10 Optimality without Regulation

10.1 Introduction

Regulation in the real world is far from optimal, and it is perhaps unrealistic to believe that it ever will be. Posner (1969, 1970) has catalogued the inefficiencies of regulation in convincing (and depressing) detail. However, academic articles are not really needed as proof: any observer of regulatory processes in the real world can easily identify areas of extensive waste, mismanagement, missed opportunities, and other social ills.

The thrust of this book so far has been to explore mechanisms of regulation that can, it is hoped, reduce these inefficiencies, or, more accurately, provide insight that will allow some of these inefficiencies to be avoided. This approach presupposes that regulation is actually needed in natural monopoly situations and that the issue is simply how to devise the best regulation.

Several authors (most prominently Demsetz 1968; Posner 1972; Baumol, Bailey, and Willig 1977; Sharkey 1982; and Baumol 1982) have argued that, contrary to standard concepts, the existence of natural monopoly is not, in itself, grounds for regulation. Under assumptions that are similar to those maintained in standard theories of regulation, these authors have shown that optimality (or at least some major aspects of optimality) can be attained without regulation, even with only one producer.

Competition among numerous firms is shown by these authors to bring about optimality even in natural monopoly situations. However, the competition arises *not* among firms that *actually* produce in the industry, but rather among firms that *could* produce. Even though, in a natural monopoly situation, only one firm actually produces the good, numerous firms *could* produce the good. Pressure from these

potential producers, exerted in somewhat different ways in the theories of different authors, forces the monopolist to produce with least-cost inputs and price as low as possible.

These theories suggest that perhaps regulation, with all its concomitant inefficiencies, is unnecessary and that market forces can be relied upon to attain optimality. This suggestion is certainly appealing. However, as with all theories of regulation, it is necessary to recognize that the abstract world of the theories is indeed just that: an abstraction. Direct application of the concepts in the real world raises numerous complications and could foster outcomes that differ substantially from those derived in theory. Williamson (1976), Schwartz and Reynolds (1983), and others have discussed these limitations. The value of the theories is that they describe important forces that operate toward optimality, forces whose power had not been recognized, or emphasized, in earlier concepts of natural monopoly. The regulator, understanding these forces, can use them along with the others to foster greater efficiency.

In the following sections, we describe the manner, power, and limitations of competition among potential producers. Section 10.2 describes the suggestion of Demsetz and Posner that the monopoly franchise (that is, the right to be the monopolist) can be auctioned off to the firm that offers to charge the lowest per-unit price to consumers. With enough noncolluding bidders, such an auction can be shown to result in efficient production and zero (economic) profits for the winning firm, at least in a world that does not change much over time. Section 10.3 introduces "contestability" as it applies to natural monopoly, a concept proposed in various forums by Bailey, Baumol, Panzar, and Willig. The notion of contestability rests on the existence of sufficiently free entry and exit of new firms into an industry. If a new firm can enter a monopolized industry without incurring costs that cannot be recouped if the firm later exits, the monopolist producing in the industry must produce efficiently and earn zero profit in order to avoid having its market taken over by an entrant.

The theory of contestability and the concept of auctions when applied in a world that changes over time, suggest that regulators should encourage, rather than prevent, entry into the market of a natural monopolist. However, while allowing entry can perhaps induce optimality in some situations, it can also actually destroy it in others. Depending on the cost structure faced by a natural monopolist, new firms might be able to enter and make a profit even if the monopolist

is acting optimally. Faulhaber (1975) and Sharkey (1981) show that if a natural monopolist charges Ramsey prices, new firms might be able to enter profitably and supply a portion of the monopolist's market at a lower price. Allowing entry in these cases would prevent the attainment of Ramsey prices in equilibrium. In fact, Panzar and Willig (1977) and Sharkey (1981, 1982) show that, under some cost structures, new firms will be able to enter and make a profit no matter what prices the natural monopolist charges. Clearly, equilibrium with one firm, which is best from a cost perspective, would be unattainable if entry were allowed in these situations. Section 10.4 discusses these ideas. The findings are derived from the definition of "sustainable prices," namely, prices that prevent entry in a contestable market.

Section 10.5 concludes with a discussion of the basic question: To what extent should the regulator rely on market forces such as entry to achieve optimality, versus direct regulation? Following Williamson, we show that the two ways of handling natural monopoly are not as different, when applied in the real world, as might appear in theory. Furthermore, the appropriate approach cannot meaningfully be determined in general. Rather, the particular circumstances of each individual situation must be considered in choosing between the approaches, or, more accurately, in determining the most effective blend of approaches for each situation.

