Introduction: The Economic Rationale and Task of Regulation

I.1 Motivation

Competition, in theory if not always in practice, is nothing short of a miracle. Each firm tries to make as much profit as possible without regard (at least directly) for social welfare. Each consumer maximizes its own utility, ignoring others. Yet the result of all this selfishness is that social welfare, in the Pareto sense, becomes as great as possible. This consistency of private goals with social goals—the existence of this "invisible hand" that molds privately motivated actions into socially desirable outcomes—serves as the basis for much of economics as a field of thought and, to a great extent, provides the rationale for "free" markets.

To work, competition requires certain conditions. Most important, the market must contain many firms with none dominant, allow free entry and exit, and exhibit no externalities.[1] Unfortunately, these conditions cannot always be met. Intervention in the market is often required to ensure that the pursuit of profit does not conflict with social welfare. Natural monopoly is the classic case. Loosely defined, a natural monopoly exists when the costs of production are such that it is less expensive for market demand to be met with one firm than with more than one. In this situation it is optimal, from a cost perspective, to have only one firm. More fundamentally, a condition required for competition (that is, numerous firms) conflicts with the attainment of the benefits of competition (namely, production at lowest possible cost, which requires one firm).

[1] Contestability theory suggests that having many firms is not necessarily required for optimality, as long as entry and exit are sufficiently "free." This theory and its implications for regulation are discussed in chapter 10.

In such cases, regulation becomes important. The purpose of regulation is to ensure socially desirable outcomes when competition cannot be relied upon to achieve them. Regulation replaces the invisible hand of competition with direct intervention—with a visible hand, so to speak.

The term "visible hand" is actually quite appropriate. The regulator must work *through* the firm, inducing the firm to produce the desired outcome. If the regulator had complete information, it could simply mandate the optimal outcome, ordering the regulated firm to produce a certain amount of output with a particular set of inputs and sell the output at a specified price. Usually, however, the regulator does not have sufficient information to determine these levels. For example, the regulator usually does not know the firm's cost function and hence does not know whether the firm is pricing at marginal cost or producing with the most efficient input combination. Instead, the regulator must establish incentive schemes or other methods of regulation that induce the firm, through its desire to earn profits, to attain the socially optimal outcome. In this sense, the regulator applies a hand that molds the private profit motive into socially optimal outcomes, just as competition does. The hand is visible rather than invisible, but the molding function is the same.

The central issue of regulatory economics is the design of mechanisms that regulators can apply to induce firms to achieve optimal outcomes. In any particular setting, this issue consists of two tasks. First, the optimal outcome must be characterized. In many situations, this characterization is a direct application of concepts from microeconomic theory, such as that price equals marginal cost at the optimal output level. However, optimality is not always so easily identified. For example, when marginal-cost pricing results in the firm losing money, what is optimal? The firm cannot lose money indefinitely and stay in business.

Once the optimal outcome is characterized, the second task is to design a regulatory mechanism that induces the regulated firm to act in a way that results in this outcome. The firm is (usually) assumed to act so as to maximize its own profits.[2] Under an effective regulatory

[2] Researchers have also examined regulatory mechanisms under the assumption that firms maximize something other than profit, such as revenue, output, rate of return to shareholders' equity, or a composite of variables. Seminal studies include Kafoglis 1969, Bailey and Malone 1970, Zajac 1970, and Bailey 1973. As Baumol and Klevorick (1970) point out, the analysis proceeds exactly as under profit maximization, only with a different maximand.

mechanism, the firm obtains greater profit when it chooses the optimal output, prices, and inputs than at any other level of these variables. That is, effective regulation establishes a situation in which the outcome that is socially optimal also generates the most profit for the firm, such that the firm chooses it voluntarily.[3] Creating this consistency between social welfare maximization and the firm's profit maximization is the crux of regulatory economics.[4]

In this book we concentrate on the regulation of natural monopoly. There are several reasons for this restriction. First, competition is clearly inappropriate in these situations, so that the introduction of a visible hand is warranted.[5] Second, there is only one firm to consider, so that interactions among firms do not complicate the analysis.[6] Third, and perhaps most important, public utilities, which are usually natural monopolies, play an essential role in the nation's economy and constitute one of the most prevalent settings for regulation in the country. Electricity, natural gas, local phone service, waste disposal, cable television, and many other goods and services are provided by public utilities subject to regulation by local or state agencies.

Although the book concentrates on natural monopolies, the prin-

[3] Profits in the optimal outcome need not be large to induce the firm to choose this outcome: they only must exceed profits at each other outcome. For example, it is possible, as demonstrated in the ensuing chapters, to establish regulatory mechanisms under which the firm just breaks even if it chooses the optimal outcome and loses money at any other outcome. Under these mechanisms, the firm's profits are higher in the optimal outcome than any other (because zero is greater than any negative number), but profits are still as low as possible for the firm to remain solvent.

