CHAPTER 2

The Theory of Natural Monopoly

2.1 THE NATURAL MONOPOLY CONUNDRUM

Historically, conventional wisdom has held that certain markets were "naturally monopolistic," which means that, due to the presence of high fixed costs whose average declines with increases in output, efficiency is best obtained when there is only one supplier. According to Kahn (Kahn, p. 15), "the public utility industries are preeminently characterized in important respects by decreasing unit costs with increasing levels of output. That is indeed one important reason why they are organized as regulated monopolies: a 'natural monopoly' is an industry in which the economies of scale are such that one company supplies the entire demand. It is a reason, also, why competition is not supposed to work well in these industries." Included here are the markets for electricity, natural gas, telephone, and water services. It has often been argued that this phenomenon is driven by the irreversibility of the initial investment required to produce a particular good or service in a naturally monopolistic industry. More specifically, the underlying production technology of this product is such that there exists a level of output for which average cost is minimized; at levels of output below this level, average costs decline, and at levels above it, they rise. This, known as *economies of scale*, is investigated further in context to its relationship with the theory of natural monopoly.

Economists have spent many years attempting to assess that level of output at which the minimum efficient scale occurs. In some industries, such as the generation of electricity, consensus has been reached that, at least in 1970, most firms were producing in and around this level, given a particular production technology (Christensen and Greene, 1976). In other words, economies of scale in the generation of electricity had been exhausted.

Until recently, no one questioned that the production of electricity was in fact a natural monopoly, since, like telephony, what is required here is a network, a complex, interactive, interdependent connection of wires (by which individuals gain access to the local distribution company, which is connected to the transmission grid at various nodes). This network

Electricity Cost Modeling Calculations
DOI: 10.1016/B978-1-85617-726-9.00002-9

represents an irreversible investment, which is characterized by both economies of scale and those of network planning, and as such yields a natural monopoly. Because this network leads to externalities, vertical integration has traditionally yielded the most efficient organization of the industry, especially for larger firms. But, due to the vertical nature of electricity production, questions have arisen concerning whether any aspect of the production process may not be a natural monopoly. And, if this is the case, the questions then become: Would the market be better served by allowing competition into that component and would the gains from competition exceed the lost economies that would result? This is the critical element that needs to be explored.

But things are not always so clear. While little work has been done in the areas of testing whether the transmission and distribution processes are natural monopolies, they are usually assumed to be so, since both are characterized by what is known as *network economies*. Network economies arise due to the interconnectedness of the national transmission grid, so that significant saving in inputs and direct routing yield both economies of scale and economies of scope. These are defined later in this chapter, along with a review of the relevant literature.

Defining natural monopoly

Older industrial organization theory cited that the presence of scale economies determines whether an industry is a natural monopoly. It is important to note that much of the *theory* of natural monopoly is concerned with the precise meaning of *increasing returns* or, equivalently, *decreasing average costs*. Scale economies exist when a proportionate increase in output leads to a less-than-proportionate increase in cost. Mathematically, a cost function (one output) is said to exhibit global (local) economies of scale if

$$C(\lambda q) < \lambda C(q) \tag{2.1}$$

for $\lambda > 1$, $q \geq 0$.

According to Marshall (1927), increasing returns can be either internal or external to the firm and, similarly, internal or external to the industry. A natural monopoly tends to arise due to high fixed costs, which tend to be asset specific and, as such, are largely sunk. As a result, average cost tends to decline as output is expanded over a large range, thus rendering a single provider socially optimal. In addition, economies of scale can be

either technical (relating to the production process) or pecuniary, related to the prices paid for inputs.

One of the difficulties in testing for natural monopoly is the practical application of testing for the subadditivity of a firm's cost function, which is critical, since local (global) subadditivity is a necessary and sufficient condition for local (global) natural monopoly (Evans, 1983). In addition, it is necessary to distinguish between single-output and multiple-output natural monopolies, which I do in the following sections.

2.2 FOR A SINGLE-OUTPUT MARKET

An industry is said to be a natural monopoly if one firm can produce the desired market demand at a lower cost than two (or more) firms. More specifically, it is defined in terms of a single-firm's efficiency relative to the efficiency of other firms in the industry (as opposed to a firm's being the controller of an essential resource or having a patent on a particular product). In other words, economies of scale may exist in the production of a particular product. Some characteristics of a natural monopoly attributable to economies of scale include

1. Decreasing long-run average cost.
2. High fixed costs.
3. Subadditivity of its cost function.

Although interrelated, the most important of these is subadditivity of the firm's cost function, which means that it is cheaper for one firm to produce the total output demanded than it would be for several firms to produce proportions of it. This can be expressed as

$$C(Y) < \Sigma C(y^i) \tag{2.2}$$

where $\Sigma y^i = Y$.

If this holds, then the cost function is strictly subadditive at output level Y (Sharkey, 1982). For a single-output firm, subadditivity is both necessary and sufficient for a natural monopoly, since subadditivity implies that it is more efficient for a single firm to produce all the output in the market. It is important to note that subadditivity is a local concept; that is, just because the cost is subadditive at one level of output does not necessarily mean that it is subadditive at all output levels, or globally subadditive. This implies that the total cost of production must be evaluated at all levels of output up to the level that satisfies market demand.

Average cost

Certainly, declining average cost throughout the relevant range of outputs is an indicator that the cost function is subadditive and it is more efficient for one firm to supply the entire industry output; that is, a natural monopoly. What this requires, however, is that the marginal cost also declines throughout a subset of this range of outputs. And necessary for this is a twice-differentiable cost function, which yields the appropriately shaped average and marginal cost curves. Such a cost function is displayed in Figure 2.1.

A cubic cost function yields the appropriately shaped average and marginal cost curves. For $Y < Y^*$, cost increases at a decreasing rate. In this range, both marginal and average cost decline. However, once diminishing returns set in, costs begin to increase at an increasing rate; it is in this range that marginal costs begin to rise and total cost increases at an increasing rate, which causes average cost to begin rising and yields the U-shaped average cost curve displayed in Figure 2.2.