10.2 Auctioning the Monopoly Franchise

The concept is simple. Suppose increasing returns exist in the production of a good such that having one firm produce the good is desirable from a cost perspective. Though only one firm will produce the good, assume that many firms are capable of producing it. The regulator is assumed to have the power to decide which firm will be allowed to produce the good; that is, the regulator is able to award the monopoly franchise. The regulator auctions off the franchise in the following way. The regulator announces that it will accept bids from all firms that are willing and able to produce the good. The bid from each firm will consist of the price that the firm agrees to charge customers if awarded the franchise. The regulator will choose the firm that offers the lowest price. The winning firm becomes the monopolist and is required to charge customers the price it bid in the auction.

If numerous firms face the same technology and production costs, and these firms do not collude when bidding, the price of the product

will be bid down to the point at which the winning firm makes (essentially) zero profits with least-cost production. This fact can be discerned by following the course of such an auction. Suppose that, at one point in the auction, the lowest price offered allows the firm presenting this bid to earn strictly positive profit. Another firm that can produce at the same costs as this first firm could offer a lower price and still make a positive (though smaller) profit. Given the choice of zero profit (which a firm earns if it does not win the franchise) and a small but positive profit, this other firm will choose the latter, bidding below the previously lowest bid. Similarly, suppose the lowest bid is made by a firm that would make zero profit, but would engage in some form of inefficiency in production. Another firm would see that it could produce without waste. With lower costs, its profits would be positive at the same price offered by the wasteful firm; it could therefore offer a slightly lower price and still make a positive profit. Again, given a choice between zero profit from not winning the franchise and a small but positive profit, this other firm would choose the latter and bid below the previously lowest bid. This process will continue until, for the winning price, profits are essentially zero with least-cost production.

It is useful pedagogically to describe the auction proceeding sequentially, as above, with firms bidding successively lower prices until profits are squeezed to zero. However, the process could actually occur all at once. If each firm knew that other firms faced the same technology and production costs, each firm would realize that it could only win the auction if it bids a price that provides zero profit with least-cost production. Each firm would make this bid, and the regulator would be faced with a multitude of identical bids, each of which consists of the lowest possible price to consumers. Which bid the regulator chooses at this point is immaterial from a welfare perspective. The point is simply that profits would be zero and production would be efficient without any direct intervention by the regulator (other than holding the auction).

It is important to note the type or extent of optimality achieved. First-best optimality is not attained. Pricing at marginal cost would, due to the presence of scale economies, require negative profits, and no firm would be willing to bid price down that far. As usual, second-best optimality is the best one can hope for. In a one-good situation, all aspects of second-best optimality are attained: price is at average cost because profits are zero, and least-cost production methods are

utilized. In a multigood situation, the auction as described need not result in the Ramsey prices, and so this aspect of second-best optimality is not necessarily attained. The auction assures that profits are zero with least-cost production, but does not determine which of the various price combinations that result in zero profit will be offered.

The intriguing aspect of this analysis is that having a monopoly producer does not result in a monopoly price. Market forces, in the form of competition among potential producers, pushes price toward cost. In principle, there is no need for regulation of the monopolist, because there is no distortion of price from cost that would require regulation.

The limitation of this approach to natural monopoly arises when it is recognized that costs and demand change over time such that a price that was optimal at one point in time might not be optimal later (Williamson 1976). In a traditionally competitive market, equilibrium price adjusts to changes in costs and demand, reaching the new optimum after each change. However, in the auctioning of a natural monopoly franchise, adjustment to changes does not occur automatically. Recall that the regulator holds the auction, awards the franchise to the firm that offers to sell at the lowest price, and then requires that the firm charge that price when it becomes the monopolist. In a static world, the regulator and the winning firm could enter a long-term contract that specified the price that the firm would charge, essentially locking the firm into the price it offered at the time of the auction. With no changes in demand or costs, this contract would insure that the optimal price is maintained.

In a changing world, however, the optimal price changes. A long-term contract that locks the firm into a price that was appropriate at one point in time might later, when circumstances change, force the firm into bankruptcy or provide windfall profits. In the face of this fact, the regulator might attempt to write a contract that contains contingency clauses for possible future events. These contingency clauses could take either of two forms. The contract might (1) specify how the price will change if certain events occur, or (2) establish a procedure by which prices are revised periodically. The first approach attempts to list all possible future events and the price that would be charged if the events came to pass. Given the vast number of possibilities, this approach would be incomplete at best. Furthermore, the regulator might not be able to observe directly whether an event occurred; for example, the regulator might not know whether a new technology

had allowed costs to drop. In an effort to increase its profits, the firm would not necessarily report events truthfully to the regulator. For example, the firm would have a clear incentive to report smaller cost savings from a new technology than actually occurred. This asymmetry of information places the regulator in the same situation as under direct regulation: needing a mechanism that would induce the firm to act optimally when the regulator does not have sufficient information to identify the optimal behavior beforehand. Hence, a contract with contingency clauses that relate price to certain events ends up being essentially the same as direct regulation.

