11 Theory of Natural Monopoly

As we discussed in Chapter 10, there are a number of market-failure arguments for economic regulation. Perhaps the most important and widely accepted is natural monopoly, and it provides the rationale for regulating electric-power and natural-gas distribution, local telephone service, water supply, and some common-carrier transportation services. We begin this chapter with a discussion of the theory of natural monopoly. Actual regulation of natural monopoly will be the subject of the next two chapters.

We will be taking an economic efficiency view of natural monopoly here. In previous chapters we have discussed various explanations for the existence of regulation, including market failure and capture theory hypotheses. In this chapter we focus exclusively on the natural-monopoly market-failure argument and various theoretical and actual solutions.

This chapter is primarily theoretical, but it also serves as an introduction to the next few chapters. Chapter 12 will be concerned with the practice of natural monopoly regulation and an evaluation of its benefits and costs. Chapters 13–15 will discuss several alternatives to regulation that are introduced only briefly here.

The Natural Monopoly Problem

An industry is a natural monopoly if the production of a particular good or service by a single firm minimizes cost. The typical example is production of a single commodity where long-run average cost (LRAC) declines for all outputs. Such a case is illustrated in Figure 11.1. Because LRAC is declining, long-run marginal cost (LRMC) necessarily lies everywhere below it.

The case shown in Figure 11.1 makes clear the public-policy dilemma. Simply stated, the problem is how society can benefit from least-cost production—which obviously requires single-firm production—without suffering from monopoly pricing. The idea, of course, is that a single firm would eventually win the entire market by continuing to expand output and lowering its costs. Having won the market, it could then set the monopoly price.[1]

Shortly, we will turn to an analysis of the variety of solutions to this problem that have been proposed. Before we do so, however, we will examine more carefully the definition and characteristics of natural monopoly.

Permanent and Temporary Natural Monopoly

An important distinction is that of permanent versus temporary natural monopoly.[2] Figure 11.1 illustrates the case of permanent natural monopoly. The key is that LRAC falls

1. Entry, induced by the monopoly price, is usually assumed to be unlikely in natural monopoly situations.

2. The term permanent is perhaps misleading inasmuch as one can never rule out dramatic technological changes that could convert a natural monopoly into a competitively structured industry.

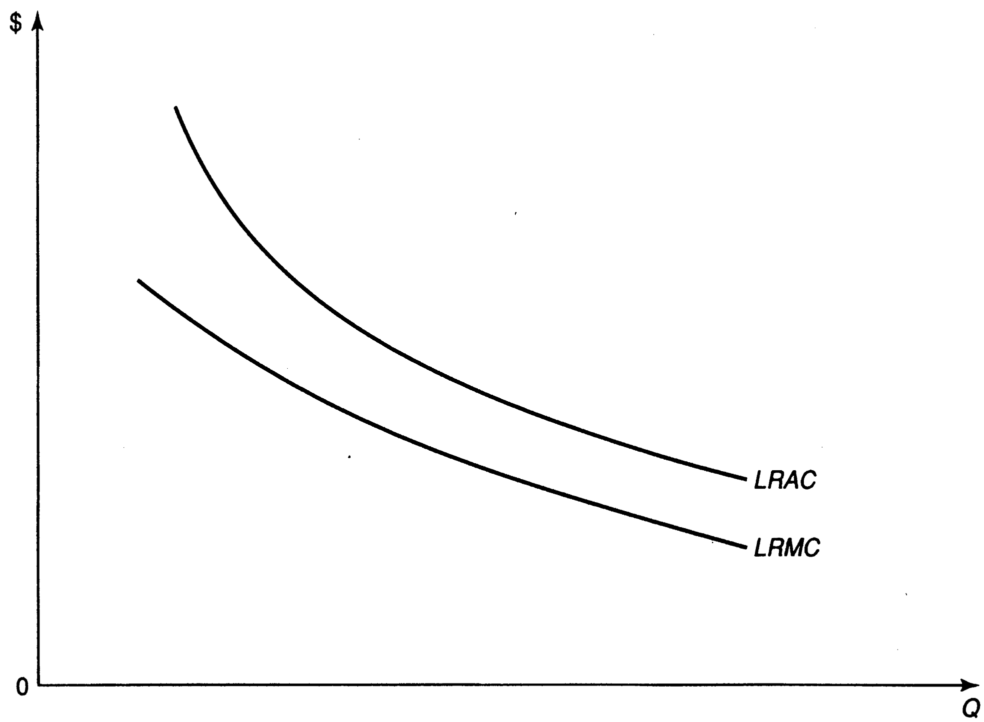

Figure 11.1
Cost Curves of Natural Monopolist

continuously as output increases. No matter how large market demand is, a single firm can produce it at least cost.

A temporary natural monopoly is shown in Figure 11.2. Observe that LRAC declines up to output Q^* and then becomes constant thereafter. Hence, as demand grows over time, a natural monopoly when demand DD prevails can become a workably competitive market when demand D_1D_1 holds.

One can argue that such a cost curve can be used to describe intercity telephone service. There are several factors that give rise to sharp unit-cost savings at low volumes of telephone calls, but they play out as volume increases.

For example, a microwave telephone system consists of a number of stations—about twenty to forty miles apart—that transmit signals of specific frequencies. Each station requires land, a building, a tower and antennas, electronic equipment, and so on. These inputs do not all increase proportionately with the number of circuits, and therefore as volume increases the fixed costs can be spread over more calls. This spreading effect becomes less and less significant, however, as volume grows.

As an example, long-distance telephone service between New York and Philadelphia required only 800 circuits in the 1940s. At this capacity, unit costs were falling and constituted a natural monopoly situation. In the late 1960s the number of circuits had risen to 79,000

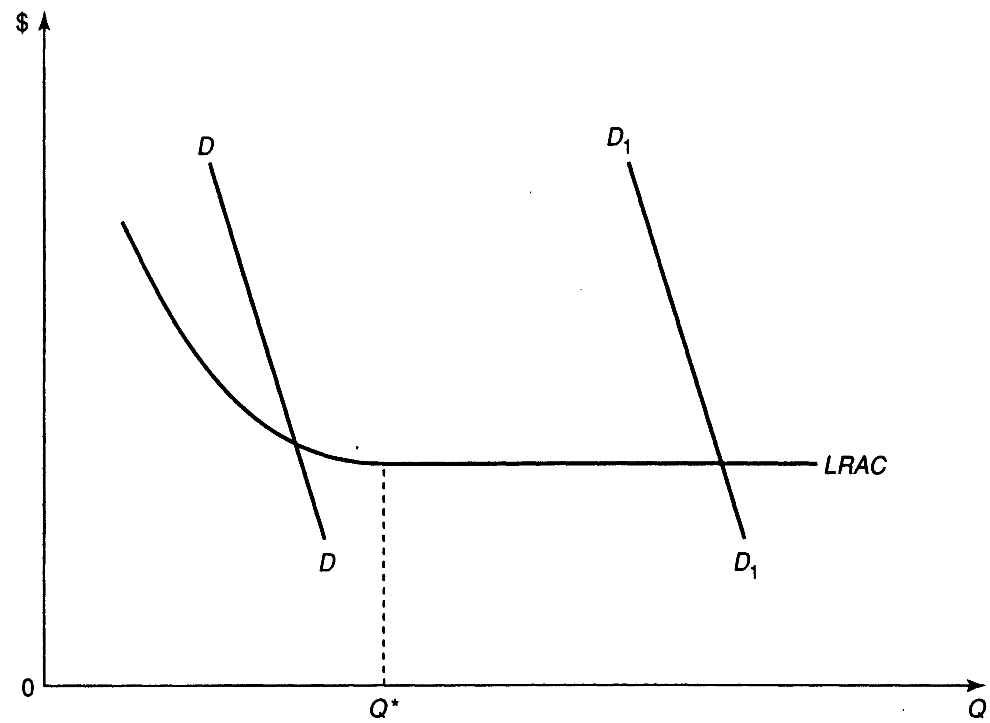

Figure 11.2
Temporary Natural Monopoly

(largely because of the requirements of television), and this volume was such that unit costs were essentially flat (beyond Q^* in Figure 11.2). Hence, by the late 1960s the temporary natural monopoly had disappeared.

This phenomenon is not rare. Railroads possessed significant cost advantages in the late 1800s, and these advantages were eroded considerably with the introduction of trucking in the 1920s. This example introduces a new element, namely, technological change.[3] That is, over long periods of time it is likely that the cost function will shift as new knowledge is incorporated into the production process. Hence, permanent natural monopoly is probably a rare category. Technical change can shift cost functions so as to render competition workable. And as we will see later, a serious deficiency of regulation seems to be that it often fails to "disappear" when the natural monopoly does.

