4

Firms

4.1 What Is a Firm?

Key ideas: firm as a transformer of inputs into outputs, production sets, production functions, net supply and net demand

In this chapter the focus switches to the transformation of commodities by firms. Within a firm, raw materials and other commodity inputs are processed by labor and managerial inputs to produce goods and services. These outputs may be for consumption (final products) or for sale as inputs to other firms (intermediate products). The amount of output that can be produced depends on the technology (machinery, buildings, etc.) held by the firm.

This is relatively straightforward. Consider, for example, a newsprint manufacturer. It transforms the primary raw materials of lumber, energy, and labor into giant rolls of paper ready for delivery to daily newspapers, using an array of machines. However, from a broader perspective, the machines are also inputs. In addition to purchasing labor inputs and raw materials, the firm can purchase additional capital equipment (for the same plant or to build a new plant) and so alter the set of available outputs. From this perspective, the technology of a firm is a set of blueprints for the transformation of commodities.

Even this is an incomplete view. Rather than purchase additional equipment, the firm might purchase another firm, or maybe all the other firms in the economy! Do regulatory constraints determine the number of firms, or are there some other natural factors that limit the scope for the agglomeration of firms into super-firms? These very important and controversial questions lie at the heart of modern industrial organization. Here we simply take the boundary of each firm (or its set of blueprints) as a given.

If there were no economies of scale there would be no technological rationale for economic agents to work together as partners or in an employer–employee relationship. Thus, gains to specialization are a core explanation for the existence of firms. However, once there is an agglomeration of

human and non-human inputs, another critical issue concerns the effective utilization of these inputs. In particular, the performance of workers must be monitored. In very small firms the owner bears this cost directly. In larger firms the monitoring of workers requires the recruiting of managers. In turn, the managers are monitored by the owners as represented by a board of directors. Thus, in either case the ultimate monitoring cost is a cost borne by the owners.[1]

In this chapter we abstract from such issues and ignore the cost of monitoring the utilization of inputs to carry out the plan selected by the owners of a firm.[2] These owners receive dividends based on their shareholdings. The higher the profit of the firm the higher is the dividend stream. Thus the owners seek to maximize the profit of the firm.

Consider a firm that can produce m possible outputs $q = (q_1, \ldots, q_m)$ using n inputs $z = (z_1, \ldots, z_n)$. The $m + n$ dimensional vector $(z, q) \geq 0$ is a feasible plan of firm f if the output vector q is feasible given the input vector z.

Let Y^f be the set of feasible input-output vectors for firm f. This is the firm's *production set*. For any input vector z, let $Q(z)$ be the set of feasible output vectors. An output vector q is output efficient for the firm if it is not possible to increase the output of one commodity without decreasing the output of some other commodity. Mathematically, for a given input vector z, an output vector is output-efficient if it lies on the boundary of $Q(z)$.

In the special case in which there is only a single output, the efficient output $q = F(z)$ is the maximum feasible output. The function $F(\cdot)$ is then referred to as the firm's *production function*.

Example 1: Cobb-Douglas Production Function

$$q = A \overset{n}{\underset{j=1}{\times}} z_j^{\alpha_j} = A z_1^{\alpha_1} \ldots . z_n^{\alpha_n}, A, \alpha > 0.$$

Example 2: CES Production Function

$$q = \left(\sum_{j=1}^{n} a_j z_j^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}, a, \sigma > 0, \sigma \neq 1.$$

With more than one output, it is often convenient to characterize the production set as the set of input-output vectors (z, q) that satisfy certain inequality constraints. That is,

$$\mathcal{Y}^f = \{(z, q) | h_k(z, q) \geq 0, k = 1, \ldots, K\}.$$

[1] In reality top managers are offered bonuses and other incentives. Thus a key role of the owners is to design effective incentive schemes. We return to this in Chapter 7.

[2] Monitoring costs can be added without affecting the analysis in any significant way if these costs are independent of the production plan. Then we simply have an additional fixed cost.

The characteristics of the production set thus depend on the characteristics of the constraint functions. For example, if each of the constraint functions is quasi-concave (so that it has a convex upper-contour set), then the production set is convex.

Example 3: Multi-Product Production Set

$$\mathcal{Y}^f = \{(z_1, q_1, q_2) | z_1 - q_1^2 - q_2^2 \geq 0, (z, q) \geq 0\}.$$

In this case the constraint function is the sum of three concave functions and is therefore concave so the production set is convex. Note that for any input z_1 the set of feasible output vectors $Q(z) = \{q | q_1^2 + q_2^2 \leq z_1\}$. This is a quarter-circle with radius $\sqrt{z_1}$.

At the level of the individual firm, the focus is often on the cost of the inputs and the revenue from the output. However, when the focus is on markets as a whole, this approach is not so useful. One reason is that the outputs of one firm are very often the inputs of another firm. A second reason is that, as relative prices change, a firm can switch from being a net demander of some commodity (such as electricity) to being a net supplier.

A more general approach focuses on the net supply of commodities by each firm. Firm f is a net supplier of commodity j if $y_j^f \geq 0$ and it is a net demander of commodity j if $y_j^f < 0$. In this case $-y_j^f$ is the firm's input demand.

Given a price vector p, the total revenue of the firm is $R^f = \sum_{j,\, y_j \geq 0} p_j y_j^f$ and the total cost is $C^f = \sum_{j,\, y_j < 0} p_j(-y_j^f)$. The profit of the firm is therefore

$$R^f - C^f = \underbrace{\sum_{j, y_j > 0} p_j y_j^f}_{\text{revenue}} - \underbrace{\sum_{j,\, y_j \leq 0} p_j(-y_j^f)}_{\text{cost}} = p \cdot y^f.$$

Thus, there is a third important reason for this alternative approach. It is far more convenient from a mathematical perspective. The set of feasible plans $\mathcal{Y}^f$ is the firm's production set,[3] and any vector $y^f \in \mathcal{Y}^f$ is a feasible production vector. To satisfy the interests of its shareholders, the firm chooses its most profitable feasible production vector.

4.2 Decision Making by Price-Taking Firms

Key ideas: input prices and marginal cost, laws of firm demand and supply, short- versus long-run response to a price change

[3] If the production vector is $(z^f, q^f) \geq 0$ where z^f and q^f are the input and output vectors then we write the production set as Y^f. When y^f is a firm's net supply we write the production set as $\mathcal{Y}^f$.

Initially we assume that the firm is a price taker in all input markets. A necessary condition for profit maximization is that, given the output choice q, the cost of the inputs should be minimized. This minimized cost of production is known as the cost function of the firm.

Cost Function

Let Y be the production set of firm f and let r be the vector of input prices. Then the cost function is

$$C(r, q) \equiv \underset{z}{\text{Min}}\{r \cdot z | (z, q) \in Y\}.$$

If the firm has a single output we define $F(z)$ to be the maximum output given the input vector z. This is called the firm's production function. In this case the cost function is

$$C(r, q) \equiv \underset{z}{\text{Min}}\{r \cdot z | F(z) - q \geq 0\}.$$

Note that from a mathematical perspective, if there is a single output, then the cost function of the firm is identical in structure to the expenditure function of the consumer. Thus, all of our analysis of the expenditure function carries over directly. In particular, the cost function is a concave function of the input price vector. We have the following further result.

Lemma 4.2-1: Gradient of the Cost Function If the cost-minimizing input vector $z(q, r)$ is a continuous function of input prices, then

$$\frac{\partial C}{\partial r_i}(r, q) = z_i(r, q) \, i = 1, \ldots, n. \tag{4.2-1}$$

This is readily proved by appealing to the Envelope Theorem. However it is best understood by making a direct argument. Suppose that input vector z^0 is optimal with input price vector r^0 and that z^1 is optimal with price vector r^1. When the input price vector changes from r^0 to r^1 one option is for the firm to produce the output vector q using the same input vector. Thus $C(r^1, q) \leq r^1 \cdot z^0$ and so we have the following upper bound for the change in cost:

$$C(r^1, q) - C(r^0, q) \leq (r^1 - r^0) \cdot z^0.$$

Also, starting from input price vector r^1, the change in minimized cost has the following upper bound:

$$C(r^0, q) - C(r^1, q) \leq (r^0 - r^1) \cdot z^1.$$

Combining these results,

$$(r^1 - r^0) \cdot z^1 \leq C(r^1, q) - C(r^0, q) \leq (r^1 - r^0) \cdot z^0. \tag{4.2-2}$$

Suppose only the ith input price changes. Then (4.2-2) can be rewritten as follows

$$z_i(r^1, q) \leq \frac{C(r^1, q) - C(r^0, q)}{r_i^1 - r_i^0} \leq z_i(r^0, q).$$

Then if the input demand function $z(r, q)$ is continuous $\lim_{r^1 \to r^0} z(r^1, q) = z(r^0, q)$.

Therefore, in the limit the lower and upper bounds converge and so $\frac{\partial C}{\partial r_i} = z_i(r^0, q)$.

Output Effect of an Input Price Increase

With this simple result, we can gain insight into the response of a firm to an input price increase. The price-taking firm chooses output so that price equals marginal cost. It is tempting to believe that an increase in r_j will raise the marginal cost of producing each commodity. Thus, as depicted in Figure 4.2-1, the output of each commodity will fall. However, as we now demonstrate, this intuition is incorrect. We assume that $z(r, q) \in \mathbb{C}^1$. Then differentiating equation (4.2-1),

$$\frac{\partial^2 C}{\partial q_i \partial r_j} = \frac{\partial z_j}{\partial q_i}.$$

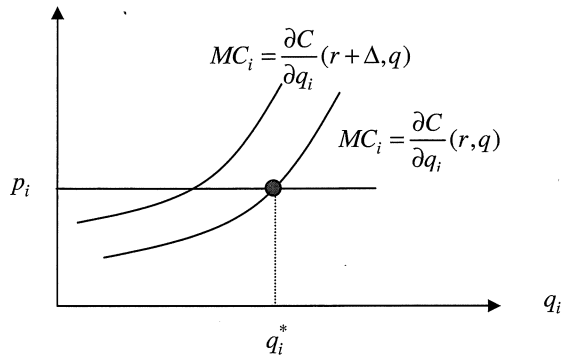

Figure 4.2-1. The normal case.