The second contracting approach recognizes that all future events cannot be foreseen and, instead of trying to list all possible events, establishes a procedure by which price is reviewed periodically. In these reviews, the events that have actually transpired are examined, and the price is adjusted to take account of these changes. However, these reviews become, just as under the first contracting approach, the same as direct regulation. The firm has more information than the regulator on the changes in demand and cost that have occurred and cannot be relied upon to report this information truthfully. The contract therefore needs to establish a procedure that induces optimal behavior when the regulator lacks sufficient information to identify this behavior directly, which is exactly the task of direct regulation. In a changing world, the distinction between direct regulation and reliance on market forces fades, at least when the market forces are harnessed though an auction with a long-term contact.

Given the problems with a long-term contract, it has been suggested (e.g., Posner 1972) that the regulator instead write a short-term contract with the winning firm. When the contract expires, a new auction is held, and the monopoly franchise is given to the winner of the new auction, again with a short-term contract. The bids in the new auction will necessarily reflect any changes in cost and demand. If all firms have access to the same technology and costs for inputs, and possess the same information, the winning bid will be a price that provides zero profit under least-cost production. The firm that won the first auction and hence has been producing in the industry need not be the winner of the second or later auctions. However, because the incumbent knows that it can win and continue its operation only if it bids as low a price as possible, the incumbent can be expected to bid to win.

Repeated auctions with short-term contacts can be expected to re-

sult in the lowest possible price only if the incumbent has no cost or other advantage with respect to other firms that might bid for the franchise. Otherwise the incumbent can win each auction by pricing above its own costs but below the costs of other bidders, thereby earning monopoly profits and/or engaging in waste indefinitely.

The conditions under which repeated auctions can be used to attain the lowest possible price are actually the same as those required for a contestable market, which is the topic of the next section. In fact, the theory of contestability, although it operates in a different way than repeated auctions, is actually, at its most fundamental level, a formalization and generalization of the idea that motivates repeated auctions. The power and limitations of repeated auctions are therefore best understood as an aspect of contestability.

10.3 Contestability

A contestable market is defined as one in which entry is "free" and exit is "costless," with both of these terms having a particular meaning. Free entry does *not* mean that a new firm need not incur any costs to enter an industry. Rather, free entry means that a new firm does not have to incur any costs that are not also incurred by a firm that is already producing in the industry; that is, the entrant is not at a cost disadvantage with respect to an incumbent. Free entry therefore requires that the entrant have access to the same technology and input sources as the incumbent, and that consumers perceive the entrant's product to be the same as the incumbent's.

Costless exit means that any firm can leave an industry (that is, stop producing) and recoup all the costs it incurred when entering. For example, if a firm had to purchase equipment to produce in the industry, costless exit means that the firm can sell the equipment at the same price (minus depreciation) it paid to purchase the equipment.

Under these conditions, a monopolist will be forced to produce efficiently and price so as to earn zero profit. If the incumbent monopolist earned strictly positive profit, a new firm could enter, charge a slightly lower price that results in a smaller but still positive profit, and capture the incumbent's entire market, thereby becoming the new monopolist. If the original monopolist retaliated by lowering its own price, the new firm could simply leave the industry, recouping all entry costs. In either case, price is reduced. Similarly, if the incum-

bent monopolist is earning zero profit but is engaging in some form of inefficiency in the production process, a new firm could enter with a lower price, produce without this waste, and earn positive profit. Either the original monopolist would lose its market or it would eliminate the inefficiency in its production so as to meet the new firm's price. In either case, the inefficiency is removed.

With free entry and costless exit, the monopolist is not able to earn positive profit or produce at higher than minimum costs even for a short period. Entry will occur, because a new firm can enter at a lower price and earn a profit for the period of time (however short) before the original monopolist adjusts, and then the new firm can exit, recouping its costs of entry. Baumol (1982) describes this process as follows: "The crucial feature of a contestable market is its vulnerability to hit-and-run entry. Even a very transient profit opportunity need not be neglected by a potential entrant, for he can go in, and, before prices change, collect his gains and then depart without cost, should the climate grow hostile." Actually, entry will never have to occur, because the *threat* of entry would keep the incumbent monopolist at zero profit with efficient production.[1]

As with the auctioning of the monopoly franchise, contestability in a natural monopoly situation guarantees only that profits are zero and production is cost minimizing. First-best optimality is not achieved, because price equaling marginal cost requires negative profit.[2] Second-best optimality is fully attained in a one-good situation. In a multigood situation, prices need not be Ramsey, such that this aspect of second-best is not necessarily achieved.