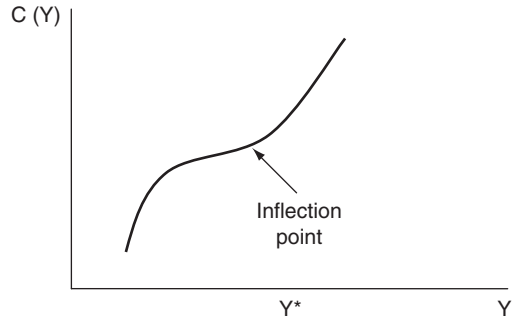

Figure 2.1 Total cost curve generated by a cubic cost function

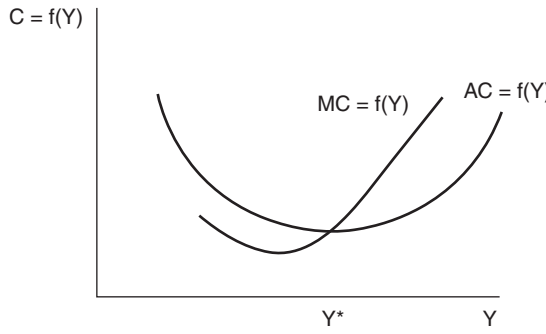

Figure 2.2 Average and marginal cost curves generated by cubic total cost function

A cubic cost function generates this particular shape and is of the general form

$$C(Y) = a + bY + cY^2 + dY^3 \tag{2.3}$$

so that average cost is given by

$$\mathrm{AC}(Y) = a/Y + b + cY + dY^2 \tag{2.4}$$

where $\mathrm{AC}(Y) = C(Y)/Y$. And marginal cost is given by

$$\partial C(Y)/\partial Y = b + 2cY + 3dY^2 \tag{2.5}$$

Note: As long as a, b, and $d > 0$, and $c < 0$, the total cost curve is as displayed in Figure 2.1, which yields appropriately shaped (U-shaped, due to diminishing returns) average and marginal cost curves; that is, as displayed in Figure 2.2.

The cubic cost function just described generates the average (AC) and marginal (MC) cost curves displayed in Figure 2.2. For $Y < Y^*$, marginal cost declines and pulls average cost down with it; this is the region of the total cost curve in which cost rises at a decreasing rate. Once diminishing returns set in, marginal costs rise and eventually cause average cost to rise as well, which occurs at Y^*, when total costs begin to increase at an increasing rate.

Economies of scale

Of the three cost concepts just described, average cost is the most important in the determination of the most efficient industry structure (i.e., number of firms supplying the market demand).

Appealing to Baumol, Panzar, and Willig (1982), scale economies are said to be present when a k-fold *proportionate* increase in every input results in a k'-fold increase in output where $k' > k > 1$. This is even stronger than declining average cost, since it implies that average costs are declining but the converse is not necessarily true. The reason is that it may be even less costly to increase output by non-proportional increases in inputs (see Baumol et al., 1982, p. 21 for more details). With this said the following propositions are offered:

> Proposition 2.1. Locally, economies of scale are sufficient but not necessary for declining average cost.
>
> Proposition 2.2. Globally, economies of scale are sufficient but not necessary for subadditivity of costs (i.e., natural monopoly).

Aside: Necessary versus sufficient conditions

In logic, the words *necessity* and *sufficiency* refer to the implicational relationships between statements. The assertion that one statement is a *necessary and sufficient* condition of another means that the former statement is true *if and only if* the latter is true. In other words,

- A *necessary* condition of a statement must be satisfied for the statement to be true. Formally, a statement P is a necessary condition of a statement Q if Q implies P.
- A *sufficient* condition is one that, if satisfied, assures the statement's truth. Formally, a statement P is a sufficient condition of a statement Q if P implies Q.

Examples

1. Given the average cost curve displayed in Figure 2.2, a necessary condition for cost minimization is that its derivative, which is equal to

$$\partial \mathrm{AC}(Y)/\partial Y \tag{2.6}$$

is equal to zero.

Does this guarantee that costs are minimized? No. There is also a sufficient condition that must be satisfied: the second derivative, which is given by

$$\partial^2 \mathrm{AC}(Y)/\partial Y^2 > 0 \tag{2.7}$$

Otherwise, a strictly negative second derivative guarantees a maximum, not a minimum as required.

2. A total revenue function of the following form:

$$PY = AY - BY^2 \tag{2.8}$$

where $P =$ price and $Y =$ output, so that $P \times Y =$ total revenue, yields a marginal revenue curve that is given by

$$\partial \mathrm{TR}(Y)/\partial Y = A - 2BY \tag{2.9}$$

A necessary condition for profit maximization is that marginal revenue equal marginal cost (thus implying that the slope of the total cost curve is equal to the slope of the total revenue curve). Solving equations (2.5) and (2.9) for Y^*, the profit maximizing level of output, we have

$$3dY^2 + 2(B - c)Y + b - A = 0 \tag{2.10}$$

(recall, $c < 0$). This requires the quadratic formula to solve and yields two distinct (and feasible) values for Y^* as long as

$$(B-c)^2 > 3d(b-A) \tag{2.11}$$

so that

$$[(B-c)^2 > 3d(b-A)]^{1/2} \tag{2.12}$$

is defined.

Given that there are two possible solutions, which are displayed in Figure 2.3, a sufficient condition must be established.

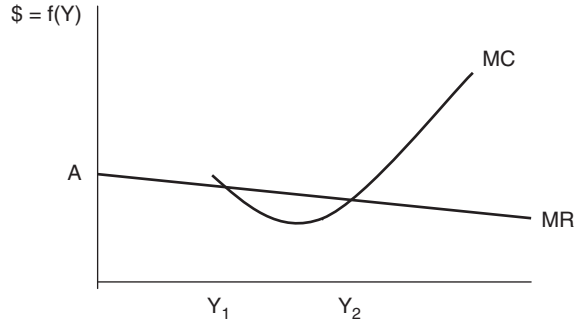

Figure 2.3 Profit maximizing level of output (Y_2 not Y_1)

Marginal cost equals marginal revenue at Y_1 and Y_2. Clearly, at output level Y_1, marginal cost exceeds marginal revenue, which results in a loss to the firm. At any Y such that $Y_1 < Y < Y_2$, marginal revenue exceeds marginal cost, so that profit is being earned. And, at Y_2, the entire profit is captured by the firm. What is the sufficient condition? Marginal cost must rise faster than marginal revenue, as displayed in Figure 2.3. In other words, the derivative of marginal cost with respect to output is higher than the derivative of marginal revenue, which is expressed as

$$\partial \mathrm{MC}(Y)/\partial Y > \partial \mathrm{MR}(Y)/\partial Y \tag{2.13}$$

That is,

$$Y > (c-B)/3d \tag{2.14}$$

These cost concepts are used extensively throughout the economics literature and are revisited in subsequent chapters. As such, it would be instructive to work through a numerical example here.