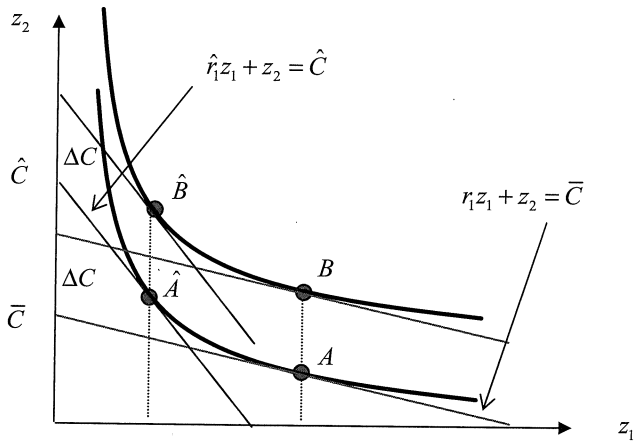

Figure 4.2-2. Isoquants are vertical translations.

The order of partial differentiation is immaterial. Moreover, $\partial C/\partial q_i$ is the marginal cost of producing commodity i. Thus,

$$\frac{\partial}{\partial r_j} MC_i = \frac{\partial}{\partial r_j}\frac{\partial C}{\partial q_i} = \frac{\partial}{\partial q_i}\frac{\partial C}{\partial r_j} = \frac{\partial z_j}{\partial q_i}. \tag{4.2-3}$$

The partial derivative $\partial z_j/\partial q_i$ is the rate at which the jth input rises when the output of commodity i rises. If this rate is positive, the input is called a normal input. If not, the input is said to be an inferior input. We therefore have the following result.

Proposition 4.2-2: Effect of Input Price Change on Marginal Cost
A rise in the price of input j raises the marginal cost of commodity i if and only if z_j is a normal input.

It seems paradoxical that marginal cost could remain constant or decline with an increase in an input price. However this result is readily illustrated.[4] Consider the production function $q = F(z) = G(h(z_1) + z_2)$, where $h(\cdot)$ is an increasing concave function and $G(\cdot)$ is increasing. For such a production function, the marginal rate of technical substitution ($MRTS$) is

$$MRTS(z_1, z_2) = \frac{\partial F}{\partial z_1} \Big/ \frac{\partial F}{\partial z_2} = h'(z_1).$$

Note that the $MRTS$ is independent of z_2 so that isoquants are vertical translations of one another as depicted in Figure 4.2-2.

Setting the price of input 2 equal to unity, the iso-cost line, $r_1 z_1 + z_2 = \overline{C}$ intersects the vertical axis at $\overline{C}$. With the price of input 1 equal to r_1, the

[4] See also Exercise 4.2-4.

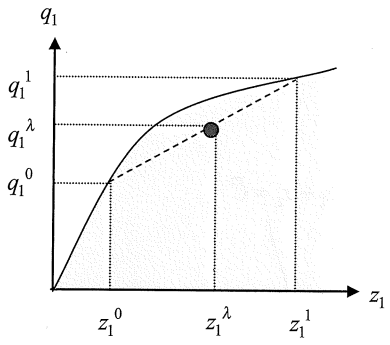

Figure 4.2-3. Convex production set.

lowest cost method of producing q units of output is given by the point A and the lowest cost method of producing $q + \Delta q$ units of output is given by the point B. The extra cost of the additional Δq units is thus the vertical distance $\Delta C = AB$.

Next suppose that the price of input 1 rises to $\hat{r}_1$ and that the new cost minimizing method of producing q units is the point $\hat{A}$. As depicted, the isoquants are vertical translations. Therefore, the lowest cost method of producing $q + \Delta q$ units is the point $\hat{B}$, which is located directly above $\hat{A}$.

Moreover, because the isoquants are vertical translations, the distance $\hat{A}\hat{B}$ is equal to the distance AB. However $\hat{A}\hat{B}$ is the extra cost ΔC of producing the additional Δq units. Thus in this particular case a change in input price has no effect on marginal cost.

Convexity of the Production Set and Increasing Marginal Cost

Consider a firm that produces a single output using a single input. If the production set is convex, as depicted in Figure 4.2-3, the extra output produced using an additional unit of input declines with output. Thus, the marginal cost of producing an additional unit of output rises with output. The following proposition generalizes this result.

Proposition 4.2-3: Convex Costs
If the production set of a firm is convex, then the cost function of the firm is a convex function of outputs.

Proof: Let z^0 be cost minimizing given an input price vector r and output q^0. Also let z^1 be cost minimizing when the output vector is q^1. Given convexity of the production set, the production vector (z^λ, q^λ), $0 < \lambda < 1$ is feasible. Then the minimized cost of output vector q^λ satisfies

$$C(q^\lambda, r) \leq r \cdot z^\lambda. \tag{4.2-4}$$

By construction, $C(q^0, r) = r \cdot z^0$ and $C(q^1, r) = r \cdot z^1$. Therefore,

$$(1-\lambda)C(q^0, r) + \lambda C(q^1, r) = (1-\lambda)r \cdot z^0 + \lambda r \cdot z^1$$
$$= r \cdot ((1-\lambda)z^0 + \lambda z^1) = r \cdot z^\lambda.$$

Substituting from (4.2-4) it follows that

$$C(q^\lambda, r) \leq (1-\lambda)C(q^0, r) + \lambda C(q^1, r), 0 < \lambda < 1. \qquad \square$$

First Laws of Firm Supply and Demand

We now extend the price-taking assumption to output prices. So far we have denoted the outputs of a firm as a positive vector q and the inputs as a positive vector z.

To simplify the mathematics we now switch to the alternative formulation in which a production plan is simply a vector $y^f = (y_1^f, \ldots, y_n^f)$. All the positive components of this plan are net outputs, and all the negative components are net inputs. The profit of a firm choosing the production vector y is

$$p \cdot y = \underbrace{\sum_{j,\, y_j \geq 0} p_j y_j}_{\text{revenue}} - \underbrace{\sum_{j,\, y_j < 0} p_j(-y_j)}_{\text{cost}}.$$

The firm chooses from among the set of possible plans or blueprints $\mathcal{Y}^f$. The profit-maximizing firm then chooses a plan $\overline{y}^f$ that solves the following problem:

$$\Pi(p) = \underset{y}{\text{Max}}\{p \cdot y | y \in \mathcal{Y}^f\}.$$

The maximized profit $\Pi(p)$ is known as the profit function. The following result also has its parallel in Section 2.3.

Proposition 4.2-4: Effect of Price Changes on Inputs and Outputs
Let p^0 and p^1 be two different price vectors and let y^0 and y^1 be profit maximizing production plans at these prices. Then $(p^1 - p^0) \cdot (y^1 - y^0) \geq 0$.

It follows immediately that if only the price of commodity j changes, then

$$\Delta p_j \Delta y_j \geq 0.$$

If the jth commodity is an output ($y_j > 0$), it follows that

$$\frac{\Delta y_j}{\Delta p_j} \geq 0. \text{ first law of supply}$$

If the jth commodity is an input ($y_j < 0$), then $-y_j$ is the number of inputs purchased and

$$\frac{-\Delta y_j}{\Delta p_j} \leq 0. \text{ first law of input demand}$$

Proof: The proof follows directly from the profit-maximization hypothesis. Because y^1 is profit maximizing at p^1,

$$\Pi(p^1) = p^1 \cdot y^1 \geq p^1 \cdot y, \quad \text{for all } y \in \mathcal{Y}^f. \tag{4.2-5}$$

In particular, this must be true for $y = y^0$. Hence,

$$p^1 \cdot (y^1 - y^0) \geq 0.$$

Also, because y^0 is profit maximizing at p^0,

$$\Pi(p^0) = p^0 \cdot y^0 \geq p^0 \cdot y, \quad \text{for all } y \in \mathcal{Y}^f.$$

Hence,

$$-p^0 \cdot (y^1 - y^0) = p^0 \cdot (y^0 - y^1) \geq 0. \tag{4.2-6}$$

Adding these inequalities,

$$(p^1 - p^0) \cdot (y^1 - y^0) \geq 0. \qquad \square$$

We actually have a stronger result. Let p^λ be a convex combination of the two price vectors. That is, $p^\lambda = (1-\lambda)p^0 + \lambda p^1, 0 < \lambda < 1$. Also let y^* be profit maximizing at p^λ. From (4.2-5),

$$\Pi(p^1) \geq p^1 \cdot y^*. \quad \text{Hence } \lambda\Pi(p^1) \geq \lambda p^1 \cdot y^*.$$

From (4.2-6),

$$\Pi(p^0) \geq p^0 \cdot y^*. \quad \text{Hence } (1-\lambda)\Pi(p^0) \geq (1-\lambda)p^0 \cdot y^*.$$

Adding these inequalities yields

$$\lambda\Pi(p^1) + (1-\lambda)\Pi(p^0) \geq ((1-\lambda)p^0 + \lambda p^1) \cdot y^* = p^\lambda \cdot y^* = \Pi(p^\lambda).$$

We thus have the following result.

Proposition 4.2-5: Convexity of the Profit Function
$\Pi(p) \equiv \text{Max}_y\{p \cdot y | y \in \mathcal{Y}^f\}$ is a convex function.

