Chapter 4

Theories of Regulation

4.1 Introduction

We saw in the last chapter that in many circumstances the competitive process provides an incentive system that impels private firms to behave in ways that are broadly consistent with efficient resource allocation. But such circumstances do not always hold, and in some industries the forces of competition are inevitably weak or nonexistent. There is then a need for regulatory policy to influence private sector behavior by establishing an appropriate incentive system to guide or constrain economic decisions. This need has arisen in several major industries involved in the U.K. privatization program, where problems of monopoly power and various kinds of externalities have been central issues. In later chapters we consider in some detail the regulatory frameworks adopted in the U.K. telecommunications, gas, and airports industries, and the framework proposed for the water industry, but the purpose of the present chapter is to examine some of the underlying principles of regulatory policy.

For the most part we will leave aside externality problems in order to focus on regulation to constrain market power. To clarify the analysis further, we will begin by assuming that competition in product and capital markets is absent and cannot be stimulated. In other words, we will suppose that the regulatory system is the only constraint upon the firm's behavior apart from the fundamental conditions of demand and technology. We can then examine how the firm would behave when faced with various regulatory systems, and we can also address the broader question of optimal regulatory policy.

It is useful to regard the problem as a game between the government (or its agency) and the firm. With this perspective we need to specify the players' possible strategies, their objectives, the move order, and the information conditions of the game. As regards *possible strategies*, the firm has to make decisions about prices, outputs, capital investment, product quality, investment in cost reduction, product innovation, and so on. The government might seek to regulate some of these variables (for example prices, product quality, or profits) but, unless it is unusually well informed about industry conditions and behavior, it is unlikely to be able to regulate (as opposed to influence) other aspects of the firm's activities. This information problem is crucial because the government can condition its policy only on what it knows. Indeed the asymmetry of information between government and firm will be a central theme of this chapter.

Turning now to decision-makers' *objectives*, there are several assumptions that can be explored. The traditional approach, which offers many useful insights, is to suppose that the firm is intent upon maximizing profits and that the government seeks to maximize social welfare defined as the (possibly weighted) sum of consumer and producer surplus in partial equilibrium analysis. However, we will also wish to pursue other approaches to company behavior—especially in view of the nature of much of the debate about the effect of privatization upon internal efficiency—by assuming that managers also attach importance to nonprofit objectives, for example the minimization of managerial effort or the enhancement of sales revenues. Similarly, we will not always suppose that governments or their regulatory bodies are imbued with the classical public interest objective. Political concerns affect governments, and the interests of regulatory agencies need not coincide with social wellbeing.

As regards *move order* and dynamics, there are again several analytical perspectives. A natural starting point is to suppose that the government has "first move" by virtue of its ability to design the regulatory framework, and that the firm then behaves as best it can in response to that framework. But this simple leader–follower approach has shortcomings. One is that regulatory policies are often more short-term in nature than some aspects of company behavior, notably investment in capital assets with long lives. In such circumstances government does not begin with a clean slate; rather, it responds to conditions shaped in part by decisions of the firm. A second and related point is that the government and firm each make a series of moves over time, and they interact strategically. Thus the firm may seek to influence the design of future regulatory policy by its current actions. Such behavior would not be surprising when—as in several U.K. privatizations—regulatory policy is explicitly temporary and periodically subject to major review. The dynamic nature of the problem also raises issues of credibility (sometimes known as "time-consistency" problems in other contexts). Thus government could not credibly adopt a policy that required it to act contrary to its interests in some future circumstances.

Finally we come to *information conditions*, in particular the asymmetry of information likely to exist between the regulator and the firm. We believe that this information problem is at the heart of the economics of regulation. A fully informed regulator equipped with suitable sanctions could simply command decision makers within the firm to behave in accordance with the

first-best outcome. But in fact there are multifarious practical limitations to what the regulator can know, and hence to what outcomes he can bring about. We will pay particular attention to the case in which decision makers in the firm know more about conditions of technology and demand than the regulator. The problem for regulatory policy is one of incentive mechanism design—how to induce the firm to act in accordance with the public interest (which will depend on the state of technology and demand) without being able to observe the firm's behavior. This problem is precisely what agency theory is about, and below we will examine in detail several recent applications of that theory to the economics of regulation.

It will be clear even from these brief remarks about various objectives, strategies, dynamics, and information conditions, that regulation is a vast subject and that a full treatment of it would take us far beyond our present scope (see Kahn, 1970; Bailey, 1973; Schmalensee, 1979; Fromm, 1981; Breyer, 1982; Crew and Kleindorfer, 1986). This chapter therefore has the more limited aim of discussing a selected set of the problems and principles of regulatory policy towards dominant enterprises. The discussion will be organized under five headings:

(i) investment problems;
(ii) internal efficiency and asymmetric information;
(iii) the regulation of multiproduct firms;
(iv) collusion and capture;
(v) some relationships between competition and regulation.

4.2 Investment Problems

Investment problems pose fundamental problems for regulators in many industries. Although the direct object of regulation is often pricing policy, which is easily measured and readily changed, the effects of regulatory policy upon social welfare depends critically upon the investment behavior that it induces. Investment—whether in capacity, R&D, or whatever—is less easily quantified and typically cannot be altered in the short run because sunk costs are involved. The magnitude of the welfare effects is illustrated by investment in industries such as telecommunications (e.g. on network development and digital exchanges), gas (transmission, exploration, etc.), electricity (power stations, transmission grids, etc.), and water (pipelines, sewers, etc.).

Two general questions arise. First, do incentives exist for productive efficiency in the sense that capital investment minimizes the cost of producing the output(s) supplied? Secondly, is the scale of investment and production appropriate to the conditions of demand and technology? We address these questions by examining the regulatory theory stemming from the famous Averch and Johnson (A–J) (1962) paper on incentives for overcapitalization under rate-of-return regulation, and by analyzing a model in which price regulation cannot credibly be committed in advance of investment decisions. We discuss some dynamic issues, including regulatory lag, and we consider incentives for strategic behavior when regulator and firm interact over time. An example is the RPI $-$ X style of regulation being adopted in the U.K., under which a bound for the path of prices (or an index thereof) is fixed for a given interval of time, at the end of which there is regulatory review. As that time approaches the firm might have an incentive to engage in socially inefficient strategic behavior designed to influence the outcome of the review. We also consider how private and social rates of discount might differ, especially when there exists the possibility of the return of the private firm to the public sector at some later date.

4.2.1 Rate-of-Return Regulation: the Averch–Johnson Effect

The fundamental problem for regulators is that they lack the information to determine what the firm's pricing and other policies ought ideally to be from the point of view of economic efficiency. Rate-of-return regulation offers the solution that price(s) should be such that an allowed "fair" rate of return on capital is earned. Three questions immediately arise. What is a "fair" rate? To what measure of the capital base should the allowed rate be applied? Will the firm make decisions affecting its capital base partly with a view to influencing the price(s) it is allowed to charge, and what distortions will result?

In their classic model of rate-of-return regulation Averch and Johnson (1962) provide an affirmative answer to this last question. Firms have an incentive to expand their capital base so as to achieve a greater absolute profit while staying within the constraint on their profit rate. An excellent review of early contributions to the debate stimulated by Averch and Johnson is provided by Baumol and Klevorick (1970), and in this section we rely heavily on their discussion.

The A–J model concerns a monopoly supplier of a single good produced with two inputs, labor L and capital K, according to production function $Q = F(L,K)$. Inverse demand is $P(Q)$, and $R(L,K) = F(L,K)\,P(F(L,K))$ in the revenue when the input levels are L and K. Labor and capital are available at factor prices w and r respectively, and profit is therefore

$$\pi(L,K) = R(L,K) - wL - rK. \tag{4.1}$$

The allowed rate of return is denoted by s, which is assumed to exceed r. (Otherwise the firm would wish to close down, at any rate in the long run.) Thus the constraint on the firm's behavior is that

$$[R(L,K) - wL]/K \leqslant s. \tag{4.2}$$

This constraint is assumed to bind—that is, s is not so generous that (4.2) is satisfied by pure monopoly behavior. The firm's problem is to maximize (4.1) subject to (4.2), and we form the Lagrangean

$$\begin{aligned} H(L,K,\lambda) &= \pi(L,K) - \lambda[R(L,K) - wL - sK] \\ &= (1-\lambda)[R(L,K) - wL] - (r - \lambda s)K. \end{aligned} \tag{4.3}$$

From the first-order conditions it follows that $\partial R/\partial L = w$, but that $\partial R/\partial K = r - \lambda(s-r)/(1-\lambda)$, which is less than r. (The second-order condition guarantees that $0 < \lambda < 1$.) Therefore excess capital is employed, and the firm produces its output in a manner that is too capital intensive and hence inefficient. The firm has no direct benefit from cost inefficiency, but it achieves a strategic gain by influencing the permitted price.

A diagrammatic method due to Zajac (1970) usefully illustrates this and related results. Figure 4.1 is a three-dimensional depiction of π as a function of L and K. The plane hinged on the L axis is the set of points such that $\pi(L,K) = (s-r)K$. The points on or beneath the plane are precisely those that meet constraint (4.2). Thus the firm's problem is to be as high as possible on the shaded "profit hill" without being above the "regulatory plane."

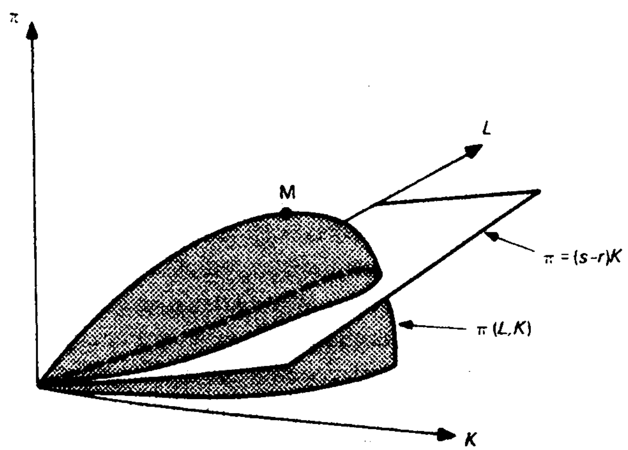

Figure 4.1 Rate-of-return regulation

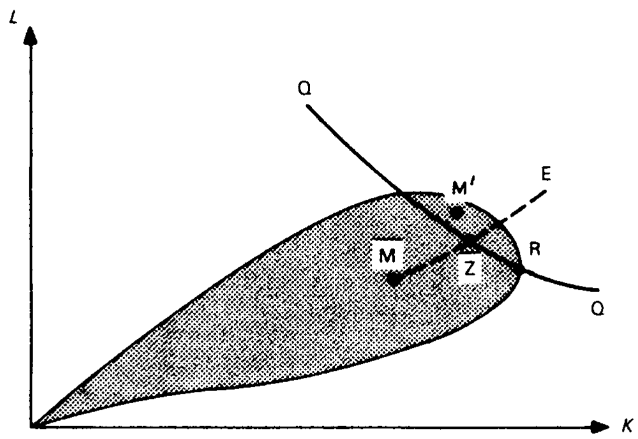

Figure 4.2 The Averch–Johnson effect

The shaded area in figure 4.2 contains the feasible (L,K) combinations that satisfy the regulatory constraint. Since profit is proportional to capital where the constraint binds ($\pi = (s-r)K$), the profit of the regulated firm is greatest at point R, the rightmost point in the shaded area. Curve QQ is the isoquant passing through R. Note that it passes through the interior of the shaded area, which shows that, given output Q, profit and welfare could be higher. The cost of producing output Q is not minimized, because the capital stock is deliberately expanded by the firm. The efficient way to produce Q is at point Z where the "efficiency locus" ME intersects the isoquant. At Z the K/L ratio is lower than at R. Thus the output of the regulated firm is produced in a manner that is too capital intensive. An unregulated monopolist would operate at point M, where efficient production occurs; the unregulated firm has every incentive to minimize production costs.