[4] The issue of how to regulate a natural monopoly is one case of a broader class of problems that is referred to generically in the literature as the "principal-agent problem." In problems of this kind, the principal must act through the agent, who has more information than the principal. A mechanism that the principal uses to activate the agent is called "incentive compatible" if the mechanism induces the agent to report information truthfully to the principal. Such a mechanism establishes incentives for the agent that make the goals of the agent consistent with those of the principal—hence the term "incentive compatible." In our context, the development of optimal regulatory procedures is equivalent to the development of incentive compatible mechanisms. If the firm reports all information on costs and demand truthfully, the regulator can determine the optimal prices, output, and inputs and mandate the firm to choose them. In this book we use terms that are descriptive of the specific case of natural monopoly, namely, "the regulator," "the firm," and "optimal regulatory procedures" for "the principal," "the agent," and "incentive compatible mechanisms," respectively.

[5] Actually, chapter 10 describes conditions under which a regulator need not intervene directly in order to attain optimality, even with a natural monopoly. However, in these situations, the regulator must still perform some functions to ensure that the conditions are maintained. These functions are in themselves a form of intervention, though indirect.

[6] Interactions among firms become relevant if entry is allowed, as in chapter 10.

ciples and lessons are relevant in all regulated settings, that is, in all situations in which people would like to harness the profit drive of firms to produce particular outcomes. The form that this relevance takes is important to recognize. The concepts and, in particular, the regulatory mechanisms that are described in this book are not intended to be applied *directly*. Effective regulation in the real world must consider so many factors—political, psychological, practical—that the application of any particular economic model would be extremely naive. Rather, the economic concepts provide insights into the process and purpose of regulation; they condition the way one thinks about regulation and the approach one takes in handling individual problems that arise in a regulatory setting. In short, they provide what Erik Erikson identifies as the true contribution of any field of thought, namely, "a way of seeing things."

Two further notes are required. Throughout the book we assume that regulators try to benefit society. This need not be the case, of course. Regulators can have their own agendas that include career advancement, self-aggrandizement, political support, and the like. At an extreme, capture theory (as described, for example, in Posner 1974) suggests that over time regulated firms gain control over the process by which they are regulated. Our emphasis on publicly motivated regulators is not intended to reflect an opinion on regulators' true motives. Rather, the emphasis reflects the current state of the field. Differences in regulators' goals (or, more precisely, in the factors that give rise to regulation) have been discussed extensively as a way of explaining the regulation that actually occurs in various settings; however, little has been said about how to design optimal procedures when regulators have these other goals. In theory, of course, the concepts in the book could be applied to regulators themselves, the issue being how to devise procedures that induce consistency between regulators' private goals and the public welfare.

More constraining is the assumption, also maintained throughout, that benefiting the public consists of maximizing total surplus (where total surplus is the sum of consumers' surplus and all firms' profits). Fairness and equity, however defined, are important social criteria that are ignored by this approach. Goals such as technical advancement, continuity with the past, conservation, and so on can also be important in certain settings and yet are not addressed.[7] Although

[7] Often these goals are actually manifestations of surplus maximization or equity considerations. For example, it might be considered unfair for large changes in prices to

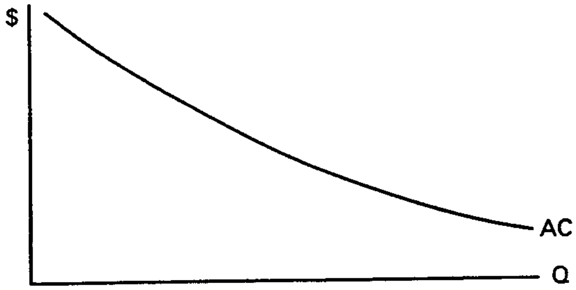

Figure I.1
Average cost curve under economies of scale

equity and other goals are clearly relevant, economists have had relatively little to say about them, certainly in the realm of designing regulatory processes. In defense of the traditional approach, experience has shown that insights obtained from the analysis of surplus maximization are helpful in examining and designing procedures that serve other goals. I hope the reader will discover the same.

The following sections provide some preliminary information. Section 2 characterizes a natural monopoly, identifying economies of scale and scope. Section 3 examines social welfare under natural monopoly, distinguishing the first-best and second-best outcomes.

I.2 Characteristics of a Natural Monopoly

A natural monopoly arises from two sources: economies of scale and economies of scope. Economies of scale exist when the average cost of production decreases as output expands. Figure I.1 illustrates such a situation. The average cost curve slopes downward, indicating that average cost falls as output increases.[8] [9]

be instituted abruptly. Continuity with the past becomes, therefore, an expression of equity. However, sometimes these goals are indeed separate concepts. For example, technical advancement can be seen as an aesthetic pursuit that expresses a desirable and basic human drive, or as evidence of preeminence, independent of the surplus it generates and the cost of its development.