Short-Run and Long-Run Adjustments to a Price Change

Suppose that a price-taking firm has a profit-maximizing production plan y^0 given the price vector p^0. Then the price of one of its inputs or outputs unexpectedly changes. A full analysis of the firm's response to this change

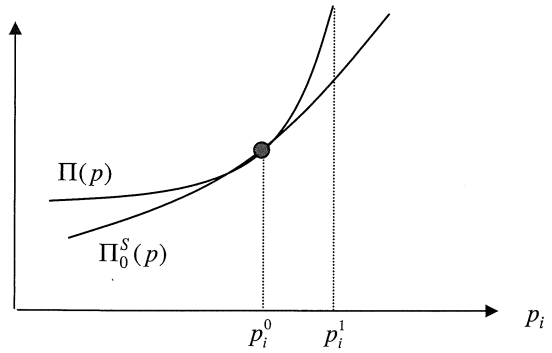

Figure 4.2-4. Short run versus long run.

would require a model of the dynamics of costly adjustment. As a first simple step toward such a model, we suppose that the firm is more limited in its feasible alternatives over the short run than over the long run.[5] Let $\mathcal{Y}$ be the set of feasible long-run plans. Starting from y^0 let $\mathcal{Y}^s(y^0)$ be those production plans that are feasible in the short run. Then $\mathcal{Y}^s(y^0) \subset \mathcal{Y}$.

Proposition 4.2-6: Le Chatelier Principle
Own price effects are larger in the long-run than in the short-run.

Proof: Let $\Pi(p)$ and $\Pi_0^S(p)$ be the long-run and short-run profit functions of the firm, starting from the long-run profit-maximizing plan at p^0. The additional short-run constraints bind more tightly so that for all price vectors $p \neq p^0$,

$$\Pi_0^S(p) < \Pi(p).$$

This is depicted in Figure 4.2-4 when it is the price of the ith commodity that changes.

Note that the additional short-run constraints bind only if the firm *changes* its production plan. Hence $\Pi_0^S(p^0) = \Pi(p^0)$. We further assume that the profit-maximizing production vector $y(p) \in \mathbb{C}^1$. Suppose that the price vector changes from p^0 to p^1. Because $y(p^0)$ is most profitable at the price vector p^0,

$$\Pi(p^1) - \Pi(p^0) = p^1 \cdot y(p^1) - p^0 \cdot y(p^0) \leq p^1 \cdot y(p^1) - p^0 \cdot y(p^1).$$

Similarly, because $y(p^1)$ is most profitable at the price vector p^1,

$$\Pi(p^1) - \Pi(p^0) = p^1 \cdot y(p^1) - p^0 \cdot y(p^0) \geq p^1 \cdot y(p^0) - p^0 \cdot y(p^0).$$

[5] This is a generalization of the traditional definition of the short run as a period over which some subset of the inputs is fixed.

Combining these results,

$$(p^1 - p^0) \cdot y(p^1) \geq \Pi(p^1) - \Pi(p^0) \geq (p^1 - p^0) \cdot y(p^0).$$

Suppose that only the ith price changes. Then

$$y_i(p^1) \geq \frac{\Pi(p^1) - \Pi(p^0}{p_i^1 - p_i^0}) \geq y_i(p^0).$$

Taking the limit;

$$\frac{\partial \Pi}{\partial p_i} = y_i(p) \quad \text{and so} \quad \frac{\partial^2 \Pi}{\partial p_i^2} = \frac{\partial y_i}{\partial p_i}.$$

An identical argument for the short-run profit function establishes that

$$\frac{\partial \Pi_0^S}{\partial p_i} = y_i^S(p) \quad \text{and} \quad \frac{\partial^2 \Pi_0^S}{\partial p_i^2} = \frac{\partial y_i^S}{\partial p_i}.$$

Thus the two profit functions have the same slope at p^0. Because $\Pi(p) \geq \Pi_0^S(p)$ it follows that $\frac{\partial^2 \Pi}{\partial p_i^2}(p^0) \geq \frac{\partial^2 \Pi_0^S}{\partial p_i^2}(p^0)$ and hence $\frac{\partial y_i}{\partial p_i} \geq \frac{\partial y_i^S}{\partial p_i}$. □

Exercises

Exercise 4.2-1: Cost Minimization

(a) Prove that the cost function is a concave function of the input price vector.
(b) Assuming that the cost-minimizing input vector $z^*(r, q)$ is continuously differentiable, show that $\partial z_i^*/\partial r_j = \partial z_j^*/\partial r_i$.

Exercise 4.2-2: Cost Function of a Cobb-Douglas Firm What is the cost function of a firm with a Cobb-Douglas production function?

Exercise 4.2-3: CES Production Function A firm has a production function $q = z_1^{\frac{1}{2}} + z_2^{\frac{1}{2}}$.

(a) Show that minimized cost is

$$C(r, q) = \frac{q^2}{\frac{1}{r_1} + \frac{1}{r_2}}.$$

(b) Solve for the cost function if the production function is instead $q = (z_1^{\frac{1}{2}} + z_2^{\frac{1}{2}})^2$.

Exercise 4.2-4: Quasi-Linear Production Function Output q is produced according to the following production function:

$$q = z_1 + 10\sqrt{z_2}.$$

(a) Show that for q sufficiently large

$$C(r, q) = r_1 q - 25 r_1^2 / r_2.$$

(b) Show that $C(r, q)$ is proportional to q^2 for sufficiently small q.
(c) Depict $C(r, q)$ for $q \geq 0$ in a neat figure.
(d) In a second figure depict marginal cost as a function of q for two different values of r_2.
(e) If the output is sold by a monopolist, describe the effect on output of an increase in r_2. Does it make a difference whether the profit-maximizing output q^* is large or small?

Exercise 4.2-5: Marginal Cost Independent of an Input Price Consider the production function $q = F(h(z_1) + z_2) \in \mathbb{C}^2$ where $F(\cdot)$ is increasing, $h(\cdot)$ is an increasing concave function, $h'(0) = \infty$, and $h'(\infty) = 0$.

(a) Given an input price vector r, show that for large enough outputs, $z_2(r, q)$, the cost-minimizing level of z_2 must be strictly positive. Henceforth assume that $z_2(r, q) > 0$.
(b) Write down total cost as a function of $F^{-1}(q)$ and z_1. Hence show that the optimal input of commodity 1, $z_1(r, q)$ is independent of q.
(c) Show that marginal cost is independent of r_1.

Exercise 4.2-6: Properties of Firm Input Demand and Output Supply Let $z^*(r)$ be the profit-maximizing input vector for a monopolist producing outputs $(q_1, \ldots, q_m)$ using inputs $(z_1, \ldots, z_n)$. The firm is a price taker in input markets. Let r be the input price vector.

(a) Assume that $z^*(r)$ is continuously differentiable and show that the matrix $\left[\frac{\partial z_i}{\partial r_j}\right]$ is negative semi-definite.
(b) If the firm is also a price taker in output markets, what can be said about the $m \times m$ matrix $\left[\partial q_i^* / \partial p_j\right]$?

Exercise 4.2-7: Short Run versus Long Run Suppose a firm is a price taker in its inputs but has monopoly power in its output markets. Does the Le Chatelier Principle still hold for inputs?

4.3 Returns to Scale

Key ideas: global and local returns to scale, implications for average and marginal cost

In the previous section we examined the implications of price-taking behavior by firms. The plausibility of the price-taking hypothesis depends, to a great extent, on whether it is technologically advantageous for a firm to be

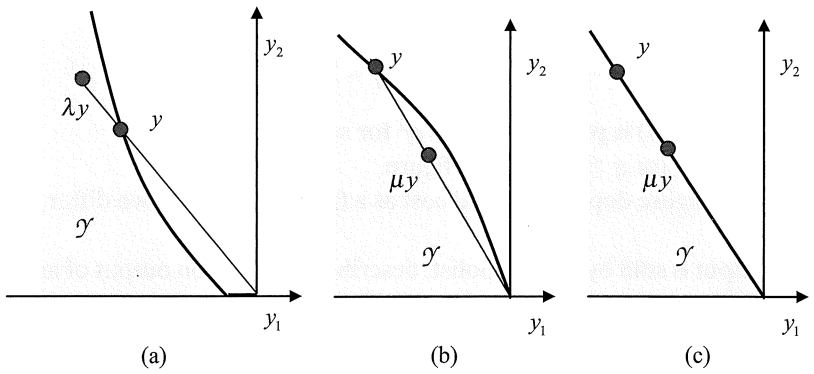

Figure 4.3-1. (a) Increasing returns to scale. (b) Decreasing returns to scale. (c) Constant returns to scale.

large. If, as a result of increasing inputs by some factor λ, outputs are scaled up by the same factor λ, then the technology exhibits constant returns to scale. If outputs are smaller than λ times the original output vector, then the technology exhibits decreasing returns to scale. If outputs exceed λ times the original output vector, then the technology exhibits increasing returns to scale. When there are decreasing returns to scale, it is hard for a large firm to compete with smaller firms, and so the number of firms in the market is likely to be relatively large. It is in such situations that a firm is most likely to lose much of its sales when it raises prices and hence the price-taking assumption is most plausible.

Definition: Returns to Scale[6] The production set $\mathcal{Y} \subset \mathbb{R}^n$ exhibits constant returns to scale if, for all $y \in \mathcal{Y}$, and any $\lambda > 0$, $\lambda y \in \mathcal{Y}$. The production set exhibits increasing returns to scale if, for any $y \in \mathcal{Y}$, such that $y_j \neq 0$, $j = 1, \ldots, n$ and any $\lambda > 1$, $\lambda y \in \text{int}\mathcal{Y}$. The production set exhibits decreasing returns to scale if, for any $y \in \mathcal{Y}$ such that $y_j \neq 0$, $j = 1, \ldots, n$ and any $\mu \in (0, 1)$, $\mu y \in \text{int}\mathcal{Y}$.

