*Chapter 23*

# OPTIMAL POLICIES FOR NATURAL MONOPOLIES

RONALD R. BRAEUTIGAM*

*Northwestern University*

## Contents

*The author would like to express appreciation to the two editors of this Handbook, and to John Panzar and Tai-Yeong Chung for a number of helpful comments and suggestions.

*Handbook of Industrial Organization, Volume II, Edited by R. Schmalensee and R.D. Willig*
*© Elsevier Science Publishers B.V., 1989*

## 1. Introduction

Over the past decade there has been substantial reform in many industries historically operating under heavy governmental control, both in the United States and abroad. In the United States, where such governmental control typically takes the form of regulation of privately owned enterprises when policy-makers believe that competition will not work well to allocate resources, remarkable changes have occurred in all or parts of the airline, railroad, motor carrier, telephone, cable television, natural gas and oil industries, among others.[1] Many other countries, including those in which such governmental intervention takes the form of nationalization, have recently been reconsidering the role of such governmental intervention as well.[2]

In many cases the basis for regulation has itself been at issue in the policy debates surrounding regulatory reform, often leading to a removal of or a reduction in the extent of governmental control of traditionally regulated industries. In other cases reform has had some effect even when the hand of regulation has not been retracted. For firms such as local electric and gas utilities, local telephone operating companies, and oil and gas pipelines (to name just a few), heavy regulation persists. Still, regulatory reform in these industries has led to a reassessment of the kinds of controls that might be utilized under regulation.

The primary purpose of this chapter is to examine some of the optimal policies that might be used to control a "natural monopoly". At the outset we must define just what a natural monopoly is from an economic perspective, and why it poses a problem that might warrant government intervention. Section 2 begins by examining these issues from a traditional perspective, which argues for regulation when there are pervasive economies of scale in a market. It then offers a more contemporary characterization of natural monopoly based on the concept of subadditivity of costs rather than on economies of scale.

Section 3 re-examines the natural monopoly problem with a thoughtful eye on the question: To regulate or not to regulate? Although the traditional view suggests that government intervention and natural monopoly go hand in hand, economic analysis since the late 1960s has suggested rather forcefully that there may be ways to introduce competition for a market, even if a natural monopoly structure exists within a market. Thus, one of the themes of this chapter is that

[1] See Weiss and Klass (1986) for a discussion of the nature and effects of regulatory reform in a number of these industries.

[2] Examples include the possible "privatization" of some railroads in Japan, the debate surrounding the sale of part of the ownership of the telephone system to the private sector in Great Britain, and the liberalized rules for interconnecting privately owned equipment to the telephone network in West Germany, among many others.

regulation is only one of several possible ways of dealing with a natural monopoly. Section 3 then provides an overview of possible competitive approaches to the natural monopoly problem.

As Section 3 will make clear, there will be circumstances when competition as a policy toward natural monopoly is not feasible or, even if feasible, may lead to a market outcome which is quite inefficient. Section 4 summarizes a number of ways in which one might improve the efficiency of resource allocation with government intervention, including external subsidies to the firm, and the regulation of tariffs with price discrimination (or "differential pricing") or the introduction of nonlinear outlay schedules (nonlinear tariffs). The concepts are introduced in the context of the single product firm. The section then discusses some of the problems encountered in the case of the multiproduct firm, including the common cost problem, i.e. the problem of pricing individual services when there are costs of production that are shared in the production of more than one output, and therefore cannot clearly be attributed to individual services.

The chapter then turns to some of the major concepts in optimal (economically efficient) pricing in regulated industries. These include peak load pricing (Section 5), Ramsey pricing (Section 6), and nonlinear outlay schedules (Section 7). Finally, Section 8 addresses a set of issues related to the "fairness" of regulated prices, often discussed in the context of "cross subsidy" or "interservice subsidy". After presenting and discussing the implications of some of the possible notions of subsidy, the section concludes by relating the concepts of subsidy free and economically efficient prices.

A chapter of this kind necessarily relies on (in fact focuses on) the work of many other researchers. Any attempt to cite the literature exhaustively would be futile, and another author attempting the same task would no doubt include a set of references somewhat different from those used here. My hope is that glaring omissions have been minimized and that readers will be understanding on this point. At the same time the author would like to acknowledge two references especially useful in the preparation of this manuscript. These are Baumol, Panzar and Willig (1982) and Brown and Sibley (1986).

## 2. The natural monopoly problem: A "traditional" view

The central economic argument for regulation of an industry is that the industry is characterized by "natural monopoly". The concept of natural monopoly has been refined over the years, particularly during the last decade. In this section we will first discuss a rather traditional view of natural monopoly and its importance with respect to the role of regulation as it might have been presented before the 1970s. We will then summarize a more recent perspective on these same issues.

In his classic treatise Kahn (1971, p. 2) describes the concept of natural monopoly to mean "that the technology of certain industries or the character of the service is such that the customer can be served at least cost or greatest net benefit only by a single firm (in the extreme case) or by a limited number of 'chosen instruments' ".[3] In Kahn's extreme case average cost declines as output increases throughout the range of production in the market; thus a single large firm serving the entire market would have a lower average cost than any smaller rival. In that case it will not be possible to have more than one firm operating in the market if the lowest possible average cost is to be achieved.

> This view is also presented by Scherer (1980, p. 482) who writes: "The most traditional economic case for regulation assumes the existence of natural monopoly – that is – where economies of scale are so persistent that a single firm can serve the market at a lower unit cost than two or more firms. Reasonably clear examples include electric power and gas distribution, local telephone service, railroading between pairs of small to medium-sized metropolitan areas, and the long-distance transportation of petroleum and gas in pipelines."[4]

The traditional story thus hinges on the existence of *economies of scale* (or *increasing returns to scale*) in an industry.

Strictly speaking, of course, the concept of economies of scale is one based on the technology of the firm.[5] In a single product production process with constant prices for factors of production, the notion of economies of scale means that the average cost schedule for the firm declines as market output increases. This can be illustrated as in Figure 23.1. The figure represents a market being served by a single firm producing a single, nonstorable output (or service), whose level is denoted by $y$. The (inverse) demand schedule for this product is shown as $p(y)$, where $p$ refers to the price of the output. The firm produces any given $y$ at the

[3] For good references on many of the topics addressed in this chapter, see Schmalensee (1978) and Crew and Kleindorfer (1986), which deal with alternatives in controlling a natural monopoly. See also "State Regulation of Public Utilities and Marginal Cost Pricing", by L.W. Weiss, Chapter 9 in Weiss and Klass (1981, p. 263).

[4] Scherer (1980, p. 482) also points out that regulation may be implemented in industries for a variety of reasons other than the existence of natural monopoly. For example, regulation might occur even in an efficiently operating market if those who hold political power are displeased with the market outcome. It might also be imposed if well organized political interest groups are able to "manipulate political levers" to realize political or economic gains that would not be achieved in an unregulated market. Because these reasons for regulation are based on political economy rather than on "natural monopoly", they are not treated further in this article. For more on these topics, see, for example, Hughes (1977), Posner (1974), and Peltzman (1976).

[5] See J.C. Panzar's contribution in Chapter 1 of this Handbook for an extensive overview of the production and cost concepts we will be using throughout this chapter.

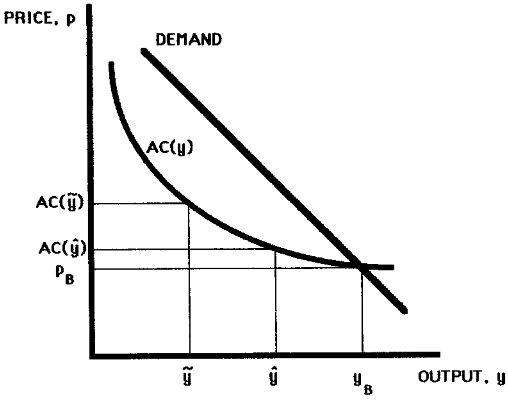

Figure 23.1. The "classic" natural monopoly problem.

minimum possible total cost, $C(y)$, and the average cost of production is denoted by the schedule $AC(y) = C(y)/y$.[6]

For the moment, assume that the firm receives no subsidy from external sources (including the government), and that it is not possible for the firm to price discriminate, so that a single, uniform price prevails in the market. The producer will need to generate total revenues that are at least as large as total costs to remain economically viable. Thus, the price charged by any firm will need to be at least as large as the average cost of production for that firm. As is clear from Figure 23.1, no firm can enter and produce $y > y_B$, since the output cannot be stored and profits would be negative for such a level of production. Furthermore, if any firm with the same technology enters the market and produces $\tilde{y} < y_B$, another firm could enter and produce $\hat{y}$, where $\tilde{y} < \hat{y} \leq y_B$; this second firm could charge a price $p$ in the range $AC(\hat{y}) \leq p < AC(\tilde{y})$, and drive the first firm from the market while remaining economically viable itself. The only production level that would preclude profitable entry by another firm charging a lower price is $y = y_B$, with $p = p_B$. In the traditional view the market is said to be characterized by a natural monopoly, since competition *within* the market is not possible.

The natural monopoly problem takes on added complexity when entry and exit are not costless and a temporal dimension is added to the problem. Firms might have incentives to enter the market, charge a price in excess of average cost to earn supernormal profits, and threaten to reduce price to a very low level (even

[6] More completely, the cost function is also a function of a vector of factor prices, $w$, $C(y, w)$. However, factor prices will be assumed constant throughout this chapter, so references to them will be suppressed to simplify notation as much as possible.

less than average cost) in the short run if any other firm should attempt to enter. As Kahn (1971, p. 2) states: "In such circumstances, so the argument runs, unrestricted entry will be wasteful... with cycles of excessive investment followed by destructive rivalry (spurred by the wide spread between marginal and average costs)". The potential for this so-called "destructive competition" has often been cited as a basis for regulating markets served by firms with substantial scale economies.

In short, the traditional notion of natural monopoly is based on the existence of economies of scale throughout the relevant range of production on the market. Such scale economies were typically taken to mean that competition might lead to greatly inefficient and even wildly fluctuating, unstable prices, so that government intervention of some sort was necessary.

What has happened to change the traditional view about natural monopoly? First, much of the regulatory experience of the past thirty years has made it clear that in many circumstances appropriate models of regulation must focus on the multiproduct nature of regulated firms. For example, during the 1960s the Federal Communications Commission began to open up so-called private line telephone service to competition, while leaving much of the intercity long distance telephone service regulated as a monopoly. Many researchers realized that the standard single product treatment of regulation in the literature was inadequate. Relatively recent research has shown that the appropriate definition of natural monopoly is one that rests on the concept of subadditivity of costs (discussed below) rather than on the more traditional notion of economies of scale; the two are related but not identical, and the difference between the two becomes particularly important when the production process involves multiple products.

To see this, first observe that a natural monopoly need not exhibit economies of scale throughout the range of production in the market. The simple single product example provided in Figure 23.2 makes this point clearly. Assume all firms that might like to provide the service in question have identical cost structures. In the figure each firm's average cost curve declines up to the production level $y^1$, and then increases (so that there are decreasing returns to scale) thereafter. The market demand schedule intersects the average cost curve at the output level $y_B > y^1$. Given the shapes of the curves in Figure 23.2, it is clear that a single supplier could serve the entire market at a lower unit cost than any industry configuration with two or more firms. In this sense the industry is therefore a natural monopoly, even though economies of scale do not exist for all levels of output up to $y_B$.

How then does subadditivity provide a better basis than economies of scale for determining when a natural monopoly exists? Consider the case in which there are $n$ different products and $k$ different firms. Each firm may produce any or all of the $n$ products. Let $y_r^i$ be the amount of output $r$ produced by firm $i$ $(i = 1,\ldots, k)$ and $(r = 1,\ldots, n)$. Also let the vector $y^i$ be the vector of outputs

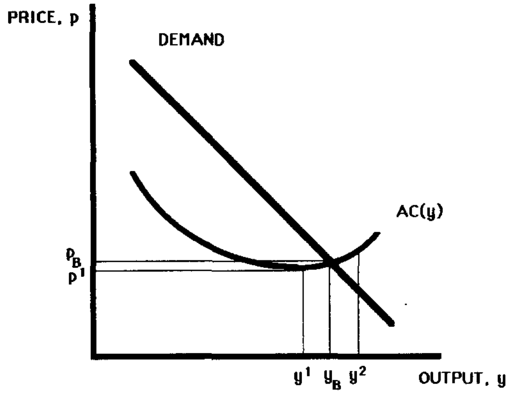

Figure 23.2. Subadditivity without global economies of scale.

$(y_1^i, y_2^i, \ldots, y_n^i)$ produced by the $i$th firm. Then, using the definition of Baumol, Panzar and Willig (1982, p. 17) a "cost function $C(y)$ is *strictly subadditive* at $y$ if for any and all quantities of outputs $y^1, \ldots, y^k$, $y^j \neq y$, $j = 1, \ldots, k$, such that

$$\sum_{j=1}^{k} y^j = y \quad \text{we have} \quad C(y) < \sum_{j=1}^{k} C(y^j).\text{''} \tag{1}$$

As (1) indicates, the vector $y$ represents the industry output. The basic question here is whether $y$ can be produced more cheaply by one firm producing $y$ all alone than it would be for a collection of two or more firms whose individual output vectors sum to the same industry output $y$.

Since costs may be subadditive at some values of $y$ but not at others, the next step toward defining a natural monopoly is to examine whether costs are subadditive at all of the "relevant" industry output vectors $y$ that might be produced; the demand for each of the outputs will help to define this relevant range of outputs. Baumol, Panzar and Willig go on to define a natural monopoly (still on p. 17) as follows: "An industry is said to be a natural monopoly if, over the entire relevant range of outputs, the firms' cost function is subadditive".

The example of Figure 23.2 illustrates that a subadditive cost structure need not exhibit economies of scale "over the entire relevant range of outputs". The example is constructed so that the output level associated with minimum cost, $y^1$, is slightly less than $y_B$, the output level at which the demand schedule intersects the average cost schedule. The average cost schedule has the typical "U" shape, and it is subadditive for $0 < y < y^2$ although economies of scale exist only over the (smaller) range of output $0 < y < y_B$. Thus, even in the single product case subadditivity does not imply economies of scale.

In the single product case it is clear that economies of scale imply subadditivity. However, it turns out that economies of scale need not imply subadditivity in the multiproduct case; this should not be a great surprise since, "given the crucial role of various forms of cost complementarity and economies of joint production, it is to be expected that economies of scale cannot tell the whole story in the multiproduce case".[7]

A comparison of Figure 23.1 with Figure 23.2 leads to another concept (sustainability) which is useful in appreciating difficulties that might be associated with the natural monopoly problem. Let us assume entry and exit are costless, that entrants will provide exactly the same service as the incumbent, and that all firms (the incumbent and all potential entrants) operate with access to the same technology, and therefore with the same cost functions. In the first graph, which depicts the traditional view of natural monopoly, it would be possible for the firm to find a price which deters entry by any other firm seeking to take away the incumbent's market by charging a lower price than the incumbent. In particular, if the extant firm charges a price $p_{\mathrm{B}}$, then any entrant charging a lower price will not be able to break even. In other words, if the incumbent charges $p_{\mathrm{B}}$, it can sustain its monopoly position against entry.

However, Panzar and Willig (1977) have pointed out that it will not always be the case that a natural monopoly can sustain itself against entry. They show that, contrary to conventional wisdom, a regulated monopolist may be vulnerable to entry, even if the incumbent produces efficiently, earns only a normal return on investment, and is confronted by an entrant operating with the same technology as its own.

Figure 23.2 presents such a case. Suppose that in serving the whole market the incumbent charges $p_{\mathrm{B}}$. Then it would be possible for an entrant to charge a lower price (say, $p = p^1$), provide $y^1$ units of service, and avoid a deficit. This is a case in which the market is unstable, and in which the natural monopoly is "unsustainable". If the whole market is to be served, it would therefore require two or more firms (since the entrant will produce only $y^1$ in the example). Furthermore, since the cost structure is subadditive in Figure 23.2, entry would be socially inefficient; yet, such entry is a real possibility, even though entrants might provide no new services and operate with no better productive technique.

Panzar and Willig have defined the concept of sustainability in a framework allowing for multiple products. Briefly, suppose that the monopolist produces $n$ different products in a product set $N$, and allow $S$ to be any subset of that

[7] See Baumol, Panzar and Willig (1982, pp. 173). For example, equation 7C1 on p. 172 represents a cost function that has globally increasing returns to scale, but is not subadditive everywhere. The cost function for that example is $C(y_1, y_2) = y_1^a + y_1^k y_2^k + y_2^a$, with $0 < a < 1$ and $0 < k < 1/2$. Sections 7C–7E of that book outline some proper tests of natural monopoly and sufficient conditions for subadditivity. See also J.C. Panzar's contribution in Chapter 1 of this Handbook for a more extensive discussion of several important concepts regarding market structure, including among others economies of scale and scope, the degrees of economies of scale and scope and product specific economies of scale.

product set ($S \subseteq N$). Let $\boldsymbol{p}^{\mathrm{m}}$ be a price vector charged by the monopolist over its product set $N$, let $\boldsymbol{p}_S^{\mathrm{e}}$ be the price vector charged by an entrant providing the product set $S$, and let the price vector charged by the monopolist over $S$ and over the services not provided by the entrant $[S]$ respectively be $\boldsymbol{p}_S^{\mathrm{m}}$ and $\boldsymbol{p}_{[S]}^{\mathrm{m}}$. Finally, denote by $\boldsymbol{Q}(\boldsymbol{p}^{\mathrm{m}})$ the vector of quantities that would be demanded if only the monopolist served the market, and let $\boldsymbol{Q}^S(\boldsymbol{p}_S^{\mathrm{e}}, \boldsymbol{p}_{[S]}^{\mathrm{m}})$ be the quantities of the product set $S$ demanded when the entrant appears. Then the price vector $\boldsymbol{p}^{\mathrm{m}}$ is sustainable if and only if (i) the monopolist earns non-negative profits at $\boldsymbol{p}^{\mathrm{m}}$, and (ii) $\boldsymbol{p}_S^{\mathrm{e}} \cdot \boldsymbol{y}_S^{\mathrm{e}} - C(\boldsymbol{y}_S^{\mathrm{e}}) < 0$ (entrants earn negative profits) for all $S \subseteq N$, with $\boldsymbol{p}_S^{\mathrm{e}} \leq \boldsymbol{p}_S^{\mathrm{m}}$, $\boldsymbol{y}_S^{\mathrm{e}} \leq \boldsymbol{Q}^S(\boldsymbol{p}_S^{\mathrm{e}}, \boldsymbol{p}_{[S]}^{\mathrm{m}})$ and $\boldsymbol{y}_S^{\mathrm{e}} \neq \boldsymbol{Q}(\boldsymbol{p}^{\mathrm{m}})$ (which excludes the trivial possibility that the entrant will exactly duplicate the entire operation of the incumbent). Then a natural monopoly is said to be sustainable if and only if there is at least one sustainable price vector.