[8] Economies of scale can be defined equivalently in terms of total cost. Suppose a firm expands its output by a given percentage (say 10%). If the total costs of the firm increase by *less* than this percentage (say, by 8%), economies of scale exist. The two definitions are clearly the same. Average cost is total cost divided by output: *TC/Q*. If total cost (the numerator) increases by a smaller percentage than output (the denominator), the ratio of these two terms must decrease.

[9] A distinction is necessary between "pecuniary" and "nonpecuniary" economies of scale. Often a large firm can negotiate with its suppliers to obtain lower prices for

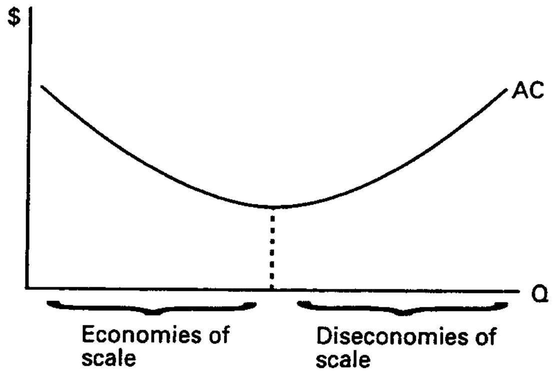

Figure I.2
Economies and diseconomies of scale

The most prevalent source of economies of scale are fixed costs, that is, costs that must be incurred no matter how many units of output are produced. Electricity production is a case in point. A generation plant is required to produce the first kilowatt-hour; yet many kilowatt-hours can be produced in the same plant.[10] When output expands, the fixed costs (in this case, the costs of the plant) are spread over more units, such that average cost declines.

Economies of scale can exist over some ranges of output but not others. For example, at low levels of production, scale economies may be present, while at larger output levels the opposite—diseconomies of scale—may occur.[11] This situation gives rise to the standard U-shaped average cost curve shown in figure I.2.

The existence of natural monopoly depends on the range of economies of scale relative to market demand. In particular, a natural monopoly exists in the production of one good only if economies of scale

inputs than would be charged if the firm were smaller. Average cost therefore declines as the size (i.e., output) of the firm increases. However, the reduction in average cost represents simply a transfer of income from the suppliers to the firm, such that the total cost to society (including both the firm and the suppliers) is unaffected. Reductions in average costs that reflect transfers only are called pecuniary, while those that represent an actual reduction in the resources used per unit of output are called nonpecuniary. From a social perspective, only nonpecuniary economies are relevant. We therefore use the term economies of scale only in reference to nonpecuniary economies. A similar distinction and usage is relevant for economies of scope.

[10] Other inputs, such as coal in a coal-burning plant, must be expanded, but the plant itself need not be.

[11] At high levels of output, management might not be able to oversee closely all the operations of the firm, giving rise to inefficiencies that can dominate any cost advantages of large-scale operation.