We begin by showing that this definition yields the familiar definitions for a firm producing a single output.

Increasing Returns to Scale Production Function

Consider a firm with a production function $F(\cdot)$. Suppose that the production set exhibits increasing returns to scale (IRS; see Figure 4.3-1a). With input vector z the maximum feasible output is $F(z)$. With input vector λz the maximum output is $F(\lambda z)$. From the definition of IRS, for any $\lambda > 1$ the input-output vector $(\lambda z, \lambda q) = (\lambda z, \lambda F(z))$ lies in the interior of the

[6] If $y \in \mathcal{Y}$ and y is not a boundary point, then y is called an interior point, and we write $y \in \text{int}\,\mathcal{Y}$.

production set. Thus, with the input λz the maximum output must be greater than $\lambda F(z)$. Therefore, for any $z > 0$,

$$\lambda > 1 \Rightarrow F(\lambda z) > \lambda F(z). \tag{4.3-1}$$

Decreasing Returns to Scale Production Function

If the production function exhibits decreasing returns to scale (DRS), then for any $\mu \in (0, 1)$, $(\mu z, \mu F(z))$ is in the interior of the production set (see Figure 4.3-1)b. Therefore, with the input μz the maximum output $F(\mu z)$ is greater. Thus

$$\mu \in (0, 1) \Rightarrow F(\mu z) > \mu F(z), \forall z > 0. \tag{4.3-2}$$

This holds for all $z > 0$. Hence, for any λ it holds for λz. That is,

$$\mu \in (0, 1) \Rightarrow F(\mu\lambda z) > \mu F(\lambda z), \lambda > 1.$$

Choose $\lambda > 1$ and $\mu = 1/\lambda$. Then $F(z) > \frac{1}{\lambda} F(\lambda z)$, $\lambda > 1$.
Rearranging this inequality we have the standard definition for DRS.

$$\lambda > 1 \Rightarrow F(\lambda z) < \lambda F(z), \forall z > 0. \tag{4.3-3}$$

Constant Returns to Scale Production Function

If a firm exhibits constant returns to scale (CRS) any input-output vector can be scaled up or down. Therefore, as depicted in Figure 4.3-1c, any scaled version of a boundary point must also be a boundary point. To demonstrate this, consider a CRS firm with a production function $F(\cdot)$. With input vector z the maximum feasible output is $F(z)$. Because $(z, F(z))$ is feasible, the CRS assumption implies that $(\lambda z, \lambda F(z))$ is feasible. With input vector λz the maximum output is $F(\lambda z)$. Thus

$$F(\lambda z) \geq \lambda F(z).$$

Moreover, because $(\lambda z, F(\lambda z))$ is feasible the CRS assumption implies that $\frac{1}{\lambda}(\lambda z, F(\lambda z)) = (z, \frac{1}{\lambda} F(\lambda z))$ is also feasible. However, with input vector z, $F(z)$ is maximum output. Therefore

$$\frac{1}{\lambda} F(\lambda z) \leq F(z).$$

Combining the two inequalities it follows that

$$F(\lambda z) = \lambda F(z).$$

Remark: Suppose that the total supply of inputs to an industry is z and that the allocation to firm k is $z^k = \theta^k z$. If technology exhibits CRS and can be freely replicated, then

$$\sum_{k=1}^{n} F(z^k) = \sum_{k=1}^{n} F(\theta^k z) = \sum_{k=1}^{n} \theta^k F(z) = F(z).$$

It follows that any industry output $q = F(z)$ can be equally well produced by a single firm or many firms. As a result the number of firms is indeterminate.

We now consider some important properties of CRS production functions. We begin by showing that the gradient vector of the production function is constant along a ray.

Proposition 4.3-1:
Suppose F exhibits constant returns to scale and is differentiable for all $z \gg 0$. Then for all $z \gg 0$, $\frac{\partial F}{\partial z}(\lambda z) = \frac{\partial F}{\partial z}(z)$.

Proof: Given CRS, for any $\hat{z}$ and $\lambda > 0$, $F(\lambda \hat{z}) = \lambda F(\hat{z})$. Differentiating this equation by $\hat{z}_j$ we obtain

$$\lambda \frac{\partial F}{\partial z_j}(\lambda \hat{z}) = \lambda \frac{\partial F}{\partial z_j}(\hat{z}).$$

□

The following proposition is a restatement of results derived in Appendix B.[7]

Proposition 4.3-2: Constant Returns to Scale and Concavity
If the function F is strictly quasi-concave and exhibits CRS, then for any z^0, z^1 and $\lambda \in (0, 1)$

$$F((1-\lambda)z^0 + \lambda z^1) \geq (1-\lambda)F(z^0) + \lambda F(z^1).$$

Moreover, the inequality is strict unless z^0 and z^1 are linearly dependent.

Returns to Scale and the Scale Elasticity of Output

Define $q(\lambda) = F(\lambda z)$. Then the proportional increase in maximum output as the scale parameter rises from 1 to λ is

$$\frac{1}{q(1)} \frac{q(\lambda) - q(1)}{\lambda - 1} = \frac{1}{F(z)} \frac{F(\lambda z) - F(z)}{\lambda - 1}.$$

Taking the limit yields the scale elasticity of output

$$\mathcal{E}(q(\lambda), \lambda)\big|_{\lambda=1} = \frac{\lambda}{F(z)} \frac{\partial}{\partial \lambda} F(\lambda z)\bigg|_{\lambda=1}.$$

[7] See Proposition B.2-9 and Corollary B.2-10.

With DRS, it follows from inequalities (4.3-2) and (4.3-3), that for all $\lambda > 0$, $\lambda \neq 1$,

$$\frac{1}{F(z)}\frac{F(\lambda z) - F(z)}{\lambda - 1} < \frac{1}{F(z)}\frac{\lambda F(z) - F(z)}{\lambda - 1} = 1.$$

Taking the limit as $\lambda \to 1$ we obtain the following result

$$\text{DRS} \Rightarrow \mathcal{E}(F(\lambda z), \lambda)\big|_{\lambda=1} \leq 1.$$

A similar argument establishes that

$$\text{IRS} \Rightarrow \mathcal{E}(F(\lambda z), \lambda)\big|_{\lambda=1} \geq 1.$$

Local Returns to Scale

The assumption of (global) increasing or decreasing returns is a very strong one. Typically firms exhibit increasing returns at low output levels because of indivisibilities in entrepreneurial setup and monitoring costs. As output grows large, the costs of monitoring a large managerial work force and providing appropriate work incentives typically rise more rapidly than output. These cost increases can more than offset any purely technological advantages to greater scale.

It is therefore helpful to consider returns to scale locally. Local returns are increasing at the input vector z if a small proportional increase in z leads to a larger proportional increase in output. Local returns are decreasing if the proportional increase in inputs leads to a smaller proportional increase in output.

Because the point elasticity $\mathcal{E}(y, x) \equiv \frac{x}{y}\frac{\partial y}{\partial x}$, it follows that

$$\mathcal{E}(q(\lambda), \lambda) = \frac{\lambda}{F(\lambda z)}\frac{\partial}{\partial \lambda}F(\lambda z) = \frac{z \cdot \frac{\partial F}{\partial z}(\lambda z)}{F(\lambda z)}.$$

Hence the scale elasticity at z can be expressed as

$$\mathcal{E}(q(\lambda), \lambda)\big|_{\lambda=1} = \frac{z \cdot \frac{\partial F}{\partial z}(z)}{F(z)}, \quad \text{where } q(\lambda) \equiv F(\lambda z).$$

Definition: Local Returns to Scale If the production function $F(\cdot)$ has scale elasticity $\mathcal{E}(q(\lambda), \lambda)\big|_{\lambda=1} = z \cdot \frac{\partial F}{\partial z}/F(z)$, greater than 1 then $F(\cdot)$ exhibits increasing returns at z. If the scale elasticity is less than 1 then $F(\cdot)$ exhibits decreasing returns at z.

In the exercises that follow, you are asked to establish that if a production function exhibits local increasing returns to scale everywhere, it exhibits (global) increasing returns.

We next show that average cost exceeds marginal cost if and only if local returns are increasing.

Proposition 4.3-3: Average and Marginal Cost
If z is cost minimizing for output q, then

$$\frac{AC(q)}{MC(q)} = \frac{z \cdot \frac{\partial F}{\partial z}}{F(z)} = \mathcal{E}\,(q(\lambda), \lambda)\big|_{\lambda=1}. \tag{4.3-4}$$

Proof: Given the input price vector r, the cost function is

$$C(q, r) = \underset{z}{\text{Min}}\{r \cdot z | q \leq F(z)\}.$$

Converting this to a maximization problem, the associated Lagrangian is

$$\mathfrak{L} = -r \cdot z + \lambda(F(z) - q).$$

By the Envelope Theorem,

$$MC(q) = \frac{\partial C}{\partial q} = \lambda.$$

Moreover the necessary conditions for an interior maximum are

$$\frac{\partial \mathfrak{L}}{\partial z_i} = -r_i + \lambda \frac{\partial F}{\partial z_i} \leq 0,\ i = 1, \ldots, n \text{ with equality if } z_i > 0.$$

Multiplying this inequality by $z_i \geq 0$, it follows that $-r_i z_i + \lambda z_i \frac{\partial F}{\partial z_i} = 0$.