Panzar and Willig (1977) have set forth a number of necessary conditions under which a regulated monopoly would be sustainable in a world with frictionless entry and exit. Among these are that the natural monopoly must produce $\boldsymbol{y}^{\mathrm{m}}$, the output vector associated with $\boldsymbol{p}^{\mathrm{m}}$, at least cost, earn only a normal return on its investment, and operate with a production structure that is subadditive. One further necessary condition requires the following definition:

*Definition (undominated price vector)*

Let $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ and $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n)$ be vectors yielding zero profits for a monopoly. The vector $\boldsymbol{p}$ is *undominated* if there exists no $\hat{\boldsymbol{p}} \neq \boldsymbol{p}$ with $\hat{p}_i \leq p_i$, $\forall i$, and $\hat{p}_i < p_i$ for at least one $i$.

In the single product example of Figure 23.1 there will be only one undominated vector (here a scalar), $p_{\mathrm{B}}$. However, in the multiproduct case there may be an infinite number of such vectors. The two product case is illustrated in Figure 23.3. Here the vectors $\boldsymbol{p}^1$ and $\boldsymbol{p}^2$ are undominated, while $\boldsymbol{p}^3$ is dominated (by $\boldsymbol{p}^1$, for example).

The price vector $\boldsymbol{p}^{\mathrm{m}}$ must also be undominated if it is sustainable. There are other necessary conditions for sustainability regarding economic efficiency and cross subsidy, concepts that will be introduced in subsequent sections. We therefore postpone comments on these until a more appropriate time.[8]

[8] Among other conclusions of Panzar and Willig are some that we will address no further other than to mention them here. First, there is no way to transform an unsustainable monopoly into a sustainable oligopoly by some regulatory act splitting the market among a number of oligopolists. Second, strong demand substitutability among the products offered by the monopolist and product specific economies of scale make it more difficult for a monopoly to be sustainable. As a related point, although it is relatively easy to identify a number of necessary conditions for sustainability, it is also relatively difficult to find rather general sufficient conditions. Vertical integration also introduces a set of interesting problems for sustainability of a natural monopoly; for an analysis of this see Panzar (1980). For another good general reference on sustainability, see Sharkey (1981).

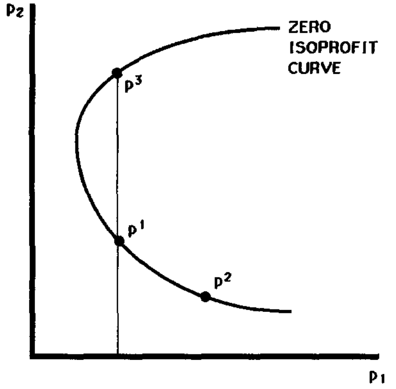

Figure 23.3. Undominated price vector.

Recent research in the characterization of natural monopoly has yielded a number of interesting results on the empirical front as well as theoretically. Much empirical work utilizing modern production and econometric theory has been directed at traditionally regulated industries in the last decade; no small part of this work casts doubt on whether some of the industries historically regulated in the United States do in fact have the structural characteristics of a natural monopoly.[9]

Finally, recent economic research has increasingly emphasized that a structure of "natural monopoly" is *not* sufficient as a basis for regulation. As the next section shows, even if an industry is characterized by natural monopoly in the sense that there is not room for competition *within* a market, under some circumstances competition *for* the market may succeed in allocating resources quite efficiently in the absence of regulation. The theoretical and empirical research on natural monopoly has contributed many economic arguments in support of deregulation and other measures of regulatory reform in a number of American industries since 1970.[10]

[9] See, for example Spady and Friedlaender (1978) and Friedlaender and Spady (1982), who reject the conclusions of earlier studies that show the motor carrier industry to have economies of scale; they show that, when empirical studies of the costs of motor carriers control for the effects of regulation, the structure of the industry is one with essentially constant returns to scale. See also Caves, Christensen and Tretheway (1983) regarding the structure of the airlines industry. For an example of an empirical test of subadditivity (as opposed to economies of scale), see Evans and Heckman (1984).

[10] For a summary of the developments in several recently deregulated industries, see Weiss and Klass (1986).

## 3. Why regulate?

Regulation is a political act. In any particular case there may be a host of possible political and economic answers to the question: Why regulate? Answers are offered by both positive and normative research. In this chapter we will focus on the latter. This is not to diminish the importance of the positive analyses of regulation; that is treated elsewhere in this Handbook.[11] On the contrary, from a political view, perhaps the most significant feature of regulation is that it redistributes income, creating winners and losers, thereby shaping interest groups and coalitions. Thus, it is not surprising that there is a large positive literature on regulation, both in economics and political science, addressing reasons for regulation far broader than natural monopoly. These writings deal both with the creation of regulatory agencies by Congress and with the behavior of regulatory bodies once they are in place.[12]

In focusing instead on normative issues from an economic perspective, we ask a narrower question in this section: When should a natural monopoly be regulated at all? In assessing the effects of regulation, and later in comparing various options for public utility pricing, we need to employ a clear measure of economic benefits to consumers and producers. While such measures do exist, they are often difficult to apply given the kinds of market data that are usually available. The work of Willig (1976) has suggested that the well-known measure of consumer and producer surplus is an adequate approximation in most circumstances, and that is the notion that is adopted in this chapter.[13]

[11] See the chapters of this Handbook by Noll (Chapter 22) and by Joskow and Rose (Chapter 25) for a discussion of many hypotheses about the reasons for and effects of regulation.

[12] See also Joskow and Noll (1981), and Noll and Owen (1983) for excellent discussions of the political economy of regulation. Stigler (1971) describes how regulatory bodies may redistribute income with activities that have effects as powerful as taxation itself. Posner (1974) and Stigler (1975) describe how organized interest groups may "capture" a regulatory agency, either by the initial design of the regulatory process or by other means as time passes. Peltzman (1976) casts the theory of regulation into a supply and demand framework, the supply of regulation being provided by politicians and agencies desiring to maximize vote margins, and the demand from interest groups who would benefit under various regulatory outcomes. Fiorina and Noll (1978) begin with the voters' demand for Congressional facilitation services to explain the congressional demand for administrative activity. Goldberg (1976) suggests that regulation may be viewed as a contract between a regulatory agency (acting as the agent of consumer groups) and regulated firms. Owen and Braeutigam (1978) describe strategies by which the regulatory process may be used to attenuate the rate at which changes in market and technological forces affect individual economic agents, effectively giving agents legal rights to the status quo. See also Hughes (1977) for an interesting historical perspective on the impetus for and transition of regulation from colonial times in the United States (and even earlier in England) until the present.

[13] As a technical point, the use of the usual Marshallian demand schedule observed from market data to measure consumer surplus will be an exact measure of the welfare change associated with a price change for an individual if there are zero income effects. However, Willig (1976) has shown that even if there are nonzero income effects, the measure of consumer surplus obtained from a Marshallian demand schedule may serve to approximate the actual welfare change quite closely.

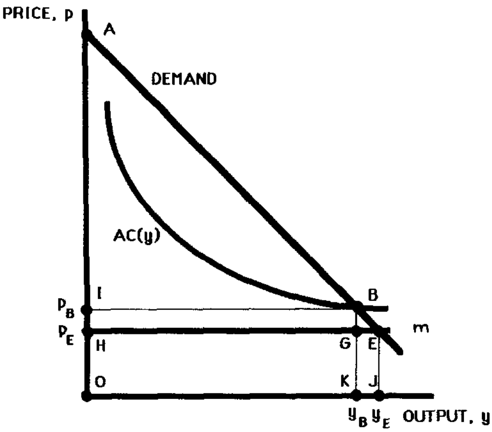

Figure 23.4. First and second best.

Consider now the case of the single product firm operating with economies of scale throughout the operating range of production as in Figure 23.4. For illustrative purposes, assume the cost structure is affine, with a positive fixed cost $F$ and constant marginal cost $m$, so that $C(y) = F + my$. In this example the average cost schedule declines everywhere since marginal cost is less than average cost.

Assume that the firm must charge a uniform tariff (i.e. the same price) to all customers, and that we seek that price that maximizes net economic benefit (alternatively, to maximize economic efficiency) as measured by the standard concept of consumer plus producer surplus.[14] Standard economic principles indicate that net economic benefit will be maximized when the level of output $y = y_E$, with service provided to all customers (and *only* to those customers) who are willing to pay at least as much as the marginal cost of producing $y_E$.[15] In that case the total surplus is represented by the area $AEH$ less the fixed cost $F$.[16] Since this is the maximum surplus that can be generated in the market, a pricing policy that leads to this allocation of resources is termed "first best".[17]

[14] There are a number of classic references dealing with the connection between economic efficiency and regulation. See, among others, Hotelling (1938), Pigou (1920), Taussig (1913), and Turvey (1969). More recent work which summarizes modern developments in the economic theory of regulation include Brown and Sibley (1986), Rees (1984), Sharkey (1982b), and Zajac (1978), all of which are excellent references in the field.

[15] See Turvey (1968, 1969) on the economics of marginal cost pricing.

[16] The fixed cost can be represented in many ways in Figure 23.4; one such measure is the area *IBGH*, so that with marginal cost pricing the total surplus is represented by the area *AEH* less the area *IBGH*.

[17] In the example here we have assumed that the firm must charge the same price for each unit sold in the market. It may be possible to achieve first best without incurring a deficit if the firm can charge different prices to different users (price discrimination) or if different units of output can be sold at different prices (nonlinear tariffs). Both of these alternatives will be addressed below.

However, in the example the firm will not break even with marginal cost pricing. In fact, given the affine cost function $C = F + my$, the profits of the firm are $\pi = -F < 0$. Thus, in order for the firm to remain economically viable, it will have to receive a subsidy of $F$.[18]

Since regulators (particularly in the United States) are not typically endowed with the powers of taxation, they may find themselves faced with a need to find a pricing policy that avoids a deficit for the firm. Without price discrimination or external subsidies to the firm, the regulator might attempt to direct the firm to set that price which maximizes net economic benefit while allowing the firm to remain viable. Since profits are negative at the first best price, there will be a net benefit loss associated with the need to satisfy a breakeven constraint for the firm (i.e. $\pi \geq 0$). Any price higher than $p_B$ will reduce total surplus below the level attainable when $p = p_B$ (the area $ABI$). Thus, the breakeven-constrained optimum (which is termed "second best") occurs at the price $p = p_B$.[19] The welfare loss associated with second best (as opposed to first best) is therefore the area $BGE$ in Figure 23.4. Such an efficiency loss is often called a "deadweight loss".

The point of this discussion is to suggest that in many circumstances it may not be possible to achieve first best without government intervention (e.g. with an external subsidy to the firm), and a program for government intervention may be quite costly. Yet, as we shall show now, it may often be possible to achieve an economic performance near second best without government intervention (even if costs are subadditive over the relevant range of outputs so that it might not be possible to have many firms competing simultaneously within a given market). *Thus, policy-makers may wish to ask whether the deadweight loss at second best is large enough to warrant intervention, especially if some form of competition can be introduced into the market that would lead to second best.*

How might there be an alternative form of competition for such a market? One answer was suggested in a classic article by Demsetz (1968). The focus of Demsetz's article is on competition *for* the market rather than *within* the market. Demsetz pointed out that much of traditional economics is directed at the notion of competition within the marketplace, which may not be possible if there are substantial economies of scale. He suggests that even if competition within the market is not possible, one might still have competition for the right to operate in the market. In other words one could envision bidding among prospective

[18] If the subsidy is provided by the government, then one must take into account not only the welfare effects in the market for $y$, but also the possible welfare losses in other markets that will be taxed in order to provide revenues for the external subsidy provided to keep the firm viable. If the taxes are levied in markets with totally inelastic demands, then the welfare loss from the tax will be zero and $p = m$ in the market for $y$ will be first best. However, if welfare losses occur as a result of the taxation, then $p = m$ may not be optimal.

[19] For more on the theory of second best and optimal taxation see Atkinson and Stiglitz (1980), Diamond and Mirlees (1971), Mirlees (1976), Lipsey and Lancaster (1956–57), and Bohm (1967). Some of these articles deal rather explicitly with the distributional issues that are central to the political debate in taxation.

entrants for the franchise rights to serve the market; this form of rivalry is often called "Demsetz competition", which may be possible if two conditions are satisfied. First, inputs must be available to all bidders in open markets at competitively determined prices. Second, the cost of collusion among bidding rivals must be prohibitively high, so that competitive bidding is in fact the outcome of the bidding process.

Demsetz competition could occur in a variety of circumstances. A relatively simple environment would be the local collection of refuse. In this example companies could bid for the right to collect refuse for a specified period of time, where the "bid" would be the price that the prospective franchisee would charge customers for the collection service, and the company with the lowest bid would win the competition. In this example, the municipal authority need not own the facilities used by the refuse collection company.

A more complicated scenario might involve the right to operate a cable television franchise for a specified time period [see Williamson (1976)]. Here the government might own the facility, but auction off the right to operate the system. The government might charge a fee to the operating company to reflect the social cost of the use of the government-owned facilities.

In the single product environment with a uniform price, Demsetz competition would lead to average cost pricing, since all excess profits would be bid away. Suppose all producers have access to the same technology and could produce efficiently, and that $p^*$ is the lowest price that would allow the firm to break even. One would expect to see bids of $p \geq p^*$, since a lower bid would leave a bidder with negative profits. If the number of bidders is large enough so that the bidding process is in fact competitive, one would expect to see a winning bid of $p^*$, since at that price a producer would earn only normal profits. As noted in the previous section, this is a second best (rather than a first best) outcome.

Demsetz competition is appealing because it suggests competition may be possible even where there are substantial economies of scale, and it is free of the usual regulatory apparatus and regulation-related incentives for firms to behave in an economically inefficient manner.[20] However, the approach is not entirely free of concern. To begin with, while it does lead to second best, there may still be substantial welfare losses relative to first best.

The outcome of Demsetz competition is in effect a contract between a franchisor (e.g. a governmental authority) and a franchisee. Since the franchisee might well adopt the short run strategy of providing the lowest quality service possible once it has won the right to serve, the franchisor may have to specify minimum quality standards for the service to be provided. The question arises:

[20] See, for example, Chapter 24 by David Baron in this Handbook which deals with the design of regulatory institutions and incentives under various regulatory mechanisms, and Chapter 25 (by Paul Joskow and Nancy Rose) which assesses the evidence on the effects of regulation. See also Owen and Braeutigam (1978) and Joskow and Noll (1981).

How does the government set the quality standards? How such standards are set is a problem common to Demsetz competition as well as to traditional regulation; neither approach resolves the problem of specification of quality.

The terms of the contract may be difficult to specify for other reasons. Since the contract may be in force over a period of years, it may be necessary to include procedures to allow for adjustments in terms of service, such as price and quality of service, as conditions in the market change. Some of these contingencies may be relatively easy to incorporate in a written contract, while others may be both unknown and unknowable at the time the franchise is established. The difficulty in writing a contract that includes all sets of possible contingencies is well known. In the context of Demsetz competition this means that a firm that wins the bidding today may attempt to renegotiate its contract tomorrow. The franchisor may then find itself deciding whether to attempt to force compliance, renegotiate, or initiate a new bidding process to find another franchisee. None of these alternatives will be costless.

Another potential difficulty with the use of Demsetz competition arises when the enterprise provides more than one service to its customers. As mentioned earlier, in the single product case the winner might be chosen on the basis of the tariff that franchisee would charge to customers, and that tariff would be second best. However, this selection criterion does not naturally generalize to the case of multiple products. Demsetz competition may lead to a number of different bids which are undominated; recall, for example, that $p^1$ and $p^2$ in Figure 23.3 both yield no excess profits and are undominated. Demsetz competition offers no obvious basis for choice among a number of undominated prices, even though some of these may be quite inefficient relative to others.

A second way in which it may be possible to introduce competition for the marketplace has been formalized with the concept of "contestability" [see Baumol, Panzar and Willig (1982), and also Panzar's Chapter 1 in this Handbook]. Although contestability and Demsetz competition are similar to one another, they are not identical. They key idea in contestability is that competition for the market can lead to second best, even if the cost structure is subadditive over the relevant range of market outputs, *as long as there are no "sunk" costs*. The assumption that there are no sunk costs is one not required by Demsetz competition, but if the additional assumption is satisfied, second best may be achieved through competition for the market *without* the need for a government supervised auction of the sort required in Demsetz competition.

To see how this works, consider first the notions of fixed cost and sunk cost. As defined by Baumol, Panzar and Willig,[21] fixed costs are those that do not vary with output *as long as output is positive*. Let $y$ and $w$ represent respectively

[21] Equations (2) and (3) in the text are respectively contained in Definitions 10A1 and 10A2 of Baumol, Panzar and Willig (1982, pp. 280–281).

vectors of outputs and factor prices, and let $C_{\mathrm{L}}$ be the long run cost of production in (2):

$$C_{\mathrm{L}}(y, w) = \delta F(w) + V(y, w), \quad \text{with } \delta = \begin{cases} 0, & \text{if } y = 0, \\ 1, & \text{if } y > 0. \end{cases} \tag{2}$$

This definition permits fixed costs to exist even in the long run, and $F(w)$ is the magnitude of that fixed cost. Fixed costs are *not* incurred if the firm ceases production.

As the usual argument goes, the long run is long enough for all costs to be avoided if the firm ceases production. However, in the shorter run, say a production period projected $s$ years into the future, a firm may have to make precommitments to incur some costs even if production ceases. If $C(y, w, s)$ is the short run cost function given the production horizon of $s$ years, then $K(w, s)$ are costs sunk for at least $s$ years, if

$$C(y, w, s) = K(w, s) + G(y, w, s), \quad \text{with } G(\mathbf{0}, w, s) = 0. \tag{3}$$

Since a sunk cost cannot be eliminated or avoided for some period of time, even if an enterprise ceases production altogether, during that period sunk costs cannot be viewed as an opportunity cost of the firm.[22]

The idea behind contestability in the single product case is as follows. If no costs are sunk, then firms operating with identical technologies and products would be free to enter the market as they please, charging whatever prices they wished. Any firm charging a price higher than average cost would find itself driven from the market by another firm charging a lower price. The consequence of competition for the market would thus be average cost pricing (and hence second best performance in the market).[23]

[22] In the long run, the usual notion that no costs are sunk means that

$$\lim_{s \to \infty} K(w, s) = 0.$$

[23] In recent years an extended discussion has developed about the meaning of contestability and the extent to which it may be appropriate to employ this concept in connection with real world markets. For example, the theory of contestability (using the notion of sustainability) focuses on prices as decision variables and models potential entrants as evaluating the profitability of entry at the incumbent's pre-entry prices. Some authors have suggested that more complicated forms of the game between entrants and incumbents might be appropriate. Alternative models might include more complicated dynamic aspects of the interactions among potential entrants and the incumbent and the use of quantities as well as prices as decision variables. A detailed discussion of this literature is well beyond the scope of this chapter. For interesting formulations of the rivalry between an incumbent and a potential entrant, see Brock (1983) for suggestions of alternative possible strategies, Dixit (1982) for a treatment of the dynamics of rivalry, Knieps and Vogelsang (1982) for an interpretation of a sustainable industry configuration as a Bertrand equilibrium, and Brock and Scheinkman (1983) for an extension of the traditional Sylos postulate to a multiproduct setting. See also Baumol (1982), Weitzman (1983), Shephard (1984), Schwartz and Reynolds (1984), and Baumol, Panzar and Willig (1984) for further discussions of strategic behavior and the role of fixed and sunk costs as barriers to entry.