Subadditivity and Multiproduct Monopoly

In the real world a single-commodity producer is rare. Electric utilities supply high and low voltage, peak and off-peak power; telephone companies provide local and long-distance

3. Strictly speaking, technical change in lowering costs was also present in the telephone service example.

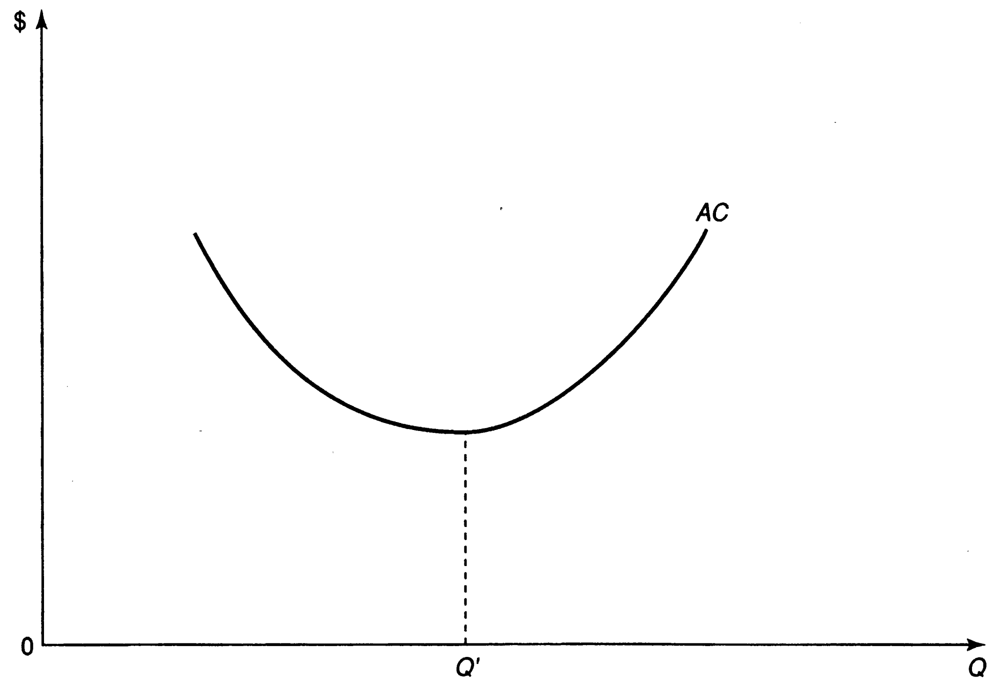

Figure 11.3
Economies of Scale up to Output Q'

service; and so on. It turns out that multiple-product natural monopoly is not only more realistic, but it also creates important theoretical issues that do not exist in the single-product case.

The definition of natural monopoly is that the cost function is subadditive.[4] We begin by explaining this concept in the single-product case because it can be illustrated graphically.

Consider the average cost curve shown in Figure 11.3. Average cost declines until the output Q' is reached, and then begins to increase. Economies of scale are said to exist at all outputs less than Q' and diseconomies at all outputs greater than Q'.

Subadditivity refers to whether it is cheaper to have one firm produce total industry output, or whether additional firms would yield lower total cost. For outputs less than Q', one firm is the least-cost solution, and therefore cost is subadditive for that range of outputs.

In order to examine the least-cost solution for outputs greater than Q', we introduce the minimum average cost function for two firms, AC_2. This curve and the single-firm AC curve from Figure 11.3 are both shown in Figure 11.4.

The curve AC_2 is obtained by construction from AC in the following manner. We know that for least-cost production, each firm must produce at the same output rate and thereby have the

4. An important article that defines natural monopoly this way is W. J. Baumol, "On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry," *American Economic Review*, December 1977.

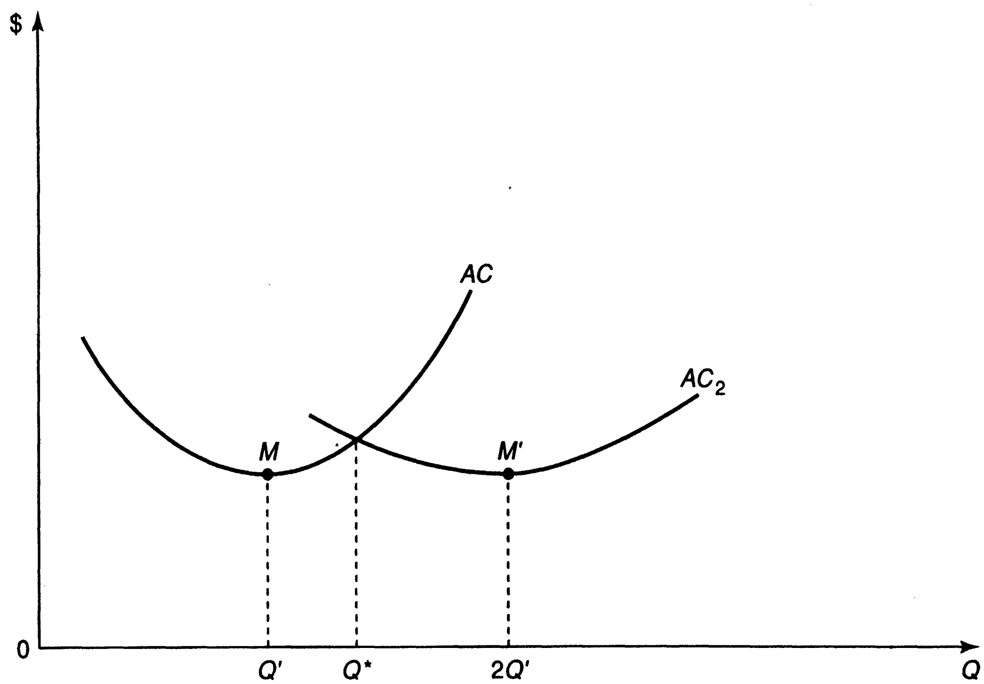

Figure 11.4
Minimum Average Cost Curve for Two Firms, AC_2

same marginal cost. Hence, for a given point on the AC curve, simply double the output rate to obtain a point on the AC_2 curve. For example, at the minimum average cost point M on AC, double Q' to get $2Q'$, which corresponds to the minimum point M' on AC_2.

The intersection of AC and AC_2 at output Q^* defines the range of subadditivity. For all outputs less than Q^*, a single firm yields least-cost production. Hence the cost function is subadditive for outputs less than Q^*. Notice that subadditivity is the best way to define natural monopoly. Even though diseconomies of scale obtain between Q' and Q^*, it would be in society's interest to have a single firm produce in that range. An important point is that economies of scale (declining average cost) are not necessary for a single-product natural monopoly (although they are sufficient).

When we turn to multiple-product natural monopoly, the distinction between subadditivity and economies of scale becomes even greater. Again, the proper definition of natural monopoly is that the cost function is subadditive. That is, whatever the combination of outputs desired (say, 85 cars and 63 trucks, or 25 cars and 78 trucks), it is cheaper for a single firm to produce that combination if the cost function is subadditive.

In the multiple-output case, it can be shown that economies of scale are neither necessary nor sufficient for costs to be subadditive! Economies of scale would hold, for example, if the total cost of producing, say, a 10-percent greater quantity of each commodity increased by some amount less than 10 percent. The reason that economies of scale are neither necessary nor sufficient for subadditivity is that in the production of multiple outputs, the interdependence among outputs also becomes important.

Although various ways have been proposed for measuring these interdependencies, the concept of economies and diseconomies of scope is appealing intuitively.[5] Economies of scope mean that it is cheaper to produce, say, 85 cars and 63 trucks within a single firm than it is for specialty firms to produce the required outputs. If you think of peak-period electric power and off-peak power as different commodities, then economies of scope are clearly present—the two commodities can share the same power plant and distribution system.

Sharkey has given an example of a cost function that possesses economies of scale for all outputs, but which is nowhere subadditive.[6] His example is

$$C(Q_1, Q_2) = Q_1 + Q_2 + (Q_1 Q_2)^{1/3}. \tag{11.1}$$

Notice that the total cost after increasing each output by 10 percent is

$$C(1.1Q_1, 1.1Q_2) = 1.1Q_1 + 1.1Q_2 + 1.1^{2/3}(Q_1 Q_2)^{1/3}$$

whereas the total cost increased by 10 percent is

$$1.1C(Q_1, Q_2) = 1.1Q_1 + 1.1Q_2 + 1.1(Q_1 Q_2)^{1/3}.$$

Because the former is less than the latter, economies of scale exist. Nevertheless, the function has diseconomies of scope that sufficiently outweigh the economies of scale to make cost nowhere subadditive.

