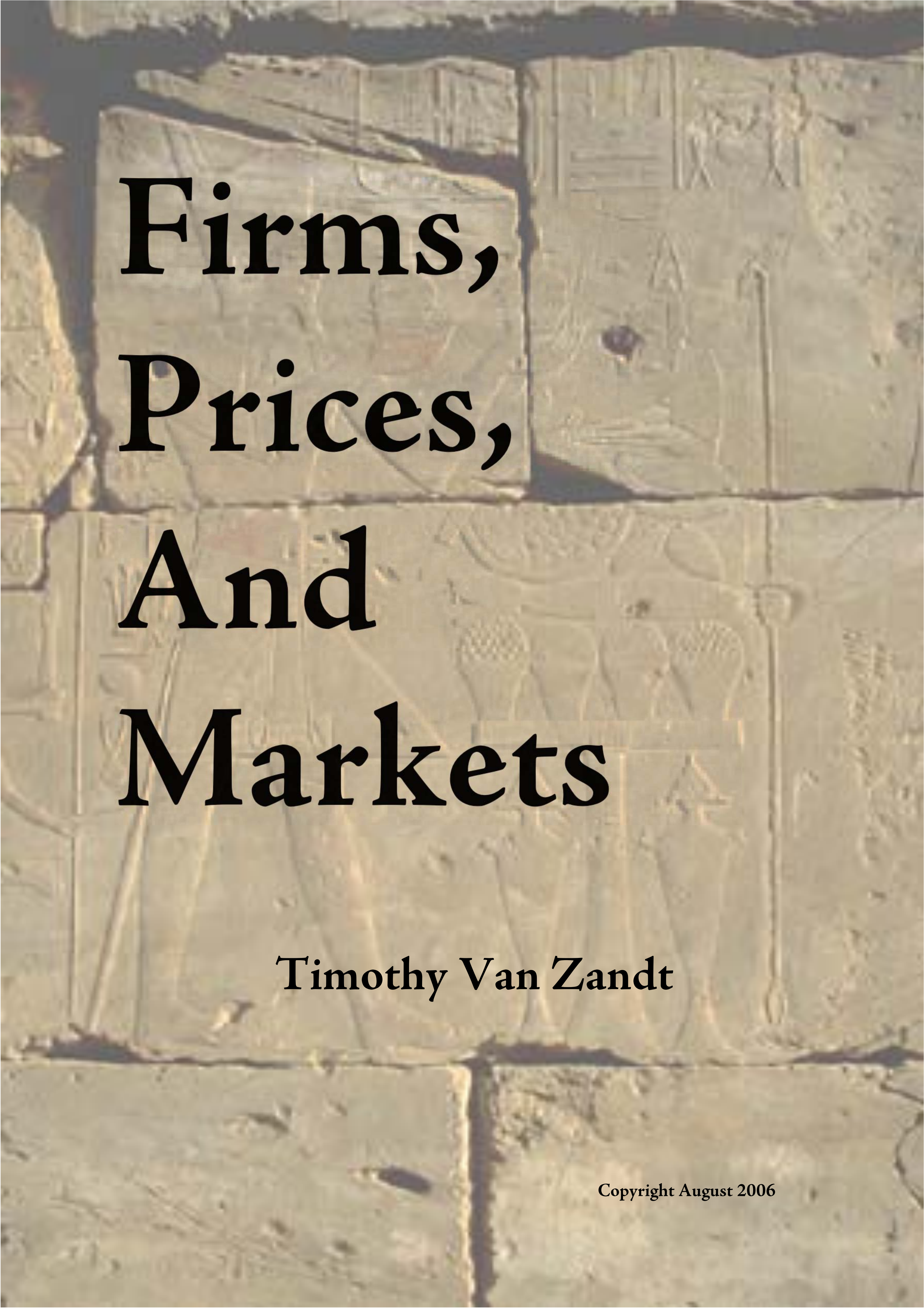


Firms, Prices, And Markets

Timothy Van Zandt

Table of Contents

Preface

This target audience of this book is MBA students taking a first course in microeconomics or managerial economics. It may also interest anyone looking for an intermediate-level manager-oriented treatment of microeconomics.

Microeconomics is an analytic discipline. This book reflects a belief that microeconomics' analytic tools, properly presented, enhance managers' ability to make decisions on their feet using soft data in complex situations. It emphasizes the use of logical analysis and simplification in order to break problems down into understandable parts. It presents numerical examples for the purpose of illustrating qualitative concepts, not because managers are likely to have the numerical data needed to replicate such examples on the job.

This book was developed while teaching MBA students at INSEAD, Northwestern University, and New York University. I am indebted to these students for putting up with early versions, providing feedback, and above all for making the teaching experience a rewarding one.

Math Review

M.1 What math?

Microeconomics can be pretty mathematical, even at the intermediate level. For example, some books use calculus extensively in order to solve numerical optimization problems. Because the focus of this book is on low-data situations, we do not use that much mathematics. (A student could otherwise get a false sense of accomplishment by calculating solutions to all kinds of problems, when in real life he or she will not have the data for such calculations.) We do work through numerical examples, but the goal is to illustrate concepts that can be applied to qualitative decision problems.

Hence, the math in this book is basic—no more than what any student would have been exposed to in high school or in a first-year college course at the latest. That said, if you have not used such tools for several years (and perhaps never really liked them back in high school), then a review is important.

M.2 Functions

We are often interested in the relationship between two variables, such as between price P and demand Q for a good or between output Q and cost C of a firm. We think of one of the variables (the dependent variable) as depending on the other (the independent variable). (If no particular meaning is ascribed to the variables, it is common to denote the independent variable by X and the dependent variable by Y.) Here are two examples.

1. Perhaps a firm's output level Q depends on the price P it observes in the market; then P in the independent variable and Q is the dependent variable.
2. Perhaps instead we think of the price P that a firm must charge for its output as depending on the amount Q that it tries to sell; then Q is the independent variable and P is the dependent variable.

The relationship between the two variables is called a *function* or *curve*. ("Function" is the standard terminology in math, but in this text we use "function" only when there are several independent variables; otherwise we use the term "curve".)

If the independent variable can take on only a few values, then we can specify a function with a table that gives the value of the dependent variable for each value of the independent variable. For example, output and revenue may be related as in Table M.1 (taking into account that you can sell more output only by charging a lower price).

Table M.1

Output	Revenue
0	0
1	57
2	108
3	153
4	192
5	225
6	252
7	273
8	288
9	297
10	300
11	297
12	288
13	273

We can also specify a function by a formula. For example, perhaps the demand Q as a function of price P is

$$Q = 2800 - 7P.$$

According to this formula, when the price is 100, the demand is $2800 - (7 \times 100) = 2100$. Perhaps revenue R as function of output Q is

$$R = 60Q - 3Q^2.$$

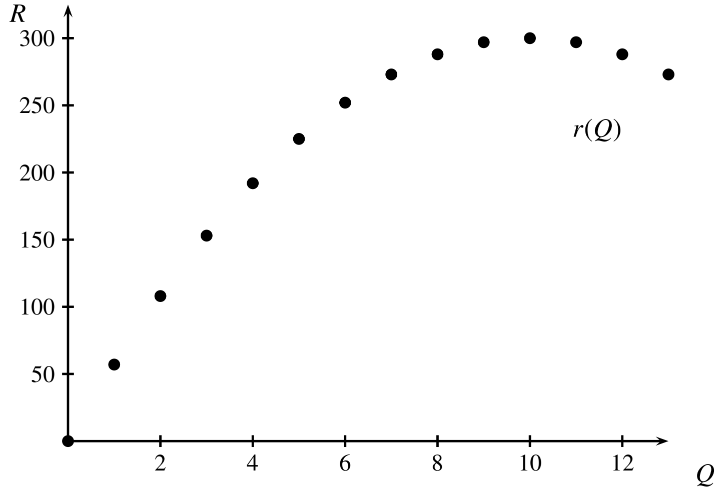
According to this formula, if output is 5, then revenue is $(60 \times 5) - (3 \times 5^2) = 300 - 75 = 225$. (This is the formula for the data in Table M.1.)

Often we want to give a function a name. For example, we might denote a demand function by d and a supply function by s. Then $d(P)$ denotes the demand when the price is P and $s(P)$ denotes the supply when the price is P. This allows us to write expressions such as the following: "If the demand curve is d and the supply curve is s, then the equilibrium price P is such that $d(P) = s(P)$." In this book, we use uppercase letters for variables and lowercase letters for functions.

M.3 Graphs

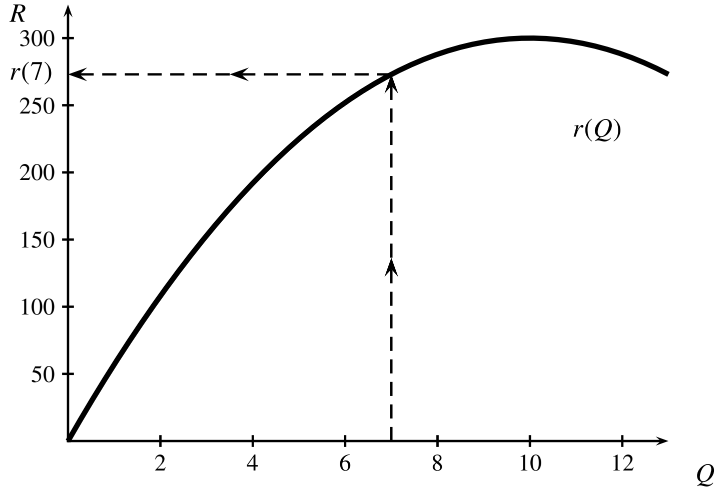
To graph a function, we let the horizontal axis measure values of the independent variable (the "X-axis") and let the vertical axis measure values of the dependent variable (the "Y-axis"). If the data is in table form, then each pair (X, Y) from the table (assuming that X is the independent variable and Y is the dependent variable) is one point on the graph. Figure M.1 shows the graph of the function in Table M.1.

Figure M.1



If the variables are continuous and we have a functional form, then the graph of the function is a smooth curve. Figure M.2 shows the graph of $r(Q) = 60Q - 3Q^2$.

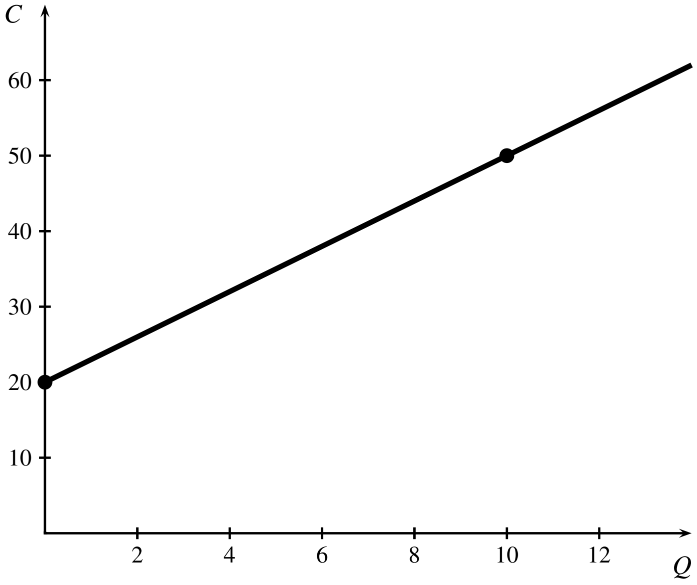
Figure M.2



From a graph, you can see the approximate value of the dependent variable for any value of the independent variable. For example, what is $r(7)$? You find 7 on the horizontal axis, move straight up to the graph of the function, and then look straight left to see what the value is on the vertical axis. This procedure is represented in Figure M.2 by the dashed line; the value of $r(7)$ is approximately 275.

You needn't be able to draw graphs like the one in Figure M.2, but only to interpret them. (One can draw a graph like this with Excel or some other software.) You should, however, be comfortable drawing linear functions, such as $Q = 2800 - 7P$ or $C = 20 + 3Q$. The easiest way is to find two points on the curve and then draw a straight line through the two points. Take the function $C = 20 + 3Q$. If $Q = 0$ then $C = 20$; if $Q = 10$ then $C = 50$. Figure M.3 shows these two points, $(0, 20)$ and $(10, 50)$, and the entire graph.

Figure M.3



M.4 Inverse of a function

Let $Q = d(P)$ be a demand function. You can think of a function as a little machine that provides answers. You plug in 7 and get $d(7)$, which is the answer to the question: "what is the demand when the price is 7?"

We can reverse the roles of the variables in a function. Rather than plugging in a price to find a quantity, we can start with a quantity and find the corresponding price. For example, we can ask: "For what price is demand equal to 3?" This is called the *inverse* of the function. It is a function that shows the same relationship between two variables, except that we reverse the roles of the dependent and independent variables. We might write the inverse of the demand function as $P = p(Q)$.

Let's see graphically how to read an inverse. Figure M.4 shows the graph of a demand curve $Q = d(P)$. The dashed line represents the procedure of starting with a price of 7 and finding what the demand is at this price. We see that it is 3.

Figure M.4

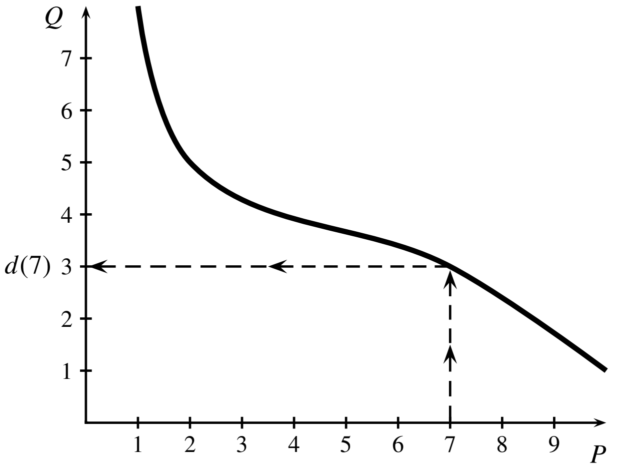
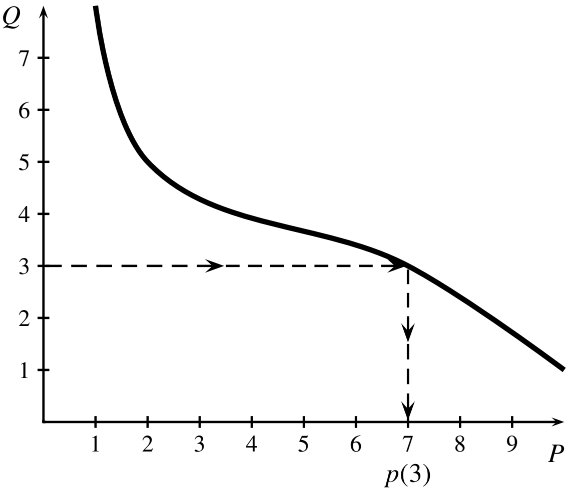


Figure M.5 shows how we can use the same graph to find an inverse. We start with a quantity of 3 on the vertical axis, We go right to the graph of the function and then look down to the value on the horizontal axis. This is the price at which demand is 3.

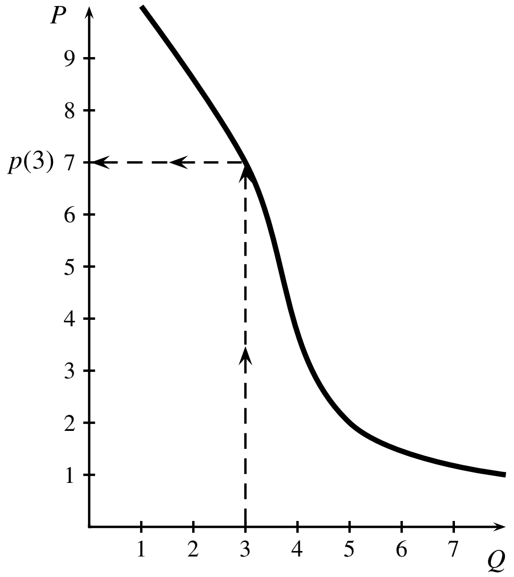
Figure M.5



Thus, since the inverse simply reverses the roles of the dependent and independent variables, we can use the same graph for a function and its inverse as long as we are willing to have the independent variable on the vertical axis for one of the cases.

However, we have the option of graphing the inverse with the axes flipped. Let $P = p(Q)$ be the inverse of the demand function shown in Figure M.4. Since Q is the independent variable of the function p, mathematical convention is to draw the graph of p with Q on the horizontal axis, as shown in Figure M.6.

Figure M.6



This is an option we prefer not to exercise if we need to refer simultaneously to a function and to its inverse: it gets very confusing to flip the graph back and forth. Instead it easier to keep the graph fixed and to just read it in different directions depending on whether we are working with the function or with its inverse.

M.5 The inverse of a linear function

The only inverses we will actually calculate are for linear functions. Suppose a demand function is $Q = 60 - 5P$, which we can also write as $d(P) = 60 - 5P$. Let $p(Q)$ be its inverse. To find the functional form for $p(Q)$, we solve the equation

$$Q = 60 - 5P$$

for P (that is, we rearrange the equation so that P is alone on the left). Here are the calculations step-by-step:

$$\begin{aligned} Q &= 60 - 5P, \\ 5P &= 60 - Q, \\ P &= 12 - Q/5. \end{aligned}$$

(We added $5P$ to both sides, subtracted Q from both sides, and then divided both sides by 5.) Thus, the inverse can be written as $P = 12 - Q/5$ or as $p(Q) = 12 - Q/5$.

From now on in this text, we graph demand and supply curves as inverses, with P on the vertical axis and Q on the horizontal axis. Economists started doing this over 120 years ago—not because they were poor mathematicians but rather because the inverses of demand and supply curves are used even more often than the demand and supply curves themselves.

Exercise M.1. We first plot a demand curve following *mathematical* convention (with price on the horizontal axis). Then we calculate and plot its inverse.

a. Draw a graph of the linear demand curve $d(P) = 20 - \frac{1}{3}P$, with price P on the *horizontal* axis and demand on the *vertical* axis. Be sure to label the units on the axes or at least the values of the intercepts.

b. Now calculate the inverse $P = p(Q)$ of this demand curve. You are solving the equation $Q = 20 - \frac{1}{3}P$ for P. What is $p(9)$?

c. Graph the inverse demand curve, with quantity on the horizontal axis and price on the vertical axis. Again, label the values of the intercepts.

d. The general form of a linear demand curve is $d(P) = A - BP$, where A and B are positive numbers. The price at which demand is 0 is $\bar{P} = A/B$, which we call the *choke price*. Write the general form for the inverse demand curve, using the coefficients $\bar{P}$ and B.

Hint: You solve $Q = A - BP$ for P. If you arrange the formula the right way, the term A/B appears; replace it by $\bar{P}$.

M.6 Some nonlinear functions

We work (infrequently) with two kinds of nonlinear functions: exponents and logarithms.

Examples of exponential functions are $Q = K^{1/2}L^{1/4}$ and $Q = 3P^{-2}$. Note that $P^{-2} = 1/P^2$ and so the second function could also be written as $Q = 3/P^2$.

Logarithms are used in the following way. When we take the log of a mathematical expression, multiplication becomes addition and exponents become multiplication. Here are some examples:

$$\begin{aligned} \log(8X) &= \log 8 + \log X\,; \\ \log(X^7) &= 7\log X\,; \\ \log(4X^2Y^{-3}) &= \log 4 + 2\log X - 3\log Y\,. \end{aligned}$$

Thus, if $Q = 3P^{-2}$ then we obtain $\log Q = \log 3 - 2\log P$ by taking the log of both sides. This function is now linear if we think of the variables as $\log Q$ and $\log P$. (This is why exponential functions are also called log-linear functions.)

Exercise M.2. Expand the following as in the previous examples.

$$\log(12P) =$$

$$\log(P^{-1/2}) =$$

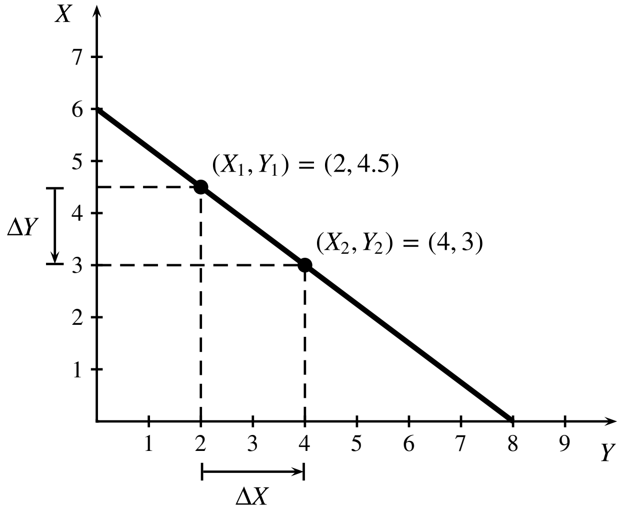
$$\log(8P^{-3}I^{0.75}) =$$

M.7 Slope of a linear function

The *slope* of a linear function measures how much the dependent variable changes per unit change in the independent variable. Slope is important to us because we study how profit, for example, changes if we adjust actions by a small amount.

Slope may be denoted $\Delta Y/\Delta X$, where ΔY means "change in Y" and ΔX means "change in X" between two points on the line. (This ratio is the same for any two points on the line.) In Figure M.7, comparing the two points $(2, 4.5)$ and $(4, 3)$ on the graph, we have $\Delta X = 2$ and $\Delta Y = -1.5$. Hence, the slope is $-1.5/2 = -3/4$.

Figure M.7



A line that slopes down has negative slope; a line that slopes up has positive slope. The steeper the line, the greater the magnitude of the slope (the opposite is true when we are working with an inverse and graph the dependent variable on the horizontal axis).

If we have the formula for a line, the slope is the coefficient of the independent variable. For example, the slope of $C = 20 + 3Q$ is 3. If Q goes up by 2, then cost goes up by 6. If Q goes down by -0.4 then cost goes down by $3 \times (-0.4) = -1.2$.

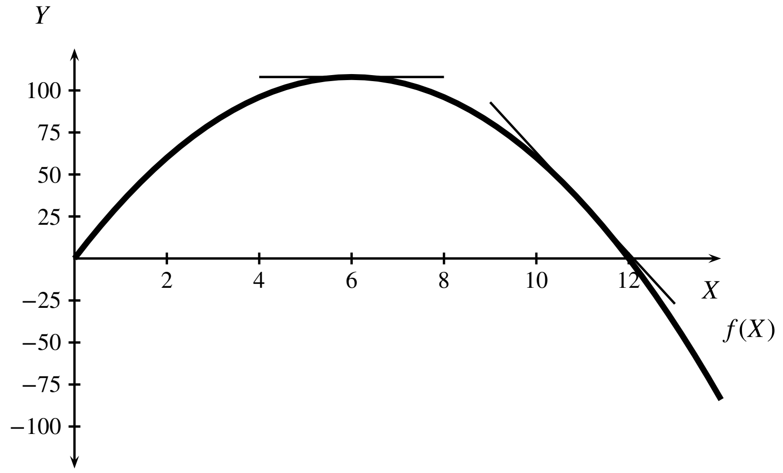
Exercise M.3. Consider the function $C = 30+4Q$. Identify the slope in two different ways: (a) from the coefficient of the independent variable Q; and (b) by calculating $\Delta C/\Delta Q$ for two points on the graph. Verify that you get the same answer.

M.8 Slope of a nonlinear function

Graphically

The slope of a nonlinear smooth function at a particular point $Y = f(X)$ is equal to the slope of the line tangent to the graph of the function. It is denoted by dY/dX or $f'(X)$ or $mf(X)$. (The notation $f'(X)$ is common in mathematics,s but we use $mf(X)$ in the main part of this book.)

Figure M.8



Consider Figure M.8, which shows the graph of a function $Y = f(X)$. The lines tangent to the curve illustrate the slope at those points. Slope is positive at first but the curve becomes less steep as X increases. Thus, slope is falling until it is zero at $X = 6$. The curve then slopes downward, meaning that slope is negative. Since it gets steeper, slope is getting more negative as X increases; that is, slope continues to fall.

For a nonlinear function like the one in Figure M.8, the slope gives an approximate measure of the rate at which the value of the function changes for small changes in X. For example, the slope is -30 at $X = 11$. This means that, if X increases by 0.1 from $X = 10$ to $X = 10.1$, then the value of the function changes by approximately $-30 \times 0.1 = -3$.

Slope as the derivative of a function

To calculate the slope of a nonlinear function, we find its derivative. This is a technique from calculus. The derivatives we use are pretty simple.

- The derivative of a constant function like $f(X) = 5$ is 0; the graph of the function is a flat line.
- The derivative of an exponential function of the form $f(X) = AX^B$ is $f'(X) = BAX^{B-1}$. That is, we multiply the function by the exponent and reduce the exponent by 1. Here are some examples:

$f(X)$	$f'(X)$
X^4	$4X^3$
$3X^5$	$15X^4$
$10X^{-2}$	$-20X^{-3}$
$4X$	4

  The last example is just the formula for the slope of a linear function; here we use the fact that a number to the 0th power, such as X^0, is equal to 1.
- The derivative of the sum of several terms is the sum of the derivatives of the terms. For example, if

$$f(X) = 4 + 10X - 3X^2$$

  then

$$f'(X) = 10 - 6X\,.$$

Exercise M.4. Calculate the derivatives of the following functions.

a. $R = 60Q - 3Q^2$.

b. $d(P) = 3P^{-2}$.

c. $f(X) = 18 - 5X + X^2$.

Preliminaries

Analytic Methods for Managerial Decision Making

This introductory chapter helps orient you in the "frame of mind" that lies behind much of this book. Some of its content will become more meaningful after you see applications. Therefore, it is recommended that you read it once before continuing (without attempting to absorb it all) and then return to it later.

P.1 Motives and objectives

Broadly

You manage a firm. The operations of the firm are very complex, with many decisions to be made about production, marketing, financing, and so on. Call a complete configuration of these decisions a *strategy*. Each strategy results in a profit for your firm. Your problem is to choose the strategy with the highest profit.

Here is how to solve this problem. Open up a spreadsheet with two columns. In one column, list the possible strategies. In the other column, list the profit for each strategy. Either by eyeballing the spreadsheet or using one of the spreadsheet program's built-in functions, pick out the highest profit level. In the same row, look at the entry in the strategy column. This is your best strategy.