Why is a lack of sunk costs critical if competition for the market is to lead to second best? If a firm incurs sunk costs, then $K(w, s) > 0$ in (3). In order for the firm to be willing to enter the market, it must charge a price that generates revenues that cover the variable costs $G(y, w, s)$ as well as the sunk costs. If the firm were assured of the right to operate in the market for $s$ years (a time period long enough to allow it to recover its sunk costs), then it could charge a price equal to average cost $(C(y, w, s)/y)$, and second best could be achieved. But under contestability the firm is not granted a franchise as it would be under Demsetz competition. The firm does not know how long it will be in the market until another firm comes along and tries to undercut its own price, and it therefore would have to charge a price higher than $C(y, w, s)/y$ to protect against the possibility that entry may occur before $s$ years have passed. Consequently, second best pricing will not be achieved under contestability if there are sunk costs.

Furthermore, the sunk costs of the incumbent would be a bygone in the event of entry by a new firm. A prospective entrant would have to contend with rivalry from a firm (the incumbent) with relatively low opportunity costs. Knowing this, an entrant might not sink its own costs in response to relatively high prices charged by an incumbent.

One might expect industries with large capital requirements, especially where the capital cannot easily be moved from one location or one use to another, to have substantial sunk costs. For example, in the railroad industry there are substantial costs associated with way and structure, including the roadbed, which might typically be regarded as sunk. The same might be said for much of the pipeline industry. Industries such as these are therefore not likely to be contestable, although one could still conceivably introduce competition for the market through some other means, such as Demsetz competition.

On the other hand, industries in which capital is highly mobile may be contestable. An example is the airlines industry. Here research has suggested that there may be "economies of density," which means that average costs will decline as more traffic is passed through a given airline network [see Caves, Christensen and Tretheway (1983)]. On the surface, this suggests that it may be efficient for only one firm (or a few firms) to operate *within* some city-pair markets. However, this is not sufficient to conclude that prices and entry in airline markets need be regulated. On the contrary, it has been argued that airline markets are contestable since entry and exit is quite easy, and that there are virtually no sunk costs in the industry [see, for example, Bailey, Graham and Kaplan (1985), and Bailey and Panzar (1981)]. These articles rely on contestability to suggest why deregulation for the airlines was an appropriate policy on economic grounds.[24]

[24] For a further discussion of the role of contestability in public policy concerning antitrust as well as regulation, see Bailey (1981).

Beyond Demsetz competition and contestability, competition can also be introduced in a third way, through Chamberlinian monopolistic competition [see Chamberlin (1962)]. For example, in the transportation sector of the economy monopolistic competition among various modes of transport is often referred to as "intermodel competition". This term is employed to describe the rivalry between railroads, motor carriers, pipelines, and water carriers, all of whom compete for freight traffic. If intermodal competition is strong enough, it might be cited as a basis for deregulation even if one or more of the modes of transport appears to have the structure of a natural monopoly.

Consider a simple example of freight transportation between two points. Suppose that a railroad and a competitive motor carrier industry can provide the required point to point service, and suppose the railroad has the cost structure of a natural monopoly.[25] If the intermodal competition between the railroad and the motor carriers is strong enough to-prevent the railroad from earning supernormal profits (even when the railroad acts as an unconstrained profit-maximizer), then the unregulated market outcome may be very nearly second best in the absence of regulation.[26] In recent years the move toward deregulation of the railroad industry no doubt partially results from pervasive intermodal competition among the railroads and other modes. In fact deregulation of the motor carrier industry in 1980 has led to declining rates in that industry, which further strengthens the extent of the intermodal competition faced by the railroads [see Moore (1986)].

In other industries similar types of competition have occurred. For example, cable television, a once heavily regulated industry, has largely been deregulated, no doubt in part because of heavy competition from over-the-air broadcasting. Currently, there is much discussion over whether oil pipelines should be deregulated. The proponents of deregulation rely on the argument that there is much competition from other transport modes, including, for example, the railroads, that would keep the pipeline industry from earning large excess profits in the absence of price regulation.

In sum, the views of conditions under which it is appropriate to regulate (or deregulate) have changed considerably during the last two decades. A (no doubt highly) simplified comparison of the older and newer views is shown in Figures 23.5 and 23.6. The more traditional view is depicted in Figure 23.5; there the existence of "natural monopoly" (as characterized by economies of scale) was the

[25] This assumption is for the sake of example in the text. A review of the literature on railroad costs is beyond the scope of the current chapter; suffice it to say here that there is mixed evidence on whether railroads operate with economies of scale, although most papers that have addressed the issue of economies of density (which, for a single product railroad, means that average costs will decline as more traffic is passed through a given network) have found evidence that they exist.

[26] For more on the theory of second best with intermodal competition, see Braeutigam (1979). Of course, if railroads have no scale economies in this example, then the unregulated outcome would be first best instead of second best.

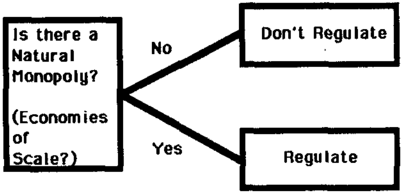

Figure 23.5. Regulation of prices and entry: "traditional" economic justification.

critical factor in determining whether an industry should be regulated. Natural monopoly was taken to preclude competition within the market, and there was very little emphasis on competition for the market as an alternative to regulation.

Although the more current view might be represented in a number of ways, the presentation of Figure 23.6 allows a convenient comparison with the more traditional view. The question of whether a natural monopoly exists is now based on the concept of subadditivity rather than on economies of scale. If there is no natural monopoly and competition within the market is possible (i.e. minimum optimal scale is small relative to the market demand), then a policy of no regulation may be used to reach first best without government intervention.

If a natural monopoly exists, then regulation may still not be warranted. Competition for the market may be possible even if competition within the market is not. If competition for the market is not possible, then some form of government intervention may be required. If competition for the market *is* possible, then performance close to second best might be reached without regulation (through Demsetz competition, contestability, or some form of monopolistic (or intermodal) competition).

It may also be possible to achieve a level of performance better than second best (perhaps even as good as first best) with regulation. One might then compare the deadweight loss at second best with the deadweight loss under a regulatory regime designed to improve performance under government intervention (including an external subsidy, some form of price discrimination, or the use of nonlinear tariffs), keeping in mind the fact that a program of government intervention is not costless.[27] If the deadweight loss at second best is intolerably large (and this requires a value judgment on the part of policy-makers), then government intervention may be warranted. To reiterate, the main point of this exercise is to indicate that even where a natural monopoly exists, government intervention may not be required to achieve economic efficiency for a number of reasons, in contrast with the more traditional view of regulation.

[27] The costs of maintaining a regulatory commission and staff, together with all of the attendant administrative support, can be quite large, as Wiedenbaum (1978) has suggested.

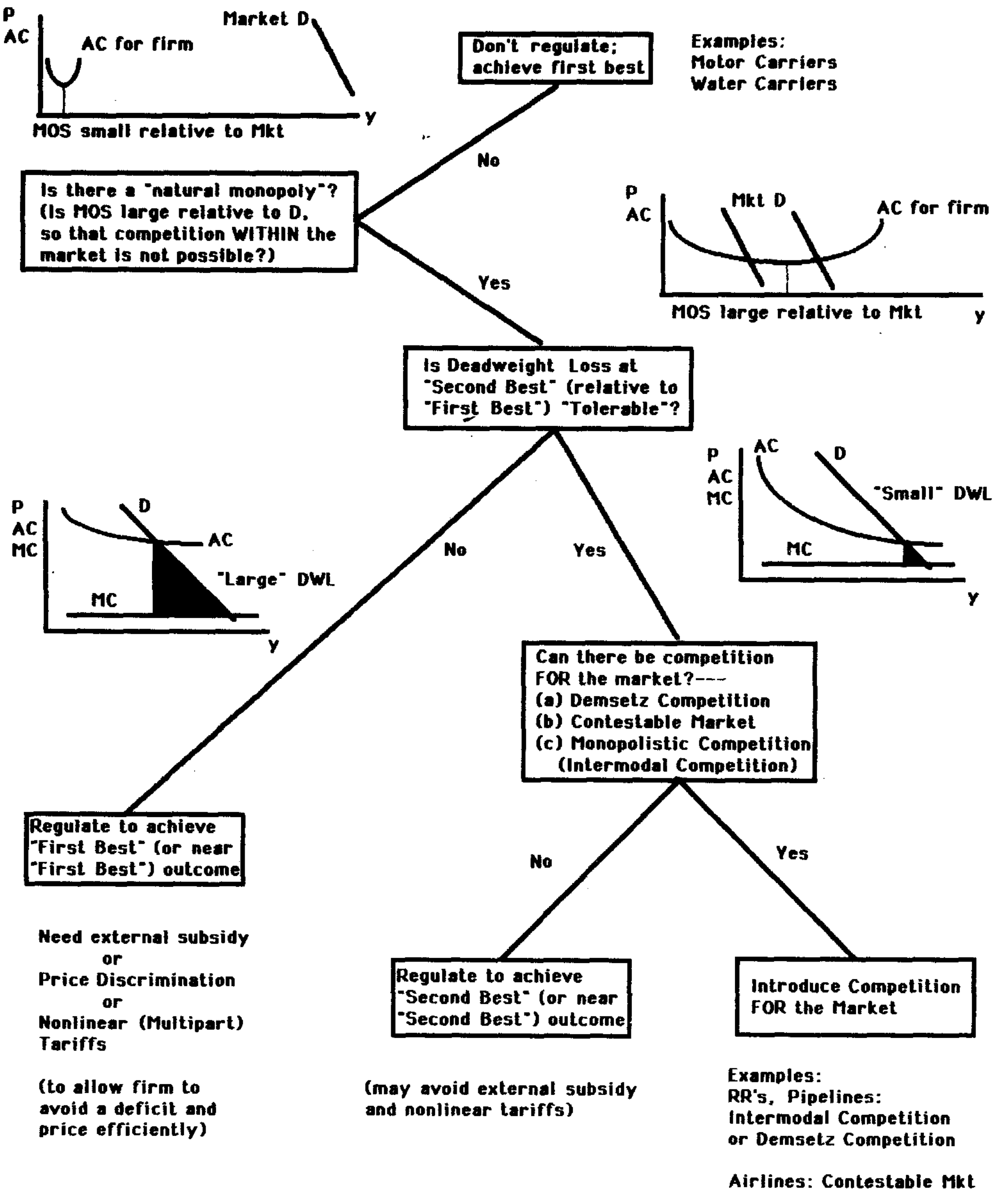
P
AC
AC for firm
Market D
y
MOS small relative to Mkt
Don't regulate; achieve first best
Examples:
Motor Carriers
Water Carriers
No
Is there a "natural monopoly"? (Is MOS large relative to D, so that competition WITHIN the market is not possible?)
Yes
P
AC
Mkt D
AC for firm
MOS large relative to Mkt
y
Is Deadweight Loss at "Second Best" (relative to "First Best") "Tolerable"?
P
AC
MC
D
AC
"Large" DWL
MC
y
No
Yes
P
AC
MC
AC
D
"Small" DWL
MC
y
Can there be competition FOR the market?---
(a) Demsetz Competition
(b) Contestable Market
(c) Monopolistic Competition (Intermodal Competition)
Regulate to achieve "First Best" (or near "First Best") outcome
No
Yes
Need external subsidy
or
Price Discrimination
or
Nonlinear (Multipart) Tariffs
(to allow firm to avoid a deficit and price efficiently)
Regulate to achieve "Second Best" (or near "Second Best") outcome
(may avoid external subsidy and nonlinear tariffs)
Introduce Competition FOR the Market
Examples:
RR's, Pipelines:
Intermodal Competition
or Demsetz Competition
Airlines: Contestable Mkt

Figure 23.6. Policy "roadmap" for regulation.

To this point we have addressed one facet of the optimal policy toward natural monopoly, namely whether to regulate at all or rely on some form of competition instead. We now turn to optimal strategies where regulation is selected as the appropriate policy. The menu of possible regulatory controls over price and entry is a rich one. The balance of this chapter will discuss some of those controls.

## 4. Pricing alternatives: Basic concepts

If regulation is undertaken as a response to the natural monopoly problem, there are several courses of action that might be followed by the regulator with respect to pricing. Of course, changes in prices have both distributive and allocative effects. In this section we will focus on the latter, that is, pricing policies designed to achieve economic efficiency.

As Figure 23.6 indicates, regulation might be implemented for a variety of reasons related to economic efficiency. For example, a natural monopoly might be regulated because no form of competition for the market is viable; here prices might be regulated to reduce the deadweight loss associated with the unregulated monopoly price, perhaps to a level associated with either second best or first best. Or, even if second best could be achieved through competition for the market, policy-makers might determine that the deadweight loss associated with second best is intolerably large, in which case regulation might be introduced to increase efficiency (perhaps even to reach first best).

Section 3 presented the basic dilemma of marginal cost pricing with a natural monopoly. In particular Figure 23.4 illustrated why marginal cost pricing will lead to a deficit for a firm operating with economies of scale if all units of output are sold at marginal cost.[28] In this case the firm will not be "revenue adequate", and would therefore require an externally provided subsidy to cover the deficit if it is to continue production. With economies of scale and a single price charged for all units of output, one can achieve first best only if an external subsidy is provided, and avoid such a subsidy only by incurring a deadweight loss. This tension between economic efficiency and revenue adequacy provides a focus for much of the literature on regulated industries.

However, it turns out that there may be other ways to achieve greater efficiency than at second best (perhaps even reach first best) without an external subsidy when there are economies of scale throughout the relevant operating range. To see this, recall that the earlier discussion of Figure 23.4 assumed that the same price is charged for all units of output sold in the market. Restated, this means

[28] Here one should keep in mind the distinction between economies of scale and subadditivity. If natural monopoly were characterized by a subadditive cost structure, but not by economies of scale over the relevant operating range, as in Figure 23.2, then marginal cost (i.e. first best) pricing would allow the firm to breakeven or even to earn some extranormal profit.

that (1) each unit purchased by an individual customer is sold at the same (i.e. uniform) price and that (2) the price per unit is the same for all customers (i.e. there is no price discrimination over customers).

### *4.1. Price discrimination (differential pricing)*

The foregoing discussion suggests that there are two ways one might further improve economic efficiency by departing from the rather restrictive assumption that the same price is charged for all units of output sold in the market. One way would be to engage in some form of *price discrimination*, sometimes referred to as *differential pricing*. As these terms suggest, a regulator could charge different prices to different customers in the market, even if each customer pays the same price for all of the units he purchases. In the simplest instance, suppose that customer $i$ must pay $p_i$ for every unit of service he purchases, and that customer $j$ must pay $p_j$ for every unit of service he purchases. Differential pricing means that $p_i \neq p_j$ for some customers $i$ and $j$. Peak load pricing and Ramsey pricing schemes fall into this category and will be discussed in greater detail in Sections 5 and 6.

Price discrimination is, of course, a subject that has received much attention in both regulated and unregulated industries. Much of that discussion surrounds the legality of the practice [see, for example, Scherer (1980, chs. 11 and 12)]. A discussion of the legality of price discrimination is not our focus here. We should observe that even if regulators wish to allow or impose price discrimination, it still may not be possible for economic reasons. As is well known, in order for differential pricing to be feasible the seller must be able to identify the price each customer (or at least different groups of customers) would be willing to pay for the service. Furthermore, resale must not be possible for either legal or technological reasons, so that a customer could not purchase the service at a low price and then sell it to another customer at a higher price. If resale is possible, arbitrage will work to eliminate price discrimination so that all customers would face the same price in the market.

To see how differential pricing might be used to improve economic efficiency while allowing the firm to avoid a deficit, consider again Figure 23.4, where the firm operates with the affine cost structure $C = F + my$. Suppose the firm knows how much each consumer is willing to pay for the service, and that resale is impossible. Now let the firm charge a price equal to $p_{\mathrm{B}}$ to all customers who would be willing to pay a price greater than or equal to $p_{\mathrm{B}}$, i.e. to all customers located to the left of point $B$ on the demand schedule. Call these "type I customers". Then let the firm charge a price equal to $p_{\mathrm{E}}$ to each of the customers who would be willing to pay a price greater than or equal to $p_{\mathrm{E}}$, but not more than $p_{\mathrm{B}}$. Call these "type II customers".

What would be the consequences of such a schedule? The revenues generated by the type I customers would cover not only the variable costs of producing $y_{\mathrm{B}}$ units, but also all of the fixed costs $F$. [Observe that $p_{\mathrm{B}} = C(y_{\mathrm{B}})/y_{\mathrm{B}}$, which means that $y_{\mathrm{B}} p_{\mathrm{B}} = F + m y_{\mathrm{B}}$.] The revenues generated by the type II customers would then cover just the variable costs from providing $(y_{\mathrm{E}} - y_{\mathrm{B}})$ units of service. Therefore, one consequence of the suggested schedule is that total costs would then cover total revenues, and there would be no need for an external subsidy to keep the firm viable. Also note that every customer who is willing to pay an amount at least equal to the marginal cost of producing the service receives it, while service is not provided to customers who are not willing to pay at least the marginal cost of production. Thus, a second consequence of the suggested schedule is that it is "first best" or economically efficient. It should also be noted that the proposed schedule leaves the firm with no extranormal profits (producer surplus), since total revenues exactly equal total costs in the example, while consumer surplus would be equal to the sum of the areas $ABI$ and $BGE$.

One could envision many other possible discriminatory tariff schedules that would accomplish the same objectives (achieving first best without an external subsidy). As a simple example, suppose each customer desires only one unit, and suppose the firm is allowed and able to price discriminate perfectly so that it charges each customer a price equal to the maximum amount that customer is willing to pay for the unit purchased. Consumer surplus is zero under this pricing schedule since each consumer is paying the maximum amount he is willing to pay in order to get the service. In the example of Figure 23.4, the firm's revenues would then equal the area represented by $AEJO$, while the costs of production would be the sum of the areas $IBKO$ and $GEJK$. Again, the firm remains viable (and in fact earns a producer surplus equal to the sum of the areas $ABI$ and $BGE$). Thus, total surplus (the sum of consumer and producer surplus) is as great as it was under the imperfectly discriminating tariff schedule that charged $p_{\mathrm{B}}$ to type I customers and $p_{\mathrm{E}}$ to type II customers, and once again first best is achieved for the same reasons as given in that earlier example. Of course, the division of the total surplus is strikingly different under the two schedules, with consumers receiving it all in the first example and producers receiving it all in the second. With still other forms of price discrimination it would be possible to achieve other distributions of the total surplus under a first best pricing structure.