To see this point, note that the third term in the cost function, equation (11.1), adds a positive amount to cost whenever both outputs are produced together. If, for example, all Q_1 was produced by firm A and all Q_2 was produced by firm B, then the sum of the total costs of the two firms would be less than if all production was carried out in a single firm, C. Specifically:

$$C_A = Q_1, C_B = Q_2, \quad \text{so} \quad C_A + C_B = Q_1 + Q_2$$

$$C_C = Q_1 + Q_2 + (Q_1 Q_2)^{1/3}.$$

Because $C_A + C_B < C_C$, production in the specialty firms, A and B, is cheaper than in a single firm, C. Thus, economies of scale are not sufficient for cost to be subadditive because of the diseconomies of scope.

In summary, the definition of natural monopoly in the multiple-output case is that the cost function must be subadditive. Subadditivity of the cost function simply means that the production of all combinations of outputs is accomplished at least cost by a single firm. It is a complex matter to specify the necessary and sufficient conditions for costs to be subadditive.

5. See, for example, J. C. Panzar and R. D. Willig, "Economies of Scope," *American Economic Review*, May 1981.
6. William W. Sharkey, *The Theory of Natural Monopoly* (New York: Cambridge University Press, 1982).

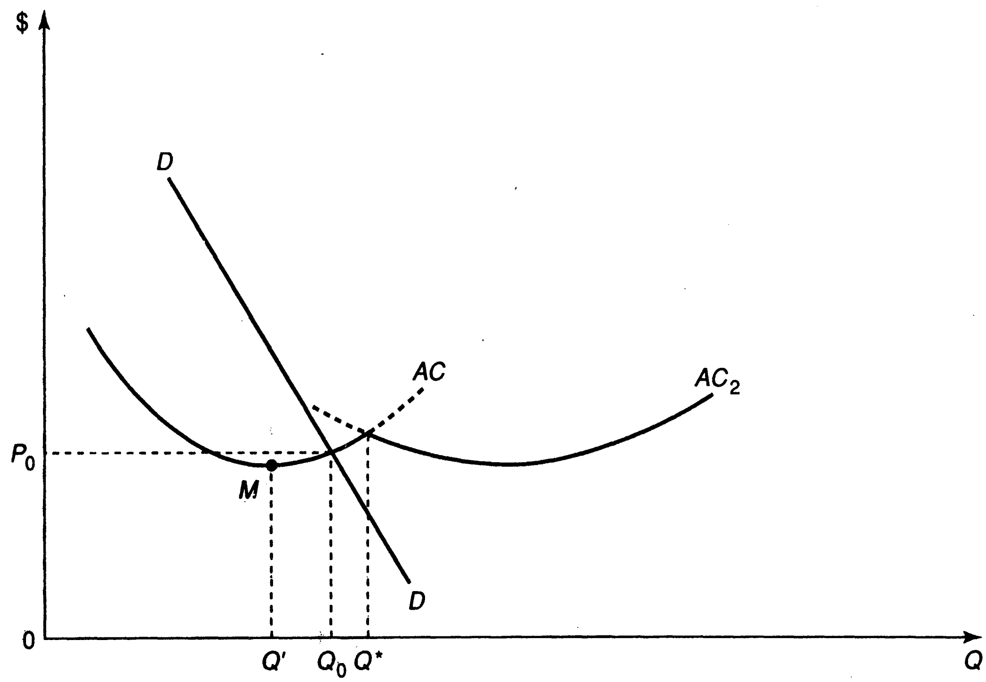

Figure 11.5
Sustainable Natural Monopoly up to Output Q'

We have shown through some simple examples, however, that it generally depends on both economies of scale and economies of scope. If both exist, then subadditivity will likely obtain.[7] Economies of scale alone, however, can be outweighed by diseconomies of scope. Thus, although economies of scale in the single-product case imply natural monopoly, this statement does not hold true for the multiple-product case.

Before turning to the various policy solutions to the natural monopoly problem, we shall briefly explain a related concept known as sustainability. It can be explained best by reference to Figure 11.5.

Figure 11.5 reproduces the cost function for the single-product case from Figure 11.4. Recall that the cost function is subadditive for outputs less than Q^*. Now consider a case in which market demand DD intersects average cost somewhere between Q' and Q^*, where AC is rising. If a single firm were to supply all output demanded at a price equal to average cost (at price P_0 and output Q_0 so that the firm would just cover all its costs), the natural monopoly would be termed unsustainable. That is, under certain assumptions, a potential entrant would have an incentive to enter the market and produce a share of total output even though doing so would increase the cost of producing the total industry output.

7. For a rigorous analysis, see W. J. Baumol, J. C. Panzar, and R. D. Willig, *Contestable Markets and the Theory of Industry Structure* (New York: Harcourt Brace Jovanovich, 1982).

The assumptions referred to in the preceding sentence are that the entrant expects the incumbent firm to keep its price unchanged for some period of time after entry, and that the incumbent will supply the residual output.[8] Under these assumptions, the entrant would perceive that it could profit by offering to sell output Q' in Figure 11.4 at some price above its minimum average cost (point M) but slightly less than the price P_0 being charged by the incumbent.

By contrast, a sustainable natural monopoly would be one where market demand intersects AC in Figure 11.5 to the left of Q'. In this case an entrant cannot undercut the incumbent and therefore has no incentive to enter. The concept of sustainability is relevant where a regulatory agency must decide whether to allow entry in a particular market of a multiple-product natural monopolist.

Alternative Policy Solutions

In this section we examine various alternatives that have been proposed (and, in some cases, implemented) to correct the natural monopoly inefficiency. These alternatives include "doing nothing"; various "ideal" solutions; competition among bidders for the right to the monopoly franchise; and, finally, actual regulation, as practiced in the United States, and public enterprise, as exemplified by the Postal Service.

The first alternative mentioned—doing nothing—might be appropriate if the potential monopoly power is not great. For example, a cable-television system might be viewed as a natural monopoly, but one with quite limited capacity for earning excess returns, for substitutes for cable television are rather close. Over-the-air broadcasting is one of them. Others are apparently becoming more important over time as new technologies are perfected.

We consider first a collection of "ideal" pricing solutions. The adjective "ideal" is employed to indicate that we are assuming that the firm is to be operated in the public interest and that the only issue is what prices produce economic efficiency.

Ideal Pricing

The most obvious candidate for the efficient price is, of course, marginal cost.[9] A natural monopolist that charges marginal cost for each product is said to practice linear (or uniform)

8. A further assumption is that the entrant perceives no entry barriers in the form of "sunk" costs. That is, the entrant believes that whatever investment is required can be recovered by transferring it elsewhere or by sale. All of these assumptions have been subject to controversy since the sustainability literature was introduced by Baumol, Panzar, and Willig.

9. See Chapter 4 for a detailed rationale. For a rigorous treatment of efficient pricing, see R. R. Braeutigam, "Optimal Policies for Natural Monopolies," in R. Schmalensee and R. D. Willig (eds.), *Handbook of Industrial Organization*, Vol. 2 (Amsterdam: North-Holland, 1989), and D. F. Spulber, *Regulation and Markets* (Cambridge, Mass.: MIT Press, 1989). For a more geometrical treatment, see K. E. Train, *Optimal Regulation* (Cambridge, Mass.: MIT Press, 1991).

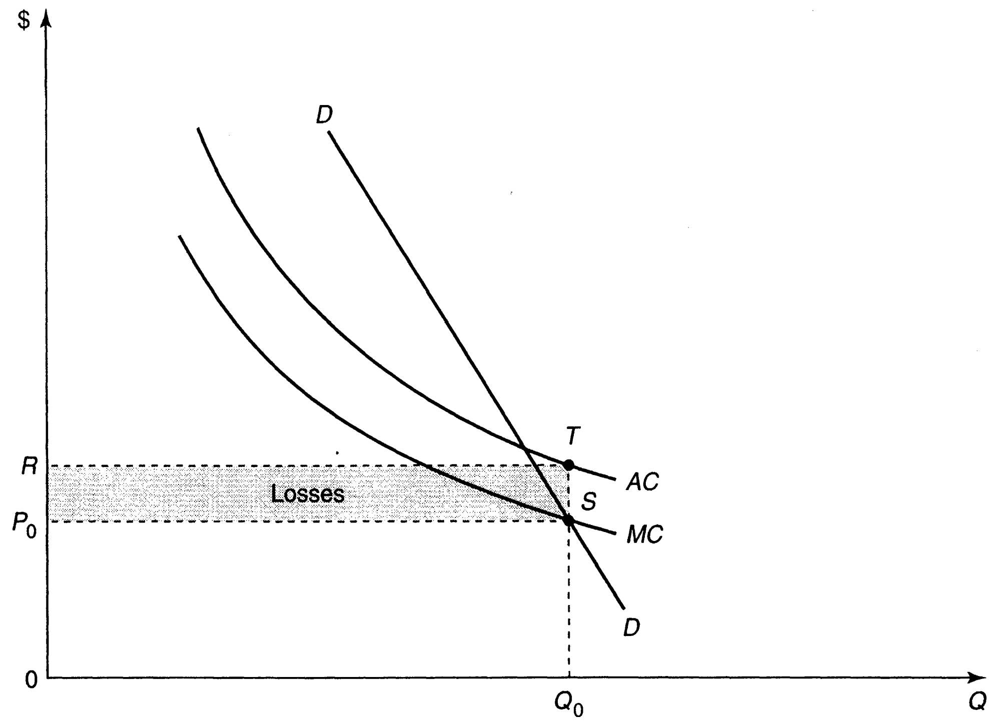

Figure 11.6
Marginal Cost Pricing Can Cause Losses

marginal cost pricing. In other words, a customer's expenditure for a product is a linear function of price and quantity sold, PQ. On the other hand, if the firm charges a fixed fee F, regardless of the amount bought, and also a per-unit charge P, nonlinear (or nonuniform) pricing would be in effect. Then the customer's expenditure would be a nonlinear function, $F + PQ$.