---

1. Costless exit is key to the argument for "hit-and-run" entry. If a potential entrant could *not* recoup all its entrance costs when exiting, it would not necessarily enter even though it could make a profit at the incumbent's current prices. The potential entrant would realize that the incumbent would probably retaliate and that it (the entrant) would lose some of the costs it incurred when entering if it were eventually forced to exit. In this case, the potential entrant would only enter if the profit it expected to earn before retaliation exceeded the unrecoupable costs of entry. With costless exit, on the other hand, a potential entrant would not fear retaliation because, even if it lost the competition for the market, it would not lose any money. A new firm would enter whenever it saw an opportunity for profit, no matter how short-lived.

2. Baumol (1982) shows that, if the least costly number of firms is more than one (e.g., a natural duopoly), price will equal marginal cost in equilibrium, such that first-best optimality is attained. An equilibrium will not necessarily exist, however, because marginal-cost pricing need not result in zero profit, and zero profit is also required for equilibrium. In a natural monopoly situation, equilibrium occurs with zero profit without marginal-cost pricing.

The reader has probably already identified a major limitation of contestability theory. As Schwartz and Reynolds (1983) point out, the power of hit-and-run entry (or, more accurately, the power of the *threat* of such entry) rests on the notion that an entrant can enter a market and earn an profit *before* the incumbent can reduce its price. In most situations, it is much easier and quicker for an existing firm to reduce its price than for a new firm to purchase necessary equipment and other production facilities, hire employees, and notify customers of its existence. In these situations, the incumbent can maintain indefinitely a price in excess of the zero-profit level. When the incumbent observes that a new firm is starting to establish operations, it lowers its price before the new firm can actually offer service. After the new firm is run out of the market, the incumbent simply raises its price again. In fact, because the potential entrant knows that the incumbent will do this, the potential entrant will not enter even though the incumbent is earning positive profit and/or producing inefficiently.

There are two ways that contestability theory can confront this argument. First, the entrant can sign long-term contracts with customers before it establishes its operations. These contracts would bind customers to buy from the entrant after the entrant established its operations and would not allow customers to switch back to the incumbent (at least until the expiration of the contract). If the entrant offers a lower price than the incumbent, customers will be willing to sign with the entrant even if they know that the incumbent will lower its price when the new firm enters. They will sign because they also know that, if they do not sign with the entrant, the incumbent will raise its prices again after the entrant is run out. Their only hope of a long-term price reduction is to sign with the entrant.

The incumbent, on observing that a potential entrant is signing up customers, could also attempt to sign up customers on terms equal to or better than those offered by the entrant. However, if the incumbent is successful in preventing customers from signing with the entrant, price is still lower in the long term, because the incumbent succeeds by offering a lower price in a long-term contract. Furthermore, because the potential entrant can start to sign up customers before it establishes its operations, it need not expend any costs on entry until it knows that it can enter profitably.

The second way that contestability theory can be maintained is essentially through mandate of the regulator. In particular, the regulator can require that the incumbent not lower its price in response to

entry. If the incumbent knows that whatever price it sets must be maintained even in the face of entry, it will choose a low price that prevents entry.[3]

10.4 Sustainable Prices in a Contestable Market

The theory of contestability and the notion of repeated auctions over time suggest that the regulator should allow entry, even in a natural monopoly situation. In fact, the theory suggests that the task of the regulator is not to oversee the incumbent's price and input decisions per se, but rather to establish policies that assure that the conditions for contestability—free entry, costless exit, and slow price response by the incumbent—are met as closely as possible.[4] When these conditions are met, the correct price and inputs will result automatically.

Allowing entry can, however, cause problems in many situations, even (in fact, especially) if the conditions for contestability are fully met. In particular, allowing entry can, depending on the cost structure of the natural monopolist, prevent equilibrium at the optimal prices. To demonstrate this possibility, and delineate the conditions under which it occurs, we introduce a new term: "sustainable prices."

The prices an incumbent firm charges are called sustainable if, under the conditions for contestability (free entry and costless exit), the incumbent earns at least zero profit and no new firm chooses to enter. "Sustainable" in this context means that the prices could be sustained over an extended period with no change in the number of firms in the industry, that is, without the existing firm leaving or new firms entering. The requirement that the incumbent earn at least zero profit reflects the fact that, if the firm loses money indefinitely, it will not be able to stay in business. Sustainability also requires, however, that potential entrants are *not* able to make a profit.