### 4.2. Nonlinear outlay schedules (nonlinear tariffs)

The second way of departing from the assumption that the same price is charged for all units of output sold would be to charge an individual customer an amount per unit purchased that varies with the total quantity he purchases. This kind of pricing is often referred to as a *nonlinear outlay schedule*, or sometimes a

***nonlinear tariff.*** The difference between a linear and a nonlinear outlay schedule can be illustrated easily. Suppose that customer $i$ must pay $p_i$ for every unit of service he purchases, and that he purchases $y_i$ units. His total outlay (expenditure) is $p_i y_i$, so that *the average outlay per unit purchases is constant.* By direct analogy, a *nonlinear outlay schedule* is one in which *the average outlay is not constant as the number of units purchased varies.*[29]

One might suspect that there are many possible ways of structuring nonlinear tariff schedules. Indeed this is so, as will be discussed in greater length in Section 7. For now we offer only a simple example of such a tariff. Consider the so-called two-part tariff; as the name suggests, the tariff has two parts here, a "fixed" and a "variable" component. Suppose, for example, there are $N$ identical consumers in the market, and that the firm operates with the affine cost structure $C = F + my$. One could envision a tariff structure that would assess each customer a fixed charge $e$ (per month), where $e = F/N$ is to be paid regardless of the number of units actually purchased. In addition customers would be required to pay a variable charge equal to $m$ for each unit actually purchased. Thus, the total expenditure by a customer would be $e + my$, which is an affine tariff schedule. First best is achieved since each additional unit consumed is priced at marginal cost. In addition the firm would remain financially viable since the total revenues would be $N(e + my) = F + Nmy$.

The reader may (correctly) suspect that income effects may introduce complexities in the way such tariffs are structured if economic efficiency is to be achieved; we address these effects in Section 7. In fact, nonlinear tariffs may involve more than two parts as in the previous example. The main point of the examples in this section is to illustrate that nonlinear tariff structures can be useful as a means of achieving greater efficiency without external subsidies.[30]

### 4.3. *The common cost problem in the multiproduct firm*

We have now suggested several ways in which one might improve economic efficiency by departing from a single price for all units of output sold in the market. The problems discussed thus far are simplified in one very important respect: the firm has been assumed to produce only one product. The problem of pricing becomes even more difficult when there is more than one output produced by the firm.

[29] One possible source of confusion in the taxonomy here should be pointed out. Since a linear outlay schedule is defined as one in which average outlay is constant, it follows trivially that total outlay is linear in output. However, a nonlinear outlay structure may also be linear in output; in particular, with the affine structure referenced in the text expenditures are linear in output. The important point is that *average* (not total) outly is not constant with respect to output purchased.

[30] As will be indicated in Section 7, nonlinear tariffs may not always lead to first best, but nonlinear tariffs can be used to increase economic efficiency relative to second best even when first best is not achieved.

To see this consider a firm which produces two products whose levels of output are respectively $y_1$ and $y_2$. Let the marginal costs of production for the services be constant and respectively $m_1$ and $m_2$, and suppose there is a fixed cost of production $F$. This describes a simple multiproduct affine cost function where the total costs are $C = F + m_1 y_1 + m_2 y_2$.

The fixed cost is said to be "common" to both services. In other words, it is a cost shared in the production of $y_1$ and $y_2$. The presence of such a common cost poses a particularly difficult problem for regulators trying to set prices so that the firm can break even. Assume that the firm must price each service uniformly, so that purchasers of service $i$ will all pay a price per unit equal to $p_i$ for that service. As in the single product affine cost case, it is clear that the firm cannot break even with marginal cost pricing. If $p_1 = m_1$ and $p_2 = m_2$, the profits of the firm will be negative (in fact, profits are $\pi = -F$).

The question then becomes: How might the regulator set rates so that the firm breaks even? This is an age-old question that has been examined in many contexts in the economic literature as well as in regulatory proceedings [see, for example, Taussig (1913), Pigou (1920) and Clark (1923) for excellent early treatises on this subject].[31]

For many years regulators had relatively little in terms of economic theory to guide their decisions in ratemaking in the face of common costs. In practice regulatory authorities such as the Interstate Commerce Commission and the Federal Communications Commission historically have determined tariffs (rates) using so-called fully distributed (fully allocated) costs, which we shall refer to here as FDC pricing. We discuss this briefly here to contrast this often used regulatory approach with those based on economic efficiency to be discussed in subsequent sections.

Under FDC pricing, as a first step regulators do (somehow) allocate the common costs among the individual services. In other words, each service is assigned a fraction $f_i$ of the common costs, so that the share of common costs for service $i$ is $f_i F$. (The fractions $f_i$ must add to 1 if the costs are fully allocated; in our example $f_1 + f_2 = 1$.) Each service is then priced so that the revenues generated from that service will cover all of the costs directly attributable to that service plus the assigned portion of the common costs (again, in the example $p_i y_i = f_i F + m_i y_i$ for $i = 1, 2$).

The issue of pricing then critically depends on the way in which the allocators $(f_i)$ are set. In principle, of course, there are an infinite number of ways one can allocate the common costs since there are an infinite number of ways one select $f_1$ and $f_2$ to sum to unity. In practice regulators have sometimes allocated common costs in proportion to (1) gross revenues (so that $f_1/f_2 = p_1 y_1 / p_2 y_2$), or

[31] See also Kahn (1970), Baumol, Panzar and Willig (1982), Brown and Sibley (1986), Faulhaber (1975), Faulhaber and Levinson (1981), Owen and Braeutigam (1978), Sharkey (1982a, 1982b), Weil (1968) and Zajac (1978) for a few among many references on the subject of the common cost problem. Some of these will be discussed further below.

(2) physical output levels (so that $f_1/f_2 = y_1/y_2$) or (3) directly attributable costs (so that $f_1/f_2 = m_1y_1/m_2y_2$).[32]

Without extending the discussion of this practice, it is rather immediately apparent that there are many potential problems with FDC pricing.[33] Regarding the arbitrariness of the method, Friedlaender (1969) notes: "Various means of prorating the common or joint costs can be used, but all of them have an arbitrary element and hence are dangerous to use in prescribing rates." It may involve circular reasoning since prices, revenues or output levels are used to determine the allocators which are used in turn to set prices. It may also lead to prices which are dominated in the sense defined in Section 2.[34] And, with respect to a point that is central to this chapter, FDC pricing will lead to prices which are in general economically inefficient, which is not surprising given the fact that the practice focuses heavily on cost and little on conditions of demand (including demand elasticities) which are important in determining the size of the deadweight losses from any pricing policy.

In connection with the common cost problem it is worthwhile to comment on a relatively new line of research called the "axiomatic" approach to common cost allocation. This work is not based on economic efficiency in its treatment of the problem (as is Ramsey pricing, discussed in Section 6); neither does it stem from an attempt to find prices which are free of cross subsidy (various notions of which are covered in Section 8). Instead, it begins with a set of features desired in a cost allocation scheme, represents them axiomatically, and derives pricing rules consistent with these desiderata. The exact specification of the axioms depends on the cost structure, and in particular whether there are fixed costs or not.

Mirman, Samet and Tauman (1983) have presented six axioms for the allocation of common costs, and analyzed pricing rules that satisfy these axioms for the case in which the firm may be operating with fixed costs. The cost function may be written $C = F + V(y)$, where $F$ is a fixed cost and $V$ is a variable cost function dependent on the level of outputs $y = (y_1, y_2, \ldots, y_n)$.[35] (This allows for the possibility that the relevant horizon for the firm or the regulator is the

[32] Friedlaender (1969, p. 32) noted that the ICC had often allocated common costs between freight and passenger services "on the basis of revenues derived from each source", and (p. 133) "the most usual basis of prorating [costs among freight services] is on the basis of ton-miles" (brackets added); of course, in this case the outputs must have a common measure of output, such as ton-miles of various types of freight (this practice would make no sense for allocating common costs among, for example, passenger service and freight service). Kahn (1970, p. 151) notes that allocation according to attributable costs has been used to some extent in the transportation industry.

[33] See Braeutigam (1980) for a more detailed analysis of FDC pricing.

[34] Sweeney (1982) considers the case of a multiproduct firm which provides some of its services in a competitive market and others in a regulated monopolistic setting. Sweeney shows that for FDC pricing rules with allocators that monotonically increase in output, prices will be on a dominated part of the isoprofit locus.

[35] As is the normal case in this chapter, factor prices are suppressed in the representation of the cost function since they are assumed constant.

short run, during which it may not be possible to adjust all factors of production to the levels that would be efficient in the long run.)

Briefly the six axioms require that (1) the prices resulting from the allocation mechanism generate revenues sufficient to cover total costs; (2) if the units of measurement for the commodities are rescaled, the prices measured with the new dimensions should be rescaled accordingly; (3) if for some subset $S$ of outputs total cost depends only on the sum of the levels of the outputs in $S$, then the prices of any two outputs in $S$ should be the same (this implies that outputs with the same marginal costs should have equal prices);[36] (4) if $C$ and $\hat{C}$ are two different cost structures with $C(\mathbf{0}) \geq \hat{C}(\mathbf{0})$ and $(C - \hat{C})$ increasing as outputs increase, then prices should be higher under $C$ than under $\hat{C}$; (5) if $V(y)$ can be written as a sum of the variable costs from $k = 1, \ldots, K$ stages of production so that $V(y) = V_1(y) + V_2(y) + \cdots + V_K(y)$, then the mechanism should allocate a fraction of the common cost $f_k F$ to each stage $k$, with $\sum f_k = 1$ so that all of the common costs are allocated; and (6) if for any two stages $i$ and $j$ described in (5) it is true that $V_i(y) > V_j(y)$, then $f_i > f_j$, so that the size of the allocation is higher when variable costs are higher.[37]

Mirman, Samet and Tauman show that the only pricing rule consistent with the six axioms is one based on the Aumann–Shapley price for each service. In the case of a general cost function, there is no obvious interpretation of this price, and we do not present a detailed statement of the pricing rule here. However, there is a case of special interest worth noting. If the cost structure can be written in an additively separable fashion $C = F + \sum_i V_i(y_i)$, then the only price rule satisfying the axioms is the allocation of common costs in proportion to directly attributable costs, which happens to be one of the fully distributed cost mechanisms discussed earlier in this section.[38] This finding is of particular interest. While the additively separable cost structure is simplistic, it has been used by some regulatory commissions in the past.[39]

We now focus on economically efficient pricing schemes that might be used where shared costs exist. The next section considers a set of pricing policies that rely on differential pricing, commonly known as peak load pricing.

[36] The third axiom makes it clear that the "axiomatic approach" bears no necessary relationship to pricing which is economically efficient. As will be clear from the discussion of Ramsey pricing in section 6, if two services have identical marginal costs, an economically efficient price will be greater for the product with the more inelastic demand.

[37] Under some circumstances a single axiom of additivity can replace the last two listed in the text (i.e. axioms (5) and (6) in the text) if the firm is operating on its long run cost function; see Mirman and Tauman (1982) and Samet and Tauman (1982) for more on this point.

[38] Braeutigam (1980) has shown that when the regulated firm operates at zero profit, two of the FDC mechanisms discussed above are equivalent. These are the allocation of common costs (1) in proportion to directly attributable costs and (2) in proportion to gross revenues.

[39] For example, Friedlaender (1969) has discussed the use of such a cost structure (Rail Form A) by the Interstate Commerce Commission in setting railroad rates.

## 5. Peak load pricing

The term "peak load" suggests a problem faced by many utilities, and one which has been treated widely in the literature. There are three essential features of the traditional peak load problem: (1) the firm must provide service over a number of time periods having perhaps greatly different demand schedules, (2) the firm must choose a single plant size (capacity) to be in place during all of the time periods over which production takes place, and (3) output is nonstorable.[40] A large number of formal models have been developed in the literature to characterize economically efficient prices for the peak load problem, all of which have led to prices that vary across time in some way. Thus, peak load pricing schemes are a form of price discrimination across time periods.

In regulatory settings the issue of peak load pricing often revolves around the fact that the plant is shared by users of all time periods. The question to be resolved is: What share of the cost of the plant should be borne by users in the various time periods? The most famous classical economic model of the peak load problem is that of Steiner (1957).[41] That work generated optimal pricing rules that are commonly known even to regulatory commissioners today, including the widely cited principle that all of the plant costs should be loaded on to the peak load period. But as we shall see, the latter conclusion is one which is very sensitive to the nature of the technology and demands.

To compare a few of the basic peak load formulations in the literature, consider the following framework. Assume the production period (e.g. a day) is divided into $T$ equal parts, indexed by $t = 1, \ldots, T$.[42] Assume that $x_t$ units of a single variable input are used in period $t$, and that $k$ represents the amount of the capital input which is chosen for all periods. Let $y_t = f(x_t, k)$ be the production function for period $t$, relating the output in that period $y_t$ to the inputs. The nature of this production function will be crucial to the form of the peak load pricing rules, and will be specified in detail in the models discussed below. Finally, let $p_t = p_t(y_t)$ represent the (inverse) demand schedule in period $t$. The demand schedule is downward sloping, so that $p_t'(y_t) < 0$.[43]

[40] If output is storable without cost, then a firm could produce and store more than is demanded in an off peak period, and then use the stored output to serve the higher demand in the peak period. This would allow the firm to pool production over all of the time periods, effectively eliminating the peak load problem. Of course, one could introduce storage costs which are positive, and still retain the essence of the peak load problem examined in this section.

[41] This classical formulation of the peak load problem is also discussed at length in Kahn (1970, ch. 5).

[42] The assumption that the production period is divided into equal parts is not necessary, but does facilitate exposition.

[43] The prime symbol will be used to denote derivatives where that can be done unambiguously in the text; thus $p_t'(y_t) \equiv \partial p_t / \partial y_t$.

Consider first the traditional formulation of Steiner. The production function has a Leontief structure, so that $y_t = f(x_t, k) = \min(x_t/a, k)$, with the constant $a > 0$. One can represent this production structure in terms of a cost function. Let $\tilde{b}$ be the cost of a unit of the variable factor, which is assumed here to be the same in each period. Then the total variable cost incurred in period $t$ will be $\tilde{b}x_t = \tilde{b}ay_t$. For simplicity in notation, let $b = \tilde{b}a$, so that the period $t$ variable cost is $by_t$. Let $\beta$ be the (rental) cost of a unit of capital over all time periods $t+1, \ldots, T$. Assume the firm must meet all demand, so that capital must be chosen to be $k = \max_t y_t$. Then the total cost for the firm will be

$$C = b \sum_{t=1}^{T} y_t + \beta \max_j y_j. \tag{4}$$

Suppose that gross economic benefit can be represented as $A(y_1, y_2, \ldots, y_T)$.[44] Then net economic benefit, $W$, can be written as (5):

$$W = A(y_1, y_2, \ldots, y_T) - C. \tag{5}$$

In *off peak* periods (in which $y_t < \max_j y_j$) the first order necessary conditions for an interior optimum (in which $y_t > 0$) of (5) would be

$$\partial W/\partial y_t = p_t - b = 0, \quad \text{for } y_t < \max_j y_j, \tag{6}$$

which implies that $p_t = b$. In other words, in off peak periods, users will be required to pay only for the variable costs of production, with no revenues being contributed toward the costs of capacity for the enterprise. In the *peak* period (in which $y_t = \max_j y_j$) the first order condition for an interior optimum of (5) would be

$$\partial W/\partial y_t = p_t - b + \beta = 0, \quad \text{for } y_t = \max_j y_j, \tag{7}$$

which implies that $p_t = b + \beta$. In other words, in peak periods, users will be required to pay for the variable costs of production *plus* the capacity costs of the enterprise.

An example using the peak load pricing principles with this Leontief technology is depicted in Figure 23.7. In the figure, the day is divided into three time

[44] One could write $A$ in terms of the usual consumer surplus integrals:

$$A(y_1, y_2, \ldots, y_T) \equiv \sum_{t=1}^{T} \int_0^{y_t} p_t(\xi)\,\mathrm{d}\xi.$$

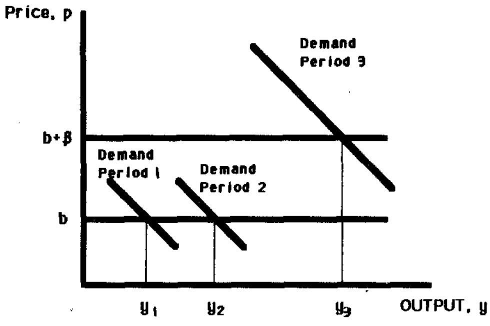

Figure 23.7. Peak load pricing with a Leontief technology.

periods, daytime ($y_3$), evening ($y_2$), and night ($y_1$). The daytime period is the peak period, with the other two being off peak periods. The Steiner model would indicate that the off peak users would pay a price of $b$, while the daytime users would pay $b + \beta$, since revenues generated from daytime service would have to cover variable costs and plant costs.

Note that in this example all of the costs of the enterprise are covered by revenues generated by the three classes of users. Revenues from the daytime users are $y_3(b + \beta)$; for evening and night users the revenues are respectively $by_2$ and $by_1$, so that all of the costs in (4) are covered. Furthermore, each class of users is paying a price equal to the marginal cost of production, since $\partial C/\partial y_1 = b$, $\partial C/\partial y_2 = b$, and $\partial C/\partial y_3 = b + \beta$, which includes the marginal cost of capacity expansion if the peak period production is increased. Therefore first best and revenue adequacy can be achieved simultaneously with this peak load pricing scheme.[45]

The peak load pricing problem can also be formulated in terms of a neoclassical production function instead of a Leontief technology. As Panzar (1976) has shown, somewhat different results follow. Again let the production function for period $t$ be $y_t = f(x_t, k)$, where, as before, $k$ is fixed across all time periods. Let $f$ be twice differentiable and quasiconcave in $x_t$ and $k$, with the partial derivatives $\partial f/\partial x_t > 0$, $\partial f/\partial k > 0$, $\partial^2 f/\partial x_t^2 < 0$, and $\partial^2 f/\partial k^2 < 0$, so that the marginal products of capital and the variable factor are positive and decreasing.

One can write the variable cost function associated with $f$, which minimizes the variable cost of producing any specified $y_t$ given the level of $k$ in place. Let the variable cost function in period $t$ be denoted by $V(y_t, b, k)$, where $b$ is the (parametric) price of a unit of the variable factor, and assume the variable cost

[45] In this example, of course, the production structure exhibits constant returns to scale, since a doubling of outputs will lead to a doubling of total production costs. Thus, it is not surprising that marginal cost pricing will lead to revenue adequacy.

function has the standard derivative properties $\partial V/\partial y_t > 0$ (marginal variable cost is positive), $\partial V/\partial k < 0$ ($k$ and $x_t$ are substitutes in production), and $\partial^2 V/\partial y_t^2 > 0$ (for fixed $k$ the marginal cost is increasing in output). In addition let $V(0, b, k) = 0$ (variable costs are zero when output is zero).