To summarize, the effect on welfare of rate-of-return regulation in this model has two parts. The level of output is affected, and so too is the efficiency with which the output is produced. If—as is usually the case—regulation increases output, the two effects work in opposite directions, and there is a conflict between internal and allocative efficiency which will appear in several contexts in this chapter.

Without further assumptions on cost and demand conditions it is not necessarily the case that regulation has the effect of increasing output. If, for example, profit were maximized at point M′, output would be higher without regulation. Nor is it necessarily true that the K/L ratio under regulation is greater than that without regulation.

A final observation on the basic A–J model, which Baumol and Klevorick (1970, pp. 175–176) emphasize, is that the amount of capital employed by the regulated firm increases as s is set closer to the cost of capital r. A reduction in s would expand the shaded area in figure 4.2, and so the optimal K would rise. The cost inefficiency due to incentives to overcapitalize may well grow as s is set closer to r. Klevorick (1971) considers the *optimal* choice of s from the point of view of social welfare. Although intuition might suggest that $s = r$ was optimal, this is not in fact so. At $s = r$ the profit-maximizing firm is indifferent between all feasible input combinations that meet the regulatory constraint, because they all involve zero profit. But even if the firm has social welfare as a secondary objective (in the sense of lexicographic preferences) there is a wide range of cases in which some $s > r$ induces a superior outcome than $s = r$. Also, if the firm's secondary objective is to maximize K (e.g. because of managerial satisfaction), then it is generally true that $s > r$ is superior. The reason has already been indicated: s close to r can cause more productive inefficiency.

An extension of the A–J model which remains within its essentially static framework is to replace profit maximization by some other objective for the firm. For example, Bailey and Malone (1970) argue that, under a wide range of conditions, a firm maximizing sales revenue subject to rate-of-return regulation would produce its output in a way that was inefficient by being too labor intensive. This contrasts with Averch and Johnson's finding. However, Atkinson and Waverman (1973) contend that the sales-maximizing firm faces a minimum profit constraint as well as the regulatory restraint, and that various outcomes are possible depending on the interaction of the constraints and the basic conditions of demand. At any rate, this work illustrates that results can be sensitive to assumptions regarding the motivation of the firm.

4.2.2 Regulatory Lag

The A–J model provides a useful starting point, but it can be criticized for being too static in its formulation. Regulation does not occur in a continuous fashion. Typically prices are set for an interval of time, during which the firm is free to choose whatever input combinations it wishes, until the next price review occurs. Review might occur at some time specified in advance—for example the formula governing the pricing of British Telecom's telecommunications services in the U.K. will be reviewed in 1989—or its timing might be uncertain. In the latter case an important distinction must be made between exogenous and endogenous uncertainty. With endogenous uncertainty, the timing of the next review depends partly upon how the firm behaves in the meantime.

Bailey and Coleman (1971) extend the A–J model by supposing that regulators set prices after an interval of T periods. The firm, making its decisions at time zero, faces a trade-off between maximizing profits by producing more efficiently during the next T periods and overcapitalizing to induce a more favorable price when review eventually occurs. The balance is struck where

$$\frac{F_K}{F_L} = \frac{r}{w} - \frac{\rho^T}{1-\rho^T}\,\frac{s-r}{w},$$

where ρ is the discount factor. It follows that it is optimal for the firm to overcapitalize to some extent (depending inversely upon T), but not as much as in the basic A–J model. A similar finding is obtained by Davis (1973) for a model in which price adjustment occurs continuously but only partially.

Baumol and Klevorick (1970, pp. 184–188) criticize the approach of Bailey and Coleman, and propose a model of regulatory lag which is of particular relevance to the style of regulation adopted in the major U.K. privatizations. They write (p. 184):

> "While Bailey and Coleman regard the period before a regulatory review as a time when the firm suffers a loss because it is carrying an excessive amount of capital, now the period between reviews is regarded as the time when the firm has the possibility of earning a profit rate exceeding that specified by the constraint. When the regulatory review occurs, this excess is eliminated by the regulators' adjustment of the prices the firm can charge."

In our view this point has great force. Regulatory lag allows the firm to appropriate the benefits of improved cost efficiency until the next review occurs. A longer lag increases the firm's incentives to reduce its costs by innovation or superior organization of factors of production, but it delays the time at which consumers benefit from this greater efficiency. On the other hand, a shorter lag means that consumers benefit sooner, but the incentive to cut costs is reduced. This trade-off between static and dynamic efficiency has a close analogy in the literature on optimal patent life, and indeed Bailey (1974) analyzes the problem of innovation and regulatory lag in exactly that spirit (see also the debate between Lesourne (1976) and Bailey (1976)).

There is, however, a further point to consider. In the framework proposed by Baumol and Klevorick, price is brought into line with current costs at the time of each regulatory review. The RPI $-$ X style of regulation implemented in Britain is likely to fit this description. Although such a system provides good incentives for efficiency immediately after a review

point, as time passes the firm's calculations will be increasingly affected by the benefit to be gained from influencing the outcome of the next regulatory review. As that time approaches, the firm will have little or no incentive to reduce costs if its future prices are positively related to its current cost level. Indeed, a point would then arise when the immediate gain from cost reduction was so short-lived as to be outweighed by the cost of having to face lower prices for the whole of the period until the following price review. In technical language, the second-order effect would be outweighed by the first-order effect, and the firm would come to favor *higher* costs when regulatory review is close at hand. We shall consider incentives for this kind of strategic behavior further in section 4.4 (see also Sappington, 1980).

These considerations suggest three lessons. First, the incentive effects of regulatory lag are not necessarily always benign. Strategic behavior designed to influence regulatory review could involve substantial losses in terms of allocative and productive efficiency, which would be offset against the initial spur to innovation provided by regulatory lag. Secondly, the potential losses from strategic behavior are reduced when regulatory review is less sensitive to current cost conditions. This points to the importance of the information available to regulators, especially information that is independent of the firm's decisions. We shall return later to this theme of the dangers of the firm's having a "monopoly of information." Thirdly, the timing of regulatory reviews is important—not only in terms of the length of regulatory lag, but also whether regulatory review occurs at regular intervals or stochastically.

We conclude this section by describing two models of stochastic regulatory review. (A discussion of further dynamic analysis is contained in section 4.4.) Klevorick (1973) examines a model in which for every period there is a given probability $\phi \in [0,1]$ of regulatory review. When review occurs, price is set so as to restore the "fair" rate of return s on the current capital stock. If ϕ were equal to 1, we would effectively have the A–J model, albeit in an explicitly dynamic setting, and if ϕ were equal to zero, regulatory lag would be infinite and the firm would have perfect incentives for productive efficiency. The intermediate case leads to overcapitalization, although not to the extent of that occurring in the A–J model.

Bawa and Sibley's (1980) more general model of *endogenous* stochastic regulatory review is more satisfactory. The probability of review in any period is a function $\phi(X)$ of current profit in excess (or deficit) of the level allowed by the rate of return s. Thus X_t is defined as

$$X_t = \pi_t - (s-r)K_t.$$

It is assumed that $\phi(0) = \phi'(0) = 0$, $\phi'(X) > 0$ for $X > 0$, and $\phi'(X) < 0$ for $X < 0$. If review does occur in period t, price is set at the level that yields rate of return s on the capital stock K_t until the next review takes place.

The model captures the idea that the firm has to balance its desire for short-run profits against the risk of jeopardizing future profits by triggering a review of its prices. Bawa and Sibley use techniques of stochastic dynamic programming to establish the following:

(i) the firm will overcapitalize, be efficient, or undercapitalize according to whether the allowed rate of return s exceeds, equals, or is less than the cost of capital r;

(ii) there is continuity in the sense that s close to r leads to approximate efficiency;

(iii) under fairly general conditions there is convergence to the price at which $X = 0$ and to cost minimization.

As well as having a more realistic formulation of the regulatory process, Bawa and Sibley's model yields more intuitive results than the basic A–J model. For example, $s = r$ leads to efficient production and $s < r$ involves undercapitalization, whereas in the A–J model we saw that $s = r$ has an outcome that is indeterminate (and, in terms of capital bias, undesirable) and $s < r$ leads to the shutdown of production. In the richer dynamic setting we therefore escape the welfare trade-off examined by Klevorick (1971) between allocative and productive efficiency as s approaches r: the conflict disappears.

4.2.3 Credibility, Commitment, and Underinvestment

So far we have paid little attention to one of the main features of much capital investment—the presence of sunk costs and adjustment costs. In the A–J model it is as though there exists a rental market for capital equipment that is a freely variable factor of production. But in fact there are typically major adjustment costs when the scale or nature of a firm's operations are changed, and capital costs are often sunk in the sense that the assets have significantly less value in their next alternative use. Much the same is true of certain types of labor when hiring, training, and firing costs are taken into account.

In contrast, variables such as price—the prime instrument of regulatory policy—are usually easier to alter. The resulting asymmetry of adjustment costs can have serious implications for regulatory policy, which we shall illustrate by way of two examples.

The first of these is a "dynamic consistency" problem (see Greenwald, 1984). Suppose for simplicity that the regulation game has three stages: (i) the regulator announces the price that the firm will be allowed to charge, (ii) the firm makes its investment decisions, which involve a large element of sunk costs, and (iii) the regulator reviews the previously announced pricing policy.

At stage (iii) a regulator seeking to maximize consumer benefits would wish to impose the lowest possible prices subject to encouraging the firm to produce (i.e. subject to covering the variable costs of production). Similarly, a regulator intent upon maximizing the sum of consumer and producer surplus would set $P = MC$, which in many regulated industries might imply that price is below (long-run) average cost. In sum there is a range of regulator's objectives for which the firm would be wary of committing large investment expenditures at stage (ii) for fear of what might happen at stage (iii), and the announcement of the price at stage (i) would then lack credibility. (The risk of renationalization on less than fair terms is a related problem, which we consider separately below.)