---

3. This approach requires that the regulator be able to distinguish price reductions in response to entry from those due to changes in costs or demand—a distinction that is often difficult in practice. The approach is also difficult politically, because the regulator must somehow explain to its constituency the seeming paradox that low prices are attained by preventing price reductions.

4. For example, the regulator might require that the incumbent share its technology with any new firm and purchase any equipment and facilities that a failed firm might need to dispose of when exiting. Of course, if the incumbent priced sufficiently low, no new firms would enter and none would hence fail. These obligations on the incumbent would therefore never be activated.

Several important results can be derived from the definition of sustainable prices.

Result 1: A firm's prices are sustainable only if its profits are exactly zero and it produces at least cost.

Profits must be at least zero for the firm to stay in business. However, for reasons given in section 10.3 on contestability, the prices that a firm charges must also result in no *more* than zero profit with cost-minimizing production in order to be sustainable. Otherwise, a new firm could enter, price below the original firm, capture the original firm's market, and earn a positive profit.

While zero profits and least-cost production are necessary, they are are not, in themselves, sufficient for sustainability. That is, the prices the firm charges might not be sustainable even if its profits are zero and it cost minimizes in production.

Suppose, for example, there are economies of scope in the production of two goods such that one firm can produce the two goods more cheaply than two firms. Suppose that a monopolist produces the two goods with least-cost inputs and prices them in a way that results in zero profit. It is possible, depending on the prices and the cost structure for production, that a new firm will be able to enter the market for *one* of the goods, undercut the monopolist's price for that one good, and earn a positive profit. In such a case, the monopolist is vulnerable to entry even though it produces efficiently and earns zero profit.

An example will suffice to demonstrate this possibility. Suppose that demand is fixed at 1,000 for each good. (The assumption of fixed demand is not essential; it simply makes the example more transparent.) Suppose the cost of producing the quantity demanded of both goods is \$50,000 when the two goods are produced together by one firm. Suppose further that if either of the two goods is produced by itself (that is, by a separate firm that produces only that one good), the cost of production is \$30,000. Economies of scope exist because the cost of producing the two goods separately (\$30,000 + \$30,000) exceeds the cost of producing them together (\$50,000). Suppose now that the firm prices good A at \$35 and good B at \$15. The firm earns zero profit because its revenues (\$35 · 1000 + \$15 · 1000) equal its costs (\$50,000). Note, however, that a new firm could enter, produce only good A, charge a lower price than the original monopolist, and

make a positive profit. If the new firm charges \$34 for good A, it would obtain revenues of \$34,000 (because all customers will buy from the new firm with the lower price). These revenues exceed the firm's costs of \$30,000, such that the new firm makes a positive profit. In short, even though the incumbent monopolist is earning zero profits, the prices charged by the monopolist would induce a new firm to enter.

The problem is that the monopolist's prices are not sustainable. The monopolist could (in this specific case) revise its prices so that it obtained revenues of no more than \$30,000 in each market. For example, the firm could charge \$28 for good A and \$22 for good B. These prices would still result in zero profit but would be sustainable: no firm could produce either good at a lower price and earn sufficient revenues to cover its costs.

The question arises: are there always prices that a natural monopolist can charge that will prevent entry, that is, are sustainable? If not, then allowing entry will prevent the attainment of an optimal equilibrium. The answer has been given by Baumol, Bailey, and Willig (1977), Sharkey (1981, 1982), and others. It constitutes our next result.

Result 2: It is possible that no sustainable prices exist for a natural monopolist.

Consider first a one-output situation. If economies of scale do not exist throughout the entire range of output, pricing at average cost will not be sustainable. Suppose, for example, that economies of scale exist up to 90% of market demand and then diseconomies set in. The average cost curve for this situation is depicted in figure 10.1. Diseconomies are not sufficient to warrant two firms producing in the industry: one firm is still cheaper than two. The only price that results in zero profit for the monopolist is p_m, which equals average cost for the entire market demand. However, a new firm could enter, charge a lower price (between p_m and p_e), sell 90% of market demand, and leave the remaining 10% to the original monopolist. Because the average cost of producing ninety percent of market demand is p_e, the new firm would make a profit at any price between p_e and p_m. Essentially, because the average cost of supplying a *portion* of the market is less than that of supplying the *entire* market, a new firm could earn a positive profit at a price below the average cost of the original monopolist.