Then total cost is

$$C(q, r) = r \cdot z = \lambda z \cdot \frac{\partial F}{\partial z}. \tag{4.3-5}$$

Thus the average cost of production is

$$AC(q) = \frac{C(q)}{q} = \lambda \frac{z \cdot \frac{\partial F}{\partial z}}{F(z)} = MC(q) \frac{z \cdot \frac{\partial F}{\partial z}}{F(z)}.$$ □

Note next that

$$MC(q) = \frac{\partial}{\partial q} C(q) = \frac{\partial}{\partial q}(q\,AC(q)) = AC(q) + q \frac{\partial AC}{\partial q}.$$

With local increasing returns, average cost is greater than marginal cost and therefore $AC(q)$ is a decreasing function of q. With local decreasing returns, on the other hand, $AC(q)$ is an increasing function of q.

Exercises

Exercise 4.3-1: Increasing Returns to Scale If the strictly increasing production function $F(\cdot)$ exhibits IRS show that for all $z \neq 0$ and $\mu \in (0,1)$, $F(\mu z) < \mu F(z)$.

Exercise 4.3-2: Returns to Scale and Average Cost Prove that if a firm exhibits increasing/decreasing returns to scale then average cost must decrease/increase with output.

Exercise 4.3-3: Local and Global Returns to Scale

(a) Show that if a production function $F(z)$ exhibits local increasing returns to scale everywhere, then the scale elasticity $\mathcal{E}(F(\mu z), \mu) > 1$ for all z and all $\mu > 0$.
(b) Hence show that $\frac{\partial}{\partial \mu} \ln F(\mu z) > \frac{1}{\mu}$.
(c) Show that for any $\lambda > 1$,

$$\ln \frac{F(\lambda z)}{F(z)} = \int_1^{\lambda} \frac{\partial}{\partial \mu} \ln F(\mu z) d\mu.$$

Then appeal to part (b) to establish that the production function exhibits (global) IRS.

Exercise 4.3-4: Modified Cobb-Douglas Production Function The production function of a firm is defined implicitly as follows:

$$q = K^{\frac{\alpha}{q}} L^{\frac{\beta}{q}}, \alpha, \beta > 0.$$

(a) Given input prices (r, w), show that the cost-minimizing input demands satisfy

$$\frac{\alpha}{rK} = \frac{\beta}{wL} = \frac{\alpha + \beta}{C(q)}.$$

(b) Hence or otherwise obtain an expression for the firm's cost function.
(c) If $\alpha + \beta = 1$, show that the average cost function is U-shaped, with a minimum at $q = 1$.
(d) Does a change in an input price have any effect on the cost-minimizing output?

4.4 Firm and Industry Analysis

Key ideas: optimality of marginal cost pricing, equilibrium with free entry, scale of competitive firms

It is efficient for an agglomeration of inputs to be organized together in a firm if and only if there are gains to specialization in some part of the production process. If the gains to specialization persist as the firm grows, it is efficient to have all production within a single firm – a "natural monopoly." This

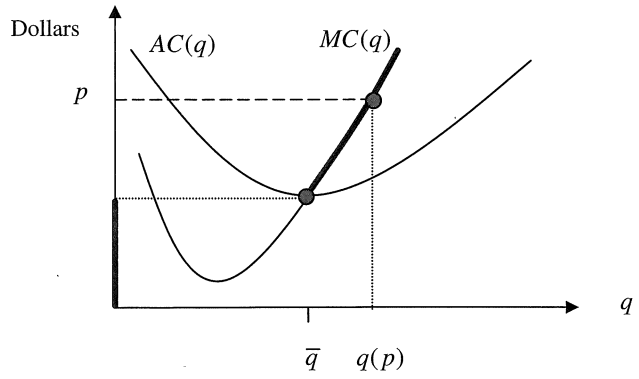

Figure 4.4-1. A firm's supply curve.

is considered in the next section. Here we consider an industry where the returns to specialization occur only at output levels that are small relative to the market. At higher output levels, large firms are less efficient, and thus we would expect to see competition between several or many small firms in such a market.

Consider a firm that produces a single output. The production function of the firm exhibits local increasing returns to scale (local IRS) at output levels below $\overline{q}$ and local decreasing returns to scale (local DRS) at higher output levels. Then the average and marginal cost curves must be as depicted in Figure 4.4-1. For $q < \overline{q}$, MC is less than AC and so AC is decreasing. For $q > \overline{q}$, MC is greater than AC and so AC is increasing.

Suppose that firm j is small enough relative to the market to make the assumption of price taking plausible. If the output price is strictly below the minimum average cost, the firm must lose money regardless of its level of production and therefore the profit-maximizing output is zero. If the price exceeds the minimum average cost, profit is maximized at the output $q_j(p)$ where marginal cost is equal to price. Thus for $p > AC(\overline{q})$, the marginal cost curve is the firm's supply curve.

The Social Optimum

Suppose there are n price-taking firms in an industry. Let $p(Q)$ be the market demand price function. The aggregate supply at the equilibrium price p must equal aggregate demand. Moreover, for each firm, marginal cost is equal to price. Thus,

$$p(Q) = MC_j(q_j),\ j = 1, \ldots, n, \quad \text{where} \quad Q = \sum_{j=1}^{n} q_j. \tag{4.4-1}$$

Define $C(Q)$ to be the minimized cost of producing the aggregate output Q. That is,

$$C(Q) = \underset{q}{\text{Min}} \left\{ \sum_{j=1}^{n} C_j(q_j) | \sum_{j=1}^{n} q_j \geq Q \right\}.$$

For all firms that produce a positive output, the first-order condition for minimizing the industry cost of production is

$$MC_j(q_j) = \lambda(Q), \text{ where } Q = \sum_{j=1}^{n} q_j.$$

Note that this is exactly the same condition as for each of the n price-taking firms. It follows that, given our price taking assumption, total industry cost is minimized. Hence for any aggregate output Q, the industry cost is the aggregate cost function $C(Q)$.

For any industry supply $\hat{Q}$ the demand price $p(\hat{Q})$ is the market-clearing price. Therefore it is the marginal willingness to pay or "marginal social benefit."

Next define $B(\hat{Q})$ to be the area under the market demand curve. That is,

$$B(\hat{Q}) = \int_0^{\hat{Q}} p(q)dq.$$

This integral of marginal social benefit is the total social benefit.[8]

We can also write total cost as the integral of the industry marginal cost:

$$C(\hat{Q}) = \int_0^{\hat{Q}} MC(q)dq.$$

Social surplus is then

$$S(\hat{Q}) = B(\hat{Q}) - C(\hat{Q}) = \int_0^{\hat{Q}} [p(q) - MC(q)]dq.$$

Social surplus is the area between the industry demand curve and the industry marginal cost curve. This is the dotted area depicted in Figure 4.4-2.

[8] See Chapter 2 for a discussion of the limitations of this measure of social benefit.

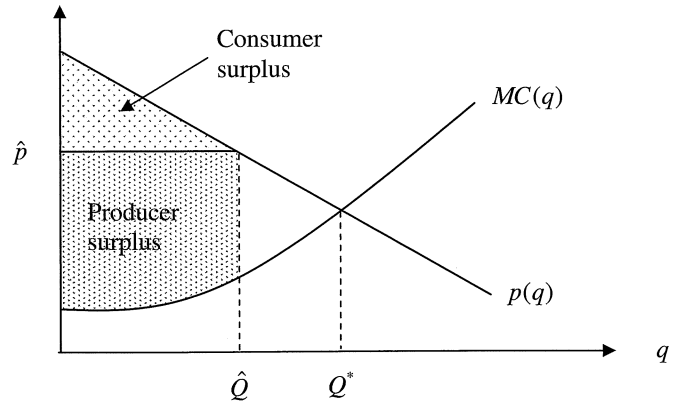

Figure 4.4-2. Social surplus.

For an output of $\hat{Q}$, demand equals supply at the price $\hat{p} = p(\hat{Q})$. It is sometimes helpful to divide the social surplus into the gains to consumers and the gains to firms:

$$S = \underbrace{B(\hat{Q}) - \hat{p}\hat{Q}}_{\text{consumer surplus}} + \underbrace{\hat{p}\hat{Q} - C(\hat{Q})}_{\text{producer surplus}}.$$

Consumer surplus is the lightly dotted region between the demand price function and the market-clearing price line. Producer surplus is the heavily dotted region between the industry marginal cost curve and the market-clearing price line.

From Figure 4.4-2, social surplus is maximized by choosing an industry output Q^* such that $p(Q^*) = MC(Q^*)$. Thus, marginal cost pricing is socially optimal.

Industry Supply with Free Entry

We now apply our methods of analysis to an industry consisting of a large number of identical firms with free entry and exit. Suppose that each firm produces a single product and that marginal cost rises with output. Furthermore, assume that there are fixed costs of production. Marginal and average costs are then as depicted in Figure 4.4-3.

With the price of output equal to p^o, each firm maximizes profit by choosing output q^o such that price equals marginal cost. As depicted, $AC(q^0) < p^o$ and thus there is an incentive for additional firms to enter the market. Entry then increases aggregate supply and so the market-clearing price must fall. Because this argument holds for any price above $\overline{p}$, industry equilibrium is

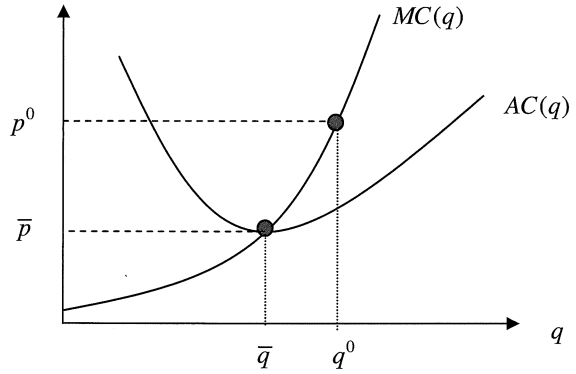

Figure 4.4-3. AC and MC.

only reached when the price is $\overline{p}$. In the long run, the industry supply curve is therefore a horizontal line.[9]

We now explore how a change in an input price affects the equilibrium output price. Suppose that the cost share of input j is k_j. A naïve answer might go as follows. A 1 percent increase in the cost of input j has a 1 percent effect on the cost of the input. Because input j's cost share is only a fraction k_j of the total input cost, the latter will rise by k_j per cent. Yet, if total cost rises by k_j percent then average cost will rise by k_j percent. In a competitive industry with free entry $p = AC$ and thus the equilibrium price must also rise by k_j percent.

