and unitary elastic demand curve for all industries and used 1924–1928 data of 73 manufacturing industries. He regarded the loss of welfare as the difference between the monopoly price and the competitive price. According to his estimates, the US economy had suffered a loss of social welfare to the extent of 0.1 per cent of its national income, which was obviously negligible. An implication of this finding was that efforts to curb monopolies was a sheer waste of time. Interest-ingly, "One economists quipped that economists might make a larger social contri-bution fighting fires and eradicating termites than attempting to curb monopolies."

Although many other studies confirmed Harberger's finding, some economists were highly critical of his data and approach. For example, Stigler criticized Harberger's results on the ground of statistics used in his study. Kamerschen used 1956–61 data in a similar study and produced a considerably different result. According to his findings, the welfare loss due to monopolies in the US was about 6 per cent of national income. Scherer and Ross find welfare loss due to monopolistic mal-allocation of resources somewhere between 0.5 and 2 per cent of the US GDP. Some other critics of Harberger have estimated the social loss to be of the order of 4-8 per cent of national income. However, the methodology and statistics used by Harberger's critics have been criticized by others. In the mean time, some other aspects of monopoly have been added to the debate, e.g., the dynamic efficiency that arises due to technological advanced made by monopolistic competition, the poor quality of goods supplied by them, advantage of the resources unused by the monopolies, and the use and abuse of monopoly profits. Finally, it may be concluded that the issue of dead-weight loss attributed to monopolies remains a controversial issue. Some even argue that the *MRTP* Acts adopted by many countries is a futile exercise.

13.4 APPLICATION OF MONOPOLY THEORY

In this section, we will discuss *two important applications of* the theory of monopoly. The topics that we will discuss here include:

(*i*) Incidence of commodity tax and effect of subsidy under monopoly; and

(*ii*) Peak load pricing under monopoly.

Measuring Tax Incidence and Subsidy Effect

The question that we will answer theoretically is: Who bears the tax burden when a tax is imposed on monopoly product and on monopoly profit? Let us first answer this question in respect of the *commodity tax.*

(i) Incidence of a Commodity Tax: A commodity tax may be in the form of a *sales tax* or *excise duty.* These taxes may be imposed either at a specific rate (a fixed tax per unit of output) or at an *ad valorem rate,* i.e., at some percentage of the value of the commodity. In our analysis, we will assume a *specific* commodity tax.

Let us suppose that cost and revenue condi-tions of a monopoly firm are given as in Fig. 13.17. Prior to the imposition of tax, the firm would be in equilibrium at point *E* where its *MC* = *MR*. Its equilibrium output is *OQ* and equilibrium price *PQ*.
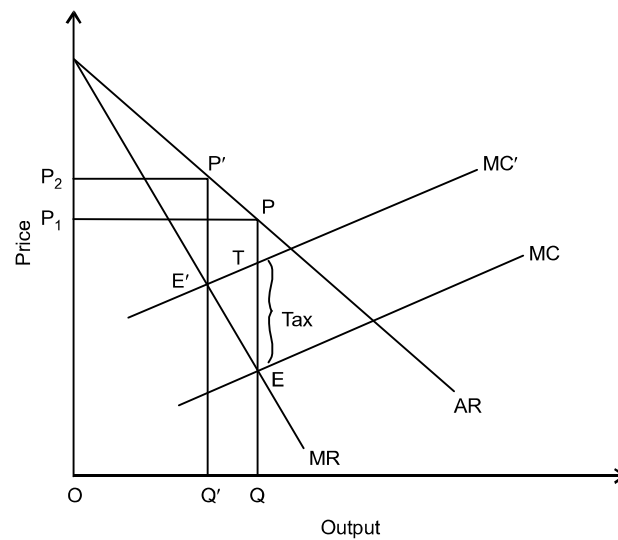
Fig. 13.17 *Effects of Specific Sales Tax on Monopoly*

Let the government now impose a specific or unit tax by an amount *ET* per unit on the monopoly product—it may be in the form of specific sales tax or excise duty per unit of output. The imposition of tax increases firm's supply price and will thereby make the *MC* curve shift upward to the position of *MC′*. As a result, firm's equilibrium shifts to point *E′*. As Fig. 13.17 shows, equilibrium output falls to *OQ′* and price rises to *P′Q′*. Note that price increases only by P_1P_2 which is less than the amount of tax *ET*. This conclusion holds even if a monopoly firm is producing at constant *MC*.

This analysis shows that the effect of commodity tax on monopoly is similar to one applicable to a competitive industry. That is, a monopoly firm is able to pass only a part of the tax on to the consumers. It can, however, be shown that output under monopoly declines by less than it does under competition.

Can a Monopolist Raise Price More than the Amount of Tax?

Monopoly's power to raise its price after tax imposition depends on the elasticity of demand for its product—the lower the elasticity, the greater the power. A monopolist can pass the entire tax burden on to the consumer when the demand for its product is perfectly inelastic. For example, consider essential goods like life-saving drugs, electricity, telephone, cooking gas, etc. Such goods have nearly zero elasticity ($e_p \simeq 0$). In case of such goods, a monopoly may raise price by the entire amount of tax—even by more than that. But, in case of products with downward sloping demand curve with $0 < e < \infty$, monopoly's power to raise price depends on the elasticity of demand as follows.

  (*i*) If $e = 1$, monopoly can raise price by half of the tax.

 (*ii*) If $e < 1$, price can be raised by more than half of the tax.

(*iii*) If $e > 1$, price can be raised only by less than half of the tax.

These conditions hold for a profit maximising monopoly only. It is important to note here that in the absence of government control of monopoly prices, however, a monopoly may raise price by the entire amount of tax—even by more than the amount of tax, if demand is highly inelastic.

(ii) Grant of Subsidy: Subsidy is opposite of tax. It may be considered as a negative tax. Therefore, the effect of subsidy on monopoly equilibrium is the reverse of the tax effect. The effects of subsidy on monopoly output and price are illustrated in Fig. 13.18. The *AR* and *MR* curves show the revenue and *LMC* shows the long run cost conditions long-run allows the time for output and price adjustment. The analysis of effects of subsidy on monopoly equilibrium is presented below.
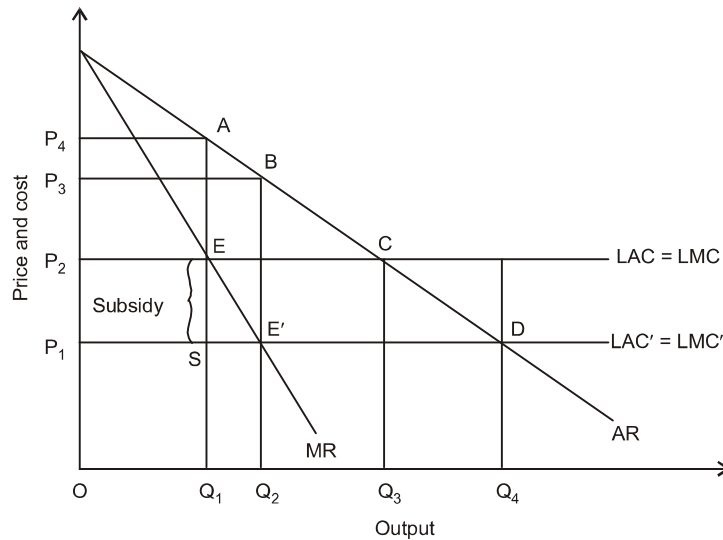
Fig. 13.18 *Subsidy and Monopoly Equilibrium*

Let the monopoly be in equilibrium at point *E* prior to the grant of subsidy. The pre-subsidy equilibrium output is OQ_1 and price is AQ_1 (= OP_4). Now, let the government provide a subsidy of an amount *ES* per unit. As a result, firm's $LAC = LMC$ shifts down to $LAC' = LMC'$ and its equilibrium shifts to point E'. At new equilibrium, monopoly output increases from OQ_1 to OQ_2 and post-subsidy price falls to BQ_2. It is important to note that monopoly price decreases by less than the amount of subsidy. It means monopolists do not pass the entire benefit of subsidy to the consumers, even if it is so intended—they retain a part to themselves.

Comparison of Subsidy under Monopoly and Competitive Industry

Figure 13.18 provides also a comparison of subsidy effect on price and output under monopoly and perfect competition. Given the revenue and cost conditions, a competitive firm will be in pre-subsidy equilibrium at point *C* where its *AR* = *LAC*. Its equilibrium output is OQ_3 and price is OP_2 (= CQ_3). After the grant of

subsidy, its equilibrium shifts to point 0 where its output OQ_4 and price is OP_1 (= DQ_4).

The effect of subsidy on the price and output of a monopoly and competitive industry is compared below assuming subsidy is granted under the same cost and revenue conditions.

(*i*) A competitive firm passes the entire benefit of subsidy to the consumers while a monopoly firms retains a part of it.

(*ii*) Increase in output is much larger under competitive conditions.

(*iii*) Decrease in monopoly price (P_4P_3) is less than (P_2P_1) under perfect competition.

Peak-load Pricing by a Monopoly

There are certain non-storable goods, e.g., electricity, telephones services, etc., which are demanded in varying quantities in different seasons and in day and night times. For example, consumption of electricity reaches its peak in day time. It is called 'peak-load' time. It reaches its bottom in the night. This is called 'off-peak' time. Electricity consumption peaks in day time because all business establishments, offices and factories come into operation. It decreases during nights because most business establishments are closed and household consumption falls to its basic minimum. Also, in India, demand for electricity peaks during summers due to use of *ACs* and coolers, and it declines to its minimum level during winters. Similarly, consumption of telephone services is at its peak at day time and at its bottom at nights. Another example of 'peak' and 'off-peak' demand is of railway and airline services. During festivals and summer holidays, 'Pooja' vacations, the demand for railway and airline services rises to its peak and they decline the off seasons.

A technical feature of such products is that they cannot be stored. Therefore, their production has to be increased in order to meet the 'peak-load' demand and reduced to 'off-peak' level when demand decreases. Had these goods been storable, the excess production in 'off-peak' period could be stored and supplied during the 'peak-load' period. But this cannot be done. Besides, given the installed capacity, their production can be increased but at an increasing marginal cost (*MC*).

Problems in Pricing: Pricing of non-storable goods like electricity is problematic. The nature of the problem in a short-run setting is depicted in Fig. 13.19. The 'peak-load' and 'off-load' demand curves are shown by D_H and D_L curves, respectively, showing high and low demand. The short-run supply curve is given by the short-run marginal cost curve, *SMC*. The problem is 'how to price electricity'.
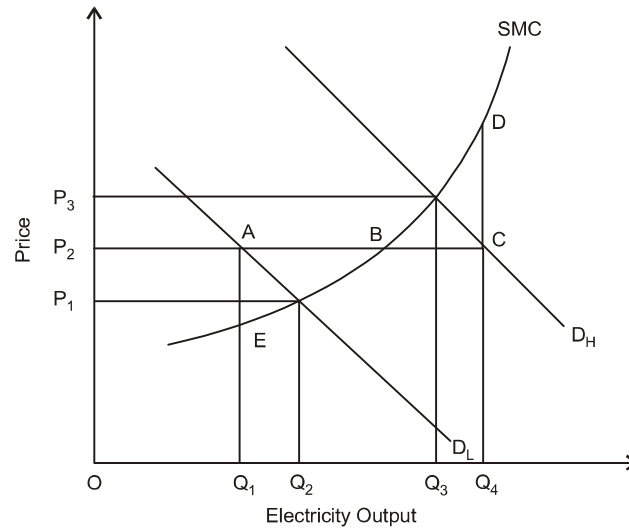
Fig. 13.19 *Peak-Load Pricing of Electricity*

As Fig. 13.19 shows, if electricity price is fixed in accordance with peak-load demand, OP_3 will be the price and if it is fixed according to off-load demand, price will be OP_1. The problem is: How to fix the price of electricity? If a 'peak-load' price (OP_3) is charged uniformly in all seasons, it will be unfair because consumers will be charged for what they do not consume. Besides, it may affect business activities adversely as it increases a variable cost of production. If electricity production is a public monopoly, the government will not allow a uniform 'peak-load' price.

On the other hand, if a uniform 'off load' price (OP_1) is charged, production will fall to OQ_2 and there will be acute shortage of electricity during peak hours. It leads to 'breakdowns' and 'loadshedding' during the peak-load period, which disrupt production and make life miserable. This is a regular feature in Delhi, the capital city of India. This is because electricity rates in Delhi are said to be one of the lowest in the country and power theft runs close to 40%—it was about 50% before 2004.

Alternatively, if an average of the two prices, say P_2 is charged, it will have the demerits of both 'peak-load' and 'off-load' prices. There will be an excess production to the extent of AB during the 'off-load' period, which will go waste as it cannot be stored. If production is restricted to OQ_1, price P_2 will be unfair. And, during the 'peak-load' period, there will be a shortage to the extent of BC, which can be produced only at an extra marginal cost of CD.

Double Pricing System: For the above reasons, generally, a double pricing system is adopted. A higher price, called 'peak-load price' (OP) is charged for the 'peak-load' period and a lower price (OP_1) is charged for the 'offpeak' period. During the 'peak-load' period, production is increased to OQ_3 at which D_H intersects *SMC,* and production is reduced to OQ_1 during the 'off-peak' period.

Advantages: Peak-load pricing system has two advantages.

(*i*) It results in an efficient distribution of electricity consumption. Housewives run their dishwashers and washing machines during the 'off-peak' period.

(*ii*) It helps in preventing a loss to the electricity company and ensures regular supply of electricity in the long-run.

Disadvantages: This system has two disadvantages too.

