1. Subject Name: Managerial Economics
2. Semester / Year: Second (Semester II)
3. Name of the Teacher: Dr. Sudip Ghosh
4. Name of the topic: Theory of Costs

4. Theory of Costs

I. GENERAL NOTES

Cost functions are derived functions. They are derived from the production function, which describes the available efficient methods of production at any one time.

Economic theory distinguishes between short-run costs and long-run costs. *Short-run costs* are the costs over a period during which some factors of production (usually capital equipment and management) are fixed. The *long-run costs* are the costs over a period long enough to permit the change of all factors of production. In the long run all factors become variable.

Both in the short run and in the long run, total cost is a multivariable function, that is, total cost is determined by many factors. Symbolically we may write the long-run cost function as

$$C = f(X, T, P_f)$$

and the short-run cost function as

$$C = f(X, T, P_f, \overline{K})$$

where C = total cost
X = output
T = technology
P_f = prices of factors
$\overline{K}$ = fixed factor(s)

Graphically, costs are shown on two-dimensional diagrams. Such curves imply that cost is a function of output, $C = f(X)$, *ceteris paribus*. The clause *ceteris paribus* implies that all other factors which determine costs are constant. If these factors do change, their effect on costs is shown graphically by a shift of the cost curve. This is the reason why determinants of cost, other than output, are called *shift factors*. Mathematically there is no difference between the various determinants of costs. The distinction between movements along the cost curve (when output changes) and shifts of the curve (when the other determinants change) is convenient only pedagogically, because it allows the use of two-dimensional diagrams. But it can be misleading when studying the determinants of costs. It is important to remember that if the cost curve shifts, this does not imply that the cost function is indeterminate.

The factor 'technology' is itself a multidimensional factor, determined by the physical quantities of factor inputs, the quality of the factor inputs, the efficiency of the entrepreneur, both in organising the physical side of the production (technical efficiency of the entrepreneur), and in making the correct economic choice of techniques (economic

efficiency of the entrepreneur). Thus, any change in these determinants (e.g., the introduction of a better method of organisation of production, the application of an educational programme to the existing labour) will shift the production function, and hence will result in a shift of the cost curve. Similarly the improvement of raw materials, or the improvement in the use of the same raw materials will lead to a shift downwards of the cost function.

The short-run costs are the costs at which the firm operates in any one period. The long-run costs are *planning costs* or *ex ante costs*, in that they present the optimal possibilities for expansion of the output and thus help the entrepreneur *plan* his future activities. Before an investment is decided the entrepreneur is in a long-run situation, in the sense that he can choose any one of a wide range of alternative investments, defined by the state of technology. After the investment decision is taken and funds are tied up in fixed-capital equipment, the entrepreneur operates under short-run conditions: he is on a short-run cost curve.

A distinction is necessary between *internal (to the firm) economies of scale* and *external economies*. The internal economies are built into the shape of the long-run cost curve, because they accrue to the firm from its own action as it expands the level of its output. (See section II below.) The external economies arise outside the firm, from improvement (or deterioration) of the environment in which the firm operates. Such economies external to the firm may be realised from actions of other firms in the same or in another industry. The important characteristic of such economies is that they are independent of the actions of the firm, they are *external* to it. Their effect is a change in the prices of the factors employed by the firm (or in a reduction in the amount of inputs per unit of output), and thus cause a shift of the cost curves, both the short-run and the long-run.

In summary, while the internal economies of scale relate only to the long-run and are built into the shape of the long-run cost curve, the external economies affect the position of the cost curves: both the short-run and the long-run cost curves will shift if external economies affect the prices of the factors and/or the production function.

Any point on a cost curve shows the minimum cost at which a certain level of output may be produced. This is the *optimality* implied by the points of a cost curve. Usually the above optimality is associated with the long-run cost curve. However, a similar concept may be applied to the short-run, given the plant of the firm in any one period.