This argument is clearly incomplete in that it ignores both the substitution effect on input as an input price rises and the change in the average cost-minimizing output. Despite this, the answer is correct!

Proposition 4.4-1:
In an industry with free entry of identical firms, the elasticity of the equilibrium output price with respect to an input price change is equal to the input cost share.

Proof: Given the assumption of free entry of identical firms the equilibrium output price is equal to the minimum average cost. Appealing to the Envelope Theorem, $\frac{\partial C}{\partial r_j}(q,r) = z_j$. Hence, $\frac{\partial}{\partial r_j} AC(q,r) = \frac{z_j}{q}$ and so the elasticity of average cost is

$$\mathcal{E}(AC, r_j) \equiv \frac{r_j}{AC}\frac{\partial}{\partial r_j} AC(q,r) = \frac{r_j z_j}{qAC} = \frac{r_j z_j}{C(q,r)} = \frac{r_j z_j}{r \cdot z}.$$

Thus, holding output constant, $\mathcal{E}(AC(q,r), r_j) = \frac{r_j z_j}{r \cdot z}$. As we will see later, the average cost-minimizing output may rise or fall when an input price

[9] This is not quite true. With m firms in the market each may be profitable, whereas with $m+1$ firms the market-clearing price may drop below the minimum average cost. Then, the equilibrium number of firms is m and equilibrium profit is small but positive. However, unless m is small the statement is a useful practical approximation.

changes. However, at the point of minimum average cost $\overline{q}(r)$, $\frac{\partial}{\partial q} AC(q, r) = 0$. Recalling that the equilibrium price is equal to minimized average cost,

$$dp = dAC(\overline{q}(r), r_j) = \frac{\partial}{\partial q} AC(\overline{q}, r) d\overline{q} + \frac{\partial}{\partial r_j} AC(\overline{q}, r) dr_j = \frac{\partial}{\partial r_j} AC(\overline{q}, r) dr_j.$$

Converting this into an elasticity,

$$\mathcal{E}(p, r_j) = \frac{r_j}{p} \frac{\partial p}{\partial r_j} = \frac{r_j}{p} \frac{\partial AC}{\partial r_j} = \mathcal{E}(AC, r_j), \quad \text{since } p = AC. \qquad \square$$

Why is the naive answer correct? First, as argued in Section 4.2, the substitution effect on input demand is a second-order effect because competitive firms are cost minimizers. (This is the Envelope Theorem at work.) Second, although it is true that the cost-minimizing output changes, equilibrium output is where average cost is minimized and is therefore where the slope of the AC curve is zero. Thus, the change in output has only a second-order effect on minimized average cost.

An input price increase results in an output price increase and so leads to a reduction in aggregate output. Yet what happens to the output of the firms remaining in the industry? An instinctive guess would probably be that each firm would cut output as well. However, the new equilibrium output is at the intersection of the MC and AC curves. Thus, output of a surviving firm falls if and only if MC cost rises more than AC when the input price rises. We have seen that there are cases when MC is not affected by an input price change; then MC lies below AC cost at the initial output level and, with a price increase, output must rise until MC and AC are equated. Figure 4.4-4 depicts the case in which both MC and AC rise but the former rises less, so again output of the surviving firms must rise.

We conclude by providing the necessary and sufficient conditions for the output of surviving firms to rise.

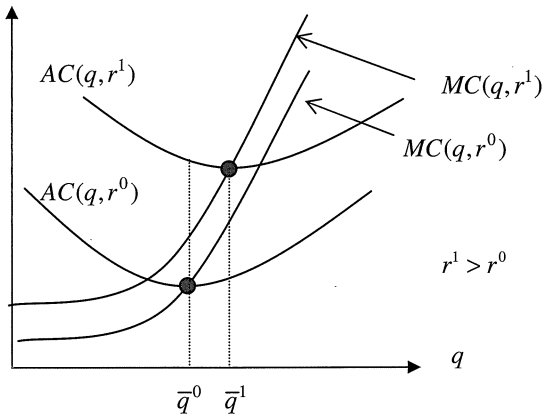

Figure 4.4-4. Input price effect on firm scale.

Proposition 4.4-2: Scale of Competitive Firms
If all firms have the same technology, an increase in the price of an input increases the output of firms remaining in a competitive industry (with free entry) if and only if the output elasticity of the demand for the input is less than unity.

Proof: From Section 4.2 we know that *(i)* $\frac{\partial C}{\partial r_j} = z_j$ and *(ii)* $\frac{\partial}{\partial r_j} MC = \frac{\partial}{\partial r_j}\frac{\partial C}{\partial q} = \frac{\partial z_j}{\partial q}$. Because $AC = C/q$, it follows that *(iii)* $\frac{\partial}{\partial r_j} AC = \frac{z_j}{q}$. Combining (*ii*) and (*iii*),

$$\frac{\partial}{\partial r_j} MC = \mathcal{E}(z_j, q)\frac{\partial}{\partial r_j} AC.$$

Suppose that at the initial lower input prices, the average cost-minimizing output is $\overline{q}^0$. At this output, average and marginal costs are equal. Suppose also that $\mathcal{E}(z_j, q) < 1$. Then, holding output constant, MC rises less than AC when an input price rises. Then $AC(\overline{q}^0, r^1) > MC(\overline{q}^0, r^1)$ and so AC is declining at $\overline{q}^0$. Then the new AC-minimizing output $\overline{q}^1$ exceeds $\overline{q}^0$. □

Exercises

Exercise 4.4-1: The Social Optimum Firm j has a cost function $C_j(q_j)\, j = 1, \ldots, n$. The demand price function is $p(Q)$.

(a) Write down the social surplus optimization problem. Then form the Lagrangian and hence obtain the first-order conditions.
(b) Show that for the social optimum,

$$\frac{dB}{dQ} \leq MC_j(q_j)\, j = 1, \ldots, n \text{ with equality if } q_j > 0.$$

(c) Hence establish the optimality of the marginal cost pricing rule.

Exercise 4.4-2: Marginal Cost Pricing with Many Commodities A consumer has utility function $U(q, x, x_0) = B(q_1, q_2) + F(x) + x_0$. Let $p = (p_1, p_2)$ be the price vector for q and let r be the price vector for $x = (x_1, \ldots, x_n)$. Normalize so that the price of x_0 is 1. Assume throughout that income I is large enough for demand for this commodity, $x_0(p, r, I)$ to be strictly positive.

(a) Show that demand for all the other commodities is independent of income. Assume, henceforth, that for all p above some upper bound $\overline{p}$, $q(p) = 0$.
(b) Explain why indirect utility is

$$V(p, r, I) = B(q(p)) + f(x(r)) + x_0(p, r, I),\ p < \overline{p}$$

and

$$V(p, r, I) = f(x(r)) + x_0(p, r, I),\ p \geq \overline{p}.$$

(c) Hence explain why the gain in utility from a price $p < \overline{p}$ can be written as follows:

$$\Delta V = \int_0^{q_1(p)} \frac{\partial B}{\partial q_1}(q_1, 0)dq_1 + \int_0^{q_2(p)} \frac{\partial B}{\partial q_2}(q_1(p), q_2)dq_1.$$

(d) Let $p_i(q)$, $i = 1, 2$ be the demand price functions. Appeal to the FOC to show that

$$\Delta V = \int_0^{q_1(p)} p_1(q_1, 0)dq_1 + \int_0^{q_2(p)} p_2(q_1(p), q_2)dq_1.$$

Exercise 4.4-3: Scale of Competitive Firms There is free entry into an industry of firms with the same U-shaped AC curve.

(a) Show that for each firm $MC = \sum_{j=1}^{n} \frac{r_j z_j}{q} \mathcal{E}(z_j, q)$. Hence show that, at the average cost-minimizing output,

$$1 = \sum_{j=1}^{n} k_j \mathcal{E}(z_j, q), \quad \text{where } k_j = \frac{r_j z_j}{r \cdot z}.$$

(b) Hence in the two-input case show that if

$$\mathcal{E}(z_2, q) > \mathcal{E}(z_1, q), \quad \text{then} \quad \mathcal{E}(z_2, q) > 1 > \mathcal{E}(z_1, q).$$

(c) A firm has production function $F(z) = (z_1 - a)^{\alpha} z_2^{\beta}$, $z_1 \geq a > 0$, $z_2 \geq 0$. The input price vector is r. Show that along the output expansion path $r_2 z_2 = \frac{\beta}{\alpha}(r_1 z_1 - a r_1)$. Hence show that along the output expansion path, $\mathcal{E}(z_2, q) > 1 > \mathcal{E}(z_1, q)$.

(d) Hence show that the scale of the active firms will rise if one of the input prices rises and the scale will fall if the other rises.