(*i*) The businesses which are by nature day-business pay higher rates than those which can be shifted to 'off-peak' period.

(*ii*) Billing system is the greatest problem. Each consumer will have to install two meters—one for 'peak-load' and another for 'off-load' period with an automatic switch-over system. This can be done.

Alternatively, the problem can be resolved by adopting a progressive tariff rate for the use of electricity as is the case in Delhi. Airlines adopt a off-season discounting system of pricing.

13.4.1 Measures of Monopoly Power

The degree of monopoly power matters a great deal in pricing and output decisions of a monopolist and in respect of control and regulation of monopolies. We discuss here the various measures of monopoly power. It must be borne in mind here that the question of measuring monopoly power arises only under monopolistic competition—not in case of *pure monopoly.* In case of pure monopoly, the degree of monopoly power is 1 and it is known to the policy makers.

Measuring monopoly power has been a very difficult problem. The efforts to devise a measure of monopoly power have not yielded any universal or non-controversial measure. As Hunter has observed, "The idea of devising a measure of monopoly power, with reference both to its general incidence and to particular situation, has been and probably always will remain an attractive prospect for economists who wish to probe in this field. If not for any other reason, then for 'sheer intellectual curiosity', economic theorists feel compelled to work on this problem, for they could not with good conscience go on talking about 'great' or 'little' monopoly power or about various degrees of monopoly without trying to ascertain the meaning of these words.

Therefore, to devise at least a 'conceivable' measure of monopoly, even if 'practical' measurement is impossible, continues to interest the economists, for at least *two reasons.*

First, apart from intellectual curiosity, people would like to know about the economy in which they live, its industrial structure, and the industries from which they get their supplies.

Second, growth of monopolies have forced governments of many countries of formulate policies and devise legislative measures to control and regulate monopolies. If the government is to succeed in its policy of restraining monopoly,

it must have at least some practicable measure of monopoly power and monopolistic trade practices.

13.4.2 Measuring Monopoly Power

The economists have suggested several devices to measure the degree of monopoly power, though none of the measures is free from flaws. Yet, the various measures suggested by the economists do provide an insight into the monopoly power and its impact on the market structure. Besides, they also help in formulating an appropriate public policy to control and regulate the existing monopolies. We have briefly discussed here some important measures of monopoly power.

(a) Number-of-firms criterion. One of the simplest measures of degree of monopoly power is to count the number of firms in an industry. The smaller the number of firms, the greater the degree of monopoly power of each firm in the industry, and conversely, the larger the number of firms, the greater the possibility of absence of monopoly power. A corollary of this criterion is that if there is a single firm in an industry, the firm has an absolute monopoly power. This criterion seems to have been derived from the characteristics of the perfect competition in which the number of firms is so large that no firm has any monopoly power.

This criterion has, however, a serious *drawback.* The number of firms alone does not reveal much about the relative position of the firms within the industry because (*i*) 'firms are not of equal size' and (*ii*) their number does not indicate the degree of control each firm exercises in the industry if an industry with a large number of firms is dominated by a large firm or by a few large firms. Therefore, the numerical criterion of measuring monopoly power is of little practical use.

(b) Concentration ratio. The **concentration ratio** is one of the widely used criterion for measuring monopoly power. The concentration ratio is obtained by calculating the percentage share of the largest group of the firms in the total output of an industry. "The number of firms chosen for calculating the ratio usually depends on some fortuitous element—normally the census of production arrangements of the country concerned." In Britain, the share of the largest three firms of a census industry, and in the USA, the share of the largest four firms is the basis of calculating concentration ratio. Apart from the share of the largest firms in the industry-output, "the size of the firms and the concentration of control in the industry may be measured ... in terms of production capacity, value of assets, number of employees or some other characteristics."

These measures too are, however, not free from drawbacks. They have *three major drawbacks.* **First**, they involve statistical and conceptual problems. For example, production capacity may not be used stragihtaway as it may inclue 'unused obsolete or excess capacity' and the value of assets invovles valuation problem as accounting method of valuation and market valuation of assets may differ. Employment figure may not be relevant in case of capital-intensive industries. The use of such figures may be misleading. The two other convenient measures

are 'gross output value' or 'net output' (value added). But the former involves the risk of double counting and the later, the omission of inter-establishment transfers.

Second, an important objection to these measures of degree of monopoly power is that these measures do not take into account the size of the market. Size of the market may be national or local. A large number of firms supplying the national market may be much less competitive than the small number of firms supplying the local market. For, it is quite likely that the national market is divided among thousand sellers so that each seller has the status of a monopolist in his own area.

Third, the most serious defect of concentration ratio as an index of monopoly power is that it does not reflect the competition from other industries. The degree of competition is measured by the elasticity of substitution between the products of different industries. The elasticity of substitution may be diferent under different classification of industries. Therefore, an industry with concentration ratio under one classification of industries may have a very low elasticity of substitution and hence a high degree of monopoly. But, if classification of industries is altered, the same industry with a high concentration ratio may have a very low elasticity of substitution, and hence may show a low degree of monopoly.

(c) Excess profitability criterion. J.S. Bain and, following him, many other economists have used *excess profit* as a measure of monopoly power. If profit rate of a firm continues to remain sufficiently higher than all opportunity costs required to remain in the industry, it implies that neither competition among sellers nor entry of new firms prevents the firm from making a pure or monopoly profit. While calculating the excess profit, the opportunity cost of owner's capital and margin for the risk must be deducted from the actual profit made by the firm. Assuming no risk, the degree of monopoly may be obtained by calculating the divergence between the opportunity costs (O) and the actual profit (P) as ($P - O$)/P. If [($P - O$)/P] = 0, there exists no monopoly, and if [($P - O$)/P] > 0, there is monopoly. The higher the value of ($P - O$)/P, the greater the degree of monopoly.

(d) Lerner's index of monopoly power. Another measure of degree of monopoly based on excess profitability has been suggested by A.P. Lerner. According to Lerner, for a competitive firm, equilibrium price equals marginal cost but for a monopoly price exceeds MC. Therefore, a logical measure of monopoly power is the ratio of $P - MC$ to P, where P is equilibrium price. Thus, the degree of monopoly power (M^P) may be measured as

$$M^P = \frac{P - MC}{P}$$

where P = price, MC = marginal cost.

Since for a profit maximising firm, $MR = MC$, Lerner's measure of monopoly power may also be expressed as,

$$M^P = \frac{P - MR}{P}$$

Since $P/(P - MR) = e$ (elasticity), $(P - MR)/P = 1/e$. It means that M^P equals the reciprocal of elasticity. Thus, Lerner's measure of monopoly power may also be expressed as $M^P = 1/e$. It may thus be inferred that lower the elasticity, the greater the degree of monopoly, and *vice versa.* According to Lerner's formula, monopoly power may exist even if firm's $AR = AC$ and it earns only normal profit.

Lerner's formula of measuring the degree of monopoly power is considered to be theoretical most sound. Nevertheless, it has been criticised on the following grounds.

First, it is suggested that any formula devised to measure degree of monopoly power should bring out the difference between the monopoly output and competitive output or the 'ideal' output under the optimum allocation of resources. The divergence between *P and MC used in Lerner's formula does not indicate the divergence between the monopoly and the* 'ideal' output. "This substitution of a price-cost discrepancy for a difference between actual and 'ideal' output is probably the greatest weakness of formula which is supposed to measure deviation from the optimum allocation of resources."

Second, price-cost discrepancy may arise for reasons other than monopoly, and price and cost may be equal or close to each other in spite of monopoly power.

Third, since data on *MC* are hardly available, this formula is of little practical use.

(e) Triffin's cross-elasticity criterion. Triffin's criterion is said to have been derived from the definition of monopoly itself—monopolist is a single seller of a product without close substitutes. According to Triffin's criterion, cross-elasticity is taken as the measure of degree of monopoly—the lower the cross-elasticity of the procluct of a firm, the greater the degree of its monopoly power. But, this criterion is based on the inter-relationships between the individual firms and indicates only the relative power of each firm—not the measure of its absolute power. It does not furnish a single index of monopoly power.

13.4.3 Government Regulation of Monopoly Prices

The monopolies are, in general, alleged to restrict production, consumption and employment, widen income and wealth disparities, exploit consumers and employees, distort resource allocation and reduce the social welfare. In most countries, therefore, government intervenes and makes laws to control and regulate monopolies to the advantage of the society. There are various measures—direct, indirect, price, non-price, legal and otherwise—to control and regulate the monopolies. We have discussed below price regulation of natural monopolies.

Price regulation is a common feature in case of natural monopolies. When the size of the market is small relative to the optimum size of the firm, market size cannot support more than one firm of optimal size. The monopoly in such a market is a natural monopoly, protected by market size itself. The government may either nationalise such monopolies or regulate their prices to eliminate the excess profits.

If government intends to regulate the monopoly price, the question arises: what price should be fixed for the monopolist to charge? The two controlled-price systems have been suggested: one that allows some *excess profit* to the monopolist, and the second that allows only *normal profit* to the monopolist. Both the systems of conrolled prices, alongwith their repercussion on output, are illustrated in Fig. 13.20.
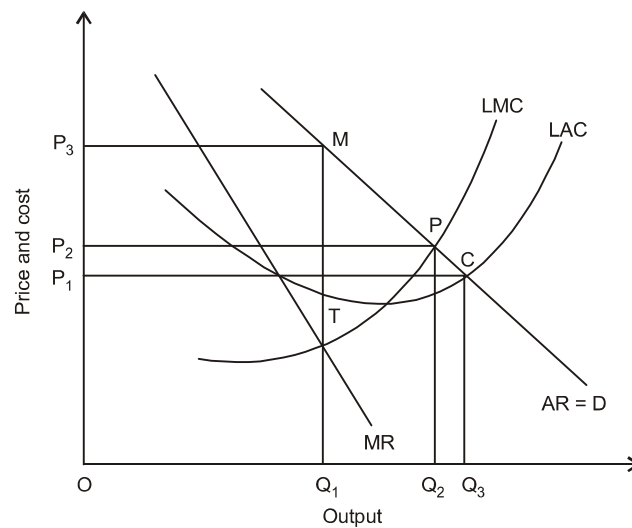
Fig. 13.20 *Government Regulated Monopoly*

An unregulated monopoly would produce OQ_1 units, charge price OP_3, and make excess profit of MT per unit. If government intends to regulate monopoly price allowing some super normal profit, then one reasonable price is $OP_2 = PQ_2$, where $LMC = AR$. Alternatively, if government intends to regulate monopoly price with only normal profit, it will fix price at $OP_1 = CQ_3$ at which $AR = LAC$. When OP_1 is the price set for the monopolist, the firm is allowed only normal profit, but the output is maximum possible under the given cost and revenue conditions. On the other hand, if price is fixed at OP_2, then the monopolist gets some excess profit, but the output is less than that at price OP_1. In both the cases, however, the total output under regulated monopoly is much greater than that under unregulated monopoly. Which of the two alternative prices (OP_1 and OP_2) is more appropriate is a matter of debate and policy.

<div style="border:1px solid black">

Check Your Progress

1. State one important feature of pure monopoly.

2. What are patent rights?

</div>

13.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. An important feature of a pure monopoly is that a monopolized industry is a single-firm industry, i.e., there is no distinction between the firm and the industry.

2. Patent rights are granted by the government to a firm to produce a commodity of specified quality and character or to use a specified technique of production.

13.6 SUMMARY

Ñ By definition, monopoly is a market situation in which there is a single seller of a commodity of 'lasting distinction' without close substitutes.

Ñ First potential source of indirect competition is the rivalry between monopoly good and other goods produced by other monopolies and competitive firms.

Ñ Finally, an important feature of a pure monopoly is that a monopolized industry is a single firm industry, i.e., there is no distinction between the firm and the industry.

Ñ The emergence and survival of monopoly is attributed to the factors which prevent the entry of other firms into the industry. The barriers to entry are, therefore, the sources of monopoly power.

Ñ According to the traditional theory of firm, a monopoly firm (or otherwise) is said to be in equilibrium where it maximises its profit. Maximisation of total profit is a matter of time required to adjust output to the price.

Ñ In general usage of the term, there are two common misconceptions about monopolies: (*a*) that monopolies always make profits, and (*b*) that they can charge any price arbitrarily.

Ñ The long-run equilibrium conditions of a monopolist differ from those faced by the competitive firms in another important respect, i.e., the entry of new firms into the industry.

Ñ When the same (or somewhat differentiated) product is sold at different prices to different set of consumers, it is called price discrimination.

Ñ Monopoly firms discriminate between consumers in regard to price charged from them. However, there are certain necessary conditions for price discrimination.

Ñ The degree of price discrimination refers to the extent to which a monopolist can divide the market and can take advantage of market division in extracting the consumer's surplus.

Ñ Monopoly's power to raise its price after tax imposition depends on the elasticity of demand for its product—the lower the elasticity, the greater the power

Ñ The degree of monopoly power matters a great deal in pricing and output decisions of a monopolist and in respect of control and regulation of monopolies.

Ñ The economists have suggested several devices to measure the degree of monopoly power, though none of the measures is free from flaws..

13.7 KEY WORDS

Ñ **Monopoly:** It is a market situation in which there is a single seller of a commodity of 'lasting distinction' without close substitutes.

Ñ **Discriminatory monopoly:** When a monopolist sells an identical product at different prices to different buyers, it is called a discriminatory monopoly.

13.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. How does existence of a close substitute affect the monopoly power? What are the sources of monopoly?

2. Why are the revenue and cost curves under monopoly different from those under perfect competition?

3. Write a note on the relationship between average revenue and marginal revenue under (*i*) perfect competition, and (*ii*) monopoly.