In our ideal pricing discussion, we begin with the linear marginal cost pricing solution. After considering nonlinear pricing we examine the so-called Ramsey pricing alternative, which applies to multiproduct cases. The section concludes with a discussion of a theoretical proposal by Loeb and Magat to induce profit-maximizing firms to price efficiently.

Linear Marginal Cost Pricing

Consider a single-product natural monopolist with decreasing average costs over the relevant output range. Figure 11.6 shows such a situation where market demand is DD.

The marginal cost price would be P_0 with output Q_0. The price does meet the well-known requirement for efficiency; however, on closer examination, several serious difficulties arise. An obvious difficulty is the loss, shown by the shaded rectangle RP_0ST.[10] Any enterprise

10. The loss is equal to the difference between price and average cost, multiplied by output.

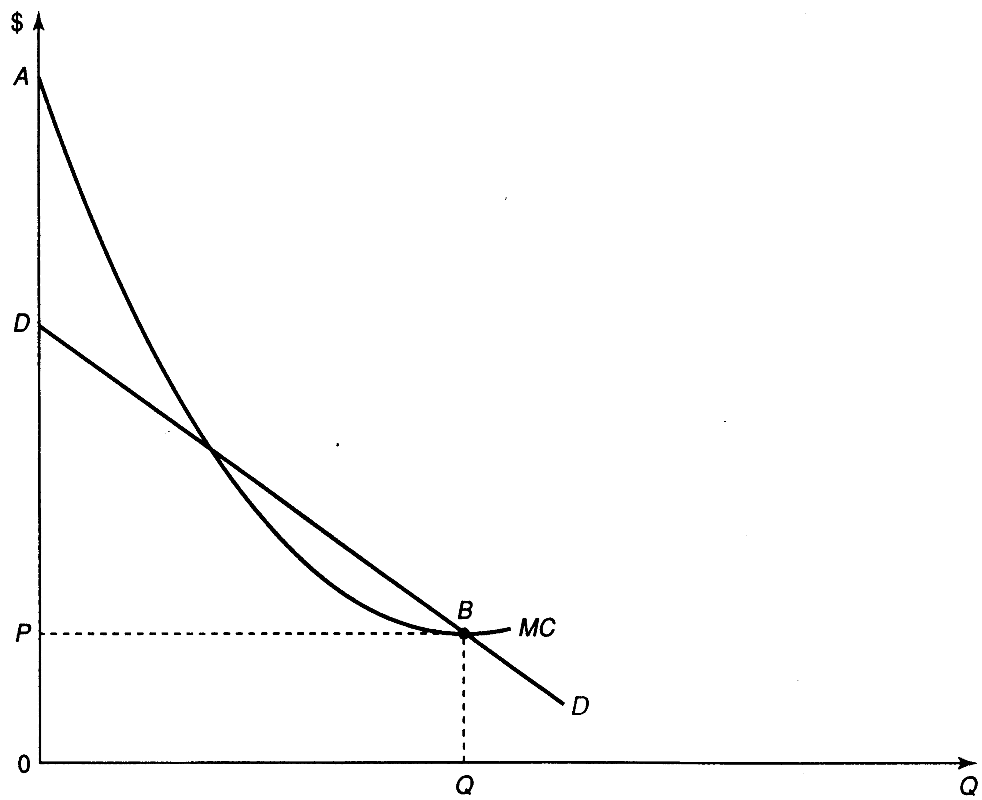

Figure 11.7
Natural Monopoly with Costs Exceeding Benefits

would need a subsidy to continue to operate at this output level, because price is less than average cost. The next question is to ask where the subsidy is to come from and what effect this will have on economic efficiency.

The only "correct" solution is for the government to raise the subsidy through a lump-sum tax, that is, a tax that would not distort other decisions throughout the economy. Such taxes are rarely, if ever, used in practice. Income taxes and sales taxes are unacceptable because they create inefficiencies themselves by introducing wedges between prices and marginal costs. Even this "correct" solution (lump-sum tax to pay subsidy) is subject to some rather persuasive opposing arguments. Three frequently mentioned arguments are as follows:

1. If total costs are not covered by consumer expenditures, it is possible that total consumer benefits (given by the area under the demand curve)[11] are less than total costs—which means the good should not be produced at all. Figure 11.7 provides such a case. Total costs $AOQB$

11. Throughout this chapter we make the common assumption that the area under the demand curve measures total willingness to pay by consumers. This requires one to assume that the income elasticity of demand is zero (or small enough to make the error unimportant). See R. D. Willig, "Consumer's Surplus without Apology," *American Economic Review*, September 1976.

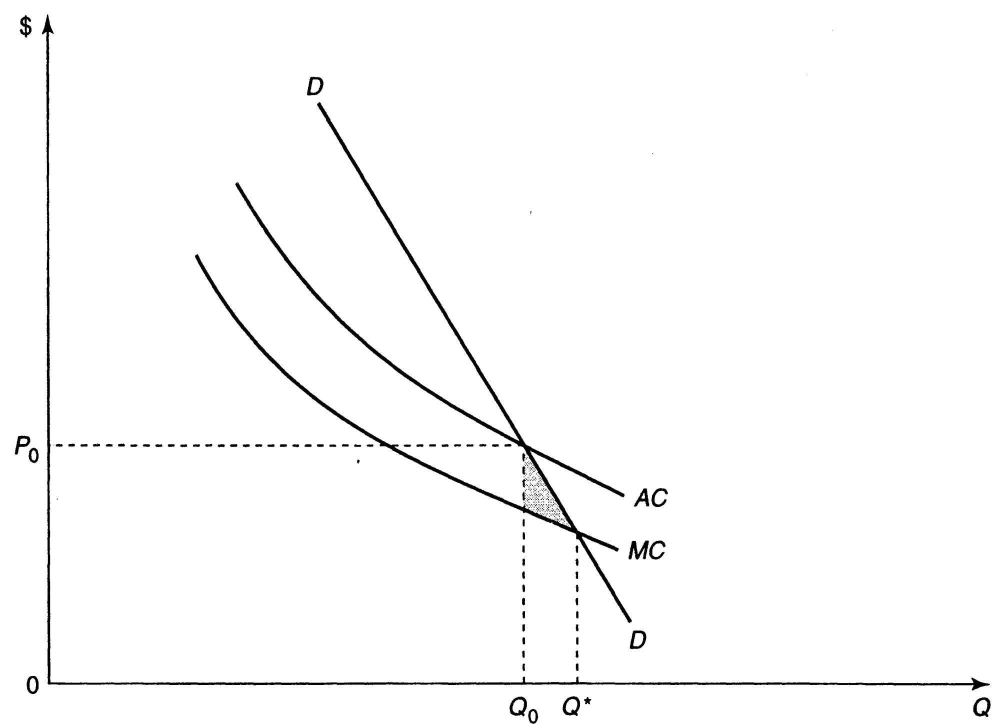

Figure 11.8
Welfare Loss with Average Cost Pricing

(the area under the MC curve) exceed total benefits $DOQB$. Only if consumers are required to actually cover total costs can we be sure that the good is socially beneficial.

2. Because the enterprise's management knows losses will be subsidized, the incentive and capacity to control costs is weakened. Postal Service employees, for example, have an advantage in bargaining with management, inasmuch as both sides know that the enterprise will not fail if revenues are less than costs. The Treasury can always be counted on to subsidize the Postal Service in a pinch. Steel industry labor unions do not have this advantage.

3. On distributional grounds, it can be argued that nonbuyers of the natural monopoly good should not be required to subsidize the marginal cost buyers. That is, why should the taxes paid by individuals without telephone service be used to subsidize individuals who purchase such service at a loss-creating price?