This credibility problem, which arises from the public interest mandate of the regulator, has the effect of undermining the public interest insofar as it inhibits investment at stage (ii), for example by increasing risk-adjusted private discount rates. The solution advocated by Greenwald (1984, p. 86) is as follows:

> "Restricting regulators with an appropriate 'fairness' criterion may, therefore, be essential to the viability of the originally optimal equilibrium. The simplest way to do this would be to require by law that past regulatory promises must be honored in future proceedings. To maintain the flexibility of regulators to respond to unforeseen circumstances, however, the set of legally binding past promises should be minimally constraining. Since investors should be concerned only with future returns, the minimum acceptable set of legal constraints need only guarantee the value of future income implied by past promises."

Greenwald argues that in the United States "properly interpreted, the present structure of rate return regulation corresponds exactly to such a system." The credibility of a commitment to fairness is no doubt enhanced by wider share ownership, because the constituency opposed to "unfairness" is larger and more vocal. But this does not solve the problem completely, because whenever the credibility of the fairness constraint is below 100 percent, there is a risk factor that managers of a profit-maximizing firm would wish to take into account.

We now turn to our second illustration. Returning to the simple schema above, it is clear that the regulator at stage (iii) will be influenced by the investment decisions made by the firm at stage (ii). Thus the firm has an opportunity to influence the regulatory regime that it faces. For simplicity, suppose that expansion of the firm's capital stock is prohibitively costly in the relevant timescale. Assume that the regulator seeks to maximize the sum of consumer and producer surplus, and that there are no externalities. Then the regulator will set P = MC given the cost curve resulting from the firm's prior investment decision.

More formally, let inverse demand be $P(Q)$ and let the cost function be $C(Q,K)$ where Q is output and K is the capital stock. It is reasonable to suppose that marginal cost C_Q is positive, increasing in Q for given K, and decreasing in K for given Q. Thus $C_Q > 0$, $C_{QQ} > 0$, and $C_{QK} < 0$. Also, we assume $C_{KK} > 0$. We are interested in at least two questions. Does the firm choose to operate on a scale that is suboptimally small? Does the firm produce its output in an efficient manner?

The regulator is assumed to impose marginal cost pricing. Thus

$$P(Q) = C_Q(Q,K). \tag{4.4}$$

The firm chooses K and Q to maximize profit $P(Q)Q - C(Q,K)$ subject to (4.4). The Lagrangean is

$$H(Q,K) = P(Q)\,Q - C(Q,K) + \mu[P(Q) - C_Q(Q,K)] \tag{4.5}$$

and the first-order conditions are

$$P + P'Q - C_Q + \mu(P' - C_{QQ}) = 0 \tag{4.6}$$

and

$$-C_K - \mu C_{QK} = 0. \tag{4.7}$$

Equations (4.4), (4.6), and (4.7) imply that

$$C_K = \frac{C_{QK}\,QP'}{P' - C_{QQ}}, \tag{4.8}$$

which is negative. It follows that the regulated firm in this context produces its output inefficiently, and with a capital-to-output ratio (and hence a capital-to-labor ratio) that is too *low*. The firm holds back its capital stock in order to induce a more profitable price from the regulator. The result is undercapitalization. Moreover, the regulated firm produces less output than at the first best.

Figure 4.3 illustrates both points. Isoprofit curves for the firm in (Q,K) space are centered on point F, and isowelfare curves for the regulator are centered on point W. That point is northeast of F, representing the firm's interest in restricting output to the monopoly level. The efficiency locus defined by $C_K = 0$ slopes up and passes through both F and W. The

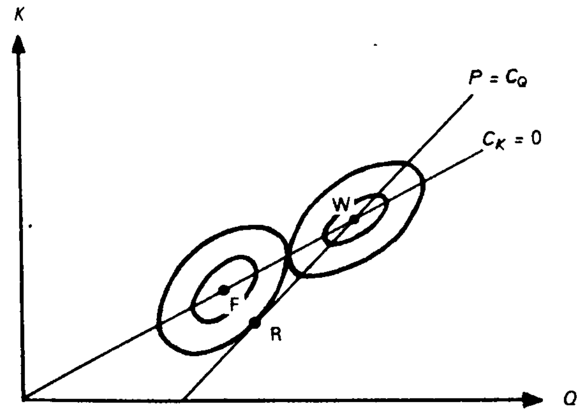

Figure 4.3 Strategic underinvestment

P = MC locus also has an upward slope; it cuts isowelfare curves where they have a zero slope. The regulated firm operates at point R. Figure 4.3 shows how output is restricted below its optimal level and produced in an undercapitalized manner.

The diagram also suggests—and indeed it can be proved—that the welfare comparison between F and R is ambiguous. The pure monopoly point F is one of allocative inefficiency because $P >$ MC, but the monopoly output is produced at minimum cost. The regulated firm does not produce its output at minimum cost, but R is a point where there is allocative efficiency because P = MC. Once again we see the conflict between internal (productive) efficiency and allocative efficiency.

Both models outlined in this section are rudimentary, but their purpose was simply to illustrate two dangers of underinvestment when considerations of commitment and dynamics are taken into account. In one case investment is inhibited by the fear of "unfair" future regulation. Unless the regulatory and/or political systems provide a credible means of commitment to future fairness, this fear is not entirely unreasonable in view of the objectives of regulators. In the second example, underinvestment was a strategic move by the firm seeking to obtain a more favorable regulatory regime. We will develop this point further, but now we address explicitly a major issue that has so far only been in the background of our discussion—the role of asymmetric information.

4.3 Regulation with Asymmetric Information

If a regulatory agency had as much knowledge about industry conditions and behavior as the firm being regulated, it could simply direct the firm to implement its chosen plan, provided that the agency possessed sufficient powers to do so. Indeed, it would then be better simply to appoint the regulatory agency to run the enterprise rather than to leave decision-making authority with the managers of the firm. But of course decision makers within the firm are generally far more knowledgeable than regulators can be about the circumstances facing them, and the regulator can neither observe nor infer all aspects of the firm's behavior. Thus asymmetric information is one of the main features of the economics of regulation (of both public and private enterprise), and in this section we examine some recent contributions to the literature, notably those of Baron and Myerson (1982) and Laffont and Tirole (1986). An excellent survey of this topic is provided by Caillaud *et al.* (1985). Owing to its more technical nature, some of our discussion is contained in starred sections, which some readers might prefer to omit.

4.3.1 Principal–Agent Theory and Regulation

Principal–agent theory was introduced in chapter 2 as a way of examining the relationship between the (public or private) owners of a firm and its managers. The theory is concerned with the design of incentives for efficiency under conditions of asymmetric information. The principal (i.e. the owners in chapter 2) is less informed than the agent (i.e. the managers) about the conditions facing the firm, and may be unable to monitor the agent's behavior with precision. Asymmetry of information gives rise to imperfect incentives, and inefficiency is the result.

Principal–agent theory can be used in exactly the same fashion to study regulation. In this context the government or the regulatory authority is the principal, and (the management of) the firm is the agent. With this perspective, a system of regulation can be regarded as an *incentive mechanism*. The firm is better informed than the regulator about cost conditions for example, and the regulator seeks to induce the firm to make its pricing, output, and investment decisions in accordance with the public interest given the cost conditions that exist. But the firm is interested in maximizing (say) its profits and, whatever the scheme of regulation may be, it will act in its own interests.

Suppose for example that the government's objective is social welfare W defined as the sum of consumer surplus S and the firm's profit π. Let the firm's objective be to maximize profit. (The firm is taken to be risk-neutral.) Let θ be the unit cost level of the firm, which the regulator cannot observe, and let Q be the firm's level of output. To begin with let us assume that Q is

observable by the regulator. He would like Q to be chosen so that price equals marginal cost, but he does not know θ. However, if he can make lump-sum transfers to (or from) the firm, it is in fact possible to bring about the optimal outcome in this example (see Loeb and Magat, 1979). The regulator should undertake to pay the firm an amount equal to the consumer surplus from output Q minus a constant sum. With this incentive scheme the firm receives its profit (i.e. producer surplus) plus the payment equal to consumer surplus minus the constant. It maximizes this objective if and only if it maximizes the regulator's objective. In this special case it is therefore possible to engineer the first-best outcome by an appropriate incentive scheme, even without knowing cost conditions.

Although this decentralized scheme might be neat in theory, it obviously has several overwhelming practical drawbacks (see Sharkey, 1979). First, although the scheme does not require cost information, the government needs to know the magnitude of consumer surplus—a much more demanding task. Although prices provide some information about *marginal* utility, they do not say much about the whole area under the demand curve (let alone income effects etc.). Secondly, the scheme runs the risk of bankrupting the firm when costs are high and consumer surplus is correspondingly low, unless the fixed element of the incentive scheme is so generous as to cushion the firm against any eventuality—in which case the firm will make huge profits in more favorable states of the world. Thirdly, the scheme depends on the government being indifferent to transfers between consumers or taxpayers and the firm, i.e. on its objective being $W = S + \pi$.

However, the government's objective cannot generally be represented simply as the sum of consumer surplus and profit (see section 2.3.2 above, and Caillaud *et al.* (1985, pp. 4–7)). A concern for distribution might cause less weight to be attached to profit than to consumer interests. In that case, an objective of the form $W = S + \alpha\pi$ would be appropriate, with $0 \leqslant \alpha < 1$. Indeed, we shall use that specification of government objectives in the following two sections. Secondly, costs to the economy of raising public funds can be represented by attaching a negative weight to lump-sum transfers to the firm. Thirdly, the interests of employees could be taken into account, although this is less likely to be an important factor for independent regulatory authorities than for politicians.

The next two sections describe versions of the models of regulation under asymmetric information examined by Baron and Myerson (1982) and Laffont and Tirole (1986). Our discussion is far from being rigorous, but it is somewhat more technical than usual. Some readers might prefer to go directly to section 4.3.4, which summarizes the main findings of the analysis.

4.3.2* Regulation with Unknown Costs

In this section we present a simplified version of Baron and Myerson's (1982) model of regulation with unknown costs. The model is one of asymmetric information. The firm knows its unit cost level, denoted θ, but the regulator does not. We assume that θ is distributed uniformly on an interval $[\underline{\theta}, \bar{\theta}]$.

Consumer utility from output level Q is denoted $V(Q)$, and $P(Q) = V'(Q)$ is the inverse demand curve. Let $R(Q) = QP(Q)$ be the firm's revenue, and let T be the transfer (possibly negative) paid to the firm. Net consumer surplus is therefore $S = V - R - T$, and profit is $\pi = R - \theta Q + T$. The regulator's objective is taken to be $W = S + \alpha\pi$, where $0 \leqslant \alpha \leqslant 1$. We saw above that $\alpha < 1$ can be interpreted as reflecting a concern for distribution.

A regulatory mechanism will induce, for each value of θ, an associated level of output $Q(\theta)$ and transfer $T(\theta)$. The *revelation principle* (see Myerson, 1979; Dasgupta *et al.*, 1979) implies that, without loss of generality, we can consider the regulator's optimization problem as equivalent to the following. The regulator requires the firm to provide a report $\hat{\theta}$ of its cost level, and determines the output $Q(\hat{\theta})$ and the transfer $T(\hat{\theta})$ as a function of that report. The firm must have no incentive to report its cost level untruthfully given that Q and T are determined in that manner.