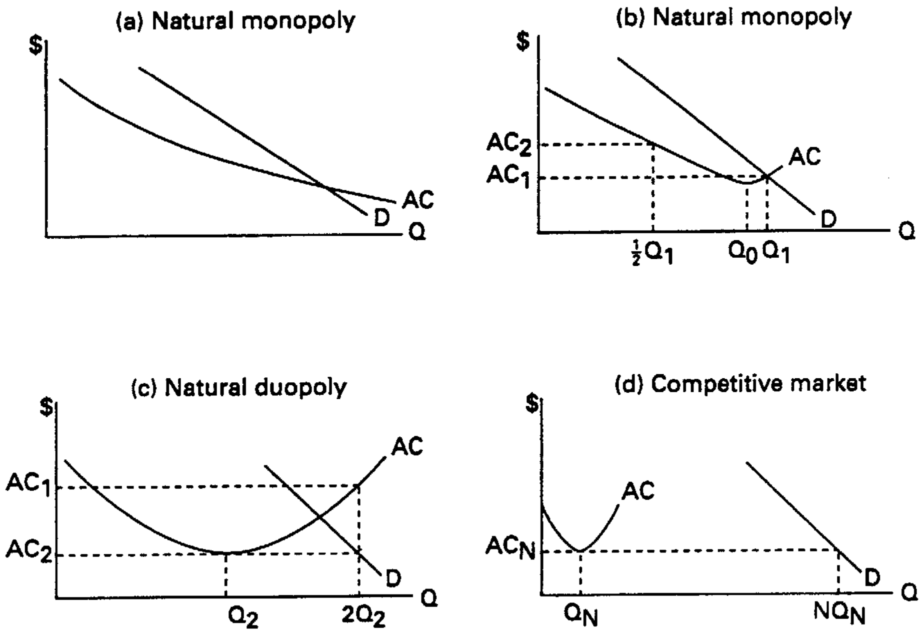

Figure I.3
Relation of average costs to demand

exist over a sufficient range of output relative to demand, where "sufficient" is defined by the situation. Four illustrative cases are shown in figure I.3. Panel (a) depicts the standard situation: average costs decline over all levels of output that would be demanded at any price, that is, over the entire range to the left of the demand curve. A natural monopoly clearly exists in this situation. A natural monopoly can exist, however, with economies of scale existing over a smaller range of output. Panel (b) depicts such a case. Economies of scale continue only to output Q_0, after which diseconomies set in. One firm could supply Q_1 output at an average cost of AC_1. If two firms supplied this output, each firm would incur average costs of $AC_2 > AC_1$ if they shared the market equally. If the two firms split the market unequally, their average costs would differ, but the total cost with two firms would always exceed that with one firm. At any division of output, production with two firms costs more than with one firm, indicating that a natural monopoly exists.

If economies of scale are exhibited over an even smaller range of output relative to demand, then a natural monopoly does not exist. In panel (c), two firms can produce output Q_2 apiece at an average cost of AC_2; one firm producing the same total output, $2Q_2$, would incur much higher average costs. This case constitutes a natural duopoly. Competition occurs when economies of scale are exhausted at a level of output that is small compared to market demand, as in panel (d), such that minimum-cost production is attained with numerous firms.

When more than one good is being produced, natural monopoly can arise from economies of scope as well as economies of scale. With several goods, there are sometimes shared equipment or common facilities that make producing them together less expensive than producing them separately. Economies of scope are said to exist if a given quantity of each of two or more goods can be produced by one firm at a lower total cost than if each good were produced separately by different firms.

This definition can be expressed in terms of the total cost function and illustrated graphically. Let the total cost to a firm of producing two goods in the quantities x and y, respectively, be represented as $f(x,y)$. The cost of producing good x only is, therefore, $f(x,0)$, because the firm produces none of good y. Similarly, the cost of producing good y only is $f(0,y)$. Economies of scope exist if $f(x,y) < f(x,0) + f(0,y)$. That is, the cost of producing both goods together, $f(x,y)$, is less than the combined cost of having one firm produce good x but none of good y, $f(x,0)$, and another firm produce good y but none of x, $f(0,y)$. Figure I.4 illustrates this possibility. The cost function facing any firm in the industry is shown as the shaded surface, which gives the cost of producing any combination of the two goods. Point A represents production of quantities x_A and y_A. The cost function evaluated at point A is $f(x_A, y_A)$, the cost to a firm of producing both goods. This cost is the distance OL on the cost axis. If a firm produced x_A only and no y, then its costs would be $f(x_A,0)$, which is the distance OM. Similarly, a firm producing y_A but no x would incur costs of $f(0,y_A)$, distance ON. The combined costs of the two firms, each producing one of the goods, is $ON + OM$ (or, as given on the graph, the distance from O to $N + M$). Because $N + M$ is higher than L, it is less costly to have one firm produce these quantities of the two goods than to have two firms produce the two goods separately.

As with economies of scale, it is possible for economies of scope to exist at some levels of outputs of the goods and not at others. For example, it may be cheaper to have one firm produce two goods when small quantities of the goods are being produced, but not for large quantities (or vice versa). Whether having one firm is desirable from