4. How is pricing under monopoly different from that under perfect competition? Can a monopoly firm fix any price for its product?

5. Suppose price charged by a monopolist at equilibrium is twice as high as its *MC*.What is the price-elasticity of demand?

6. What are prerequisites of price discrimination by a monopolist?

7. Write notes on the following criteria of monopoly power.

    (*a*) Concentration ratio,

    (*b*) Excess profitability criterion, and

    (*c*) Cross-elasticity criterion.

Long Answer Questions

1. Explain the equilibrium of a monopoly firm in the short-run. Why is monopoly price always higher than the competitive price?

2. For a profit maximising monopoly, price is greater than marginal cost and it remains so over a large range of output. Why does then a monopolist not produce more than an output at which its *MC =MR?*

3. A monopoly firm may earn normal or abnormal profits or may even incur losses in the short-run. Do you agree with this statement? Give reasons for your answer.

4. Will a monopolist remain in business in the short-run if it is just covering its average variable costs? Explain with the help of a diagram.

5. Compare monopoly and perfect competition with regard to the following: (*i*) price, (*ii*) output (*iii*) welfare cost, and (*iv*) relationship between*MC* and price.

6. What are the necessary conditions of price discrimination under monopoly? Show how a profit maximising discriminating monopolist allocate his output between two markets and charge different prices?

7. Explain how discriminating monopolist's equilibrium price-output configuration compares with that of a simple monopolist's equilibrium behaviour?

8. Discuss the important measures of monopoly power.

13.9 FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.

UNIT 14 PRICE AND OUTPUT DETERMINATION UNDER OLIGOPOLY

Structure

14.0 INTRODUCTION

As noted in preceding unit, perfect competition and monopoly are uncommon cases of market strucutre. During the early 20th century monopolistic competition emerged as a major kind of market structure. Over time, however, monopolistic competition too gave way to oligopoly kind of market structure. Therefore, the attention of economists was drawn towards the analysis of price and output determination under oligopolistic market structure in which there are 'a few seller of a product', although the first oligopoly model (in the form a duopoly model) was developed by a classical economist, Augustin Cournot, as early as 1838.

In this unit, we will discuss theories of price and output determination in oligopoly. A number of complex theories were developed over time to analyse price and output determination in an oligopolistic market. This, however, added more complication and confusion to the subject rather than offering a non-controversial theory. In this unit, however, we will confine to theories considered to have made significant contribution to the theory of oligopoly. We begin our discussion with meaning and characteristics of oligopoly.

14.1 OBJECTIVES

After going through this unit, you will be able to:

- Define oligopoly
- Identify the factors causing oligopoly

- List the Characteristics of oligopoly
- Discuss the oligopoly models with the help of diagrams
- Explain price and output determination under oligopoly

14.2 OLIGOPOLY: AN OVERVIEW

Oligopoly is a form of market in which there are a few sellers selling homogeneous or differentiated products. Economists do not specify how few are the sellers in an oligopolistic market. However, two sellers is the **limiting case** of oligopoly. When there are only two sellers, the market is called **duopoly**.

In any case, if oligopoly firms sell a homogeneous poduct, it is called *pure* or *homogeneous oligopoly.* For example, industries producing bread, cement, steel, petrol, cooking gas, chemicals, aluminium and sugar are industries characterised by *homogeneous* oligopoly. And, if firms of an oligopoly industry sell *differentiated products,* it is called *differentiated* or *heterogeneous oligopoly.* Automobiles, television sets, soaps and detergents, refrigerators, soft drinks, computers, cigarettes, etc. are some examples of industries characterized by *differentiated* or *heterogeneous oligopoly.*

In the opinion of some authors, "Oligopoly is the most prevalent form of market organization in the manufacturing sector of the industrial nations .... " In non-industrial nations like India also, a majority of big and small industries have acquired the features of oligopoly market. The market share of 4 to 10 firms in 84 big and small industries of India is given below.

Market share (%) of 4-10 firms	No. of industries
1 – 24.9	8
25 – 49.9	11
50 – 74.9	15
75 – 100	50
Total	**84**

As the data presented above shows, in India, in 50 out of 84 selected industries, i.e., in about 60 per cent industries, 4 to 10 firms had a 75 per cent or more market share which gives a *concentration ratio* of 0.500 or above. All such industries can be classified under oligopolies.

Factors Causing Oligopoly

The main sources of oligopoly are described here briefly.

1. **Huge capital investment.** Some industries are by nature capital-intensive, e.g., manufacturing automobiles, aircraft, ships, TV sets, refrigerators, steel and aluminium goods, etc., and hence require huge investment. Therefore, only a few big firms can enter these kinds of industries. In fact, a huge investment requirement works as a natural barrier to entry to the oligopolistic industries.

2. Economies of scale. By virtue of huge investment and large scale production, large units enjoy *absolute cost advantage* due to economies of scale in purchase of industrial inputs, market financing, and sales organization. This gives the existing firms a comparative advantage over the new firms, especially in price competition. This works as a deterrent for the entry of new firms.

3. Patent rights. In case of *differentiated oligopoly,* firms get their differentiated product patented which gives them monopoly power, i.e., an exclusive right to produce and market the patented commodity. This prevents other firms from producing the patented commodity. Therefore, unless new firms have some thing new to offer and can match the existing products in respect of quality and cost, they cannot enter the industry. This keeps the number of firms limited.

4. Control over certain raw materials. Where a few firms acquire control over almost the entire supply of important inputs required to produce a certain commodity, new firms find it extremely difficult to enter the industry. For example, if a few firms acquire the right from the government to import certain raw materials, they control the entire input supply.

5. Merger and takeover. Merger of rival firms or takeover of rival firms by the bigger ones with a view to protecting their joint market share or to put an end to waste of competition is working, in modem times, as an important factor that gives rise to oligopolies and strengthens the oligopolistic tendency in modern industries.

14.2.1 Characteristics of Oligopoly

Let us now look at the important characteristics of oligopolistic industries.

1. Small number of sellers. As already mentioned, there is a small number of *sellers* under oligopoly. How small is the number of sellers? There is no precise answer to this question: it depends largely on the size of the market. Conceptually, however, the number of sellers is so small and the market share of each firm is so large that a single firm can influence the market price and the business strategy of its rival firms. The number may vary from industry to industry. Some examples of oligopoly industries in India and market share of the *dominant* firms in 1997–98 is given below.

Industry	No. of firms	Total market share (%)
Ice-cream	4	100.00
Bread	2	100.00
Infant Milk Food	6	99.95
Motorcycles	5	99.95
Passenger cars	5	94.34
Cigarettes	4	99.90
Fruit Juice, pulp & cone	10	98.21
Fluorescent lamps	3	91.84
Automobile tyres	8	91.37

Source: CMIE, *Industries and Market Share,* August 1999.

2. **Interdependence of decision-making.** The *most striking feature* of an oligopolistic market structure is the *interdependence* of oligopoly firms. The characteristics fewness of firms under oligopoly brings the firms in keen competition with each other. The competition between the firms takes the form of action, reaction and counteraction in the absence of collusion between the firms. Since the number of firms in the industry is small, the business strategy of each firm in respect of pricing, adevertising, product modification is closely watched by the rival firms and it evokes imitation and retaliation. What is equally important in strategic business decisions is that firms initiating a new business strategy anticipate and take into account the counteraction by the rival firms. This is called *interdependence* of oligopoly firms.

An illuminating example of strategic maneuvering is cited by Robert A. Meyer. To quote the example, one of the US car manufacturing companies announced in one year in the month of September an *increase* of $ 180 in the price list of its car model. Following it, a second company announced a few days later an increase of $ 80 only and a third announced an increase of $ 91. The first company made a counter move: it announced a *reduction* in the enhancement in the list price from $ 180 to $ 71. This is a pertinent example of interdependence of firms in business decisions under oligopolistic market structure. In India, when Maruti Udyog Limited (MUL), announced a price cut of ₹ 24,000 to ₹ 36,000 in early 1999 on its passenger cars, other companies followed the suit. However, *price competition* is not the major form of competition among the oligopoly firms as price war destroys the profits. A more common form of competition is *non-price competition* on the basis of product differentiation, vigorous advertising and provision of services.

3. **Barriers to entry.** Barriers to entry to an oligopolistic industry arise due to such market conditions as (*i*) huge investment requirement to match the production capacity of the existing ones, (*ii*) economies of scale and absolute cost advantage enjoyed by the existing firms, (*iii*) strong consumer loyalty to the products of the established firms based on their quality and service, and (*iv*) resistance by the established firms by price cutting. However, the new entrants that can cross these barriers can and do enter the industry, though only a few, that too mostly the branches of MNCs.

4. **Indeterminate price and output.** Another important feature, though controversial, of the oligopolistic market structure is the indeteminateness of price and output. The characteristic fewness and interdependence of oligopoly firms make derivation of the demand curve a difficult proposition. Therefore, price and output are said to be indeterminate. However, price and output are said to be determinate under collusive oligopoly. But, collusion may last or it may breakdown. *An opposite view* is *that price under oligopoly* is *sticky,* i.e., if price is once determined, it tends to stabilize.

14.3 THE OLIGOPOLY MODELS: AN OVERVIEW

As already mentioned, under oligopolistic conditions, rival firms adopt an intricate pattern of actions, reactions and counteractions showing a variety of behaviour patterns. The uncertainty arising out of unpredictable behaviour, actions and reactions of oligopoly firms makes systematic analysis of oligopoly an extremely difficult task. As Baumol puts it, "Under [these] circumstances, a very wide variety of behaviour pattern becomes possible. Rivals may decide to get together and cooperate in the pursuit of their objectives, ... or, at the other extreme, may try to fight each other to the death. Even if they enter an agreement, it may last or it may breakdown." Economists have, therefore, found it extremely difficult to make a systematic analysis of price and output determination under oligopoly. This has, however, not deterred the economists from their efforts to find a reasonable explanation to the problem of price determination in an oligopolistic market.

In accordance with the wide variety of behaviour patterns, economists have developed a variety of analytical models based on different behavioural assumptions. The famous models of oligopoly include Cournot's duoply model (1838), Bertrand's leadership model (1880), Edgeworth's duoply model (1897), Stackelberg's model (1933), Sweezy's kinked demand curve model (1939), Neumann and Margenstern Game Theory model (1944), Baumol's sales maximization model (1959). None of these models, however, provides a universally acceptable analysis of oligopoly, though these models do provide an insight into firms' oligopolistic behaviour.

In this unit, we will discuss some selected classicial and modern oligopoly models with the purpose of showing the behaviour of oligopoly firms and working of the oligopolistic markets. The analytical models discussed here are selected on the basis of how price and output are determined under price competition, cartel system and the dilemma that oligopoly firms face in their price and output decisions. Specifically, we will discuss the following oligopoly models.

   (*i*)   Cournot's Model of Duopoly: A limiting case of oligopoly,

  (*ii*)   Chamberlin's oligopoly model,

 (*iii*)   Sweezy's kinked-demand curve model,

 (*iv*)   Price leadership models:

       (*a*)   Price leadership by low-cost firm,

       (*b*)   Price leadership by dominant firm, and

       (*c*)   Price leadership by barometric firm,

   (*v*)   Collusive model: The Cartel Arrangement,

  (*vi*)   Baumol's sales revenue maximization model,

 (*vii*)   The Game Theory model of oligopoly, and

(*viii*)   Prisoner's Dilemma model.

A Classical Model of Duopoly: Cournot's Model

Augustine Cournot, a French economist, was the first to develop a formal oligopoly model in 1838. He formulated his oligopoly theory in the form of a *duopoly model* which can be extended to oligopoly model. To illustrate his model, Cournot made the following assumptions.

(*a*) There are two firms, each owning an artesian mineral water well;

(*b*) Both the firms operate their wells at zero marginal cost;

(*c*) Both of them face a demand curve with constant negative slope;

(*d*) Each seller acts on the assumption that his competitor will not react to his decision to change his *output*—Cournot's behavioural assumption.

On the basis of this model, Cournot has concluded that each seller ultimately supplies one-third of the market and both the firms charge the same price. And, one-third of the market remains unsupplied.

Cournot's duopoly model is presented in Fig. 14.1. The demand curve for mineral water is given by the *AR* curve and firm's *MR* by the *MR* curve. To begin with, let us suppose that there are only two sellers *A* and *B,* but initially, *A* is the only seller of mineral water in the market. By assumption, his *MC* = 0. Following the profit maximizing rule, he sells quantity *OQ* where his *MC* = 0 = *MR,* at price *OP*$_2$. His total profit is *OP*$_2$*PQ*.

Fig. 14.1 *Price and Output Determination under Duopoly: Cournot's Model*

Now let *B* enter the market. He finds that the market open to him is *QM* which is *half* of the total market. That is, he can sell his product in the remaining half of the market. *B* assumes that *A* will not change his *output* because he is making maximum profit. Specifically, *B* assumes that *A* will continue to sell *OQ* at prices *OP*$_2$, Thus, the market available to *B* is *QM* and the relevant part of the demand curve is *PM.* Given his demand curve *PM,* his *MR* curve is given by the curve *PN* which bisects *QM* at point *N* where *QN* = *NM.* In order to maximize his revenue, *B* sells *QN* at price *OP*$_1$. His total revenue is maximum at *QRP′N* which equals his total profit. Note that *B* supplies only *QN* = 1/4 = (1/2)/2 of the market.