Then with the same demand structure as used in the Leontief model, net economic benefit, $W$, can be written as (8):

$$W = A(y_1, y_2, \dots, y_T) - \sum_{t=1}^{T} V(y_t, b, k) - \beta k. \tag{8}$$

Let the output levels (or, equivalently, prices) and the level of capital be chosen to maximize $W$. At an interior optimum ($y_t > 0$ and $k > 0$), first order conditions require that (1) $p_t = \partial V/\partial y_t$, and (2) $\sum_t \partial V(y_t, b, k)/\partial k = -\beta$. The second condition shows that capital is employed until the total variable cost savings from an added unit of capital equals the cost of that added unit of capital. The first condition indicates that the price equals the marginal variable cost in each period. Here, too, with constant returns to scale, marginal cost pricing will lead to revenue adequacy.

Finally, recall that $\partial^2 V/\partial y_t^2 > 0$, which means that marginal costs are rising for any given size of plant. Consider any two periods, and denote them by $t = 1$ and $t = 2$ without loss of generality. Suppose $y_2 > y_1$. Then $p_1 = \partial V/\partial y_1 < p_2 = \partial V/\partial y_2$. Thus, prices will not be equal in periods with different demands; in fact price will be higher in the period with the higher demand.

Still a third possible technology, having elements of both the Leontief and the neoclassical production structure, is examined by Waverman (1975), with some interesting conclusions. Assume that *any* output-variable factor ratio can be chosen, but that once the ratio is chosen, it is then applicable in all periods. (By contrast, the Leontief technology assumes that the ratio $y_t/x_t$ is fixed and not freely chosen, while the neoclassical structure allows the ratio to be chosen at different levels in different time periods.) To illustrate this formulation, consider a three period model, with period three having the peak demand. As before, assume there is a single variable factor, whose levels are $x_1$, $x_2$, and $x_3$ in the three periods respectively, and whose unit price is $\tilde{b}$. With the same demand structure as used previously, assume the firm chooses $y_t$, $x_t$, and $k$ to maximize net economic benefit, $W$, as follows:

$$\max_{(y_t, x_t, k)} W = A(y_1, y_2, y_3) - \tilde{b} \sum_{t=1}^{T} x_t - \beta k \tag{9}$$

$$\text{subject to} \quad x_1/y_1 = x_2/y_2 = x_3/y_3$$

$$\text{and } f(x_3, k) \geq y_3.$$

Waverman's analysis indicates that in the two off peak periods prices will be equal to one another and equal to the (marginal) variable cost of production, a conclusion much like that of the Steiner model. Furthermore, the ratio of the peak price to the off peak price does depend on the distribution of outputs across time periods in the Waverman model, whereas in the Steiner model that ratio does not depend on the distribution of output.

Without belaboring these models further, it can be concluded that the optimal pricing policy does depend on the nature of the underlying technology, as suggested earlier. This has an important implication for applications of economic theory to peak load problems; one might be advised to examine the properties of estimated cost or production functions to find out what kind of technology exists before advocating any particular optimal pricing rule.

Finally, there are a number of other articles that address other problems related to peak load pricing. For example, Bailey and White (1974) show that a peak period price can actually be less than the price in an off peak period under a variety of circumstances. Among others these include pricing for a welfare maximizing firm operating with a decreasing average cost in production. Here the firm needs to satisfy a breakeven constraint while maximizing welfare over all periods. For example, a higher off peak price might result in the off peak period if the demand in the off peak period is inelastic relative to the elasticity of demand in the peak period.[46] One must also be careful when trying to identify which period is a peak period; when one moves from a high price to a lower price, demand schedules for two periods may intersect one another, so that the peak period may change. Carlton (1977) has addressed the problem of peak load pricing when demands are stochastic, in contrast to the survey of this section in which demands are known with certainty. Crew and Kleindorfer (1976) have introduced the possibility that firms may operate with diverse technologies, including several types of plants, as is often observed in industries such as the electric utility industry. Additional discussions of peak load pricing models can be found in Littlechild (1970), which applies the theory to the telephone industry, in Brown and Sibley (1986) and in Rees (1984).

## 6. Ramsey pricing

The discussion of peak load pricing in the previous section indicated how differential pricing might be used to improve economic efficiency when a single plant size must be chosen to provide service over more than one time period. The nonstorability of the service and the variation in demand across time periods were identified as crucial aspects of the peak load problem. In the standard

[46] The addition of the breakeven constraint in the face of increasing returns to scale is a problem that will be discussed below in greater detail in the section on Ramsey pricing, where economically efficient prices depend on the elasticities of demand as in Bailey and White (1974).

presentation of the peak load problem, returns to scale are constant; thus optimal pricing schemes lead to first best while allowing the firm to break even.

Let us now turn to the case in which the firm is unable to break even when a uniform price is set equal to marginal cost for each of the services offered by the firm. The outputs of the firm might be essentially the same product provided in different periods (as with electricity in the peak load case), or, unlike the peak load problem, they might be services which are entirely different from one another (e.g. passenger and freight transportation services). Suppose the regulator has determined that the firm (1) charge a uniform price for each of its services, and (2) price its services so that it breaks even without an external subsidy, i.e. the firm must remain viable with no subsidy from the government or from some other source outside the firm. Under these circumstances the firm will need to charge prices that deviate from marginal costs in some or all of its markets in order to avoid a deficit.

In Section 3 we indicated how a single product monopoly would set the price in order to maximize economic efficiency while allowing the firm to avoid negative profits. We showed that this problem of second best was solved by pricing at average cost for the single product firm because no greater net economic benefit can be achieved if the breakeven constraint for the firm is to be satisfied. Recall that for any price less than average cost, the firm will incur a deficit, which violates the breakeven constraint. For any price greater than average cost, the firm will remain profitable, but the size of the deadweight loss will be larger than when price equals average cost. As suggested in Section 3, the second best price can be viewed as simultaneously maximizing net economic benefits (total surplus) and minimizing the deadweight loss given the constraint on non-negativity of profits for the firm.

The notion of second best pricing becomes more complicated for the case of the multiproduct firm. In general the concept of average cost will not be well defined for a multiproduct technology; if there are shared costs of production, in the sense defined in Section 4, then there is no unambiguous way to allocate the common costs. Thus, there is no clear way to determine an economically meaningful measure of the average cost associated with each service.

The name "Ramsey pricing" stems from the work of the English economist Frank Ramsey, who developed the concept in the context of optimal taxation in 1927 [see Ramsey (1927)]. It was later extended to the problem of public monopolies by Boiteux [see the original version in French, Boiteux (1956) and the English language version, Boiteux (1971)], and further developed by Baumol and Bradford (1970).[47]

To facilitate the exposition, we adopt the following notation. Consider the case of the $N$ product firm, where $y_i$ is the level of output of the $i$th service produced by the firm, $i = 1, \ldots, N$. Let $p_i$ be the price of the $i$th output, $y$ the vector of

[47] See also Sorenson, Tschirhart and Winston (1978).

outputs $(y_1, y_2, \ldots, y_N)$, and $\boldsymbol{p}$ the vector $(p_1, p_2, \ldots, p_N)$. Let $y_i(\boldsymbol{p})$ be the demand schedule for the $i$th service, $i = 1, \ldots, N$, and $\psi(\boldsymbol{p})$ be the consumer surplus at the price vector $\boldsymbol{p}$.[48] Let $w_i$ be the factor price of the $i$th input employed by the firm, $i = 1, \ldots, l$, $\boldsymbol{w}$ be the vector factor prices $(w_1, w_2, \ldots, w_l)$, and $C(\boldsymbol{y}, \boldsymbol{w})$ represent the firm's long run cost function. Finally, note that $\pi = \boldsymbol{p} \cdot \boldsymbol{y} - C(\boldsymbol{y}, \boldsymbol{w})$ corresponds to the economic profit of the firm.

Formally one can represent the Ramsey pricing problem as follows. Ramsey optimal (second best) prices will maximize the sum of consumer and producer surplus, $T$, subject to a constraint on the non-negativity of profits, $\pi \geq 0$:

$$\max_{\boldsymbol{p}} T = \psi(\boldsymbol{y}) + \boldsymbol{p} \cdot \boldsymbol{y} - C(\boldsymbol{y}, \boldsymbol{w}) \tag{11}$$

$$\text{subject to} \quad \pi = \boldsymbol{p} \cdot \boldsymbol{y} - C(\boldsymbol{y}, \boldsymbol{w}) \geq 0. \tag{12}$$

Let $\lambda$ be the non-negative Lagrange multiplier associated with the profit constraint (12). At an interior optimum (in which $p_i > 0$), the constraint will be binding when marginal cost pricing for all outputs would lead to a deficit; thus $\lambda > 0$. In addition the following conditions must hold:

$$\partial T/\partial p_i + \lambda \partial \pi/\partial p_i = 0, \quad \forall i, \tag{13}$$

which can be rewritten as:

$$-\lambda y_i = (1 + \lambda) \sum_{j=1}^{n} \left[p_j - \partial C/\partial y_j\right]\left(\partial y_j/\partial p_i\right), \quad \forall i. \tag{14}$$

In general, of course, the terms $\partial y_j/\partial p_i$ need not be zero for $i \neq j$. In fact, this cross derivative will be positive when products $i$ and $j$ are substitutes, negative when they are complements, and zero when the demands are independent. For simplicity, consider the special (and most famous) case in which all demands are independent, and let the price elasticity of demand for output $i$ with respect to price $p_j$ be denoted by $\varepsilon_{ij}$ and defined in the usual way as $(\partial y_i/\partial p_j)(p_j/y_i)$. Then after some algebra the conditions for optimality can be expressed in the following form:

$$\left\{\frac{p_i - \partial C/\partial y_i}{p_i}\right\} \varepsilon_{ii} = \left\{\frac{p_j - \partial C/\partial y_j}{p_j}\right\} \varepsilon_{jj} = -\frac{\lambda}{1 + \lambda}, \quad \forall i, j. \tag{15}$$

[48] The consumer surplus measure was discussed in Section 3; here one could represent it in terms of the familiar integral form as

$$\psi(p) = \int_p^{\infty} y(\hat{p}) \, \mathrm{d}\hat{p}.$$

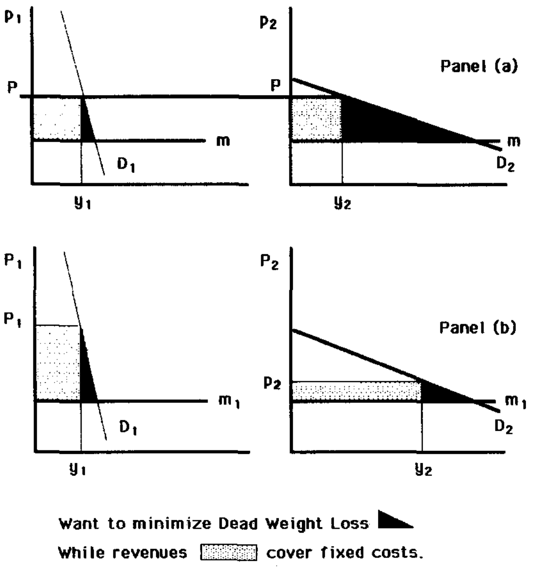

Figure 23.8. Ramsey pricing.

This relationship is the most well-known form of the Ramsey pricing rule. The terms in brackets in (15) represent the extent to which price deviates from marginal cost in the indicated (subscripted) markets, and is often referred to as the "markup" of price over marginal cost. The product of this markup and the corresponding elasticity of demand is known as the "Ramsey number"; for example, $(p_i - \partial C/\partial y_i)\varepsilon_{ii}/p_i$ is the Ramsey number for market $i$. The Ramsey number will be negative at an optimum in which the breakeven constraint is binding ($\lambda > 0$), since its numerical value is $-\lambda/(1+\lambda)$, which lies between zero and minus one; it will be zero when the breakeven constraint is not binding ($\lambda = 0$). When the demands are independent, the second best price in each market will be above marginal cost (i.e. the markup is positive) when the breakeven constraint is binding, and equal to marginal cost (i.e. first best) when the breakeven constraint is not binding.

Equation (15) indicates that the Ramsey number in each market must be equal. This relationship represents the famous "inverse elasticity rule", since it indicates that a lower markup must be associated with a more elastic demand when the breakeven constraint is binding. For an intuitive explanation of this result, consider the example illustrated in Figure 23.8. In this example the cost structure is affine, with equal marginal costs in each of the two markets served by the firm. Let the cost function be $C(y) = F + m(y_1 + y_2)$, and suppose the demands are independent.

Suppose first that the markups in the two markets are identical, instead of being based on the inverse elasticity rule. Since the marginal costs in the two markets are equal, equal markups mean equal prices in the two markets. This situation is depicted in panel (a) of Figure 23.8 (the top panel). The lightly shaded area in each market represents the revenues in excess of variable costs in that market; in each market that area can be thought of as a contribution toward covering the firm's fixed cost ($F$). The idea in panel (a) is to have equal markups (at a price $p$) which are large enough to have the dollar sum represented by the lightly shaded areas just equal to $F$. The demand in market 2 is drawn to be more elastic than the demand in market 1 when the price in each market is equal to $p$. Since the price in each market exceeds marginal cost, there is a deadweight loss, in each market represented by the black triangle. The sum of the areas of these triangles will be the dollar measure of the total economic inefficiency introduced by charging the prices $p_1 = p_2 = p$ instead of the first best prices $p_1 = p_2 = m$.[49]

The approach of requiring equal markups is but one of many possible ways of achieving non-negative profits. The question is: Is there another set of prices that would leave the firm without a deficit and make the sum of the deadweight losses smaller than the one indicated in panel (a), and in fact smaller than any other possible set of prices $(p_1, p_2)$? The inverse elasticity rule suggests how one might go about the task of finding that set of second best prices. It shows that the markup in market 1, the one with the more inelastic market, should be higher than in the (more elastic) market 2. Therefore one could adjust the markups accordingly, as represented in panel (b) of Figure 23.8. In panel (b) the sum of the lightly shaded areas in the two markets is intended to be the same as in panel (a), so that the revenues generated from the two markets once again just cover the fixed costs $F$. At the Ramsey optimal prices $(p_1, p_2)$ in panel (b), the sum of the areas of the black deadweight loss triangles is smaller than in panel (a), and in fact is as small as possible given that the firm must break even.

As the formulation of the Ramsey optimal problem (11)–(14) suggests, the inverse elasticity rule (15) is valid for much more general cost and demand structures than the linear ones illustrated in Figure 23.8. In fact the demands need not be independent, although the inverse elasticity rule (15) needs some modification in that case. Rohlfs (1979) has developed the Ramsey optimal rules in some detail for the case of interdependent demands.[50] The rule (15) must be altered to incorporate the effects of the cross partial derivatives $\partial y_i/\partial p_j$; this can be done in a straightforward fashion. For example, in the two product case, define Rohlfs' "superelasticity" as follows; $E_1 = \varepsilon_{11} - \varepsilon_{12} p_1 y_1/p_2 y_2$ and $E_2 = \varepsilon_{22} - \varepsilon_{21} p_2 y_2/p_1 y_1$, and then restate (15) to (16) to include the effects of demand

[49] The simple exercise of adding the welfare triangles in the two markets will not be valid if the demands in the two markets are interdependent. For more on welfare measurement in this case, see Braeutigam and Noll (1984).

[50] See also Zajac (1974).

interdependencies:

$$\left\{\frac{p_1 - \partial C/\partial y_1}{p_1}\right\}E_1 = \left\{\frac{p_2 - \partial C/\partial y_2}{p_2}\right\}E_2, \quad \forall i, j. \tag{16}$$

Observe that (16) simplifies to (15) when the cross elasticities of demand are zero.

The standard formulation of the Ramsey pricing problem [such as in the work of Baumol and Bradford (1970) and others cited above] assumes that the regulator operates with certainty about cost and demand relationships. That work is also typically developed in a static framework, and assumes that the regulated firm has a monopoly position in each of its markets. One might easily envision a host of additional modifications in the problem of second best in any particular industrial setting.[51] While we cannot hope to treat all of these extensions in detail, we do indicate the general nature of and provide selected references to some of this work.

Ramsey pricing principles have been developed for the case of uncertainty about the demand structure by Sherman and Visscher (1978).[52] Brock and Dechert (1983) and Braeutigam (1983) have shown how the principles can be extended to find optimal prices (and plant size) in a dynamic setting.[53]

The theory of Ramsey pricing has also been applied to cases in which the multiproduct firm does not have a monopoly in each of its markets. Braeutigam (1979) noted this problem in connection with the regulation of intermodal competition in surface freight transportation. Suppose one were interested in characterizing second best prices in the following setting. There are two modes of transport, each providing only a single service. Mode 1 is comprised of a single firm operating with economies of scale.[54] Mode 2 is comprised of a set of atomistic other firms which are competitive with one another.[55] All of the mode 2

[51] Examples of applications of second best pricing include among others Owen and Willig (1981), who apply Ramsey pricing to postal services, Willig and Bailey (1979), who examine AT & T's long distance rates by miles and time of day as well as postal rates, Willig (1979), who examines the problem of determining prices for access to a network (such as the telephone network), and Winston (1981), who examines the welfare losses from observed surface freight transportation rates relative to the losses that would have been observed at second best prices.

[52] The earlier work of Visscher (1973) is also of interest on this point.

[53] As one might suspect, there are interesting alternative ways of specifying both objective functions and constraints in these more complicated models. For example, in a dynamic formulation the exact form of optimal pricing rules will depend on whether the firm must break even at each point in time, or whether the firm must simply satisfy a constraint that requires the present value of profits over the relevant time horizon be non-negative.

[54] For example, mode 1 might be a railroad or a pipeline; this is stated here merely for illustration, and does not assert that any given railroad necessarily operates under economies of scale, since that is an empirical issue.

[55] An example of such a mode might be water carriers or motor carriers.

firms produce exactly the same service, and that service is an imperfect substitute for the service produced by mode 1. That paper shows that second best prices would in principle have to be set for *all* of the firms with interacting demands, not for just the mode with economies of scale. It also suggests why a Ramsey optimum might not be sustainable, since second best rates might typically be above marginal costs for mode 2.[56]

To be sure, each of these additional complexities leads to some modification in the exact form of the appropriate Ramsey rules. However, it seems fair to say that the essential principles of Ramsey pricing emerge in a robust fashion from the analysis, particularly as embodied in an inverse elasticity rule in some form.[57]

In closing this section, it is appropriate to point out that there is a fundamental difference between the approaches to pricing represented by Ramsey pricing and fully distributed cost pricing described in Section 4. As that earlier discussion indicated, FDC pricing proceeds with an ex ante allocation of common cost to all of the services, and then sets prices so that the revenues generated by each service will cover all of the costs allocated to that service. In other words, an allocation of common costs is the first step taken in a process that ultimately leads to a determination of prices.

Under Ramsey pricing, no allocation of common costs is made on the way to determining economically efficient prices. *After* the efficient prices are found, it may be possible to determine how the common costs would have to be allocated

[56] One could envision a kind of "third best" model in which the regulator allows the competitive mode 2 to clear its markets without regulation, thereby focusing only on the rates charged by the mode with economies of scale. This concept of regulation is called "partially regulated second best" (PRSB) in Braeutigam (1979), in contrast to "totally regulated second best" (TRSB) in which all rates for *all* competing modes are set by the regulator, a formidable task indeed. PRSB rates look very much like the Ramsey rules developed by Baumol and Bradford (1970), except that the elasticities of demand are those facing the firm instead of an industry (there is no well-defined industry demand since there are imperfect substitutes in the market).