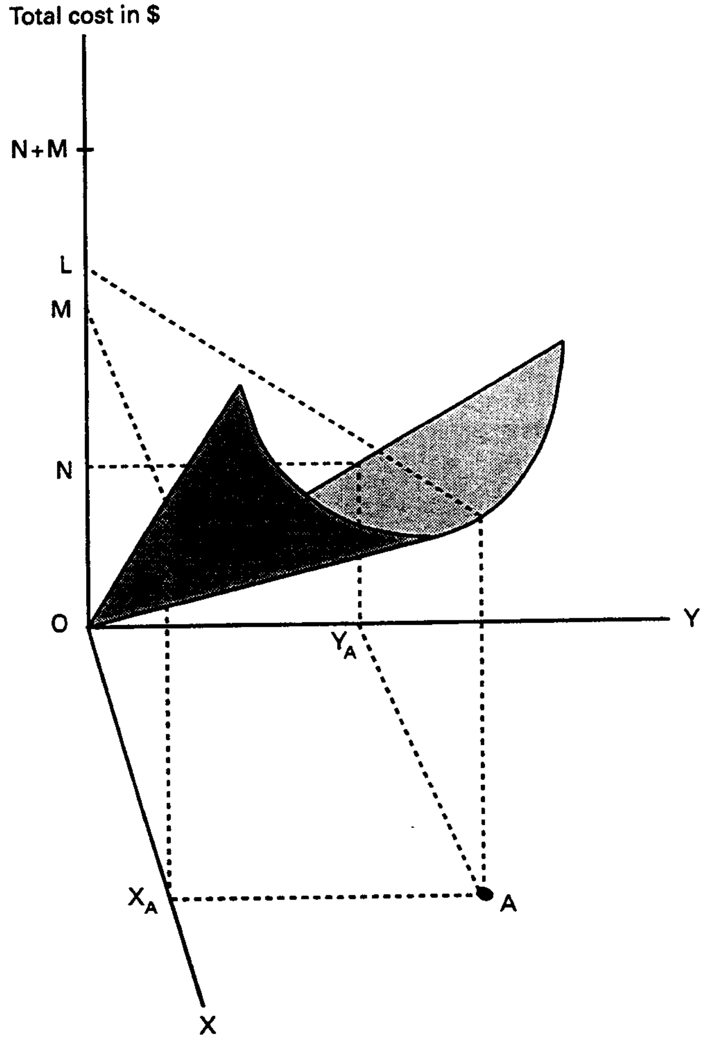

Figure I.4
Economies of scope

a cost perspective depends on how these regions of economies and diseconomies of scope relate to the demand for the two goods.

The existence, or relevance, of economies of scope often depends on how goods are defined. Local and long-distance telecommunication service is a case in point. If local and long-distance service are considered to be the two goods, there are strong grounds for believing that economies of scope exist. The wires that connect the phone to the "local exchange unit" (that is, the switchboard that directs the call to its destination) are used for both local and long-distance calls. Having two companies provide the two services separately entails redundant equipment: two sets of wires going to each phone, one for each company. Under this definition of goods, it would seem preferable to have one company provide both local and long-distance service, so as to obtain the benefits of the economies of scope. This was the rationale for AT&T, prior to the divestiture, being allowed a monopoly franchise for telecommunication service in most areas of the United States.

The relevant goods can be defined differently, however, in which case the argument for economies of scope is not as strong. Consider the provision of a long-distance call from a phone in one city (the origin city) to a phone in another city (the destination city.) The call consists of three parts: first, the call moves along a wire from the originating phone to the local exchange unit in the origin city; it is there combined with other calls and placed on a larger wire to the local exchange unit in the destination city; at that point it is disentangled from the other calls and moved along a wire from the local exchange unit to the phone being called. Three services are being provided in moving the call: service from the phone to the local exchange unit in the origin city, service between local exchange units, and service from the local exchange unit in the destination city to the phone receiving the call. Having three separate firms—a local phone company in the origin city providing service between phones and local exchange units, a long-distance carrier providing service between local exchange units in different cities, and a local phone company in the destination city providing service between phones and local exchange units in that city—is not necessarily more costly than having one firm provide all three services. Only one wire goes to each phone, as provided by the local phone company, such that redundancy in these facilities does not occur; and no other redundancies are immediately obvious. In fact, the concept that economies of scope do not seem to exist in the provision of services defined in this way is the economic justification for the divestiture of AT&T. Today, service between phones and local exchange units is provided by a local phone company in each area, and service between local exchange units is provided by long-distance carriers. Furthermore, since economies of scale in the provision of service between local exchange units is thought to be exhausted at a level of output that is small compared to market demand, competition is permitted and encouraged in the long-distance market (rather than allowing AT&T to hold a mandated monopoly, as would have been appropriate if a natural monopoly existed in long-distance service).

Economies of scope can exist with or without economies of scale, and vice versa. For example, it is possible that joint facilities can be used in the production of two goods and yet expanding production of both raises costs more than proportionately. Whether a natural monopoly exists depends on the overall cost situation, considering both economies or diseconomies of scope and/or scale. Economists use the term "subadditivity" for this purpose. A cost curve is said to exhibit subadditivity at a given level of one or more outputs if the cost of producing these outputs is lower with one firm than with more than one firm, regardless of how the output might be divided among the multiple firms.

Consider, for example, a situation with two goods, labeled *A* and *B*, and the possibility of production by two firms, labeled I and II, instead of one. Several different divisions of output between the two firms are possible. Firm I could produce all of good *A* and firm II all of good *B*. This arrangement would be appropriate if economies of scale existed, but not economies of scope. Or, each firm could produce both goods, with each supplying half of the total output of each good. This arrangement would be cost-effective if economies of scope existed but diseconomies of scale started to arise at half the market output of each good. Or, firm I could supply one-third of the units of good *A* and two-thirds of the units of good *B*, while firm II produced the remaining two-thirds of good *A* and one-third of good *B*, and so on, for numerous other possible arrangements. Costs exhibit subadditivity only in the event that one firm producing all of goods *A* and *B* is cheaper than any of these, or any other, arrangements with two or more firms. Thus, the concept of subadditivity incorporates considerations of both scope and scale and identifies whether, given all considerations, one firm is cheapest.