In a multigood situation, the same type of problem could occur. A

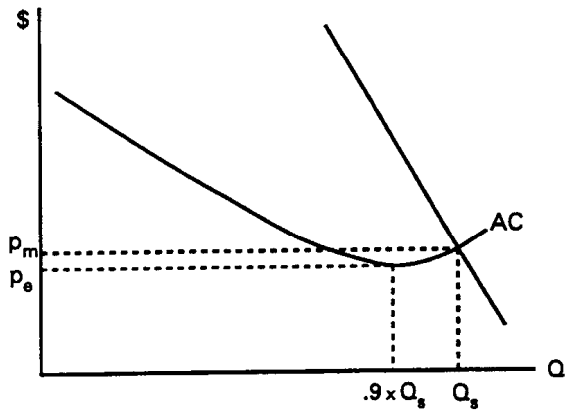

Figure 10.1
No sustainable prices for a one-good natural monopolist

convenient example is given by Zajac (1978). Suppose three services are provided and demand is fixed at 1,000 for each service. Suppose that each service could be provided by one firm at a cost of $30,000 apiece, for a total cost of $90,000 for all three services supplied by three separate firms. Suppose that economies of scope exist such that providing the services jointly is less expensive than separately. For example, suppose that any two of the services could be provided by one firm at a cost of $48,000, and all three services could be provided by one firm at a cost of $75,000. With three firms, total cost is $90,000; with two firms (one firm providing two services and the other firm providing the third), total cost is $78,000 ($48,000 + $30,000); and with one firm, total cost is $75,000. A natural monopoly exists because it is cheaper to meet market demand for these three services with only one firm than with more than one. However, there are no sustainable prices that a monopolist can charge.

To see this fact, suppose the monopolist prices each service at $25, earning $25,000 revenues on each service for a total revenue of $75,000. The monopolist's profit is zero because its costs are also $75,000. At these prices, however, a new firm could enter and provide *two* of the services at a lower price and make a profit. The new firm could charge, say, $24.50 for each of two services, which we can call services *A* and *B*. Because this price is below that charged by the monopolist, customers would buy services *A* and *B* from the new firm, providing it with $49,000 in revenue. Because the cost of providing two services is $48,000, the new firm would earn a profit of $1,000. The original monopolist would be left with only service *C*.

Consider the possibility of the monopolist lowering the price of services *A* and *B* to prevent the new firm from entering. Because the cost of providing two services is \$48,000, the monopolist must price each of the two services at \$24 to prevent the entrant from being able to make a profit on these two services. However, if the monopolist charges \$24 for services *A* and *B*, it must price the third service at \$27 to break even overall. (Revenue from services *A* and *B* is \$48,000, and revenue from service *C* is \$27,000, which just covers the firm's costs of \$75,000.) At these prices, however, a new firm could enter and provide services *A* and *C* at a profit. The new firm could charge \$23.50 for service *A* and \$26.50 for service *C*, earning revenues of \$50,000 (i.e., \$23,500 from service *A* and \$26,500 from service *C*). With costs of \$48,000, the new firm would make a profit of \$2,000.

Any other price combination that the monopolist tried would encounter the same problem: a new firm would be able to provide two of the services at a lower price and make a profit. Even though a natural monopoly exists and the incumbent monopolist is earning zero profit without waste, a new firm will be able to enter at whatever prices the incumbent charges. Clearly, in this situation, an equilibrium with one firm, which is optimal from a cost perspective, is not possible if entry is allowed.

Result 2 states that there *may* be no sustainable prices for a natural monopolist. It does not state that sustainable prices do not exist in *all* natural monopoly situations. Depending on the cost structure of the firm, sustainable prices may or may not exist. We have seen, in figure 10.1, a situation in which there is no sustainable price for a one-good natural monopolist: the average cost curve is such that an entrant can supply a portion of the market at a price that is lower than the monopolist can possibly charge for the entire market. If, on the other hand, the average cost curve is continuously downward sloping (that is, economies of scale exist throughout the entire range of output), then a sustainable price exists in a one-good situation. The situation is depicted in figure 10.2. The incumbent can price at p_m, which is the second-best optimum. Because average cost is continuously decreasing, an entrant cannot price below p_m and make a profit from selling a portion of market demand.

The important implication of result 2 is that the regulator, without knowing the cost and demand of the firm, cannot be assured that the desire for entry by another firm necessarily means that the incumbent is pricing too high. The incumbent may be operating efficiently and

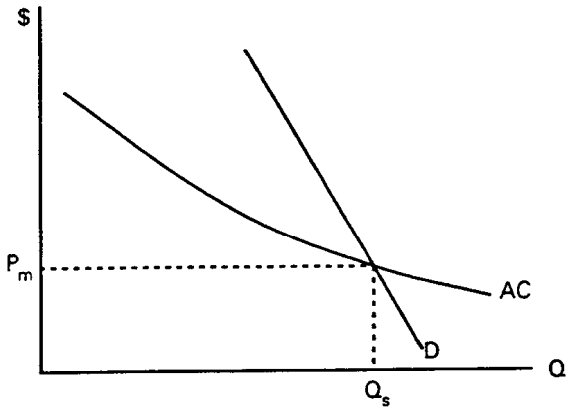

Figure 10.2
Sustainable price for one-good natural monopolist

pricing as low as possible, and yet the cost structure may be such that no sustainable prices exist. This in turn implies that a regulator, without knowing cost and demand, cannot necessarily rely on entry and the threat of entry to induce optimality.