If only life were so simple! In practice: (a) the problem is very complex—for example, you could not even list all the strategies; and (b) you do not have the hard data needed to determine the profit of each strategy. (Luckily for you—this is why firms hire human managers rather than computers to make decisions.)

The goal of this book is to enhance your ability to make decisions *on your feet* using *soft data* in *complex situations*. In pursuit of this goal we introduce several methods, central to which are *logical analysis* and *simplification*.

1. Logical analysis is one of the important tools a manager brings to bear on a problem (others include intuition, experience, and knowledge of similar cases).
2. Good managers are smart but do not have infinite information-processing ability. Therefore, like good physicists, good doctors, and good economists, they simplify problems in order to reason logically about them.

More specifically

We cover the following methods, which are applied to every topic covered in this book.

- *Models.* A model is a simplified or stylized description of a problem that isolates the most important features.
- *Smooth functions.* Even when the underlying data is discrete, we may use smooth approximations to simplify our analysis.
- *Decomposition of decision problems.* Decomposing a decision problem means to divide it into smaller and simpler subproblems.
- *Marginal analysis.* Marginal analysis considers the incremental effects of small changes in decisions. Under the right conditions, marginal analysis provides a simple way to find an optimal decision or to check whether a decision is optimal.

P.2 The economist's notion of models

> *Although this may seem like a paradox, all exact science is dominated by the idea of approximation.* — Bertrand Russell

Models

A model is an artificial situation (a "metaphor") that is related to, but simpler than, the real-world situations that the model is meant to help us understand. Every topic in this book is studied by constructing models.

Simplification is a goal of modeling, not an unintended negative consequence. This simplicity has two roles.

1. The human brain cannot comprehend all aspects of a real-world situation simultaneously; hence, it is useful to decompose it into various simpler situations. Only then can we apply logical analysis.
2. Our goal is not to derive conclusions from a wealth of data about a few cases. Instead, we wish to say as much as possible using as little information as possible. This will make it more likely that the conclusions apply to a broad range of future experiences in which you will often have limited information.

So you should judge a model by what is in it, not by what has been left out. The components we ignore in a model might introduce new relationships, but they will not invalidate those we have identified. Of course, all conclusions drawn from models must be taken with a grain of salt rather than applied dogmatically.

In Chapter 2, for example, we construct a model of a simple market. The model does not include any details about how traders interact and settle on trades. Ignoring such details not only makes our model simpler, it also allows us to draw *robust* conclusions that are relevant to a wide variety of trading mechanisms.

We next discuss three of the simplifying assumptions that appear in most models in this book (though they are not part of all economic models): rationality, no uncertainty, and partial equilibrium. Simplifying assumptions are not "correct"—otherwise they would not be simplifying. Therefore, to discuss such assumptions means to explain what is lost and what is gained by making them.

Rationality

Microeconomics, like the other disciplines that underlie your business education, is a social science. We are interested in understanding the interaction between people. For this, we need to specify how each individual behaves.

People cannot handle unlimited amounts of information and they make mistakes, but this does not mean that their behavior is arbitrary or irrational. In important economic interactions, people are goal oriented and work hard to pursue these goals. Not all entrepreneurs are equally good at managing a company but they seek to earn a profit, and such goal-oriented behavior is the largest determinant of their actions.

The simplest way to capture such goal-oriented behavior is to ignore the imperfections and limitations in people's abilities. Economists call this the *rationality* assumption, though "rationality" has a stronger meaning here than in everyday discourse because it implies that each person is infinitely smart and makes no mistakes. (Economists use the term "bounded rationality" for behavior that is goal-oriented and reasonable but with the usual human limitations on processing information.)

The rationality assumption is very powerful. Managers are much more likely to err by underestimating the cleverness of their competitors than by overestimating their cleverness. Furthermore, there are so many ways to make mistakes that the "best guess" of likely behavior is the behavior of a rational agent. When aggregating over many players in a market, the random effects of nonrational behavior are likely to average out at the aggregate level. Selection favors those actors who are the most rational: Firms that persistently make mistakes are likely to shut down; investors that persistently make mistakes are likely to become small players in financial markets. Finally, besides being a simple way to capture real-life goal-oriented behavior, the rationality assumption also helps us learn how we should act in economic situations. To make good decisions, we should learn how a hypothetical "rational" person would behave.

No uncertainty

In most of this book, we assume that people know the information relevant to their decision problems. Models that incorporate uncertainty are more complex and should only be studied after understanding the corresponding model without uncertainty.

Partial equilibrium

We study *partial equilibrium* models. This means that we concentrate on one market at a time while keeping other activities and prices fixed. Economics also has *general equilibrium* models, which consider the simultaneous determination of all prices. These are more complex and are not well suited to the study of pricing by firms with market power or of strategic interaction between market players.

P.3 Smooth approximations

Most economic variables are not perfectly divisible. For example, prices must usually be quoted in a smallest currency unit, and most goods come in multiples of a smallest quantity (fax machines, bars of soap, sheets of paper). However, we can safely ignore these indivisibilities if the smallest unit is small compared to the relevant scale of prices or output. We thus obtain a smooth relationship between economic variables that simplifies much analysis.

Consider Figure P.1 (on page 15), which shows a firm's profit as a function of the level of output. The quantity is denoted by Q, the profit by Π, and the profit function (which relates each quantity to its profit) by $\pi(Q)$. (We denote profit by the Greek letter "pi": lowercase π and uppercase Π.) You can see that the optimal output level is only 8 units. Perhaps this firm is a small builder of single-family homes (and profit is measured in \$1000s).

Suppose instead that the firm builds in-ground swimming pools. The size of the unit is smaller compared to the optimal scale of output, so the graph of the profit function could look like Figure P.2.

Once the scale of output is up to, say, 200 or 300 units, which is the case for most firms, this graph will look almost like a smooth line. It is then a reasonable approximation to treat the good as perfectly divisible (like oil or water), so that the graph looks like Figure P.3.

P.4 Decomposition of decision problems

Decomposing a decision problem means dividing it into several smaller and simpler problems. This is useful because it is easier to understand one simple problem at a time than an entire complex problem all at once. Furthermore, one can distribute the task of making decisions among the many members of an organization, thereby making use of more diversified brain power.

For example, suppose your task is to choose the output level Q of your single-product firm. As a simple accounting identity, profit equals revenue minus cost: $\Pi = R - C$. To express this as a function of Q, we let $r(Q)$ be the highest possible revenue when selling

Figure P.1

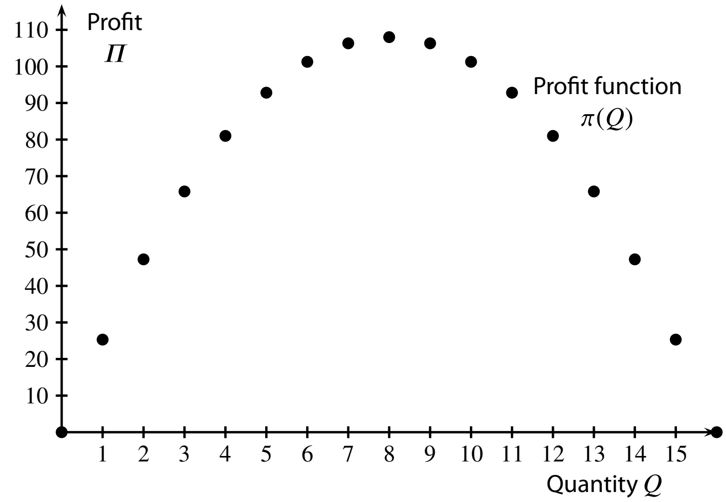


Figure P.2

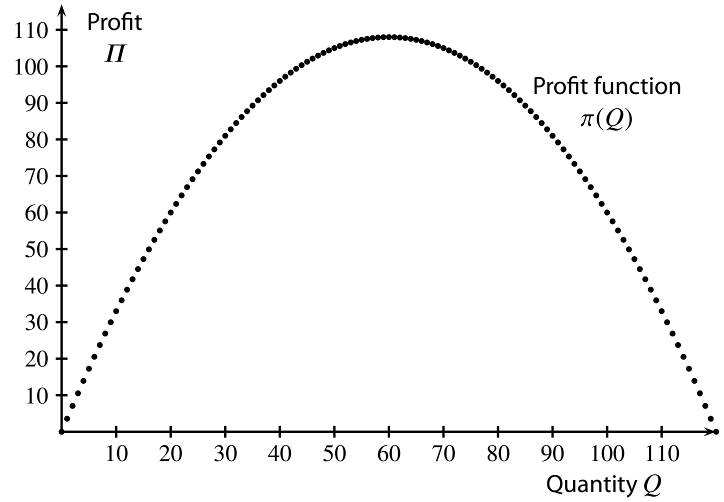
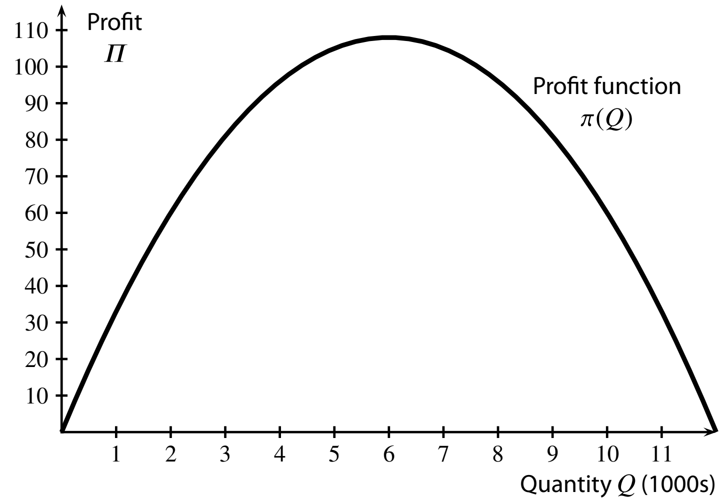


Figure P.3



Q units and let $c(Q)$ be the lowest possible total cost when the output level is Q. Then $\pi(Q) = r(Q) - c(Q)$.

We can see how maximizing profit involves trading off revenue and cost. Increasing output from Q_1 to Q_2 raises profit if and only if the increase in revenue exceeds the increase in cost: $r(Q_2) - r(Q_1) > c(Q_2) - c(Q_1)$.

Furthermore, we can delegate the task of determining revenue to the marketing department and can delegate the task of determining cost to the production department. For example, the marketing department must determine how to market the output and what price to charge in order to achieve the highest revenue when selling Q units. We can then set the output level knowing only the functions $r(Q)$ and $c(Q)$, and without having to know the details of the tasks we delegated to the marketing and production departments.

P.5 Marginal analysis

> *Firms are not frictionless reflections of their momentary environments, but rather highly inertial action repertoires, responding to—indeed perceiving—today's environment largely in terms of lessons learned from actions in days gone by.*
>
> Workshop at the Santa Fe Institute[1]

> *As with any new technology, the early years of the Internet have been a learning process—and here's what we now know. First, the Internet was supposed to change everything. That's just plain wrong. Clearly, in much of the economy, the Internet offers incremental payoffs without substantially altering core businesses.*
>
> Article in *Business Week Online*[2]

Suppose you come into an organization as a consultant or as a new CEO. Your job is to make this company perform as well as possible. You should start from scratch, right? Don't even look at the current state of affairs. That's what the "change gurus" say: "The larger the scale of change, the greater the opportunity for success" (James Champy).

This sounds exciting, but it ignores the complexity of large-scale change. Shying away from complexity is not a sign of weakness. It is a sign of wisdom, of recognizing an iron law: The more complex a problem is, the more likely mistakes are and the more costly it is to avoid them. This section is about how to make use of the simplicity of incremental change and how to recognize when it works.

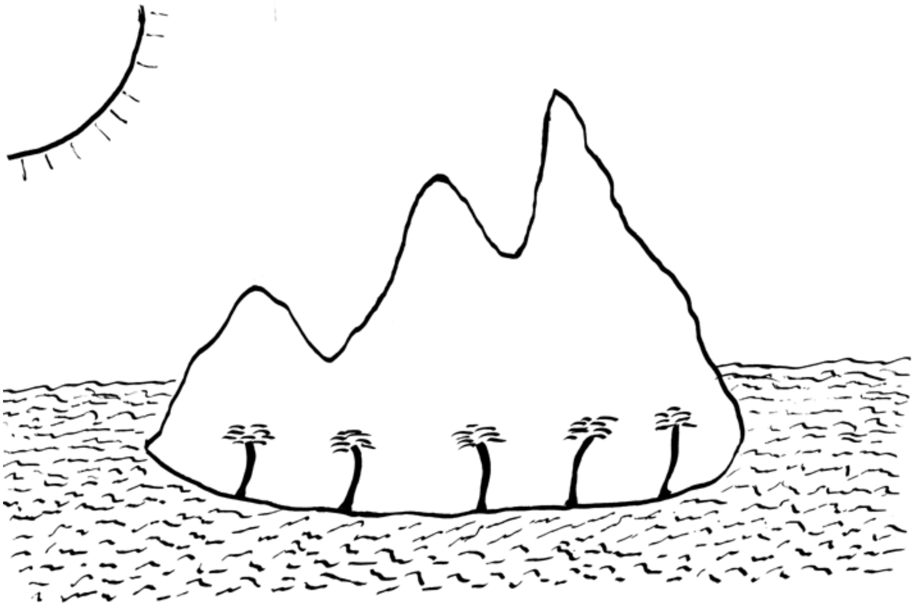
[1] Michael Cohen, Roger Burkhart, Giovanni Dosi, Massimo Egidi, Luigi Marengo, Massimo Warglien, and Sidney Winter. "Routines and Other Recurring Action Patterns of Organizations: Contemporary Research Issues." *Industrial and Corporate Change*, 1996.

[2] Michael Mandela and Robert Hof. "Rethinking the Internet." *Business Week Online*, 26 March 2001.

Marginal conditions: An allegory

Suppose a blind man wants to get to a peak on an island, such as the one drawn in Figure P.4.

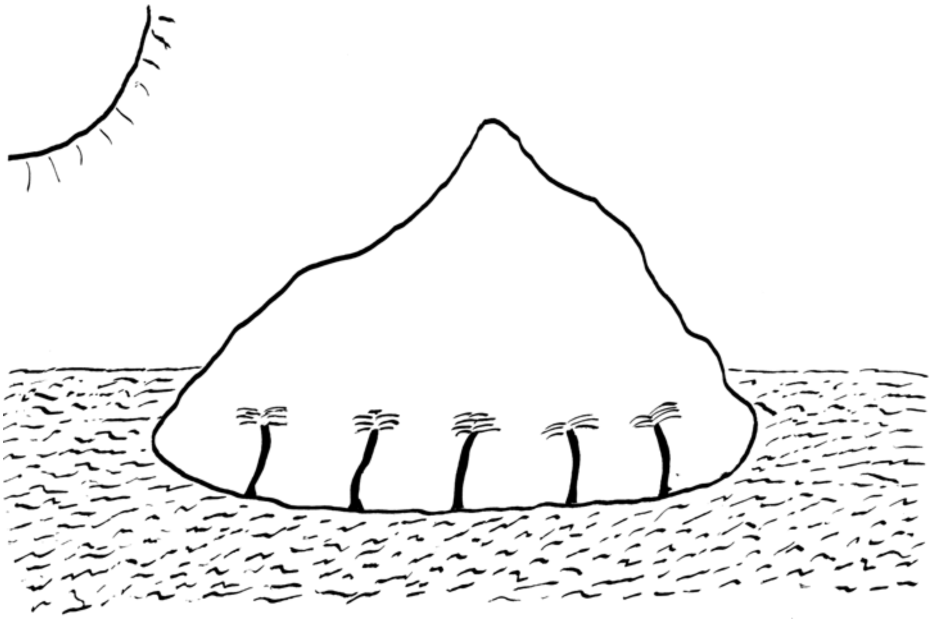
Figure P.4



To *check* whether he is at a peak, he need only tap his cane to see if any point around him is higher—if not, then he is at the top of a peak. This is called the *local* or *marginal condition for optimality*. To *reach* a peak from any point on the island, he can tap his cane to see in which direction the slope goes up and then take his next step in that direction, repeating until there is nowhere higher to go. This is called *local* or *incremental search*.

Such local methods are useful to the blind man. However, on the island in Figure P.4, the blind man cannot be sure to find the highest possible peak. On the other hand, local methods work perfectly on the island in Figure P.5 because it has a single peak, with the rest of the island sloping up toward it.

Figure P.5



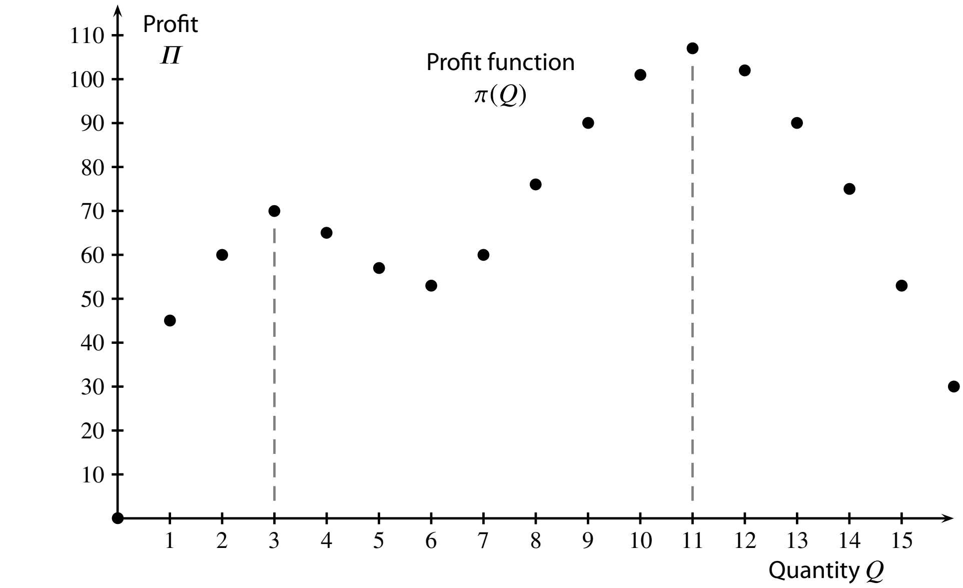
So, although local methods are useful on both islands, only on the second island can they be used on their own to find the highest point on the island. Therefore, we say that *marginal conditions are sufficient* on the second island but not on the first.

Marginal conditions: A manager

Let's relate the story about the blind man to managerial decision making. Suppose you manage a firm that produces a single indivisible good and you must decide how much to produce and sell in order to maximize the firm's profit $\pi(Q)$. Profit is zero if you produce and sell nothing (zero cost and zero revenue). The firm takes a loss if you produce too much because, in order to sell the output, you have to lower your price to below the average cost of production. For intermediate output levels, you can make a profit. The question is: how much should you produce?

Suppose that, for the first 16 units, the graph of the profit function is as shown in Figure P.6.

Figure P.6



The two peaks, at $Q = 3$ and $Q = 11$, are called *local optima* because each gives a higher profit than nearby quantities. The quantity $Q = 11$ is the *global optimum*—it gives you the overall highest profit.

Realistically, you do not actually know the exact relationship between quantity and profit. You can determine the profit for particular quantities either by experimenting or by acquiring information and performing calculations, but doing so for all quantities would be costly. Thus, you are like the blind man: you can feel the terrain around you but you do not know the topography of the entire island.

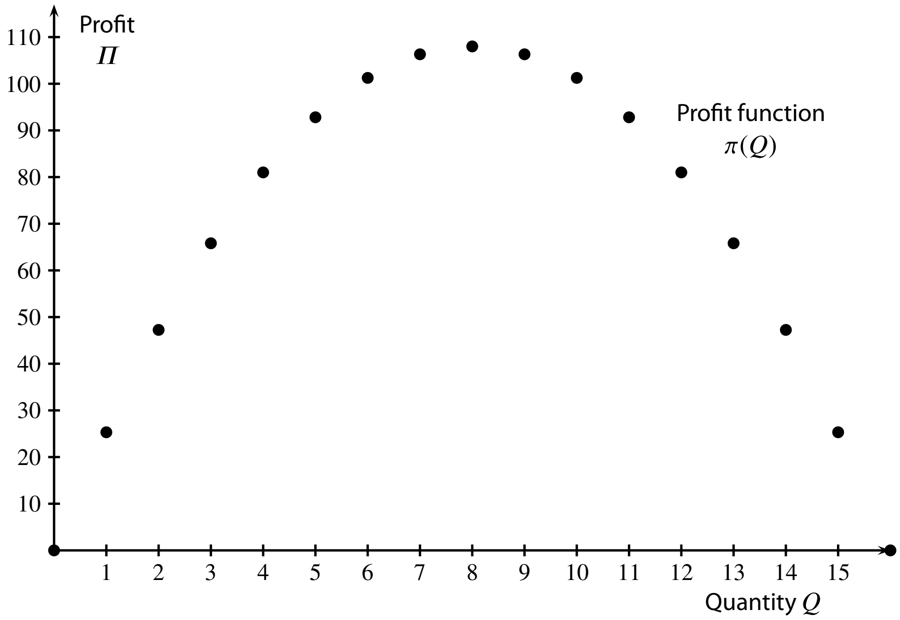
Still, you can at least reach a local optimum by using marginal conditions and local search.

- *Marginal conditions*: To check whether a level Q is a local optimum, it suffices to check that the profit is not higher at the *nearby* alternatives $Q - 1$ and $Q + 1$. Only at $Q = 3$ and $Q = 11$ will you find that neither of the two nearby quantities yields a higher profit.
- *Local search*: To find a locally optimal level of sales, you can search *nearby* for improvements until you find no better alternatives. Suppose you start with output equal to 5. You check the quantities 4 and 6 and determine that 4 gives you a higher profit. Then you check 3 and find it is also better. Continuing, you find that decreasing your output to 2 does not lead to a further improvement. Hence, you have found that 3 is a local optimum.

If your profit function is as shown in Figure P.6, such local or marginal analysis is useful for finding and identifying local optima. Yet it does not guarantee, by itself, that you find the profit-maximizing output level (the global optimum). In the preceding example of local search, you set the output to 3, unaware that a jump in your output to 11 would give a higher profit.

However, marginal analysis works perfectly if your profit function has a single peak, as in Figure P.7. Then every local optimum is a global optimum and we say that *marginal conditions are sufficient*.

Figure P.7



In this book we maintain an unstated assumption that marginal conditions are sufficient, except in a few cases in which we state otherwise: a competitive firm with a U-shaped average cost curve; a firm with market power that has a fixed cost; and a consumer that faces a two-part tariff. In these cases, the only required supplement to marginal analysis is that you have to check a shut-down option.

P.6 The mathematics of marginal analysis

Marginal profit

Consider again the problem of choosing output that will maximize profit $\pi(Q)$. "Marginal conditions" mean conditions that must be satisfied in order for a quantity Q to be a local optimum. We state marginal conditions in terms of the marginal profit, which is defined differently depending on whether we have a discrete or smooth function.

$$\begin{aligned} \text{Discrete:} \quad & m\pi(Q) = \pi(Q) - \pi(Q-1). \\ \text{Smooth:} \quad & m\pi(Q) = \pi'(Q). \end{aligned}$$

Marginal conditions in the discrete case

Consider the marginal conditions in the discrete case. In Figure P.6, $\pi(2) = 60$, $\pi(3) = 70$, and $\pi(4) = 65$. Therefore,

$$m\pi(3) = \pi(3) - \pi(2) = 70 - 60 = 10,$$
$$m\pi(4) = \pi(4) - \pi(3) = 65 - 70 = -5.$$

The fact that $m\pi(3) \geq 0$ tells us that lowering output from 3 to 2 would cause profit to fall (or stay the same); the fact that $m\pi(4) \leq 0$ tells us that raising output from 3 to 4 would cause profit to fall (or stay the same). Thus, together these conditions tell us that $Q = 3$ is at least a local optimum.

In summary, the marginal conditions for Q to be a local optimum are as follows:

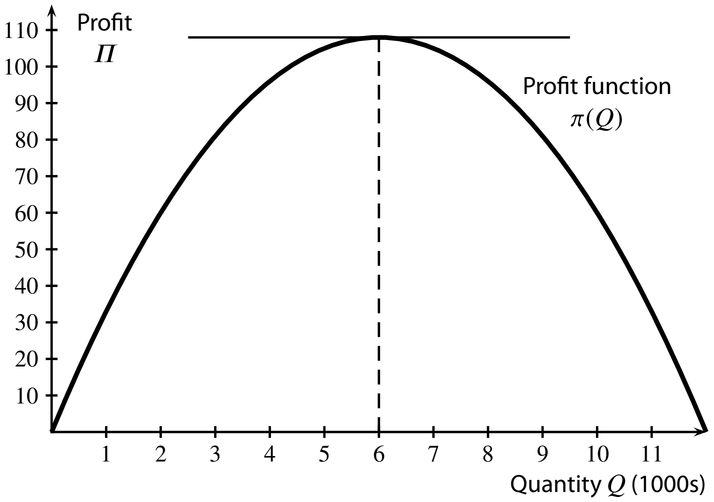
$$m\pi(Q) \geq 0 \quad \text{and} \quad m\pi(Q+1) \leq 0.$$

Thus, a local optimum is a point where marginal profit shifts from positive to negative. In Figure P.7, this happens at $Q = 8$; in Figure P.6, this happens at $Q = 3$ and $Q = 11$.

Marginal conditions in the smooth case

When the curve is smooth, as in Figure P.8, the slope of the profit function is exactly zero at a local optimum; this is illustrated by the horizontal line drawn in Figure P.8.

Figure P.8



In other words, the marginal condition for an optimum is $m\pi(Q) = 0$. This is simpler than the pair of inequalities $m\pi(Q) \geq 0$ and $m\pi(Q + 1) \leq 0$ in the discrete case, which is one reason that smooth functions are easier to work with.