Let us now see how *A's* profit is affected by the entry of *B*. With the entry of *B*, price falls to OP_1. Therefore, *A'*s expected profit falls to OP_1RQ. Faced with this situation, *A* assumes, in turn, that *B* will not change his output *QN* and price OP_1 as he is making maximum profit. Since $QN = 1/4$th of the market, *A* assumes that he has $3/4 (= 1 – 1/4)$ of the market available to him. To maximize his profit, *A* supplies 1/2 of the unsupplied market (3/4), i.e., 3/8 of the market. It is noteworthy that *A'*s market share has fallen from 1/2 to 3/8.

Now it is *B'*s turn to react. Following Cournot's assumption, *B* assumes that *A* will continue to supply only 3/8 of the market and the market open to him equals $1 – 3/8 = 5/8$. To maximise his profit under the new conditions, *B* supplies $1/2 \times 5/8 = 5/16$ of the market. It is now for *A* to reappraise the situation and adjust his price and output accordingly.

This process of action and reaction continues in successive periods. In the process, *A* continues to lose his market share and *B* continues to gain. Eventually, a situation is reached when their market share equals 1/3 each. Any further attempt to adjust output produces the same result. The firms, therefore, reach their equilibrium where each one supplies one-third of the market and both charge the same price.

The actions and reactions and equilibrium of the sellers *A* and B, according to Cournot's model, are presented in Table 14.1.

Table 14.1 Determination of Market Share

Period	Seller A		Seller B
I	$\frac{1}{2}(1) = \frac{1}{2}$	→	$\frac{1}{2}\left(\frac{1}{2}\right) = \frac{1}{4}$
II	$\frac{1}{2}\left(1 - \frac{1}{4}\right) = \frac{3}{8}$	← →	$\frac{1}{2}\left(1 - \frac{3}{8}\right) = \frac{5}{16}$
III	$\frac{1}{2}\left(1 - \frac{5}{16}\right) = \frac{11}{32}$	← →	$\frac{1}{2}\left(1 - \frac{11}{32}\right) = \frac{21}{64}$
IV	$\frac{1}{2}\left(1 - \frac{21}{64}\right) = \frac{43}{128}$	← →	$\frac{1}{2}\left(1 - \frac{43}{128}\right) = \frac{85}{256}$
...	...		...
...	...		...
N	$\frac{1}{2}\left(1 - \frac{1}{3}\right) = \frac{1}{3}$	← →	$\frac{1}{2}\left(1 - \frac{1}{3}\right) = \frac{1}{3}$

Note: Arrows show the direction of actions and reactions of sellers *A* and *B*.

Cournot's equilibrium solution is stable. For, given the action and reaction, it is not possible for any of the two sellers to increase their market share as shown in the last row of the table.

Cournot's model of duopoly can be extended to a general oligopoly model. For example, if there are three sellers in the industry, each one of them will be in equilibrium when each firm supplies 1/4 of the market. The three sellers together supply 3/4 of the total market, 1/4 of the market remaining unsupplied. Similarly, when there are four firms each one of them supply 1/5th of the market and 1/5th of the market remains unsupplied. The formula for determining the share of each seller in an oligopolistic market is: $Q \div (n + 1)$ where Q = market size, and n = number of sellers.

Algebraic Solution of Duopoly: Cournot's model can also be presented algebraically. Let us suppose that the market demand function is given by linear function as

$$Q = 90 - P \qquad (14.1)$$

As noted above, under zero cost condition, profit is maximum where $MC = MR = 0$ and when $MR = 0$, the profit maximizing output is $1/2\ (Q)$.

Let us suppose that when A is the only seller in the market, his profit-maximising output is Q_A which is determined by the profit maximising rule under zero cost condition. $A's$ market share can be written as

$$Q_A = 1/2\ (90 - P) \qquad (14.2)$$

When seller B enters the market, his profit maximising output is determined as follows.

$$Q_B = 1/2\ [(1/2(90 - P)] \qquad (14.3)$$

Thus, the respective shares of sellers A and B are fixed at Q_A and Q_B. The division of market output may be expressed as

$$Q = Q_A + Q_B = 90 - P \qquad (14.4)$$

The demand function for A may now be expressed as

$$Q_A = (90 - Q_B) - P \qquad (14.5)$$

and for B as

$$Q_B = (90 - Q_A) - P \qquad (14.6)$$

Given the demand function (14.5), the market open to A (at $P = 0$) is $90 - Q_B$. The profit maximising output for A will be

$$Q_A = \frac{90 - Q_B}{2} \qquad (14.7)$$

and for B, it will be

$$Q_B = \frac{90 - Q_A}{2} \qquad (14.8)$$

The equations (14.7) and (14.8) represent the reaction functions of sellers A and B, respectively. For example, consider equation (14.7). The profit maximising output of A depends on the value of Q_B, i.e., the output which B is assumed to produce. If B chooses to produce 30 units (i.e., $Q_B = 30$), then $A's$ profit maximizing output = $[(90 - 30)1/2] = 30$. If B chooses to produce 60 units, $A's$

profit maximizing output = (90 – 60) 1/2 = 15. Thus, equation (14.8) is *A′*s reaction function. It can similarly be shown that equation (14.8) is *B′*s reaction function.
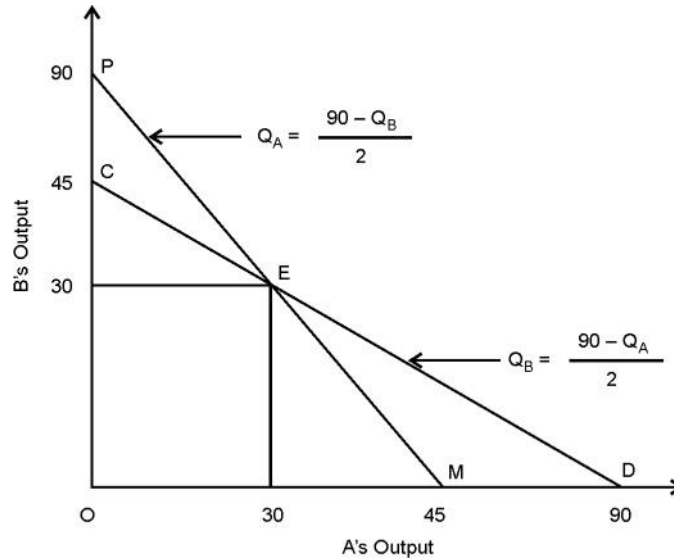
Fig.14.2 *Reaction Functions and Equilibrium: Cournot's Model*

The reaction functions of *A* and *B* are graphed in Fig. 14.2. The reaction function *PM* shows how *A* will react on the assumptions that *B* will not react to changes in his output once *B*'*s* output is fixed. The reaction function *CD* shows a similar reaction of *B*. The two reaction functions intersect at point *E*. It means that the assumptions of *A* and *B* coincide at point *E* and here ends their action and reaction. Point *E* is, therefore, the point of stable equilibrium. At this point, each seller sells only 30 units.

The same result can be obtained by equating the two reaction equations (14.7) and (14.8). The market slope of *A* and *B* can be obtained by equating *A′*s and *B′*s reaction functions (14.7) and (14.8), respectively. That is, market equilibrium lies where

$$\frac{90 - Q_B}{2} = \frac{90 - Q_A}{2}$$

Since, $Q_B = (90 - Q_A)/2$, by substitution, we get first term as

$$Q_A = \frac{90 - (90 - Q_A)/2}{2}$$

$$Q_A = 30$$

Thus, both the sellers are in equilibrium at their respective output of 30. The maket output will be 60 units. Given the market demand curve, market price will be $P = 90 - Q = 90 - 60 = ₹\ 30$.

As mentioned above, the dupoly model can be extended to oligopoly market.

Criticism of Cournot's Model: As we have seen earlier, Cournot's model is logically sound and yields a stable equilibrium solution. His model has, however, been criticized on the following grounds.

First, Cournot's behavioural assumption, specifically assumption (*d*) above, is said to be naive as it implies that firms continue to make wrong calculations about the behaviour of the rival firms even though their calculations are proved wrong. For example, each seller continues to assume that his rival will no change his output even though he finds frequently that his rival does change his output.

Secondly, Cournot assumed zero cost of production, which is not realistic. However, even if this assumption is ignored, Cournot's results are not affected.

Chamberlin's Model of Oligopoly: The Small Group Model

The *classical models* assume, as shown in case of Cournot's model, that the actions taken by the firms in order to maximise their profits are independent of possible reactions of the rival firms. Chamberlin rejected the assumption of independent action by competing firms. He has developed a model of oligopoly assuming **interdependence** between the competitors. He argues that firms do not act independently. They do recognise their mutual interdependence. Firms are not as 'stupid' as assumed in the classical models of Cournot, Edgeworth and Bertrand. In his own words, "when a move by one seller evidently forces the other to make a counter-move, he is very stupidly refusing to look farther than his nose if he proceeds on the assumption that it will not." Chamberlin suggests that each seller seeking to maximise his profit reflects well and looks into the consequences of his move. The total consequence of a seller's move consists of both its direct and indirect effects. *Direct effects* are those that result from a seller's own action, rival sellers not reacting. *Indirect effects* are those which result from the reaction of the rival sellers to the action of a seller.

Chamberlin suggests in his oligopoly model that, if rival firms are assumed to recognise their mutual interdependence and act accordingly, a stable equilibrium can be reached where each firm charges a monopoly price and shares equally the monopoly equilibrium output. When all firms are in equilibrium, industry profit is maximised. Since Chamberlin follows Cournot to develop his own model, Chamberlin's oligopoly model of 'small group' can be best understood if presented in the framework of Cournot's duopoly model.

Cournot's model is reproduced in Fig. 14.3, added with ordinate *JK*. Suppose there are two firms, *A* and *B*. Let the firm *A* enter the market first as a monopolist. Following the profit maximization rule, firm *A* produces *OQ* and charges monopoly price OP_2 (= *PQ*). When the firm *B* enters the market, it considers *PM* segment of the demand curve as its own demand curve. Under Cournot's proft maximization assumption, firm *B* will sell output *QN* at price OP_1. As a result, market price falls from OP_2 to OP_1. It is now *A*'s turn to appraise the situation. At this point, Chamberlin *deviates from Cournot's model.* While Cournot's model assumes that firm *A* does not recognise the *interdependence* between them and

acts independently, Chamberlin assumes that firm *A* does recognise the interdependence between them and it does recognise the fact that *B* will react to its decisions. Therefore, firm *A* decides to compromise with the existence of firm *B* and decides to reduce its output to *OK* which is half of the monopoly output, *OQ*. Its output *OK* equals *B*'s output *ON* (= *KQ*). In its turn, firm *B* also recognises their interdependence. It realises that *KQ* is the most profitable output for it. Thus, the industry output is *OQ* which equals *monopoly output* and market price is *OP*$_2$ (= *PQ*) which equals *monopoly price*. Thus, by recognissing their interdependence, the firms reach a stable equilibrium. This equilibrium is stable because under the condition of interdependence, firms do not gain by changing their price and output.
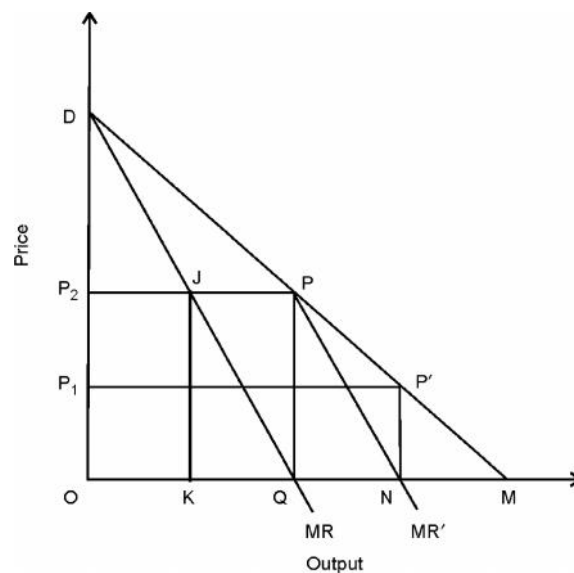
Fig.14.3 *Chamberlin's Model of Stable Equilibrium*

Criticism. Chamberlin's model is regarded as an improvement over the earlier models, at least in respect of its behavioural assumption of interdependence. His model has, however, been criticised on the grounds that the idea of joint profit maximization implicitly assumes that demand and cost functions are known which, in reality, may not be. Unless demand and cost functions are fully known to the competitors, joint profit maximization is doubtful. However, this criticism is not strong enough to challenge Chamberlin's theory.

Sweezy's Kinked-demand Curve Model

The origin of kinked-demand curve can be traced into Chamberlin's theory of monopolistic competition. Later, Hall and Hitch used kinked-demand curve to explain rigidity of prices in oligopolistic market. But, neither Chamberlin nor Hall and Hitch used kinked-demand curve as a tool of analysis in their respective theories. It was Paul M. Sweezy who used the kinked-demand curve in his model of price stability in oligopolistic market. Sweezy's Model is described subsequently in the text.

The kinked-demand curve model developed by Paul M. Sweezy has features common to most oligopoly pricing models. This is the best known model to explain, relatively more satisfactorily, the behaviour of the oligopolistic firms. It must, however, be noted at the outset that *kinked-demand curve analysis does not deal with price and output determination. Rather, it seeks to establish that once a price-quantity combination is determined, an oligopoly firm does not find it profitable to change its price even when there is a considerable change in the cost of production and change in demand for the product.*

The logic behind the proposition that price once determined remains stable runs as follows. An oligopoly firm believes that if it reduces the price of its product, the rival firms would follow and neutralise the expected gain from price reduction. But, if it raises the price, the firms would either maintain their prices or even go for price-cutting, so that the price-raising firm loses a part of its market to the rival firms. This behaviour is true of all the firms. The oligopoly firms would, therefore, find it more desirable to maintain the prevailing price and output. This is the basic theme of Sweezy's theoretical model. This model is explained and illustrated below.