Result 3: Ramsey prices may not be sustainable, even if sustainable prices exist for a multi-output natural monopolist.

Consider first the meaning of this result in relation to previous findings. Result 2 states that, in some situations, a natural monopolist may find that there are no prices that it can charge to prevent entry. If the regulator allows entry in these cases, then an equilibrium with one firm, which is optimal from a cost perspective, would not be possible. The question arises: What about situations in which sustainable prices *do* exist? Will allowing entry in these situations induce optimality? Result 3 implies that allowing entry could actually prevent optimal pricing by the incumbent. In particular, result 3 states that, even in situations in which sustainable prices exist, the Ramsey prices may not be sustainable. If the regulator allowed entry in these cases, equilibrium would occur with one firm, but with the firm charging some prices other than Ramsey.

Faulhaber (1975) and Sharkey (1981) demonstrate the result. An example of nonsustainable Ramsey prices is sufficient proof. Consider two goods, labeled A and B. Assume that demand for good A is fixed at 1,000, whereas demand for good B is price sensitive with demand being $Q = 1{,}280 - 10P$. If good A is produced by a firm on its own (that is, without producing good B), the firm incurs a fixed cost of

$20,000 and a marginal cost of $2 per unit. Similarly, good B can be produced on its own for a fixed cost of $30,000 and a marginal cost of $3. If the two goods are produced by one firm, certain equipment can be shared, such that economies of scope exist. In particular, the fixed cost of joint production is $40,000 as opposed to the combined fixed costs of $50,000 ($20,000 for good A and $30,000 for good B). With joint production, marginal costs are still $2 per unit of good A and $3 for good B.

Sustainable prices clearly exist for one firm that produces both goods. The monopolist could, for example, set the price of good A at $17 and the price of good B at $28. It would obtain revenues of $17,000 for good A and $28,000 for good B (i.e., $Q = 1{,}280 - 10(28) = 1{,}000$ times $P = 28$), for a total revenue of $45,000. The costs of the firm are the $40,000 fixed costs plus variable costs of $2,000 for good A and $3,000 for good B, such that total costs are $45,000. The monopolist breaks even. And, because revenue from each good is less than the fixed cost of producing that good alone,[5] no firm can produce either of the two goods at a lower price and make a profit. The prices are therefore sustainable.

The prices are *not*, however, the Ramsey prices. The demand for good A is fixed, while that for good B is price sensitive. The Ramsey rule requires, therefore, that the price of good B (the good with price-sensitive demand) be set at its marginal cost, while the price of good A (with fixed demand) be set sufficiently high for the firm to break even. That is, all fixed costs for both goods are to be loaded onto good A. The Ramsey prices are $42 for good A and $3 for good B.

At these Ramsey prices, a new firm could provide only good A, charge a lower price than the original monopolist, and make a profit. The new firm could charge, say, $40. Its revenues would be $40,000, and its costs would be $22,000 (fixed costs of $20,000 plus variable costs of $2,000), providing the new firm with a profit of $18,000. In short, while sustainable prices exist, the Ramsey prices are not sustainable. In this situation, the monopolist would not choose the Ramsey prices if the regulator allowed entry. To achieve Ramsey prices in this situation, the regulator must not allow new firms to enter the monopolist's markets.

Result 3 states that Ramsey prices *might* not be sustainable. It does not state the Ramsey prices are never sustainable. Whether Ramsey

5. Also, the elasticity of demand for each good is below one (in magnitude) in the relevant range of prices, such that decreasing price does not increase revenue.

prices are sustainable depends on costs and demand for the particular situation. Baumol, Bailey, and Willig (1977) provide conditions under which Ramsey prices are sustainable. In a one-good situation, we have already observed in figure 10.2 the conditions that are sufficient for the second-best optimum to be sustainable: when economies of scale exist for all levels of output up to market demand, the second-best price (i.e., p_m in the figure) is sustainable. For multigood situations, the conditions are more complex, involving a certain type of economies of scope in addition to economies of scale. We do not describe these conditions here: as well as being complex, it would be hard to ever determine whether the conditions are actually met in a given situation. The important point is simply that there *are* situations in which Ramsey prices are definitely sustainable.