Using marginal conditions for a numerical example

Though the real purpose of the marginal analysis (and our use of calculus for marginal analysis) in this text is to draw qualitative conclusions, marginal conditions can also be used to solve numerical examples. Suppose that a firm has the following revenue and cost curves:

$$\begin{aligned} r(Q) &= 60Q - 3Q^2, \\ c(Q) &= 24Q. \end{aligned}$$

Then the profit curve is

$$\pi(Q) = r(Q) - c(Q) = (60Q - 3Q^2) - 24Q = 36Q - 3Q^2.$$

Taking the derivative yields the marginal profit curve:

$$m\pi(Q) = 36 - 6Q.$$

We solve the marginal condition (also called the first-order condition in mathematics) as follows:

$$\begin{aligned} m\pi(Q) &= 0, \\ 36 - 6Q &= 0, \\ 36 &= 6Q, \\ Q &= 6. \end{aligned}$$

Thus (assuming that marginal conditions are sufficient, which is true in this case), the profit-maximizing quantity is $Q = 6$. This is the data that lies behind Figure P.8, in which one can see that $Q = 6$ maximizes profit and that $m\pi(6) = 0$.

Decomposition of marginal conditions

When we decompose a decision problem, we can also decompose the marginal conditions. Since $\Pi = R - C$, it follows that $M\Pi = MR - MC$ and that the marginal conditions can be restated as in Table P.1.

Table P.1

	Marginal conditions	
	In terms of marginal profit	In terms of marginal revenue and cost
Discrete case	$m\pi(Q) \geq 0$ $m\pi(Q+1) \leq 0$	$mr(Q) \geq mc(Q)$ $mr(Q+1) \leq mc(Q+1)$
Smooth case	$m\pi(Q) = 0$	$mr(Q) = mc(Q)$

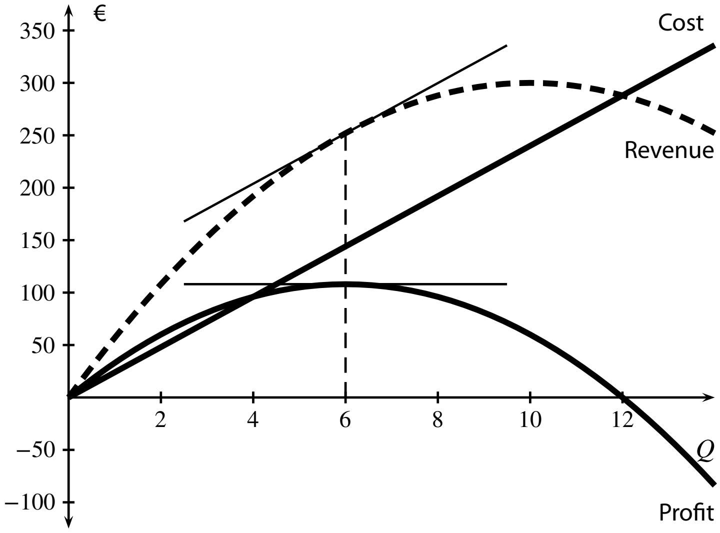
We usually state the marginal conditions as "marginal revenue equals marginal cost". This statement is exact for the smooth case and approximate for the discrete case.

Let's illustrate such decomposition using data from our previous example:

$$r(Q) = 60Q - 3Q^2,$$
$$c(Q) = 24Q,$$
$$\pi(Q) = 36Q - 3Q^2.$$

Figure P.9 shows these three curves on the same graph.

Figure P.9



Observe that at the profit-maximizing quantity $Q = 6$, where $m\pi(Q) = 0$, it is also true that the revenue and cost curves have the same slope—that is, marginal revenue equals marginal cost.

We can solve numerical examples by solving the marginal condition $mr(Q) = mc(Q)$. Taking the derivative of the revenue and cost curves yields

$$mr(Q) = 60 - 6Q,$$
$$mc(Q) = 24.$$

(The cost curve is linear and so has the constant slope 24.) We solve:

$$mr(Q) = mc(Q),$$
$$60 - 6Q = 24,$$
$$36 = 6Q,$$
$$Q = 6.$$

Of course, we get the same answer $Q = 6$ as when we solved $m\pi(Q) = 0$.

Exercise P.1. Suppose a firm's revenue and cost curves are

$$
\begin{aligned}
r(Q) &= 84Q - 3Q^2,\\
c(Q) &= 4Q + Q^2.
\end{aligned}
$$

a. Write the formula for the profit curve.

b. Using a spreadsheet, create a table with four columns: Output (ranging from 0 to 20), revenue, cost, and profit. By visual inspection of the profit column, determine the quantity Q^* that maximizes profit.

c. Using your spreadsheet program (or another graphing program), graph $r(Q)$, $c(Q)$, and $\pi(Q)$. Print out the graph. Mark the point where profit is maximized. Draw lines tangent to the graphs of $r(Q)$, $c(Q)$, and $\pi(Q)$ at the profit-maximizing Q^*. You should see visually that $m\pi(Q^*) = 0$ and that $mr(Q^*) = mc(Q^*)$.

d. Calculate $m\pi(Q)$ and find the profit-maximizing quantity by solving $m\pi(Q) = 0$.

e. Calculate $mr(Q)$ and $mc(Q)$. Find the profit-maximizing quantity by solving $mr(Q) = mc(Q)$.

P.7 Wrap-up

Our use of models and our analysis of decision problems are based on *simplification*. Only through simplification can we carefully describe the key elements of a situation and then apply logical reasoning.

The key tools for simplifying decision problems are decomposition and marginal analysis. These tools are important not only for analyzing other people's decisions but also as methods for making decisions ourselves.

Such logical analysis can be used to develop quantitative decision-making tools. However, it is also useful for making decisions on your feet using soft data. Improving such decision making is the ultimate goal of this book.

Chapter 1

Gains from Trade

1.1 Motives and objectives

Broadly

Let us step back from a specific managerial decision and consider the markets in which we make our transactions—whether we are managers (who hire workers, buy other inputs, and sell products) or consumers (who sell labor for a wage, trade houses, and buy food and automobiles).

Most of these markets have many (or at least several) sellers and buyers, none of whom controls the markets. The sellers and buyers have fairly narrow self-interested goals and are not looking out for the collective interests of all the participants. There are many different ways in which trade takes place. Markets are indeed complicated, disorderly beasts.

Yet we can build a simple model that captures the essential features of a variety of markets. This model helps us understand what transactions end up taking place and at what price. We can measure how buyers and sellers benefit from trade and we can answer the question of whether such markets are an efficient way to conduct trade. We can also determine the impact of a tax or of a trade restriction on buyers and sellers.

This is the subject of Chapters 1 and 2. We study a market for an *indivisible* good (e.g., paintings, refrigerators, houses, cars, books) in which each buyer purchases at most one unit (*unit demand*) and each seller supplies at most one unit (*unit supply*). The good is homogeneous, meaning that all units of the good are identical.

This does not describe many real markets. For example, though it is a good approximation to say that the each household wants to purchase at most one refrigerator or one furnace, the firms that sell such goods typically produce more than just one unit. In the used-housing market, both buyers and sellers are buying or selling at most one unit, but the goods in such a market are far from identical to each other.

However, as usual a model does not have to be realistic in order to help us understand the real world. The model of a simple market studied in Chapters 1 and 2 provides an introduction to basic concepts that appear throughout this book and that are relevant to realistic and complicated markets: (a) valuation and cost, (b) surplus, (c) market equilibrium, and (d) efficiency.

More specifically

We begin (in this chapter) by studying how to measure gains from trade and how to determine whether trade is efficient (in the sense that it would be impossible to make anybody better off without making someone worse off). Then (in Chapter 2), we consider what actually happens in markets and look at the effect of a tax.

Studying what is "socially optimal" (efficient) and then what actually happens or what a profit-maximizing firm would do is a pattern that repeats itself in this book.

1. Understanding the socially efficient outcome is of interest in itself. For example, when markets are inefficient, any unrealized gains from trade are simply lost and go to no one. Even a profit-maximizing firm would like to figure out how to generate such gains from trade and appropriate at least part of them for itself.
2. Understanding efficiency is—pedagogically—a good first step toward understanding market outcomes or the profit-maximizing solution.

We first introduce the valuation of a buyer, the cost of a seller, and the surplus that a trader gets from a transaction. Then we consider what it takes for trade between a single buyer and a single seller to be efficient. The answer is easy: trade should take place if the buyer's valuation exceeds the seller's cost and should not take place if the opposite is true.

When there are several buyers and sellers, trade is efficient if it maximizes the total surplus: the total valuation of the buyers minus the total cost of the sellers. The marginal condition for maximizing total surplus is that marginal valuation be equal to marginal cost. We illustrate these ideas graphically.

1.2 Efficiency

We often take the point of view of one of the market participants in our models. For example, what would or should you do if you were the manager of one of the firms? However, we will also find it useful to evaluate outcomes from the point of view of a third-party observer. We use a fairly weak and inconclusive criterion called "efficiency" (for short) or "social efficiency" (to emphasize that we are taking into account everyone's preferences rather than just those of a particular market participant) or "Pareto efficiency" (named after the 19th-century Italian economist Vilfredo Pareto, 1848–1923).

For example, suppose that Yakov lives next to a factory that is owned by Zahra. (To keep things simple, there are no other residents or factories in the vicinity.) Zahra's factory could make a lot of noise using a low-cost technology or less noise using a technology that costs $200/month extra. Suppose that the noise of the low-cost technology is so much that Yakov would be willing to pay $300/month to reduce the noise. Consider these two outcomes:

X. Zahra uses the low-cost technology;
Y. Zahra uses the high-cost technology and Yakov pays Zahra \$250.

Both Zahra and Yakov prefer Y over X. Given such unanimity among the participants, we, as outside observers, should also agree that Y is better than X.

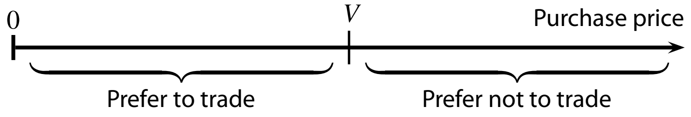
We therefore say that X is *inefficient* because there is an alternative that benefits at least one party without hurting another party. On the other hand, as long as Zahra uses the noise-reducing technology the outcome is *efficient*, meaning that it is impossible to make someone better off without hurting someone else.

This efficiency criterion is weak and inconclusive because we rank only those outcomes that the market participants themselves unanimously rank the same way. In our example, there are many efficient outcomes, which differ by the amount (if any) that Yakov pays Zahra. Of course, Yakov prefers to pay less or nothing; Zahra prefers to be paid more. In this book, we do not judge which of these outcomes is better—doing so would require more information about the circumstances as well as contentious criteria about fairness.

1.3 Valuation, cost, and surplus

Now let's turn to the market that we want to study. Recall that each seller has one unit to sell and each buyer is interested in buying only one unit. We use the pronouns "he" for buyers and "she" for sellers.

Suppose that the buyer is faced with the following two options: Either he purchases the good at a price P or he must walk away and not purchase the good at all. Which option does he prefer? For a range of low prices, he prefers to trade (purchase the good); for a range of higher prices he prefers to walk away. There is some cutoff point V that divides the prices at which he trades and the prices at which he does not:



This cutoff V is called the buyer's *valuation*. (Equivalent terms are *willingness to pay* and *reservation price*.)

If he purchases the good, then $V - P$ is called the *buyer's surplus*. If he does not, then his surplus is 0. The buyer prefers outcomes that give him the highest surplus.

The seller has a cutoff price above which she prefers to sell the good and below which she prefers not to. This cutoff is called the seller's *cost* and is denoted by C. It might literally be an amount the seller must pay to produce or obtain the good. Alternatively, it may be how much she values keeping the good for herself: her "opportunity cost" of giving up the good.

If she sells the good at a price P, then $P - C$ is called the *seller's surplus*. If she does not, then her surplus is 0. The seller prefers outcomes that give her the highest surplus.

1.4 One buyer and one seller

Suppose there are a single buyer and a single seller: Yakov and Zahra. They must decide whether or not to trade and at what price. If $C < V$ then it is efficient to trade and inefficient not to trade: both Yakov and Zahra are better off trading at a price P such that $C < P < V$. There are many efficient trades, which differ only in the price that is paid and hence in the distribution of surplus. (If instead $C > V$, then *not* trading is efficient.)

1.5 Many buyers and many sellers

Conditions for efficiency

Suppose there are many buyers and many sellers. Trade is efficient (i.e., it would be impossible to make one person better off without making someone else worse off) if and only if it maximizes total possible surplus (gains from trade): the total valuation of the buyers who buy the good minus the total cost of the sellers who sell the good. The prices at which trade takes place determines the distribution of the surplus but not its total value.

To maximize total surplus, the units traded should go to the highest-valuation buyers and come from the lowest-cost sellers. Taking this as given, we can measure surplus as a function of trading volume Q as follows.

- Total valuation of the buyers is the sum of the Q highest valuations; denote this by $v(Q)$.
- Total cost of the sellers is the sum of the Q lowest costs; denote this by $c(Q)$.
- Total surplus is $v(Q) - c(Q)$.

For efficiency, the volume Q should maximize $v(Q) - c(Q)$. Let's state the marginal conditions, as described in Section P.5. First, we define marginal valuation $mv(Q)$ and marginal cost $mc(Q)$ of the Qth unit traded:

1. $mv(Q) = v(Q) - v(Q - 1)$, which is the Qth-highest valuation of the buyers;
2. $mc(Q) = c(Q) - c(Q - 1)$, which is the Qth-lowest cost of the sellers.

Then the marginal conditions for volume Q to maximize total surplus are that

1. $mv(Q) \geq mc(Q)$ (surplus does not go up by trading one fewer unit), and
2. $mv(Q + 1) \leq mc(Q + 1)$ (surplus does not go up by trading one more unit).

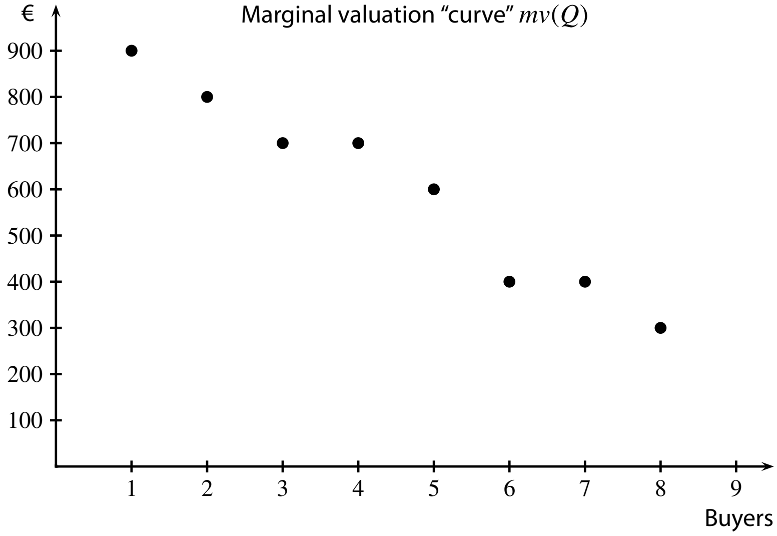
Stated more succinctly in a way that is approximate here but exact for the case of a smooth model, the marginal condition is $mv(Q) = mc(Q)$.

We will now illustrate in detail (a) marginal and total valuation, (b) marginal and total cost, and (c) gains from trade and efficiency.

Illustrating total valuation and marginal valuation

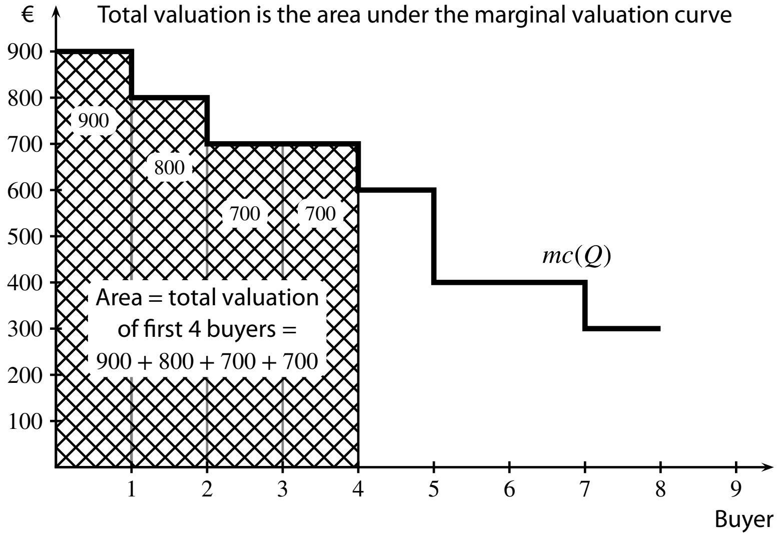
We can graph the marginal valuation curve by plotting the consumers' valuations from highest to lowest. For example, suppose there are 8 consumers with valuations 300, 400, 400, 600, 700, 700, 800, and 900. The marginal valuation curve is shown in Figure 1.1.

Figure 1.1



Let's redraw this graph as a solid line such that the area under the curve up to Q is the sum $v(Q)$ of the Q highest valuations.

Figure 1.2

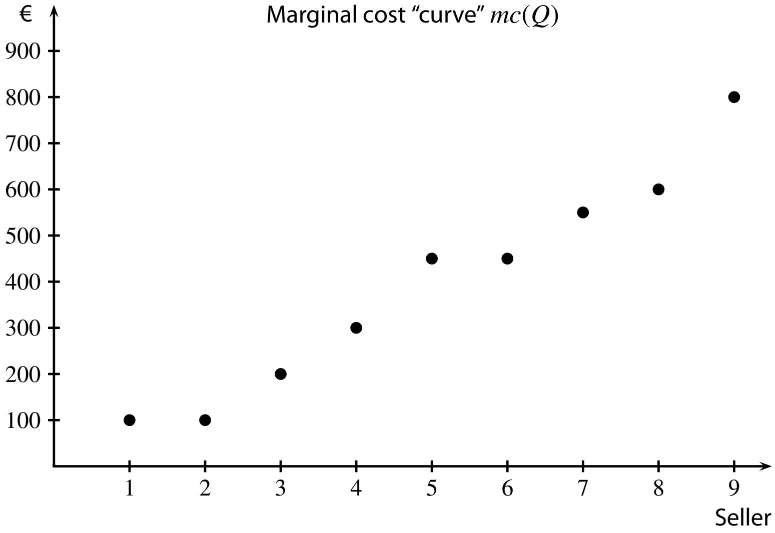


This way, we can visualize marginal valuation and total valuation on the same graph. For example, the total valuation $v(4)$ of the first 4 buyers is $900+800+700+700 = 3100$, shown in Figure 1.2.

Illustrating total cost and marginal cost

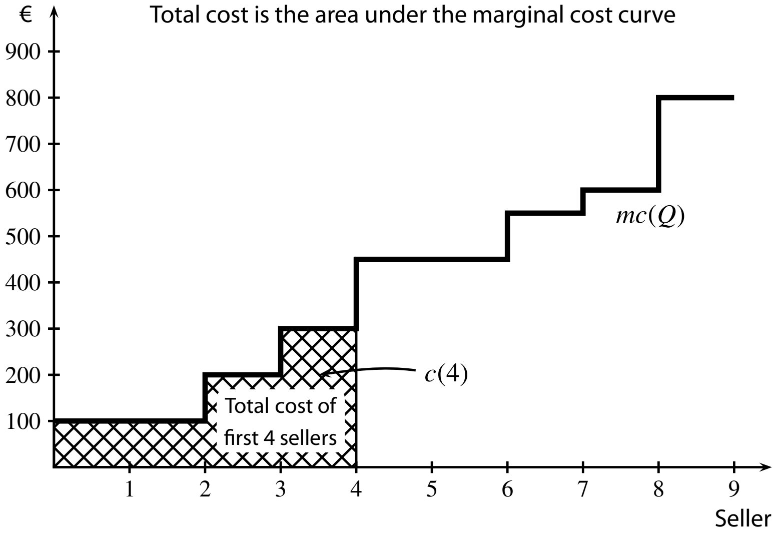
We can similarly illustrate marginal and total cost on the same graph. Suppose that there are 9 potential sellers whose costs are 100, 100, 200, 300, 450, 450, 550, 600, and 800. Number the sellers from lowest cost to highest cost. Figure 1.3 shows the graph of their costs; it is the graph of the marginal cost curve $mc(Q)$.

Figure 1.3



By connecting the dots in the graph, we see that $c(Q)$ is the area under the graph up to Q. For example, the total cost $c(4)$ of the first 4 sellers is $100 + 100 + 200 + 300 = 700$, as illustrated in Figure 1.4.

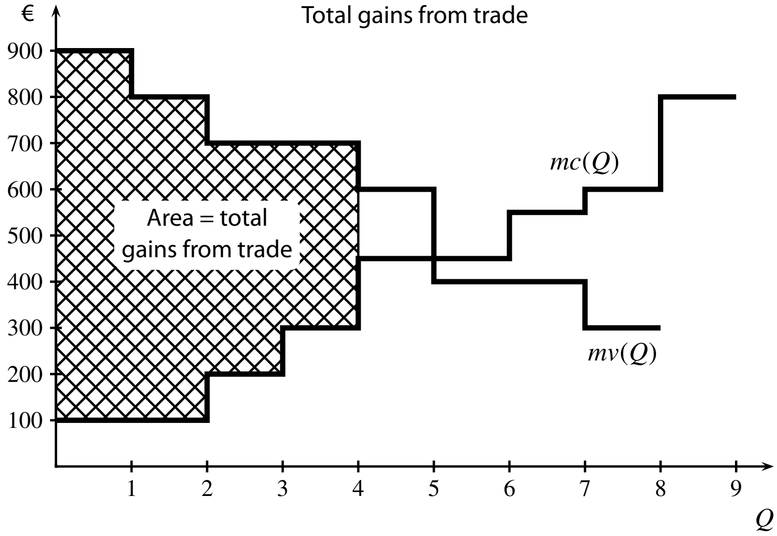
Figure 1.4



Illustrating gains from trade

Let's put the marginal valuation and marginal cost curves on the same graph. Figure 1.5 illustrates the total gains from trade when 4 units are traded. The total valuation $v(4)$ generated for the buyers is the area under the marginal valuation curve up to $Q = 4$. In order to find the total gains from trade, we should subtract $c(4)$, which is the area under the marginal cost curve up to $Q = 4$. This leaves the area between the two curves.

Figure 1.5

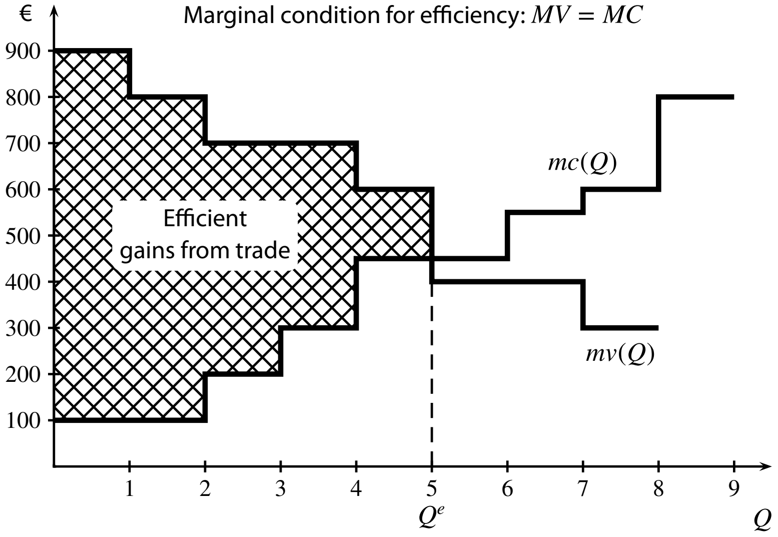


Illustrating efficiency

We can see in Figure 1.5 that $Q = 4$ does not maximize the total gains from trade. There is a buyer with a valuation of 600 and a seller with a valuation of 450 who could trade, generating an extra 150 of surplus. Hence, the efficient quantity is instead $Q = 5$. Beyond that amount, any buyer and seller who have not yet traded are such that the buyer's valuation is lower than the seller's cost; forcing them to trade would simply reduce the surplus.

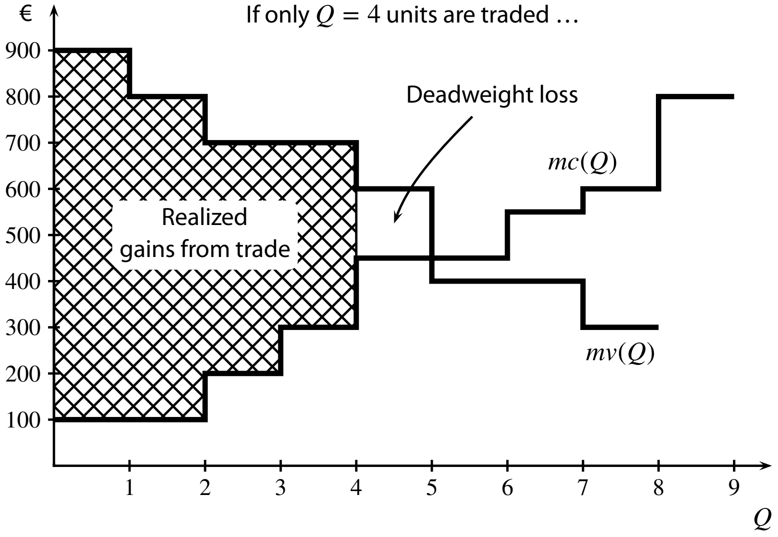
The efficient quantity occurs where the marginal cost and marginal valuation curves (with dots connected in our particular way) intersect. This is a graphical indication that the marginal conditions are satisfied. Figure 1.6 shows the gains from trade when the efficient quantity Q^e is traded.

Figure 1.6



If (for some reason) trade is not efficient, then the difference between the total possible gains from trade and the realized gains from trade is called the *deadweight loss*. For example, if only 4 units are traded then the deadweight loss is 150. This is illustrated in Figure 1.7.