In order to analyse the effects of possible reactions of the rival firms on the demand for the product of the firm initiating the change in price, let us make the following assumptions.

  (*i*) There are four oligopoly firm—*A, B, C* and *D*;

 (*ii*) Market demand curve is given by *dd'* in Fig. 14.4; and

(*iii*) All the firms are in equilibrium at point *P*.

Let us suppose that firm *A* takes lead in changing its price and examine the effect of various kinds of reactions of the rival firms on demand for *A*'s product.
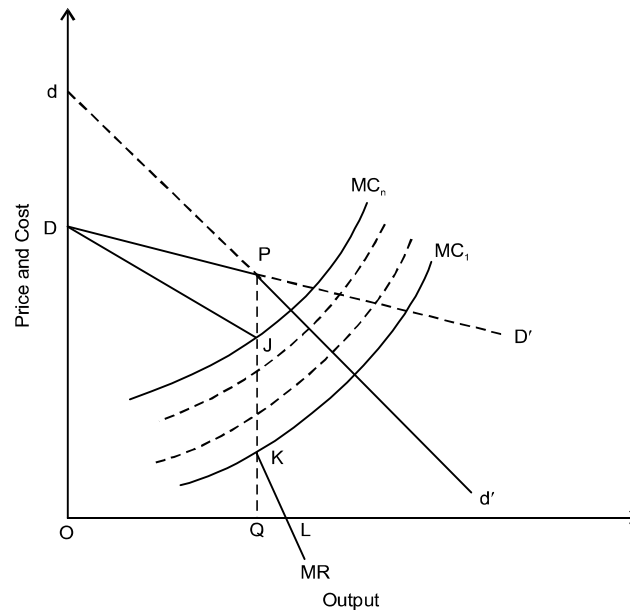

Fig. 14.4 *Kinked-Demand Curve Analysis*

Reaction (i). When firm *A* increases or decreases its price, the rival firms follow the suit. Then firm *A* finds itself moving along the demand curve *dd'*. It does not gain nor does it lose.

Reaction (ii). When the rival firms do not react to price changes made by the firm *A*, its demand curve becomes highly elastic as shown by the curve *DD'* in Fig. 14.4. To explain it further, when firm *A* raises its price and rival firms do not follow, firm *A* loses a part of its market to the rival firms and moves along *PD* part of the demand curve. But, when firm *A* cuts its price and rival firms do not follow, then it captures a part of the rival's market share and finds itself moving along the *PD'* part of the demand curve. This is what firm *A* would like to achieve. Note that *PD'* part of demand curve is more elastic than *Pd'*.

Reaction (iii). When firm *A* raises its price and rival firms do not follow, then firm *A* loses a part of its market share to the rival firms. Then the relevant demand curve for firm *A* is *DP*. But, when firm *A* decreases its price, rival firms react by cutting down their own prices by an equal amount or even more. This is a more realistic reaction. This counter move by the rival firms prevents firm *A* from taking any advantage of price cut. Therefore, the relevant segment of demand curve for firm *A* (below point *P*) is *Pd'*. If the two relevant segments of the two demand curves are put together, the demand curve for *A*'s product takes the form of the curve *DPd'*. Note that this demand curve has a **kink** at point *P*. It is, therefore, called **a kinked-demand curve.**

Let us now derive *MR* curve. We know that given the demand function as $D = a - bP$, marginal revenue (*MR*) function is given as $MR = a - 2bP$. The derivation of the *MR* curve on the basis of this *MR* function is shown in Fig. 14.4 under the condition of kinked demand (or *AR*) curve. The segment *DJ* of the *MR* curve corresponds to *DP* segment of the demand curve and *KL* segment of *MR* curve corresponds to *Pd'* segment of the demand curve. By joining the two segments of the *MR* curves, we get the full *MR* curve as *DJKL*.

Let us suppose that the marginal cost curve is given as MC_1 which intersects *MR* at point *K*. Point *K* satisfies the necessary condition for profit maximization ($MR = MC$). Therefore, oligopoly firms are in equilibrium at output *OQ* and they are making maximum profit. Now, if marginal cost curve shifts upwards to MC_n or to any level between points *J* and *K,* their profit would not be affected because profit maximization condition remains undistributed. Therefore, they have no motivation for increasing or decreasing their price. It is always beneficial for them to stick to the price *PQ* and output *OQ*. Thus, both price and output are *stable*. The oligopoly firms would think of changing their price and output only if *MC* rises beyond point *J*. The same analysis applies to decrease in *MC* below point *K*. The firms would not cut their prices down unless *MC* decreases below point *K* (Fig. 14.4).

Criticism of Sweezy's Model

As mentioned earlier, Sweezy's model is considered to be the best known model that explains relatively more satisfactorily, the behaviour of the firms in oligopoly. On the face it, it appears to be logically sound and realistic. However, economists have criticized his model on both theoretical and empirical grounds as follows.

1. Sweezy's model does not explain price determination. The basic function of price theory is to explain price and output determination in a particular kind of market. Sweezy's model, however, does not explain price and output determination. His model only assumes the price to be given at a point of time. It explains only why price once determined tends to be sticky even if there are changes in cost conditions to a certain extent. Sweezy's model is, therefore, regarded as an *ex-post rationalization* rather than *ex-ante* explanation of market equilibrium.

2. This model does not determine the point of kink. This is a criticism related to non-determination of price. The kinked demand curve anlaysis explains why 'kink' appears on the demand curve. It does not explain how and at what level of price and output, the point of kink is determined. George Stigler doubts even the existence of the kinked-demand curve. Stigler's view is supported by Julian Simon. This makes the model a purely hypothetical one, not as realistic as it appears on the face of it. However, Cohen and Cyert argue that kink in the demand curve and price rigidity may exist for a short period, for lack of inter-firm information, especially when new and unknown rivals enter the market. They are of the opinion that kink is clearly not a stable long-run equilibrium.

3. Price rigidity is not supported by empirical facts. Sweezy's claim of price rigidity in oligopoly does not stand the test of empirical verification. Empirical facts reveal a surprising lack of price stability in oligopoly markets. Empiricially, monopoly prices have been found to be more stable than oligopoly prices. Economists' opinion is, however, divided on the issue of price rigidity in oligopoly. While Stigler has questioned price rigidity in oligopoly market, Liebhafsky finds considerable evidence of price rigidity in oligopolistic industries of the US.

4. Sweezy's conclusion conflicts with marginal productivity theory. In Sweezy's model, *MC* curve can shift up and down (say, between finite points *J* and *K* in Fig. 14.4), while *MR* remains the same. This argument is in conflict with *marginal productivity theory* of factor pricing as this means that factor prices do not necessarily equal the marginal revenue productivity.

14.4 PRICE AND OUTPUT DETERMINATION IN COLLUSIVE OLIGOPOLY

The oligopoly models discussed in the previous sections are based on the *assumption* that the oligopoly firms act *independently;* they are in competition with one another; and there is *no collusion* between the firms. The oligopoly

models of this category are called *non-collusive models.* In reality, however, oligopoly firms are found to have some kind of collusion or agreement—open or secret, explicit or implicit, written or unwritten, and legal or illegal—with one another for at least **three major reasons. First,** collusion eliminates or reduces the degree of competition between the firms and gives them some monopolistic powers in their price and output decisions. **Second,** collusion reduces the degree of uncertainty surrounding the oligopoly firms and ensures profit maximisation. **Third,** collusion creates some kind of barriers to the entry of new firms.

The models that deal with the collusive oligopolies are called **collusive oligopoly models.** Collusion between firms may take many forms depending on their relative strength and objective of collusion, and on whether collusion is legal or illega1. There are, however, two major forms of collusion between the oligopoly firms: (*i*) cartel, i.e., firms' association, and (*ii*) price leadership agreements.

Accordingly, the collusive oligopoly models that economists have developed to explain the price determination under oligopoly can be classified as:

 (i) Cartel Models, and

(*ii*) Price Leadership Models.

 In this section, we will discuss these two types of oligopoly models.

The Cartel Models: The Collusive Models

Oligopoly Cartels: A Form of Collusion: A *cartel* is a formal organisation of the oligopoly firms in an industry. A general purpose of cartels is to centralise certain managerial decisions and functions of individual firms in the industry, with a view to promoting common benefits. Cartels may be in the form of *open* or *secret collusion.* Whether open or secret, cartel agreements are *explicit* and formal in the sense that agreements are enforceable on the member firms not observing the Cartel Rules or dishonouring the agreements. Cartels are, therefore, regarded as **the perfect form of collusion.** Cartels and cartel type agreements between the firms in manufacturing and trade are illegal in most countries. Yet, cartels in the broader sense of the term exist in the form of trade associations, professional organisations and the like.

A cartel performs a variety of services for its members. The two services of central importance are (*i*) fixing price for joint profit maximization; and (*ii*) market-sharing between its members. Let us now discuss price and output determination under the cartel system.

(i) Joint Profit Maximization Model

Let us suppose that a group of firms producing a homogeneous commodity forms a cartel aiming at joint profit maximization. The firms appoint a central management board with powers to decide (*i*) the total quantity to be produced; (*ii*) the price at which it must be sold; and (*iii*) the share of each firm in the total output. The cartel board is provided with cost figures of individual firms. Besides, it is supposed to obtain the necessary data required to formulate the market

demand (*AR*) curve. The cartel board calculates the marginal cost (*MC*) and marginal revenue (*MR*) for the industry. In a sense, the cartel board holds the position of a multiplant monopoly. It determines the price and output for each firm in the manner a multiplant monopoly determines the price and output for each of its plants.

The model of price and output determination for each firm is presented in Fig. 14.5. It is assumed for the sake of convenience that there are only two firms, *A* and *B*, in the cartel. Their respective cost curves are given in the first two panels of Fig. 14.5. In the third panel, *AR* and *MR* curves represent the revenue conditions of the *industry*. The *MC* curve is the summation of *mc* curves of the individual firms. The *MC* and *MR* curves intersect at point *C* determining the industry output at *OQ*. Given the industry output, the market price is determined at *PQ*.

Now, under the cartel system, the industry output *OQ* has to be so allocated between firms *A* and *B* that their individual *MC* = *MR*. The share of each firm in the industry output, *OQ*, can be obtained by drawing a line from point *C* and parallel to X-axis through mc_2 and mc_1. The points of intersection c_1 and c_2 determine the profit maximizing output for firms *A* and *B*, respectively. Thus, the share of firms *A* and *B*, is determined at OQ_A and OQ_B, respectively, where $OQ_A + OQ_B = OQ$. At these outputs, they maximize their respective profits.
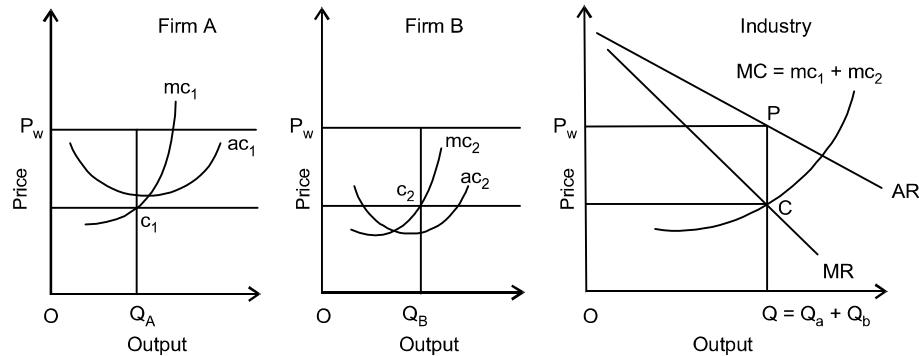
Fig. 14.15 *Price and Output Determination Under Cartel*

Problems in Joint Profit Maximization: Although the above solution to joint profit maximization by cartel looks theoretically sound, William Fellner gives the following reasons why profits may not be maxmized jointly.

First, it is difficult to estimate market demand curve 'accurately since each firm thinks that the demand of its own product is more elastic than the market demand curve because its product is a perfect substitute for the product of other firms.

Secondly, an accurate estimation of industry's *MC* curve is highly improbable for lack of adequate and correct cost data. If industry's *MC* is incorrectly estimated, industry output can be only incorrectly determined. Hence joint profit maximization is doubtful.

Thirdly, cartel negotiations take a long time. During the period of negotiation, the composition of the industry and its cost structure may change. This may render demand and cost estimates irrelevant, even if they are correct. Besides, if the number of firms increases beyond 20 or so, cartel formation becomes difficult, or even if it is formed, it breaks down soon.

Fourthly, there are 'chiselers' who have a strong temptation to give secret or undeclared concessions to their customers. This tendency in the cartel members reduces the prospect of joint profit maximisation.

Fifthly, if cartel price, like monopoly price, is very high, it may invite government attention and interference. For the fear of government interference, members may not charge the cartel price.

Sixthly, another reason for not charging the cartel price is the fear of entry of new firms. A high cartel price which yields monopoly profit may attract new firms to the industry. To prevent the entry of new firms, some firms may decide on their own not to charge the cartel price.

Lastly, yet another reason for not charging the cartel price is the desire to build a public image or good reputation. Some firms may, to this end, decide to charge only a fair price and realise only a fair profit.

(ii) Cartel and Market-Sharing

The market-sharing cartels are more common because this kind of collusion permits a considerable degree of freedom in respect of style and design of the product, advertising and other selling activities. There are two main methods of market allocations: (*a*) non-price competition agreement, and (*b*) quota system.