A major point of the preceding analysis is that enterprises should price so that their revenues cover costs. Furthermore, in the United States, because most public utilities are privately owned firms, it is politically unrealistic to imagine government subsidizing the losses of private firms. Hence we conclude that there are compelling reasons to accept the constraint that natural monopolies should operate such that total revenues and total costs are equated.

In the single-product case, linear pricing implies that price must equal average cost if total revenues must equal total costs. This relationship is shown in Figure 11.8 as price P_0 and output Q_0. This departure from marginal cost pricing leads, of course, to the welfare loss given by the shaded triangular area.[12]

This argument refers to linear pricing; that is, the buyer pays a single price per unit, and therefore the buyer's total expenditure is proportional to total consumption. An important alternative is nonlinear pricing.

Nonlinear Pricing

A two-part tariff is nonlinear and consists of a fixed amount or fee, regardless of consumption, plus a price per unit. If the price per unit equals marginal cost, then it is possible to have efficient pricing and have total revenues of the firm equal to its total costs.

For example, if the loss under linear marginal cost pricing is estimated to be K (the shaded rectangle in Figure 11.6), the fixed fee of the two-part tariff could be set so that the sum over all customers equals K. There are various ways for this equality to hold—the simplest is to set the fixed fee equal to K/N, where N equals the number of consumers.

There are possible problems with this nondiscriminatory two-part tariff. Because consumers usually vary considerably in terms of their demands for the good, it is possible for some consumers to be driven from the market if K/N exceeds their consumer surpluses at price equal to marginal cost. One might expect this outcome to be more likely for, say, telephone service than for such "necessities" as electricity and water. Hence, efficiency losses will occur if these excluded consumers would have been willing to pay marginal cost. It is also true that in some markets it is not feasible to enforce a fixed fee for the "right-to-buy" at a price per unit. Consumers would have an incentive to have one person purchase for all, thereby paying only one fixed fee. This is not a problem for most public utilities.

The obvious thing to do to avoid excluding consumers is to charge different fixed fees to different consumers, or classes of consumers. In short, discriminatory two-part tariffs could tailor the fixed fees to the consumers' willingnesses to pay where the sum of the fixed fees should add up to K. Although this solution is best in terms of efficiency, it may be illegal to so discriminate.

If all consumers must be charged the same fixed fee, it will still be more efficient to use a two-part tariff than to use linear pricing (which in the case of a single product implies average cost pricing). The reason is simply that by using a fixed fee to make a contribution to revenues, the price per unit can be lowered toward marginal cost—thereby reducing deadweight losses. (In principle, one can pick some fixed fee, no matter how small, that will not drive anyone from the market and will permit a lowering of the price.)

The next logical question is, What is the optimal two-part tariff? Here, we explain only the economic principle involved.[13] Suppose initially that the fee is zero and price equals marginal

12. For a discussion of welfare loss determination, see Chapter 4.

13. See Stephen J. Brown and David S. Sibley, *The Theory of Public Utility Pricing* (New York: Cambridge University Press, 1986), p. 93, for a formal analysis.

cost. The result is, of course, a deficit that must be covered by increasing either the fee or the price per unit, or both. In essence, the derivation depends on a balancing of efficiency losses because of exclusion of additional consumers as the fixed fee rises against the increased consumption losses as price per unit increases above marginal cost. Hence the optimal two-part tariff generally will involve a price per unit that exceeds marginal cost and a fixed fee that excludes some consumers from the market.

Multipart tariffs are often used by public utilities. Consider the following example of the type of tariff sometimes used for local telephone service (such tariffs are often referred to as declining-block tariffs).

Fixed fee per month—$5

+10 cents per call for up to 100 calls

+5 cents per call for all calls between 100 and 200

+0 cents per call for all calls above 200

Notice that the marginal price falls as one moves to successively larger calling "blocks"—from 10 cents to 5 cents to 0 cents. This multipart tariff is plotted in Figure 11.9 as the bold segmented line $ABCD$. (The reason for the extensions of these segments in Figure 11.9 will become clear shortly.) Hence the figure shows "total consumer expenditure" vertically as a function of total "calls per month" horizontally.

A rationale often given for the declining blocks is that utilities are characterized by economies of scale, and falling marginal prices stimulate consumption—in turn permitting the construction of larger, lower-unit-cost plants. An alternative rationale is to view the declining-block tariff as a *self-selecting* set of two-part tariffs, and a set of such tariffs can increase economic efficiency along the lines discussed earlier.

Recall that discriminatory two-part tariffs permit the firm to tailor the tariffs to fit the differences in willingnesses to pay across consumers. The efficient solution can be achieved if no consumers are excluded from the market and all pay marginal cost per unit. As an approximation to this "ideal," one can use the multipart tariff in Figure 11.9 to cause consumers to self-select a two-part tariff that they prefer—wherein consumers with high willingnesses to pay pay high fixed fees in return for low prices per unit.

The three "self-selecting" two-part tariffs are

Fixed Fee	Price/Unit
$5	10 cents
$10	5 cents
$20	0 cents

One can represent a two-part tariff by a vertical intercept (for the fixed fee) and a straight line with slope equal to the price per unit. The three such lines in Figure 11.9 represent the three two-part tariffs that we have referred to. (Notice that no consumer would wish to

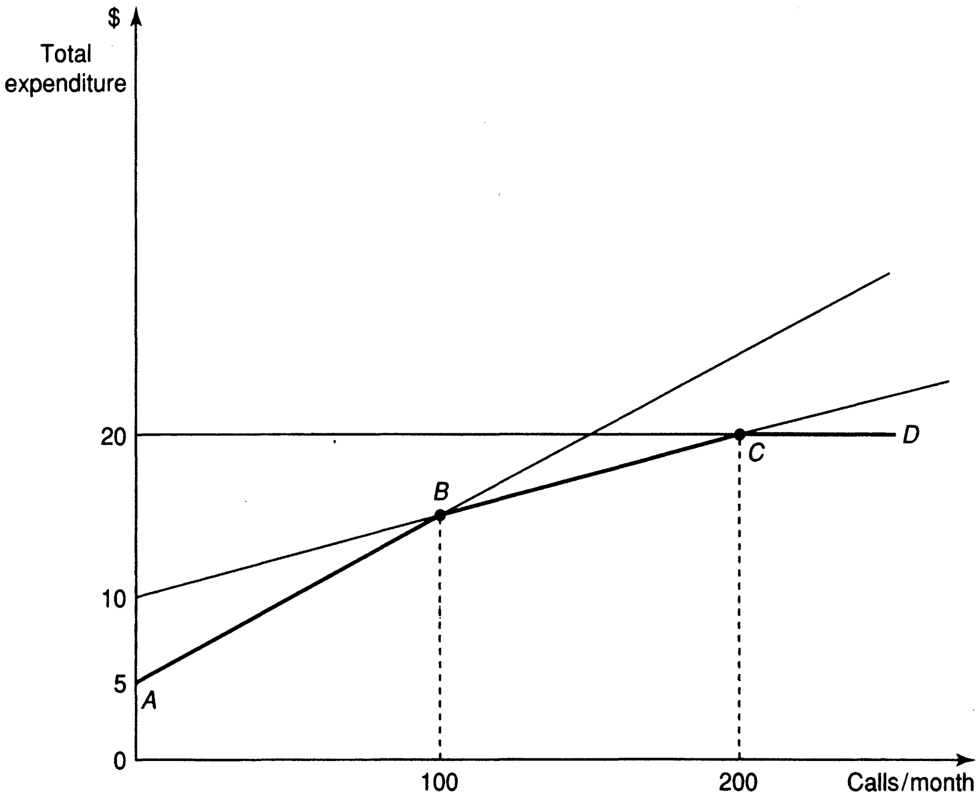

Figure 11.9
Multipart Tariff for Local Telephone Service

consume on portions of the tariffs other than the lower boundary *ABCD*. Hence it does not matter that these "dominated" portions of the two-part tariffs are not actually part of the declining-block tariff.) The point is that the declining-block tariff has the same effect as confronting consumers with two-part tariffs that are tailored to their demands. And, of course, all consumers are free to choose the particular tariff that they prefer, so that there is no discrimination involved that is likely to be disallowed.

Up to this point our discussion of ideal pricing has been limited to a single-product natural monopolist. We now turn to the case of a multiple-product natural monopolist and describe what has become known as Ramsey pricing.

Ramsey Pricing

In a famous article published in 1927, Frank Ramsey suggested the following pricing (and taxing) method.[14] It is applicable to a multiple-product natural monopolist that would generate losses if linear marginal cost pricing were used. In essence, Ramsey prices are those linear

14. Frank Ramsey, "A Contribution to the Theory of Taxation," *Economic Journal*, March 1927.

prices that satisfy the total-revenues equal-total cost constraint and minimize the deadweight welfare losses. Note that Ramsey prices are linear prices—one for each product—so that we are implicitly ruling out multipart tariffs.