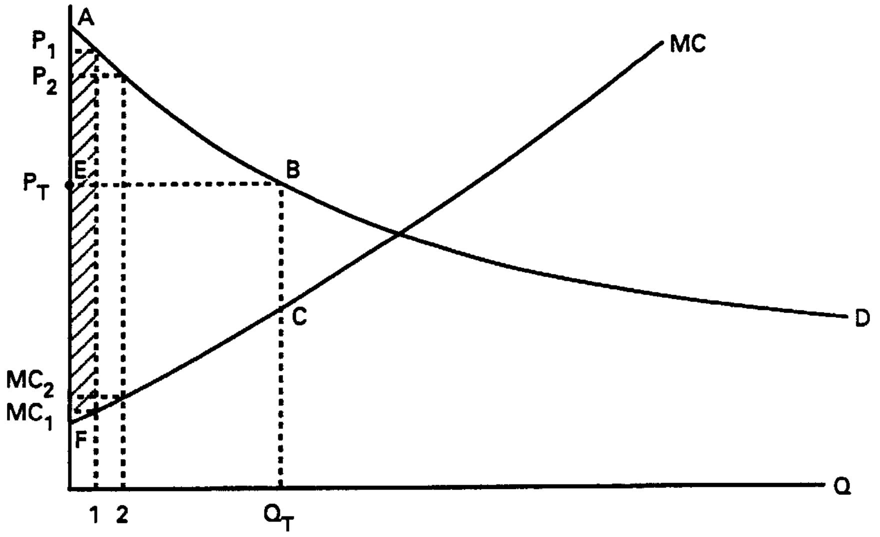

Figure I.5
Total surplus

Just as economies of scope and scale can exist at certain levels of output and not at others, so can subadditivity. A natural monopoly exists when the cost curve exhibits subadditivity in the relevant range of market demand. Because subadditivity essentially means that natural monopoly exists, we simply use the latter term throughout the book.

I.3 Welfare Concepts with Natural Monopoly

Given that a natural monopoly exists, what output, price, and inputs should the regulator try to induce it to choose? That is, what is the optimal outcome?

The definition of optimal outcome relies critically on the concept of total surplus, a term readers will recall from microeconomics. To refresh the memory, a brief discussion of the term is useful. Total surplus is the dollar amount by which the benefits from consumption of a good exceed the cost of producing it. Consider figure I.5, which illustrates typical demand and marginal cost curves for a good. The total surplus that accrues from Q_T units of the good is the area *ABCF*, the area above the marginal cost curve and below the demand curve, up to Q_T units of output. To see this, consider the benefits and costs of producing each unit of output up to Q_T. The first unit is labeled 1 in the graph. Consumers are willing to pay P_1 for this first unit. (At a price P_1 consumers demand one unit, which means that they value that unit at P_1.) The cost of producing this unit is MC_1. The benefits to consumers exceed the cost of this unit by $P_1 - MC_1$, which is the shaded column in the graph. This shaded column is the total surplus from the first unit.

Consider now the second unit. Consumers are willing to pay P_2 for the second unit and the unit costs MC_2 to produce. Total surplus from this second unit is $P_2 - MC_2$, namely, the area below demand and above marginal cost. Continuing for all units up to Q_T, total surplus from all units is the area *ABCF*.

Total surplus consists of both consumer surplus and producer's profit. Suppose price is P_T. Consider only the first unit of production. Consumers benefit by P_1 for this unit of output and must pay P_T for it. Their net benefit, or surplus, is therefore $P_1 - P_T$. The firm obtains revenues of P_T from this first unit and must pay MC_1 to produce it. Its profits are $P_T - MC_1$. Total surplus on this unit ($P_1 - MC_1$) is the sum of consumer surplus ($P_1 - P_T$) and profit ($P_T - MC_1$).

Consider now the total output Q_T. Using the same logic as above for each unit of output, consumer surplus for all Q_T units is the area *ABE*, the area above price and below the demand curve. Profit is the area *EBCF*, the area above the marginal cost curve and below the demand curve up to Q_T. Total surplus (*ABCF*) is the sum of consumer surplus (*ABE*) and profit (*EBCF*).[12]

Optimality can now be defined. The optimal outcome is that which provides the greatest total surplus, that is, the largest dollar value of benefits in excess of costs.[13]

From microeconomics, we know that total surplus is maximized when the firm prices at marginal cost, sells the output demanded at this price, and uses the least costly input combination to produce the

[12] No fixed costs are included in this example. If there are fixed costs, total surplus is the area *ABCF* minus these fixed costs. Consumer surplus is the same as before (*ABE*), and profit is *EBCF* minus the fixed costs. Alternatively, fixed costs can be incorporated in the graph as part of the marginal cost of the first unit of production. In this case, the marginal cost curve would be very high for the first unit and lower for the second and subsequent units.