Numerical example 2.1

Let the demand and cost curves be given by

$$P = 20 - 0.5Y \tag{2.15}$$

and

$$C = 0.04Y^3 - 1.94Y^2 + 32.96Y \tag{2.16}$$

In the absence of regulation, the monopolist's profit maximizing levels of output are determined by setting marginal revenue equal to marginal cost. In this case, marginal revenue and marginal cost are given by

$$\partial \mathrm{TR}(Y)/\partial Y = 20 - Y \tag{2.17}$$

and

$$\partial \mathrm{TC}(Y)/\partial Y = 0.12Y^2 - 3.88Y + 32.96 \tag{2.18}$$

Using the quadratic formula to solve for Y^*, the profit maximizing level of output, yields two solutions:

$$Y^* = (18, 6)$$

Which solution is correct? A check of the second-order (or sufficient) conditions is now required, which involves evaluating the second derivatives of the cost and revenue functions. This yields

$$\partial \mathrm{MC}(Y)/\partial Y = 0.24Y - 3.88 \tag{2.19}$$

Evaluating at Y^* yields two solutions:

$$\partial \mathrm{MC}(Y)/\partial Y = (0.44, -2.44)$$

and

$$\partial \mathrm{MR}(Y)/\partial Y = -1$$

Only one solution ($Y^* = 18$) satisfies the sufficient condition for profit maximization; that is, that marginal cost rises faster than marginal revenue (i.e., $\partial \mathrm{MC}(Y)/\partial Y > \partial \mathrm{MR}(Y)/\partial Y$).

Efficient industry structure

For now, let us move on to the fundamental concept in determining the most efficient industry structure in single-output markets.

Degree of scale economies

The degree of scale economies (SCE) at output Y, is the elasticity of output at Y with respect to the cost to produce it. Formally, it is defined as

$$\mathrm{SCE}(Y) = \mathrm{C}(Y)/Y\mathrm{C}'(Y) \tag{2.20}$$

where

$$C'(Y) = \partial C(Y)/\partial Y \tag{2.21}$$

Equation (2.20) is equivalent to the ratio of average cost to marginal cost. Returns to scale are said to be increasing, constant, or decreasing as SCE is greater than, equal to, or less than unity.

Economies of scale applied to the electric utility industry

Due to the nature of electricity, which is not storable and flows along the path of least resistance; the high, largely sunk required capital investment (which yields declining long-run average costs over the relevant range of output); and the well-established presence of economies of scale and vertical integration (see the literature review later in this chapter for details), for many years, conventional wisdom held that competition was infeasible and price regulation was necessary to ensure that consumers pay a fair price and producers and owners are appropriately compensated for any risks associated with supplying the electricity. Clearly characterized as natural monopolies, Figure 2.4 displays the theoretical construct of a naturally monopolistic market and the reason that utilities have historically been subjected to price regulation.

Without price regulation, the consumer would pay the monopoly price (denoted P^M) and the output in the market would be Y^M. Although allocative efficiency would dictate that price equal marginal cost (denoted P^C in Figure 2.4), which would yield an industry output of Y^C, this is not

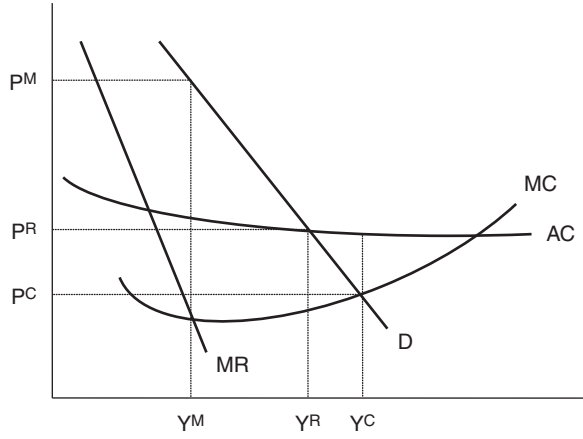

Figure 2.4 Traditional price regulation in the United States

feasible since the firm cannot cover its costs. Thus, the regulator would set the price equal to average cost, which yields a higher level of output (Y^R) than the monopoly output and a lower price would prevail in the marketplace (denoted by P^R). In addition, the regulated price allows the firm to recover its prudent costs and earn an acceptable return on equity (in the case of investor-owned utilities, which supply 66% of the U.S. market as indicated in the introductory chapter to this book).

2.3 LITERATURE REVIEW—ECONOMIES OF SCALE IN GENERATION: SINGLE-OUTPUT MODELS

In the electric utility industry, several papers in the literature treat electricity as a single (homogeneous) output. However, they pertain mostly to the generation of electricity only, the cost of which has been estimated using increasingly sophisticated econometric techniques, which typically employ either production functions or cost functions to reach their findings. Among those studies that estimate economies of scale in generation are Nerlove (1963), who used 1955 data on 145 utilities and found that the cost function was characterized by increasing returns to scale but that returns to scale tended to decline with the size of the firm. His study is discussed in more detail in Chapter 4.

Concerning the generating stage, it is necessary to distinguish between studies that use the plant as the sample unit and those that use the firm itself. In their seminal paper, Christensen and Greene (1976), used both Nerlove's 1955 data and also 1970 data on the same firms and found that, by 1970, most firms were generating electricity at a point on the average cost curve in which economies of scale had been exhausted. Using a different cost specification than Nerlove, they found that the minimum efficient scale was attained at 3800 MW and some firms were producing even beyond this level of output (i.e., in the diseconomies of scale region of the long-run average cost curve). This implied that the generation of electricity was not a natural monopoly and led to the realization that competition in the generation component was not only feasible but may also be more efficient. This realization precipitated the eventual deregulation of the generation component of the industry, which is discussed in a subsequent chapter (Chapter 3). In addition, the translogarithmic cost function employed in this particular study is the subject of further analysis in Chapters 4, 5, and 6.