This truth-telling constraint involves no loss of generality, because if the firm found it optimal to lie by reporting $\hat{\theta}(\theta)$ when the truth was θ, the regulator could simply amend the mechanism to be $\bar{Q}(\theta) = Q[\hat{\theta}(\theta)]$ and $\bar{T}(\theta) = T[\hat{\theta}(\theta)]$, and the firm would then find it optimal to report the truth.

The revelation principle therefore allows us to consider the regulator's problem as one of choosing $Q(\theta)$ and $T(\theta)$ to maximize the expected value of W subject to (i) the firm's finding it optimal to report θ truthfully, and (ii) the firm's always being willing to operate—i.e. receiving nonnegative profits in all states of the world.

More formally, let

$$S(\theta) = V[Q(\theta)] - R[Q(\theta)] - T(\theta)$$

be the net consumer surplus in state θ, and let

$$\pi(\hat{\theta},\theta) = R[Q(\hat{\theta})] - \theta Q(\hat{\theta}) + T(\hat{\theta})$$

be the profit in state θ when $\hat{\theta}$ is reported. Define $\pi(\theta) = \pi(\theta,\theta)$. Then we can state the regulator's problem as follows.

Choose $Q(\theta)$ and $T(\theta)$ to maximize

$$\text{EW} = \int_{\underline{\theta}}^{\bar{\theta}} [S(\theta) + \alpha\pi(\theta)]\,\mathrm{d}\theta \tag{4.9}$$

subject to

$$\pi(\theta) \geqslant \pi(\hat{\theta},\theta) \text{ for all } \theta \text{ and } \hat{\theta} \tag{4.10}$$

and

$$\pi(\theta) \geqslant 0 \text{ for all } \theta. \tag{4.11}$$

Conditions (4.10) and (4.11) correspond to (i) and (ii) above. We shall assume that (4.10) is characterized by the first-order condition

$$R'Q' - \theta Q' + T' = 0. \tag{4.12}$$

The constraint (4.11) is binding only at $\bar{\theta}$, because for $\theta < \bar{\theta}$ we have

$$\pi(\theta) \geqslant \pi(\bar{\theta},\theta) > \pi(\bar{\theta}). \tag{4.13}$$

The rent accruing to the firm from its monopoly of information derives from this fact.

The Lagrangean associated with the regulator's problem is

$$\begin{aligned} H &= \int [S + \alpha\pi + \mu(R'Q' - \theta Q' + T')]\,\mathrm{d}\theta \\ &= \int [V - (1-\alpha)(R+T) - \alpha\theta Q + \mu(R'Q' - \theta Q' + T')]\,\mathrm{d}\theta. \end{aligned} \tag{4.14}$$

For notational simplicity in (4.14) we suppress the dependence of Q, T, and the multiplier μ upon θ, and the range of integration $[\underline{\theta}, \bar{\theta}]$. Let I be the integrand [.]. Then the Euler optimization conditions are

$$\frac{\partial I}{\partial X} = \frac{\mathrm{d}}{\mathrm{d}\theta}\frac{\partial I}{\partial X'} \quad \text{for } X = Q, T \tag{4.15}$$

The condition with respect to Q is

$$V' - (1-\alpha)R' - \alpha\theta + \mu R''Q' = \mu(R''Q' - 1) + \mu'(R' - \theta) \tag{4.16}$$

and the condition with respect to T is

$$-(1-\alpha) = \mu'. \tag{4.17}$$

Since we have a free-boundary problem we can choose $\mu(\underline{\theta}) = 0$, and (4.17) therefore implies that

$$\mu(\theta) = -(1-\alpha)(\theta-\underline{\theta}). \quad (4.18)$$

Using (4.16) to (4.18), and recalling that $V' = P$, we have the central result that

$$P[Q(\theta)] = \theta + (1-\alpha)(\theta-\underline{\theta}). \quad (4.19)$$

Under the optimal regulatory mechanism, price is equal to unit (and marginal) cost *plus* a mark-up depending on α and $(\theta - \underline{\theta})$. Note that optimality always involves marginal cost pricing when $\alpha = 1$, in keeping with the Loeb–Magat mechanism described above. But in general, when $\alpha < 1$, there is a loss of allocative efficiency because price exceeds marginal cost, except in the best state of the world $\underline{\theta}$.

It is optimal for the regulator to forego allocative efficiency to some extent because he is also concerned to minimize the size of the transfer T, which has a net cost of $(1-\alpha)T$. He could induce marginal cost pricing, but only at the expense of a greater expected transfer to the firm. Optimality requires that a balance be struck between allocative efficiency and the minimization of the transfer.

If the regulator were as well informed as the firm—i.e. if he could observe θ—his problem would be to maximize (4.9) subject only to (4.11). The solution to this problem is $P[Q(\theta)] = \theta$ and $T = 0$ for all θ. In that event there is always allocative efficiency, and the firm always exactly breaks even. Therefore the partial loss of allocative efficiency is not the only reason why the regulator is adversely affected by the presence of asymmetric information. He also loses from the fact that the firm obtains a strictly positive payoff (in all but one state of the world). The asymmetry of information therefore causes two kinds of inefficiency to the detriment of consumers and the regulator's objective. However, the firm gains from the regulator's imperfect information because it obtains *money rent* in the form of transfers more than sufficient to meet its break-even constraint.

4.3.3* Regulation with Unobservable Effort

We now present a model based (somewhat loosely) on the work of Laffont and Tirole (1986) which adds another dimension to the regulatory problem. In the Baron–Myerson model above it was assumed that the level of costs was given to the firm but that the regulator could not observe it. In contrast, we now suppose that costs are influenced by the firm's cost-reducing *effort*, and that the regulator can observe the cost level. However, costs are determined jointly by two factors—the state of nature and the firm's effort—neither of which is observable by the regulator. He therefore cannot tell whether (say) low costs are due to great efforts by the firm or to a favorable state of nature.

More specifically, let unit costs c depend upon the state of nature θ and the level of effort a as follows:

$$c = \theta - a.$$

The cost of effort is denoted $z(a)$, where $z(0) = z'(0) = 0$, $z'(a) > 0$ for $a > 0$, and $z''(a) > 0$. As before, θ is taken to be distributed uniformly on $[\underline{\theta}, \bar{\theta}]$. The notation for consumer utility, price, output, revenue, and the transfer is also as in the previous section. The regulator is again assumed to be concerned with welfare defined as $W = S + \alpha\pi$ with $0 \leqslant \alpha \leqslant 1$.

As before, a regulatory mechanism will induce an output level $Q(\theta)$, a cost level $c(\theta)$, and a transfer $T(\theta)$ for each value of θ. Invoking the revelation principle, we can consider the regulator's problem as choosing the three functions $Q(\theta)$, $c(\theta)$, and $T(\theta)$ to maximize expected welfare subject to (i) the firm's finding it optimal to report θ truthfully, and (ii) the firm's always being willing to operate in the sense of achieving nonnegative profits. It might be thought more natural to view the problem as one of choosing Q and T as functions of observed c, but the revelation principle is more convenient analytically and anyway accommodates the point. For if $\bar{Q}(c)$ and $\bar{T}(c)$ were an optimal regulatory scheme, and if $\bar{c}(\theta)$ was optimal for the firm facing that scheme, then by defining $Q(\theta) = \bar{Q}[\bar{c}(\theta)]$ and $T(\theta) = \bar{T}[\bar{c}(\theta)]$ we would have an optimal scheme satisfying (ii) and expressed in a more convenient form. Therefore no generality is lost.

The net consumer surplus in state θ is

$$S(\theta) = V[Q(\theta)] - R[Q(\theta)] - T(\theta),$$

and the profit in state θ when $\hat{\theta}$ is reported is

$$\pi(\hat{\theta}, \theta) = R[Q(\hat{\theta})] - c(\hat{\theta})Q(\hat{\theta}) - z[\theta - c(\hat{\theta})] + T(\hat{\theta}).$$

Define $\pi(\theta) = \pi(\theta, \theta)$. Then the regulator's problem is as follows.

Choose $Q(\theta)$, $c(\theta)$, and $T(\theta)$ to maximize

$$\mathrm{EW} = \int_{\underline{\theta}}^{\bar{\theta}} [S(\theta) + \alpha\pi(\theta)]\,\mathrm{d}\theta \quad (4.20)$$

subject to

$$\pi(\theta) \geqslant \pi(\hat{\theta}, \theta) \text{ for all } \theta \text{ and } \hat{\theta} \quad (4.21)$$

and

$$\pi(\theta) \geqslant 0 \text{ for all } \theta. \quad (4.22)$$

We shall assume that (4.21) is characterized by

$$R'Q' - c'Q - cQ' + z'c' + T' = 0. \tag{4.23}$$

Condition (4.21) binds only at $\bar{\theta}$, and $\pi(\theta) > 0$ for all $\theta \neq \bar{\theta}$ (see (4.12) above).

The Lagrangean associated with the regulator's problem is

$$\begin{aligned} H &= \int [S + \alpha\pi + \lambda(R'Q' - c'Q - cQ' + z'c' + T')]\,\mathrm{d}\theta \\ &= \int [V - (1-\alpha)(R+T) - \alpha c Q - \alpha z + \\ &\qquad \lambda(R'Q' - c'Q - cQ' + z'c' + T')]\,\mathrm{d}\theta. \end{aligned} \tag{4.24}$$

For notational convenience we suppress the functional dependence of Q, c, T, and λ upon θ, and the range of integration $[\underline{\theta}, \bar{\theta}]$.

The Euler conditions with respect to Q, c, and T respectively are given in the following three equations:

$$V' - (1-\alpha)R' - \alpha c + \lambda(R''Q' - c') = \lambda(R''Q' - c') + \lambda'(R' - c) \tag{4.25}$$

$$-\alpha Q + \alpha z' + \lambda(-Q' - z''c') = \lambda[-Q' + z''(1-c')] + \lambda'(-Q + z') \tag{4.26}$$

$$-(1-\alpha) = \lambda'. \tag{4.27}$$

Since this is a free-boundary problem we can choose $\lambda(\underline{\theta}) = 0$, and (4.27) therefore implies

$$\lambda(\theta) = -(1-\alpha)(\theta - \underline{\theta}). \tag{4.28}$$

Equations (4.25) to (4.28) now imply the two central equations

$$P[Q(\theta)] = c(\theta) \tag{4.29}$$

and

$$\begin{aligned} z'[\theta - c(\theta)] &= Q(\theta) - (1-\alpha)(\theta - \underline{\theta})z''[\theta - c(\theta)] \\ &< Q(\theta) \text{ except at } \underline{\theta}. \end{aligned} \tag{4.30}$$

At the first best, where the regulator can observe effort, we have $P = c$ and $z' = Q$ for all θ. Equation (4.29) states that the optimum with asymmetric information has price equal to marginal cost, which is allocatively efficient *given* the level of costs. But (4.30) implies that cost-reducing effort is generally *less* than that required at the first best. Therefore costs are too high, and so price is higher than at the first best.

