The procedural reformation that I have suggested is worth making for its own sake, and regardless of any changes which might be made in the substantive doctrines of antitrust. Needless to say, however, it would also facilitate the introduction into antitrust trial procedure of the kind of flexible inquiry into the economic consequences of challenged practices advocated in earlier chapters of this book.

In closing, I would like to recapitulate very briefly the main theme of the book. As a result of neglect of economic principles, the judges, lawyers, and enforcement personnel who are responsible for giving meaning to the vague language of the antitrust statutes have fashioned a body of substantive doctrine and a system of sanctions and procedures that are poorly suited to carrying out the fundamental objectives of antitrust policy—the promotion of competition and efficiency. The per se rule against price fixing, the merger rules, the rules governing competition in the distribution of goods, the tie-in rule, the use of structural remedies, the trial of antitrust cases according to methods of proof developed hundreds of years ago—these and the other features of the antitrust system examined in this book reflect above all an endeavor, sometimes ingenious and sometimes pathetic, to set antitrust free from any dependence on economic principles. The endeavor has failed; the system is in disarray. The time has come to rethink antitrust with the aid of economics. This book is offered as a contribution to the process of rethinking.

Appendix: An Introduction to the Formal Analysis of Monopoly

The main purpose of this appendix is to introduce the reader to some of the simpler geometrical and mathematical methods by which the monopoly problem can be analyzed, in the hope that lawyers, law students, and other noneconomists can be helped to read the economic literature relating to antitrust. A subsidiary purpose of the appendix is to establish a little (but only a little) more rigorously some of the propositions asserted in previous chapters.

The first question we consider is how the monopolist decides at what price to sell his product. His decision process is illustrated in figure 2. Three curves are drawn in Figure 2. The first is the demand curve, *dd*. A demand curve shows the different prices at which the monopolist's product will sell, depending on how much he produces. The negative (downward) slope of the demand curve

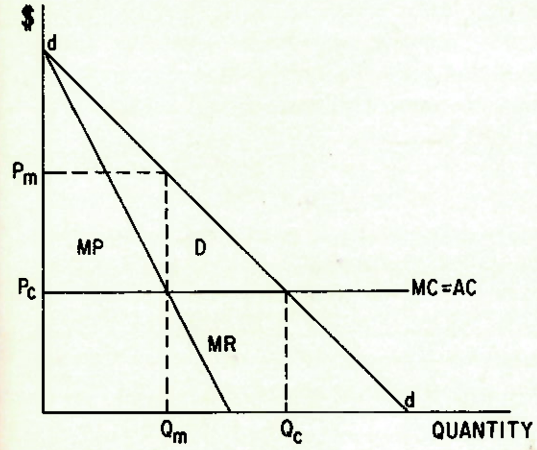

Fig. 2

as one moves from left to right reflects the fact that lower prices are associated with larger quantities demanded and higher prices with smaller quantities demanded. The higher the price of the product relative to other products, the greater is the incentive of the consumer to substitute other, and now cheaper, products. Conversely, the lower the price of the product relative to substitutes, the greater is the incentive of the consumer to switch to the product from the now dearer substitutes. The relationship between price and quantity operates in both directions. A reduction in quantity supplied will lead to an increase in price in order to ration the smaller quantity among consumers, while an increase in the quantity supplied will lead to a reduction in the price in order to attract the marginal consumer.

Several further points should be noted about the demand curve. First, it is a schedule of relative, not absolute, prices. A rise in the average price level in the economy (i.e., inflation) will not cause a movement up the demand curve for particular products; the relative prices of products are not affected by inflation. Second, the negative relationship between price and quantity assumes that other things which might affect the demand for the good are being held constant. For example, if the demand for a particular product rises as people's incomes rise, one might well observe a simultaneous increase in both the price and the quantity sold. Third, while the demand curve is to the consumer a schedule of prices, to the producers it is a schedule of average revenue. The total revenues of the industry are simply prices times output. Total revenue divided by output—i.e., average revenue—is thus equal to price. Fourth, the demand curve need not be linear as in figure 2. The mathematical (and economic) properties of nonlinear demand curves are, as we shall see, somewhat different.

The second curve in figure 2 is the marginal-revenue curve (*MR*). Marginal revenue is the contribution to the industry's total revenues made by selling another unit of output. Since the negative slope of the demand curve means that an increase in output is associated with a decline in price, marginal revenue will be positive (above the horizontal axis in figure 2) or negative (below the axis) depending on whether the change in output is proportionately smaller or greater than the change in price. Marginal revenue is everywhere below price in figure 2 because as the industry increases its output it sells not only the marginal output at a lower price but *all* of its output at that price. This assumes,

of course, that the industry cannot sell the marginal output at the price that the marginal purchaser will pay and the rest of its output at the former, and higher, price. We discuss a bit later the consequences of relaxing this assumption; for now, our analysis focuses on the monopolist who is constrained either by law (or more precisely by the costs of violating the law) or by the (other) costs of preventing arbitrage from selling at more than one price.