[13] Note that this definition ignores the issue of equity, namely, which consumers obtain benefits and the allocation of surplus between consumers and firms. When total surplus is maximized, it is theoretically possible to distribute the surplus in a way that makes every party (each consumer and each firm) better off than under any other possible arrangement. (Essentially, with a larger pie, each person can get a larger slice.) In practice, of course, implementing such a distribution is difficult.

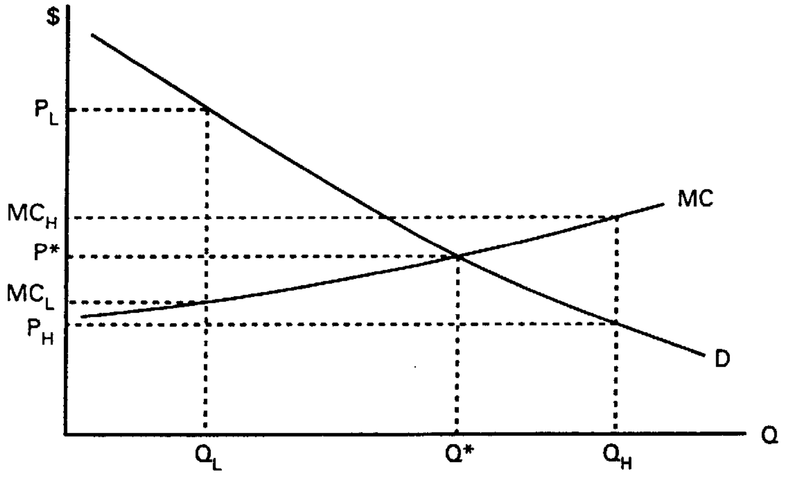

Figure I.6
Optimal outcome: first-best

output. For a one-output firm, the situation is illustrated in figure I.6. The optimal output is Q^*. The logic indicating that this output is optimal can be observed by considering any other output level, say Q_L. At Q_L, consumers are willing to pay P_L for an extra unit of the good. The cost of an extra unit, given that Q_L is produced, is MC_L, which is less then P_L. Because consumers are willing to pay more than it costs the firm to produce one extra unit, surplus increases when the unit is produced. That is, surplus increases as production is expanded from Q_L toward Q^*. A similar argument holds for levels of output above Q^*. At Q_H, the cost of an extra unit (MC_H) is greater than the amount that consumers are willing to pay for the unit (P_H) such that expanding production decreases total surplus. Stated conversely, decreasing output toward Q^* increases surplus. Only at Q^* can surplus not be increased by expanding or contracting output. Given that Q^* is the optimal quantity, the optimal price is P^*. At this price, any consumer who is willing to pay at least the marginal cost of the good obtains it, and those not willing to pay the marginal cost do not.

In the presence of economies of scale, the firm necessarily loses money when pricing at marginal cost. Economies of scale imply that the average cost curve of the firm is downward sloping. Declining average costs mean that marginal cost is below average cost.[14] There-

[14] If an extra unit costs less to produce than the average of all previous units, then producing an extra unit lowers the average cost. Stated conversely, if average cost is declining, marginal cost is necessarily below it.

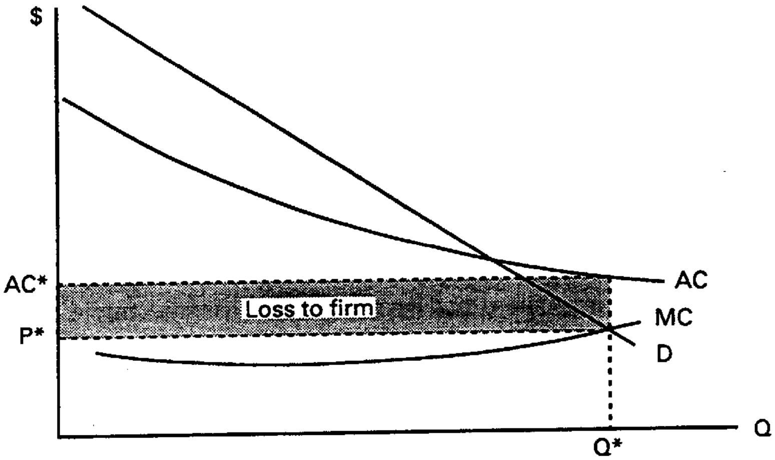

Figure I.7
Firm loses money at first-best price under economies of scale

fore, in the presence of economies of scale, marginal cost is below average cost. When price is set at marginal cost, as required for optimality, the firm loses money on each unit sold. Figure I.7 illustrates the problem. At Q^* and P^*, the firm loses the amount given in the shaded area: the amount by which average cost exceeds price, times the number of units sold.