Joskow and Schmalensee (1983) present a summary of studies carried out in the United States based on econometric estimations and engineering

methods. At the time, the minimum efficient scale (MES) for conventional electricity generation was around 800 MW and around 2000 MW for nuclear energy. In a later study, Huettner and Landon (1978), using yet another cost specification and 1971 data on 74 electric utilities, confirmed the Christensen and Greene results, although they found that scale economies were exhausted at an even lower level of output. As they pointed out, the relationships observed at plant level, particularly scale economies, are often modified by interrelationships at higher levels of decision making, such as the firm level. Greene (1983) studied economies of scale using panel data on investor-owned utilities from 1955 to 1975 and found that scale economies actually declined over that period of time. Technical change, he argues, was a significant factor in the decreasing average costs that firms were experiencing throughout the majority of the study period. Thermal efficiencies were being exhausted while the demand for electricity was rising, thanks to declining power prices. Atkinson and Halvorsen (1984) employed yet a different cost model and found that, using 1970 data on 123 privately owned firms, most of the firms in the sample were operating in the downward sloping portion of the long-run average cost curve.

For the most part, these studies consider cost functions in which the output is the kilowatt-hours of electricity generated. (Christensen and Greene, 1976; Huettner and Landon, 1978; Atkinson and Halvorsen, 1984). But others, namely, Kamerschen and Thompson (1993) and Thompson and Wolf (1993), studied possible cost differences between conventional electricity generating technology (fossil-fuel generation) and nuclear electricity generation.

This was not the case, however, in the years (and decades) that followed, which were extremely turbulent ones for the industry. Rapidly rising fuel prices, double-digit inflation, and rising capital prices led to a decline in the demand for electricity, causing financial distress for a number of utilities, which were saddled with excess capacity. (Thompson, 1995).

Later studies include Maloney (2001), who estimated the MES at 321 MW and 260 MW for coal- and gas-fired plants, respectively, but found that the average cost curve is flat at this level. Kleit and Terrell (2001) and Hiebert (2002) found increasing scale economies for most observations. Hiebert found that the degree of scale economies was 20% in coal-fired plants and 12% for natural-gas-fired plants for average sample values (780 MW and 284 MW, respectively). This work also found that

major economies can be attained by producing with more than one plant for each kind of generation. This latter aspect highlights the importance of distinguishing between plant and company in generation.

Whereas these studies focus on the generation component, a few studies focus on either transmission or distribution alone, two of the three components, and all three components. Of those that focus on some combination of the components, most do so to study the economies associated with vertical integration, which is discussed later in this chapter. In virtually all these studies, the consensus is that distributed electricity is not a homogenous good. This is discussed further in Chapter 4, but for now suffice it to say that different end users have different elasticities of demand and some users are more costly to serve than others.

Economies of scale and density in transmission and distribution

Some studies estimated the economies of scale for the transmission and distribution elements, like Huettner and Landon (1978), who found that the minimum efficient scale occurred at around 2600 MW capacity. Kaserman and Mayo (1991) also found specific economies of scale for these phases, and they situate the minimum efficient scale at around 5 GWh. And Greer (2003) found that none of the rural electric cooperatives distributed anywhere near the minimum efficient scale in 1996.

The network elements and the costs involved in these activities can be studied in greater depth by studying economies of density. This concept explains the evolution of average costs when production is increased and some of the characteristics that define the product are maintained constant, for example, the size of the service area or the number of consumers.

Network economies

For electricity, a quintessential element is the transmission network grid by which electricity, once generated, is transmitted to local distribution companies then to end users. Because of economies of scale, the per mile cost of transmitting electricity along a longer, interconnected grid is much less than doing so along a series of shorter grids (assuming, of course, that line losses are minimal). Furthermore, because some electricity is sold in bulk while the rest is sold to various classes of end users, both economies of scale and of scope arise as these multiple outputs jointly utilize this interconnected transmission grid. Furthermore, the very nature of this grid yields additional savings due to the network economies or economies of

density, which play a critical role in such an industry. According to Salvanes and Tjotta (1994), who examined the distribution function in Norway, "the characteristics of the network affect all costs and should be included by a measure of the number of nodes supplied." They asserted that, "In industries where output is delivered via a network to spatially distributed points with distinct demand characteristics and thus a continuum of outputs exists, a traditional approach with a single output to represent firm size to facilitate econometric estimation may have serious implication for measuring productivity differences."

Hence, no longer is it sufficient to measure only returns to scale; returns to density must also be considered if one is to obtain precise and relevant measures of industry structure and form appropriate public policy. Employing the definition of Caves, Christensen, and Tretheway (1984), returns to density (for the translogarithmic cost specification) are given by

$$\mathrm{RTD} = 1/(\partial \ln C/\partial \ln Y) \tag{2.22}$$

where $\partial \ln C/\partial \ln Y$ is the cost elasticity with respect to output.

Returns to density are increasing, constant, or decreasing for RTD greater than, equal to, or less than unity. Therefore, returns to density measure the economies of increasing the number of kilowatt-hours produced where the size of the network is fixed.

While only a few studies attempted to measure economies in the transmission and distribution functions, few dispute that they exist. Schmalensee (1978) asserted that: "Total distribution cost depends on the cost of transmitting services and on the spatial pattern of demand. Everywhere-decreasing average cost of transmission is found to be sufficient, but not necessary, for natural monopoly."