4.5 Monopoly Pricing

Key ideas: natural monopoly, price discrimination, indirect price discrimination

If there are enough firms in a market, the influence of any one firm's decisions on the market equilibrium price is small. Thus, the price-taking assumption of the Walrasian equilibrium is a good first approximation. If AC rises with output, a small firm is more efficient than a large firm, and so monopolization of a market is impossible (without government regulation or collusion). Conversely, if average cost decreases with higher output, a firm can cut prices to the point where the profit of any smaller competitor is negative. Thus, if increasing returns are in effect up to market output levels, then monopoly is natural.

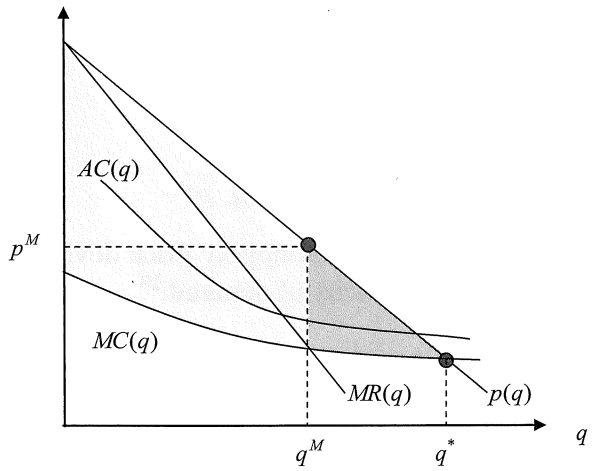

Figure 4.5-1. Social optimum with decreasing AC.

From a cost perspective, production by a natural monopoly is efficient. To maximize profit, the monopoly minimizes its costs, and given increasing returns to scale, these costs are lower than the costs of producing the same level of output by two or more firms. However, on the demand side, the monopolist is not a price taker; as it increases its output the market-clearing price falls. Thus, the marginal revenue from an output increase is less than the market price.

The profit-maximizing firm equates MR and MC. As we have seen, social surplus is maximized if the demand price (marginal social benefit) and marginal cost are equated. Thus, the profit-maximizing monopolist reduces social surplus by undersupplying the market.

The social optimum and the monopoly outcome are depicted in Figure 4.5-1. The social benefit from producing q units is the area under the demand price function. The total cost is the area under the marginal cost function. Thus, the social surplus is

$$S(q^M) = B(q^M) - C(q^M) = \int_0^{q^M} p(q)dq - \int_0^{q^M} MC(q)dq.$$

This is the lightly shaded region in the figure. The social cost of a monopoly is the extra surplus generated by increasing to q^* and is represented by the more heavily shaded area.

Note that

$$MR(q) = \frac{d}{dq}(p(q)q) = p(q) + q\frac{dp}{dq} = p(q)\left(1 + \frac{q}{p}\frac{dp}{dq}\right)$$
$$= p(q)\left(1 + \frac{1}{\mathcal{E}(q, p)}\right),$$

where $\mathcal{E}(q, p)$ is the price elasticity of demand. Hence, the profit-maximizing price is

$$p(q) = \frac{MC(q)}{1 + \dfrac{1}{\mathcal{E}(q, p)}}.$$

Therefore the degree to which the monopoly price deviates from marginal cost depends on the market elasticity of demand.[10]

Price Discrimination by Group

Implicit in this discussion is the idea that the monopolist can charge only a single price. If the cost of reselling by market intermediaries is low, this is a reasonable assumption. However, for many commodities, the cost of resale is significant, and a monopolist can charge different prices to different groups of customers.[11] Shipping costs are one example of a barrier to resale. If a firm locates outlets in two disparate locations and shipping costs by consumers (or middlemen) are sufficiently high, a monopoly firm can charge different prices in each location. Let $p_i(q_i)$ be the demand price function for group i. If there are n groups, the monopoly chooses output for each group to solve

$$\underset{q}{\text{Max}} \left\{ \sum_{i=1}^{n} p_i(q_i)q_i - C(q_1 + \cdots + q_n) \right\}.$$

From the first-order conditions,

$$MR_i = p_i \left(1 + \frac{1}{\mathcal{E}_i(q_i, p_i)}\right) = MC.$$

Hence

$$p_i = \frac{MC}{1 + \dfrac{1}{\mathcal{E}_i(q_i, p_i)}}.$$

Thus, the more negative is the price elasticity of group i, the lower is the price set by the monopoly.

[10] Obviously this condition fails if the elasticity of demand lies between 0 and -1. However, in this case higher prices increase revenue so a monopolist always increases price until the elasticity of demand, $\mathcal{E}(q, p)$, is less than -1.

[11] Seniors and students are offered price "discounts" on all sorts of commodities. Certainly seniors can buy any number of cheap movie tickets, but they and their companions must present plausible IDs at the entrance to prove that they all qualify.

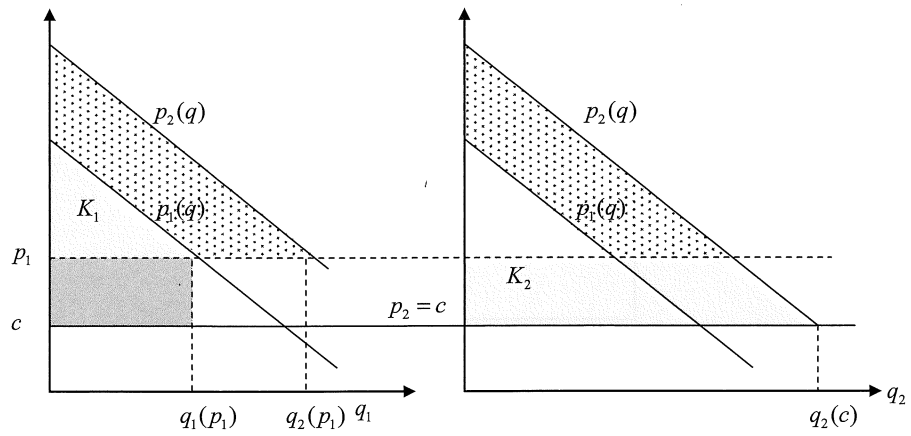

Figure 4.5-3. Price discrimination by type.

advance. Despite these barriers, a monopoly still has a powerful incentive to price discriminate. Instead of offering a single plan to each customer type, it is profitable to offer multiple plans and let each consumer choose between them. Intuitively, by carefully designing the different plans, the monopoly can give each consumer type an incentive to choose the plan designed for it.

Consider the simplest case of two consumer types: low demanders (type 1) and high demanders (type 2). Suppose that the monopolist offers a plan with a use fee p_1 intended for type 1 buyers. Consider the left-hand diagram in Figure 4.5-3. If both types of buyer choose plan 1 their demands are $q_1(p_1)$ and $q_2(p_1)$. The consumer surplus for the low demanders is the lightly shaded area. For the high demanders it is the sum of the lightly shaded and dotted areas. The entire surplus for the low demanders, $S_1(p_1)$ can be extracted by charging an access fee K_1 equal to the lightly shaded area.

The total profit on each type 1 consumer is the sum of the access fee and the profits made from the use fee $(p_1 - c)q_1(p_1)$, that is,

$$\Pi_1(p_1) = S_1(p_1) = CS_1(p_1) + (p_1 - c)q_1(p_1).$$

This is the sum of the heavily and lightly shaded regions.

Suppose then that the monopoly offers this access and use fee so that the total payment for q units is $K_1 + p_1 q$. A type 2 (high) demander who was to select this plan would purchase $q_2(p_1)$ units. Type 2's consumer surplus is then the sum of the lightly shaded and dotted regions in the left-hand diagram. Because all customers pay an access fee K_1 on this plan, each high-demand consumer has a net payoff equal to the dotted area.

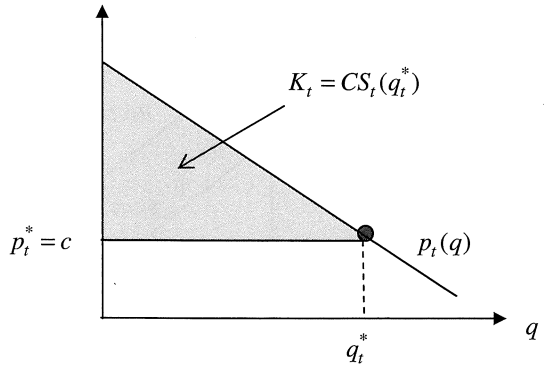

Figure 4.5-2. Extracting all the surplus.

Price Discrimination by Buyer Type

In the previous discussion, price discrimination is possible because of sufficiently high resale costs between groups. The implicit assumption is that resale is easy within groups. A buyer who is offered the low price can go on the Internet and cheaply resell to those who are precluded from direct purchase at the low price.

However, for many commodities, the cost of resale is high between all pairs of consumers. Commodities that are flows of services (such as electricity, phone services, or satellite TV) are particularly difficult to resell. In such cases a monopoly has a much more powerful incentive to price discriminate, because it can personalize both the services offered and the fees that it charges to each consumer. Let $p_t(q)$ be the demand price function for the t th type of consumer. For simplicity, suppose that the marginal cost of production is a constant, c. Then social surplus is maximized at output q_t^* where price $p_t(q_t^*) = c$ as depicted in Figure 4.5-2.

The very best that the monopoly can do is to extract the entire buyer surplus. This is the shaded region in the figure. Consider the following two-part pricing plan. Each type t buyer is charged both a use fee equal to p_t and an access fee K_t. The total cost of purchasing q units is then $K_t + p_t q$. If the monopoly charges a use fee equal to marginal cost, it induces the efficient output q_t^*. Then, by charging an access fee equal to the total buyer surplus, the monopoly profit is equal to the maximized total surplus.