The firm in this example also enjoys some rent from its monopoly of information, but this rent comes partly in the form of *slack*—i.e. from

suboptimally low levels of cost-reducing effort. Thus there is a precise sense in which the optimal regulatory mechanism involves X-inefficiency (see Leibenstein, 1966).

If the regulator did not have available the possibility of making a lump-sum transfer, he would of course have more problems. Price would have to exceed unit cost in order to cover the cost of effort. This would necessitate a departure from allocative efficiency, and would further attenuate incentives for cost reduction because gains from reducing unit costs would be spread across fewer units of output.

The analysis in the last two sections has been more technical than most of this book, but we must emphasize that it has not been at all rigorous. Our aim has been simply to try to convey the flavor of some of the methods used to analyze asymmetric information. A more exhaustive and rigorous treatment is given by Caillaud *et al.* (1985). Next we summarize the main findings of recent work on regulation under asymmetric information, and discuss some important extensions of the analysis.

4.3.4 Regulation with Asymmetric Information: Conclusions

Asymmetric information is at the heart of the economics of regulation. If the government and the firm's managers had access to the same information about industry conditions and the firm's behavior, then the regulatory problem could be solved by simply directing the managers to implement the socially optimal plan given the (common) information available. In reality, however, managers are much better informed about industry conditions than are the firm's owners and regulators, and their behavior can be monitored only imperfectly. The question is how to motivate managers to exploit their superior information to advantage despite the problem of imperfect monitoring. Note here the very close analogy between (a) the problem that a firm's owners (public or private) have in giving managers incentives to act in the owners' interests, and (b) the problem that government regulators have in giving a regulated firm (or its managers) incentives to act in the public interest.

Chapter 2 on ownership considered problem (a), while the present chapter is concerned with problem (b). Ideally we would like to combine (a) and (b) since the incentives of the managers of a regulated firm are influenced by both its owners and its regulators. However, that would raise very complex issues, and for the present we leave aside problem (a) by supposing that the managers of a regulated private firm act as profit maximizers.

Theories of regulation in the Averch–Johnson tradition do not explicitly take account of asymmetric information. Their purpose is to examine the consequences for firm behavior of given (and not necessarily optimal) regulatory schemes. The recent work reviewed in this section, which explicitly models asymmetries of information, addresses the question of what is the *optimal* regulatory mechanism given the information available. In doing so, it illuminates the trade-offs between internal and allocative efficiency that result from asymmetric information, and it reveals how the effectiveness of regulation depends critically upon the information available to the regulators.

In the model proposed by Baron and Myerson (1982) the government cannot observe the (exogenously given) cost structure of the firm. The government attaches more weight to consumer interests than to producer interests, and a scheme of the type suggested by Loeb and Magat (1979)—in which the firm receives consumer surplus minus a fixed amount—is therefore undesirable on distributional grounds because the firm would tend to make large profits. The government would like price to equal unit (variable) cost, but it cannot observe cost. If it imposed a low price, there would be some circumstances in which the firm would refuse to supply the market. In order to avoid this unpleasant result, the government's regulatory scheme must strike a compromise, and it turns out that price generally exceeds unit costs at the optimal compromise. Allocative inefficiency is the result. Furthermore, the firm generally makes a positive profit thanks to its "monopoly of information."

Laffont and Tirole (1986) extend the model by allowing costs to depend upon the firm's efforts as well as on given circumstances. The government is assumed to be able to observe the level of costs, but not the extent of cost-reducing effort. It cannot tell whether low costs are due to good luck or effort. The trade-off between internal efficiency (i.e. optimal effort given output) and allocative efficiency (i.e. optimal output given effort) is clear. Setting price equal to unit cost gives perfect incentives for allocative efficiency but no incentive for cost reduction. Setting price equal to a given constant gives perfect incentives for internal efficiency but poor allocative efficiency. The optimal compromise involves output being lower, and price higher, than at the optimum with symmetric information. The degree of cost reduction is too low, and so there is internal inefficiency. Once again the firm benefits from its "monopoly of information," and the government is doubly disadvantaged by the asymmetry of information. The outcome is inefficient, and the firm extracts a profit from its informational advantage.

Analyses of this kind can be extended in various ways. Baron and Besanko (1984) introduce the possibility of costly *ex post auditing* of the

conditions facing the firm, which can enhance efficiency by diminishing the asymmetry of information between regulator and firm. The same authors (Baron and Besanko, 1987) examine regulation under asymmetric information in a *dynamic* setting. Over time the regulator may be able to learn about the cost conditions facing the firm, and choose a regulatory mechanism that uses the information that emerges. Much also depends on whether the regulator can commit his strategy in advance (see also Freixas and Laffont, 1985). Baron and Besanko examine intermediate degrees of commitment, in particular a "fairness" condition. For an excellent survey of all these matters and more, we again refer the interested reader to Caillaud *et al.* (1985).

4.4 Regulation of Multiproduct Firms

The economics of regulating multiproduct firms is central to an assessment of policy towards companies such as BT, British Gas, and the electricity supply industry (ESI), irrespective of how they are owned. It is obvious that BT supplies a wide range of products (telephone handsets, mobile phone services, private branch exchanges, etc.), and its principal activity (supplying telephone calls) is also a complex business. A call made at 10 a.m. on Monday is a separate product from one made at 4 a.m. on Sunday. A local call within Oxford is a different product from a long-distance call from Oxford to Glasgow (or to Washington). BT's pricing structure must reflect these differences between time and place, and their associated costs. This task is complicated by the fact that many costs are shared between various types of call, and the question arises of which consumers should bear them. Very similar issues are faced by energy utilities such as the electricity industry. Demand fluctuates between times of the day and year, and is influenced by the weather. Given the limitations on capacity and the difficulty of storing output, the electricity pricing structure must be sensitive to demand variation if rationing is to be avoided. Again the question arises of how to cover common costs (e.g. generating and distribution capacity).

Multiproduct pricing and investment problems are of course the subject of a large body of theory on public enterprise, which we have no wish to replicate here (see, for example, Atkinson and Stiglitz, 1980, chapter 15; Baumol and Bradford, 1970; Bös, 1986; Diamond and Mirrlees, 1971; Rees, 1984a,b). Rather, our aim is briefly to describe some work on the behavior of multiproduct profit-maximizing firms subject to regulatory constraint. In particular we will outline the dynamic regulatory adjustment mechanism proposed by Vogelsang and Finsinger (1979), but we begin by looking at the problem in its simpler static form.

Consider a multiproduct firm producing outputs $Q_1, Q_2, \ldots, Q_n$ for n markets. Let Q denote this vector of outputs. Let $P = (P_1, P_2, \ldots, P_n)$ be the vector of prices in the various markets. The demand Q_i for product i will depend on the price vector P. The firm's costs $C(Q)$ will depend on the output vector Q, which in turn depends on the prices P. We assume natural monopoly cost conditions. We can write the firm's profit as a function of prices:

$$\pi(P) = \Sigma P_i Q_i(P) - C[Q(P)] \quad (4.31)$$

The first term on the right-hand side is the sum of the firm's revenues in the various markets that it serves, and the second term is its costs. Note that we have not made very restrictive assumptions about the dependence of demand on prices, or the dependence of cost on outputs.

Let consumer surplus (the sum of the areas under each demand curve) plus profit be the social welfare objective. Consumer surplus will depend on the prices charged, and we will denote it by $S(P)$. A useful fact is that

$$\frac{-\partial S(P)}{\partial P_i} = Q_i(P). \quad (4.32)$$

If the price of product i is increased by a small unit, then the loss of consumer surplus is equal to that unit times the quantity of product i demanded.

Which pricing and production plan maximizes social welfare? The ideal solution (the first-best) has marginal cost pricing for each product:

$$P_i = \mathrm{MC}_i = \frac{\partial C}{\partial Q_i}. \quad (4.33)$$

But marginal cost pricing entails losses when there are scale economies. If transfers from the government to the firm are impossible or undesirable, social welfare must be maximized subject to a break-even constraint. The problem then is equivalent to choosing P to

maximize $S(P)$ subject to $\pi(P) \geqslant 0$.

Under fairly mild assumptions about cost and demand conditions, the solution to this second-best problem requires that the term

$$\frac{P_i - \mathrm{MC}_i}{\mathrm{MR}_i - \mathrm{MC}_i} \quad (4.34)$$

is the same in each market (MR_i denotes marginal revenue in market i). This is known as Ramsey pricing. The Ramsey formula implies that price–cost mark-ups are higher in markets where demand is less elastic.

Now let us turn to the profit-maximizing decision of a regulated private firm. Suppose that regulation takes the form of an average price constraint (this is roughly how BT is regulated). Suppose that a weighted average of its prices must be less than a given level $\bar{P}$:

$$\Sigma w_i P_i \leqslant \bar{P}, \quad (4.35)$$

where the w_i are positive weights that add up to unity. The firm maximizes $\pi(P)$ subject to (4.35). Everything now depends on how the weights and the price limit $\bar{P}$ are chosen. An important special case is when they are chosen in such a way that the firm can just break even, and when the weights are proportional to the demands for each product when there is Ramsey pricing. Given some assumptions about cost and demand conditions, it then turns out that *the regulated private firm chooses Ramsey pricing*.

Figure 4.4 attempts to illustrate why this is so in the two-product case. (See Vogelsang and Finsinger (1979) for a more rigorous account. The arguments depend on assumptions (e.g. about concavity) that we do not detail here.) The shaded region contains the price combinations that satisfy constraint (4.35) for a particular choice of the weights w_i and the limit $\bar{P}$. The diagram shows a consumer surplus indifference curve ($S(P)$ = constant) and an isoprofit curve tangential to the line representing the price constraint. Consumer surplus rises, but profit falls, nearer the origin. It is

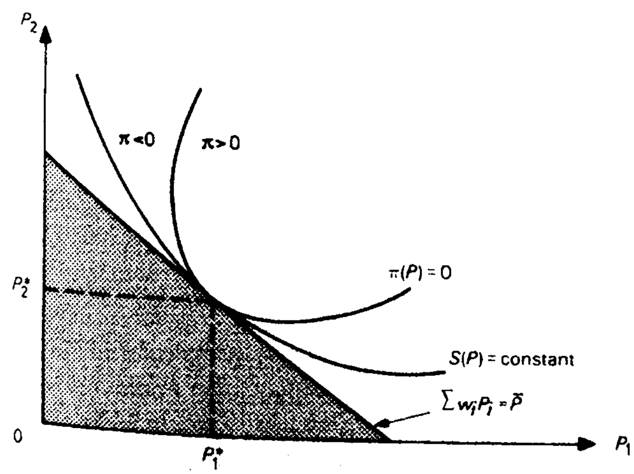

Figure 4.4 Regulation of multiproduct pricing

evident from the way that the diagram has been drawn that the price pair (P^*_1, P^*_2) maximizes consumer surplus subject to $\pi(P) \geqslant 0$. These are the Ramsey prices. The price constraint has been set so that the same price combination maximizes profit subject to prices obeying the regulatory price constraint. Thus the price constraint has induced Ramsey pricing.