Figure 1.7



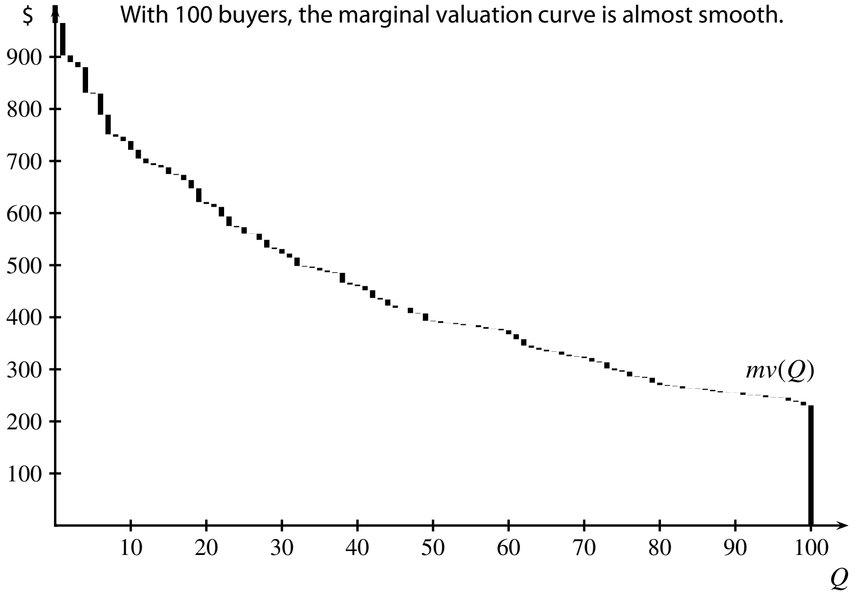
1.6 Very many buyers and very many sellers

The indivisibility of the good in the previous section is quite apparent in our graphs. However, this becomes less important at the aggregate level when the market is large. Suppose, for example, that in the market for refrigerators there are a hundred buyers who have the following valuations.

230.89	237.81	239.86	245.71	246.14	246.50	250.26	250.79	251.20	255.56
255.71	255.91	258.10	260.67	262.74	263.19	263.43	267.70	268.42	269.80
274.48	283.93	285.67	286.57	295.72	298.26	301.92	313.59	315.13	321.84
324.58	325.34	328.36	334.22	334.68	337.79	341.63	345.80	357.98	367.57
375.19	377.90	378.19	381.24	384.95	385.02	387.21	389.12	389.35	392.61
393.47	407.49	408.27	418.39	418.41	422.65	434.30	437.32	452.05	460.17
462.96	466.70	485.48	486.74	490.15	495.12	497.47	498.76	514.92	522.55
531.31	534.14	548.98	560.77	561.06	573.19	575.58	593.94	612.22	618.00
621.65	647.98	663.85	673.79	675.50	688.30	692.74	696.20	705.23	721.95
738.85	747.13	751.75	789.41	830.05	831.67	880.84	890.60	903.43	965.69

The marginal valuation curve (the graph of these valuations from highest to lowest) begins to look smooth, as is seen in Figure 1.8.

Figure 1.8

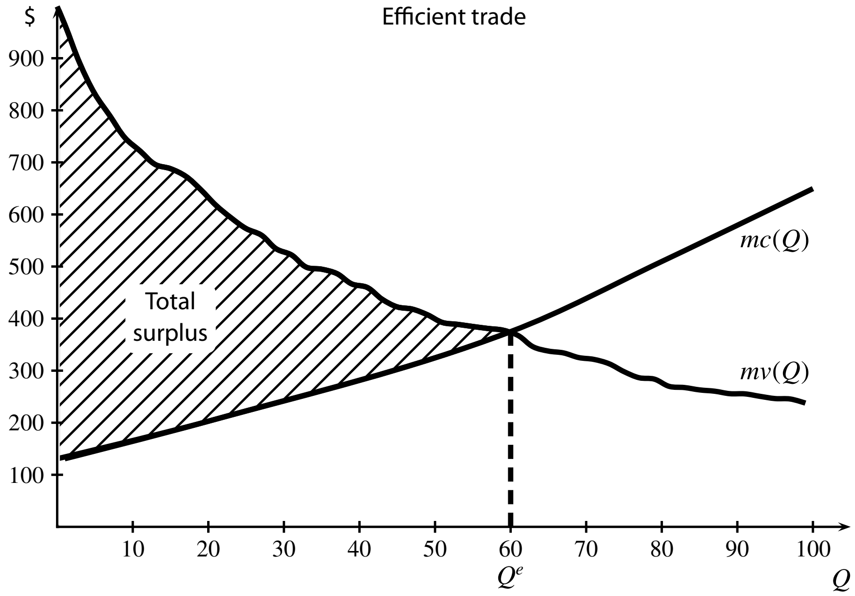


With 100 buyers, the marginal valuation curve is almost smooth.

Hence, even when a good is indivisible, we can model marginal and total valuation as smooth functions as long as there are many buyers with diverse valuations. Similarly, we can model the marginal and total cost as smooth functions as long as there are many sellers with diverse costs.

Our graphical analysis of the previous section still applies. Figure 1.9 illustrates the efficient quantity Q^e and the total gains from trade.

Figure 1.9



1. The area under the marginal valuation curve up to Q equals the total valuation $v(Q)$.
2. The area under the marginal cost curve up to Q equals the total cost $c(Q)$.
3. When Q units are traded between the lowest-cost sellers and highest-cost buyers, the total gains from trade are the area between the marginal cost and marginal valuation curves up to Q.
4. The efficient quantity is where the marginal cost and marginal valuation curves intersect, illustrating the marginal condition $mv(Q) = mc(Q)$.

1.7 Sources of gains from trade

One of the sources of gains from trade is production and specialization. Modern production is organized such that firms, as organizations with separate legal identities, purchase inputs from people (and from other firms) and sell their output to people (and to other firms).

Such gains from trade are quite clear. However, there are other sources of gains from trade that are independent of production. They come from diversity: there are gains from trade because people have or want different things.

1.8 Wrap-up

This chapter introduced some basic concepts that are used frequently in this book.

1. A buyer's preferences can be summarized by his *valuation*; a seller's preferences can be summarized by her *cost*.
2. If a buyer with valuation V and a seller with cost C trade, then the *gains from trade* are $V - C$. If they trade at price P, then the buyer's share of these gains is $V - P$ and is called the *buyer's surplus*, while the seller's share is $P - C$ and is called the *seller's surplus*.
3. The graph of the valuations of buyers in a market is also the graph of the marginal valuation curve. The area under the marginal valuation curve up to quantity Q is the total value generated for the buyers when the Q buyers with the highest valuations obtain the good.
4. The graph of the costs of the sellers is also the graph of the marginal cost curve. The area under the marginal cost curve up to quantity Q is the total cost incurred by the sellers when the Q sellers with the lowest costs sell the good.

Furthermore, we defined and analyzed the efficiency of trade. The marginal condition for the efficient level of trade is $mv(Q) = mc(Q)$.

Chapter 2

Supply, Demand, and Markets

The immediate "common sense" answer to the question "What will an economy motivated by individual greed and controlled by a very large number of different agents look like?" is probably: There will be chaos.

— Kenneth Arrow and Frank Hahn, *General Competitive Analysis*, 1971

2.1 Motives and objectives

Broadly

We continue our study, started in Chapter 1, of a market for an indivisible good in which each buyer is interested in purchasing only one unit and each seller has only one unit to sell. Whereas in Chapter 1 we studied gains from trade and what kind of trade would be efficient, here we study what actually happens in such a market.

More specifically

We first consider the outcome of the bargaining problem between one buyer and one seller. Then we consider markets with many buyers and sellers. A common market price emerges. The equilibrium market price is the one at which supply equals demand.

We introduce the demand curve, which measures the number of buyers who would like to buy at each price. It is the inverse of the marginal valuation curve. We also introduce the supply curve, which measures the number of sellers who would like to sell at each price. The condition that supply equals demand is illustrated graphically as the intersection of the demand and supply curves.

We can then illustrate graphically how the total gains from trade are divided among the buyers ("consumer surplus") and among the sellers ("producer surplus"). We can also see that the equilibrium trade is efficient.

Next we consider the equilibrium when a per-unit tax is imposed on any transaction. This restricts trade to below the efficient level, resulting in a deadweight loss.

2.2 Bargaining: One buyer and one seller

Main idea

Recall the situation in which a single buyer named Yakov, with valuation V, and a single seller named Zahra, with cost C, have to decide whether to trade and at what price.

Yakov and Zahra will not agree to trade if $C > V$: there is no trade that would not imply a loss for one of the traders. If instead $C < V$, then we would expect them to trade: why would they walk away when both know that they could find a mutually beneficial trade? Thus, we expect trade to be efficient, no matter how it is conducted.

The distribution of the gains from trade will depend on the bargaining procedure and on subtle traits, such as stubbornness and patience, that give each trader bargaining power. If the bargaining procedure does not favor either party then a rule of thumb is that the buyer and seller divide the gains from trade, setting a price that is midway between the seller's cost and the buyer's valuation. However, a more patient party or a party with a reputation for intransigence may extract greater gains.

There are other bargaining procedures that give one party more bargaining power. Being able to credibly make a take-it-or-leave-it offer, meaning that negotiations are irrevocably broken off if the other party rejects the offer, confers the highest bargaining power. How should Yakov proceed if he has such bargaining power? He should offer the lowest price that Zahra would accept: *Yakov offers $P = C$, Zahra accepts, and Yakov gets all the gains from trade.*

A messy detail and a way around it

There is a little detail to discuss because it comes up so often.

We just claimed that Zahra would accept a price equal to her cost, even though doing so gives her no gains from trade and leaves her no better off (nor worse off) than if she rejected the offer. In fact, we cannot say for sure how Zahra would respond to such an offer. If Yakov is concerned that Zahra would not accept $P = C$, then he should offer C plus the smallest currency unit, which guarantees that the Zahra does accept.

This precise answer is hardly different from our first description of the outcome. Either way, Yakov offers a price very close to C, Zahra accepts, and Yakov gets all or nearly all the gains from trade. However, the precise answer muddles the analysis and the conclusion.

In order to avoid such unnecessary detail, we state—once and for all—a simplifying assumption that applies throughout this book: *If a party to a trade is indifferent between accepting and rejecting an offer, then she accepts it.* In our example, this means that Zahra accepts $P = C$, so our first and simpler conclusion is correct.

2.3 Many buyers and many sellers: Equilibrium

Consider the case in which there are many buyers and many sellers. There are various mechanisms for conducting trade. Trade may be unstructured as in a pit market, with buyers and sellers intermingling and conducting bilateral negotiations for trade; if a pair choose not to trade then they break up and search for new trading partners. There may be a more organized market in which proposed trades are handled electronically or manually by a central clearing house; this is how many financial transactions are conducted. There may be intermediaries who buy and sell and who hold temporary inventories as a means of transferring goods from sellers to buyers.

In all these cases, if the trading process is visible to all the traders, if the transaction costs and other frictions are small, and if the trade is recurrent, then eventually all trade should take place at the same price. Why should a buyer pay $10 knowing that he could pay $8? This is true even if trade takes place bilaterally. The price at which a party is willing to trade is no longer determined solely by her valuation, but also by what terms of trade she expects to obtain by breaking off and looking for a new trading partner. The market price that emerges defines this outside option.

We say that a market is in equilibrium when a market price emerges and everyone who wants to trade at this price does so. For any price P, let $d(P)$ be the number of buyers willing to purchase at this price (the *demand*) and let $s(P)$ be the number of sellers willing to sell at this price (the *supply*). The function $d(P)$ is called the demand curve, and $s(P)$ is called the supply curve. At the equilibrium price P^*:

1. trade is optional;
2. everyone who wants to trade does so; and
3. the number of purchases must equal the number of sales.

Hence demand equals supply: $d(P^*) = s(P^*)$.

We can also see that the equilibrium trade is efficient. Among those who trade, any buyer's valuation is at least P and any seller's cost is at most P; hence, all such trades generate surplus. Among those who do not trade, any buyer's valuation is at most P and any seller's cost is at least P; hence, there are no extra gains from trade to be generated.

2.4 Demand and supply: Graphical analysis

In order to graphically relate the demand and supply curves to consumer and producer surplus, we first show that the demand curve is the inverse of the marginal valuation curve and that the supply curve is the inverse of the marginal cost curve.

Recall from Section M.2 of the Math Review that we can represent a discrete function by a table. Suppose, for example, that the marginal valuation "curve" is the one in Table 2.1.

Table 2.1

	Q	MV	
	1	900	
	2	800	
	3	750	
	4	600	
→	5	550	←
	6	475	
	7	400	

Recall also (from Section M.4) the relationship between a function and its inverse:

- As a function, we look up, say, $Q = 5$ in the left column and find the marginal valuation $mv(5) = 550$ in the right column.
- As an inverse, we would look up a marginal valuation, say 550, in the right column and find the corresponding quantity 5 in the left column.

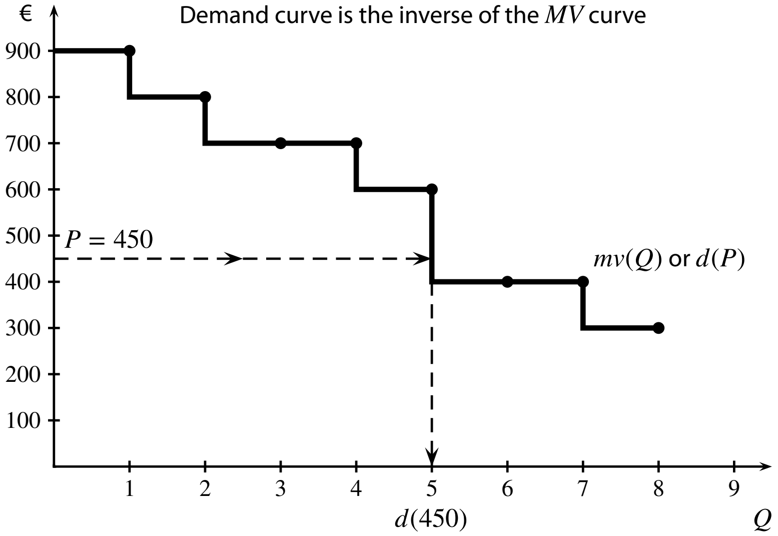
This inverse is also the demand $d(550)$:

$$\begin{aligned} mv(5) = 550 &\implies \text{5th-highest valuation is 550} \\ &\implies \text{there are 5 buyers whose valuations are at least 550} \\ &\implies d(550) = 5. \end{aligned}$$

The argument works as well for prices that are not equal to one of the valuations and also if there are multiple buyers with the same valuation. We just need to connect the dots in the graph of the valuations as we did in Section 1.5.

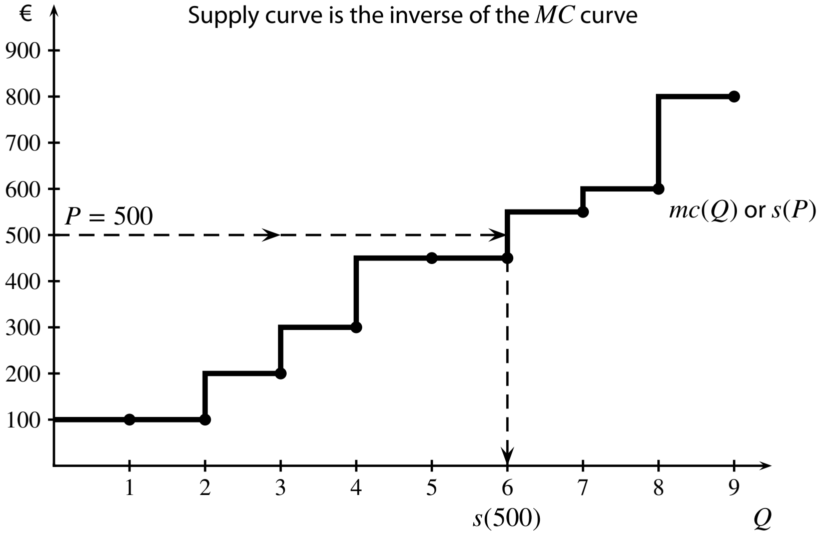
Figure 2.1 shows the marginal valuation curve from Section 1.5. To graph its inverse we could redraw the graph so that price or valuation is on the horizontal axis, but this is not necessary—we can simply treat the vertical axis as the *independent* variable and interpret it as price while treating the horizontal axis as the *dependent* variable. For example, $d(450)$ is the number of people willing to pay at least 450. From the graph, we see that this number equals 5.

Figure 2.1



Analogously, the supply curve is the inverse of the marginal cost curve. Consider the numerical example from Section 1.5; the graph of the marginal cost curve is shown in Figure 2.2. For example, $s(500)$ is the number of sellers willing to sell at 500 or less, which in turn is the number of sellers whose costs are at most 500. From Figure 2.2, we see that this number is 6.

Figure 2.2



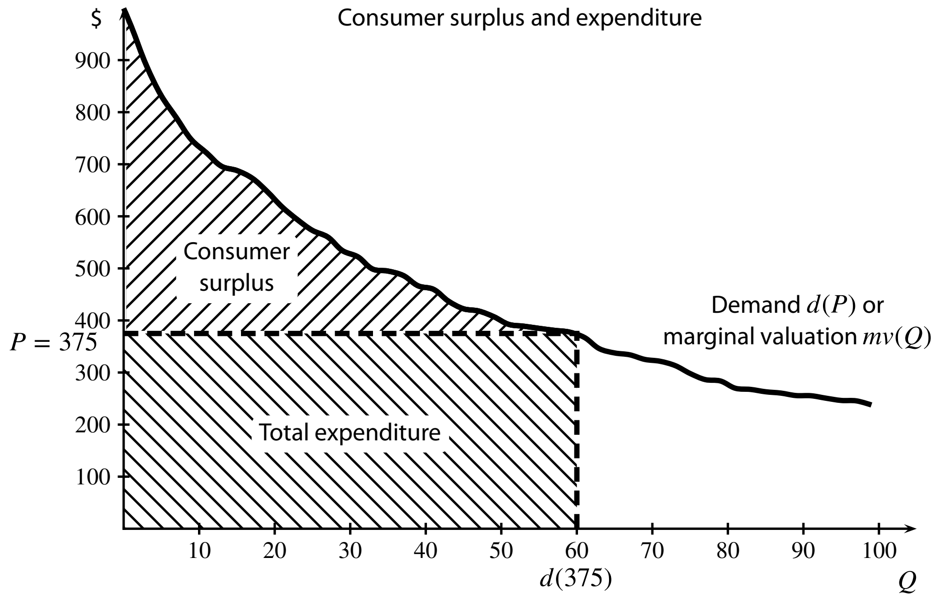
2.5 Consumer and producer surplus

We will continue, throughout this book, to draw demand and supply with price on the vertical axis and quantity on the horizontal axis. This is a standard practice in economics, adopted long ago because it saves us the confusion of redrawing the graph as we switch between the interpretation as a marginal valuation curve (or marginal cost curve) and the interpretation as a demand curve (or supply curve).

Furthermore, we can use the same graph to visualize demand, marginal valuation, total valuation, expenditure, and consumer surplus (or supply, marginal cost, total cost, revenue, and producer surplus). We illustrate this with the smooth marginal valuation and marginal cost curves from Section 1.6.

Figure 2.3 shows the marginal valuation curve (i.e., the demand curve). Suppose the price is $P = 375$. We can see that $d(375) = 60$.

Figure 2.3

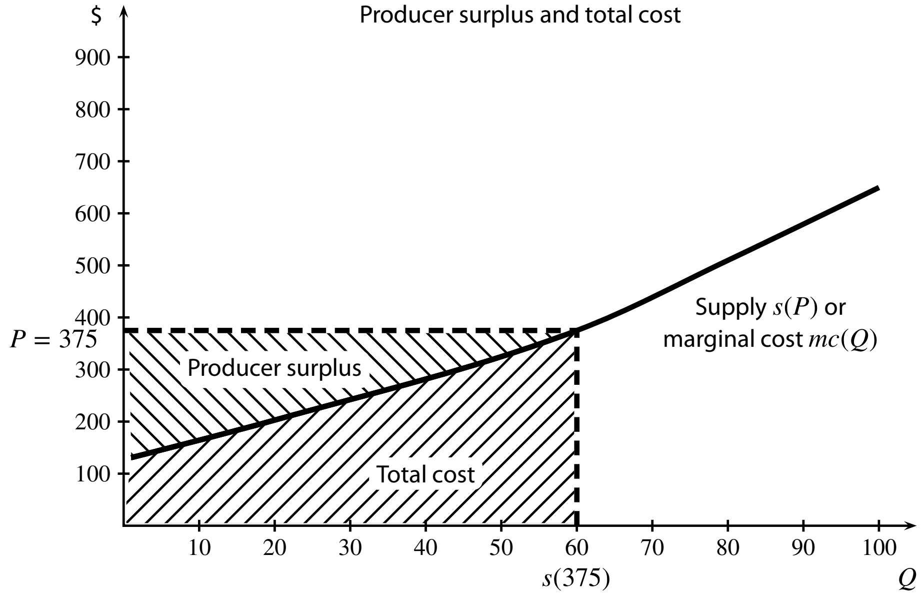


This graph also shows the division between expenditure and consumer surplus as follows.

1. The buyers' total expenditure, 60 × \$375, equals the area of the rectangle bounded by $P = 375$ and $Q = 60$.
2. The surplus for each buyer is the distance between his valuation and the price P. Hence, total consumer surplus is the area below the marginal valuation curve (i.e., the demand curve) and above the line $P = 375$.
3. Equivalently, consumer surplus equals total valuation (area under the marginal valuation curve) minus total expenditure.

We can illustrate the surplus of sellers when they sell at a price P. Figure 2.4 shows the marginal cost curve (i.e., the supply curve). Suppose the price is $P = 375$, and observe that $s(375) = 60$.

Figure 2.4



1. The sellers' total revenue, 60 × \$375, equals the area of the rectangle bounded by $P = 375$ and $Q = 60$.
2. The surplus for each seller is the distance between P and her cost. Producer surplus is thus the area below the line $P = 375$ and above the marginal cost or supply curve.
3. Equivalently, producer surplus equals total revenue minus total cost (the area under the marginal cost curve).

Exercise 2.1. Suppose a market for commercial water purification systems has 13 buyers with the following valuations (in €1000s), from highest to lowest:

Buyer	1	2	3	4	5	6	7	8	9	10	11	12	13
Valuation	57	54	51	48	45	42	39	36	33	30	27	24	21

These valuations are graphed in Figure E2.1 and again in Figure E2.2.

a. What is the total valuation of the first 5 buyers? Illustrate this as the area under the valuation curve in Figure E2.1.

b. Suppose the price is 43. How much is the demand? How much is the total expenditure? How much is the surplus of each of the buyers? How much is the total consumer surplus? Illustrate the demand, the total expenditure, and the consumer surplus using the valuation curve in Figure E2.2.

Figure E2.1

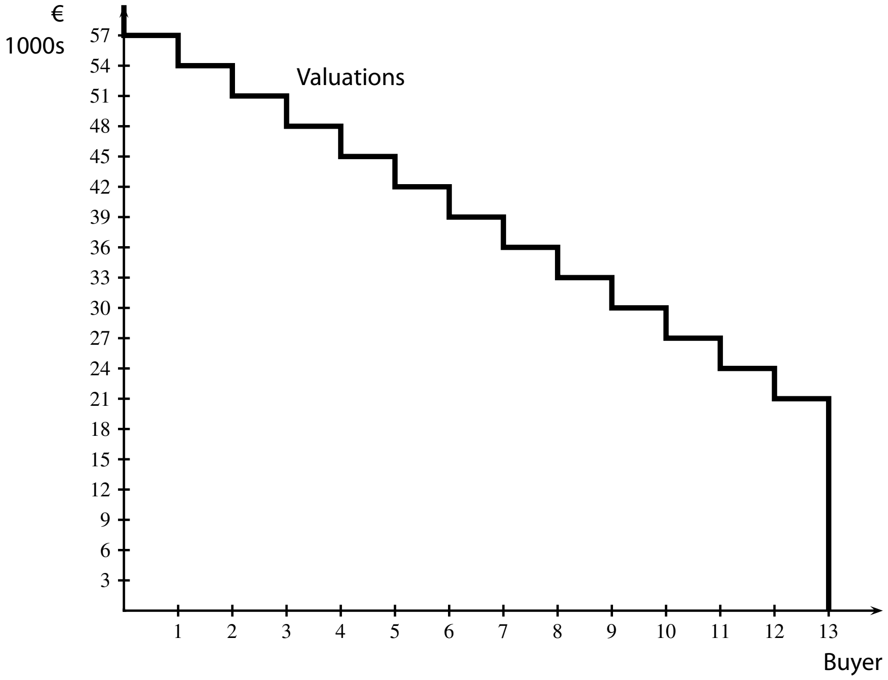
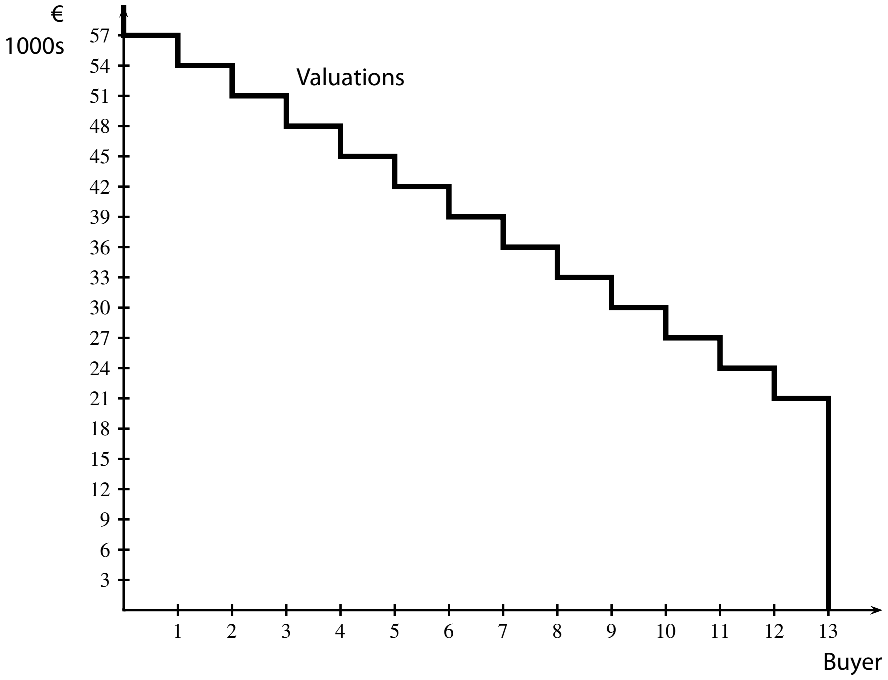


Figure E2.2



2.6 Gains from trade in equilibrium

The equilibrium condition $d(P^*) = s(P^*)$ means that graphs of the demand curve and the supply curve intersect. Continuing with the same demand and supply curves, we have Figure 2.5. The equilibrium price is about $P^* = 375$, and $Q^* = 60$ units are traded at this price.