In the section II of this chapter we examine the traditional theory of U-shaped costs. In section III we examine some recent developments in the theory of costs which reject the strict U shape of the short-run cost curves on the grounds that its assumptions are not realistic, and question the 'envelope' long-run cost curve on the grounds that diseconomies are not a necessary consequence of large-scale operations.[1] In section V we examine the main types and sources of economies of scale. In section VI we summarise the available empirical evidence on the shape of the long-run and the short-run cost curves. Finally, in section VII we discuss briefly the relevance of the shape of cost curves in decision-making.

II. THE TRADITIONAL THEORY OF COST

Traditional theory distinguishes between the short run and the long run. The short run is the period during which some factor(s) is fixed; usually capital equipment and entrepreneurship are considered as fixed in the short run. The long run is the period over which all factors become variable.

[1] In section IV we discuss the engineering cost curves.

A. SHORT-RUN COSTS OF THE TRADITIONAL THEORY

In the traditional theory of the firm total costs are split into two groups: total fixed costs and total variable costs:

$$TC = TFC + TVC$$

The fixed costs include:

(*a*) salaries of administrative staff
(*b*) depreciation (wear and tear) of machinery
(*c*) expenses for building depreciation and repairs
(*d*) expenses for land maintenance and depreciation (if any).

Another element that may be treated in the same way as fixed costs is the normal profit, which is a lump sum including a percentage return on fixed capital and allowance for risk.

The variable costs include:

(*a*) the raw materials
(*b*) the cost of direct labour
(*c*) the running expenses of fixed capital, such as fuel, ordinary repairs and routine maintenance.

The total fixed cost is graphically denoted by a straight line parallel to the output axis (figure 4.1). The total variable cost in the traditional theory of the firm has broadly an inverse-S shape (figure 4.2) which reflects the law of variable proportions. According to this law, at the initial stages of production with a given plant, as more of the variable factor(s) is employed, its productivity increases and the average variable cost falls.

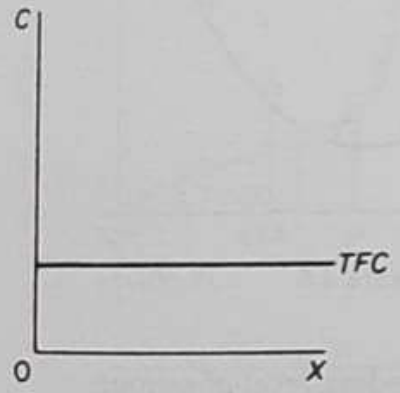

Figure 4.1

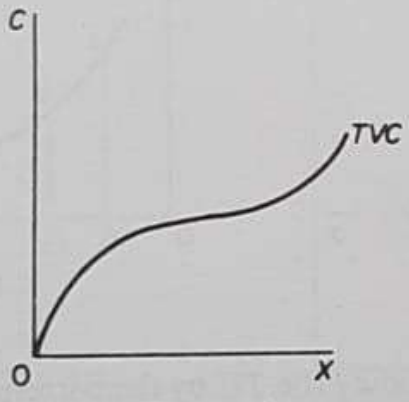

Figure 4.2

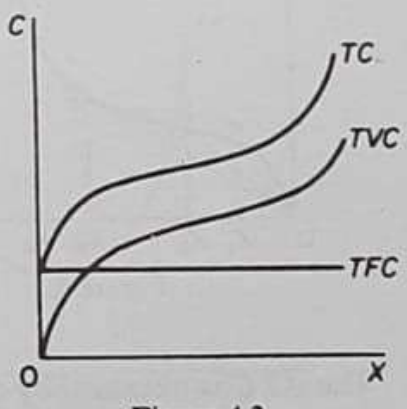

Figure 4.3

This continues until the optimal combination of the fixed and variable factors is reached. Beyond this point as increased quantities of the variable factor(s) are combined with the fixed factor(s) the productivity of the variable factor(s) declines (and the AVC rises). By adding the TFC and TVC we obtain the TC of the firm (figure 4.3). From the total-cost curves we obtain average-cost curves. The average fixed cost is found by dividing TFC by the level of output:

$$AFC = \frac{TFC}{X}$$

Graphically the AFC is a rectangular hyperbola, showing at all its points the same magnitude, that is, the level of TFC (figure 4.4). The average variable cost is similarly obtained by dividing the TVC with the corresponding level of output:

$$AVC = \frac{TVC}{X}$$

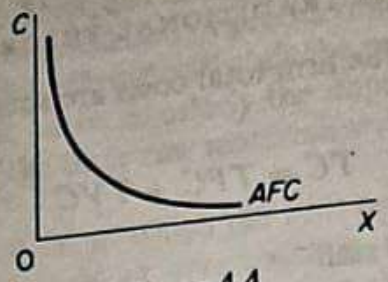

Figure 4.4

Graphically the AVC at each level of output is derived from the slope of a line drawn from the origin to the point on the TVC curve corresponding to the particular level of output. For example, in figure 4.5 the AVC at X_1 is the slope of the ray $0a$, the AVC at X_2 is the slope of the ray $0b$, and so on. It is clear from figure 4.5 that the slope of a ray through the origin declines continuously until the ray becomes tangent to the TVC curve at c. To the right of this point the slope of rays through the origin starts increasing. Thus the $SAVC$ curve falls initially as the productivity of the variable factor(s) increases, reaches a minimum when the plant is operated optimally (with the optimal combination of fixed and variable factors), and rises beyond that point (figure 4.6).

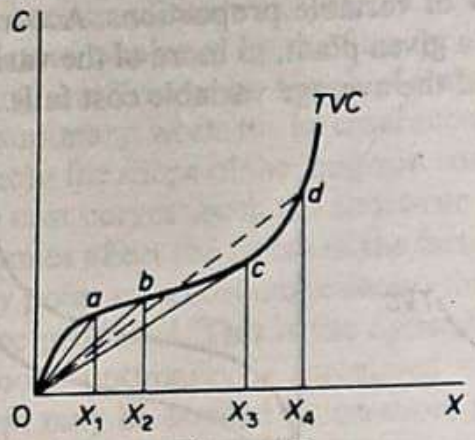

Figure 4.5

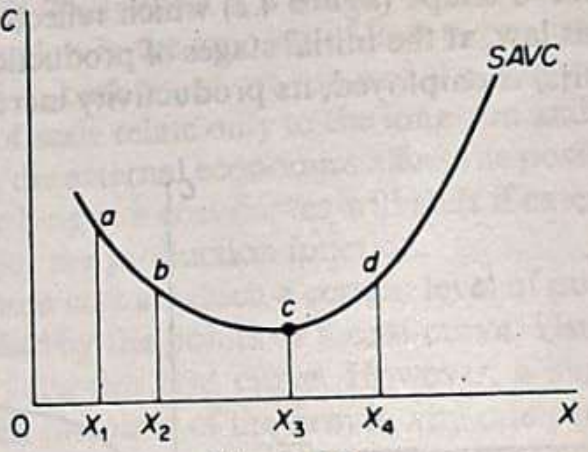

Figure 4.6

The ATC is obtained by dividing the TC by the corresponding level of output:

$$ATC = \frac{TC}{X} = \frac{TFC + TVC}{X} = AFC + AVC$$

Graphically the ATC curve is derived in the same way as the $SAVC$. The ATC at any level of output is the slope of the straight line from the origin to the point on the TC curve corresponding to that particular level of output (figure 4.7). The shape of the ATC is similar to that of the AVC (both being U-shaped). Initially the ATC declines, it reaches a minimum at the level of optimal operation of the plant (X_M) and subsequently rises again (figure 4.8). The U shape of both the AVC and the ATC reflects the *law of variable proportions* or *law of eventually decreasing returns* to the variable factor(s) of production (see Chapter 3). The marginal cost is defined as the change in TC which results from a unit change in output. Mathematically the marginal cost is the first derivative of the TC function. Denoting total cost by C and output by X we have