This line of work has been extended still further. Baumol, Panzar and Willig (1982, ch. 11) suggest a concept of "viable firm Ramsey optimum" for Ramsey pricing in the case in which two or more firms, each operating with economies of scale, provide outputs which are perfect substitutes for one another. Braeutigam (1984) has developed Ramsey pricing rules for the case in which two or more firms, each operating with economies of scale, provide outputs which are *im*perfectly substitutable with one another.

[57] It turns out that, although it is not obvious, there is also a connection between prices that are sustainable and Ramsey optimal. Baumol, Bailey and Willig (1977) have stated a "Weak Invisible Hand Theorem" which points out that under a set of assumptions including a cost structure which exhibits both economies of scale and transray convexity (see Panzar's description of these concepts in Chapter 1 of this Handbook), Ramsey optimal prices are sufficient to guarantee sustainability. However, Faulhaber (1975) has generated a simple example in which a cost function not satisfying both economies of scale and transray convexity yields Ramsey optimal prices which are not sustainable; in fact a simple affine cost structure in which there are product specific fixed costs ($C = F_0 + F_1 + m_1 y_1 + F_2 + m_2 y_2$) is not transray convex if $F_i$ can be avoided when $y_i = 0$, and thus the Weak Invisible Hand theorem will not generally hold with such a structure.

in order for the second best prices to be generated from an FDC process.[58] However, this is an ex post exercise in allocating common costs. Although an allocation that is entirely cost-based may be desirable from an accounting perspective, it is not useful in the determination of efficient prices.

## 7. Nonlinear outlay schedules

In the previous two sections we have examined ways of increasing economic efficiency by charging different prices to customers in different markets served by the firm. For example, with peak load prices, daytime customers of electricity might be charged a price different from that charged to users of electricity in the nighttime. With Ramsey pricing as discussed in Section 6, shippers of different kinds of freight might be charged different rates by a railroad. However, in each case (Ramsey and peak load), users are still paying uniform prices within each market. For example, in the peakload case, daytime users are all paying the same (average) amount per unit purchased. We now extend the analysis of pricing to allow for tariffs which are not uniform as a way of improving economic efficiency still further. There is a rich literature on nonlinear outlays [see, for example, Oi (1971), Leland and Meyer (1976), Mirman and Sibley (1980), Schmalensee (1981), Spence (1981a), and Stiglitz (1977)], to name only a few important contributions. A particularly useful reference on this topic is Brown and Sibley (1986).

As was suggested in Section 4, there are many possible ways of structuring nonlinear outlay schedules; there the two part tariff was considered as one example. Recall that this kind of tariff has a "fixed" component and a "variable" component, as is illustrated in Figure 23.9. Suppose the customer must pay a fixed charge (sometimes called an entry charge) of $\$e$ per month to have access to the service in question (e.g. electricity or telephone service), where $e$ is to be paid regardless of the number of units actually purchased. In addition customers would be required to pay a variable charge equal to $m$ for each unit actually purchased during the month. The customer's total outlay would be $E = e + my$, an affine tariff schedule which is illustrated in Figure 23.9. The marginal outlay schedule (i.e. the schedule showing the *additional* expenditure $m$ incurred with the purchase of an *additional* unit of service) is constant; the average outlay schedule, which is nonlinear, is also shown in the second panel of Figure 23.9.

[58] Consider the simple case of a two product firm operating at a Ramsey optimum (with zero economic profits) under an affine cost structure, $C = F + m_1 y_1 + m_2 y_2$; let the Ramsey optimal prices be $(p_1, p_2)$. Then the contribution of revenues above the attributable costs for services 1 and 2 respectively would be $(p_1 y_1 - m_1 y_1)$ and $(p_2 y_2 - m_2 y_2)$; these two contributions must sum to $F$, since the firm is earning zero economic profits. Thus, the decimal fraction of the common cost $F$ allocated to service 1 is $(p_1 y_1 - m_1 y_1)/F$.

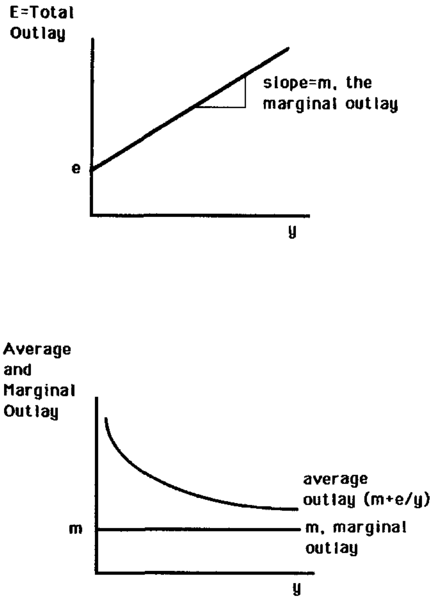

Figure 23.9. Affine tariff structure.

Note that a two part tariff with a zero fixed charge is therefore just a uniform tariff.

One could extend this approach to tariffs with more than two parts. For example, a four part tariff could be constructed with a fixed charge $e$, and three variable charges as follows:

$$E = \begin{cases} e + m_1 y, & \text{if } y \leq y_1, \\ e + m_1 y_1 + m_2(y - y_1), & \text{if } y_1 \leq y \leq y_2, \\ e + m_1 y_1 + m_2(y_2 - y_1) + m_3(y - y_2), & \text{if } y_2 \leq y. \end{cases} \tag{17}$$

This can be generalized to construct an "$n$ part tariff", which consists of a fixed charge $e$ and $(n - 1)$ variable charges, $m_1, m_2, \ldots, m_{n-1}$.

A nonlinear outlay schedule need not have a fixed charge. For example, suppose a tariff structure assesses each customer a charge of $m_1$ for each unit purchased up to some limit, $y_1$, and then a different amount per unit $m_2$ for each unit purchased in excess of $y_1$. Then the total outlay for the customer, $E$, would be as follows, where $y$ is the number of units the customer purchases:

$$E = \begin{cases} m_1 y, & \text{if } y \leq y_1, \\ m_1 y_1 + m_2(y - y_1), & \text{if } y > y_1. \end{cases} \tag{18}$$

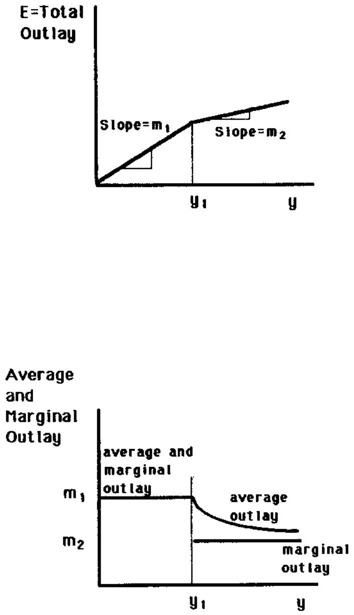

Figure 23.10. Nonlinear outlay structure.

The total, average and marginal outlay schedules for this tariff are shown in Figure 23.10. With this tariff, the average outlay is constant for output up to $y_1$, and declines thereafter, and is therefore nonlinear in $y$.

One could construct an $n$ part tariff in which the number $n$ becomes very large. In the limit, as $n$ approaches infinity, the tariff schedule would result in a smooth nonlinear outlay schedule of the kind illustrated in Figure 23.11. This tariff involves a total outlay $E = e + G(y)$, where $e$ is a fixed charge per month and $G(y)$ is the total variable charge per month. Here the slope of the total outlay schedule is continuously changing as output increases; since the slope of the total outlay schedule represents the value of the marginal outlay, the marginal outlay schedule is nonlinear everywhere in this example.

### *7.1. Pareto improving nonlinear outlay schedules*

How might a nonlinear outlay schedule lead to improved economic efficiency over a uniform tariff? Willig (1978) has demonstrated that any uniform price not equal to marginal cost can be Pareto dominated by a nonlinear outlay schedule.

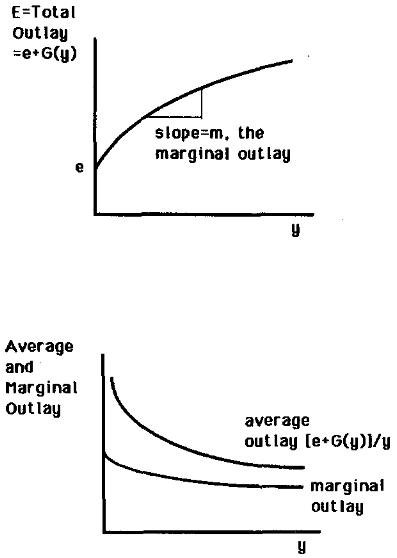

Figure 23.11. Nonlinear outlay schedule.

This important result can be illustrated with the aid of Figure 23.12.[59] Suppose a firm provides a product or service with an affine cost function, so that the marginal cost is constant. Consider the very simple example in which there are two consumers in the market, one with a "low" demand for the service, with a demand schedule $D_{\mathrm{L}}$, and one with a "high" demand for the service, with a demand schedule $D_{\mathrm{H}}$. If the firm must charge a uniform price to both consumers, the price must exceed marginal cost if the total revenues are to cover total costs

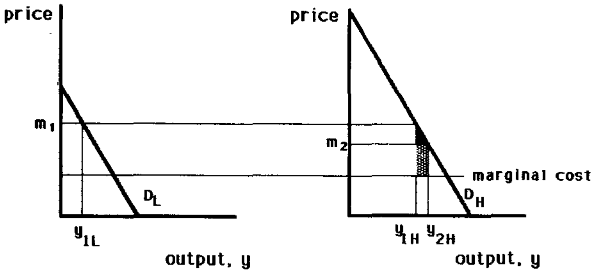

Figure 23.12. Pareto superior nonlinear outlay.

[59] This simple explanation was suggested to me in a conversation with John Panzar.

including the fixed cost of production; let the lowest uniform price that allows the firm to break even be $m_1$. At that price the quantities purchased by the low and high demand users will be $y_{1L}$ and $y_{1H}$ respectively, so that the total quantity demanded will be $(y_{1L} + y_{1H})$.

Now introduce some nonlinearity. Suppose that a tariff schedule like (18) were put in place. Figure 23.12 illustrates the demands of the two consumers, each demand being represented in a different panel of the figure. Suppose the tariff states that when consumer $i$ purchases $y_i \leq y_{1H}$, his total outlay will be $m_1 y_i$. If the consumer purchases $y_i > y_{1H}$, then his total outlay will be $m_1 y_{1H} + m_2(y_i - y_{1H})$, where $m_2 < m_1$, and $m_2$ is assumed greater than marginal cost in the figure. Note that the large consumer will be better off, since his consumer surplus has increased by the amount represented by the area of the solid black triangle in Figure 23.12. The small consumer is unaffected by the change in the tariff schedule. Finally, the firm is strictly better off since its profits have increased by the amount represented by the area of the dotted rectangle in Figure 23.12. Thus, the large user and the firm are strictly better off and the small user is no worse off under the nonlinear tariff, and the new tariff is therefore Pareto superior to the uniform tariff. In fact the firm could take a portion of the excess profit it has generated with the nonlinear tariff and lower $m_1$ by some amount so that even the small users are better off.

In the example just considered, the nonlinear tariff constructed included no fixed charge. It is also possible that economic efficiency can be improved over the level achievable with a uniform tariff by introducing an $n$ part tariff, which, as described earlier, has a fixed charge and $(n - 1)$ variable components.

To see how this might work, once again suppose a firm provides a product or service with an affine cost structure (with a fixed cost $F$) to a market with two consumers. Figure 23.13 illustrates the demands of the two consumers, with, as before, each demand being represented in a different panel. As before let the lowest uniform price that allows the firm to break even be $m_1$. Under this tariff low and high demand customers realize consumer surpluses represented respec-

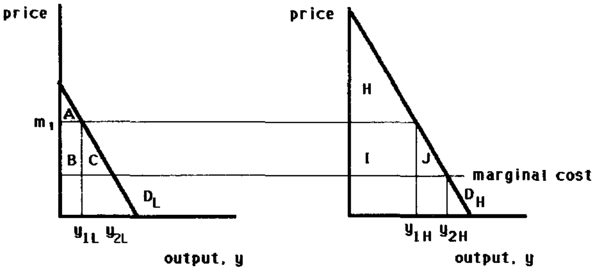

Figure 23.13. Pareto superior nonlinear outlay with entry fee.

tively by the areas $A$ and $H$. The sum of the areas $B$ and $I$ will have the same magnitude as the fixed cost $F$ since the firm is just breaking even when the tariff is $m_1$. The deadweight loss under $m_1$ is the sum of the areas $C$ and $J$.

One possible way of introducing a two part tariff is to charge each customer a fixed charge $e = F/2$ and a variable component of the tariff equal to marginal cost.[60] As long as the area $(A + B + C)$ is greater than the fixed charge $F/2$, then both consumers will remain in the market.[61] Furthermore, the firm is still just breaking even under the two part tariff and the market is operating at first best since the deadweight loss $(C + J)$ has been eliminated.

As noted, this scheme is qualified by the condition that the area $(A + B + C)$ be greater than the fixed charge $F/2$. If this is not satisfied, then the smaller customer will drop out of the market since he would be better off with no service than with service under the two part tariff. One might be tempted to split the coverage of the fixed cost somewhat differently, perhaps assigning a smaller fixed component $e_{\mathrm{L}}$ to the smaller customer and a larger entry fee $e_{\mathrm{H}}$ to the larger user (still requiring that $e_{\mathrm{L}} + e_{\mathrm{H}} = F$). This may even be feasible if the firm can discriminate between the two users. However, in order to implement this discriminatory scheme, the firm must know the identity of the two customers (who is large and who is small) so that a large customer can not pretend to be small, thereby incurring only $e_{\mathrm{L}}$, and leaving the firm with a deficit.[62] The problem arises here since the firm has established two different tariff schedules with the two entry fees, but has no way of forcing the high demand user to admit he is a high demand user in order to collect the higher entry fee from him.

The example illustrates that the limit on the efficiency of uniform entry fees is the elasticity of membership in the system with respect to the entry fee. Once users are recognized as being on the margin with respect to the entry fee, the entry fee becomes another price to be set with Ramsey pricing principles.

### 7.2. *Asymmetric information*

This brings us to one of the central ideas in the literature on nonlinear pricing: pricing under asymmetric information. Information is asymmetric here because the customer knows his own type, but in practice the firm often does not. If more

[60] This is the form of the two part tariff originally suggested by Coase (1946). The idea extends simply enough to the case of $n$ consumers; each customer would pay a fixed fee equal to $F/n$, and a variable component equal to marginal cost.

[61] Since $D_{\mathrm{H}}$ is a "larger" demand than $D_{\mathrm{L}}$ (i.e. $D_{\mathrm{H}}$ would lie to the right of $D_{\mathrm{L}}$ if drawn on the same graph), the area $(H + I + J)$ exceeds the area $(A + B + C)$; if the low demand customer remains in the market under the two part tariff, so will the high demand customer.

[62] Even if the firm knows the identity of the two users, there is also a possible problem with entry fees since one user can resell the output to the other customer in a way that would make it more attractive than buying from the firm directly. This restriction on resale is a standard condition for price discrimination to be possible.

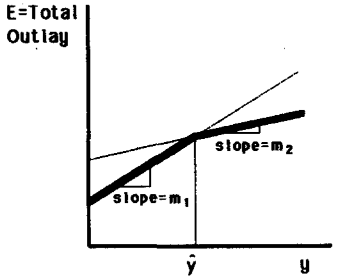

Figure 23.14. Self-selecting two part tariff.

than one tariff is announced by the firm, each consumer will choose ("self-select") that tariff schedule which is best for himself. In the case discussed above, each customer will find that the two part tariff with the lower entry fee dominates the one with the higher entry fee (since the variable components to the tariffs are identical), and no rational customer would ever pay the higher entry fee. It is therefore often not feasible to implement a pricing strategy which offers different tariff schedules to different customers.

Of course, this does not rule out a strategy of offering more than one tariff option to *all* customers. For example, the firm might announce two options that any customer may choose. The outlay schedules might take the form $E_i = e_i + m_i y$, where $y$ is the amount purchased by an individual. Suppose there are two such options, with $e_1 < e_2$. Then in order for tariff schedule 1 not to be dominated by tariff schedule 2 for all customers, it must be the case that $m_1 > m_2$. Some customers (presumably the "high" demand users) may find their optimal consumption to be with a high entry fee and a low variable fee (schedule 2), while other customers (presumably the "low" demand users) might prefer a low entry fee and a high variable fee (schedule 1). Such an arrangement is sometimes referred to as a self-selecting two part tariff. This is illustrated in Figure 23.14. A customer planning on consuming $y < \hat{y}$ would find his total outlay lower under tariff schedule 1 than under schedule 2. If consumption is greater than $\hat{y}$, a customer would find it less expensive to purchase under schedule 2. The lower envelope of the outlay schedules (represented by the heavy line segments in Figure 23.14) indicates that outlay schedule that would be chosen by a rational consumer since it minimizes the outlay in purchasing any given quantity of the service.

In the example above, welfare was improved by offering a tariff with two options since there were two types of customers. In the example if three options were introduced, one of the options would not be utilized since there are only two types of customers. However, in general there may be many "types" of consumers, instead of just the "low" and "high" demand users considered in the examples above. With more types of customers one can improve welfare by allowing consumers to self-select among more options.

Consider an example in which there are $j = 1, \ldots, J$ consumer types. Assume that if a consumer of type $j$ purchases service when confronted with an option $(e_i, m_i)$, he will purchase $y^j(m_i)$. Assume the consumer types can be ordered from smallest to largest so that $y^1(m_i) < y^2(m_i) < \cdots y^J(m_i)$ for any $m_i$.[63] Now construct an $n$ part tariff (with $n + 1 < J$) which is comprised of a collection of $n$ two part tariffs $(e_1, m_1), (e_2, m_2), \ldots, (e_n, m_n)$ from which the consumer can select the one optimal for himself. Let the entry fees be ordered so that $e_1 < e_2 < \cdots e_n$, and the variable fees be ordered so that $m_1 > m_2 > \cdots m_n$, so that no option is always dominated by another for all customers. An extension of the reasoning of Willig (1978) leads to the conclusion that a Pareto improvement can be achieved by introducing still another option $(e_{n+1}, m_{n+1})$ with $m_{n+1} < m_n$ and $m_{n+1}$ no less than marginal cost.[64] Although we do not treat it in detail here, the idea is as follows.