Nonetheless, Schmalensee developed a model to show that economies in transmission at all service flows are sufficient, but not necessary, for distribution to be a natural monopoly. Furthermore, pricing at marginal cost fails to cover total cost, and even in the presence of economies of scale in transmission, average distribution cost may rise with total demand. Of those (few) studies that attempt to quantify such economies, Huettner and Landon (1978) employed nonconventional (in that some variables are in natural logarithms while others enter as quadratics) cost functions for both transmission and distribution. They found that, for transmission, both the long-run and the short-run average variable cost curves (oddly, they do not include the fixed costs of transmitting electricity) were inverted U-shaped with the maximum occurring at a capacity of

4000 trillion MW (long-run curve) and utilization rate of 94% (short-run curve). Neither the capacity nor the utilization variables' coefficients were significant, however. For distribution, another nonconventional cost function was utilized, with the finding that the coefficients of the capacity variables were statistically significant and of the appropriate sign, generating the appropriate U-shaped long-run average variable cost curve with the minimum point occurring at a firm size of 2600 MW. However, they go on to indicate that "this U-shaped curve is somewhat L-shaped over the range of observed firm sizes." On examining the short-run average variable cost curve for distribution, they again found an inverted U-shaped curve with its maximum occurring at a 54% utilization rate. Finally, they included a measure for the density of the distribution network and found higher unit costs for more densely populated areas (this is not what they expected to find). They concluded that higher congestion costs associated with higher density overwhelm any economies that may have been present. What is interesting is that they included fixed costs in the generation component but not in either the transmission or distribution function. As previously stated, economies of scale, scope, and density are primarily the result of the highly sunk capital investments required in both transmission and distribution.

Another study that sought to measure scale economies in the distribution of electricity is that of Giles and Wyatt (1989), who examined the presence of economies of density in New Zealand. They found that the number of firms operating in the industry at the time was greater than that which was consistent with average cost minimization. They found that the cost minimizing level of output was 2315 GWh, which could have been produced efficiently by about 20 firms, 40 fewer than there actually were at the time of this study.

2.4 FOR A MULTIPLE-OUTPUT NATURAL MONOPOLY

It is well-established that distributed electricity is not a homogeneous good; that is, the electricity distributed to different types of end users can be differentiated by voltage level. For example, many industrial customers can accept electricity at much higher voltage levels than either commercial or residential customers, which is one reason why rates are set in the fashion that they are; that is, different rate and revenue classes pay different base rates (i.e., energy charges) and often residential customers do not pay demand charges. The structure of rates is discussed in more detail in the chapters on pricing and regulation. Given this, numerous studies recognize that distributed electricity should be modeled as

a multiproduct industry, which motivates the concepts described in the following section.

Multiproduct natural monopoly

While single-output scale economies imply single-output natural monopoly, this is not necessarily the case for multiple-output (or multiproduct) firms. The subadditivity conditions for a multiple-output natural monopoly are far more complex than are those of a single-output monopolist. In this case, economies of scale are not equivalent to decreasing average cost, since the firm may not operate along a linear expansion path. For a multiproduct firm, cost analysis requires the examination of not one but several concepts.

Ray average costs

Ray average costs (RAC) describe the behavior of the cost function as output is expanded proportionally along a ray emanating from the origin. Baumol et al. (1982) offer the following definition: In the two-product case, one considers the behavior of costs along a cross-section of the total cost surface. Defining a *composite* good, this measure allows a calculation of the average cost of this particular bundle and is given by

$$\text{RAC} = C(tY^0)/t \tag{2.23}$$

where Y^0 is the unit bundle for a particular mix of outputs and t is the number of units in the bundle such that $Y = tY^0$ (Baumol et al., 1982, p. 49). This is displayed in Figure 2.5.

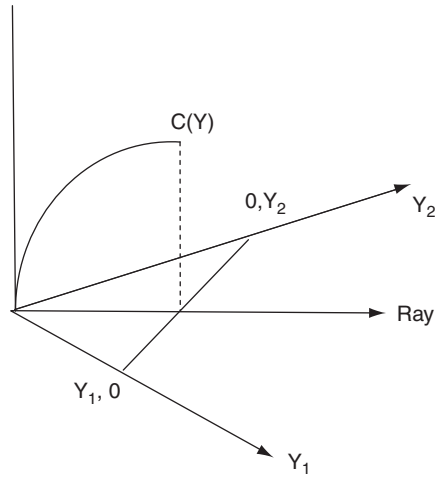

Figure 2.5 Ray average *(Source: Baumol et al., 1982, Figure 3A1. It is reproduced here with the consent of the authors.)*

Consider the behavior of costs along a cross-section of the total cost surface obtained by dropping a perpendicular plane along a ray that emanates from the origin. The ray average cost at any point on $C(Y)$ is equal to the slope of the cost function at that point. Note: In the case of Figure 2.5, *as drawn,* the slope of the cost function $C(Y)$ at $C(Y_1, Y_2) = 0$.

Degree of scale economies

As the analog to the single-output concept of economies of scale, the degree of scale economies, S_N, is equal to the ratio of average cost to marginal cost. In the multiple-output case, we have

$$S_N(Y) = C(Y)/Y_i C_i(Y), \quad \text{for } i = 1, \ldots, n \tag{2.24}$$

where $C_i(Y)$ is the marginal cost with respect to Y_i. Baumol et al. (1982, p. 51) show that

$$S_N = 1/(1 + e) \tag{2.25}$$

where e is the elasticity of RAC (tY) with respect to t at a point Y (t is a scalar).

Corollary: Returns to scale at the output point γ are increasing, decreasing, or locally constant ($S_N > 1$, $S_N < 1$, $S_N = 1$, respectively) as the elasticity of RAC at γ is negative, positive, or 0, respectively. Moreover, increasing or decreasing returns at γ imply that RAC is decreasing or increasing at Y, respectively.

As such, S_N (the degree of scale economies) may be interpreted as a measure of the percentage rate of decline or increase in ray average cost with respect to output (Baumol et al., 1982).

Cost concepts applicable to multiproduct cases for nonproportionate changes in output

As said, ray average costs are relevant when outputs move in fixed proportions, which is quite often not the case in the distribution of electricity. For this, several concepts are required to establish subadditivity of the cost function, which are discussed here in more detail.

Product-specific economies of scale

Because output is not always expanded proportionally for a multiproduct firm, the concept of product-specific economies of scale must be examined. That is, to assess the impact on cost of a change in one output, holding other outputs constant, one must examine the average incremental cost (AIC) of the product of which output is being varied. This is defined as

$$\mathrm{AIC}(\gamma_i) = [C(Y_N) - C(Y_{N-i})]/Y_i \tag{2.26}$$

where $C(Y_{N-i})$ is the cost of producing all N of the multiproduct firm outputs except product i.