The third curve in figure 2 is the industry's marginal cost (*MC*) at various levels of output. Marginal cost is the increase in total cost if output is expanded by one unit. There are also fixed costs, that is, costs that are independent of output. Even if the industry stopped producing, the fixed costs (e.g., the cost of servicing the industry's long-term debt) would remain. We are not especially interested in the fixed costs, however. By definition they do not enter into the producer's decision as to what quantity to sell, and therefore what price to charge, for he cannot alter these costs by his decision on quantity or price; they are independent of the scale of his activity. So we shall assume for convenience that all of the costs of our industry are variable costs reflected in the marginal-cost curve.

Another way of justifying the exclusion of fixed costs is to observe that, in the long run, *all* costs are marginal; all, that is, depend ultimately on what decisions are made with regard to the output of the product. When a plant wears out, it will be replaced only if the expected revenues from the plant are greater than its total costs. From this perspective, the cost of building the plant is a variable rather than a fixed cost (and will enter into marginal cost) since it will determine how much (if anything) will be produced. We can therefore think of *MC* as the industry long-run marginal-cost curve.[1]

In figure 2, *MC* is a horizontal line. This implies that the cost of producing a unit of output is the same regardless of the number of units produced. Although the principal reason for assuming a constant-costs supply curve is expository convenience—and nothing vital to the analysis would be altered by dropping the assumption—it probably approximates the cost conditions facing many, or even most, industries within a broad range of possible

1. On the difference between short-run and long-run marginal costs see also pp. 191–92 supra.

outputs. But this is *not* because individual firms typically have constant costs. That would imply that firm size was indeterminate —a firm that produced one automobile a year would have the same long-run marginal costs as a firm that produced 250,000 automobiles, or for that matter 10,000,000.

To reconcile a constant-costs industry curve with a U-shaped firm-cost curve, it is necessary to introduce the concept of average cost, i.e., total cost divided by output. With a horizontal marginal-cost curve and no fixed costs, average cost is identical to marginal cost at all outputs. But if marginal cost (say) rises with output, average cost will rise too (though more slowly, just as price, i.e., average revenue, falls more slowly than marginal revenue).

In an industry that has many firms, all facing identical cost conditions and therefore having identical average-cost curves, each firm will produce the output at which its average costs are minimized, for at any other level of output its average costs would be higher than those of other firms and it would be in danger of being underpriced. Thus the industry marginal-cost curve will be a horizontal line lying along the locus of the minimum points of the firms' (identical) average-cost curves. This is the industry's marginal-cost curve because, were it necessary to expand the output of the industry, this would be accomplished by the entry of one or more firms which would produce at the point of their lowest average costs, the same point at which the existing firms are producing. Thus the cost of the additional units would be the same as those of the existing units, and the industry's marginal costs are constant.[2]

2. If there are only a few firms in the industry, the analysis becomes a bit more complicated. Unless the total demand for the industry's output can be divided by the cost-minimizing output of each firm to yield a whole number, it will be impossible to satisfy that demand at a cost equal to the minimum average cost of each firm. If, for example, the total demand for the industry's output is 100,000 units but the cost-minimizing output of each firm in the industry is 15,000 units, there is no way in which the total demand can be satisfied by production at the lowest average cost of each firm. One firm will have to produce at a level of output at which its average costs are not minimized. Moreover, even if the current output of the industry is being produced at an average cost equal to the lowest average cost of each of the (identical) firms comprising the industry, should it become necessary to expand the output of the industry to satisfy a growth in demand it may be impossible to do so without incurring higher costs. The increase in demand may not be great enough to justify the entry of a new firm that would produce the level of output at which its average costs would be minimized. Rather, it may be necessary to satisfy the increase in demand by an expansion in the output of one or more of the existing firms in the market, resulting in an increase in the average and marginal costs of the industry.

The optimal monopoly price in figure 2 is given by the intersection of *MR* and *MC*. Profit is simply the difference between total revenues and total costs and is maximized by carrying production to the point where the last unit produced contributes just as much to total revenues as to total costs—that is, where marginal revenue and marginal cost are equal. At any output to the left of Q_m (the output corresponding to the profit-maximizing price, P_m) the monopolist could increase his profits by expanding output, for at such points marginal revenue lies above marginal cost, meaning that additional output adds more to total revenue than to total cost and thus generates additional profit. To the right of Q_m the relationship is reversed. Producing quantity Q_m thus maximizes the difference between the area under the marginal-revenue curve and the area under the marginal-cost curve—i.e., maximizes profits.