It is useful to illustrate Ramsey pricing with a numerical example. Let the natural monopoly be a two-product firm with total cost

$$C = 1800 + 20X + 20Y.$$

The market demands for the two goods X and Y are given by

$$X = 100 - P_x$$

$$Y = 120 - 2P_y.$$

An important assumption that we will make for our example is that the demands are independent—the demand for X does not depend on the price of Y, and vice versa. The more general case of interdependent demands involves much more complex mathematics and is beyond the scope of the discussion here.[15]

It should be obvious that the marginal costs of X and Y are each \$20, and that marginal cost prices would exactly cover the variable costs but not the fixed cost of \$1,800. Because the firm must cover its total costs, it is clear that the prices will necessarily exceed their respective marginal costs. One possibility would be to raise the prices by the same proportion above marginal costs until total costs are covered. This is shown in Figure 11.10a.

The figure shows that prices would need to be raised from \$20 to \$36.1 to generate sufficient revenues just to cover total costs.[16] In particular, the contribution that product Y makes toward fixed cost equals the rectangle $CEFD$. This is just price minus the constant unit variable cost of \$20, multiplied by the output of 47.7. Similarly, the contribution that product X makes equals rectangle $CEKJ$. The sum of these two rectangles is \$1,800. (The fact that the demands intersect at the price equals marginal cost point for each is not necessary, and was chosen merely to make the graphical exposition simpler.)

Now consider the deadweight losses that this proportionate price increase method causes. The deadweight loss triangle for product Y is triangle DFH, and it is JKH for product X. The actual numerical values are \$260 and \$130, respectively, or a total of \$390. Hence, one way of summing up this method is to observe that it "costs" \$390 in deadweight welfare losses to generate the \$1,800 necessary for the firm to break even. The question becomes whether one can find another method for raising prices to generate the \$1,800 that entails a lower welfare cost.

15. The interested reader should consult Brown and Sibley, *The Theory of Public Utility Pricing*, p. 42.

16. Because P_x and P_y must be equal under the assumption that the marginal costs are both \$20, the \$36.1 value can be found by solving the equation that equates total revenues and total costs.

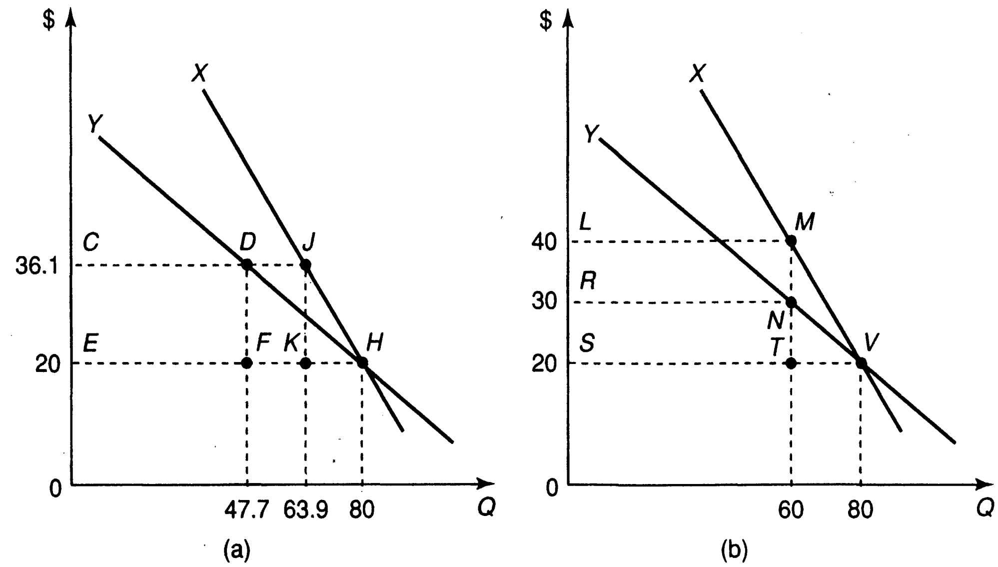

Figure 11.10
Proportionate Price Increase versus Ramsey Pricing

A bit of reflection while examining Figure 11.10a might suggest differential price increases. That is, it is clear that the same price increase produces a smaller contribution to fixed cost from product Y at a higher cost in terms of deadweight loss. This observation is not surprising when one realizes that product X has a more inelastic demand (at point H) than does product Y. This difference suggests that it would be better to raise the price of X more than the price of Y.

The Ramsey pricing "rule" that gives the prices that minimize the deadweight losses is to raise prices in inverse proportion to demand elasticities. Mathematically, the rule[17] is

$$\frac{P_i - MC_i}{P_i} = \frac{\lambda}{\eta_i}$$

where P_i is the price of good i, MC_i is the marginal cost of i, η_i is the absolute value of the elasticity of demand of good i, and λ is a constant. Using this rule, one can derive the actual Ramsey prices.[18] They are shown in Figure 11.10b. Hence the firm would minimize the welfare losses by charging \$40 for good X and \$30 for good Y. At these prices, the

17. See Brown and Sibley, 1986, p. 39, for a formal derivation.

18. Computations are made simpler by using the alternative rule for Ramsey prices that will be given shortly involving proportionate quantity changes. That rule implies that the two products will have equal outputs. Hence this fact together with the total-revenues-equal-total-costs equation yields the Ramsey prices.

demand elasticities are 0.67 and 1.0, respectively. The deadweight loss triangles are $200 for good X (triangle MTV) and $100 for good Y (triangle NTV) for a total of $300. This is, of course, a lower "cost" in terms of welfare by $97 than the proportionate method of Figure 11.10a.

Another interesting fact about Ramsey prices is apparent in Figure 11.10b. The proportionate decrease in output from the price-equals-marginal-cost output (outputs of 80 for both) is the same for the two goods. That is, both outputs are cut by (80–60)/80, or 25 percent. This is an alternative way of describing Ramsey pricing: cut output of all goods by the same proportion until total revenue just equals total cost. This way of stating the rule for Ramsey pricing is more general than the inverse elasticity rule, and holds true for the case of interdependent demands.

The Ramsey pricing rule can be viewed as providing theoretical justification for so-called *value of service* pricing that has been used for years in the railroad industry. It has been common for rail rates for shipping gravel, sand, potatoes, oranges, and grapefruits to be lower relative to shipping costs than for liquor, electronic equipment, cigarettes, and the like. The reason is that the elasticities of demand for shipping products that have low values per pound are higher than for products that have relatively high values per pound. (We are assuming that the actual costs of shipping are proportional to weight.)

In summary, all of the ideal pricing schemes discussed have problems (except for the two-part tariff with price equal to marginal cost and no exclusion of consumers by the fixed fee). It should be kept in mind that we have assumed away the very real difficulty of designing incentive systems that will induce enterprise managers to implement these pricing schemes. In short, managers of private firms are presumably interested in maximizing profits, not total economic surplus. Managers of public enterprises may also have objectives other than economic efficiency. Economists have recently begun to explore theoretical models of how regulatory agencies might provide incentives for natural monopolies to price efficiently. We will briefly describe the Loeb-Magat proposal in the next subsection.

Loeb-Magat Proposal

Of course, if regulators had perfect information as to the monopolist's costs and demands, the ideal pricing schemes that we have discussed could be put into effect by command. However, such is not the case. Although the monopolist may not have perfect information itself, most people would probably agree that the monopolist has much better knowledge of its costs than the regulators do. Because the firm's profits will increase with higher prices, the firm has an incentive to overstate its costs (which is the usual basis that a regulator uses to set prices).