Indirect Price Discrimination (Two-Part Pricing)

The previous argument is rather extreme for two reasons. First, governments often legislate against pricing policies that discriminate by personal characteristics. Second, even if the monopoly knows that there are "big buyers" and "little buyers" in its market, it may not be easy to identify them in

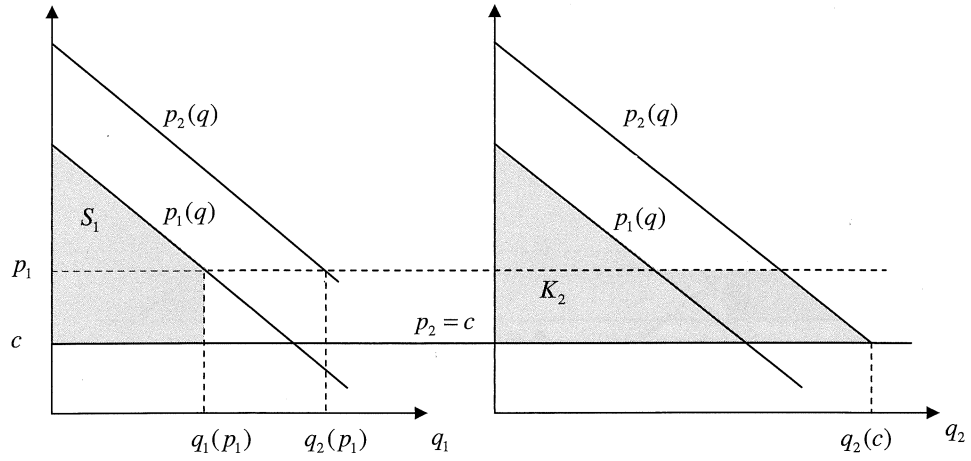

Figure 4.5-4. Profit on each buyer type.

These observations are crucial. Whereas the monopolist is able to extract all of the social surplus from low demanders, it cannot do so for high demanders because they have the option of switching to the low plan. The dotted area in the left-hand diagram is therefore an upper bound on the consumer surplus that the monopoly can extract from high demanders.

In the right-hand diagram of Figure 4.5-3, the maximum possible gain to buyer and seller is the sum of the shaded and dotted regions (maximized social surplus). The monopoly induces a purchase by each high demander equal to $q_2(c)$ by setting the use fee in plan 2 equal to marginal cost. Each high demander is better off switching to plan 1 unless offered a net gain at least equal to the dotted area. Thus, profit is maximized by choosing an access fee K_2 that extracts all the rest of the consumer surplus (the shaded region in the right-hand diagram). Because the use fee is equal to marginal cost, the shaded region is also the profit made by the firm from each high demander. The total profit made on each buyer type is therefore the sum of the shaded regions depicted in Figure 4.5-4.

This leaves open the question of what use fee, p_1, to set in the first plan. Suppose that p_1 is raised to $\hat{p}_1$ as shown in Figure 4.5-5.

At the new higher price, the reduction in low demander total surplus and hence the reduction in profit from low demanders, ΔS_1, is the heavily shaded region in the left diagram. The unshaded region in the right diagram is the surplus to high demanders if they switch to plan 1. Comparing this with the dotted region of the right diagram in Figure 4.5-4, the surplus associated with switching has been reduced by the heavily shaded region. Thus, the monopoly can increase its access fee in plan 2 by this heavily shaded region

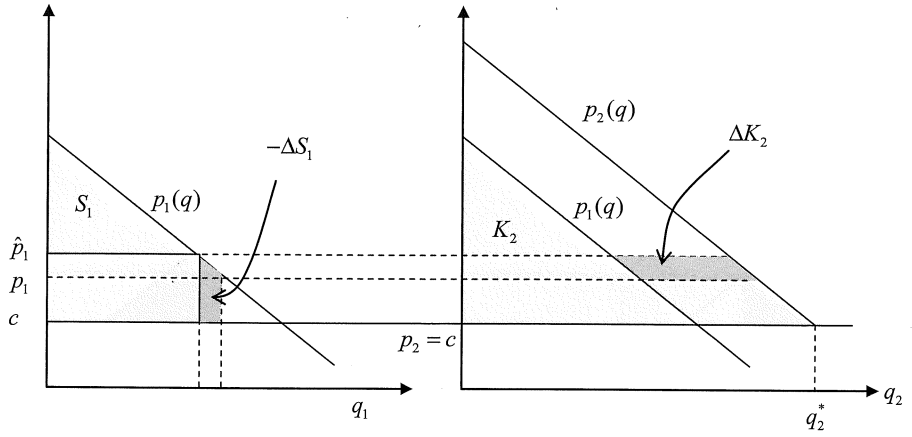

Figure 4.5-5. Choosing the use fee for plan 1.

without giving the high demanders an incentive to switch to plan 1. Thus the increase in profit from high demanders is ΔK_2.

Note that when p_1 is close to c, the heavily shaded region in the left diagram is small relative to the heavily shaded region in the right diagram. Thus when p_1 is raised, the loss in profit ΔS_1, from each low demander is small relative to the gain in profit ΔK_2, from each high demander. Therefore, regardless of the number of buyers of each type, the monopoly is better off setting the plan 1 price, p_1, higher than c. Just as with simple monopoly pricing, indirect price discrimination is inefficient in that it undersupplies the low demanders.

As we see in the discussion of "mechanism design" in Chapter 11, this is a general principle. The basic idea is that to extract more consumer surplus from the high demanders, the monopoly makes the alternative option (plan 1) less attractive. It does so by raising the price offered on plan 1. The larger the proportion of high demanders, the greater the incentive to extract surplus from them and thus the higher is the price offered on plan 1. Moreover, if the ratio of high to low demanders is sufficiently high, the low demanders are squeezed from the market completely.

Example: Linear Demands

Suppose that $p_i(q) = a_i - q$, $i = 1, 2$, and that the fraction of type i buyers is f_i. Because the slope of the demand price functions is -1, $\Delta q_1 = -\Delta p_1$, $\Delta S_1 = (p_1 - c)\Delta q_1 = -(p_1 - c)\Delta p_1$. Moreover, for any p_1, the difference in demand by the two types of buyer if they both choose plan 1 is $a_2 - a_1$. Thus, $\Delta K_2 = (a_2 - a_1)\Delta p_1$. The total change in profit is therefore

$$\Delta\Pi = f_1(\Delta S_1) + f_2\Delta K_2 = [-f_1(p_1 - c) + f_2(a_2 - a_1)]\Delta p_1.$$

Dividing both sides by Δp_1 and taking the limit,

$$\frac{d\Pi}{dp_1} = -f_1(p_1 - c) + f_2(a_2 - a_1) = f_1\left(c + \frac{f_2}{f_1}(a_2 - a_1) - p_1\right).$$

Case (i) $c + \frac{f_2}{f_1}(a_2 - a_1) \leq a_1$. Setting $\frac{d\Pi}{dp_1} = 0$, yields the optimal price

$$p_1^* = c + \frac{f_2}{f_1}(a_2 - a_1).$$

Case (ii) $c + \frac{f_2}{f_1}(a_2 - a_1) > a_1$. Then $\frac{d\Pi}{dp_1} > 0$, $p_1 \leq a_1$ and thus the profit-maximizing price drives low demanders out of the market. The monopolist then extracts the entire surplus from type 2 buyers by setting a price $p_1 \geq a_2$. This is equivalent to dropping plan 1 and only offering plan 2.

Exercises

Exercise 4.5-1: Price Discrimination by Region In region 1 the market demand price function is $p_1 = a_1 - q$ whereas in region 2 it is $p_2 = a_2 - q$, where $a_2 > a_1$. There are equal numbers in each region. The unit cost of production is c.

(a) Solve for the profit-maximizing prices if firms can price discriminate across regions.
(b) Suppose such price discrimination is impossible. Compare the profit-maximizing price with the prices in part (a) if a_1 and a_2 are sufficiently close for it to be optimal to sell to both regions. Show also that if a_2 is sufficiently large relative to a_1, region 1 buyers are priced out of the market completely. Is there a discontinuity in the profit-maximizing price function $p^*(a_2)$?

Exercise 4.5-2: Indirect Price Discrimination There are two types of buyers. Each type 1 buyer has a low demand price function $p_1 = a_1 - q$, whereas each type 2 buyer has a high demand price function $p_2 = a_2 - q$. The number of type t buyers is n_t. The unit cost of production is c.

(a) Explain why, if it is most profitable to sell only to the high demanders, the profit-maximizing two-part pricing plan is $(p_2, F_2) = (c, \frac{1}{2}(a_2 - c)^2)$.
(b) Alternatively, the monopolist offers two plans and serves both types (so $p_1 < a_1$). Show that the gain to a type 1 buyer who purchases plan 1 is $U_1 = \frac{1}{2}(a_1 - p_1)^2 - F_1$.
(c) Hence show that if the monopoly extracts the entire surplus from type 1 buyers, the access fee is $\frac{1}{2}(a_1 - p_1)^2$ and so the profit from a type 1 buyer is $(p_1 - c)(a_1 - p_1) + \frac{1}{2}(a_1 - p_1)^2$. Show also that with this access fee, type 2 buyers have a gain of $U_{21} = (a_2 - a_1)(a_1 + a_2 - 2p_1)$ if they choose plan 1.
(d) Explain why the maximum profit that the monopoly can extract from type 2 is $\frac{1}{2}(a_2 - c)^2 - \frac{1}{2}(a_2 - a_1)(a_1 + a_2 - 2p_1)$.

(e) Appeal to your answers to (c) and (d) to obtain an expression for the total profit. Then solve for the profit-maximizing price p_1.

(f) Confirm that as the number of type 2 buyers increases, the plan 1 use fee rises.

References and Selected Historical Reading

Debreu, G. (1959). *Theory of Value*. Cowles Foundation Monograph. New York: John Wiley & Sons.

Marshall, A. (1920). *Principles of Economics*. New York: Macmillan.