Figure 2.5

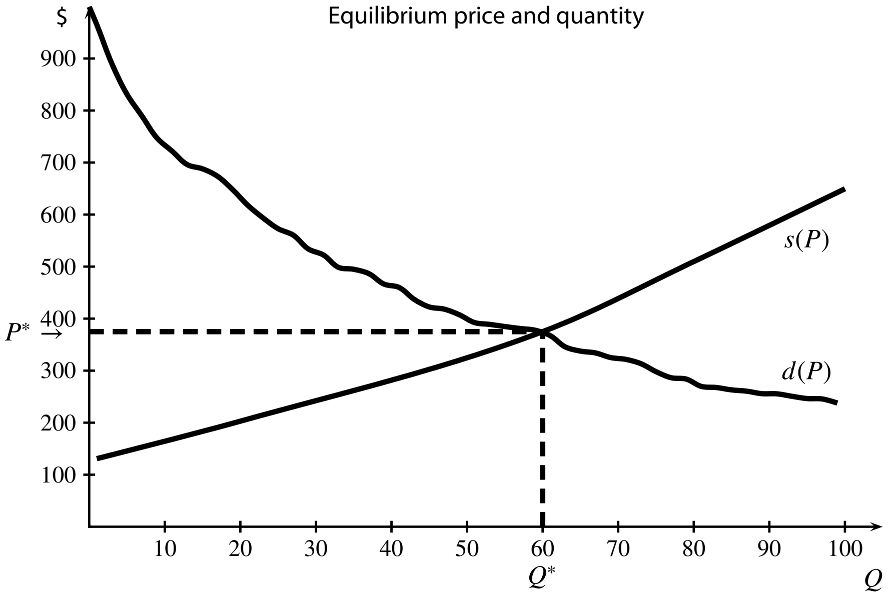
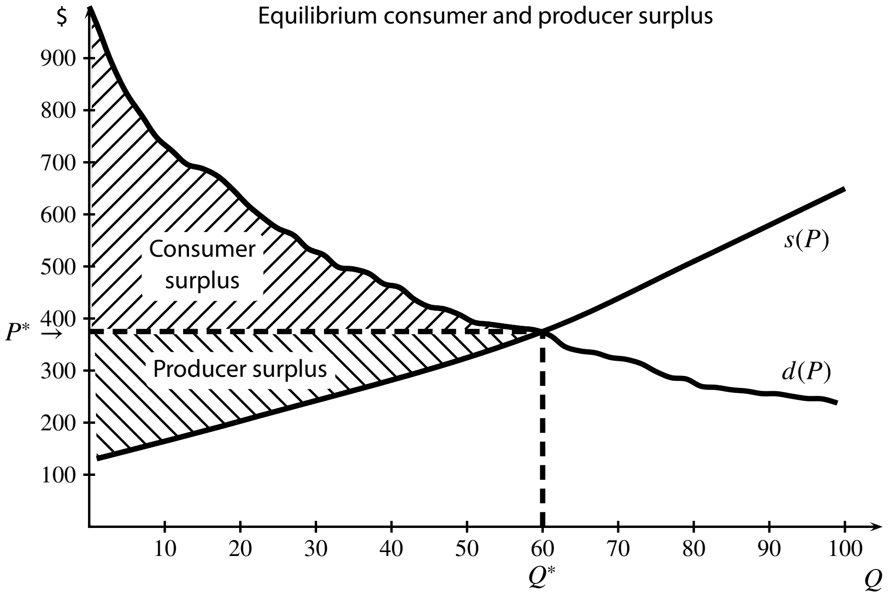


Figure 2.6 illustrates the total gains from trade, divided between the consumer surplus and the producer surplus.

Figure 2.6



The market in our model serves to reallocate the good and money among the traders. Initially, there are buyers who value the good more than it is worth to some of the sellers; they can make a trade that leaves both better off but does not affect any of the other traders.

Is the market equilibrium efficient? Yes. Recall that the efficient quantity is where the marginal valuation and marginal cost curves intersect. The equilibrium quantity is where the demand and supply curves intersect. These two intersection points are the same because the curves are the same, that is, because the demand curve is the inverse of the marginal valuation curve and the supply curve is the inverse of the marginal cost curve.

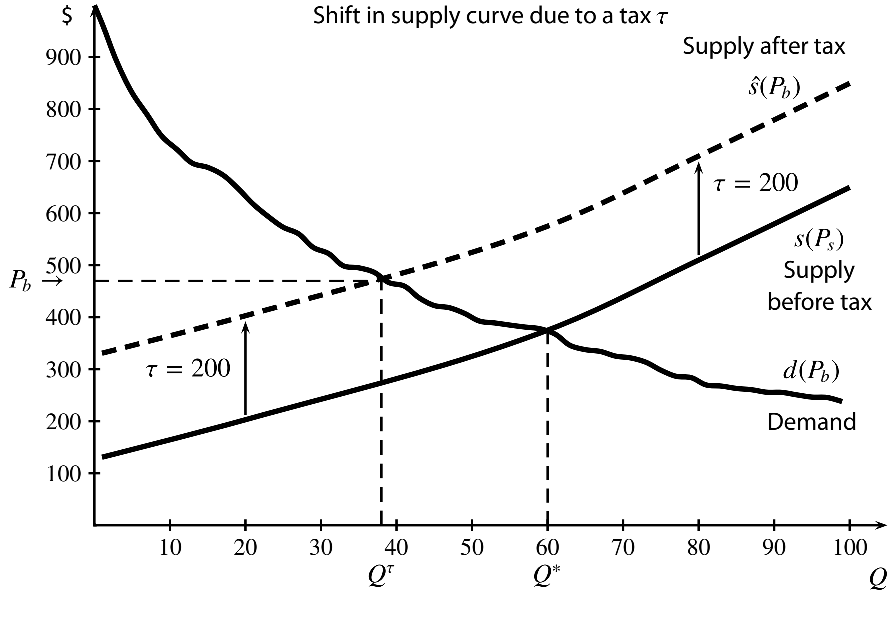
2.7 Taxes on transactions

Equilibrium with a sales tax

Suppose the government imposes a tax τ on each sale. (We denote a tax by τ, the Greek letter "tau".) It does not matter whether buyers or sellers are responsible for handing the tax to the government and whether prices are quoted with or without the tax. Either way, we can identify the effective price P_b that buyers pay and the effective price P_s that sellers keep. The difference $P_b - P_s = \tau$ is the tax. Such prices are an equilibrium if $d(P_b) = s(P_s)$.

To calculate the equilibrium, whether graphically or by solving an equation, it is easiest to work with just one price. Suppose we work with the buyers' price P_b—a posted price that includes the tax. Let $\hat{s}(P_b)$ be supply as a function of this price. Since the sellers receive $P_b - \tau$, we see that $\hat{s}(P_b) = s(P_b - \tau)$. Graphically, the curve $\hat{s}(P_b)$ is obtained by shifting the supply curve $s(P_s)$ *upward* by τ, as seen in Figure 2.7. The intersection of $\hat{s}(P_b)$ and $d(P_b)$ gives the equilibrium value of P_b.

Figure 2.7



We can alternatively work with the sellers' price P_s—a posted price that does not include the tax. We thus shift the demand curve *downward* by τ (as you will do yourself in Exercise 2.2). These two approaches yield the same equilibrium values of P_b, P_s, and Q.

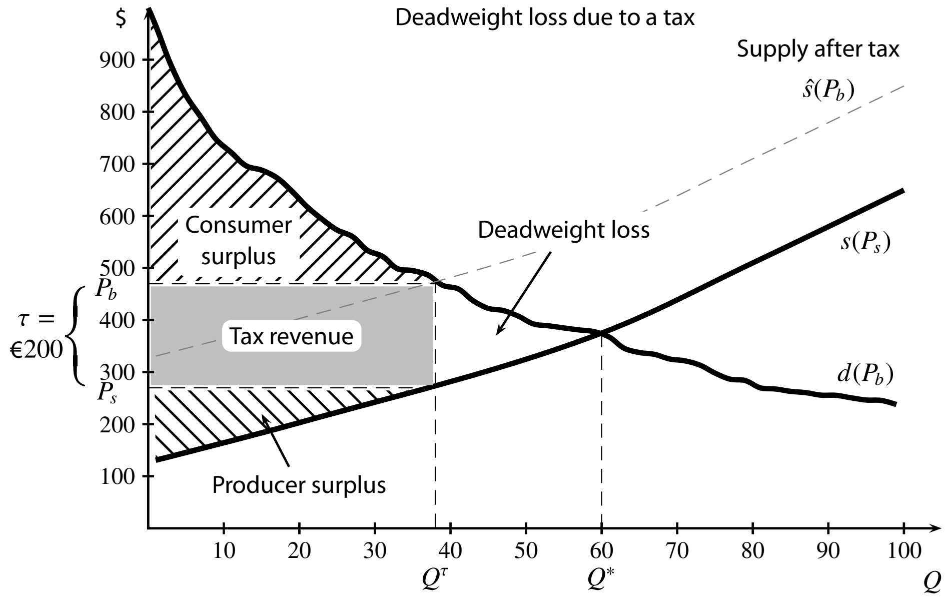
The following three examples show the irrelevance of who must give the tax to the government and of whether prices are quoted with or without tax.

1. Equilibrium wages (take-home pay for workers and costs for employers) and employment levels are the same whether social security taxes are paid by the workers, paid by the firms, or split.
2. France has a 6.5% tax on real-estate transactions that is paid by the buyer; the 5%–6% real-estate agent fee is customarily included in the price quoted by the seller. The equilibrium cost of housing to a buyer and sales revenue for a seller would be the same if the seller paid the tax (adjusted so that it remains a percentage of the before-tax transaction price) or if the buyer paid the real-estate fee.
3. In most countries, any value-added tax or sales tax at a retail store is included in the posted price (as if paid by the store). However, in the United States, it is added at the cash register (as if paid by the consumer). Such conventions affect neither the equilibrium price paid by consumers (including taxes) nor the equilibrium price received by retailers (net of taxes). The posted prices differ, of course, because one includes the tax and the other does not.

Deadweight loss

To illustrate the gains from trade, we need to use the marginal cost curve (the unshifted supply curve) and the marginal valuation curve (the unshifted demand curve). Figure 2.8 shows how the gains from trade are divided into consumer surplus, producer surplus, and tax revenue.

Figure 2.8



Compared to the equilibrium without the tax, the total gains from trade have fallen by the area of the triangle marked "Deadweight loss". The tax has caused the amount traded to fall to Q^τ from its efficient level Q^*, because a buyer and a seller execute a trade only if the gains from trade exceed the tax.

That taxes create a deadweight loss does not mean that there should be no taxes. Rather, it has two implications.

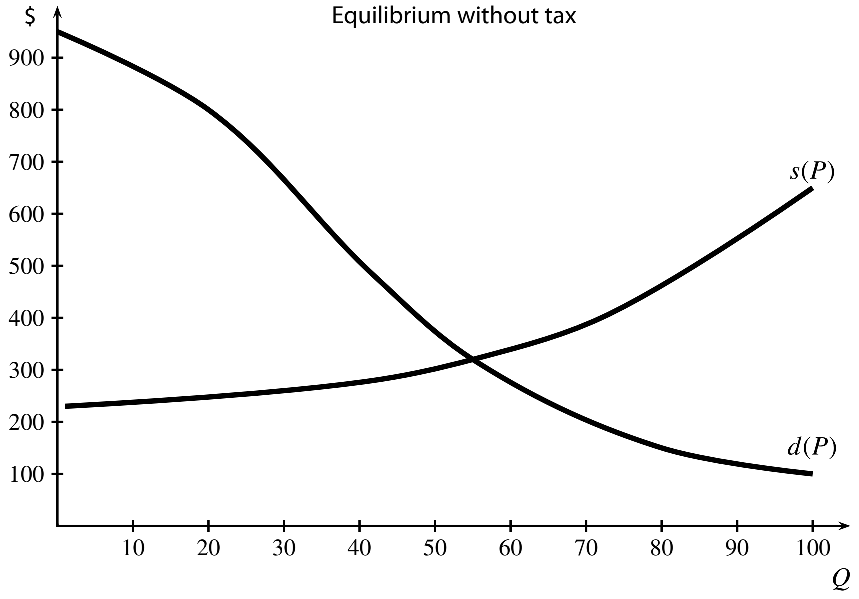
1. When deciding how much to spend, the government should weigh the benefits of the expenditures with their total cost, including the deadweight loss from taxation.
2. For a given amount of revenue that a government wants to raise, it should design the tax system to minimize the deadweight loss.[1]

[1] In this simple model, a better way to raise revenue is through *lump-sum* taxes: taxes that do not depend on trade and hence that do not distort trade. In practice, optimal tax policy is much more complicated. The government may want to tax the more fortunate in order to help the less fortunate. However, it cannot tell who is more fortunate except imperfectly: by observing peoples' income or expenditures. Hence the government uses a combination of progressive income taxes, wealth taxes, and sales taxes.

Exercise 2.2. This exercise asks you to show the effect of a tax by working with the seller's price P_s (not including the tax) and shifting the demand curve.

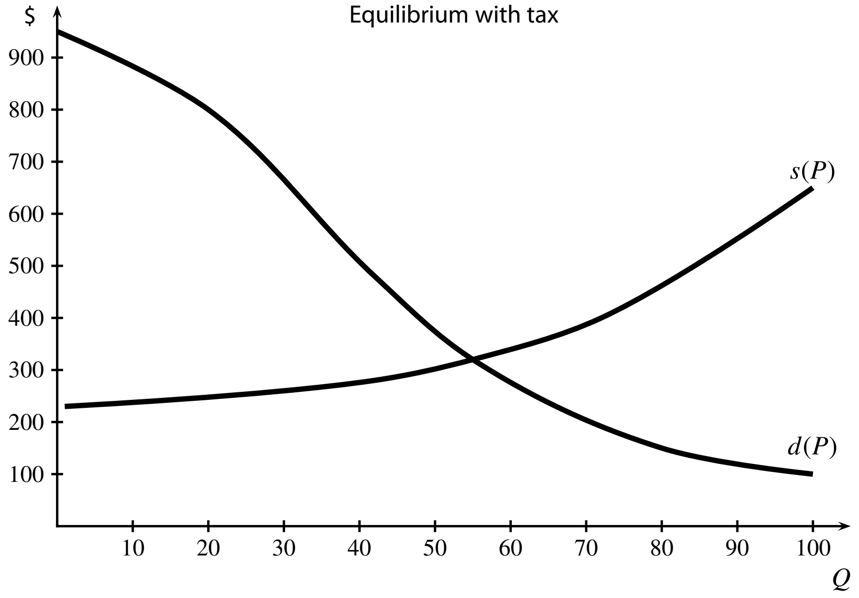
a. Figure E2.3 shows a demand and supply curve. Show the equilibrium price and quantity. Illustrate also the producer surplus and consumer surplus.

Figure E2.3



b. Suppose a tax of €200 is imposed. Let the vertical axis measure the price received by sellers. Redraw the demand curve on Figure E2.4 as a function of the seller's price. Illustrate the new equilibrium price and quantity as well as the deadweight loss.

Figure E2.4



2.8 Wrap-up

We have seen how people trade to make themselves better off. Even decentralized markets can achieve efficient allocations.

To summarize the main points:

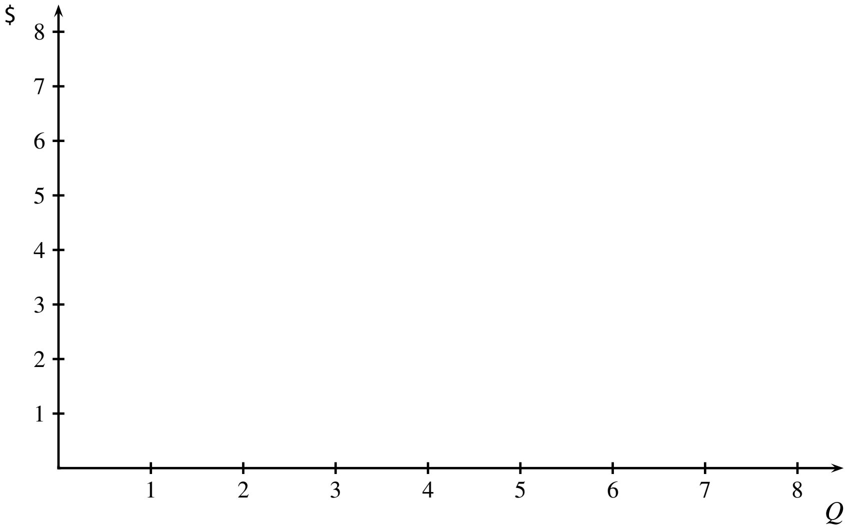
1. The demand curve (demand as a function of price) is the inverse of the marginal valuation curve (marginal valuation as a function of quantity).
2. The supply curve (supply as a function of price) is the inverse of the marginal cost curve (marginal cost as a function of quantity).
3. The equilibrium quantity and price are given by the intersection of the demand and supply curve.
4. The consumer surplus when the equilibrium price is P^* equals the area below the demand curve and above the horizontal line at P^*.
5. The producer surplus when the equilibrium price is P^* equals the area above the supply curve and below the horizontal line at P^*.
6. The equilibrium is efficient.
7. A tax on a good will cause the price that buyers pay to rise and the amount that sellers receive to fall. The quantity traded then declines and is not efficient. The loss in total surplus is called the deadweight loss.

Additional exercises

Exercise 2.3. Suppose there are 8 buyers with valuations 2, 3, 3, 4, 5, 5, 7, and 8.

a. Graph the valuations as points on the axes in Figure E2.5, from highest to lowest.

Figure E2.5



b. Suppose the market price is $4.50. How many buyers will purchase at this price? Connect the points on your graph so that it becomes a demand curve, and illustrate the quantity purchased when the price is $4.50.

c. What is the total valuation of the buyers who would purchase when the price is $4.5? Illustrate this as the area of a region on your graph.

d. What is the total expenditure by those who purchase? What is the total surplus? Illustrate these as the areas of regions on your graph.

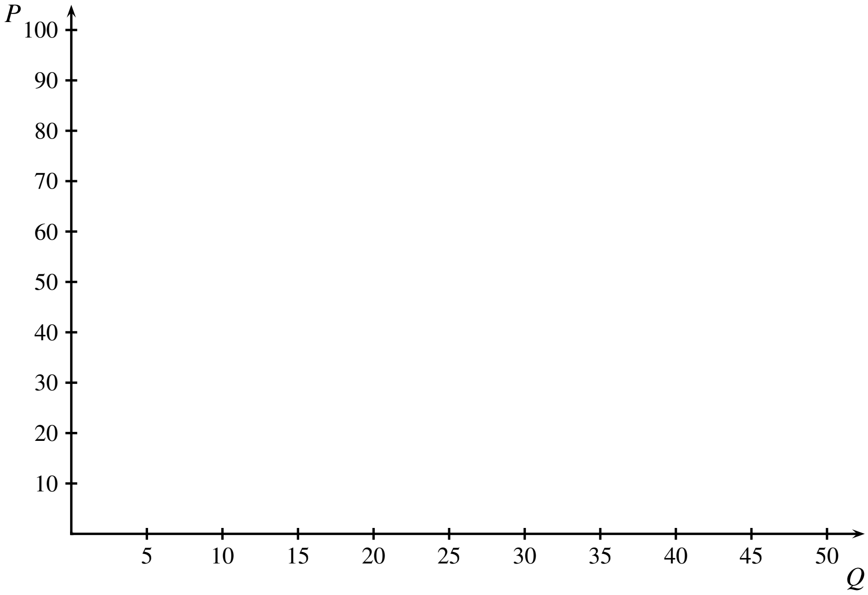
Exercise 2.4. Consider a market with the following demand and supply curves:

$$\text{Demand:} \quad d(P) = 50 - (1/2)P\,,$$
$$\text{Supply:} \quad s(P) = -10 + P\,.$$

a. Compute the equilibrium price and quantity.

b. Graph the demand and supply curves on Figure E2.6. Illustrate on the graph the equilibrium price and quantity, the consumer surplus, and the producer surplus.

Figure E2.6

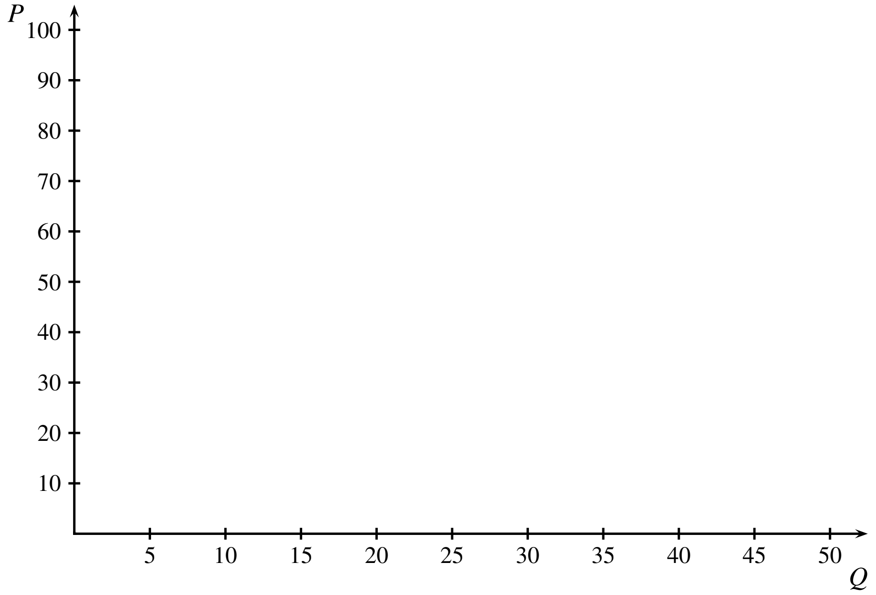


(*Note*: With such linear demand and supply curves, it would appear that demand or supply would be negative for some prices. Of course, this is impossible. If we were to be pedantic, we would write them as $d(P) = \max\{0, -10 + P\}$ and $s(P) = \max\{0, 50 - 2P\}$. This is unnecessary as long as we remember not to extend the graphs into the negative quantities.)

c. Calculate the consumer surplus and the producer surplus. [You are calculating the areas of two triangles. Remember: area of a triangle = (1/2)(base × height).]

Exercise 2.5. Consider the demand and supply curves from Exercise 2.4. Assume a per-unit tax of 15 is imposed. You are to find the price the buyers pay (including the tax), the price the sellers receive (net of the tax), and the quantity transacted. Some of the questions require that you graph curves within Figure E2.7.

Figure E2.7



a. You should find the new equilibrium values by shifting the demand curve. What is the demand $d^{\tau}(P_s)$ as a function of the price (net of tax) received by the sellers?

b. Graph this function and the original demand curve $d(P)$ on Figure E2.7. In which direction does the demand curve shift by 15?

c. Calculate the equilibrium price paid by buyers (including the tax), price received by sellers (net of tax), and quantity.

d. Illustrate the equilibrium on the graph. Show the regions that correspond to (a) producer surplus, (b) consumer surplus, (c) tax revenue, and (d) deadweight loss.

Chapter 3

Consumer Choice and Demand

3.1 Motives and objectives

Broadly

We study choices by consumers and properties of consumer demand. It is important to understand consumer behavior in order to set the price of your own product or to predict how market prices will evolve as a result of changing market conditions. Either way, you need information on how demand depends on prices and other variables. The analysis of how consumers make choices can help you predict how consumers will respond to price changes that have never before been observed. Alternatively, if the proposed prices are within a range of past prices, then you can estimate demand from past data.

More specifically

This chapter has three parts.

Interpretation of demand functions and demand curves. A demand function is a mathematical formula that relates demand for a good to its determinants, such as the price of the good, prices of other good, income of the consumers, and advertising expenditures. A demand curve describes demand merely as a function of the good's own price. We present some forms of demand functions, we categorize how prices and income affect demand, and we relate demand functions to demand curves.

Elasticity of demand. How responsive is demand to changes in price or other variables? We measure this in terms of percentage changes, using the concept of *elasticity*.

A model of consumer choice with one good. We then extend the model of consumer behavior developed in Chapters 1 and 2. In those chapters, each consumer purchases at most one unit of an indivisible good. In our extension, the consumer may choose any quantity of the good. We model consumer preferences via total valuation: how much a consumer is willing to pay for each amount of the product, as compared to not trading at all. We relate this to marginal valuation, to demand, and to consumer surplus. This model of consumer behavior and demand will be a workhorse for subsequent chapters on competitive, monopolistic, and oligopolistic markets.

3.2 Interpretation of demand functions and curves

Examples of demand functions

The demand for a good depends on many factors besides the price of the good, such as the prices of other goods, advertising expenditures, and seasonal variations. The relationship between demand and these factors is called a *demand function*.

To estimate demand functions, you can use econometric regression on either historical data (demand observed in the past) or data generated by consumer surveys, focus groups, or market experiments. Such estimation involves using a limited amount of data to predict demand in market situations (prices of goods, income, etc.) that have not been observed. To make best use of the limited data, we must start with only a few parameters to estimate. This means:

1. we should focus on only a few prices or other factors that affect demand; and
2. we should restrict attention to a simple parametric relationship between the variables.

The simplest relationship is linear:

$$Y = a_0 + a_1 X_1 + \cdots + a_n X_n,$$

where Y is the dependent variable, $X_1, \ldots, X_n$ are the independent variables, and $a_0, a_1, \ldots, a_n$ are the parameters we need to estimate. We can use the methods of linear regression (taught in statistics courses) to estimate these parameters.