$$MC = \frac{\partial C}{\partial X}$$

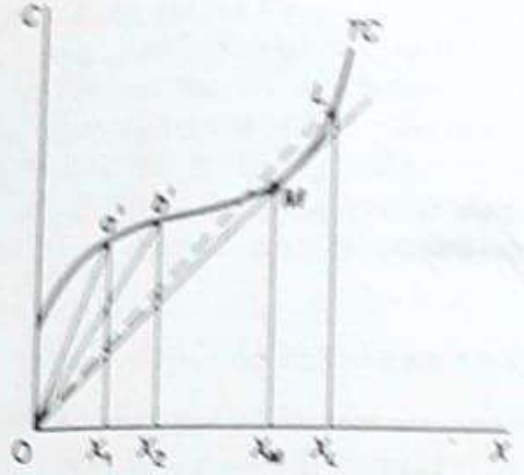

Figure 4.7

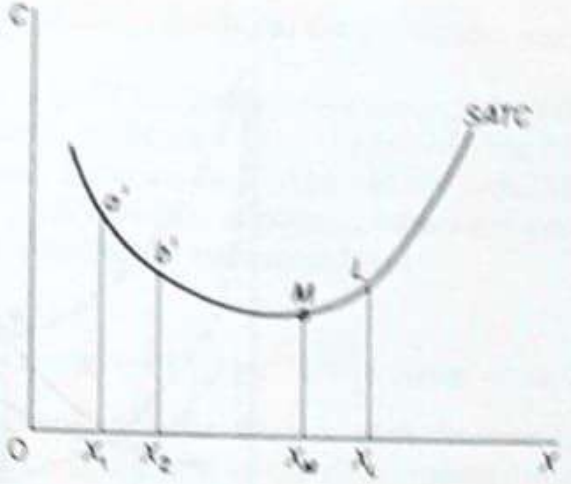

Figure 4.8

Graphically the *MC* is the slope of the *TC* curve (which of course is the same at any point as the slope of the *TVC*). The slope of a curve at any one of its points is the slope of the tangent at that point. With an inverse-S shape of the *TC* (and *TVC*) the *MC* curve will be U-shaped. In figure 4.9 we observe that the slope of the tangent to the total-cost curve declines gradually, until it becomes parallel to the X-axis (with its slope being equal to zero at this point), and then starts rising. Accordingly we picture the *MC* curve in figure 4.10 as U-shaped.

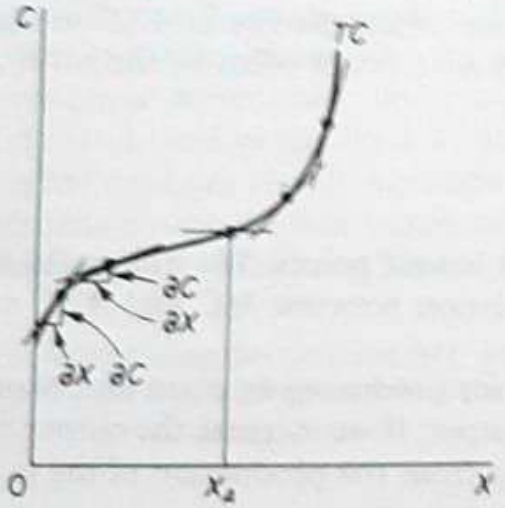

Figure 4.9

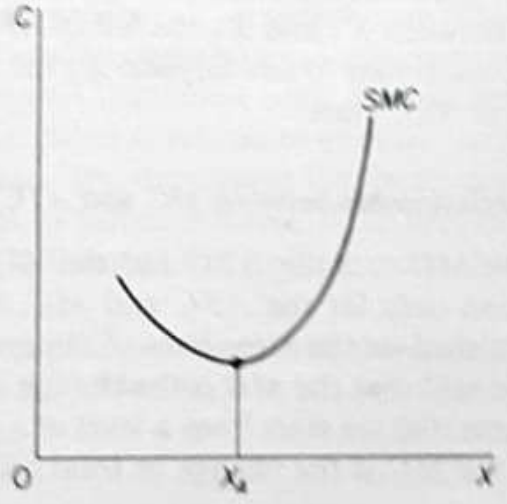

Figure 4.10

In summary: the traditional theory of costs postulates that in the short run the cost curves (*AVC*, *ATC* and *MC*) are U-shaped, reflecting the law of variable proportions. In the short run with a fixed plant there is a phase of increasing productivity (falling unit costs) and a phase of decreasing productivity (increasing unit costs) of the variable factor(s). Between these two phases of plant operation there is a *single point* at which unit costs are at a minimum. When this point on the *SATC* is reached the plant is utilised optimally, that is, with the optimal combination (proportions) of fixed and variable factors.