When Ramsey prices are sustainable, the monopolist in a contestable market might charge Ramsey prices to prevent entry. Baumol, Bailey, and Willig call this tendency for Ramsey pricing a "weak invisible hand." An "invisible hand" is operating because market forces in the form of potential entry could induce a monopolist to attain second-best optimality. The invisible hand is "weak" because the firm will not *necessarily* charge the Ramsey prices. Although Ramsey prices are sustainable, other prices might also be sustainable, and the firm might choose these other prices instead.

A summary of the discussion can now be made. If certain conditions are met (such as economies of scale throughout the relevant range for a one-good monopolist), Ramsey prices are sustainable. The regulator can perhaps, in these situations, rely on the threat of entry as a way of inducing the monopolist to charge Ramsey prices. It must be remembered, however, that this inducement is weak, because the firm may charge other sustainable prices instead. If the conditions are *not* met, Ramsey prices may not be sustainable, and the regulator, by allowing entry, could be preventing the attainment of Ramsey prices. Furthermore, there is no guarantee than *any* sustainable prices exist. Allowing entry could therefore prevent the attainment of equilibrium with one firm, which is optimal from a cost perspective.

If the regulator knew the costs and demand faced by a natural monopolist, the regulator could determine whether sustainable prices exist and whether the Ramsey prices are sustainable. However, without knowing demand and cost, the regulator cannot know whether allowing entry will secure or prevent optimality. The power of potential entry, while indeed strong, needs to be used guardedly.

10.5 Market Forces versus Regulation for a Natural Monopoly

The basic question is: Should the regulator institute direct regulation or rely on market forces in a natural monopoly situation? The answer is of course less clear-cut than the question might suggest. First, as Williamson has pointed out and as we discussed in section 10.2, direct regulation and reliance on market forces, when applied over time in a changing world, are not as different as they might seem. If market forces are harnessed through an auction of the monopoly franchise with a long-term contract for the winning firm, the contract must somehow account for the fact that costs and demand change over time. The regulator will usually have less information than the firm on the way in which costs and demand have actually changed over time and cannot necessarily rely on the firm to report this information truthfully. The regulator therefore will want to establish in the contract some procedure that results in prices being adjusted optimally over time in the face of this informational asymmetry. Yet this is the task of direct regulation: the establishment of a regulatory mechanism that induces optimality when the regulator does not have sufficient information to know what prices are optimal.

The same convergence occurs if we consider the possibility of harnessing market forces through repeated auctions with short-term contracts or by allowing entry. Both of these procedures result in zero profit with cost-minimizing production under changes in costs and demand over time, provided the market is contestable at each point in time. To rely on these procedures, the regulator must assure itself that the market is indeed contestable and remains so over time. In particular, the regulator needs to know (among other things) whether the incumbent monopolist possesses a cost or demand advantage relative to potential entrants. This information allows the regulator to determine whether (or the extent to which) the incumbent can maintain prices above their optimal level while still preventing entry or winning the repeated auctions. Yet the regulator usually does not possess this information and must obtain it from the incumbent and potential entrants themselves, who cannot necessarily be relied upon to report truthfully. A mechanism is therefore needed that induces the firms to report information truthfully at each point in time, so the regulator can determine whether the repeated auctions or threat of entry are actually resulting in the lowest possible prices. Again, this is essentially the task of direct regulation.

The second point to be made about the question of direct regulation versus market forces is that the regulator would usually not want to choose one or the other of these approaches exclusively, but rather utilize some aspects of each. It is rare that a market will be sufficiently contestable for the regulator to be able to confidently rely on market forces exclusively. Some form of oversight of the incumbent's prices and costs would usually be desirable even if the market is considered to be fairly contestable. On the other hand, in markets that are clearly not contestable, the power of potential entry can still be utilized fruitfully, if only to guard against extremes. The regulator can always hold as an option—and make sure that the incumbent knows of this option—the possibility of negotiating with other producers to supply all or part of the incumbent's market. Even if potential entrants are at a cost or demand disadvantage, the threat posed by these firms still places a *limit* on the extent of waste and/or overpricing by the incumbent.

In light of these ideas, our original question regarding direct regulation versus reliance on market forces can be restated more meaningfully. As we have described, direct regulation and reliance on market forces are not so very different in practice, and the regulator will probably find it advantageous to use a mix of both approaches. The question becomes therefore: What is the appropriate combination of practices? The answer to this, as to all questions of how to regulate in the real world, is different in different settings. Even in one setting, no clear-cut answer is available. Judgment is essential. By understanding the power and limitations of market forces—knowing when the power is most effective and when limitations come into play—this judgment can perhaps be improved.