Also simple are *exponential* functions:

$$Y = a_0 X_1^{a_1} \cdots X_n^{a_n}.$$

Again, $a_0, a_1, \ldots, a_n$ are the parameters we wish to estimate. This is not a linear equation, but we can turn it into one by taking the *logarithm*, which turns multiplication into addition and exponents into multiplication. Thus, we have

$$\log(Y) = \log(a_0) + a_1 \log(X_1) + \cdots + a_n \log(X_n). \tag{3.1}$$

Because there is a linear relationship between the logarithms of the variables, this functional form is also called *log-linear*. By treating the data as $\log(X_1), \ldots, \log(X_n)$ and $\log(Y)$, we can again use the standard methods of linear regression.

Classification of how price and income affect demand

The following terms classify the direction of change in demand for a good in response to a change in other variables. They can be applied to the demand of an individual consumer or the aggregate demand in a market. We group these terms by the variable that changes.

Own price. Because demand for a good nearly always goes down when its price goes up, there is no special term for this case.

Prices of other goods. If an increase in the price of good *B* causes the demand for good *A* to increase, then these goods are *substitutes*. If it causes the demand for good *A* to decrease, then they are *complements*.

Income. If an increase in income causes demand to fall, then the good is *inferior*. Otherwise, it is *normal*. If expenditure on the good increases at a faster rate than income (i.e., the percentage of income spent on the good rises as income rises), then the good is a *luxury* good. (Luxury goods are a subcategory of normal goods.)

For example, the competing video gaming consoles—Sony's Playstation2, Nintendo's Gamecube, and Microsoft's Xbox—are substitutes. On the other hand, video game hardware (a console) and video game software (a game title) are complements.

Different methods of transportation in a city are substitutes. In the city of London, for example, the London Underground "Tube" system, the public buses, and individual cars are substitutes. In February 2003, London began imposing a £5 congestion charge for any private vehicle entering the city center. Within six months, the number of Tube journeys went up by 17,000 per day—equivalent to an extra 8,000 passengers. The number of bus passengers increased by 15,000. Road congestion dropped by 30%. Encouraged by the success of the scheme, several big cities in and outside of the United Kingdom are considering replicating it.

Exercise 3.1. Consider the U.S. demand Q for minivans (measured in hundred thousands of units) as a function of the price P of minivans (measured in thousands of dollars), the price P_s of station wagons (measured in thousands of dollars), the price P_g of gasoline (measured in dollars), and per capita income I (measured in thousands of dollars). Suppose the demand function is linear, as follows:

$$Q = 12 - 0.6P + 0.2P_s - 3P_g + 0.2I. \tag{E3.1}$$

Based on the form of this demand function (instead of on your prior knowledge about the minivan market), answer the following questions.

a. Are minivans normal goods?

b. Are minivans and station wagons substitutes or complements?

c. Are minivans and gasoline substitutes or complements?

Demand as a function of a single variable

When working analytically with demand, it helps to focus on the relationship between demand and a single explanatory (independent) variable. To illustrate how we go from a multi-variable demand function to a single-variable function, let's start with the hypothetical demand function for minivans that was presented in Exercise 3.1:

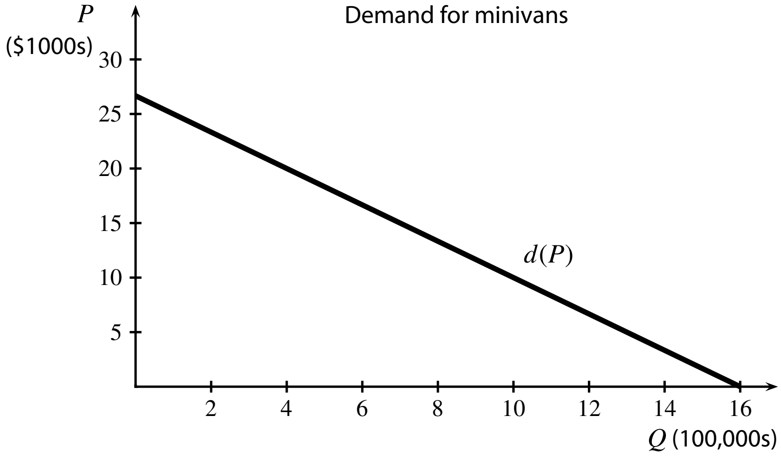
$$Q = 12 - 0.6P + 0.2P_s - 3P_g + 0.2I.$$

Suppose we want to focus on the relationship between demand for minivans and their own price P. This relationship depends on the values of the other variables. If the price of station wagons is \$15K (we use K to denote "thousand"), the gallon price of gasoline is \$1, and per capita income is \$20K, then we have

$$\begin{aligned} Q &= 12 - 0.6P + (0.2 \times 15) - (3 \times 1) + (0.2 \times 20) \\ &= 16 - 0.6P. \end{aligned}$$

The relationship between demand for a good and the good's price—keeping other variables fixed—is called the good's *demand curve*. We typically use the symbol $d(P)$ and graph demand curves with price on the vertical axis. Figure 3.1 shows the demand curve $Q = 16 - 0.6P$.

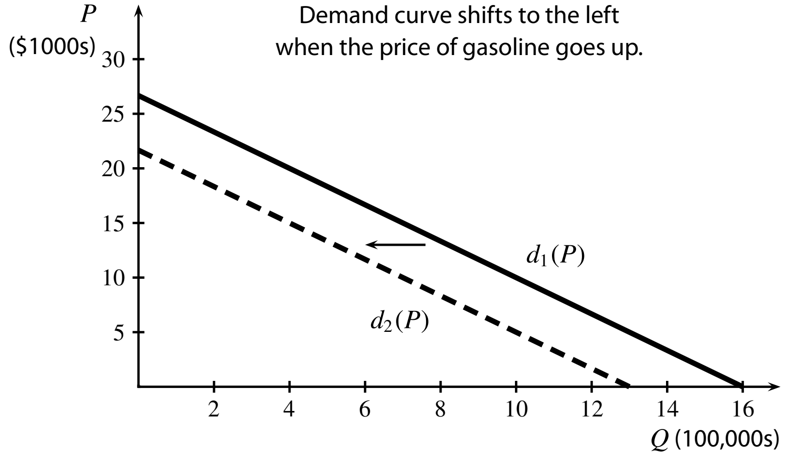
Figure 3.1



When working with demand curves instead of demand functions, we have to distinguish between a *movement along the demand curve*, meaning that demand changes because of a change in the good's own price, and a *shift in the demand curve*, meaning that the demand curve changes because of a change in some other variable that affects demand.

For example, if the price of gasoline rises from \$1 to \$2, then the demand curve becomes $Q = 13-0.6P$. In the graph, it shifts to the *left* by 3 units because 3 fewer units are consumed at each price. The new demand curve $d_2(P)$ is drawn as a dashed line in Figure 3.2.

Figure 3.2



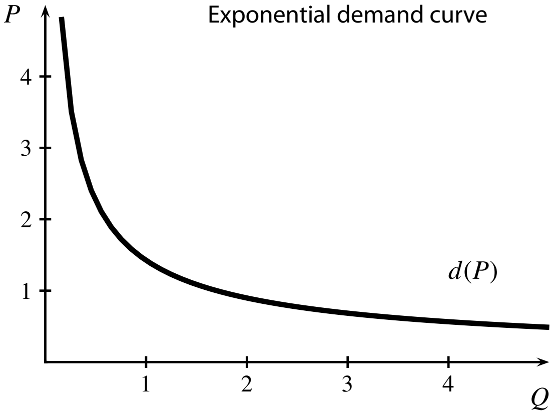
Exercise 3.2. Consider the demand function for minivans shown in Exercise 3.1. What is the demand curve when the price of station wagons is \$16K, the price of gasoline is \$3 per gallon, and per capita income is \$25K?

Linear and exponential demand curves

The general form of a linear demand curve is $Q = A - BP$, where A and B are positive numbers. Demand is zero at the price $\bar{P} = A/B$ (the vertical intercept of the demand curve as drawn with P on the vertical axis). We call $\bar{P}$ the *choke price*. For example, Figure 3.1 shows the demand curve $Q = 16 - 0.6P$. The choke price is $16/0.6 = 26.67$.

We write an exponential demand curve as $Q = AP^{-B}$, where A and B are positive numbers. The log-linear form is $\log Q = \log(A) - B\log(P)$. Figure 3.3 shows the demand curve $Q = 1.7P^{-1.5}$.

Figure 3.3



3.3 Elasticity of demand

Overview

Here are some cases in which sensitivity of demand is important.

1. For a firm with market power, its optimal price depends on the price sensitivity of the demand for its good. For example, the benefit of raising its price depends on how abruptly demand would fall.
2. The impact of a tax on the competitive equilibrium price—and also the tax revenue generated and the deadweight loss—depend on the price sensitivity of demand and supply.
3. If a supply curve shifts, perhaps because of change in technology, the effect on the price depends on the price sensitivity of demand.

In particular, price sensitivity will be an important concept during our study of pricing by imperfectly competitive firms.

You might expect to measure price sensitivity by the slope of a demand curve. But for many applications, a more useful measure of the sensitivity of a dependent variable Y to an independent variable X is *elasticity*: the *percentage* change in Y divided by the *percentage* change in X. This is true for all the applications of price sensitivity just described.

This will become evident as the book progresses, but here is a little teaser from Chapter 8. The following statement is intuitive: "Consider a firm that can segment its market. The firm should charge a higher price in the market segment in which demand is less price sensitive." We will see that this statement is correct if we measure price sensitivity by elasticity, while it is false if we measure it by slope.

Thus, we measure the responsiveness of demand to changes in a good's own price by the *(own-price) elasticity of demand*, which we denote by E:

$$E = -\frac{\%\text{ change in }Q}{\%\text{ change in }P}. \tag{3.2}$$

Because the change in demand goes in the opposite direction of the change in price, we have inserted a negative sign to make sure that elasticity is a positive number. This convention, which is common practice within the discourses of economists, makes it much easier to discuss elasticity because a higher value means more responsive demand. However, the formal definition of elasticity does not have this sign change, so you should not be surprised to see own-price elasticity as a negative number in some articles. (Then, if elasticity has gone *up* from -3 to -2, demand has become *less* elastic!)

Elasticity is "unit free". A percentage change is the same whether we measure quantity in liters or gallons or whether we measure price in euros or bahts, whereas any such variation in units changes the numerical value of a demand curve's slope.

Discrete changes: Arc elasticity

Elasticity varies along a demand curve; that is, the responsiveness of demand to a change in price depends on the initial price that is being charged. Furthermore, quantitatively it measures responses to small changes in price. We therefore say that elasticity is a *local* property of demand. This introduces a few technical issues when translating the intuitive equation (3.2) into specific formulas.

Consider first the elasticity that we measure based on a discrete change in the price level. Suppose that the price changes from P_1 to P_2 and that, as a result, demand changes from Q_1 to Q_2 as a result. We end up with different numbers for the elasticity depending on whether we measure changes as percentages of the initial or of the final values. Which price level should we treat as the status quo?

We give the two points equal status by measuring changes as percentages of the averages of the initial and final values. This yields the formula

$$E = -\left. \frac{Q_2 - Q_1}{\frac{1}{2}(Q_1 + Q_2)} \middle/ \frac{P_2 - P_1}{\frac{1}{2}(P_1 + P_2)} \right.,$$

which is called the *arc elasticity* between the points (P_1, Q_1) and (P_2, Q_2) on the demand curve.

Example 3.1 Let the demand function be

$$Q = 16 - 0.6P.$$

Now suppose the price is initially \$20K and hence demand is 4. Suppose the price increases to \$20.2K and hence demand falls to 3.88. Then the arc elasticity is

$$\left. \frac{0.12}{\frac{1}{2}(4 + 3.88)} \middle/ \frac{0.2}{\frac{1}{2}(20 + 20.2)} \right. = 3.06.$$

Smooth changes: Point elasticity

If the demand function is smooth, then we have a nice definition of the elasticity at a point on the demand curve:

$$E = -\frac{dQ}{dP}\frac{P}{Q}.$$

This is called the *point elasticity*. The slope dQ/dP is "normalized" by multiplying by P/Q (and by -1 so that it is positive). The point elasticity is approximately equal to the arc elasticity for small price changes.

Point elasticity is a local measure: When you use it to predict a change in demand due to a discrete change in price, your answer will only be approximate—but with the percentage error going to zero as the size of the price change decreases.

Example 3.2 In Example 3.1, the demand curve is $Q = 16 - 0.6P$. Let's calculate the point elasticity at $P = \$20$ and $Q = 4$. The slope dQ/dP of the demand curve is -0.6. Hence, the point elasticity is

$$E = 0.6\,\frac{20}{4} = 3.$$

Elasticity with respect to other parameters

It is possible to measure how demand responds to changes in other parameters, such as income and prices of other goods. For example, the responsiveness to changes in the price P_s of another good is measured by

$$\frac{\%\text{ change in }Q}{\%\text{ change in }P_s}.$$

This is called the *cross-price elasticity of demand.* Cross-price elasticity is positive if the two goods are substitutes and is negative if the two goods are complements.

The responsiveness to changes in income is measured by

$$\frac{\%\text{ change in }Q}{\%\text{ change in }I}$$

and is called the *income elasticity of demand.* It is positive if the good is normal and it is negative if the good is inferior.

The elasticity with respect to changes in the good's own price is called the *own-price elasticity of demand* (when it is necessary to distinguish it from these other elasticities). In this text, we make use mainly of own-price elasticity of demand and so refer to it simply as the elasticity of demand.

3.4 Elasticity: Why 1 is a special value

Elasticity and expenditure

Own-price elasticity can range from 0 to ∞ (the symbol for infinity). We have the following terminology for the different values of the elasticity.

Table 3.1

We say demand is ...	if ...
perfectly inelastic	$E = 0$
inelastic	$E < 1$
unit-elastic	$E = 1$
elastic	$E > 1$
perfectly elastic	$E = \infty$

Why is $E = 1$ the right division between elastic and inelastic demand? Here is one reason. Suppose the price rises. What happens to the consumers' expenditure (and hence the firms' revenue)? Does it go up, go down, or stay the same?

1. Consumers pay more per unit. This causes expenditure to go up—in percentage terms, by the percentage increase in price.
2. Consumers buy less. This causes expenditure to go down—in percentage terms, by the percentage decrease in demand.

Which effect dominates? It depends on whether the percentage decrease in demand is larger than the percentage increase in price.

- *Inelastic demand:* $E < 1$. The percentage decrease in demand is *smaller* than the percentage increase in price. Expenditure rises.
- *Elastic demand:* $E > 1$. The percentage decrease in demand is *larger* than the percentage increase in price. Expenditure falls.
- *Unit-elastic demand:* $E = 1$. The percentage decrease in demand just offsets the percentage increase in price. Expenditure stays roughly the same.

In Example 3.1, the elasticity is 3.06, meaning that demand is elastic. When the price rose from \$20K to \$20.2K, expenditure fell (from $20 \times 4 = 80$ to $20.2 \times 3.88 = 78.38$), as predicted.

Worldwide demand for coffee is inelastic (as it is for many commodities). The inflation-adjusted price of coffee fell by half from 1998 to 2002 owing to increases in supply from Vietnam (and some other new producers) and Brazil (which expanded production in frost-free zones). (Vietnam's exports grew from 78 million kilos in 1991 to 418 million kilos in 1998 and then 896 million kilos in 2000; since 1999, it has displaced Colombia as the world's second-largest exporter of coffee.) The Association of Coffee Producing Countries tried to get producing countries to retain 20% of their production. The UN Food and Agricultural Organization estimated that the elasticity of demand for coffee was about 0.6 (inelastic), so that such a program would have raised the price of coffee by about 32% and resulted in an increase in revenue of 5.5%.

(However, as is common with such cartels, getting countries to participate is subject to a "free rider" problem. Only Brazil, Colombia, Costa Rica, and Vietnam pledged to participate. Their reductions would have raised the price by only 17%. Although total revenue would still have risen, the non-participating countries would have profited at the expense of the participating countries. Brazil, Colombia, Costa Rica and Vietnam would have seen their revenues fall by 6.5%, whereas the revenues of the other countries would have risen by 17%. Eventually, the program was abandoned.)

Income elasticity and luxury goods

Recall that a good is a luxury good if the fraction of income spent on the good rises when income rises. This happens if, following a rise in income, the percentage increase in demand is larger than the percentage increase in income—that is, if the income elasticity is greater than 1.

In European countries, where bicycles are often used for leisure, the income elasticity of the demand for bicycles is greater than 1. In China, on the other hand, bicycles are an inferior good (negative income elasticity) because wealthier people swap their bicycles for cars.

3.5 Elasticity of special demand curves

Elasticity of linear demand curves

The slope of a linear demand curve $Q = A - BP$ is $dQ/dP = -B$. Thus, point elasticity when the price is P equals

$$E = -\frac{dQ}{dP}\frac{P}{Q} = B\frac{P}{A - BP} = \frac{P}{A/B - P} = \frac{P}{\bar{P} - P}.$$

Hence, elasticity at a given price depends on the parameters A and B of the demand function only through their ratio A/B, which is the choke price $\bar{P}$.

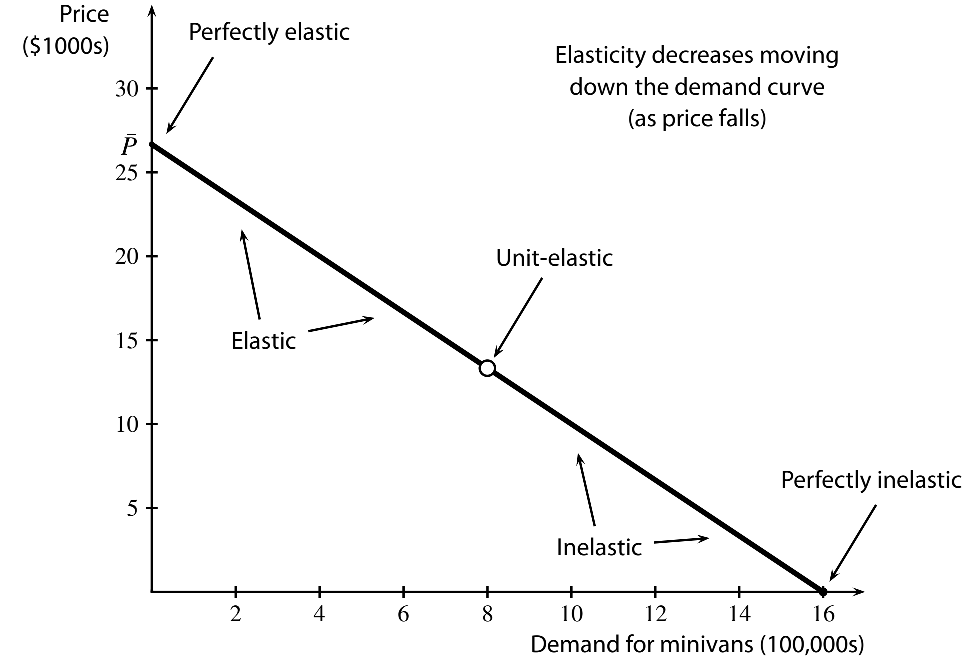
Observe that demand becomes less elastic as the price falls. In fact, if the price is close to zero then the elasticity is close to zero, whereas elasticity of demand increases without bound as P approaches the choke price $\bar{P}$. Demand has unit elasticity when $P/(\bar{P} - P) = 1$, that is, when $P = \bar{P}/2$.

Example 3.3 Recall our linear demand curve for minivans:

$$Q = 16 - 0.6P.$$

The choke price is $\bar{P} = 16/0.6 = 26.67$, so demand has unit elasticity when $P = 13.33$. Figure 3.4 shows the demand curve and labels the regions of elastic, unit-elastic, and inelastic demand.

Figure 3.4



Exercise 3.3. Market research revealed that the market demand function for home exercise equipment is

$$Q = 2400 - 2P - 15P_v,$$

where P is the price of exercise equipment and P_v is the price of exercise videos. The current price of exercise equipment is 300 and the current price of exercise videos is 20.

a. Given these prices, calculate the own-price elasticity of demand for exercise equipment.

b. Are exercise videos and exercise equipment complements or substitutes?

c. Suppose the price of exercise videos increases to 40. Does the own-price elasticity of demand increase or decrease?

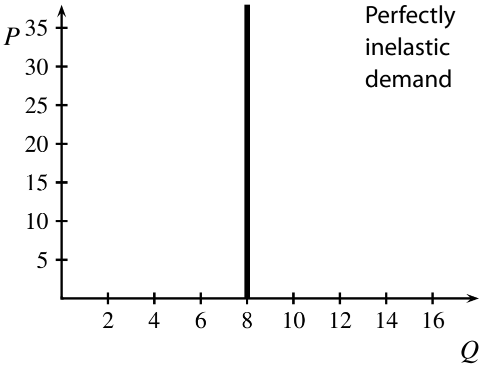
Constant-elasticity demand curves

For an exponential demand curve $Q = AP^{-B}$, elasticity equals B everywhere on the demand curve. Hence, a third name for exponential or log-linear demand curves is *constant-elasticity* demand curves.

Perfectly inelastic and elastic demand

Sometimes demand is extremely inelastic. We can approximate such inelastic demand by the limiting case of a *perfectly inelastic* demand curve: demand is the same no matter what price is charged, so the demand curve is a vertical line. Figure 3.5 shows an example.

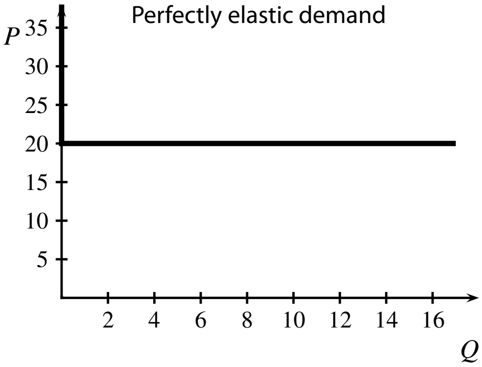
Figure 3.5



A 1996 study of the demand for outpatient services in Japan measured elasticities for different categories of service, with demand measured as a function of the patients' out-of-pocket expenses.[1] Elasticities ranged from 0.12 to 0.54 for most categories. Demand for outpatient services is typically not perfectly inelastic because many of the services are elective and are not required for survival, or a person can postpone a visit hoping that an illness subsides on its own. However, for genitourinary disorders the elasticity was not statistically distinguishable from zero. For this category, the main service provided was kidney dialysis, which patients need at fixed intervals in order to survive.

At the other extreme, demand may be very elastic. There is a price P such that (a) when the price is a little higher than P, demand drops off quickly; and (b) when it is little below P, demand rises quickly. It can be useful to approximate such elastic demand by the limiting case in which demand is zero when the price is higher than P, is infinite when the price is below P, and is any amount when the price equals P. The demand curve is a horizontal line at P. Such a demand curve is *perfectly elastic*; Figure 3.6 shows an example.

Figure 3.6



[1] J. Bhattacharya, W.B. Vogt, A. Yoshikawa, and T. Nakahara, 1996, "The Utilization of Outpatient Medical Services in Japan." *Journal of Human Resources*, 31:450–476.

3.6 The reality of estimation of demand

Linear demand versus constant-elasticity demand

Linear demand functions and exponential demand functions are the simplest classes of demand functions. Neither is a true representation of real-world demand functions (which are too complicated to work with or measure exactly), but each is a useful approximation. One might say that linear functions are too straight and exponential functions are too curved, with the curvature of real-world demand functions lying between the two.

We can now see for what purposes each form is useful.

Linear demand: for simple graphical and algebraic illustrations. A linear demand curve is easy to draw and work with. Furthermore, it has the property that demand becomes more elastic moving up the demand curve, which holds in the real world and is important for certain qualitative conclusions pursued in this book.

Exponential demand: for empirical estimation. One cannot accurately estimate an entire demand curve—instead, the goal is to obtain a good local estimate in the region of the data used for the estimation. Furthermore, one is typically interested in estimating the elasticity of demand (you will see why in subsequent applications of elasticity). This is best done using an exponential demand function in its log-linear form. When you estimate the linear regression equation

$$\log(Q) = \log(A) - B\log(P),$$

the coefficient B is simply the elasticity.

Any such estimation is an approximation, because (a) there are other variables that affect demand and (b) the relationship between the included variables and demand is not perfectly linear or log-linear (or whatever functional form is used). These effects are picked up in the "error term" of the regression. Adding more variables or more parameters to the functional form might seem to give a more accurate description of the market, but it will weaken our empirical estimates of the coefficients for the other variables.[2]

Interpreting the results of demand estimation

In fact, nearly all estimates of demand functions use the log-linear form or some variant thereof. Researchers typically report only the elasticities. It can be surprising to see no description of the units by which quantities and prices are measured, but these are not reported because elasticities are unit-free.

[2] In regression theory, we say that there is a loss of "degrees of freedom". The goodness of fit (R^2) goes up as we add more explanatory variables, but the accuracy of the estimates of the coefficients for the explanatory variables eventually goes down.

For example, a 2005 econometric study[3] of the demand for tobacco and other addictive goods in India summarized its results for rural India as shown in Table 3.2.

Table 3.2

Own- and cross-price elasticities

Item	Bidi	Cig	Tleaf	Pan	Alcohol
Bidi	−0.997	−0.100	−0.010	−0.026	0.023
Cigarette	−0.187	−0.626	−0.018	0.010	0.150
Tleaf	−0.093	0.212	−0.848	−0.129	−0.030
Pan	−0.075	−0.021	−0.010	−0.600	−0.023
Alcohol	−0.258	0.114	−0.022	0.084	−1.032

Bidi is made by rolling a piece of temburini leaf around flaked tobacco into a cone shape. Pan is a composite of betel leaf, areca nut, slaked lime, catechu, and tobacco. Tleaf stands for unprocessed leaf tobacco.

The own-price elasticities are all negative. This is the way own-price elasticities are reported in written documents. We have adopted a convention in this textbook to give own-price elasticities as magnitudes (positive numbers), which is common in discourse. Hence, we would say that the own-price elasticity of demand for Bidi is 0.997; it is close to unit-elastic. If a tax on bidi is imposed, consumption will fall but expenditure on bidi will remain roughly constant. The demand for cigarettes is less elastic, only 0.626. If the tax on cigarettes is raised, consumption will fall but people will still spend more on cigarettes.