The relationship between *ATC* and *AVC*

The *AVC* is a part of the *ATC*, given $ATC = AFC + AVC$. Both *AVC* and *ATC* are U-shaped, reflecting the law of variable proportions. However, the minimum point of the *ATC* occurs to the right of the minimum point of the *AVC* (figure 4.11). This is

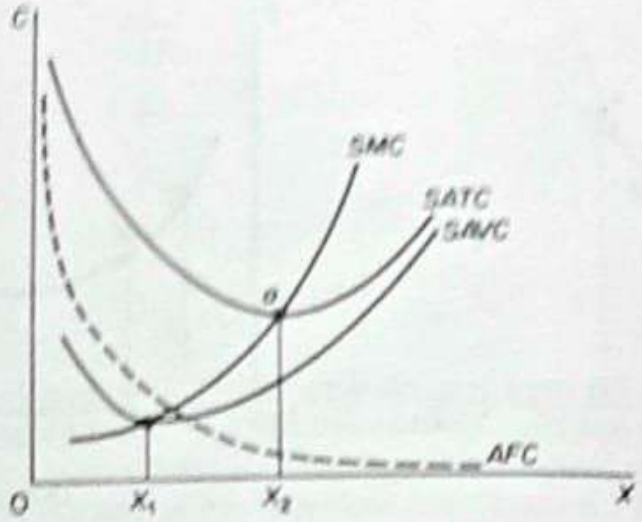

Figure 4.11

due to the fact that ATC includes AFC, and the latter falls continuously with increases in output. After the AVC has reached its lowest point and starts rising, its rise is over a certain range offset by the fall in the AFC, so that the ATC continues to fall (over that range) despite the increase in AVC. However, the rise in AVC eventually becomes greater than the fall in the AFC so that the ATC starts increasing. The AVC approaches the ATC asymptotically as X increases.

In figure 4.11 the minimum AVC is reached at X_1 while the ATC is at its minimum at X_2. Between X_1 and X_2 the fall in AFC more than offsets the rise in AVC so that the ATC continues to fall. Beyond X_2 the increase in AVC is not offset by the fall in AFC, so that ATC rises.

The relationship between MC and ATC

The MC cuts the ATC and the AVC at their lowest points. We will establish this relation only for the ATC and MC, but the relation between MC and AVC can be established on the same lines of reasoning.

We said that the MC is the change in the TC for producing an extra unit of output. Assume that we start from a level of n units of output. If we increase the output by one unit the MC is the change in total cost resulting from the production of the $(n+1)^{th}$ unit.

The AC at each level of output is found by dividing TC by X. Thus the AC at the level of X_n is

$$AC_n = \frac{TC_n}{X_n}$$

and the AC at the level X_{n+1} is

$$AC_{n+1} = \frac{TC_{n+1}}{X_{n+1}}$$

Clearly

$$TC_{n+1} = TC_n + MC$$

Thus:

(*a*) If the MC of the $(n+1)^{th}$ unit is less than AC_n (the AC of the previous n unit the AC_{n+1} will be smaller than the AC_n.

(*b*) If the MC of the $(n+1)^{th}$ unit is higher than AC_n (the AC of the previous n units) the AC_{n+1} will be higher than the AC_n.

So long as the MC lies below the AC curve, it pulls the latter downwards; when the MC rises above the AC, it pulls the latter upwards. In figure 4.11 to the left of a the MC lies below the AC curve, and hence the latter falls downwards. To the right of a the MC curve lie above the AC curve, so that AC rises. It follows that at point a, where the intersection of the MC and AC occurs, the AC has reached its minimum level.[1]

B. LONG-RUN COSTS OF THE TRADITIONAL THEORY: THE 'ENVELOPE' CURVE

In the long run all factors are assumed to become variable. We said that the long-run cost curve is a *planning curve*, in the sense that it is a guide to the entrepreneur in his decision to plan the future expansion of his output.

The long-run average-cost curve is derived from short-run cost curves. Each point on the LAC corresponds to a point on a short-run cost curve, which is tangent to the LAC at that point. Let us examine in detail how the LAC is derived from the SRC curves.