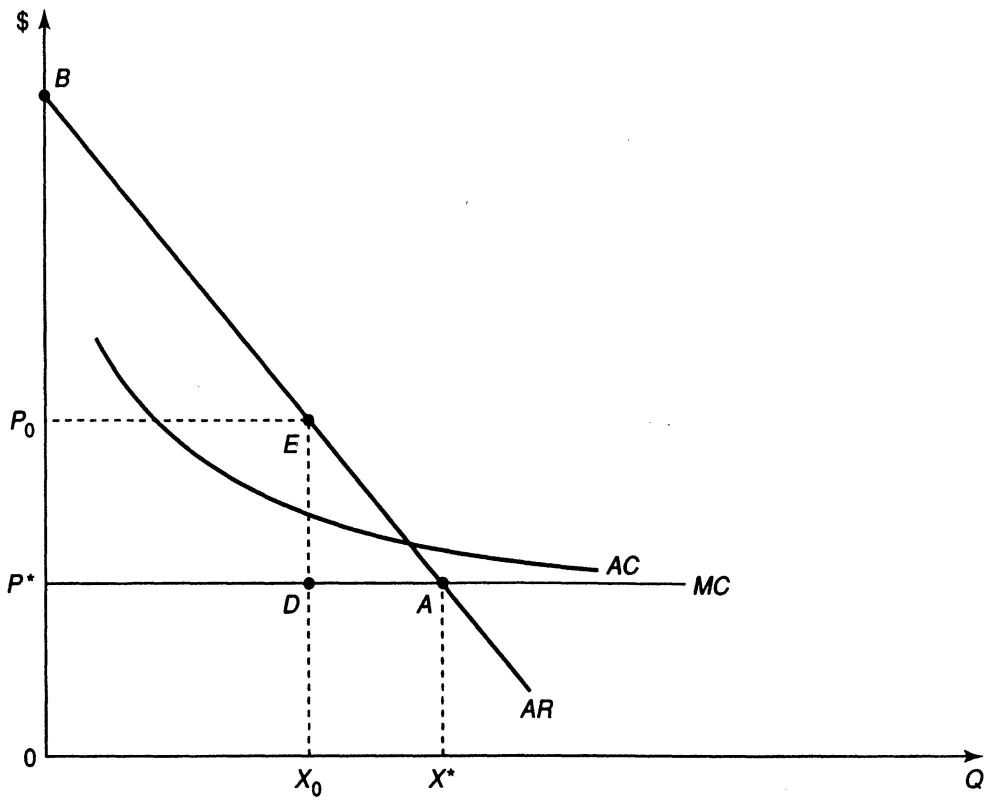

Figure 11.11
Loeb and Magat Incentive Scheme

Loeb and Magat (L-M) assumed that the monopolist knows costs and demand information perfectly, but that the regulator knows demand only.[19] Hence, given this asymmetry of information and the assumption that the monopolist's objective is to maximize profit, what might the agency do to induce efficient pricing? The L-M scheme can be explained easily with the aid of Figure 11.11, which shows a single-product natural monopolist.

The monopolist has declining average cost (AC) and demand curve (AR). For simplicity, we assume the total cost function is $K + vX$; hence, marginal cost (MC) is constant and equal to v. The L-M proposal is to allow the monopolist to choose its own price—this differs

19. Martin Loeb and Wesley Magat, "A Decentralized Method for Utility Regulation," *Journal of Law and Economics*, 1969. Some additional research on this same issue can be found in Ingo Vogelsang and Jorg Finsinger, "A Regulatory Adjustment Process for Optimal Pricing by Multiproduct Monopoly Firms," *Bell Journal of Economics*, 1979; D. P. Baron and R. B. Myerson, "Regulating a Monopolist with Unknown Costs," *Econometrica*, 1982; D. Sappington, "Optimal Regulation of a Multiproduct Monopoly with Unknown Technological Capabilities," *Bell Journal of Economics*, 1983. A comprehensive though difficult recent survey is D. P. Baron, "Design of Regulatory Mechanisms and Institutions," in R. Schmalensee and R. D. Willig (eds.), *Handbook of Industrial Organization*.

from the usual practice of the regulatory agency setting the price. However, they propose to have the agency subsidize the firm by an amount equal to consumer surplus at the selected price.

Suppose that the monopolist selects the price P_0. Its profits will be $P^*DEB - K$. The firm collects $0X_0EP_0$ from customers and P_0EB from the regulatory agency. Its variable cost is $0X_0DP^*$, leaving a variable profit of P^*DEB. Subtracting the fixed cost of K leaves the profit just asserted. Observe, however, that the firm can do better by lowering price. For example, if the monopolist selected P^*, it is easy to show that its profits will increase to $P^*AB - K$. That is, profits increase by the usual deadweight loss triangle DAE. This is, in fact, the profit-maximizing solution for the monopolist! Convince yourself that any other price will reduce profits. (Alternatively, note that the proposal causes the demand curve AR to become the monopolist's marginal revenue curve, and setting MC equal to marginal revenue is the profit-maximizing solution.)

The explanation for this price-equal-to-marginal-cost result is simply that the regulator has changed the firm's objective function by the subsidy. Now, in effect, the monopolist is maximizing total surplus—the total area under the demand curve minus costs.

The solution is economically efficient, but most people would find it objectionable on distributional grounds. The monopolist is appropriating the total economic surplus! To rectify this problem L-M suggest that a franchise bidding scheme (or a tax scheme) could recover some of the subsidy for the general treasury. In the case shown in Figure 11.11, the regulatory agency would auction off the right to operate the monopoly franchise. The key idea is that above-normal returns (of amount $P^*AB - K$) are available to the firm that operates the monopoly and that bidding for the franchise would continue until that amount is bid. Note that the subsidy is not completely recovered—there remains a net subsidy of an amount equal to fixed cost, K.[20]

Obviously, the L-M proposal is not the perfect solution to natural monopoly. Informational problems about the demand curve and the existence of a subsidy make it an unlikely substitute for the present regulatory process. It has, however, stimulated research by economists toward the goal of understanding how the regulatory process might be improved with respect to providing better incentive structures for natural monopolists.

In the next section we return to the discussion of alternative policy solutions to the natural monopoly problem. In contrast to the ideal pricing solutions that we have been examining heretofore, we now turn to actual solutions that have been used. The first is franchise bidding.

---

20. For a variation on the Loeb and Magat proposal that eliminates the net subsidy and the need of the regulator to know demand, see D. A. Graham and J. M. Vernon, "A Note on Decentralized Natural Monopoly Regulation," *Southern Economic Journal*, July 1991.

Franchise Bidding

Harold Demsetz has argued that the "theory of natural monopoly is deficient for it fails to reveal the logical steps that carry it from scale economies in production to monopoly price in the market place."[21] His point is that it may be possible to have bidding for the right to supply the entire demand (in effect, bidding for a franchise to serve a certain market). Even though only the single firm submitting the low bid would actually produce, there could be competition among potential suppliers. For example, given the situation shown in Figure 11.8, the low bid presumably would be a price of P_0 for Q_0 units.

Note that P_0 is not the efficient price. Nevertheless, P_0 would be an improvement over the natural monopoly price (a price above P_0). Then P_0 would be the lowest price bid for the right to supply the market inasmuch as any lower price would result in losses. At P_0 the winning bidder would just cover costs, including a normal return on investment.

This bidding for the franchise argument has stimulated a great deal of useful thinking about alternatives to natural monopoly regulation. However, the highly abstract example here oversimplifies many of the problems that such bidding would raise. A detailed discussion will be provided in Chapter 13.

Actual Solutions

In this section we briefly consider actual solutions that have been implemented in response to the natural monopoly problem. There are basically two distinct solutions: the regulatory agency and public enterprise. Extensive discussions of each will be presented in subsequent chapters; only a short treatment is given here.

Regulation

The typical natural monopoly in the United States is a private firm: Consolidated Edison, Bell Atlantic, and so on. The firm is controlled by a regulatory agency that must approve the prices the monopolist can charge. A key goal is that the firm's revenues just cover its costs.

The measurement of costs is obviously a major task for the agency. Indeed, the attempt by the agency to estimate the proper return on capital investment is perhaps its most time-consuming activity. For example, a typical regulatory hearing involves testimony by numerous experts as to the "true" cost of capital for the firm.

In contrast, relatively little of the agency's resources are expended on the issue of the correct pricing structure. However, this situation is changing and agencies are becoming more interested in, for example, marginal cost pricing. In short, regulatory agencies try very hard to

21. Harold Demsetz, "Why Regulate Utilities?" *Journal of Law and Economics,* April 1968.

ensure that the monopolist's revenues equal its costs, and historically have been less concerned with the pricing structure used.

As a result, there is no simple way to describe the pricing structures used under regulation. Price discrimination is often employed both across customer groups (industrial, commercial, residential, and so on) and within groups (declining block rates, for instance, 5 cents per unit for the first 300 units, 4 cents per unit for the next 500 units, and so on).

Richard Schmalensee has observed,

> To the extent that utility regulators in the United States have been concerned with rate structures, they have tended to focus on prices paid by different classes of users. But this focus has typically been motivated and informed by considerations of equity or fairness rather than efficiency.[22]

Hence, regulatory agencies often try to prohibit undue discrimination across customer groups. They require the firm to allocate its total costs to customer groups and then adjust their prices if the revenues by groups do not correspond to the groups' "fully distributed costs."

There is a serious problem implicit in this procedure, however, because a large proportion of a firm's costs are usually common costs. For example, high-voltage power lines are used in common by all customer groups. And although arbitrary accounting rules can be made up to apportion these costs among groups (for instance, in proportion to their respective annual purchases of the product), none are meaningful in an economic sense as a basis for setting prices.