How did the choice of w_i and $\bar{P}$ bring this about? The weights w_i determine the slope of the price constraint line, and $\bar{P}$ determines its distance from the origin. For given weights w_i it is easy to choose a level of $\bar{P}$ that causes the price constraint line to be a tangent to the $\pi(P) = 0$ locus. Weights proportional to the quantities demanded at Ramsey prices ensure that this point of tangency is also the point where the consumer surplus indifference curve is a tangent to the price constraint line. That is because the slope of the consumer surplus indifference curve at that point is

$$\frac{dP_2}{dP_1} = \frac{-\partial S(P^*)}{\partial P_1} \div \frac{\partial S(P^*)}{\partial P_2} = \frac{-Q_1(P^*)}{Q_2(P^*)}. \tag{4.36}$$

The first equality in (4.36) follows from totally differentiating $S(P^*) =$ constant, and the second equality follows directly from (4.32). If the weights are chosen so that w_i is proportional to $Q_i(P^*)$, it follows that the points of tangency coincide: maximizing profit subject to the appropriate price constraint delivers the same result as maximizing welfare subject to a break-even constraint.

The general reason why this form of price control produces the (constrained) optimal outcome can be outlined as follows. The problem of maximizing profit subject to the price constraint has the same solution as the problem of minimizing the cost of purchasing the consumption bundle demanded at Ramsey prices subject to the break-even constraint. (This follows from the weights' being proportional to Ramsey quantities.) The latter problem is equivalent to maximizing consumer surplus subject to the break-even constraint, because both involve minimizing the expenditure needed to obtain the consumers' preferred consumption bundle.

To summarize, private ownership of a multiproduct monopolist is no bar to allocative efficiency provided that the price control formula is aptly chosen. This proviso must be emphasized strongly, because the information about cost and demand needed to set the w_i and $\bar{P}$ correctly is very difficult to obtain. If the government possessed the information, it could just as well run the industry itself and implement optimal pricing directly. Before turning to dynamics there is one last point to make about the static case. There is a major difference between the average *price* constraint (4.35) and an average *revenue* constraint of the form

$$\frac{\Sigma P_i Q_i}{\Sigma Q_i} \leqslant \bar{P}. \tag{4.37}$$

The difference is that the weights in (4.35) are exogenous to the firm, whereas the weights (i.e. $Q_i/\Sigma Q_i$) in (4.37) depend on the firm's behavior. The optimality result for the average price constraint does not carry over to the average revenue constraint. In the latter case the firm has an incentive to behave strategically to alter the weights in its price control formula, and allocative inefficiency is the usual result (see section 9.2.4 for a discussion of this point in relation to the average revenue constraint facing British Gas).

The dynamic price control mechanism proposed by Vogelsang and Finsinger (V–F) (1979) is motivated by the limitation on the government's information that was mentioned in the paragraph above (and analyzed at some length in section 4.3). The government simply cannot know enough to design a price control formula that induces Ramsey pricing. Indeed, it does not know what the Ramsey prices are. The V–F mechanism is designed to enforce the eventual adoption of Ramsey pricing despite this lack of information about cost and demand functions.

The V–F mechanism allows the monopoly firm to choose its product prices in each time period subject to the condition that a weighted average of its prices should not exceed a given level. In particular, the prices charged in period t must be such that the revenues from setting the previous period's outputs at those prices must not exceed the total costs incurred in the previous period. In notation, the constraint is that the prices charged in period t must satisfy

$$\Sigma_i P^t_i Q_i(P^{t-1}) \leqslant C[Q(P^{t-1})], \tag{4.38}$$

where the superscripts (t and $t-1$) indicate time periods. It is assumed that the government can observe last period's prices, outputs, and costs, though it does not know the cost and demand functions. For a single-product firm the constraint is that the price in period t must not exceed the average cost in period $t-1$. (It is assumed that there are scale economies. Otherwise the firm would not be able to find a price that meets this constraint and makes a profit. Similarly, in the multiproduct case, it is assumed that there are decreasing ray average costs, i.e. the average cost of supplying a given bundle of goods decreases as the scale of output increases.)

It is assumed that the firm knows its cost and demand functions (which are constant across time), and that it maximizes profits in each period. It follows that its behavior subject to constraint (4.38) leads to an increasing level of social welfare over time. In the limit welfare converges to a level

$W^* = W(P^*)$ such that the optimal (Ramsey) conditions hold at P^*. Thus the V–F mechanism appears to have very desirable incentive properties, despite the limited information available to government. The mechanism gives the firm freedom over relative prices and uses a kind of regulatory lag. In the words of Vogelsang and Finsinger (1979, p. 170) "the regulated firm ... is encouraged to exploit both the potential for cost decreases and the consumers' willingness to pay. The firm converts these into profits. But both these advantages are turned over to the consumers in the next period."

However, the V–F mechanism has weaknesses, and a major problem is analyzed by Sappington (1980). The desirable welfare properties of the V–F mechanism are based on the assumption that the firm is a short-run profit maximizer, but (as Vogelsang and Finsinger themselves recognize) the firm may respond to the V–F mechanism strategically when it has longer-term aims. Sappington (1980, p. 360) argues that "pure waste, inefficient factor utilization, excessive research and development, and overinvestment in demand-increasing expenditures may be employed by a firm to increase long-run profits." Indeed, he shows that V–F regulation can be worse than no regulation at all. Sappington shows that the V–F mechanism can encourage the firm to engage in "pure waste," i.e. the deliberate raising of costs. The strategic advantage of pure waste today is that it increases the permitted level of prices tomorrow. The idea of a firm engaging in pure waste might be thought rather implausible, but the point applies more generally. For example, the firm might strategically slacken its efforts to cut X-inefficiency and there are many instruments that the firm might use in order to manipulate the prices that it can charge under the regulatory scheme.

In this section we have considered the *structure* (as well as the overall level) of prices chosen by a multiproduct monopolist subject to regulation. In a static framework we saw that a profit-maximizing firm can be induced to adopt a desirable pricing structure when regulation takes the simple form of a limit on a suitably weighted average of the prices of the firm's various products. The problem is that the authorities generally do not possess enough information to set the weights and the price limit at the right levels. The ingenious V–F dynamic mechanism can bring about the optimal constraint endogenously over time, but it relies on myopic behavior by the firm. Many regulatory schemes would work well if firms were myopic, but in fact they tend not to be, and the pervasive danger of inefficient strategic manipulation arises once again because asymmetric information rules out its effective prevention.

Many points made in this section are directly relevant to an assessment of

the RPI $-$ X method of price control that has been adopted for many privatized firms in Britain. This system essentially lays down a limit on the average price that the multiproduct firm can charge, and there is long regulatory lag. In principle the system could encourage an efficient price structure if the parameters of the allowed pricing formula were set correctly, but the authorities' relative paucity of information prevents this from happening, except perhaps by chance. Secondly, the system is vulnerable to inefficient strategic manipulation of costs by the regulated firm, especially as the time of regulatory review draws near (a problem which is shared by many regulatory schemes).

Finally, regulating a multiproduct firm by the RPI $-$ X method faces further problems when the firm faces competition in some of its regulated product markets. In this section we have assumed that the firm is a pure monopolist, but companies like BT have some competitors even though they enjoy great market power. If the average price constraint covers markets in which there is some actual or potential competition, incentives for an efficient pricing structure can become distorted. The average price constraint encourages the firm to undercut its rivals in the competitive business segments by allowing it to recoup the costs of doing so elsewhere. This problem, which arises from averaging, calls for more product-specific regulation and for safeguards against anticompetitive conduct. These questions go beyond the scope of the present section, and we will return to them when discussing regulation in practice in later chapters.

4.5 Collusion between Regulator and Firm

So far we have discussed what are often called *public interest* theories of regulation. These theories take it as given that the purpose of regulation is the enhancement of economic welfare via improved efficiency in resource allocation, and that the established agencies faithfully pursue the implied allocative objectives. There is a second major strand in the U.S. literature, however, which explicitly challenges these assumptions. Work in this second tradition—often labelled the *economic* theory of regulation (to emphasize that it is concerned with the determinants of the supply of, and demand for, regulatory activities)—has developed from seminal papers by Stigler (1971), Posner (1971), and Peltzman (1976). This work has focused heavily upon the income-distribution consequences of regulatory processes and the incentives faced by the regulators themselves. The theories are intended to be nonnormative, and seek to explain how particular forms of regulation emerge and change by evaluating the gains and losses implied by alternative institutional arrangements for the various interest groups involved.

Some of the interest group pressures on regulators fixing the prices of monopolistic firms are clear enough: consumers benefit from lower prices and producers favor higher prices (up to the unconstrained monopoly level). There are other potentially important aspects of the problem, however, that may be relevant: trade unions may align themselves with management on the pricing issue, hoping to appropriate some of the monopoly returns in the form of higher wages or better working conditions; consumers tend to be less well organized as a lobby group than either management or labor; the greater frequency of contact between management and regulators could, over a period, make the latter more receptive to the firms' arguments; regulators may be influenced by the prospect of remunerative employment in the industry once their public service days are over.

The ways in which these various pressures filter through into regulatory policies are affected by the institutional arrangements of the agencies, and by the constraints placed on the latter in the form of delegated mandates and judicial decisions. In the United States most regulatory agencies have relatively vague mandates (requiring, say, that their rates be "just, reasonable and nondiscriminatory"), thus leaving commissioners with significant discretion as to their interpretation and thereby opening up more opportunities for pressure group lobbying.

The effect of such lobbying will also depend upon the terms of appointment of the commissioners. Factors such as length of service, whether commissioners are appointed or elected, restrictions on re-appointment or re-election, etc. vary considerably from agency to agency in the United States, illustrating the range of alternatives that have been considered appropriate in different circumstances.

There is particular reason to be concerned about the potential influence of producer groups on regulations dealing with new entry into the industry. The effects on consumers of entry restrictions are less visible than the effects of price fixing, and there is a public interest argument in favor of control of entry that could be used in self-serving ways by producer groups. Simply stated, it is that natural monopoly implies that efficiency is improved by having the goods or services in question supplied by a single firm, and that entry prohibition is necessary to guarantee this outcome.

We are, to say the least, highly skeptical of this argument in favor of entry restriction. In the first place, an efficient dominant firm, with significant sunk costs and subject to price regulation, is unlikely to be highly

vulnerable to substantial entry threats. More importantly, given the inherent difficulty of actually establishing whether or not a given industry is subject to natural monopoly conditions, it is probable that entry restrictions would in many cases lead to supply by a single firm when the goods or services in question could be more efficiently provided by separate firms. This is particularly significant when technological change is rapid and cost conditions are constantly changing. Finally, if entry threats are removed there will be a corresponding loss of incentives for production efficiency and innovation on the part of the dominant firm.