This specification allows the identification of returns to scale that are specific to a particular output. Hence, product-specific returns to scale are given by

$$S_i(\gamma) = \mathrm{AIC}(\gamma_i)/(\partial C/\partial \gamma_i) \tag{2.27}$$

where $\partial C/\partial \gamma_i$ is the marginal cost with respect to product i.

Returns to scale of product i at γ are said to be increasing, decreasing, or constant as $S_i(\gamma)$ is greater than, less than, or equal to unity, respectively.

If product 2(Y_2), as shown in Figure 2.6, has no output-specific fixed costs, then the total cost surface rises continuously above ST (curve AE). However, if there exists some special fixed cost that must be incurred to begin production of Y_2 as an addition to the firm's line of other products,

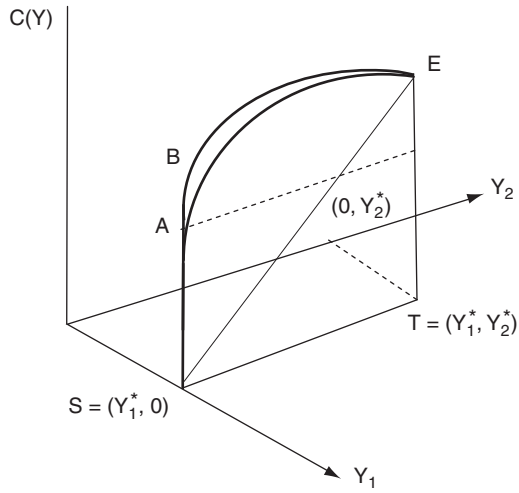

Figure 2.6 Product-specific returns to scale *(Source: Baumol et al., 1982, Figure 4A2. It is reproduced here with the consent of the authors.)*

then the cross-section of the cost surface contains a vertical fixed cost segment, AB, which results in a jump discontinuity of $C(Y)$ above the Y_1 axis. Thus, the height CE in Figure 2.6 measures the total incremental cost of Y_2 at output vector **T,** which is the addition to the firm's total cost resulting from the decision to add Y_2 to the firm's product mix. The average incremental cost of Y_2, $\mathrm{AIC}_2(Y_1^*, Y_2^*)$, is clearly given by the slope of the line from A to E. What is also clear from Figure 2.6 is that the average incremental costs of product 2 decline with Y_2, at least between 0 and Y_2^*. This suggests, by analogy to the single-output case, the novel and useful concept of product-specific scale economies.

Economies of scope

The multiproduct cost concepts discussed prior to this relate to the behavior of cost along a cross-section of the cost-output space. In addition to economies that result from the size or scale of a firm's operations, other cost savings can result from the production of several outputs at the same time; that is, in many cases and certainly in the case of electricity, there are fixed costs that are jointly utilized in the production of the firm's outputs. These common costs, as they are also known, give rise to the concept of economies of scope (or economies of horizontal integration) and provide a basis for determining whether an industry is a multiproduct natural monopoly.

Mayo (1984) argues that: "In addition to measures of scale, efficient industry structure is determined by the behavior of costs as the scope of the firm is altered. The cost savings or dissavings that result from multiproduct versus specialized firm operations are given by the notion of economies and diseconomies of scope."

Therefore, economies of scope (also known as *economies of joint production*) are said to exist if a given quantity of each of two or more goods can be produced by one firm at a lower cost than if each good were produced by two different firms or even two different production processes. That is, for a two-product case, weak economies of scope are given by

$$C(Y_1, Y_2) \leq [C(Y_1, 0) + C(0, Y_2)] \tag{2.28}$$

for all Y_1, $Y_2 > 0$. If not, then there are diseconomies of scope, and separate production of outputs is more efficient.

As in the single-output case, we define the degree of economies of scope, which is given by

$$S_c = [C(Y_1, 0) + C(0, Y_2) - C(Y_1, Y_2)]/C(Y_1, Y_2) \tag{2.29}$$

The importance of economies of scope cannot be overstated: Economies of scope are a necessary condition for natural monopoly in a multiple output firm.

Both economies of scale and of scope tend to occur due to specialization. As stated previously, the latter can arise from the sharing or joint utilization of inputs. According to Panzar and Willig (1977), if a given input is imperfectly divisible, production of a small set of goods may leave excess capacity in the utilization of that input. Another way that economies of scope can arise is that the input may have some properties of a public good, so that when it is purchased for one production process, it can then be freely available to another. A third way is that economies of scope can arise due to the economies of networking (recall the discussion of network economies and returns to density).

Subadditivity of the cost function

However, even if a cost function exhibits both economies of scale and economies of scope, it is not necessarily subadditive. A sufficient condition must now be established for natural monopoly in a multiproduct industry. Cost complementarity, which requires that marginal or incremental costs of any output decline when that output or any other outputs increase, provides such a condition. Mathematically, cost complementarity for a twice differential multiproduct cost function exists if

$$\partial^2 C(Y)/\partial Y_i \partial Y_j < 0, \text{for } i \neq j \tag{2.30}$$

and for all Y_i, $Y_j > 0$.

If this is satisfied, then the cost function exhibits cost complementarity, which is a sufficient condition for subadditivity in a multiproduct cost function. An industry is said to be a natural monopoly if, over the entire relevant range of outputs, the firm's cost function is subadditive.

2.5 ELECTRICITY AS A MULTIPLE-OUTPUT INDUSTRY AND ECONOMIES OF SCOPE AND SUBADDITIVITY

While treating generated electricity as a homogeneous good may seem appropriate, it is certainly not appropriate treatment for distributed electricity. A quote from Joskow and Schmalensee (1983, pp. 54–55) summarizes this principle nicely: "treating diverse power systems as single-product firms is likely to produce error. The cost of an optimally designed power system depends in complex ways on the distribution of demand over time and

space. No two power systems produce the same mix of products and product mix differences affect the magnitude and form of optimal investments in transmission and in distribution."

Among the first to actually model electricity as a multiple-output function was Neuberg (1977), who examined the market for distribution employing four interdependent outputs. The predominant output in his analysis was the number of customers served; the other outputs were the number of megawatt-hours sold, the size of distribution territory, and the miles of overhead distribution line. Karlson (1986) tested for and found that the multiproduct characterization of electricity was appropriate, having treated residential and commercial electricity as distinct outputs. He found that the marginal cost of any one output depended on the levels of all other outputs and all other inputs. Furthermore, he rejected the hypothesis of separability between inputs and outputs, which implies that the marginal rate of substitution between any two inputs is not independent of the quantities of outputs nor is the marginal rate of transformation between any two outputs independent of the quantities of inputs.