The cross-price elasticities of the demand for bidi with respect to the price of cigarettes is negative: −0.100. Thus, these two goods are complements. This is common for some pairs of addictive goods. Unusually, however, these data show that cigarettes and alcohol are substitutes in India.

3.7 A model of consumer choice and welfare

We have analyzed the properties of demand functions. Now we study the consumer behavior that lies behind these demand functions. We limit ourselves to a simple but powerful model in which a consumer chooses quantities of a single good that depend on the price of that good. We keep other prices fixed. The purpose is twofold: (a) to enable evaluation of the gains from trade realized in markets; and (b) to have a model that will work for a firm's fancy pricing strategies, where a demand curve does not summarize what the firm needs to know about potential buyers.

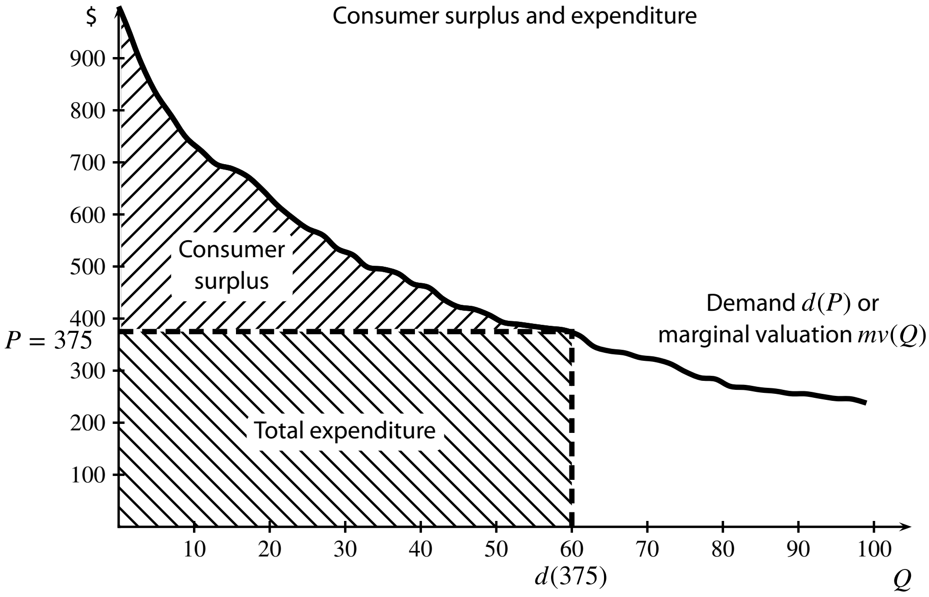
As you read this section, it should look similar to the model of valuation and demand presented in Chapters 1 and 2. Analytically, the following models are the same:

[3] "Price Elasticity Estimates for Tobacco and Other Addictive Goods in India", by Rijo M. John, Indira Gandhi Institute of Development Research.

(a) many consumers, each of whom buys one unit;
(b) one consumer, who may buy any number of units; and
(c) many consumers, each of whom can buy any number of units.

You have already studied (a), so the models (b) and (c) presented in this section contain little that is new other than their interpretation. In particular, the familiar picture presented in Figure 3.7 is valid for all three models.

Figure 3.7



This picture illustrates the following points.

1. The inverse of the demand curve is the marginal valuation curve of the consumer or consumers, and vice versa.
2. When graphed with price on the vertical axis, the area under the demand curve (i.e., under the marginal valuation curve) up to a quantity Q is equal to the total valuation of the consumer or consumers who receive the Q units.
3. Total consumer surplus at price P equals the area between the horizontal line at P and the demand curve.

A consumer's valuation and demand

For each quantity Q that the consumer might purchase, we can define the valuation $v(Q)$ for this quantity in the same way as we defined the valuation of a single unit. That is, $v(Q)$ is the maximum amount the consumer would pay for Q units if the alternative is to buy none of this good. The difference between his valuation for Q and the amount he spends on Q is his *consumer surplus*. It measures the consumer's gain from the trade compared with not trading at all.

We assume that the consumer ranks the possible quantities by the consumer surplus they generate. The consumer's decision is analogous to a firm's output decision. Whereas the firm chooses how much to produce and sell based on trade-offs between revenue and cost, the consumer chooses how much to demand based on trade-offs between valuation and expenditure.

Marginal valuation $mv(Q)$ is the extra amount the consumer would pay to have Q units instead of $Q-1$ units (or, in the smooth case, it is the extra amount per unit that the consumer would pay to increase consumption by a small amount). It is typically the case that marginal valuation is decreasing. The marginal condition for maximizing surplus is that the marginal valuation equal the marginal expenditure.

The marginal expenditure as consumption rises is just the price P of the good. Thus, the marginal condition is $mv(Q) = P$: the consumer buys up to the point that his marginal valuation equals the price. (If instead $MV > P$, he can increase his surplus by consuming a little more; if $MV < P$, he can increase his surplus by consuming a little less.) Therefore, to find the demand curve, we solve the equation $mv(Q) = P$ for Q. This means that the demand curve is the inverse of the marginal valuation curve.

For example, suppose a consumer's valuation curve is $v(Q) = 6Q^{1/2}$. Then his marginal valuation is $mv(Q) = 3Q^{-1/2} = 3/Q^{1/2}$. We solve $mv(Q) = P$ for Q to obtain the demand curve:

$$\begin{aligned} 3/Q^{1/2} &= P\,, \\ 3/P &= Q^{1/2}\,, \\ Q &= 9/P^2\,. \end{aligned}$$

From individual demand to market demand

Consider now a market with many consumers. The demand at a price P is just the total demand of all consumers at this price. This aggregate or market demand curve can be used—in the same way as individual demand curves—to measure total valuation, marginal valuation, and consumer surplus as follows.

1. *The inverse of the aggregate demand curve is the marginal valuation curve of the consumers*: At a given price, all consumers choose a quantity that equates their marginal valuation to this price.
2. *The area under the aggregate demand curve up to a quantity Q is equal to the collective total valuation of the consumers*: Suppose we distribute Q efficiently among the consumers (so that their marginal valuations are equal) and suppose we then sum the consumers' valuations of their allocated amounts. This total valuation of Q is then equal to the area under the demand curve.
3. *Total consumer surplus is equal to the area between the horizontal line at P and the demand curve:* Total consumer surplus when the price is P is equal to total valuation (area under the demand curve, as noted previously) minus the total expenditure (area under the horizontal line at P up to its intersection with the demand curve).

3.8 Wrap-up

This chapter had three main objectives. First, we considered how to interpret demand functions and we categorized how prices and income affect demand. Second, we introduced elasticity, a measure of the sensitivity of demand. Third, we developed a model of consumer behavior in which the consumer decides how much of a single good to consume given a per-unit price.

Additional exercises

Exercise 3.4. What will be the effect (increase or decrease) of the following events on the demand for French wine? Be sure to distinguish between shifts of the demand curve and movements along the curve.

a. A decrease in the price of French wine.

b. A new study linking longevity with moderate amounts of red wine.

c. An increase in the price of California wine.

d. A severe drought in the wine-growing regions of France.

Exercise 3.5. Calculate the price elasticity at current prices in the following examples. If you do not have enough information, say so.

a. The firm's demand curve is $Q = 2000 - 5P$, and the firm's output is 500.

b. The firm's demand curve is $Q = 5P^{-1.55}$; the firm's price and output are unobserved.

Exercise 3.6. (Valuation) A newspaper poll in Columbus showed that two thirds of the voters rated an excellent school system as one of the city's important assets. However, in an election the voters turned down a school bond issue. Does this mean the poll was faulty or that voters are irrational?

Exercise 3.7. (Inferior goods) It has been observed that the amount consumed of the services of domestic servants declined in most Western countries during the first half of the 20th century, while per capita income was increasing. Does this mean that domestic servants are an inferior good?

Exercise 3.8. (Elasticity) The 23 January 1992 issue of *The Economist* stated that, owing to a wet spring, truffle production in France was expected to reach 16 tons—up from the previous year's production of 8 tons. The price of truffles, which reached $690/pound in the previous year, was expected to fall to approximately $290/pound this year.

a. Assuming that the demand function for truffles has not changed, what is the arc elasticity of demand for this price change?

b. Do you think that truffle producers are happy about the good truffle-growing weather?

Exercise 3.9. (Elasticity, shifts in demand) Table E3.1 shows actual data about the prices of Model T touring cars in different years and the sales volumes at those prices.

Table E3.1

Year	Retail price	Sales volume
1908	850	5,986
1909	950	12,292
1910	780	19,293
1911	690	40,402
1912	600	78,611
1913	550	182,809
1914	490	260,720
1915	440	355,276
1916	360	577,036

a. Assuming that these data represent points on a fixed demand curve, calculate the arc elasticity of demand by comparing the data (i) for the years 1910 and 1911 and (ii) for the years 1915 and 1916.

b. Give two reasons why we might not want to consider these data to be points on a fixed demand curve.

Chapter 4

Production and Costs

4.1 Motives and objectives

Broadly

This book focuses on the profit-maximizing decisions of single-product firms (or of a single product line within a firm). Profit equals revenue minus cost. Revenue depends on consumer decisions, which we studied in Chapter 3. In this chapter, we study cost, which depends on the production technology and input prices.

The *cost curve* $c(Q)$ measures the total cost of producing Q units for any Q. Behind a cost curve are decisions on how to produce each level of output at the lowest cost. We can decompose this decision problem into two stages as follows.

Engineering. For given levels of inputs, we need to design *efficient* production processes that maximize the total output. The production design problem is not the subject of this book (it is covered in courses on operations and in engineering disciplines). The outcome of this production design problem is summarized by a *production function*: it relates total output to each combination of inputs once your engineers and operations managers have made optimal use of those resources.

Choice of input mix. For any level of output level, we should choose the least expensive way to produce the output according to the production function and the input prices. The cost curve $c(Q)$ then relates the minimum cost of production to each output level. The input prices affect both the optimal choice of inputs and the minimum cost, so a shift in these prices causes the cost curve to shift.

We will not go into the details of these two stages, but it is useful to understand that they are sitting in the background as we explore the properties of the cost curve.

More specifically

We first explore more carefully what we mean by the cost curve. Then we study the different properties that a cost curve—and the associated marginal and average cost curves—may have. One important concept is economies of scale, which is related to market concentration and natural monopoly.

4.2 Short run and long run in production

Production is an ongoing process. At each moment, a firm produces a certain level of output using a certain production process and input mix. When deciding to change the output level, the minimum cost of production depends on how much time is available to adjust the production process and the input mix to their efficient values.

For example, suppose you manage an oil refinery that currently produces one million liters of gasoline per day. You are contemplating cutting production to half its current level in response to a decrease in the price of gasoline. To see whether this change increases your profit, you must ask yourself: "What is the minimum cost of producing only 500,000 liters of gasoline per day?" The answer depends on whether you intend to change the output in a week, in a month, or in a year, because each input requires a different amount of time to vary its usage.

1. You can modify the usage of electricity and water almost instantly.
2. You can change the use of crude oil within a week's notice, provided the refinery is not locked into long-term delivery contracts.
3. It takes months to fire or hire workers. You can reduce labor costs by attrition, gradual layoffs, and early retirement packages, but these take time to implement. It also takes time to find and train new workers.
4. You may need over a year to purchase and install (or dismantle and sell) major capital equipment.

Thus, if you are responding to a drop in gasoline prices that is expected to last a long time, then you should consider how much your cost will go down over the course of a year—after you have had a chance to adjust usage of all inputs. If instead you are responding to an anticipated two-week drop in gasoline prices, then you should consider only the cost savings that can be achieved during that period by reducing your use of electricity, water, crude oil, and other inputs whose usage can be modified quickly.

To capture these differences among inputs and planning horizons in a simple way, economists often distinguish between only two horizons: the long run and the short run. The *long run* is sufficiently long that the usage of all inputs can be varied and the firm can even shut down. The *short run* is sufficiently short that the usage of one or more important inputs cannot be significantly modified. These horizons are metaphors whose interpretation depends on the actual speed with which inputs can be adjusted within a given industry.

We briefly discuss the short-run horizon in Chapter 6. Otherwise, you should presume that all production decisions and costs in this book are long-run and you should not let short-run fixed inputs sneak into the picture. For example, in long-run production decisions by a power company, the company can change the number of production and can choose among (say) small gas-fired plants and large nuclear plants. (The long run is very long in this industry.) One way to think of long-run production decisions is to imagine a new company that starts production from scratch; it is completely unconstrained by installed capacity and other prior input decisions. This is a particularly useful view when considering, for example, production of music CDs. The initial recording cost is incurred once and that decision cannot later be reversed. As a long-run production decision, whether or not to record the CD must be decided *before* any production has started. (If we study the long-run production *after* the CD has been recorded, then the recording cost is sunk and we disregard it entirely.)

4.3 What to include in the costs

We measure cost differently from the way it appears in a firm's accounting books because economists and accountants treat opportunity costs and sunk costs differently, as described in this section. Our measure is called *economic cost* as compared to *accounting cost*. The profit so calculated is called *economic profit* as compared to *accounting profit*. (Since we consistently use economic cost and economic profit, we refer to these merely as "cost" and "profit" once we pass this section.)

Opportunity cost

Suppose a restaurant chain owns an old piece of property in downtown Paris, already depreciated on its accounting books, which it uses for one of its restaurants. The company wants to determine whether the restaurant is breaking even or should be shut down. The accountants tell them that the revenues exceed the costs (salaries, maintenance, heating). Does this mean that the restaurant should continue to be operated?

The accountants have told the company that the accounting profit is positive. However, missing from the calculation is the rent that the firm could obtain if it shut down the restaurant and leased the property. This is called the *opportunity cost* of the property. If the accounting profit does not exceed the opportunity cost, then the restaurant should be shut down. If the opportunity cost is figured into the cost when calculating the profit, then the heuristic "operate the restaurant if and only if the profit is positive" is restored.

The assets of a firm do not represent a financial outlay and do not figure as a cost in the accounting. The return those assets could obtain outside the firm is the opportunity cost of those assets. Typical opportunity costs are the value that an entrepreneur's time would have in other activities, the return on capital of equity holders, and the return that real estate owned by the firm would have if rented to others.

Thinking in terms of opportunity costs, and looking carefully for them, helps one to avoid forgetting the value that resources have in alternative uses. For example, firms that fail to take into account the opportunity cost of their assets may be subject to takeover. An outsider may realize that a firm has assets that could be put to better use elsewhere even though the firm is profitable with its current operations. The outsider might buy the firm and sell those assets.

Taking into account opportunity cost when calculating profit also avoids spurious variations in profit such as the following. Suppose that a firm uses capital that comes both from equity and from debt. The interest payments on the debt appear as a cost in the accounting books. Now suppose that the equity holders contribute more equity to pay off the debt. The interest payments disappear from the accounting cost and the firm's "profit" goes up. However, the overall return on capital has not changed.

In this book, we will often talk about a firm earning "zero profit". This would be quite bad if profit were calculated using accounting cost, for then the return on equity and on the entrepreneur's time would be zero. Such capital and human resources would be better used elsewhere and the firm should be shut down. However, zero economic profit means that these resources are getting a fair return equal to the best return they could get elsewhere; hence, there is no gain in shutting down the firm.

Opportunity cost is a concept that is useful for more than the measurement of a cost curve. For example, when calculating the net financial return of an MBA education, on the income side is the present discounted value of the increased salary over a student's lifetime. The cost side includes tuition and the opportunity cost of the student's time—that is, the wages lost while studying for the degree.

Sunk costs

Sunk costs are costs that have already been incurred and can never be recovered. It is unimportant whether the costs have already appeared on the accounting books; it is rather a question of whether they are affected by the available options. For example, if you are locked into a two-year lease that cannot be broken, then even future payments on this lease are sunk costs. If you have recently set up a new plant that has no salvage value, then the entire cost is sunk even though only a fraction of that cost will be amortized in the accounting books in each of the coming years.

Because sunk costs are immutable, they do not affect the comparison between any alternatives or between different output levels. Therefore, it is simplest to ignore them (i.e., not include them in the calculation of the cost curve).

Ignoring sunk costs also avoids accidentally letting them influence decisions ("fallacy of sunk costs"). For example, imagine you work in marketing for a pharmaceutical company. You are asked to explain your price for Endostatin. You reply: "We are trying to recover the research and development cost." No! This cost is sunk and is not affected by your marketing and pricing strategies; your goal is to maximize the revenue from sales minus the incremental production cost.

An improper treatment of sunk costs can lead you either to stick with projects that should be abandoned or to abandon projects that you should stick with. As an example of the latter, suppose that you open a restaurant that generates less income than you were earning in your salaried job or insufficient income to cover the cost of renovations. Then you might close the restaurant. However, if you cannot get your old job back (a sunk opportunity cost!) or if the renovations are particular to your restaurant and you can never recover their cost through resale, then you should ignore these costs. You should stay open as long as your revenue exceeds the costs that are not sunk.

The other type of error comes from a psychological need to justify erroneous past decisions. Imagine you purchase a nonrefundable airplane ticket for a personal trip and then discover that the weather will be horrible at your destination. If you had never purchased the ticket, then you would not go even if the airfare were free. However, you say that you will go anyway because you already bought the ticket and would feel terrible wasting the money you paid. No! The cost of the ticket is sunk, so your decision about whether to proceed with the trip should ignore that cost.

Likewise, suppose you decide to invest $500,000 to develop a chip set for DSL modems. After the development is finished, a competitor beats you to the market, softening demand for your own product. Should you bring your product to the market? You should proceed if and only if the future revenue exceeds the future cost of production. However, there is often a bias to proceed regardless because you do not want to admit that the original investment was a mistake.

Keep in mind that each cost at some time was not sunk. A question such as “Are R&D expenses a sunk cost?” is meaningless—you need to specify the context and time perspective of the decision problem. A drug’s R&D cost should not affect its pricing because it is sunk at the time the firm markets the medicine. However, when deciding whether to invest in the R&D of a new drug (before that R&D cost becomes sunk), a firm needs to anticipate whether post-R&D profit will exceed the R&D cost.

4.4 Economies of scale

Marginal and average cost

The cost curve $c(Q)$ is also called *total cost*. Various other curves derived from $c(Q)$ are useful for applications. The two most important ones are average and marginal cost.

$$\begin{aligned} \text{Average cost:} \quad & ac(Q) = c(Q)/Q; & \\ \text{Marginal cost:} \quad & mc(Q) = c(Q) - c(Q-1) & \text{(discrete case)}, \\ & mc(Q) = dC/dQ & \text{(smooth case)}. \end{aligned}$$

The marginal cost measures the extra cost of increasing output by one unit (or the extra savings of decreasing output by one unit). When we apply marginal analysis to output decisions, marginal cost will be the key cost measure.

However, the average cost curve is important because its shape determines industry concentration and the benefits of mergers. We use the terminology in Table 4.1 to categorize whether the average cost curve increases or decreases.

Table 4.1

We say that a firm has …	if $ac(Q)$ is …
no economies of scale	constant
diseconomies of scale	increasing
economies of scale	decreasing
U-shaped average cost	first decreasing, then increasing

Suppose that two firms produce the same good with the same technology and hence have the same average cost curve $ac(Q)$. Suppose that initially the firms produce quantities Q_1 and Q_2 at average costs $ac(Q_1)$ and $ac(Q_2)$. How would a horizontal merger between the firms affects the cost of production? The total output of the merged firm is $Q_1 + Q_2$ and hence the average cost is $ac(Q_1 + Q_2)$.

- *Diseconomies of scale.* If $ac(Q)$ is increasing then $ac(Q_1 + Q_2)$ is higher than $ac(Q_1)$ and $ac(Q_2)$. Hence, a merger would raise costs; the firms will not merge except perhaps to gain monopoly profits through increased industry concentration.
- *Economies of scale.* If $ac(Q)$ is decreasing then $ac(Q_1 + Q_2)$ is lower than $ac(Q_1)$ and $ac(Q_2)$. Thus, the merger reduces costs. Such efficiency gains lead the industry toward monopoly.

Antitrust authorities might block a merger when there are economies of scale, but in so doing they would be trading off the inefficiencies of fragmented production against the inefficiencies of market concentration.

Sources of economies and diseconomies of scale

Suppose two firms with access to the same technology merge. If there are economies of scale, then the merged firm produces the combined output more cheaply. This may happen for several reasons.

1. *Returns to specialization.* At a larger scale of production, the merged firm can divide tasks more finely and so allow workers to concentrate on specialized tasks.
2. *Indivisible inputs.* The merged firm may be able to adopt production processes that require large indivisible inputs and hence are efficient only at high output levels. For example, an electric generation company can switch from gas-fired to coal-fired and then to nuclear generation as it expands output.
3. *Elimination of duplicate fixed (setup) costs.* Suppose the technology requires a setup cost in order to start production. The two firms if independent would each incur this cost, whereas the merged firm would incur this cost only once. For example, selling a single copy of a software package entails a large development and maintenance cost. This cost is incurred separately by two firms who each develop a comparable software package, but is incurred only once if the firms merge and develop a single package.

It is more difficult to explain how there may be *diseconomies* of scale. How could the costs rise after a merger? As a worst case, could not the merged firm achieve the same costs by *replicating* the activities of the two independent firms? If this is correct and if any of the other gains to mergers just listed are present, then all industries are natural monopolies. Yet we observe many competitive markets that do not seem to be driven to consolidation.

The resolution to this paradox is that a merged firm cannot replicate the managerial processes of the independent firms without becoming a collection of independent firms (divesting). Hence, the limits to firm size may derive from differences in the organizational properties of one large firm compared with those of several small firms. One of these differences is that the scale of coordination and centralized decision making is greater within a single large firm—with its headquarters and tight bureaucratic procedures for coordinating the parts and making such common decisions as total output. (This is documented in the work of the business historian Alfred Chandler.) Such centralization could lead to the slower and otherwise poorer decision making that hinders large firms.

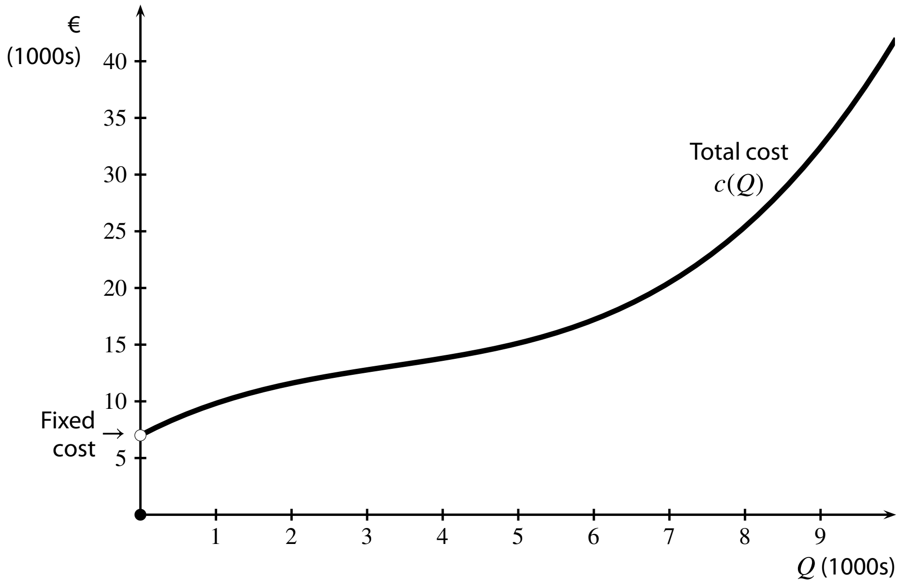
For example, Deutsche Telekom (DT) ran into trouble when its revenues stagnated in 1998. Until 1996, the government-owned company enjoyed a virtual monopoly in the German telecom market, but the scene changed in 1996 when DT was privatized and new players were allowed to compete with it on an equal basis. At the time of privatization, DT already had its operations in almost all segments of telecommunication (e.g., the Internet, mobile, and telephone networks) and also had a consultancy division. But owing to the managerial diseconomies that come with huge size, the company found itself mired in a debt whose servicing was having a negative effect on net profits. Deutsche Telekom had acquired many companies whose integration was a major issue. The company had to move quickly in order to counter the threat of new entrants, but its large size made this difficult to do. The share price started to drop and analysts began to write off the stock of DT.

In 1999, Deutsche Telekom underwent a restructuring that divided its business into four segments: T- Mobile, T-Com, T-Systems, and T-Online. This made the organization leaner and helped it adapt to the dynamic nature of the four individual segments. It also hived off some of its previous acquisitions. The result was a surge in revenue: whereas stagnant in 1998 and 1999, revenue grew by 8.6% and 11.7% in 2000 and 2001 and then by 30.9% and 24.5% in 2002 and 2003.

4.5 A typical cost curve

Main features

Figure 4.1

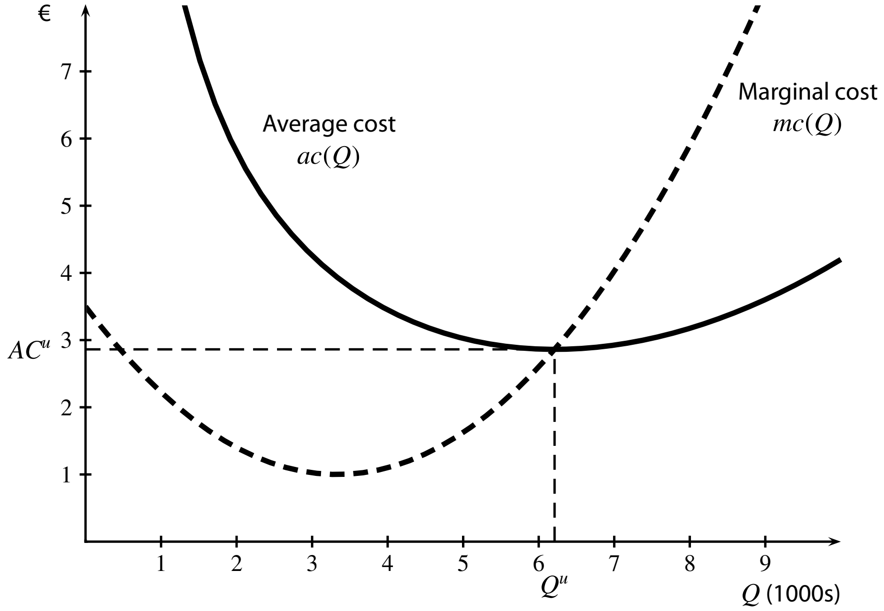


A typical real-world cost curve has the features illustrated in Figure 4.1. We can observe the following.