Assume, as a first approximation, that the available technology to the firm at a particular point of time includes three methods of production, each with a different plant size: a small plant, medium plant and large plant. The small plant operates with costs denoted by the curve SAC_1, the medium-size plant operates with the costs on SAC_2 and the large-size plant gives rise to the costs shown on SAC_3 (figure 4.12). If the firm plans to produce output X_1 it will choose the small plant. If it plans to produce X_2 it will choose the medium plant. If it wishes to produce X_3 it will choose the large-size plant. If the firm starts with the small plant and its demand gradually increases, it will produce at lower costs (up to level X_1'). Beyond that point costs start increasing. If its demand reaches the level X_1'' the firm can either continue to produce with the small plant or it can install the medium-size plant. The decision at this point depends not on costs but on the firm's expectations about its future demand. If the firm expects that the demand will expand further than X_1'' it will install the medium plant, because

[1] The relationship between the MC and AC curves becomes clearer with the use of simple calculus. Given $C = zX$, where $z = AC$. Clearly $z = f(X)$. The MC is

$$\frac{\partial C}{\partial X} = \frac{\partial (zX)}{\partial X}$$

Applying the rule of differentiation of 'a function of a function' (which states that if $y = uv$, where $u = f_1(x)$ and $v = f_2(x)$, then $dy/dx = dy/du \cdot du/dx$), we obtain

$$MC = \frac{\partial C}{\partial X} = z\frac{\partial X}{\partial X} + X\frac{\partial z}{\partial X}$$

or

$$MC = AC + (X)(\text{slope of } AC)$$

Given that $AC > 0$ and $X > 0$, the following results emerge:

(*a*) if (slope of AC) < 0, then $MC < AC$

(*b*) if (slope of AC) > 0, then $MC > AC$

(*c*) if (slope of AC) $= 0$, then $MC = AC$

The slope of the AC becomes zero at the minimum point of this curve (given that on theoretical grounds the AC curve is U-shaped). Hence $MC = AC$ at the minimum point of the average-cost curve.

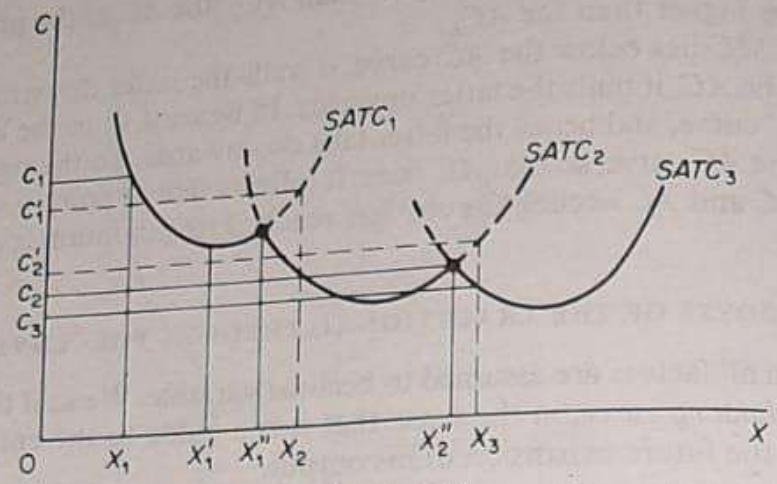

Figure 4.12

with this plant outputs larger than X_1'' are produced with a lower cost. Similar considerations hold for the decision of the firm when it reaches the level X_2''. If it expects its demand to stay constant at this level, the firm will not install the large plant, given that it involves a larger investment which is profitable only if demand expands beyond X_2''. For example, the level of output X_3 is produced at a cost c_3 with the large plant, while it costs c_2' if produced with the medium-size plant ($c_2' > c_3$).