Some studies attempted to identify the existence of scope economies in this industry. Mayo (1984) employed a multiproduct quadratic cost function to estimate the cost of producing both electricity and gas for 200 public utilities. Using 1979 data, he confirmed the presence of economies of scope for smaller firms. However, as output is expanded, the absence of competitive pressure leads to cost inefficiencies and eventual diseconomies of scope. His finding led to the realization that the regulated utilities in his sample were characterized by interproduct discomplementarities, since his empirical results confirmed that

$$\partial^2 C(Y)/\partial Y_1 \partial Y_2 > 0 \tag{2.31}$$

This particular result, which was anticipated by Kahn, can be attributed (at least in part) to the *type* of regulation imposed on these firms; that is, average cost pricing in which firms are certainly not incentivized to minimize costs. Furthermore, the Averch-Johnson effect,[1] which is also a result of

[1] Traditional rate making provides an incentive to overinvest in capital (i.e., the rate base). For an investor-owned utility, this is a large component of the revenue requirement on which the utility is allowed to earn a return to its investors. Known as the *Averch-Johnson effect*, this is the tendency of companies to engage in excessive amounts of capital accumulation to expand the volume of their profits. If a firm's profits-to-capital ratio is regulated at a certain percentage, then there is a strong incentive for companies to overinvest to increase profits overall. This goes against any optimal efficiency point for capital that the company may have calculated as higher profit is almost always desired over and above efficiency.

rate-of-return regulation, is still at play, since firms have an incentive to overinvest in capital as a mechanism to increase rates and hence profits.

Like Mayo's 1984 study, Sing (1987) employed a different cost specification (a generalized translog cost function) to estimate whether a sample of U.S. electric and gas utilities were natural monopolies. He found that the average combination utility exhibited diseconomies of scope, but other output combinations were associated with economies of scope.

Roberts (1986) and Thompson (1997) differentiated output according to voltage level. Their results suggest that there are economies of density; that is, for a given network size and a fixed number of customers, average costs fall when the quantity of power supplied increases. Roberts defined a measure of economies of output density as

$$R_D = 1/(\partial \ln C/\partial \ln Y_L + \partial \ln C/\partial \ln Y_H) \tag{2.32}$$

where Y_L and Y_H denote low-voltage and high-voltage output, respectively.

He rejected the hypothesis of separability of the generation and transmission functions from distribution. This was predominantly due to the lack of separability between the inputs required to perform all three functions, which is the reason that a majority of the utilities in the United States are vertically integrated. (This confirms Karlson's finding.) The concept of economies of vertical integration is explored later in this chapter and throughout this book, since it is integral to the appropriate cost modeling and public policy making for electric utilities. In addition, it provides the subject of a case study presented in Chapter 8.

As previously stated, most of the studies of this nature focus on investor-owned utilities. However, and as stated in the introductory chapter, other types of entities are worthy of such analysis. Yatchew (2000) estimated the costs of distributing electricity using data on municipal electric utilities in Ontario, Canada, for the period 1993–1995. The data reveal substantial evidence of increasing returns to scale with minimum efficient scale being achieved by firms with about 20,000 customers. Larger firms exhibit constant or decreasing returns. Utilities that deliver additional services (such as water and sewage), have significantly lower costs, indicating the presence of economies of scope.

Greer (2003) estimated economies of scale and scope for U.S. distribution cooperatives. Distributed electricity (i.e., output) was differentiated by voltage level, with 1000 kVA being the distinction between "small" users and "large" users. She found that the cost function exhibits

product-specific economies as well as economies of scope, and substantial cost savings could be realized via mergers between distribution coops. This study and the cost model used in the analysis are the subject of much more detail as well as the case studies examined in Chapters 7 and 8.

Fraquaelli, Piancenza, and Vannoni (2004) studied Italian public utilities that provided the combination of gas, water, and electricity. They confirmed the presence of global scope and scale economies only for multiutilities, with output levels lower than the ones characterizing the "median" firm. This indicates that relatively small specialized firms would benefit from cost reductions by evolving into multiutilities, providing similar network services such as gas, water, and electricity. However, for larger-scale utilities, the hypothesis of null cost advantages is not rejected. Therefore, it is possible that the recent diversification waves of leading companies are explained by factors other than cost synergies, so that the welfare gains that can be reasonably expected from such examples of horizontal integration, if any, are likely to be very low.

2.6 ECONOMIES OF VERTICAL INTEGRATION AND SEPARABILITY

The issue with which we are dealing is the appropriate modeling of costs to formulate public policy that maximizes the total welfare of the players (both consumers and producers) involved. Thus far, we have been concerned with the separate stages (or processes) required to supply electricity to end users and whether each stage (or process) may be a natural monopoly. What has been established is that the generation component, due mostly to technological change, is no longer a natural monopoly and there could be societal gains from allowing competition into that component of the process, which is what the deregulation of the industry was all about. Unfortunately, what was essentially ignored was the *network* (or wires) aspect of the business; that is, unlike telephony (voice, data, fax—more on this in Chapter 5, a case study on deregulation and the breaking up of the Bell System), water, and natural gas, electricity cannot be economically stored and, once generated, flows according to Kirchoff's law (i.e., the path of least resistance).[2]

[2] This is a critical point that needs to be kept in mind. In my opinion, it is the reason that deregulation of the industry was such an abject failure.

Given this, what now needs to be established is the relationship between these three functional components. After all, part of the notion of deregulation was that generation could be separated (lack of scale economies, so competition was deemed feasible) from the transmission and distribution functions (irrefutably natural monopolies). More specifically, what is necessary is to establish the existence (or lack thereof) of economies of vertical integration, which are another critical and distinguishing aspect of this industry?

Vertical integration of electric utilities

Landon (1983) argued that "the electricity industry has special characteristics such as close coordination of each process, transaction costs, and idiosyncratic capital requirements, which all favour vertical integration."