Consider a consumer of the highest demand type $J$, who is choosing the tariff whose parts are $(e_n, m_n)$ under the $n$ part tariff. Under the $n$ part tariff, customer $J$'s demand for the good was $y^J(m_n)$, and his total outlay was $e_n + m_n \cdot y^J(m_n)$. He will surely be induced to purchase under the new tariff option if his total outlay for $y^J(m_n)$ under $(e_{n+1}, m_{n+1})$ is less than it was under $(e_n, m_n)$; in other words he will purchase under $(e_{n+1}, m_{n+1})$ if $e_{n+1} + m_{n+1} \cdot y^J(m_n) < e_n + m_n \cdot y^J(m_n)$. Restated, since the consumer's demand schedule is downward sloping, he will achieve new consumer surplus from the new units he will purchase at the new option $(e_{n+1}, m_{n+1})$.[65] The firm is no worse off since the total outlay on $y^J(m_{n+1})$ is as great as it was under the $n$ part tariff, and the firm gets to keep any revenues above marginal cost on the new sales $y^J(m_{n+1}) - y^J(m_n)$. Thus, the $n + 1$ part schedule is Pareto superior to the $n$ part tariff since both the firm and consumers of type $J$ are better off under the new schedule, and consumers of other types are no worse off by having the new option available to them as well.

Although we have not yet addressed the optimality of a nonlinear outlay schedule, the arguments on Pareto superiority indicate that, at an optimum, the value of the variable component of the tariff $(m_i)$ available to the largest class user will be equal to marginal cost. This important result follows from the fact that the Willig argument can be used to generate Pareto improvements whenever

[63] The assumption that demands can be ordered in the strongly monotonic fashion indicated by $y^1(m_i) < y^2(m_i) < \ldots < y^J(m_i)$ for any $m_i$ is not innocuous, but it is the assumption utilized in most of the literature on nonlinear pricing. In particular it rules out the possibility that the demand schedules for any two types of consumers may intersect or cross one another at some price $m_i$.

[64] See chapter 4 of Brown and Sibley (1986) for an extended discussion of this point.

[65] The consumer gains from a lower $m_{i+1}$ on the purchases of the $y^J(m_i)$ units he is already purchasing but those inframarginal gains are taxed away by the higher entry fee $e_{m+1}$ under the new option. However, the consumer does get to keep the surplus on the additional units $[y^J(m_{i+1}) - y^J(m_i)]$ he purchases under $(e_{n+1}, m_{n+1})$.

$m_i$ exceeds marginal cost, and therefore a value of $m_i$ greater than marginal cost for the largest type user is not optimal.

One exception to the Pareto superiority arguments described in this section is worth noting. Ordover and Panzar (1980) have developed a model of a monopoly selling output to a single downstream industry. Thus, "consumers" of the product of the monopoly in this case are firms rather than direct end users, as we have considered above. Ordover and Panzar consider the case in which the downstream industry is competitive in its own product market, but that the firms in the industry differ with respect to their cost structure. Some of the firms produce with higher costs than others. Ordover and Panzar point out that if a nonlinear outlay schedule is introduced for the product of the monopoly, it may not be optimal to sell the final unit to the largest producer at marginal cost. This could occur since such a sale could lower the equilibrium price in the competitive downstream industry by enough so that "too many" higher cost firms are driven from the market, thereby eliminating a source of demand for the regulated product. They thereby demonstrate why it may be optimal for the final unit of the regulated product to be sold at a price greater than marginal cost.

### 7.3. *Optimal nonlinear outlay schedules*

Up to this point the discussion has addressed the Pareto superiority of nonlinear outlay schedules. The presentation has depended rather crucially on the ability to tailor nonlinear tariffs according to the desires of consumers of different types. It is worth stating that the results summarized so far do not depend on the distribution of consumer types. In other words, the Pareto superiority arguments depend on the existence of consumers of different types, they do not require information on the *number* of consumers of each type.

In the case of the determination of the exact values of the parameters of an optimal nonlinear outlay schedule, the distribution of consumer types (although not the identity of the type of any particular customer) must be known. The distribution may be discrete or continuous, and pricing formulations in the literature have treated both cases. [See, for example, Goldman, Leland and Sibley (1984) and Brown and Sibley (1986, chs. 4 and 5) for theoretical discussions of the problem of distribution of consumer types, and Mitchell (1978) for an empirical study of optimal pricing of local telephone service, which employs a lognormal distribution.]

In this subsection we briefly present one of the approaches that might be taken for the case in which consumers are distributed continuously. [This is treated in more detail in Brown and Sibley (1986, appendix to chapter 5).] Let $\theta$ be a parameter that indexes consumer type where observed types are bounded so that $\theta_L \leq \theta \leq \theta_U$, and let the (inverse) demand schedule for a type $\theta$ customer be

$p(y, \theta)$, with $p_y < 0$ and $p_\theta > 0$, the latter representing the strong monotonicity assumption that requires that demands of consumers of different types not cross one another. Denote the number of consumers of type $\theta$ by $g(\theta)$, with a cumulative measure $G(\theta)$. Also, assume the cost structure is affine with a fixed cost $F$ and a constant marginal cost $c$.

Let the tariff schedule be $p(y)$. This schedule indicates the price a consumer must pay for the $y$th (marginal) unit; thus $p(y)$ is said to be the marginal price for any unit of output. For a given quantity $y$, there will be a critical value of $\theta, \hat{\theta}$, such that a consumer of type $\hat{\theta}$ just has an incentive to purchase the $y$th unit under the tariff schedule $p(y)$. Thus, the marginal consumer type at $y$ given $p(y)$ is defined by the self-selection condition $p(y) = p(y, \hat{\theta})$, since consumers of type $\theta > \hat{\theta}$ will purchase the unit while those of type $\theta < \hat{\theta}$ will not. The self-selection condition implies that $\partial\hat{\theta}/\partial p(y) = 1/[\partial p(y, \hat{\theta})/\partial\theta] > 0$, a fact that will be useful in a later substitution. The total consumer and producer surplus over all $y$ can be written (ignoring the fixed cost $F$):

$$T = \int_0^\infty \left\{ \int_{\hat{\theta}}^{\theta_U} [p(y,\theta) - p(y)]\, g(\theta)\, d\theta + \left[1 - G(\hat{\theta})\right] \cdot \left[p(y) - c\right] \right\} dy, \tag{19}$$

where, for a differential (small) market $dy$ around a given $y$, $\int [p(y,\theta) - p(y)] g(\theta)\, d\theta$ represents the consumer surplus for customers in the market (with $\theta \geq \hat{\theta}$) and $[1 - G(\hat{\theta})] \cdot [p(y) - c]$ represents producer surplus. Thus, integration over all $y$ yields the total surplus associated with the schedule $p(y)$. The breakeven constraint for the firm (including the fixed cost $F$) is then:

$$\pi = \int_0^\infty \left\{ \left[1 - G(\hat{\theta})\right] \cdot \left[p(y) - c\right] \right\} dy - F \geq 0. \tag{20}$$

One can then characterize the outlay schedule $p(y)$ that maximizes (19) subject to (20). This leads to an expression of the following kind:

$$\frac{p(y) - c}{p(y)} = \frac{\lambda}{1+\lambda} \cdot \frac{1 - G(\hat{\theta})}{p(y) g(\hat{\theta}) \partial\hat{\theta}/\partial p} = \frac{\lambda}{1+\lambda} \cdot \frac{1}{\varepsilon(y, p(y))}, \tag{21}$$

where $\lambda$ is the multiplier associated with the constraint (20), the quantity $[1 - G(\theta)]$ is the quantity demanded in the differential market $dy$, and $\varepsilon(y, p(y))$ is the absolute value of the price elasticity of demand in that differential market.[66]

[66] The second order conditions for an optimum require that the marginal price schedule $p(y)$ cut the willingness to pay schedule $p(y, \theta)$ from below [i.e. $p(y, \theta)$] must have a more negative slope in $y$ than $p(y)$. For more on this see Goldman, Leland and Sibley (1984).

The relationship in (21) is of interest for several reasons. First, it is a kind of Ramsey rule now derived for a nonlinear outlay schedule instead of for the linear outlay schedules of Section 6. The deviation of price from marginal cost in each differential d$y$ market is inversely related to the price elasticity of demand in that market. Although the actual calculation of optimal prices may be difficult, the notion of Ramsey optimality unifies the literature on linear and nonlinear outlays. Equation (21) also indicates that for the final unit purchased by the largest customer ($\theta = \theta^{\mathrm{U}}$), $G(\hat{\theta}) = 1$, so that price equals marginal cost. This verifies a principle of optimality suggested earlier in this section for the case in which all customers are end users (rather than businesses).

In addition to the points just noted, one can summarize some of the important ideas from the literature on nonlinear outlays as follows. If a firm cannot break even under uniform marginal cost pricing, nonuniform tariffs can be used to improve welfare in a Pareto superior fashion. Nonuniform prices do this by tailoring tariffs according to the preferences of various types of consumers. They are typically implemented in a setting of asymmetric information, since a consumer knows his type but the firm does not. If there are more types of consumers than two part options within a tariff structure, then a Pareto improvement is possible with the addition of still another two part option. Finally, an economically efficient nonlinear outlay schedule covers total costs by requiring consumers with the greatest demands to make larger contributions on the inframarginal units they purchase. An optimal pricing relationship can be interpreted as a kind of Ramsey pricing rule.

## 8. Interservice subsidy

The discussions of pricing in the last four sections have focused on the economic efficiency of various pricing alternatives under regulation. Yet it has often been argued that the historical emphasis in regulatory rate-making has been on the "fairness" of rates rather than whether rates are economically efficient. Parties to regulatory hearings as well as commissions themselves have often asked whether a proposed rate is "fair", even in cases in which a party argues that a rate is economically efficient. The frequent tension between pricing to achieve economic efficiency and pricing to avoid interservice subsidy have been effectively summarized by Zajac (1978).

In this section we discuss the concept of a fair rate. It is usually raised in connection with the prices charged by a multiproduct firm for its different services. It is also often cast in terms of a question as to whether a rate is free of "cross subsidy" or its synonym "interservice subsidy". Crudely speaking interservice subsidy is said to occur when some service (or group of services) is either (i) not generating revenues sufficient to cover its fair share of the costs or (ii) generating revenues that cover more than its fair share of the costs. The problem becomes particularly interesting and difficult when there are common costs of production in the sense defined in Section 4. Recall that common costs are those that are shared in the production of two or more services; it is therefore impossible to allocate these costs in an unambiguous fashion among the services of the firm. Since tests for cross subsidy typically relate revenues for a service (or group of services) to the costs of providing that service, attempts to base tests of subsidy on fully distributed costing methods are themselves fraught with ambiguity. Therefore in this section we will confine ourselves to tests of cross subsidy which avoid the allocations of common costs as a procedural matter.

One could still envision a number of tests. One possibility would be to require that a service be priced no lower than marginal cost if it is to avoid subsidy. This has the virtue of avoiding any allocation of common costs, but it is a rather weak test. To see this, suppose that the cost structure of the firm is affine with total costs $C = F + F_1 + m_1y_1 + F_2 + m_2y_2$, where $(F_i + m_iy_i)$ are costs unambiguously attributable to service $i$ $(i = 1,2)$ and $F$ is a common cost. In this simple case a service that is priced to pass the marginal cost test may not even generate revenues sufficient to cover the costs directly attributable to that service. For example, if $p_1 = m_1$ (which passes the marginal cost test for service 1), the revenues from service 1 will not cover any of the fixed cost $F_1$ directly attributable to that service. Furthermore, if the firm earns zero economic profits, the revenues from service 2 will have to cover the balance of the costs $(F + F_1 + F_2 + m_2y_2)$; thus service 2 is generating revenues sufficient to cover not only all of its own attributable costs and all of the common costs $F$, but also all of the fixed cost $F_1$ directly attributable to service 1.

For these reasons the marginal cost test has not received widespread attention in the literature on regulation. Yet, if price were below marginal cost, one might well argue that at least the consumer of the marginal unit is being subsidized, since the price received for that unit would not cover the added costs of producing it. For that marginal unit the difference between price and marginal cost would have to be covered by revenues from other customers if the firm were to remain revenue adequate.

For a number of reasons discussed below the literature has focused on two other tests for subsidy. These are the *incremental cost test* and the *stand alone test*. To begin with, assume that the firm produces $N$ products under a cost structure $C(y) = C(y_1, y_2, \ldots, y_N)$. Consider now any subset of these services $S \subseteq N$. Let $C(y_S)$ denote the cost of producing the given levels of products in the subset $S$, and let $C(y_{N-S})$ be the cost of providing the given levels of products other than those in the subset $S$.

The incremental cost test [as defined by Faulhaber (1975)] requires that the revenues from the subset $S$ at least cover the increment to total cost that occurs when $S$ is produced as opposed to not being produced at all, holding constant the levels of the outputs in $y_{N-S}$. Formally this test can be stated as follows:

$$\sum_{i \in S} p_i y_i \geq C(y) - C(y_{N-S}) \equiv IC_S, \tag{22}$$

where $IC_S$ is the incremental cost of producing the product set $S$. If revenues from the product set $S$ do not satisfy (22), then service $S$ is said to be subsidized by revenues from other services.[67]

By contrast the stand alone test sets an upper (rather than a lower) bound on the revenues generated by services in the set $S$. The idea behind this test is that if the revenues generated by services in the subset $S$ exceed the cost of providing those services alone, then users of the services in $S$ are subsidizing users of other services. In other words suppose users of products in $S$ are paying more revenues when $S$ is provided in conjunction with other services not in $S$ than they would have to pay if only the products in $S$ are offered. Then the customers of $S$ could in principle withdraw from the production process that generates $S$ and the other services, form their own productive enterprise producing only $S$, and be better off, since the total revenues they would have to generate in a stand alone operation could be reduced relative to what they are currently paying. Formally the stand alone test can be represented as follows:

$$\sum_{i \in S} p_i y_i \leq C(y_S). \tag{23}$$

Several interesting observations can be made about these two tests. First, it can be shown that when profit for the firm is zero, then set $S$ passes the incremental cost test if and only if the remaining product set $(N - S)$ passes the stand alone test. This can be demonstrated rather easily. Consider the condition that the firm is just breaking even:

$$\sum_{i \in N} p_i y_i = C(y). \tag{24}$$

Suppose $S$ passes the incremental cost test, so that (22) is satisfied, and that the

[67] For example, under the affine cost structure

$$C = \begin{cases} F + F_1 + m_1 y_1 + F_2 + m_2 y_2, & y_1 > 0 \text{ and } y_2 > 0, \\ F + F_1 + m_1 y_1, & y_1 > 0 \text{ and } y_2 = 0, \\ F + F_2 + m_2 y_2, & y_1 = 0 \text{ and } y_2 > 0, \end{cases}$$

the incremental cost test on service 1 would require that $p_1 y_1 \geq F_1 + m_1 y_1$.

firm is just breaking even. Then subtracting (22) from (24) implies that

$$\sum_{i\in(N-S)} p_i y_i \leq C(y_{N-S}), \tag{25}$$

which is the condition that the stand alone test on $(N-S)$ is satisfied. This connection between the incremental cost test on $S$ and the stand alone test on $(N-S)$ is valid for any partition of the product set $N$ as long as the firm is earning zero economic profits.

A second observation about the subsidy tests is that it is not enough to test for subsidy only at the level of the individual services. In fact, when profits are zero either the incremental cost test or the stand alone test must be passed for *all* possible subsets $S$ if subsidies are to be avoided [see Faulhaber (1975)].[68] With $N$ services, this means that one would have to carry out $(2^N - 1)$ tests (including a test on all $N$ services taken together) in order to be sure that all possible groups of services are free of subsidy.[69]

Third, in a contestable market, one would expect entry to occur if any of the subsidy tests (on any subset of services) were not satisfied. This follows directly from two observations. First, in a contestable market one would expect to see the firm just breaking even; otherwise entry would occur or service would disappear. Second, given zero economic profits, if any of the subsets of services fails one of the subsidy tests, there is some subset of products which is generating revenues in

[68] Faulhaber also contributed the important insight that for a multiproduct firm with a subadditive cost structure, there may be *no* prices that are subsidy free according to the incremental cost and stand alone cost tests for all subsets of services. Thus subsets of consumers might find it attractive to purchase from alternative suppliers, even though the natural monopoly structure indicates that it would be socially efficient to have only a single supplier. Panzar and Willig (1977) showed that cost complementarities eliminate this possibility.

[69] To see why this might be a problem, consider a three product affine cost structure as follows:

$$C = \begin{cases} F + F_{12} + F_1 + m_1 y_1 + F_2 + m_2 y_2 + F_3 + m_3 y_3, & y_1 > 0,\ y_2 > 0 \text{ and } y_3 > 0, \\ F + F_{12} + F_1 + m_1 y_1 + F_2 + m_2 y_2, & y_1 > 0,\ y_2 > 0 \text{ and } y_3 = 0, \\ F + F_{12} + F_1 + m_1 y_1 + F_3 + m_3 y_3, & y_1 > 0,\ y_2 = 0 \text{ and } y_3 > 0, \\ F + F_{12} + F_2 + m_2 y_2 + F_3 + m_3 y_3, & y_1 = 0,\ y_2 > 0 \text{ and } y_3 > 0, \\ F + F_{12} + F_1 + m_1 y_1, & y_1 > 0,\ y_2 = 0 \text{ and } y_3 = 0, \\ F + F_{12} + F_2 + m_2 y_2, & y_1 = 0,\ y_2 > 0 \text{ and } y_3 = 0, \\ F + F_3 + m_3 y_3, & y_1 = 0,\ y_2 = 0 \text{ and } y_3 > 0. \end{cases}$$

Then the incremental cost of producing $y_1$ is $C(y_1, y_2, y_3) - C(0, y_2, y_3) = F_1 + m_1 y_1$. Suppose that incremental cost test is just passed so that $p_1 y_1 = F_1 + m_1 y_1$. Similarly the incremental cost of producing $y_2$ is $C(y_1, y_2, y_3) - C(y_1, 0, y_3) = F_2 + m_2 y_2$. Suppose that incremental cost test is just passed so that $p_2 y_2 = F_2 + m_2 y_2$. Then the total revenues from services 1 and 2 will be $(F_1 + m_1 y_1 + F_2 + m_2 y_2)$; yet this falls short of the incremental costs of services 1 and 2 taken together by an amount $F_{12}$, since that incremental cost would be $C(y_1, y_2, y_3) - C(0, 0, y_3) = F_{12} + F_1 + m_1 y_1 + F_2 + m_2 y_2$. Therefore, passing the incremental cost test for individual services does not guarantee that the incremental cost test for a group of services collectively will be passed.

excess of stand alone costs. In a contestable market this subset of products would be a target for entrants who would be satisfied with normal returns on that subset.[70]

Finally, much has been written about the relationship between subsidy-free prices and economically efficient (particularly Ramsey optimal) prices. One important result is the Weak Invisible Hand Theorem of Baumol, Bailey and Willig (1977). These authors showed that, under a set of assumptions including (among others) economies of scale and transray convexity, Ramsey optimal price vectors are sufficient (but not necessary) for sustainability.[71] Since sustainable prices must be subsidy-free, then under the conditions of the Weak Invisible Hand Theorem, Ramsey optimal prices would be subsidy-free.