A theoretical perspective on the possibility of collusion between regulator and firm is provided by work on *hierarchies*, i.e. principal–agent relationships consisting of several levels (see the discussion in chapter 2, and Caillaud *et al.* (1985, section 7)). In the previous section we examined the principal–agent relationship

regulator → firm,

and we supposed that the regulator had the public interest at heart. A generalization is the scheme

government → regulator → firm.

If members of the regulatory agency have interests that do not coincide with the public interest, we should consider the first link in this chain as well as the second. For example, the government might wish to limit the discretion of the regulator. This is what happens under the RPI $-$ X schemes of price regulation being introduced in Britain. The regulator has the duty of seeing that the firm complies with this general formula, but has no duty to intervene on specific pricing decisions.

A further generalization might be appropriate if it was felt that members of government are not necessarily fervent champions of the public interest. Then we might have

voters → government → regulator → firm.

Indeed it has been suggested in connection with the U.K. privatization program that it is government, rather than regulators, which has been partly "captured" by firms when designing regimes of competition and regulation for them. Related criticisms can be made of the control of nationalized industries. In chapter 5 below we will describe how the long-term development of nationalized industries in Britain has often suffered from government intervention and constraints motivated by short-term considerations. One of the advantages sometimes claimed for privatization is that it avoids problems of this kind. Whether or not it is the best way of doing so is another matter, to which we shall return later.

4.6 Competition and Regulation

In this section we examine three respects in which competition and regulation interact. The first is competition *for* monopoly, or franchising, which has indeed been used in a number of areas, for example local authority services in Britain where private operators have opportunities to outbid and displace public suppliers. Franchising has many attractive features especially where the product in question has a simple specification, but in industries of any complexity its merits are likely to be outweighed by problems of uncompetitive bidding, the handover of fixed assets, and contract monitoring. The second theme is competition via regulation, or "yardstick competition," in which regulated units in submarkets that are distinct (e.g. geographically) are brought into competition by the regulatory mechanism. For example, the price increase allowed in region A might be a function partly of cost performance in regions B and C. Thirdly we look at regulation in industries where there is some competition (actual or potential), as in the U.K. telecommunications industry for instance. The presence of competition influences the appropriate form of regulation, and regulation in turn affects the effectiveness of measures to permit or promote competition. Thus it is an important part of the task of regulatory authorities in the U.K. to try to guard against anticompetitive behavior. Here the overlap between regulatory and antitrust policies is most evident, and indeed it should always be remembered that regulatory mechanisms are just one element in the overall combination of public policies toward industrial organization and behavior.

4.6.1 Franchising

The dilemma for policy regarding natural monopolies is how to enjoy the cost benefits of single-firm production without suffering from monopolistic behavior. One answer is to have a competition—in the form of an auction—*for* the monopoly, with several firms competing to be the one that actually operates in the market. We will concentrate on an attractive form of franchising that was originally advocated by Edwin Chadwick, the Victorian social reformer, and developed by Demsetz (1969) in his famous

article "Why regulate utilities?" According to the Chadwick–Demsetz (C–D) proposal, the franchise is awarded for a period of time to the competitor offering to supply the product or service at the lowest price(s), or, more generally, the best price–quality package (for a review of franchising in relation to natural monopoly, see Sharpe (1982)).

On the face of it, franchising appears to provide a very attractive way of combining competition and efficiency without any arduous burden for regulators. The competition for monopoly appears to destroy the undesirable monopoly of information that hinders traditional regulation, and price is set by competition, not by administrators. In practice, franchising has been successful in a number of fields. For example, a study in 1986 by the London Business School and Institute for Fiscal Studies showed that local authorities in Britain using private contractors have reduced costs by 22 percent on average while maintaining the standard of services. Local authorities have successfully used competitive tendering, which is a form of franchising, for subsidized bus services (see section 10.4). It goes almost without saying that franchising is widely used within the private sector.

However, there are many industries where franchising cannot work, at any rate in this simple form, and the industries described later in this book (energy, telecommunications, water, etc.) provide leading examples. We shall focus on three sources of difficulty—the danger that bidding for the franchise will be uncompetitive, problems of asset handover, and, most important, the difficulties of contract specification and monitoring.

There are two reasons why bidding for the franchise might fail to be competitive. First, there is a danger of *collusion* between bidders, especially if they are few in number (e.g. because the requisite skills or resources are rare) or if the firms are effectively in a repeated game with one another by virtue of frequent contacts of various kinds.

The second reason is that one firm might enjoy such *strategic advantages* in the competition for the franchise that other firms would be unwilling to compete with it. Suppose, for example, that firm A has recently been the holder of a franchise that is now up for renewal. If the experience gained by A from its past operation of the franchise has had the effect of reducing its costs of operation, then the future franchise is worth more to firm A than to other firms. This fact might deter the other firms even from competing with A for the future franchise because they know that they are unlikely to win the competition.

Another source of incumbent advantage can arise from asymmetries of information. If A is the incumbent operator of the franchise, then A's knowledge of cost and demand conditions is likely to be superior to that of any other firm. This will tend to deter others from competing with A for the future franchise. For if firm B outbids A for the franchise, it is likely that B has bid too much. The fact that the relatively ignorant firm B wins against the knowledgeable firm A is itself an indication that B has paid over the odds. This problem is sometimes known as the "winner's curse." Its effect is to deter competition with the knowledgeable firm, i.e. the incumbent. (Precisely how the effect operates depends, of course, upon the exact nature of the competition for the franchise. Although we do not present a formal model here, we believe that the verbal argument is sufficient to establish the general point. (See further Englebrecht-Wiggans *et al.* (1983) on the value of information in auctions.))

We now turn to problems of *asset handover*. Suppose that A has held the franchise until now, but that B has just defeated A in the competition for the franchise for the next interval of time. What happens to the assets hitherto used by A to operate the franchise? Unless sunk costs are zero (an extremely unlikely event) efficiency requires that B, the new operator of the franchise, takes over these assets from A. Otherwise there will be inefficient duplication of the assets. But how are the assets to be valued for this purpose? Here there is a problem of bilateral monopoly. If A had no alternative, it would accept as little as the scrap value of the assets. If B had no alternative, it would pay as much as their replacement value. The gap between replacement value and scrap value is likely to be large if the assets involve sunk costs, and the expense of bargaining or arbitration regarding the appropriate transfer price might well be considerable.

This fact in turn has implications for the nature of competition for the renewal of the franchise itself. Let X and Y denote the values to A and B respectively of operating the franchise in the future, aside from the cost of transferring the assets and bargaining costs. Let Z be the amount paid by B to A for the assets if B wins the future franchise, and let C_A and C_B be the bargaining costs of the two firms in that event. If A wins the franchise it receives X, and if A loses it receives $Z - C_A$. A's incentive to win is therefore $X - Z + C_A$. If B wins, it receives $Y - Z - C_B$ (which we initially assume to be positive), and if B loses it receives zero. Therefore A has a greater incentive than B if and only if

$$X + C_A + C_B > Y. \tag{4.39}$$

The condition for A to be a more efficient franchise operator than B is simply

$$X > Y. \tag{4.40}$$

A comparison of these two inequalities shows that the costs of bargaining $(C_A + C_B)$ have the effect of giving the incumbent firm A an advantage, because bargaining costs are avoided if the franchise does not change hands.

Note that Z, the amount paid to A for the assets if B wins, does not affect the *difference between* the incentives of A and B in the franchise competition (provided that $Z < Y - C_B$). This is because a higher level of Z reduces A's incentive to win just as much as it reduces B's incentive. However, if $Z \geqslant Y - C_B$, the level of Z *does* effect competition for the franchise, because B cannot make a positive profit whatever it bids and so would not compete. Thus A would be the only contestant. This consideration indicates that some form of regulation of the level of Z may be required.

Moreover the level of Z certainly influences the *level* of the bids that would be made in the auction for the franchise. If B could purchase A's assets at low cost, then B would be prepared to bid more than if the assets were more costly for B to acquire from A. Similarly, the incumbent firm A would compete less vigorously with B if B were required to pay more for the assets. Therefore the level of Z is bound to influence the size of incentives (if not the difference between them) and hence the efficiency of resulting pricing arrangements, especially if the auction is of the C–D type.

The level of Z, the amount paid for the assets of the displaced franchise, is also a critical determinant of the *investment* decisions of an incumbent firm. If it is thought likely that Z will be low (e.g. because the assets are of minimal value to an outgoing incumbent), then the existing incumbent will have an incentive to underinvest if there is any chance that he will fail to win future competitions for franchise renewal. On the other hand, if he were to receive an inflated price for the assets being passed on, he might have an incentive to overinvest.

More generally, there is likely to be considerable *uncertainty* about the level of Z *ex ante*. With risk-averse firms, this will affect investment strategies, bidding behavior, and perhaps even the decision to enter the competition for the franchise.

These numerous problems of asset valuation and handover perhaps suggest that investment decisions should be left to public authority and that the competition should be simply for an *operating* franchise. However, operating franchises allow market forces to act only to a limited extent, and the divorce of investment and operating decisions can lead to undesirable losses of coordination.

Finally we come to the important question of the *specification and* *administration of franchise contracts* (see in particular Williamson, 1976; Goldberg, 1976). If a franchise contract is for the provision of a well-defined product or service—for example the production of a thousand taxi license plates of a given specification at a given time—then the contract between franchisor and franchisee is a relatively simple affair that requires little effort to administer. But if there is technological or market uncertainty in relation to the product, then the specification of the franchise contract can be a very complex task, and the need to monitor and administer the contract during its life is certain to arise.

Williamson (1976) draws important distinctions between different types of franchise contract. A *complete* contract requires a franchise bidder to specify the terms on which he will supply the product or service at each future date during the life of the contract, and for every future contingency that might arise. A complete contract sensitive to future events would be impossibly expensive to write, negotiate, and enforce if uncertainty is present. But a complete contract does not have to take a complex form. For instance, a contract might simply say that the price charged will be such-and-such in all circumstances—i.e. whatever happens to demand, production costs, inflation, and so on. But an unconditional contract of this form faces two severe problems. First, the firm might be unable to fulfill the contract under some circumstances. The threat of inability or refusal to supply would probably lead to flexibility *ex post*, even though the original contract had been specified unconditionally. Therefore, unconditional contracts, especially if they are longer term, are likely to be infeasible. Moreover, unconditional contracts are undesirable. Considerations of efficiency require that price and quality adapt in response to changes in demand and technology.

Thus we are left with *incomplete* contracts, which do not make explicit what is to happen in every possible circumstance. With incomplete contracts there is a need for administration and monitoring of the (partly implicit) contract as time unfolds; a continuing contractual relationship exists, and this inevitably involves continuing costs. The alternative is for the franchisor to be left at the mercy of the franchisee.

The duration of the franchise contract must also be considered. The difficulties of contract specification and administration alluded to in the previous paragraph perhaps suggest that short-term contracts have advantages, because fewer future contingencies then need to be catered for. But the organization of frequent contests for the franchise also involves major costs. As well as the direct costs of holding more auctions, all the problems of asset valuation and handover (see above) occur more often, and the industry would frequently be in a state of turmoil.