Now if we relax the assumption of the existence of only three plants and assume that the available technology includes many plant sizes, each suitable for a certain level of output, the points of intersection of consecutive plants (which are the crucial points for the decision of whether to switch to a larger plant) are more numerous. In the limit, if we assume that there is a very large number (infinite number) of plants, we obtain a continuous curve, which is the planning *LAC* curve of the firm. Each point of this curve shows the minimum (optimal) cost for producing the corresponding level of output. The *LAC* curve is the locus of points denoting the least cost of producing the corresponding output. It is a planning curve because on the basis of this curve the firm decides what plant to set up in order to produce optimally (at minimum cost) the expected level of output. The firm chooses the short-run plant which allows it to produce the anticipated (in the long run) output at the least possible cost. In the traditional theory of the firm the *LAC* curve is U-shaped and it is often called the 'envelope curve' because it 'envelopes' the *SRC* curves (figure 4.13).

Let us examine the U shape of the *LAC*. This shape reflects the *laws of returns to scale* (see Chapter 3). According to these laws the unit costs of production decrease as plant

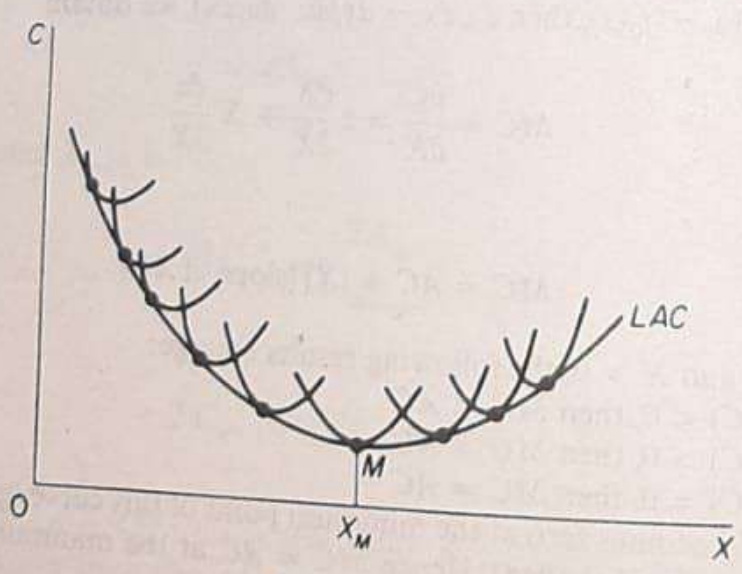

Figure 4.13

size increases, due to the economies of scale which the larger plant sizes make possible. The nature of economies of scale is discussed in section V below. The traditional theory of the firm assumes that economies of scale exist only up to a certain size of plant, which is known as the *optimum plant size*, because with this plant size all possible economies of scale are fully exploited. If the plant increases further than this optimum size there are diseconomies of scale, arising from managerial inefficiencies. It is argued that management becomes highly complex, managers are overworked and the decision-making process becomes less efficient. The turning-up of the LAC curve is due to managerial diseconomies of scale, since the technical diseconomies can be avoided by duplicating the optimum technical plant size (see section V).

A serious implicit assumption of the traditional U-shaped cost curves is that each plant size is designed *to produce optimally a single level of output* (e.g. 1000 units of X). Any departure from that X, no matter how small (e.g. an increase by 1 unit of X) leads to increased costs. The plant is completely inflexible. There is no reserve capacity, not even to meet seasonal variations in demand. As a consequence of this assumption the LAC curve 'envelopes' the $SRAC$. Each point of the LAC is a point of tangency with the corresponding $SRAC$ curve. The point of tangency occurs to the falling part of the $SRAC$ curves for points lying to the left of the minimum point of the LAC: since the slope of the LAC is negative up to M (figure 4.13) the slope of the $SRAC$ curves must also be negative, since at the point of their tangency the two curves have the same slope. The point of tangency for outputs larger than X_M occurs to the rising part of the $SRAC$ curves: since the LAC rises, the SAC must rise at the point of their tangency with the LAC. Only at the minimum point M of the LAC is the corresponding SAC also at a minimum. Thus at the falling part of the LAC the plants are not worked to full capacity; to the rising part of the LAC the plants are overworked; only at the minimum point M is the (short-run) plant optimally employed.