(a) Non-price competition agreement. The non-price competition agreements are usually associated with loose cartels. Under this kind of arrangement between firms, a uniform price is fixed and each firm is allowed to sell as much as it can at the cartel price. The only requirement is that firms are not allowed to reduce the price below the cartel price.

The cartel price is, however, a bargain price. While low-cost firms press for a low price, the high-cost firms press for a higher price. But the cartel price is so fixed by mutual consent that all member firms are able to make a reasonable profits. However, firms are allowed to compete with one another in the market on a non-price basis. That is, they are allowed to change the style of their product, innovate new designs and to promote their sales without reducing their price below the level of cartel price.

Whether this arrangement works or breaks down depends on the cost conditions of the individual firms. If some firms expect to increase their profits by violating the price agreements, they will indulge in cheating by charging a lower price. This may lead to a price-war and the cartel may break down.

(b) Quota system. The second method of market-sharing is *quota system*. Under this system, the cartel fixes a quota of market-share for each firm. There

is no uniform principle by which quota is fixed. In practice, however, the main considerations are (*i*) bargaining ability of a firm and its relative importance in the industry, (*ii*) the relative sales or market share of the firm in pre-cartel period, and (*iii*) production capacity of the firm. The choice of the base period depends on the bargaining ability of the firm.

Fixation of quota is a difficult problem. Nevertheless, some theoretical guidelines for market sharing are suggested as follows. A reasonable criterion for ideal market-sharing can be to share the total market between the cartel members in such proportions that the industry's marginal cost equals the marginal cost of individual firms. This criterion is illustrated in Fig. 14.6 assuming an oligopoly industry consisting of only two firms, *A* and *B*. The profit maximizing output of the industry is *OQ.* The industry output *OQ* is so shared between the two firms *A* and *B* that their individual *MC* equals industry's *MC*. As shown in Fig. 14.6, at output OQ_A, *MC* of firm *A* equals industry's marginal cost, *MC,* and at output OQ_B, *MC* of firm *B* equals industry's *MC*. Thus, under quota system, the quota for firms *A* and *B* may be fixed as OQ_A and OQ_B, respectively. Given the quota allocation, the firm may set different prices for their product depending on the position and elasticity of their individual demand curves. This criterion is identical to the one adopted by a multiplant monopolist in the short-run, to allocate the total output between the plants.
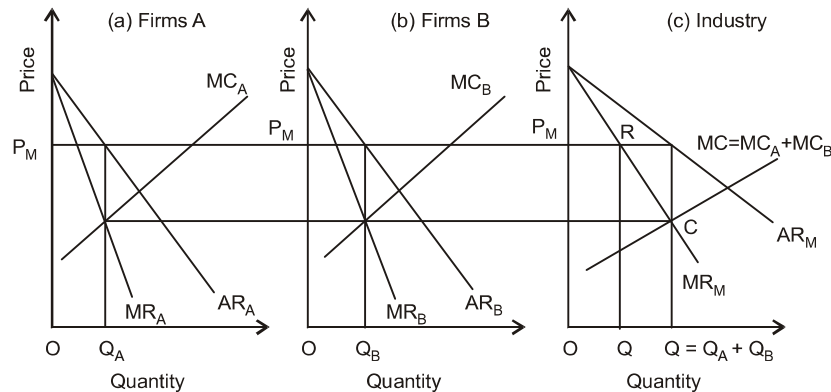
Fig. 14.6 *Quota Allocation under Cartel Agreements*

Another reasonable criterion for market-sharing under quota system is *equal market-share for equal firms.* This criterion is applicable where all firms have *identical* cost and revenue curves. This criterion also leads to a monopoly solution. It resembles Chamberlin's duopoly model.

To illustrate equal market sharing through quota allocation, let us assume that there are only two firms, *A* and *B*. Their *AR, MR* and *MC* curves are presented in Fig. 14.6 (a) and 14.6 (b). The market revenue and cost curves, which are obtained by summing the individual revenue and cost curves, respectively, are presented in panel (c) of the figure. The industry output is determined at *OQ.* The share of each firm, which maximises their profits, is so determined that $OQ = OQ_A + OQ_B$, Given the identical cost and revenue conditions, $OQ_A = OQ_B$. That is,

market is divided equally between firms *A* and *B*. This result can be obtained also by drawing an ordinate from the point where price line (P_M) intersects the MR_M i.e., from point *R*. The market output *OQ* is divided equally between firms *A* and *B*.

It may be noted at the end that cartels do not necessarily create the conditions for price stability in an oligopolistic market. Most cartels are loose. Cartel agreements are generally not binding on the members. Cartels do not prevent the possibility of entry of new firms. On the contrary, by ensuring monopoly profits, cartels create conditions which attract new firms to the industry. Besides, 'chiselers' and 'free-riders' create conditions for instability in price and output.

Price Leadership Models of Oligopoly

Price leadership is an imperfect form of collusion betweeen oligopoly firms. Price leadership is an informal position given to or attained by a firm in an oligopolistic setting to lead other firms in pricing. This leadership may emerge spontaneously due to technical reasons or out of tacit or explicit agreements between the firms to assign leadership role to one of them.

The *spontaneous price leadership* may be the result of such technical factors as size, efficiency, economies of scale or firm's ability factors. The most typical case of price leadership is the leading role played by the dominant firm of the industry. The dominant firm takes lead in changing the price and the smaller ones follow. Sometimes price leadership is *barometric*. In the barometric price leadership, one of the firms, not necessarily dominant one, takes lead in announcing change in price, particularly when such a change is due but is not brought into effect due to uncertainty in the market.

The price leadership is possible under the conditions of both *product homogeneity* and *product differentiation.* There may, however, be price differentials on account of product differentiation. Price differentials may also exist on account of cost differentials.

Another important aspect of price leadership is that it often serves as a means to price discipline and price stabilisation. Achievement of this objective establishes an effective price leadership. Such a price leadership can, however, exist effectively only under the following conditions.

(*i*) number of firms is small;

(*ii*) entry to the industry is restricted;

(*iii*) products are, by and large, homogeneous;

(*iv*) demand for industry is inelastic or has a very low elasticity; and

(*v*) firms have almost similar cost curves.

There are three common types of price leaderships: (*i*) Price leadership by a low-cost or most efficient firm; (*ii*) Price leadership by a dominant firm; and (*iii*) Barometric price leadership. Let us discuss price and output determination under the three kinds of price leaderships.

(i) Price Leadership by a Low-Cost Firm

The price and output decisions under price leadership of a low-cost firm is illustrated in Fig. 14.7. Suppose all the firms face identical revenue curves as shown by *AR* and *MR* curves. But the largest firm or the low-cost firm, has its cost curves as shown by AC_1 and MC_1 whereas all the rival firms, smaller in size, have their cost curves as shown by AC_2 and MC_2. The largest firm has greater economies of scale and, therefore, its cost of production is lower than that of other firms. Given the cost and revenue conditions, the low-cost firm would find it most profitable to produce and sell OQ_2 and fix its price at OP_2 ($= LQ_2$), Since at this level of output, its $MC = MR$, its profit is maximum. On the other hand, the high-cost firms would be in a position to maximise their profit at price OP_3 and quantity OQ_1. But, if they charge a higher price, OP_3, they would lose their customers to the low-cost firm. The high-cost firms are, therefore, forced to accept the price OP_2 and recognise the price leadership of the low-cost firm.
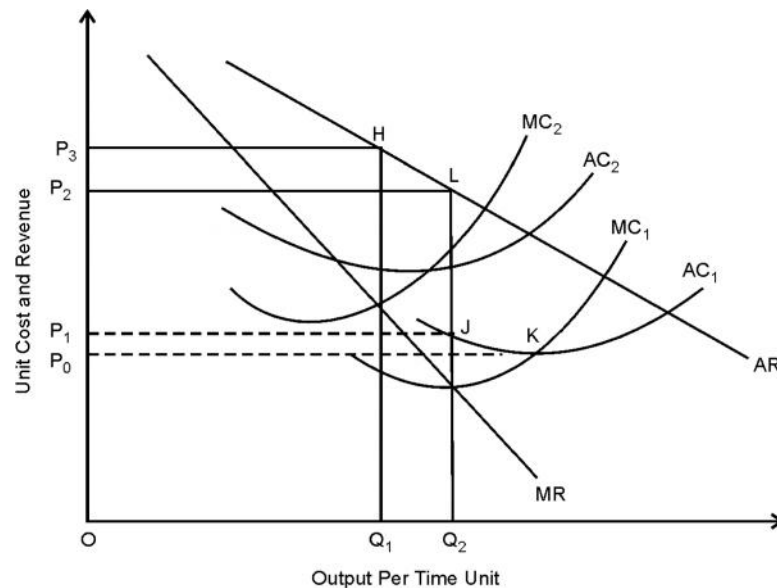
Fig. 14.7 *Price Leadership by a Low-Cost Firm*

Note that the low-cost firm can eliminate the high-cost firms and become a monopolist by cutting the price down to OP_1 ($= JQ_2$). The low-cost firm can sell its entire output OQ_2 at price OP_1 and make only normal profit. If necessary, it can cut its price further down to OP_0 and still make normal profits. It will, however, not do so as it would avoid falling under anti-monopoly laws.

(ii) Price Leadership by a Dominant Firm

Price leadership by a dominant firm is more common than by a low-cost firm. In the analysis of price leadership by a dominant firm, it is assumed that there exists a large size firm in the industry, which supplies a large proportion of the total market. The dominance of the large firm is indicated by the fact that it could possibly eliminate all its rival firms by price-cutting. In that case, the large firm

gains the status of a monopoly which may invite legal problems. The dominant firm, therefore, compromises with the existence of small rival firms in the market. It uses its dominance to set its price so as to maximise its profit. The smaller firms recognise their weak position and behave like a firm in a perfectly competitive market, i.e., smaller firms accept the price set by the dominant firm.

The price leadership and market sharing between the dominant firm and the rival small firms as a group is illustrated in Fig. 14.8. Suppose that the market demand curve is given by DD_M and the supply curve of the small firms together is given by the curve S_s in panel (a) of the figure. The problem confronting the dominant firm is to determine its price and output that will maximise its profit, leaving the rest of the market to be jointly supplied by the small firms. To solve this problem, the dominant firm finds its demand curve by deducting the quantity supplied jointly by the small firms at different prices from the corresponding market demand. The dominant firm considers the residual of the market share as the demand for its own product. Thus, at a given price the market share of the dominant firm equals the market demand *less* the share of small firms.
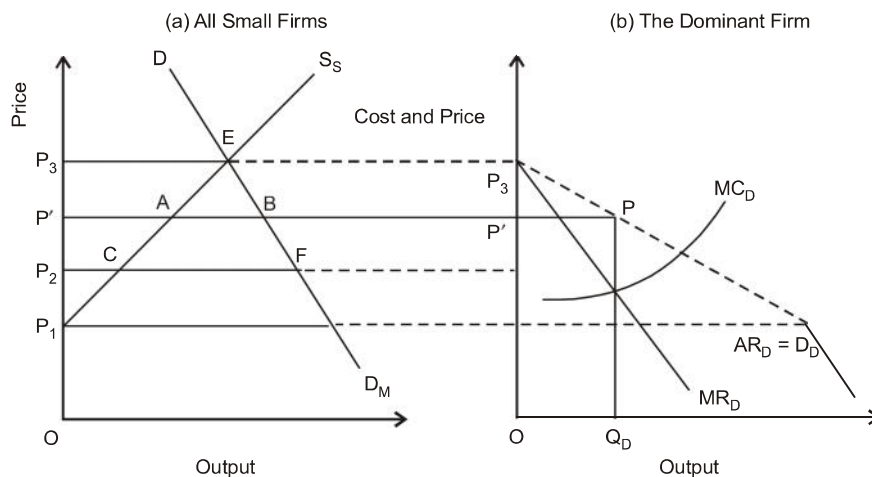
Fig. 14.8 *Price Leadership by a Dominant Firm*

For example, when market price is set at OP_3, the total supply by the smaller firms is P_3E which equals the market demand. Therefore, at price OP_3, the market left for the dominant firm is zero. When price falls to OP_2, the demand for dominant firm's product is $CF = P_2F – P_2C$. Following this process, the market-share of the dominant firm at other prices can be easily obtained.

Note that the gap between demand curve DD_M and supply curve P_1S_s below point E in Fig. 14.8 (a) measures the demand for the dominant firm.

The information so derived and plotted graphically gives P_3D_D as the demand curve for the dominant firm [Fig. 14.8 (b)]. Since the relation *between AR* and *MR* is known, the *MR* curve for the dominant firm can be dervied as MR_D [Fig. 14.8 (b)]. If *MC* curve of the dominant firm is assumed to be given as MC_D, its profit maximising output will be OQ_D and price PQ_D.

Once the dominant firm sets its price at $PQ_D = OP'$, the small firms have to accept this price, and then their joint market demnad curve is the horizontal straight line $P'B$ [in Fig. 14.8 (a)], because they can sell at this price as much as they can produce. But, in order to maximise their joint profits, small firms will produce only $P'A$. For small firms, therefore, profit maximizing joint output is $P'A$.

Criticial Appraisal: The dominant-firm price-leadership model, as presented above, yields a stable solution to the problem of oligopoly pricing and output determination, only if the small firms faithfully follow the leader. That is, small firms produce the right quantity and charge the price set by the dominant firm. Besides, the model requires that the dominant firm should be both large and a low-cost firm. For, if a firm does not enjoy the advantage of large size and, consequent upon it, the advantage of low-cost, it cannot act as a price leader.