A firm cannot lose money indefinitely and remain in business. In theory, the firm could be subsidized by the amount of its loss each period. In the United States, however, the tradition has been not to subsidize public utilities directly, under the belief that customers should pay the full costs of production. More important, if the firm is subsidized, the procedure by which the funds are raised (such as taxing income or property) distorts prices elsewhere in the economy away from marginal cost.

Without a subsidy, the only solution is for prices to be raised sufficiently for the firm to break even.[15] In a one-output situation, the requirement is clear: price must be raised to average cost. This price

[15] Chapters 2 and 7 suggest that if the firm charges a different price for different levels of consumption (e.g., a higher price for consumption up to a certain number of units, and then a lower price for consumption beyond that number), the price for marginal consumption can sometimes, depending on various factors, be retained at marginal cost without causing the firm to lose money. In these cases, the higher price for low levels of consumption provides the needed subsidy: essentially the firm is taxing its customers for the additional funds required to break even. For the present purpose, however, we assume the firm charges one price for each good independent of consumption level.

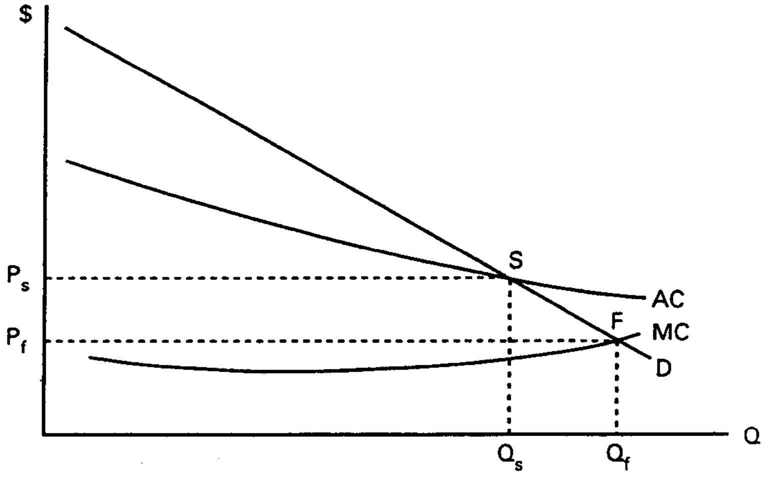

Figure I.8
First- and second-best outcomes

is optimal in the absence of subsidy because any lower price would result in negative profits, which is infeasible, and any higher price would distort price further away from marginal cost than necessary. Figure I.8 illustrates the situation. Production at *F* provides the greatest total surplus, but results in the firm losing money. *S* is the closest point that allows the firm to break even.

Points *F* and *S* represent two different concepts, or definitions, of optimality. Total welfare is as high as possible at *F*, where price equals marginal cost. This is called the "first-best" outcome, or first-best pricing, to indicate that no other outcome provides greater surplus. If at all possible, this is the outcome that the regulator would like to achieve. In the case of natural monopoly, the firm obtains revenues under first-best pricing that are insufficient to cover its costs.

At point *S*, total surplus is greater than at any other outcome that allows the firm to earn at least zero profits. This is called the "second-best" outcome, reflecting the fact that it provides less surplus than the first-best outcome. The regulator would like to achieve the second-best outcome if first-best is infeasible.[16]

[16] Stated alternatively, *F* is the unconstrained maximum of total surplus, while *S* is the constrained maximum, with the constraint being that profits be at least zero. *F* provides greater surplus than *S* even though the firm's profits at *F* are negative because consumers obtain sufficiently greater surplus at *F* compared to *S* to compensate for the loss in profits.

In competition, the distinction between these two concepts of optimality is not re-

If the firm produces two or more goods, the second-best outcome is not immediately obvious. Unlike a one-output firm where zero profit is attained only when price equals average cost, many combinations of prices for a multi-output firm can result in zero profits. For example, an energy utility that provides gas and electricity can make zero profits by pricing electricity at marginal cost but gas far above marginal cost (earning profits on gas to make up for the losses on electricity), or by pricing electricity far above its marginal cost with gas at marginal cost, or by pricing both moderately above their marginal costs. There is, in fact, an infinite number of possible price combinations for the two goods that would result in zero profits. Of these combinations, the one that provides the greatest total surplus is the second-best outcome.

While the definition of second-best pricing is straightforward in a multi-output situation, the identification of which price combination actually constitutes the second-best is not. One of the major accomplishments in the field of regulatory economics has been to determine, or characterize, the second-best prices for multi-output natural monopolies. We address this issue in chapter 4. For the present, distinguishing the concepts of first- and second-best is sufficient.

quired. In equilibrium, each firm produces at the minimum of its average-cost curve, where marginal cost equals average cost..Points *F* and *S* are therefore the same. These points being different only arises in a natural monopoly situation where the firm does not produce at the minimum of the average-cost curve (because, usually, the minimum is beyond market demand).