We stress once more the optimality implied by the LAC planning curve: each point represents the least unit-cost for producing the corresponding level of output. Any point above the LAC is inefficient in that it shows a higher cost for producing the corresponding level of output. Any point below the LAC is economically desirable because it implies a lower unit-cost, but it is not attainable in the current state of technology and with the prevailing market prices of factors of production. (Recall that each cost curve is drawn under a *ceteris paribus* clause, which implies given state of technology and given factor prices.)

The long-run marginal cost is derived from the $SRMC$ curves, but does not 'envelope' them. The $LRMC$ is formed from points of intersection of the $SRMC$ curves with vertical lines (to the X-axis) drawn from the points of tangency of the corresponding SAC curves and the LRA cost curve (figure 4.14). The LMC must be equal to the SMC for the output at which the corresponding SAC is tangent to the LAC. For levels of X to the left of tangency a the $SAC > LAC$. At the point of tangency $SAC = LAC$. As we move from point a' to a, we actually move from a position of inequality of $SRAC$ and $LRAC$ to a position of equality. Hence the change in total cost (i.e. the MC) must be smaller for the short-run curve than for the long-run curve. Thus $LMC > SMC$ to the left of a. For an increase in output beyond X_1 (e.g. X_1'') the $SAC > LAC$. That is, we move from the position a of equality of the two costs to the position b where SAC is greater than LAC. Hence the addition to total cost ($=MC$) must be larger for the short-run curve than for the long-run curve. Thus $LMC < SMC$ to the right of a.

Since to the left of a, $LMC > SMC$, and to the right of a, $LMC < SMC$, it follows that at a, $LMC = SMC$. If we draw a vertical line from a to the X-axis the point at which it intersects the SMC (point A for SAC_1) is a point of the LMC.

If we repeat this procedure for all points of tangency of $SRAC$ and LAC curves to the left of the minimum point of the LAC, we obtain points of the section of the LMC

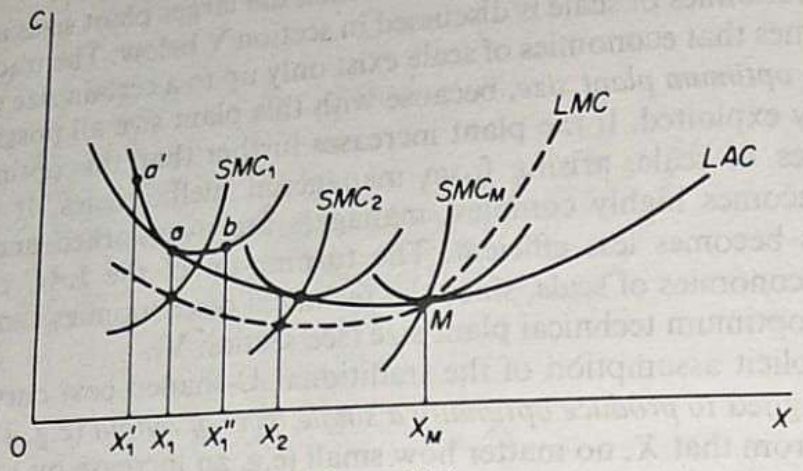

Figure 4.14

which lies below the LAC. At the minimum point M the LMC intersects the LAC. To the right of M the LMC lies above the LAC curve. At point M we have

$$SAC_M = SMC_M = LAC = LMC$$

There are various mathematical forms which give rise to U-shaped unit cost curves. The simplest total cost function which would incorporate the law of variable proportions is the cubic polynomial

$$\begin{aligned} C &= \underbrace{b_0}_{} + \underbrace{b_1X - b_2X^2 + b_3X^3}_{} \\ TC &= TFC + \qquad\quad TVC \end{aligned}$$

The AVC is

$$AVC = \frac{TVC}{X} = b_1 - b_2X + b_3X^2$$

The MC is

$$MC = \frac{\partial C}{\partial X} = b_1 - 2b_2X + 3b_3X^2$$

The ATC is

$$\frac{C}{X} = \frac{b_0}{X} + b_1 - b_2X + b_3X^2$$

The TC curve is roughly S-shaped (figure 4.3), while the ATC, the AVC and the MC are all U-shaped; the MC curve intersects the other two curves at their minimum points (figure 4.11).