In practice, however, one finds many cases of price leadership by a firm which is neither large nor is a low-cost firm. But such cases are found mostly under recessionary conditions when a relatively smaller firm reduces its price, to survive in the market.

Furthermore, if a leading firm loses its cost advantages, it also loses its leadership. Such cases are frequent in the real business world. Leadership also changes following the innovation of products and techniques of production by the small firms.

Besides, where there are many large firms of equal size and have some cost advantage, price leadership by any firm or group of firms becomes less probable, particularly when the number of small firms is smaller than that of larger firms. Under such conditions, another kind of price leadership, i.e., barometric leadership, emerges.

Lastly, it is assumed that the entry of new firms is prevented either by low-cost of the existing firms or by initial high cost of new firms. In practice, however, many firms having the capacity to diversify their products enter the industry with relatively initial low-cost.

For these reasons, dominant-firm leadership model is not considered to be a very realistic one.

(iii) The Barometric Price Leadership

Another form of price leadership is *barometric price leadership.* In this form of price leadership, a firm initiates well publicised changes in price which are generally followed by the rival firms. This kind of price leadership comes from a firm whose activities are taken as the barometer for the industry—it may not necessarily come from the largest firm of the industry. The barometric firm is supposed to have a better knowledge of the prevailing market conditions and has an ability to predict the market conditions more precisely than any of its competitors. These qualities of the barometric firm should have been established and recognized over time by the rival firms. The firm having the qualifications of price leadership, is regarded

as a barometer, which reflects the changes in business conditions and environment of the industry. The price changes announced by the barometric firm serve as a barometer of changes in demand and supply conditions in the market.

The barometric leadership evolves for various reasons. The major ones are the following.

First, the rivalry betweeen the large firms may lead to cut-throat competition to the disadvantage of all the firms. On the other hand, rivalry between the larger firms may make them unacceptable as a leader. So a firm which has better predictive ability emerges as the price leader.

Secondly, most firms in the industry may have neither the capacity nor the desire to make continuous calculations of cost, demand and supply conditions. Therefore, they find it advantageous to accept the price changes made by a firm which has a proven ability to make reasonably good forecasts.

Thirdly, Kaplan *et. al.,* observe that barometric price leadership often develops as a reaction to a long economic warfare in which all the firms are losers.

Nature of Non-Price Competition in Oligopoly Markets

It is obvious from the foregoing discussion that oligopolists may be reluctant to wage price war between themselves and encroach upon each other's market share. There is therefore generally an absence of price competition in the oligopolistic market structure. The absence of price competition should not mean the absence of competition among oligopoly firms. In fact, the competition among oligopoly firms takes different forms of *non-price competition.* Yet, there are two important techniques of non-price competition.

First, non-price competition involves product differentiation which is intended to attract new customers and the customers of the other firms by creating preference for the new design and variety of product.

Second, perhaps the most important technique of non-price competition is advertisement. The primary objective of advertising is to make the demand curve for the product shift upward. The sellers try to capture the market of other sellers through advertising. Advertising is also necessary to retain market-share in the face of tough competition between the firms.

Baumol's Theory of Sales Maximization

We have so far discussed the theories of price determination in different kinds of market under the assumption that profit maximization is the basic objective of the business firms. Modern economists have questioned the profit maximization assumption and have formulated theories of firms' behaviour—known as alternative theories of firm. Baumol's theory of sales maximization is one of the important theories of this category. In this section, we discuss the theory of firm's behaviour under the assumption that firms seek to maximize their sales—not profit. This theory was propounded by William J. Baumol.

Baumol's theory of sales maximization is an important alternative theory of the firm's behaviour. The basic assumption of this theory is the *sales maximization,* rather than *profit maximization.* This seems to be the most plausible goal of the business firms. Baumol argues that there is no reason to believe that all firms seek to maximize their profits. Business firms, in fact, pursue a number of incompatible objectives and it is not easy to single out one as the most common objectives pursued by the firms. However, from his experience as a consultant to many big business houses, Baumol finds that most managers seek to maximize sales revenue rather than profits. He argues that in modern business, management is separated from ownership and managers enjoy discretion to pursue goals other than profit maximization. Their discretion eventually falls in favour of sales maximization.

According to Baumol, business managers pursue the goal of sales maximization for the following reasons.

First*,* financial institutions consider sales an an index of performance of the firm and prefer to finance only the firm with growing sales.

Second*,* while profit figures are available only annually, periodic sales figures can be obtained easily and more frequently, to assess the performance of a management. Maximization of sales is more satisfying for the managers than the maximization of profits, which go to the pockets of the shareholders.

Third*,* the routine evaluation of managers' performance, salaries and slack earnings of the top managers are linked more closely to sales than to profit.

Fourth*,* the routine personnel problems are more easily handled with growing sales. Higher payments may be offered to employees if sales figures indicate better performance. Profits are generally known after a year. To rely on profit figures means, therefore, a longer waiting period for both employees and management.

Fifth*,* if profit maximization is the goal and it rises in one period to an unusually high level, this becomes the standard profit target for the shareholders, which managers find very difficult to maintain in the long-run.

Finally*,* sales growing more than proportionately to market expansion, indicate growing market share and a greater competitive strength and bargaining power of a firm in a collusive oligopoly.

To formulate his theory of sales maximisation, Baumol has developed two basic models: (*i*) Static Model, and (*ii*) Dynamic Model, each with and without advertising. His static models with and without advertising are discussed below.

(i) Baumol's Model Without Advertising

Baumol assumes cost and revenue curves to be given as in conventional theory of pricing. Suppose that the total cost (*TC*) and the total revenue (*TR*) are given as shown in Fig. 14.9. The total profit curve, *TP,* is obtained by plotting the difference between the *TR* and *TC* curves. Profit is zero where *TR* = *TC*. Given

the *TR* and *TC* curves, there is a unique level of output at which total sales revenue is maximum. The total sales revenue is maximum where elasticity of demand equals one (i.e., $e = 1$ and the slope of the *TR* curve (i.e., $MR = \partial TR/\partial Q$) is equal to zero. Such a point lies at the highest point of *TR*. The point *H* on the *TR* in Fig. 14.9 represents the total maximum revenue, HQ_3. In Baumol's theory, the *profit maximizing rule, MR = MC*, is replaced by *sales maximization* rule.

At output OQ_3, the firm maximises its total revenue and makes profit $HM = TQ_3$. If this profit is enough or more than enough to satisfy the stockholders, the firm will produce output OQ_3 and charge a price $= HQ_3/OQ_3 = TR$/Total sale. But, if profit at output OQ_3 is not enough to satisfy the stockholders, then the firm's output must be changed to a level at which it makes a satisfactory profit. Suppose, the shareholders fix a minimum profit target at $OP_2 = LQ_2$. Then the manager will set output target at OQ_2 which yields a profit LQ_2 which is greater than profit TQ_3 – the profit with sales, revenue maximization.
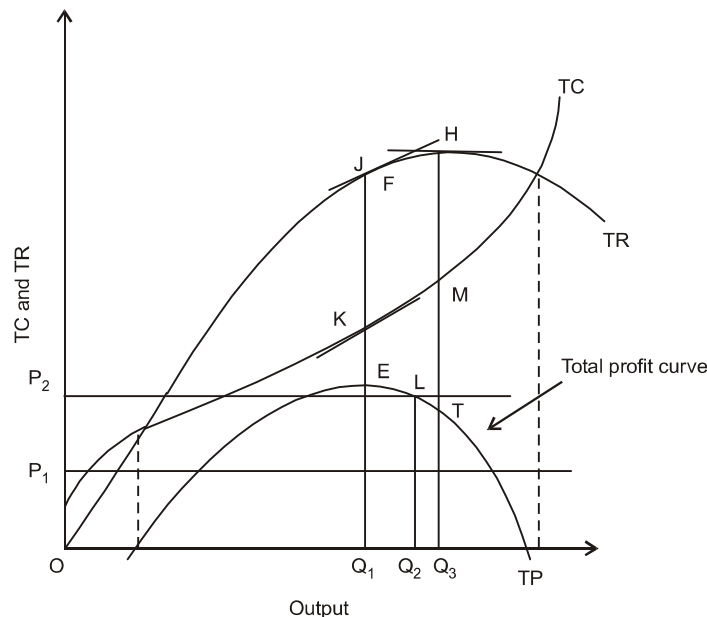
Fig. 14.9 *Sales Revenue Maximization*

Thus, there are two types of equilibria which appear to be possible: *first* in which the profit constraint does not provide an effective barrier to sales maximization, and *second,* in which profit maximization does provide an effective barrier to sales maximization. In the second type of equilibrium, the firm will have to produce an output which yields a satisfactory profit. It may be any output between OQ_1 and OQ_3. For example, if minimum required profit is OP_1, then the firm will stick to its sales maximization goal and produce output OQ_3 which yields a profit much greater than the required minimum.

Profit constraint and revenue maximization: In case actual profit (TQ_3) is much greater than the minimum required (say, OP_1), the minimum profit constraint

is not operative. But, if the required minimum profit level is OP_2, output OQ_3 will not yield sufficient profit to meet the profit requirement. The firm will, therefore, have to produce an output which yields the required minimum level of profit OP_2 ($= LQ_2$). Thus, the firm will produce OQ_2 where its profit is just sufficient to meet the requirement of minimum profit. This output OQ_2 is less than the sales maximization output OQ_3. Evidently the profit maximization output, OQ_1 is less than the sales maximization output OQ_2 (with profit constraint).

(ii) Baumol's Model with Advertising

We have explained above Baumol's model of price and output determination in a static model without advertising. In an oligopolistic market structure, however, prices and output are subject to non-price competition. Baumol considers, in his model, *advertising* as the typical form of non-price competition and suggests that the various forms of non-price competition may be analysed on similar lines.

In his analysis of advertising, Baumol makes the following assumptions.

(*a*) Firm's objective is to maximize sales, subject to a minimum profit constraint;

(*b*) Advertising causes a rightward shift in the demand curve and, hence, the total sales revenue (*TR*) increases with increase in advertisement expenditure (*A*), i.e., $\partial TR/\partial A > 0$;

(*c*) Price remains constant—a simplifying assumption; and

(*d*) Production costs are independent of advertising. This, too, is an unrealistic assumption since increase in sales may put output at a different cost structure.

Baumol's model with advertising is presented in Fig. 14.10. The *TR* and *TC* are measured on the Y-axis and total advertisement outlay on the X-axis. The *TR* curve is drawn on the assumption that advertising increases total sales in the same manner as price reduction.
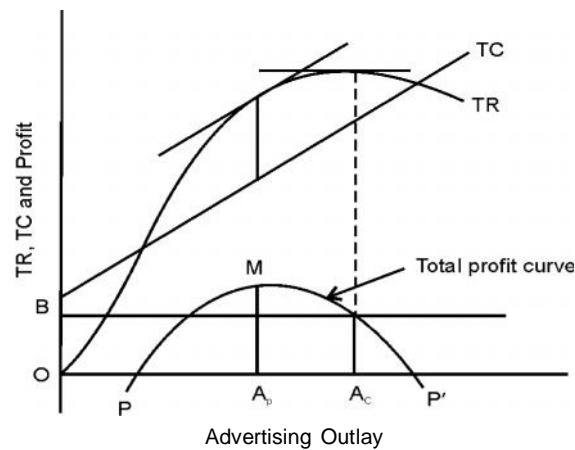

Fig. 14.10 *Baumol's Model with Advertising*

The *TC* curve includes both production cost and advertising outlays. The total profit curve *PP'* is drawn by substracting the *TC* from the *TR*. As shown in

Fig. 14.10, profit maximising expenditure is OA_p where the maximum profit is MA_p. Assuming that minimum profit required is *OB,* the sales maximizing advertisement outlay would be OA_c. This implies that a firm increases its advertisement outlay until it reaches the profit constraint level, i.e., until profits are reduced to the minimum acceptable level. This also means that sales maximisers advertise no less, rather more, than the profit maximisers.

Criticism of Baumol's Theory

Baumal's theory of sales maximization has been critcized on the following grounds.

First*,* it has been argued that in the long-run, Baumol's sales maximization hypothesis and the conventional hypothesis would yield identical results, because the *minimum required* level of profits would coincide with the normal level of profits.

Second*,* Baumol's theory does not distinguish between firm's equilibrium and industry equilibrium. Nor does it establish industry's equilibrium when all the firm's are sales maximisers.

Third*,* it does not clearly bring out the implications of inter-dependence of firm's price and output decisions. Thus, Baumol's theory ignores not only actual competition between the firms but also the threat of potential competition in an oligopolistic market.

Finally*,* Baumol's claim that his solution is preferable to the solutions offered by the conventional theory of price and output determination is not necessarily valid at least, from social welfare point of view.

Oligopoly and the Game Theory

In the preceding sections, we have discussed the classical models of strategic action and reaction among the oligopoly firms and the cartel system of price and output determination. We have also noted that none of the models explains satisfactorily the strategic actions and reactions of and interaction among the oligopoly firms mainly because of uncertainty and unpredictability of firms' behaviour. But the search for a reasonable solution to his problem does not end there. Classical theories show, in fact, only the beginning of the effort to analyse the determination of the profit maximising price and output in an oligopolistic market setting.

In this section, we discuss the *game theory approach* to explain the **strategic interaction** among the oligopoly firms. This approach uses the apparatus of *game theory*—a mathematical technique—to show how oligopoly firms play their game of business. The first systematic attempt was made in this field by John Von Neumann and Margenstern. Though their work was followed by many others, Martin Shubik is regarded as the 'most prominent proponent of the game-theory approach' who 'seems to believe that the only hope for the development of a general theory of oligopoly is the games theory'. However, his hope does not seem to be borne out by further attempts in this area. Nevertheless, the usefulness

of game theory in revealing the intricate behavioural pattern of the oligopoly firms cannot be denied. Here, we present an elementary description of the game theory as applied to oligopoly. We will first illustrate the nature of the problem faced by the oligopoly firms in their strategy formulation.