While the assumptions required for the Weak Invisible Hand Theorem may be plausible for many cases, they are not totally innocuous. Early on Zajac (1972) pointed out that Ramsey optimal prices need not be subsidy-free according to the incremental cost test. This is intuitively easy to understand. Consider a two product firm operating with an affine cost structure. One of the markets it serves has a demand that is highly elastic (call this market 1) and the other has a rather inelastic demand (market 2). Then the inverse elasticity rule (see Section 6) would indicate that the Ramsey optimal markup of price over marginal cost would be relatively small in market 1. However, suppose there are fixed costs that are directly attributable to service 1, and which are avoidable if that service is discontinued.[72] Then the incremental cost of service 1 would include that attributable fixed cost, which might not be covered by revenues under Ramsey optimal prices sufficiently close to marginal cost. An alternative characterization of the example just given is that the demand in market 2 is so inelastic that Ramsey optimal prices would yield a price in that market which violates the stand alone test in market 2.[73]

In a contestable market, such a price could not be sustained without entry since entry would occur in the market or set of markets that fail the stand alone test. In regulated markets which are not contestable, one could think of modify-

[70] This view of subsidy has been generalized to the industry level (as opposed to the level of the firm) in markets that are contestable. Faulhaber and Levinson (1981) point out that any (and all) groups of consumers will pay an amount at least equal to industry wide incremental cost and no more than their own stand alone cost, regardless of their identities or consumption choices; Faulhaber and Levinson therefore call this distributive property "anonymous equity".

[71] Among the other assumptions the Weak Invisible Hand Theorem in the form presented above does not apply when there are demand complementarities. The requirements of transray convexity and no demand complementarities can be relaxed to some extent [see appendix 11 to chapter 8 in Baumol, Panzar and Willig (1982)].

[72] The Weak Invisible Hand Theorem does not apply in this example because the cost function is not transray convex. This occurs because the directly attributable fixed cost for service 1 creates a discontinuity of the cost function when service 1 disappears.

[73] Concern over a situation like the one described here might occur if, for example, service 2 is essential to some group of users. If our two product firm is the sole supplier of this service, then the provision of the service might constitute a "bottleneck" to users who need this product.

ing the second best Ramsey optimal formulation of (11) and (12) in Section 6 by appending additional constraints to ensure that the resulting prices are as efficient as possible while both being subsidy-free and allowing the firm to break even. These additional constraints would contribute to dynamic efficiency by guiding prices to send appropriate signals on entry.

## 9. Conclusion

This chapter has examined a number of optimal policies that might be used to control a natural monopoly. It has indicated why the traditional view of natural monopoly, which argues for regulation when there are pervasive economies of scale in a market, has been extensively questioned and modified in the literature since the late 1960s. It provides a summary of the contemporary literature characterizing a natural monopoly and shows how economic analysis has suggested rather forcefully that there may be ways to introduce competition for a market, even if a natural monopoly structure exists within a market. Competition for the market in these instances will lead to economically efficient prices. The possible optimality of such competition (at least in the sense of second best) in dealing with a natural monopoly is one of the main themes pursued here.

The chapter has also indicated that there are circumstances under which competition as a policy toward natural monopoly may not be feasible, or, even if feasible, may not lead to an economically efficient market outcome. It has summarized a number of ways in which one might improve the allocation of scarce resources if price regulation is imposed. These included peak load, Ramsey, and nonlinear pricing schemes.

While most of the discussion has dealt with efficiency, the chapter has also addressed a set of issues related to the "fairness" of regulated prices. It presented and discussed a set of possible notions of "cross subsidy" or "interservice subsidy", and related these concepts and economically efficient prices to one another.

Research described in this chapter has no doubt contributed to the many economic arguments that have supported deregulation or other regulatory reform in a number of American industries since 1970. Examples include the deregulation of airlines, motor carriers and cable television. They also include the efforts of the postal service to eliminate cross subsidies among postal services, the Federal Communications Commission's use of peak load pricing principles for telephone services, changes in structure and pricing in the electric power industry under the Public Utility Regulatory Policy Act of 1978, and the decision of the Interstate Commerce Commission to use Ramsey pricing principles and interservice subsidy tests in the railroad industry. A better understanding of natural monopoly will no doubt lead to improved theoretical and empirical work in the future, and should contribute still more to enlightened policy.

## References

Atkinson, A.B. and Stiglitz, J.E. (1980) *Lectures in public economics*. New York: McGraw-Hill.

Bailey, E.E. (1981) 'Contestability and the design of regulatory and antitrust policy', *American Economic Review*, 71:178–183.

Bailey, E.E. and Panzar, J.C. (1981) 'The contestability of airline markets during the transition to deregulation', *Law and Contemporary Problems*, 44:125–145.

Bailey, E.E. and White, L.J. (1974) 'Reversals in peak and off-peak prices', *Bell Journal of Economics*, 5:75–92.

Bailey, E.E., Graham, D.R. and Kaplan, D.P. (1985) *Deregulating the airlines*. Cambridge: MIT Press.

Baumol, W.J. (1982) 'Contestable markets: An uprising in the theory of industry structure', *American Economic Review*, 72:1–15.

Baumol, W.J. and Bradford, D.E. (1970) 'Optimal departures from marginal cost pricing', *American Economic Review*, 60:265–283.

Baumol, W.J. and Willig, R.D. (1981) 'Fixed cost, sunk cost, entry barriers and sustainability of monopoly', *Quarterly Journal of Economics*, 95:405–431.

Baumol, W.J., Bailey, E.E. and Willig, R.D. (1977) 'Weak invisible hand theorems on the sustainability of prices in a multiproduct monopoly', *American Economic Review*, 67:350–365.

Baumol, W.J., Panzar, J.C. and Willig, R.D. (1982) *Contestable markets and the theory of industry structure*. New York: Harcourt Brace Jovanovitch.

Baumol, W.J., Panzar, J.C. and Willig, R.D. (1984) 'Contestable markets: An uprising in the theory of industry structure: Reply', *American Economic Review*, 73:491–496.

Bohm, P. (1967) 'On the theory of 'second best',' *Review of Economic Studies*, 34:301–314.

Boiteux, M. (1956) 'Sur la gestion des monopoles publics astreint á l'équilibre budgetaire', *Econometrica*, 24:22–40.

Boiteux, M. (1971) 'On the management of public monopolies subject to budgetary constraints', *Journal of Economic Theory*, 3:219–240.

Braeutigam, R.R. (1979) 'Optimal pricing with intermodal competition', *American Economic Review*, 69:38–49.

Braeutigam, R.R. (1980) 'An analysis of fully distributed cost pricing in regulated industries', *Bell Journal of Economics*, 11:182–196.

Braeutigam, R.R. (1983) 'A dynamic analysis of second best pricing', in: J. Finsinger, ed., *Public sector economics*. London: Macmillan, 103–116.

Braeutigam, R.R. (1984) 'Socially optimal pricing with rivalry and economies of scale', *Rand Journal of Economics*, 15:124–131.

Braeutigam, R.R. and Noll, R.G. (1984) 'The regulation of surface freight transportation: The welfare effects revisited', *The Review of Economics and Statistics*, 56:80–87.

Brock, W.A. (1983) 'Contestable markets and the theory of industry structure', *Journal of Political Economy*, 91:1055–1066.

Brock, W.A. and Dechert, W. (1983) 'Dynamic Ramsey pricing', manuscript, Department of Economics, University of Wisconsin–Madison.

Brock, W.A. and Scheinkman, J.A. (1983) 'Free entry and the sustainability of natural monopoly: Bertrand revisited by Cournot', in: D.S. Evans, ed., *Breaking up Bell: Essays on industrial organization and regulation*. Amsterdam: North-Holland.

Brown, S.J. and Sibley, D.S. (1986) *The theory of public utility pricing*. Cambridge: Cambridge University Press.

Carlton, D. (1977) 'Peak load pricing with stochastic demands', *American Economic Review*, 67:1006–1010.

Caves, D., Christensen, L. and Tretheway, M. (1983) 'The structure of airline costs and prospects for the U.S. airline industry under deregulation', SSRI workshop series paper 8313, University of Wisconsin–Madison.

Chamberlin, E. (1962) *The theory of monopolistic competition*, 8th edn. Cambridge: Harvard University Press.

Clark, J.M. (1923) *Studies in the economics of overhead costs*. Chicago: University of Chicago Press.

Coase, R. (1946) 'The marginal cost controversy', *Economica*, 13:169–189.

Crew, M. and Kleindorfer, P. (1976) 'Peak load pricing with a diverse technology', *Bell Journal of Economics*, 7:207–231.

Crew, M. and Kleindorfer, P. (1986) *The economics of public utility regulation*. Cambridge: MIT Press.

Demsetz, H. (1968) 'Why regulate utilities?', *Journal of Law and Economics*, 11:55–65.

Diamond, P. and Mirlees, J. (1971) 'Optimal taxation and public regulation', *American Economic Review*, 61:261–278.

Dixit, A. (1982) 'Recent developments in oligopoly theory', *American Economic Review*, 72:12–17.

Evans, D.S. and Heckman, J.J. (1984) 'A test for subadditivity of the cost function with an application to the Bell system', *American Economic Review*, 74:615–623.

Faulhaber, G.R. (1975) 'Cross-subsidization: Pricing in public enterprises', *American Economic Review*, 65:966–977.

Faulhaber, G.R. and Levinson, S. (1981) 'Subsidy free prices and anonymous equity', *American Economic Review*, 71:1083–1091.

Fiorina, M.P. and Noll, R.G. (1978) 'Voters, bureaucrats and legislators: A rational choice perspective on the growth of bureaucracy', *Journal of Public Economics*, 9:239–254.

Friedlaender, A.F. (1969) *The dilemma of freight transport regulation*. Washington, D.C.: Brookings Institution.

Friedlaender, A.F. and Spady, R. (1982) *Freight transport regulation*. Cambridge: MIT Press.

Goldberg, V. (1976) 'Regulation and administered contracts', *Bell Journal of Economics*, 7:250–261.

Goldman, M.B., Leland, H.E. and Sibley, D.S. (1984) 'Optimal nonuniform prices', *Review of Economic Studies*, 51:305–319.

Hotelling, H. (1938) 'The general welfare in relation to problems of taxation and railway and utility rates', *Econometrica*, 6:242–269.

Hughes, J.R.T. (1977) *The governmental habit: Economic controls from colonial times to the present*. New York: Basic Books.

Joskow, P.L. and Noll, R.G. (1981) 'Regulation in theory and practice: An overview', in: G. Fromm, ed., *Studies in public regulation*. Cambridge: MIT Press, 1–65.

Joskow, P.L. and Schmalensee, R. (1981) *Markets for power: An analysis of electric utility deregulation*. Cambridge: MIT Press.

Kahn, A.E. (1970) *The economics of regulation: Principles and institutions*, vol. I. New York: Wiley.

Kahn, A.E. (1971) *The economics of regulation: Principles and institutions*, vol. II. New York: Wiley.

Knieps, G. and Vogelsang, I. (1982) 'The sustainability concept under alternative behavioral assumptions', *Bell Journal of Economics*, 13:234–241.

Leland, H. and Meyer, R. (1976) 'Monopoly pricing structure with imperfect discrimination', *Bell Journal of Economics*, 7:449–462.

Lipsey, R.G. and Lancaster, K. (1956–57) 'The general theory of second best', *Review of Economic Studies*, 24:11–32.

Littlechild, S.C. (1970) 'Peak-load pricing of telephone calls', *Bell Journal of Economics and Management Science*, 1:191–200.

Mirlees, J.M. (1976) 'Optimal tax theory: A synthesis', *Review of Economic Studies*, 38:175–208.

Mirman, L.J. and Sibley, D. (1980) 'Optimal nonlinear prices for multiproduct monopolies', *Bell Journal of Economics*, 11:659–670.

Mirman, L.J. and Tauman, Y. (1982) 'Demand compatible, equitable, cost sharing prices', *Mathematics of Operations Research*, 7:40–56.

Mirman, L., Samet, D. and Tauman, Y. (1983) 'Axiomatic approach to the allocation of a fixed cost through prices', *Bell Journal of Economics*, 14:139–151.

Mitchell, B.M. (1978) 'Optimal pricing of local telephone service', *American Economic Review*, 68:517–537.

Moore, T.G. (1986) 'Rail and trucking deregulation', in: L.W. Weiss and M.W. Klass, eds., *Regulatory reform: What actually happened*. Boston: Little, Brown, 14–39.

Noll, R.G. and Owen, B.M. (1983) *The political economy of deregulation: Interest groups in the regulatory process*. Washington, D.C.: American Enterprise Institute.

Oi, W.Y. (1971) 'A Disneyland dilemma: Two part tariffs for a Mickey Mouse monopoly', *Quarterly Journal of Economics*, 85:77–90.

Ordover, J.A. and Panzar, J.C. (1980) 'On the nonexistence of Pareto superior outlay schedules', *Bell Journal of Economics*, 11:351–354.

Owen, B.M. and Braeutigam, R.R. (1978) *The regulation game: Strategic use of the administrative process*. Cambridge: Ballinger.

Owen, B.M. and Willig, R.D. (1981) 'Economics and postal pricing policy', in: J. Fleishman, ed., *The future of the Postal Service*. New York: Praeger.

Panzar, J.C. (1976) 'A neoclassical approach to peak load pricing', *Bell Journal of Economics*, 7:521–530.

Panzar, J.C. (1980) 'Sustainability, efficiency and vertical integration', in: P. Kleindorfer and B.M. Mitchell, eds., *Regulated industries and public enterprise*. Lexington: Heath.

Panzar, J.C. and Willig, R.D. (1977) 'Free entry and the sustainability of natural monopoly', *Bell Journal of Economics*, 8:1–22.

Peltzman, S. (1976) 'Toward a more general theory of regulation', *Journal of Law and Economics*, 19:2111–2140.

Pigou, A.C. (1920) *The economics of welfare*. London: MacMillan.

Posner, R.A. (1974) 'Theories of economic regulation', *Bell Journal of Economics and Management Science*, 5:335–358.

Ramsey, F.P. (1927) 'A contribution to the theory of taxation', *Economic Journal*, 37:47–61.

Rees, R. (1984) *Public enterprise economics*. London: Weidenfeld and Nicolson.

Rohlfs, J.H. (1979) 'Economically efficient Bell system pricing', Bell Laboratories Economic Discussion Paper 138.

Samet, D. and Tauman, Y. (1982) 'A characterization of price mechanisms and the determination of marginal cost prices under a set of axioms', *Econometrica*, 50:895–910.

Scherer, F.M. (1980) *Industrial market structure and economic performance*. Chicago: Rand McNally.

Schmalensee, R. (1978) *The control of natural monopolies*. Lexington: Lexington Books.

Schmalensee, R. (1981) 'Monopolistic two-part pricing arrangement', *Bell Journal of Economics*, 12:445–466.

Schwartz, M. and Reynolds, R. (1984) 'Contestable markets: An uprising in the theory of industry structure', *American Economic Review*, 73:488–490.

Sharkey, W.W. (1981) 'Existence of sustainable prices for natural monopoly outputs', *Bell Journal of Economics*, 12: 144–154.

Sharkey, W.W. (1982a) 'Suggestions for a game theoretic approach to public utility pricing and cost allocation', *Bell Journal of Economics*, 13:57–68.

Sharkey, W.W. (1982b) *The theory of natural monopoly*. Cambridge: Cambridge University Press.

Shephard, W. (1984) '"Contestability" vs. competition', *American Economic Review*, 74:572–587.

Sherman, R. and Visscher, M. (1978) 'Second best pricing with stochastic demand', *American Economic Review*, 68:41–53.

Sorenson, J., Tschirhart, J. and Winston, A. (1978) 'A theory of pricing under decreasing costs', *American Economic Review*, 68:614–624.

Spady, R. and Friedlaender, A.F. (1978) 'Hedonic cost functions for the regulated trucking industry', *Bell Journal of Economics*, 9:159–179.

Spence, A.M. (1981a) 'Multi-product quantity dependent prices and profitability constraints', *Review of Economic Studies*, 47:821–841.

Spence, A.M. (1981b) 'Nonlinear prices and welfare', *Journal of Public Economics*, 8:1–18.

Steiner, P.O. (1957) 'Peak loads and efficient pricing', *Quarterly Journal of Economics*, 71:585–610.

Stigler, G. (1971) 'The theory of economic regulation', *Bell Journal of Economics and Management Science*, 2:3–21.

Stigler, G. (1975) *The citizen and the state: Essays on regulation*. Chicago: University of Chicago Press.

Stiglitz, J.E. (1977) 'Monopoly, nonlinear pricing and imperfect information', *Review of Economic Studies*, 44:407–430.

Sweeney, G. (1982) 'Welfare implications of fully distributed cost pricing applied to partially regulated firms', *Bell Journal of Economics*, 13:525–533.

Taussig, F.W. (1913) 'Railway rates and joint costs', *Quarterly Journal of Economics*, 27:692–694.

Turvey, R. (1968) *Optimal pricing and investment in electricity supply*. Cambridge: MIT Press.

Turvey, R. (1969) 'Marginal cost', *Economic Journal*, 79:282–299.

Visscher, M. (1973) 'Welfare maximizing price and output with stochastic demand: Comment', *American Economic Review*, 63:224–229.

Waverman, L. (1975) 'Peak-load pricing under regulatory constraint: A proof of inefficiency', *Journal of Political Economy*, 83:645–654.

Weil, Jr., R.L. (1968) 'Allocating joint costs', *American Economic Review*, 58:1342–1345.

Weiss, L.W. and Klass, M.W. (1981) *Case studies in regulation: Revolution and reform*. Boston: Little, Brown.

Weiss, L.W. and Klass, M.W. (1986) *Regulatory reform: What actually happened*. Boston: Little, Brown.

Weitzman, M. (1983) 'Contestable markets: An uprising in the theory of industry structure: Comment', *American Economic Review*, 73:486–487.

Wiedenbaum, M.W. (1978) *The cost of government regulation of business*. Washington, D.C.: U.S. Congress, Joint Economic Committee, Subcommittee on Economic Growth and Stabilization.

Williamson, O.E. (1966) 'Peak load pricing and optimal capacity under indivisibility constraints', *American Economic Review*, 56:810–827.

Williamson, O.E. (1976) 'Franchise bidding for natural monopolies – in general and with respect to CATV', *Bell Journal of Economics*, 7:73–104.

Willig, R.D. (1976) 'Consumer's surplus without apology', *American Economic Review*, 66:589–597.

Willig, R.D. (1978) 'Pareto superior nonlinear outlay schedules', *Bell Journal of Economics*, 9:56–59.

Willig, R.D. (1979) 'The theory of network access pricing', in: H.M. Trebing, ed., *Issues in public utility regulation*. East Lansing, Michigan: Michigan State Public Utilities Papers, 109–152.

Willig, R.D. and Bailey, E.E. (1979) 'The economic gradient method', *American Economic Review*, 69:96–101.

Winston, C. (1981) 'The welfare effects of ICC rate regulation revisited', *Bell Journal of Economics*, 12:232–244.

Zajac, E.E. (1972) 'Some preliminary thoughts on subsidization', presented at the Office of Telecommunications policy research conference on communication policy research, Washington, D.C.

Zajac, E.E. (1974) 'Note on an extension of the Ramsey inverse elasticity of demand pricing or taxation formula', *Journal of Public Economics*, 3:181–184.

Zajac, E.E. (1978) *Fairness or efficiency: An introduction to public utility pricing*. Cambridge: Ballinger.