If the industry were organized competitively and its marginal-cost curve were identical to that of the monopolist, output would be carried to point Q_c in figure 2, the point at which price (P_c) is just equal to marginal cost. Assume that the market contains many firms of roughly equal size so that the output of any single firm is small relative to the total output of the market, and that the firms do not coordinate their price and output decisions. To a firm whose decisions to increase or reduce output do not (measurably) affect the market price because the firm's output is very small relative to that of the market as a whole, the demand curve appears as a horizontal line at the market price and the firm's marginal-revenue curve is identical to its demand curve. That is, the firm assumes that every additional sale is made at the same price as the existing sale (which is why the demand curve appears to the firm as a horizontal rather than as a downward-sloping line) and increases its total revenues by the full amount of the sales price (which is why its marginal-revenue curve is a horizontal line identical to the demand curve). Each seller in a market of many sellers that are not colluding will continue expanding his output along what he perceives to be his horizontal demand curve until the combined effect of all of the sellers' actions in increasing the supply of the industry's product depresses price to the point at which it is equal to the industry's marginal-cost curve. At this point, the sellers will cease to expand their output, since if they produced more units they would have to sell below their costs.

It is plain from figure 2 that monopoly results in a lower output than competition, as well as generating monopoly profits which are equal to the rectangle denoted *MP*. The reduction in output generates a subtle form of social loss measured by the triangle labeled *D*. This represents the loss in value to those consumers who at the competitive price would buy the product, but at the monopoly price are deflected to substitutes. The fact that the demand curve lies above the marginal-cost curve in this region indicates that the value of the product to consumers who no longer purchase it exceeds the opportunity costs of producing it. This extra value is lost when the product is monopolized, and it is not recouped by the monopolist (or anyone else), for the monopolist obtains no revenues from output that he does not produce.

Thus far we have assumed that the monopolist is constrained to sell at a single price. He would, of course, prefer to vary his price with the intensity of the consumer desire for his product, charging more to those who have poor substitutes for it and less to those who have good substitutes. Suppose that the monopolist can discriminate in price perfectly and costlessly: every sale is made at a price equal to the value that the consumer places on the purchase and no costs are incurred in ascertaining those values or in dealing separately with each consumer over each unit of output purchased. Then the monopolist will proceed down the demand curve from its intersection with the vertical axis to its intersection with the marginal-cost curve, and the demand curve in that interval will become a schedule of different prices for each unit of output. The results of perfect price discrimination are compared with those of single-price monopoly in figure 3.

Observe that the output of the perfectly discriminating monopolist is identical to that of the competitive industry. Thus, perfect price discrimination eliminates the social loss of monopoly that we earlier denoted by *D*. However, *MP*, monopoly profits, are greater under perfect price discrimination than under single-price monopoly. Indeed *MP* in figure 3(b) is larger than $MP + D$ in figure 3(a), implying, according to our assumption that expected monopoly profits are transformed into social costs,[3] that the social costs of perfect price discrimination are actually greater than those of single-price monopoly even though the output of the perfectly price-discriminating monopolist is identical to that under competition.