In summary, an important solution to natural monopoly in the United States is regulation. The regulatory solution is not an attempt to implement the ideal pricing schemes discussed earlier. Regulators do not see as their primary objective achieving economic efficiency. Rather, they appear to seek a set of prices that are not unduly discriminatory but that permit total revenues to cover total costs. However, regulatory agencies have become more interested in pricing schemes that promote economic efficiency. For example, peak pricing—charging more when demand presses on capacity, and, therefore, marginal cost is higher—is being implemented by electric utilities in various parts of the country.

Public Enterprise

The second actual solution to natural monopoly is public enterprise, or government ownership and operation of the monopoly. This is not as common in the United States as it is in other countries. The Postal Service is an example in the United States. Other examples include various government-owned electric utilities (for instance, the Tennessee Valley Authority) and Amtrak, the government-owned passenger service railroad.

---

22. Richard Schmalensee, *The Control of Natural Monopolies* (Lexington, Mass.: Lexington Books, 1979).

In principle, public enterprise would appear to be a sensible alternative. Managers would be directed to maximize economic surplus—there would be no need for regulators to try to channel the decisions of profit-maximizing firms closer to the public interest. The efficacy of public enterprise as compared to regulation, however, is a complex issue and will be examined further in Chapter 14.

Summary

This chapter has been an introduction to natural monopoly. Theoretical issues have been introduced and discussed. First, the definition of natural monopoly was developed in both the single-product and the multiple-product cases. Second, alternative policy solutions and their difficulties were discussed. The solutions included "doing nothing," various efficient pricing solutions, competition among bidders for the right to the monopoly franchise, actual regulation, and public enterprise.

In the next chapter we will elaborate extensively on the regulation alternative. Chapter 15 will examine further issues in natural monopoly regulation, with an emphasis on telecommunications.

Questions and Problems

1. Consider a single-product natural monopoly situation with the usual U-shaped long-run average cost curve. Is the range of output over which natural monopoly holds from zero to the output corresponding to minimum average cost? If not, explain how to determine the appropriate range. Use the total cost function $C(q) = 1 + q^2$ to answer this question.
2. Assume a natural monopoly with total costs $C = 500 + 20Q$. Market demand is $Q = 100 - P$.

   a. If price is set at marginal cost, what is the monopolist's profit?

   b. The answer to part a implies that linear (or uniform) marginal cost pricing has a serious problem in natural monopoly situations. Suppose that average cost pricing is employed. Find price, output, and the deadweight loss compared to part a.

   c. Now consider two-part pricing—a type of nonlinear (or nonuniform) pricing. Each consumer must pay a fixed fee regardless of consumption level plus a price per unit. Assume that the market consists of ten consumers with identical demand curves for the product. If the price is set equal to marginal cost, what is the largest fixed fee that a consumer would pay for the right to buy at that price? What fixed fee would permit the monopolist to break even? What is the deadweight loss in this case?

3. Assume the same facts as in question 2 but that now there are six "rich" consumers with each having inverse demands: $p = 100 - 6.3q$; also, there are four "poor" consumers each with demands: $p = 100 - 80q$.

a. What is the largest fixed fee that a poor consumer would pay for the right to buy at marginal cost?

b. Because the poor consumers would not be willing to pay the uniform fixed fee of $50 necessary for the monopolist to break even, the rich consumers would have to pay a fixed fee of $83.33. What is the deadweight loss in this case?

c. Third-degree price discrimination could be a solution. That is, if it is legal, resales are not feasible, and consumers could be identified by the monopolist as being rich or poor, the monopolist could charge different fixed fees to the two consumer types. If the price per unit is still equal to marginal cost, what are two fixed fees that are feasible? In this case, what is the deadweight loss?

4. If third-degree price discrimination is not a feasible alternative in question 3c, consider the optimal two-part tariff. That is, what is desired is the two-part tariff that minimizes deadweight loss—or that maximizes total surplus. One way to think about it is to imagine the case of a zero fixed fee and price equal to marginal cost. This causes a loss of $500 that must be covered. Imagine raising both the fixed fee and the price simultaneously—both can cause losses: the fee by excluding poor consumers and the price by causing deadweight consumption losses. One possibility is to exclude poor consumers and go to solution 3b. The other possibility is to keep all consumers in the market; this implies that the fixed fee should equal the consumer surplus of a poor consumer. It is optimal to take all of the poor consumers' surpluses as a fee. To see why, consider the opposite case where the poor have some excess of surplus over the fee. Then the price could be lowered, reducing deadweight losses and the surplus could be used to offset the reduction in revenues without excluding the poor from the market.

a. Find the sum of consumer and producer surplus minus the $500 fixed cost (that is, find total surplus) for case 3b where the poor are excluded.

b. Find total surplus for the case of all consumers retained in the market. Hint: An equation in P can be defined that equates to $500 the total contributions to fixed cost (10 times the fixed fee, equal to the consumer surplus of a poor consumer, plus the revenues net of variable cost generated by consumption). Hence, what is the optimal two-part tariff where all are retained in the market?

c. Compare the efficiency of the tariffs in parts a and b.

5. A multipart tariff can be superior to the optimal two-part tariff found in question 4. A multipart tariff involves a fixed fee plus multiple prices per unit, which depend upon predefined blocks of consumption.

a. Show that by making an additional two-part tariff available to the consumers that they can use at their option, the "two" two-part tariffs are Pareto superior to the optimal tariff in question 4 (that is, $F = \$38.55$, $P = \$21.50$). Let the optional two-part tariff be $P = \$20.50$ and $F = \$51$. These two two-part tariffs are equivalent to a multipart tariff that has a fixed fee of $38.55 and a price of $21.50 for the first 12.4 units and a price of $20.50 for all units above 12.4. Show this result by plotting the two tariffs on a graph that has total expenditure on the vertical axis and total units on the horizontal axis. The two straight lines representing the tariffs intersect at 12.4 units. Because consumers will always operate on the lowest line that they can attain to minimize expenditure, the multipart tariff is just the lower boundary (that is, the kinked line defined by $F = \$38.55$ and the marginal prices of $21.50 for the first 12.4 units and $20.50 thereafter).

b. Demonstrate that the two two-part tariffs are Pareto superior to the optimal two-part tariff in question 4b. Note that the optional tariff will not change the poor consumers' behavior at all. Why?

c. As a result we can focus solely on the rich consumers and the monopoly. If both are made better off by the optional tariff and the poor are kept the same, then the optional tariff results in a Pareto improvement—which is a stronger welfare statement than simply saying one tariff yields a higher total surplus. (That is, if we focus on total surplus comparisons, we ignore the fact that some people may be made worse off even though total surplus is higher.) Find the consumer surplus of a rich consumer under the two-part tariff of question 4b.

d. Find the consumer surplus of a rich consumer under the multipart tariff.

e. Find the change in profit of the monopolist. Hence a movement from two-part tariffs to multipart tariffs clearly has the potential for gains in efficiency. The intuition is that the more the "parts," the better the tariff can be tailored to the differences in willingness to pay across consumers.

6. Assume that a water distribution monopoly serves two consumer types: industrial and residential. The demands by the two classes are as follows. Industrial: $Q_I = 30 - P_I$ and Residential: $Q_R = 24 - P_R$. The company has no costs other than the fixed cost of the pipeline, which is \$328. Find the Ramsey prices. Hint: See note 18.

7. Assume a natural monopoly with total cost $500 + 20Q$ facing a demand of $Q = 100 - P$.

a. Find the price that enables the monopolist to break even. (This is the same problem as 2b.) Call this price P^*.

b. Loeb and Magat show that if the monopolist is allowed to choose its own price and to have the regulatory agency subsidize the firm by an amount equal to consumer surplus at the selected price, the monopoly will select price equal to marginal cost. What is the price and amount of government subsidy?

c. Loeb and Magat also note that a bidding process for the monopoly franchise would enable the government to recover some of the subsidy. What is the amount recovered and what is the net subsidy after bidding?

d. An alternate proposal would make use of two-part tariffs. For example, assume that the current regulated price is P^*. Now assume that the regulatory agency offers the firm the right to select any two-part tariff that it wishes as long as the consumer continues to have the option of buying at P^*. (For simplicity, assume a single consumer.) What is the two-part tariff that the monopolist will choose and what is its profit? What is the deadweight loss?

e. Assume that the government uses a bidding process to eliminate the monopoly profit in part d. The bid is in the form of a single price, like P^*, that the consumer will always have as an option to the two-part tariff. That is, the same rules are in effect as in part d except that now the bidding is for the right to offer a two-part tariff optional to some P^* that the bidding will determine. What is the low bid?

f. Compare the Loeb and Magat proposal in part c with the proposal in part e. Do both proposals give efficient prices? Are there any substantive differences?