The conclusion to be drawn is that, in industries where there is significant uncertainty about technology and demand, competition for monopoly by franchising does not have many of the advantages over regulation that it superficially appears to possess. Indeed franchising involves an implicit regulatory contract for all but the simplest products and services. As Goldberg (1976, p. 426) writes: "Many of the problems associated with regulation lie in what is being regulated, not in the act of regulation itself."

4.6.2 Yardstick Competition

One of the main themes emphasized in this chapter has been the importance of information for effective regulation. If the regulator is relatively uninformed about industry conditions, and especially if the firm being regulated has a monopoly of information, the regulatory mechanism is liable to become a blunt instrument that is insensitive to the basic parameters of cost and demand. Economic efficiency (in both allocative and internal terms) becomes impaired, and the firm extracts monetary or slack rent from its monopoly of information.

Yardstick competition is a method of promoting competition between regulated units indirectly via the regulatory mechanism. It has been proposed in the Littlechild Report (1986) on regulation of the U.K. water authorities. To take the simplest example, suppose that a national monopolist was split into separately owned northern and southern units, denoted N and S respectively, each with a natural monopoly in its geographical area. Suppose further that cost and demand conditions were very similar in the two regions, although the regulator might not know (say) the scope for cost reduction in either region. The two regional units could be brought into competition by the following kind of regulatory mechanism. The price that N could charge in a given period of time would depend on the level of costs achieved by S, and vice versa for the price allowed to firm S. Provided that N and S face very similar circumstances, and that they do not collude in any way, a method of this kind offers the prospect of combining both internal and allocative efficiency, and therefore of escaping the dilemma that usually exists between the two. Good incentives for internal efficiency exists because N keeps the benefits of its cost-reducing activities, for its price is linked to the cost performance of firm S. Allocative efficiency results if there is symmetry between firms, because industry prices are kept in line with industry costs. The promotion of competition via regulation overcomes the informational disadvantage of the regulator in an economical fashion, and shows again how competition can act as an efficient incentive mechanism.

Yardstick competition illustrates the general proposition that under asymmetric information, when a principal has many agents under his control, it is almost always the case that the optimal incentive scheme involves the reward of each agent's being contingent upon the performance of other agents as well as his own performance. The theoretical literature on this point includes Holmstrom (1982), Mookherjee (1984), and Nalebuff and Stiglitz (1983). It is particularly desirable to make reward contingent partly upon the performance of others when the uncertainties facing different agents are correlated to a high degree. If such correlation is absent, there is no advantage in linking reward to others' performance. Indeed, to do so would serve only to add "noise" in an undesirable way: the risk facing any agent would increase, and he would not be encouraged to behave as his circumstances warranted.

Shleifer (1985) examines a model of yardstick competition. In the basic version of the model there are n identical risk-neutral firms operating in a certain environment. Each faces demand curve $Q(P)$ in its market (the n markets are separate). A firm spending z on cost-reducing effort achieves unit cost level $c(z)$, with $c(0) = c_0$. The lump-sum transfer to the firm (if any) is denoted by T. Profit is therefore given by

$$\pi = [P - c(z)]Q(P) - z + T. \tag{4.41}$$

If the social welfare objective is the sum of consumer and producer surplus (and so is not affected by considerations of distribution or the cost of raising public funds), then the optimum subject to the nonnegative profit constraint has

$$P^* = c(z^*), \tag{4.42}$$

$$-c'(z^*)Q(P^*) = 1, \tag{4.43}$$

and

$$z^* = T^*. \tag{4.44}$$

In sum, price equals unit (and hence marginal) cost, efforts to reduce unit costs occur up to the point where their marginal cost (= 1) equals marginal benefit (= $-c'Q$, i.e. degree of cost reduction times volume of output), and the cost of effort is reimbursed by the lump-sum transfer.

However, this first-best outcome cannot be achieved if the regulator does not know the function $c(z)$, which describes the scope for cost reduction. Shleifer supposes that each firm is run by managers who like profits π but dislike effort z. In particular it is supposed that their preference ordering is lexicographic with profits preferred over leisure. This is the minimal extent

to which some weight can be given to leisure in managers' preferences. Even so, a regulatory regime in which $P = c$ and $z = T$ (which is sometimes known as "cost of service regulation") induces no cost-reducing effort whatsoever. Profit is the same (i.e. zero) for all z, and so managers prefer to minimize z by setting it at zero, and the cost level is therefore c_0.

The key to efficiency is to break the dependence of the price for firm i upon its cost level. Let

$$\bar{c}_i = \sum_{j \neq i} c_j/(n-1) \tag{4.45}$$

and

$$\bar{z}_i = \sum_{j \neq i} z_j/(n-1) \tag{4.46}$$

be the average cost and effort levels of firms other than i. These provide yardsticks against which to compare i's performance. Shleifer (1985, proposition 1) shows that the following regulatory mechanism for all firms i induces first-best behavior:

$$P_i = \bar{c}_i \tag{4.47}$$

and

$$T_i = \bar{z}_i. \tag{4.48}$$

The profit of firm i is then

$$\pi_i = [\bar{c}_i - c(z_i)]Q(\bar{c}_i) - z_i + \bar{z}_i, \tag{4.49}$$

and the first-order condition is therefore

$$-c'(z_i)Q(\bar{c}_i) - 1 = 0. \tag{4.50}$$

There is a symmetric Nash equilibrium in which all firms choose $c_i = c^*$, in which case $P^* = c^*$ and $T^* = z^*$, and Shleifer shows that there exists no asymmetric Nash equilibrium. Therefore the equilibrium that sustains the first-best outcome is unique. This result in fact holds with pricing rules considerably more general than (4.47), but the simple example suffices to establish the main point. A related result holds even when lump-sum transfers are impossible. Yardstick competition can then induce the second-best outcome, i.e. the social optimum subject to $T = 0$.

The main shortcoming of the version of the model described so far is that it assumes that firms operate in *identical* environments. Of course this is quite unrealistic. The economics of water supply, for example, differs substantially between geographical regions. The water authority operating in the Welsh mountains faces conditions and uncertainties that are correlated by no means perfectly with those encountered by the authority operating in the plains of East Anglia. Shleifer examines the use of regression analysis based on observable characteristics to screen out at least part of the heterogeneity between firms that occurs in practice. *"Reduced-form" regulation* operates roughly as follows. Let θ be the vector of exogenous observable characteristics with respect to which firms differ. The regulator estimates a regression of unit costs against θ. With a linear functional form the regression equation is $c = \alpha + \beta\theta$. With $\hat{\alpha}$ and $\hat{\beta}$ denoting the estimated coefficients, the "predicted" unit cost level for firm i is $\hat{c}_i = \hat{\alpha} + \hat{\beta}\theta_i$, where θ_i is the observable characteristic of that firm. The regulator then imposes the price rule $p_i = \hat{c}_i$ and an associated transfer rule.

Reduced-form regulation works well if θ captures almost all of the variation between firms and if it is truly exogenous. If θ fails to capture the full extent of diversity, reduced-form regulation does not provide perfect incentives, and it causes there to be undesirable noise. The problem is especially acute if characteristics that are correlated with the observed characteristics θ are omitted, because omitted-variable bias is then introduced. Further difficulties arise if firms are able to manipulate the observed characteristics, because incentives then exist for strategic rent-seeking behavior and signaling. Similarly, reduced-form regulation has the disadvantage of encouraging endless argument about the appropriate way to conduct the regression analysis, which variables to include, and so on.

More problems arise if firms are able to collude and thereby frustrate competition via regulation. If firms tacitly agreed to slack to an equal extent, inefficiency would persist. Incentives to cheat might be weak if, as is probable, firms were few in number and well informed about each others' behavior.

Despite these difficulties, however, we believe that competition via regulation can provide good incentive systems in a number of industries. We know that the best regulatory mechanisms will exploit information from comparative performance in some form, but the question is *how* to do so in any particular case. It would be foolish to attempt generalization of this, since the degree of homogeneity between regional units differs from industry to industry, but the benefits to be gained from breaking the monopoly of information of a regulated firm could be substantial in individual cases. This is a factor which should be taken into account when considering the regional break-up of privatized companies.

4.6.3 Regulation to Maintain Effective Competition

Regulation is not only called for when competition is absent. Regulation designed to maintain freedom of *entry* is sometimes essential if threats of potential competition are to have force. We showed in section 3.2.2 that an incumbent firm with market power usually has at its disposal a variety of instruments of strategic entry deterrence, and that incentives for predatory behavior are likely to exist. Unless this sort of conduct (and the threat of it) are checked by suitable policy measures, market "liberalization" in the legal sense can be quite ineffective.

It can be argued that the sanctions of ordinary competition policy are sufficient to strike down anticompetitive behavior of this kind, but we disagree for several reasons. First, competition policy in the U.K. (and elswhere) evolved at a time when dominant utility companies were in public ownership. The competition problems that arise in those industries were therefore not envisaged when policy was made, and so there is little reason why it should be expected to cope with them. Secondly, it can be argued that U.K. competition policy has weaknesses generally (see Sharpe, 1985). Certainly it has usually been less vigorous than U.S. antitrust policy. Thirdly, where the danger of particular anticompetitive practices can be foreseen, it makes sense to legislate against them in advance, and to give the specialist regulatory agency the duty of monitoring and enforcing the policy. This also reduces uncertainty. Finally, the agency has greater knowledge and expertise regarding industry conditions than a generalist competition authority can have. (A separate question, which we do not pursue here, is whether there should be one regulatory authority for all privatized utility companies, or one for each.)

In sum, we believe that the task of "regulation" to promote and maintain competition in industries with dominant privatized firms should belong to the regulatory authority for that industry. This is not to say that the general competition authority (the Monopolies and Mergers Commission (MMC) in the U.K.) has no role to play: the regulator should be able to refer cases to it, but he should also have sufficient power to deal swiftly with anticompetitive conduct if and when it occurs.

In section 3.2.2 we described some of the economics of anticompetitive behavior to deter entry. As regards practical policy measures to combat such conduct, we will pursue this question further when we come to consider the frameworks of competition and regulation that have been established for privatized industries in Britain.

4.7 Concluding Remarks

The purpose of this chapter has been to provide a theoretical perspective on the economics of regulation that will guide our assessment of U.K. regulatory policy in part II of the book.

We have focused on the *incentive* properties of various regulatory mechanisms to encourage both internal and allocative efficiency. We have seen how the regulator's relatively *imperfect information* can lead to an awkward trade-off between the two, and how the firm and its managers can enjoy rewards from their monopoly of information. This suggested that the social return from having better informed regulatory bodies could be high, and it indicated that the benefits from greater competition (potential if not actual) could extend to internal as well as allocative efficiency. We also emphasized the *dynamics* of regulation, and the *strategic interaction* between firm and regulator (or government) that can occur over time. We shall bear in mind all these themes when we come to the case studies of regulation in practice in the following chapters.

Part II

THE BRITISH PRIVATIZATION PROGRAM