Vertical integration makes sense when a product is produced sequentially, such that the output from the first stage of production is employed as an input in successive stages, which is the case of electricity. When a firm is vertically integrated, it owns the entire production process, controlling both the upstream (input) supply and the downstream (output) production processes. Needless to say, the electric utility industry in the United States was organized in this fashion for a number of years by investor-owned firms, who were willing to supply power to the larger, more densely populated areas of the country. Vertical integration makes sense since it provides an alternative to market transactions, which tend to be costly given the nature of the industry, which requires specialized assets and sunk costs. It would have been extremely difficult to foresee the input-price increases experienced since the mid-1970s; were the industry not vertically integrated but rather contractually related, the financial difficulties experienced by utilities in the late 1970s–1990s would have been far greater, since it is unlikely that these price increases were foreseen and could be written into the contracts, which were typically of longer duration.

Vertical integration is especially appealing in industries characterized by bottlenecks, which tend to occur with exclusive ownership of a resource necessary to the production of the good but whose cost is prohibitive, so that it is not economically feasible for separate firms to invest. This type of investment yields a market that approximates a natural monopoly in the sense that its cost is sunk and its duplication would be wasteful. In the case of electricity, the bottleneck is that which yields access to the transmission mechanism that delivers electricity from generation to the local distribution system.

Additional benefits are attributable to vertical integration as well:

1. The elimination of the "wedge" that results when the upstream firm sells its product to the downstream firm at a price above economic cost.
2. The mitigation of certain problems that arise due to the separation of ownership of the firm from whoever actually controls it, which is also known as a *principal-agent problem*.

For the presence of economies of vertical integration in the supply of electricity, it must be the case that successive stages of production (generation, transmission, and distribution) are less costly for a single firm to perform than for these functions to be performed by separate producers. Both issues are relevant in the production of electricity, whose underlying production technology not only lends itself to economies of scale but also to economies of vertical integration.

Defining vertical integration

Mathematically, economies of vertical integration exist if the following is satisfied:

$$C(G, D) < C(G, 0) + C(0, D) \tag{2.33}$$

where $G > 0$ represents the first stage of production (upstream production) and $D > 0$ represents the latter stage (downstream production), so that $C(G, D)$ is the cost of production for a vertically integrated firm. If this is less than the sum of the cost of separate production by separate entities, given by $C(G, 0) + C(0, D)$, then it is said that there exist economies of vertical integration. Or, expressed in percentage terms,

$$S_v = [C(G, 0) + C(0, D) - C(G, D)]/C(G, D) \tag{2.34}$$

where

$S_v > 0$, there are economies of vertical integration.

$S_v < 0$, there are no economies of vertical integration.

Separability

Because the marginal cost of any one output depends on the levels of all other outputs and all other inputs, the issue of separability must be considered on the formation of appropriate policy. Karlson (1986, p. 78) states that: "Separability between inputs and outputs requires that the marginal rate of substitution between any two inputs is independent of the quantities of outputs, and the marginal rate of transformation between any two

outputs is independent of the quantities of inputs . . . The rejection of the hypothesis of separability between inputs and outputs implies that the relative marginal costs of electricity sold to different consumer classes depend on the product and input mixes; furthermore, it is impossible to construct some homogeneous aggregate output called 'electricity' to be sold to consumers."

Karlson rejects the hypothesis of such separability, as do Henderson (1985), Roberts (1986), and Lee (1995). These are discussed in more detail next.

2.7 RELEVANT LITERATURE REVIEW—VERTICAL INTEGRATION AND SEPARABILITY

Several studies tested for the presence of vertical economies in the supply of electricity. Virtually all of them test for and reject the separability of the functional components. In fact, it has been empirically demonstrated that there exist economies of vertical integration in the production of electricity. Such studies include Henderson (1985), who finds downstream costs are dependent on input usage at the generation stage, hence the cost function (which is translogarithmic) fails the test for separability between generation and distribution. Roberts (1986) concurs, as do Hayashi, Yeoung-Jia Goo, and Chamberlain (1997) and Thompson (1995). Other studies include those by Kaserman and Mayo (1991) and Gilsdorf (1994, 1995). As an extension to their testing for vertical economies, both Kaserman and Mayo (1991) and Gilsdorf (1994, 1995) employ a multiproduct cost function to determine whether vertical integration and economies of scale together constitute a natural monopoly. In fact, Kaserman and Mayo also test for multistage economies between generation and transmission/distribution. They too reject the separability of inputs and outputs (what is generated is an input to what is transmitted or distributed) in the cost function. It is important to note that separability is not the same thing as economies of vertical integration, where output-output interactions matter. Byung-Joo Lee (1995) estimated a production function and performed more direct tests for vertical integration and economies of scale. All reject separability of all three functional components of electricity production. Kwoka (1996) employed the Kaserman and Mayo approach to test for multistage economies between generation and distribution. He too rejects separability (especially for "larger" systems) and argues that these vertical economies are precisely the reason that most investor-owned utilities are

vertically integrated, while most "smaller" systems (i.e., publicly owned utilities and rural electric cooperatives) are not. In addition he found that vertical integration achieves significant cost efficiencies, in some cases, sufficient to offset diseconomies of scale in generation and distribution separately.

More recent studies include Goto and Nemoto (2004), who test the technological externality effects of generation assets on the costs of transmission and distribution stages in their study of vertically integrated Japanese utilities. Their results show that downstream costs depend on the generation capital, suggesting significant economies of vertical integration. Fraquaelli, Piancenza, and Vannoni's (2005) analysis of Italian municipal electric utilities finds significant vertical economies for average size and large utilities while failing to find any significant effects for smaller than average size utilities. Efficiencies associated with vertical integration are largest for fully integrated utilities, confirming results found in most other studies. Greer (2008) estimated the lost economies of vertical integration due to the rural electric cooperatives' choice of market structure. As indicated in the introductory chapter, cooperatives are organized as either generation and transmission or member coops (distribution only). Greer found that cost savings of close to 40% could be realized had they adopted a truly vertically integrated structure. This paper and the cost models used to generate these results are the basis for the case study presented in Chapter 8.

2.8 CONCLUSION

This chapter provides an overview of electric utility industry structure and some relevant cost concepts as well as a brief survey of the literature pertaining to this industry. In subsequent chapters, these concepts are expounded on and examined in much more detail.