The Nature of the Problem: Prisoners' Dilemma Model

The nature of the problem faced by the oligopoly firm is best explained by the *Prisoners' Dilemma Game.* To illustrate *prisoners' dilemma,* let us suppose that there are two bookies, Ranga and Billa, who are partners in an illegal activity of match fixing. On a tip-off, the CBI arrests Ranga and Billa on suspicion of their involvement in fixing cricket matches. They are arrested and lodged in separate jails with no possibility of communication between them. They are being interrogated separately by the CBI officials with the following conditions disclosed to them in isolation.

1. If you confess your involvement in match fixing, you will get a 5-year imprisonment.

2. If you deny your involvement and your partner denies too, you will be set free for lack of evidence.

3. If you confess and turn approver, and your partner does not, then you get 2-year imprisonment, and your partner gets 10-year imprisonment.

Given these conditions, each suspect has two options open to him: (*i*) to confess, and (*ii*) not to confess. Now, both Ranga and Billa face a dilemma on how to decide whether or not to confess. While taking a decision, both have a common objective, i.e., to minimise the period of imprisonment. Given this objective, the option is quite simple that both of them deny their involvement in match fixing. But, there is no certainty that if one denies, the other will also deny: the other may confess and turn approver. With this uncertainty, the dilemma in making a choice still remains. For example, if Ranga denies his involvement, and Billa confesses (settles for a 2-year imprisonment), and turn approver, then Ranga gets a 10-year jail term. So is the case with Billa. If they both confess, then they get a 5-year jail term each. Then what to do? That is the dilemma. The nature of their problem of decision making is illustrated in the following Table 14.2 in the form of a 'pay-off matrix'. The pay-off matrix shows the pay-offs of their different options in terms of the number of years in jail.

Table 14.2 *Prisoners' Dilemma*

		Billa's Options			
		Confess		*Deny*	
		Ranga	Billa	Ranga	Billa
	Confess	5	5	2	10
Ranga's Options		Ranga	Billa	Ranga	Billa
	Deny	10	2	0	0

Given the conditions, it is quite likely that both the suspects may opt for 'confession', because neither Ranga knows what Billa will do, nor Billa knows what Ranga will do. When they both confess, each gets a 5 year jail term. This is the second best option. For his decision to confess, Ranga might formulate his strategy in the following manner. He reasons out: if I confess (though I am innocent), I will get a maximum of 5 year's imprisonment. But, if I deny (which I must) and Billa confesses and turns approver, I will get 10 year's imprisonment. And, that will be the worst of the worst. It is quite likely that suspect Billa also reasons in the same manner, *even* if he too is innocent. If they both confess, they would *avoid* 10 year's imprisonment, the maximum possible jail sentence under the law. This is the best they could achieve under the given conditions.

Application of Game Theory to Oligopolistic Strategy

Relevance of prisoners' dilemma to oligopoly

The prisoners' dilemma illustrates the nature of problems oligopoly firms are confronted with in the formulation of their business strategy with respect to strategic advertising, price cutting and cheating in case of a cartel. Look at the nature of the problems an oligopoly firm is faced with when it plans to increase its *advertisement* expenditure (ad-expenditure for short). The basic issue is whether or not to increase the ad-expenditure. If the answer is 'do not increase', then the questions are: Will the rival firms increase ad-expenditure or will they not? And if they do, what will be the consequences for the firm under consideration? And, if the answer is 'increase', then the following questions arise: What will be the reaction of the rival firms? Will they increase or will they not increase their ad-expenditure? What will be the pay-off if they do not and what if they do? If the rival firms do increase their advertising, what will be the pay-off to the firm? Will the firm be a net gainer or a net loser? The firm will *have* to find the answer to these queries under the conditions of uncertainty. It will *have* to anticipate actions, reactions and counteraction by the rival firms and chalk out its own strategy. It is in case of such problems that the case of prisoners' dilemma becomes applicable to oligopoly.

Let us now apply the game theory to our example of 'whether or not to increase ad-expenditure', assuming that there are only two television companies, Sony and Suzuki, i.e., the case of a duopoly. We know that in all the games, the players have to anticipate the move made by the opposite player(s) and formulate their own strategy to counter the different possible moves by the rival. To apply the game theory to the case of 'whether or not to increase ad-expenditure' a company needs to know or anticipate the following:

   (*i*) counter moves by the rival company in response to increase in ad-expenditure by this company, and

   (*ii*) the *pay-offs* of this strategy when (*a*) the rival company does not react, and (*b*) the rival company does react and makes a counter move by increasing its ad-expenditure.

After this data is obtained, the company will have to decide on the best possible strategy for playing the game and achieving its objective of, say, increasing sales and capturing a larger share of the market. The best possible strategy in game theory is called the 'dominant strategy'. *A dominant strategy* is *one that gives optimum pay-off, no matter what the opponent does.* Thus, the basic objective of applying the game theory is to arrive at the *dominant strategy*.

Suppose that the possible outcomes of the ad-game under the rival's alternative moves are given in the pay-off matrix presented in Table 14.3. In the figure, *A* indicates Sony's gain and *B* indicates Suzuki's gain in terms of increase in sales.

Table 14.3 *Pay-off Matrix of the Ad-Game (Increase in sales in million ₹)*

		Suzuki's Options			
		Increase Ad		*Don't increase*	
		A	B	A	B
	Increase Ad	20	10	30	0
Sony's Strategy		A	B	A	B
	Don't increase	10	15	15	5

As the matrix shows, if Sony decides to increase its ad-expenditure and Suzuki counteracts Sony's move by increasing its own ad-expenditure, Sony's sales go up by ₹ 20 million and that of Suzuki by ₹ 10 million. And, if Sony increases its advertisement and Suzuki does not, then Sony's sales gain is ₹ 30 million and no gain to Suzuki. One can similarly find the pay-offs of the strategy 'Don't increase' in case of both of firms.

Given the pay-off matrix, the question arises as to what strategy should Sony choose to optimize its gain from extra ad-expenditure, irrespective of moves of the rival Suzuki. It is clear from the pay-off matrix that Sony will choose the strategy of increasing the ad-expenditure because, no matter what Suzuki does, its sales increase by at least ₹ 20 million. This is, therefore, the *dominant strategy* for Sony. A better situation could be that when Sony increases its expenditure on advertisement, Suzuki does not. In that case, Sony's sales could increase by ₹ 30 million and sales of Suzuki do not increase. But there is a greater possibility that Suzuki will go for counter-advertising, in anticipation of losing a part of its market to Sony in future. Therefore, a strategy based on the assumption that Suzuki will not increase its ad-expenditure, involves a high degree of uncertainty.

Nash equilibrium. In the preceding section, we have used a very simple example to illustrate the application of game theory to an oligopolistic market setting, with simplifying assumptions: (*i*) that strategy formulation is a one-time affair, (*ii*) that one company initiates the competitive warfare and other companies only react; and (*iii*) that there exists a *dominant strategy—a* strategy which gives an optimum solution. The real-life situation is, however, much more complex. There is a continuous one-to-one and tit-for-tat kind of warfare. Actions, reactions and counteractions are regular phenomena. Under these conditions, a *dominant*

strategy is often non-existent. To analyse this kind of situation, John Nash, an American mathematician, developed a technique, known as *Nash equilibrium. Nash equilibrium technique* seeks to establish that each company does the best it can, given the strategy of its competitors. Thus, the *Nash equilibrium is one in which none of the players can improve their pay-off given the strategy of the other players.* In case of our example, Nash equilibrium can be defined as one in which none of the companies can increase its pay-off (sales) given the strategy of the rival company.

The Nash equilibrium can be illustrated by making some modifications in the pay-off matrix given in Table 14.3. Now we assume that action and counter-action between Sony and Suzuki is a regular phenomenon and that the pay-off matrix that appears finally is given in Table 14.4. The only change made in the modified pay-off matrix is that if neither Sony nor Suzuki increases its ad-expenditure, then pay-offs change from (15, 5) to (25, 5).

Table 14.4 *Pay-off Matrix of the Ad-Game (increase in sales in million ₹)*

A's Strategy		B's Options			
		Increase Ad		*Don't increase*	
		Sony	Suzuki	Sony	Suzuki
	Increase Ad	20	10	30	0
	Don't increase	Sony	Suzuki	Sony	Suzuki
		10	15	25	5

It can be seen from the pay-off matrix (Table 14.4) that Sony has no more a *dominant strategy.* Its optimum decision depends now on what Suzuki does. If Suzuki increases its ad-expenditure, Sony has no option but to increase its advertisement expenditure. And, if Sony reinforces its advertisement, Suzuki will have to follow the suit. On the other hand, if Suzuki does not increase its ad-expenditure, Sony does the best by increasing its ad-expenditure. Under these conditions, the conclusion that both the companies arrive at is to increase ad-expenditure if the other company does so, and 'don't increase', if the competitor 'does not increase'. In the ultimate analysis, however, both the companies will decide to increase the ad-expenditure. The reason is that if none of the companies increases advertisement, Sony gains more in terms of increase in its sales (₹ 25 million) and the gain of Suzuki is much less (₹ 5 million only). And, if Suzuki increases advertisement expenditure, its sales increase by ₹ 10 million. Therefore, Suzuki would do best to increase its ad-expenditure. In that case, Sony will have no option but to increase its ad-expenditure. Thus, the *final conclusion* that emerges is that both the companies will go for advertisement war. In that case, each company finds that it is doing the best, given what the rival company is doing. This is what the Nash equilibrium establishes.

However, there are situations in which there can be more than one Nash equilibrium. For example, if we change the pay-off in the south-east corner from (25, 5) to (22, 8) each firm may find it worthless to wage advertisement war and may settle for 'don't increase' situation. Thus, there are two possible Nash equilibria.

Concluding Remarks

What we have presented above is an elementary introduction to the game theory. It can be used to find equilibrium solution to the problems of oligopolistic market setting under different assumptions regarding the behaviour of the oligopoly firms and market conditions. However, despite its merit of revealing the nature and pattern of oligopolistic warfare, game theory often fails to provide a determinate solution.

Check Your Progress

1. Define oligopoly.
2. List any two features of oligopoly.

14.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Oligopoly is a form of market in which there are a few sellers selling homogeneous or differentiated products.

2. Two features of oligopoly are the following;
   - Small number of sellers
   - Barriers to entry

14.6 SUMMARY

- Oligopoly is a form of market in which there are a few sellers selling homogeneous or differentiated products.

- In any case, if oligopoly firms sell a homogeneous product, it is called pure or homogeneous oligopoly.

- The most striking feature of an oligopolistic market structure is the interdependence of oligopoly firms.

- The famous models of oligopoly include Cournot's duoply model (1838), Bertrand's leadership model (1880), Edgeworth's duoply model (1897), Stackelberg's model (1933), Sweezy's kinked demand curve model (1939), Neumann and Margenstern Game Theory model (1944), Baumol's sales maximization model (1959).

- Augustine Cournot, a French economist, was the first to develop a formal oligopoly model in 1838. He formulated his oligopoly theory in the form of a *duopoly model* which can be extended to oligopoly model

- The origin of kinked-demand curve can be traced into Chamberlin's theory of monopolistic competition. Later, Hall and Hitch used kinked-demand curve to explain rigidity of prices in oligopolistic market.

- The models that deal with the collusive oligopolies are called collusive oligopoly models. Collusion between firms may take many forms depending on their relative strength and objective of collusion, and on whether collusion is legal or illega1.

- A cartel performs a variety of services for its members. The two services of central importance are (*i*) fixing price for joint profit maximization; and (*ii*) market sharing between its members.

- The market-sharing cartels are more common because this kind of collusion permits a considerable degree of freedom in respect of style and design of the product, advertising and other selling activities. There are two main methods of market allocations: (*a*) non-price competition agreement, and (*b*) quota system.

- Price leadership is an imperfect form of collusion between oligopoly firms. Price leadership is an informal position given to or attained by a firm in an oligopolistic setting to lead other firms in pricing.

- Baumol's theory of sales maximization is an important alternative theory of the firm's behaviour. The basic assumption of this theory is the *sales maximization,* rather than profit maximization.

- The Nash equilibrium is one in which none of the players can improve their pay-off given the strategy of the other players.

14.7 KEY WORDS

- **Duopoly:** It is a form of oligopoly where only two sellers exist in one market.

- **Cartel:** It is a formal organisation of the oligopoly firms in an industry.

14.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short Answer Questions

1. Write a short note on the main sources of oligopoly.

2. Mention the features of oligopoly.

3. What are the criticisms raised against Sweezy's Kinked-demand curve model?

4. What are the price leadership models of oligopoly?

Long Answer Questions

1. Prepare algebraic solution of duopoly.

2. Critically analyse Chamberlin's model of oligopoly.

3. Discuss price and output determination in collusive oligopoly.

4. Describe Baumol's models with and without advertising.

5. Evaluate application of game theory to oligopolistic strategy.

14.7 FURTHER READINGS

Dwivedi, D. N. 2008. *Principles of Economics*, Seventh Edition. New Delhi: Vikas Publishing House.

Weil. David N. 2004. *Economic Growth*. London: Addison Wesley.

Thomas, Christopher R. and Maurice S. Charles. 2005. *Managerial Economics: Concepts and Applications*, Eighth Edition. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Mankiw, Gregory N. 2002. *Principles of Economics*, Second Edition. India: Thomson Press.