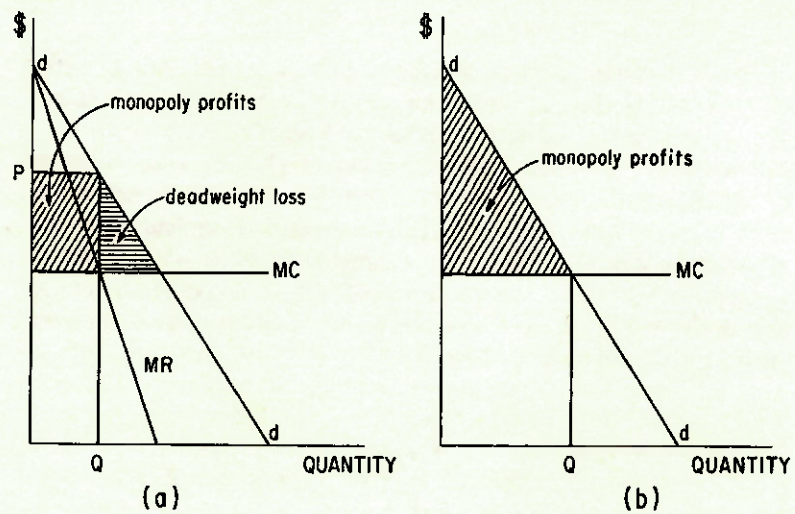

Fig. 3

Let us now derive the major results of the analysis of monopoly pricing algebraically, beginning with the simple case in which the demand curve can, as in figures 2 and 3, be approximated by a straight rather than a curved line. The algebraic form of a negatively sloped straight line such as *dd* in figure 2 is $a - bQ$, where a is the intercept on the vertical axis, $-b$ is the slope of the curve (and is negative because the slope of a downward-curving line is negative), and Q is the independent variable. Since the demand curve is a price schedule we can write $P = a - bQ$. The marginal-cost curve is simply C, by our assumption of constant costs. Since total revenue equals price times quantity, since total cost (under our assumption that there are no fixed costs) equals marginal cost times quantity, and since profit is simply the difference between total revenue and total cost, we can write

3. See pp. 11–13 supra.

$$\pi = (a - bQ)Q - CQ, \tag{1}$$

where π is profit. The monopolist's goal is to sell that quantity (Q) of the product at which his profits (π) will be maximized, and we can use elementary calculus to determine what value of Q will maximize π. Since π is a function of Q (i.e., our choice of Q will determine the value of π), we set the first derivative of π with respect to Q equal to zero. The geometric equivalent of this is shown in figure 4. The point where the function π of Q attains its highest point (i.e., maximum profits) is the point where the first derivative (i.e., slope) of the profit function is equal to zero (a horizontal line has no slope).[4]

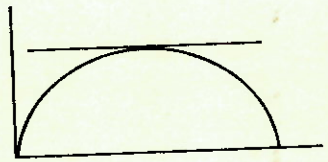

Fig. 4

Using a standard formula which can be found in any elementary calculus text,[5] we differentiate π with respect to Q to obtain the first derivative, yielding

$$\frac{d\pi}{dQ} = a - 2bQ - C. \tag{2}$$

Setting this equal to zero and adding C to each side of the resulting equation we obtain

$$a - 2bQ = C. \tag{3}$$

4. As a detail, but an important one, note that merely knowing that the slope of a function is equal to zero doesn't tell you whether the function is at a maximum or minimum point. For a maximum, the second derivative of the function must be negative (i.e., changing from positive to negative at its extreme point). That is, the slope of the slope must be negative. (One's upward movement slows as one approaches the top of a hill and turns negative as one begins to descend the hill on the other side.) This means that, in the area of the putative maximum, profits as a function of quantity produced must be increasing at a decreasing rate. Since the second derivative of our profit function (equation [1]) is negative (–2b), the "second order" condition for a maximum is indeed satisfied.

5. Particularly good for people interested in economics is Alpha C. Chiang, *Fundamental Methods of Mathematical Economics* (2d ed. 1974).

The left-hand side of equation (3) is the first derivative of total revenue, or marginal revenue; the right-hand side is the first derivative of total cost, or marginal cost (this can be verified by applying the formula for obtaining a first derivative directly to our expressions for total revenue and total cost, $(a - bQ)Q$ and CQ). In short, profits are maximized when $MR = MC$. Observe that the slope of the marginal-revenue curve, $-2b$, is exactly twice the slope of the demand curve ($-b$). At a price of zero, Q is equal to a/b, while at a marginal revenue of zero Q is equal to $a/2b$. That is, when the demand curve is linear, the marginal-revenue curve intersects the horizontal axis exactly midway between the origin and the intersection of the demand curve with that axis—as shown in figure 2.

It will be useful to have a more general expression for the optimal price and quantity under monopoly, one that does not depend on the assumption that the demand curve has a particular shape (other than that it be negatively sloped and differentiable at all relevant points). The more general expression for the monopolist's profit function is

$$\pi = P(Q)Q - CQ, \tag{4}$$

where $P(Q)$ (or P for short) is the demand curve (i.e., price is written as a function of the quantity demanded). Taking the first derivative of π with respect to Q and setting the result equal to zero, we have

$$\frac{d\pi}{dQ} = \frac{dP}{dQ}Q + P - C = 0, \tag{5}$$

or

$$\frac{dp}{dQ}Q + P = C. \tag{6}$$

The left-hand side of the equation again denotes marginal revenue and the right-hand side marginal cost.[6]

6. Now let us check the second-order condition. The second derivative of (4) with respect to Q is

$$\frac{d^2\pi}{dQ^2} = \frac{d^2P}{dQ^2}Q + 2\frac{dP}{dQ}. \tag{i}$$

How are we to know whether this expression is negative? To begin with, we know that the second term is negative since the slope of the demand curve (dP/dQ) is negative. To be sure that the entire expression is negative,