1. The cost of producing 0 units is 0. This point on the cost curve is the solid circle at the origin of the graph. In the long run, a firm can shut down or not start up.
2. The cost of starting production is large compared to the marginal cost of subsequent units. This is called a *fixed cost* because it does not depend on exactly how much is produced (only on whether or not the product is produced at all).
3. As production gets started, marginal cost decreases (the cost curve becomes flatter) because of returns to specialization and the ability to use large-scale production processes.
4. Eventually, marginal cost increases (the cost curve becomes steeper) as managerial diseconomies of scale kick in.

Marginal and average cost

Figure 4.2



As just described and as seen in Figure 4.2, the marginal cost curve is U-shaped (initially decreasing and then increasing). The average cost curve is also U-shaped.

- Average cost initially decreases because (a) the fixed cost is spread over more units of output and (b) marginal cost is decreasing. The average cost for low levels of output is extremely high because the entire fixed cost is spread over only a few units or perhaps a mere fraction of a unit; it then decreases quickly as the average cost is spread over more units.
- Average cost eventually rises because of the increasing marginal cost.

The quantity with the minimum average cost is called the *efficient scale of production*. This is denoted by Q^u in Figure 4.2, where the superscript u stands for the U-shape of the average cost curve. The minimum average cost is denoted by AC^u. Observe that the average and marginal cost curves intersect at this point. This is always true when the average cost curve is U-shaped. Here's why.

- Average cost decreases when output goes up by a unit if and only if the cost of the extra unit (the marginal cost) is lower than the average cost of all the preceding units, thus bringing the average cost down. Therefore, the MC curve must lie *below* the AC curve wherever the AC curve is *decreasing*.
- Likewise, the MC curve must lie *above* the AC curve wherever the AC curve is *increasing*.

Therefore, at the point where the AC curve shifts from decreasing to increasing (at its minimum), the MC curve must cross the AC curve—starting below and ending above it.

Variable cost

It can be useful to separate the fixed cost from the total cost, leaving the variable cost: $vc(Q) = c(Q) - FC$. The profit calculated using the variable cost (ignoring the fixed cost) is called the variable profit.

Graphically, the variable cost curve looks like the total cost curve except that it is shifted down by the fixed cost so that the graph starts at the origin. This is seen in Figure 4.3.

Figure 4.3

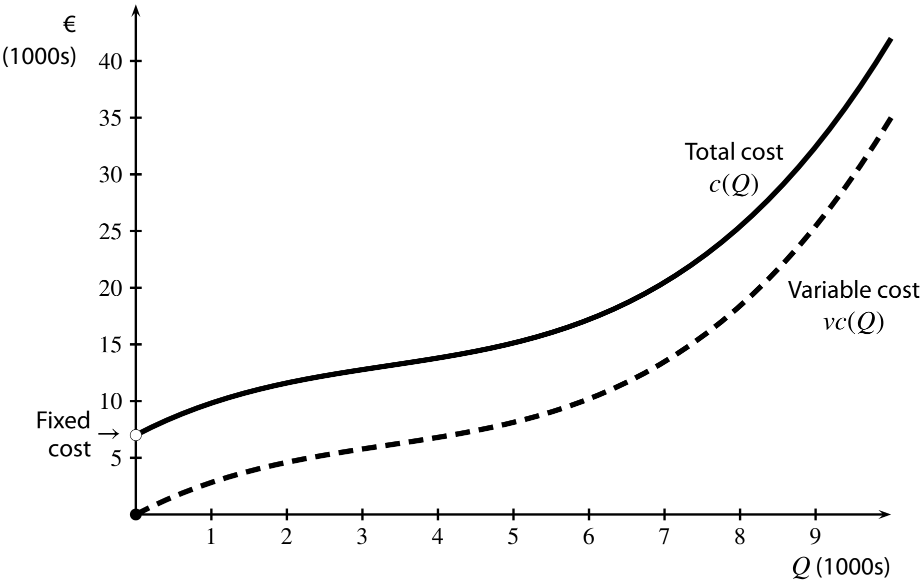
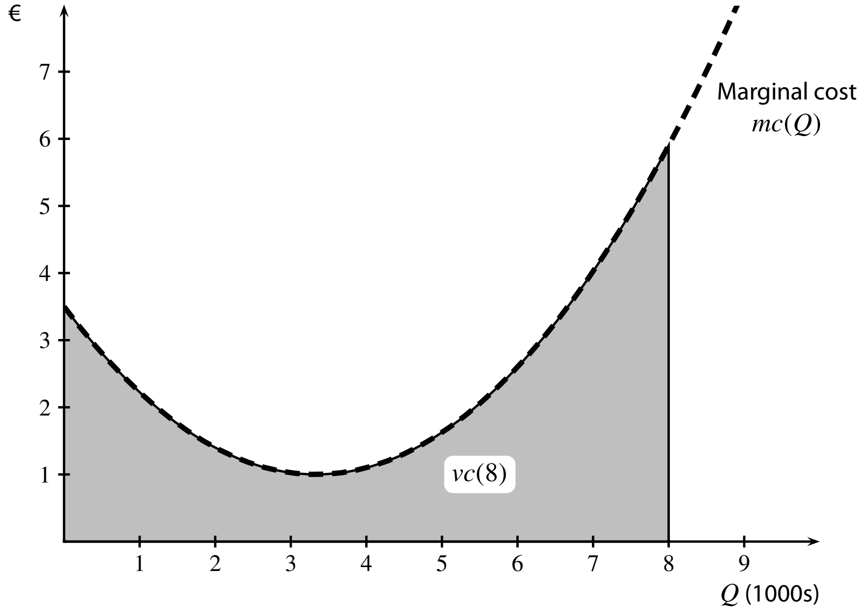


Figure 4.4 illustrates that the area under the marginal cost curve up to Q is equal to the variable cost of Q. (The fixed cost is not reflected in the marginal cost, which is why this area is equal to the variable cost rather than the total cost.) This is the analogue of the following fact from our model of a single consumer: the area under the marginal valuation curve up to Q is equal to the total valuation of Q.

Figure 4.4



From this canonical cost curve to simpler special cases

In models (and in verbal discourse), it is useful to focus on simpler cost curves that emphasize the most important features of a particular application. In subsequent models, we will not use a cost curve that combines all the features seen in this section. The next two sections describe the choices one makes when simplifying the cost curve. Section 4.6 examines when it is important to model the fixed cost. Section 4.8 outlines the simplest cost curves that capture each of the four cases of economies of scale listed in Table 4.1.

4.6 When are fixed costs important?

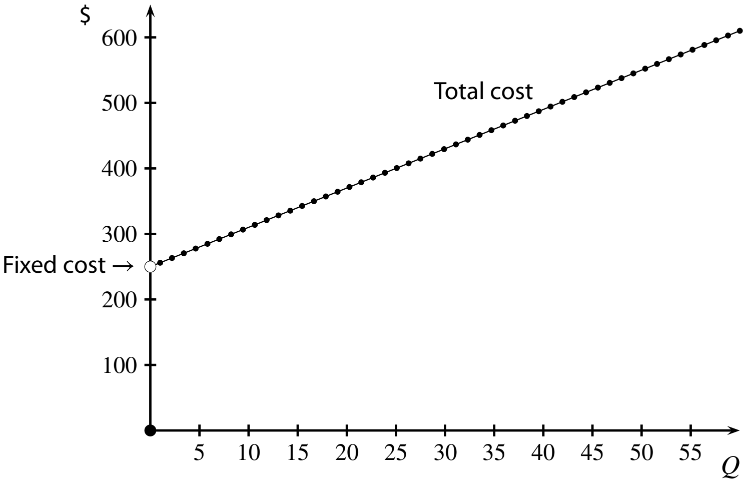
Fixed costs can be a particularly strong source of sustained industry concentration owing to (a) mergers, because such mergers avoid the duplication of these costs, and (b) a first-mover advantage, because an entrant may not recover these costs once it is competing with the incumbent. However, this is true only when the fixed costs are significant enough. Though every production technology has some fixed cost, such a cost is often—in fact, usually—not an important source of economies of scale. It is then not worth modeling.

Everything is relative

Whether the fixed cost is significant depends on how it compares to other costs at the firm's likely scale of production.

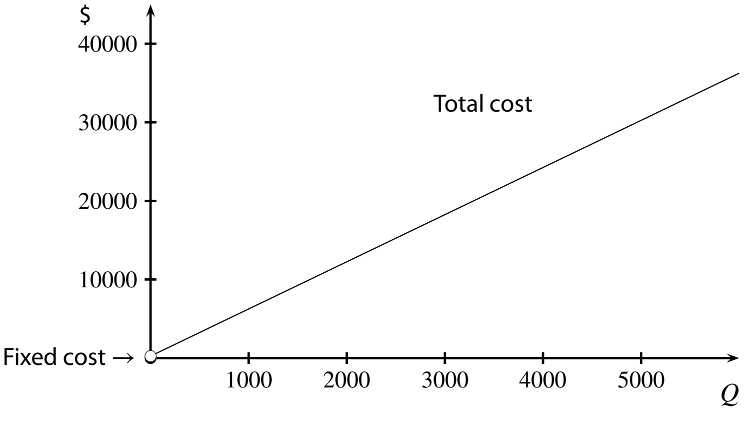
Suppose that a small school with 60 students plans to sell a school photo as a fund-raiser; the potential customers are the students' families. The photographer charges \$250 for the sitting and \$6 for each copy. The cost function for the first 60 units is graphed in Figure 4.5 as a series of dots; a smooth approximation is drawn in the same figure as a solid line. The fixed cost is \$250; the marginal cost is constant and equal to \$6. The fixed cost is significant. If there were two competing student groups each producing such a photo, a significant fraction of the total cost could be saved if the two groups merged.

Figure 4.5



Suppose the school has 6000 rather than 60 students. Then the cost curve should be drawn out to $Q = 6000$ as in Figure 4.6. Here the fixed cost is insignificant and cannot even be discerned in the graph. If two competing student groups merged in order to eliminate the duplicate fixed cost, just a small fraction of the total expense would be saved. Hence, the fixed cost does not create a strong pressure to merge and can be ignored.

Figure 4.6



Cases in which fixed costs are important

There is no hard-and-fast rule about when to worry about fixed costs. However, it is fair to say that fixed costs are important sources of economies of scale mainly for knowledge or information goods. We can subdivide these into two cases as follows.

1. *R&D costs.* R&D produces knowledge needed to start production of a good and this cost may be large relative to the per-unit production cost. For example, the R&D cost of a new cholesterol medication is high and must be incurred even if only one dose is subsequently manufactured, but the per-unit production cost of such medication is typically low. Another example is a software package: with on-line distribution, the per-unit production cost is close to zero.
2. *First-copy costs of published information and entertainment.* The information itself is the product. For example, it is expensive to develop music for and to record a CD, even if only a single CD is subsequently produced. Other examples include movies, books, and newspapers. In each case, the per-unit duplication cost of the information is quite low (and is getting lower owing to digital technology).

Thus, most metropolitan markets today have a single "upmarket" newspaper, such as the *London Times* or the *New York Times*. (There are other newspapers, but with different content and style.) This is also *one* reason why there is one dominant firm (Intel) making microprocessors for PCs.

There are two other cases—more complicated and less important than the preceding two—in which fixed costs may be significant.

3. *Distribution network costs in densely populated areas.* By "distribution network costs" we mean the costs of supplying utilities such as water, fixed-line telephone, and cable television to each household or business. There is a large cost to stringing a television cable in order to supply a single household, whether or not other households that the cable passes by also subscribe to the service. Similarly, there is a large cost to burying a sewer in a street, whether one or all households on the street get connected to the sewer. The reason for the cumbersome "in densely populated areas" qualifier is that the economies of scale are present only if, by serving one customer, you must pass your network by other customers whom you could also serve at little extra cost.
4. *Capital costs of providing quality of service.* Consider a public library. The more diverse its collection of periodicals and books, the more valuable the library is to each user. Buying and storing a copy of many books and magazines involves a large cost regardless of whether one person or many use the library. Similarly, for a retailer to offer a large variety of home electronics, it must incur a large capital cost (for the display space and a minimal inventory of each product) regardless of how many customers who frequent the store. A third example is the provision of cellular phone service. The more coverage of a country or region that a provider has, the more valuable is the service. Providing extensive coverage is expensive even if just one customer uses the system.

4.7 Pitfalls to avoid regarding fixed costs

Don't mix up the effects of quantity and quality on costs

Fixed costs are those that do not vary with the volume of production *for a given product of a given quality*. It may be possible to vary the fixed cost by increasing or decreasing the *quality* of the product, but this does not change the importance of the fixed cost.

It would be expensive to produce the *New York Times*, with its current format, content, and other quality-defining characteristics, even if only one copy of the paper were produced. One could produce a two-page newsletter with a small first-copy cost, but such a newsletter would be a different product with a different quality.

Similarly, it may be possible to produce an ineffective and low-quality word-processing program with a low development cost, but this does not negate the fact that a high-quality program has a large long-run fixed cost.

Exercise 4.1. Suppose that the costs of a large newspaper company consist of

1. workers and machines who print the newspaper;
2. newsprint and ink; and
3. reporters/editors/typesetters who prepare the content.

For each of the listed costs, determine if is a long-run fixed cost. You can answer this by examining the following scenario: the newspaper decides to keep the same quality newspaper (i.e., sell the same product) while cutting its circulation in half. A long-run fixed cost is one that could never be reduced without changing the quality of the product and hence that is not linked to the size of circulation.

Don't mix up long-run and short-run fixed costs

Other words for the fixed cost, depending on the industry or situation, are "setup cost", "first-copy cost", or "entry cost". We stick to the more common and generic term "fixed cost" even though it has the following problem: The short-run cost curve has a property that is also called a "fixed cost"—the cost of inputs that are fixed in the short run—which is unrelated to the fixed cost of a long-run cost curve. When we work with short-run costs in Chapter 6, we will have to be more careful about the distinction between these two concepts. For now, remember that the fixed cost is what it would cost to serve one customer or to sell one unit, if the firm had unlimited time to optimize production for such low volume or if the firm were starting production from scratch.

In the short run, there are always substantial costs that are fixed. However, substantial fixed costs for a long-run cost curve are the exception, not the rule. For a large electric power company, the cost of its installed capacity of power plants is a large short-run fixed cost. However, this industry has small long-run fixed costs. A firm that chooses to serve just one customer or to produce a single kilowatt-hour per day could use a tiny gasoline generator (rather than the large coal and nuclear power plants that are used to serve many customers).

Exercise 4.2. Recall the example of a large newspaper company from Exercise 4.1. The costs of the company are:

1. workers and machines who print the newspaper;
2. newsprint and ink; and
3. reporters/editors/typesetters who prepare the content.

Which costs are fixed in the short run, that is, which cannot be adjusted quickly even if the firm shuts down (but meets its financial obligations rather than entering into bankruptcy)? This is a question of degree, not a black-and-white categorization, so it is better to give a qualitative discussion of how long it would take to reduce or increase each input.

(In Exercise 4.1, you identified the long-run fixed costs. Remember that this question about short-run fixed costs bears no direct relationship to Exercise 4.1.)

Fixed costs and sunk costs

There is no particular relationship between fixed costs and sunk costs, a point we make because people sometimes confuse themselves by trying to find one.

It is true that many fixed costs are one-time expenses that become sunk once incurred. Examples are the R&D expense of developing a new computer chip or the cost of writing a book. However, other fixed costs are ongoing and do not become sunk. An example is the R&D expense of maintaining antivirus software. Such software is usually sold on a subscription basis, and the developers must constantly update the product to respond to new threats. The cost of such updates is large and must be incurred no matter how many subscribers the firm has (as long as the firm continues to supply the same-quality product). Another example is the expense of putting together a major newspaper each day (the costs of journalists, editors, and setting up the printing).

(The only useful relationship is that fixed costs that become sunk cannot simultaneously be both. Once sunk, such costs should merely be ignored; they are no longer a source of economies of scale. For example, after two firms have developed similar competing digital media technologies, it is too late to merge in order to save on the development costs already incurred.)

Exercise 4.3. A firm may develop software for blocking pop-up windows. The software is to be sold on a subscription basis, giving users access to regular updates that are needed to adapt to changing tactics of advertisers. The product has the following costs:

1. Initial development of software: $10 million.
2. Development of updates: $200K per month.
3. Distribution costs (e.g., payment processing): $4 per customer per year.

a. Decision problem 1: You have not yet decided to develop the software and enter the market, but you are putting together a business plan to decide whether to do so and what price to charge. What are the long-run fixed costs? What are the sunk costs?

b. Decision problem 2: You have already developed the product and have been operating for a year. A new competitor has entered the market, shifting your demand curve. You are now deciding how to adjust your pricing and whether to shut down. What are the long-run fixed costs? What are the sunk costs?

4.8 The leading examples of cost curves

In subsequent chapters, we will work with four types of cost curves. Each is simpler than the one studied in Section 4.5, and each allows us to focus on one of the cases of economies of scale. Table 4.2 shows the assumptions on the fixed cost and marginal cost for each case.

Table 4.2

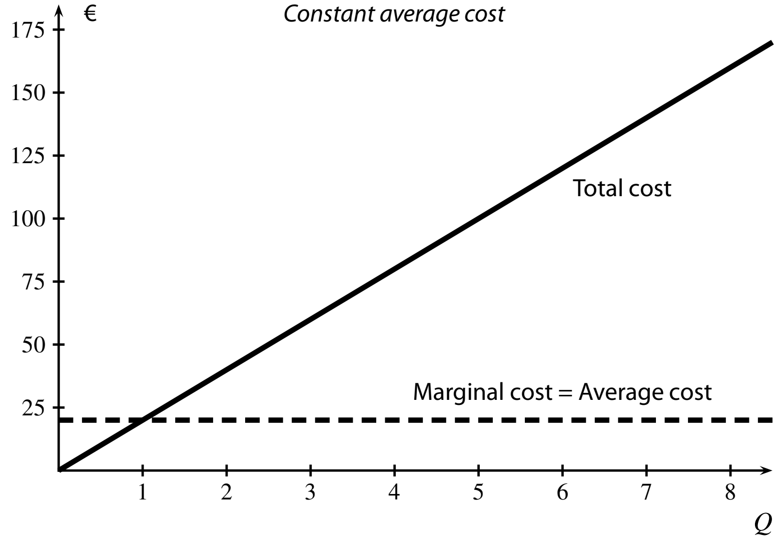
Economies of scale		Assumptions	
In words	shape of $ac(Q)$	FC	$mc(Q)$
no economies of scale	constant	$= 0$	constant
diseconomies of scale	increasing	$= 0$	increasing
economies of scale	decreasing	> 0	constant
U-shaped average cost	first decreasing, then increasing	> 0	increasing

No economies of scale

Assuming no economies of scale is a good approximation if the average cost curve is very flat over the range of relevant output levels. It is also a useful assumption if we want to focus on things that are not driven by the properties of the cost curve. (The case of no economies of scale is the simplest and blandest cost structure.)

The average cost curve is constant if and only if there is no fixed cost and the marginal cost curve is constant. The cost curve is simply a straight line that starts at the origin and whose slope equals the marginal cost. Average cost equals marginal cost. Figure 4.7 shows the graph of $c(Q) = 20Q$. The marginal and average cost are equal to 20 at all output levels.

Figure 4.7



Diseconomies of scale

We might assume that a firm has diseconomies of scale if, in reality, the initial economies of scale occur over only a small range of output and so the firm is likely to end up operating where there are diseconomies of scale. Also, when we study firms operating in perfectly competitive markets, the case of diseconomies of scale is simplest.

Sufficient conditions for there to be diseconomies of scale over all output levels are that $FC = 0$ and that marginal cost be increasing. This means the cost curve starts at the origin and becomes steeper as output increases; hence it has a shape like the one in Figure 4.8. The corresponding marginal and average cost curves are shown in Figure 4.9.

Figure 4.8

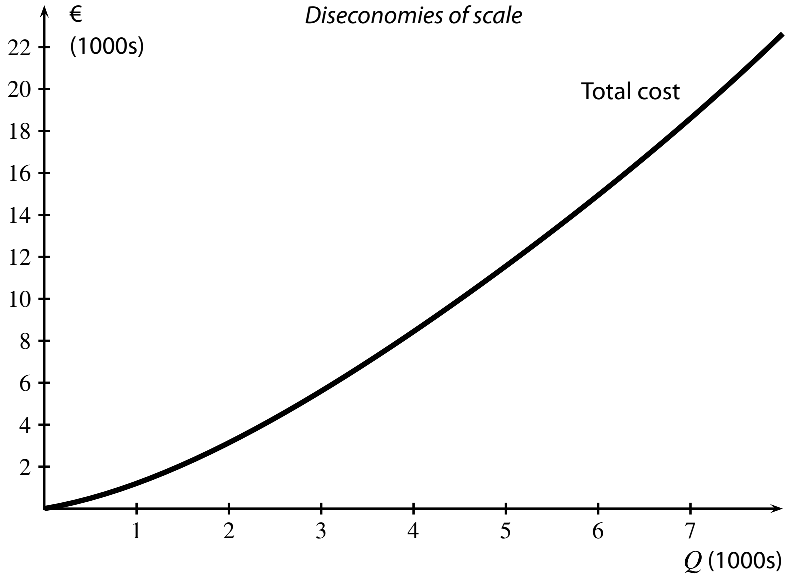
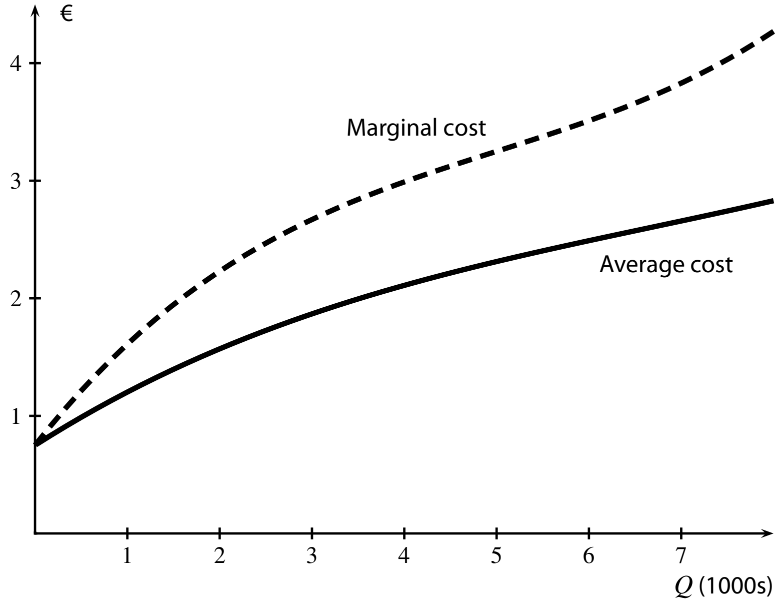


Figure 4.9



Economics of scale

We might assume that a firm has economies of scale if, in reality, the eventual diseconomies seen in Section 4.5 occur at such a large output level that the firm is never likely to encounter them—even if the firm were the only one in the market. Such an industry is a natural monopoly because cost savings from mergers remain as long as there is more than one firm.

There are economies of scale if either (a) marginal cost is always decreasing or (b) there is a fixed cost and marginal cost is constant. The insights we seek are the same in either case and, since case (b) is simpler, it is the one we use. Furthermore, constant marginal cost is a good approximation for most cases in which the economies of scale are driven mainly by large fixed cost, such as the following.

- Once the printing presses are running, each copy of a book that comes off the press costs about the same amount.
- Once a computer or memory chip is developed, actually producing them occurs at about constant marginal cost.
- Once a new medicine is developed, producing each pill occurs at about constant marginal cost.
- Once software is developed, the duplication cost is nearly zero and hence nearly constant.

With a fixed cost and constant marginal cost, the cost curve is a line starting at the fixed cost. For example, suppose it costs €20,000 to record a CD and €3 per copy to reproduce it. Then

$$\begin{aligned} c(Q) &= 20{,}000 + 3Q\,, \\ FC &= 20{,}000\,, \\ mc(Q) &= 3\,, \\ ac(Q) &= 20{,}000/Q + 3\,. \end{aligned}$$

Figure 4.10 (page 92) shows the cost curve and Figure 4.11 shows the marginal and average cost curves.

Figure 4.10

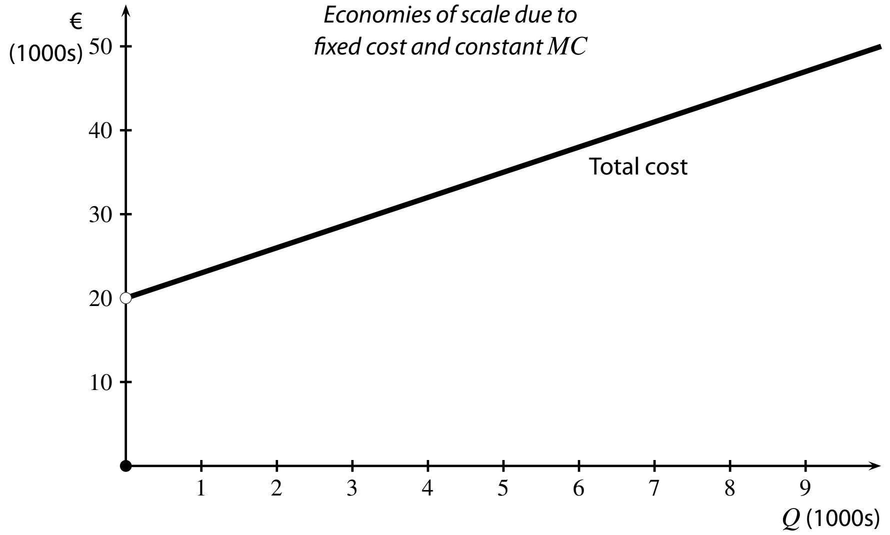
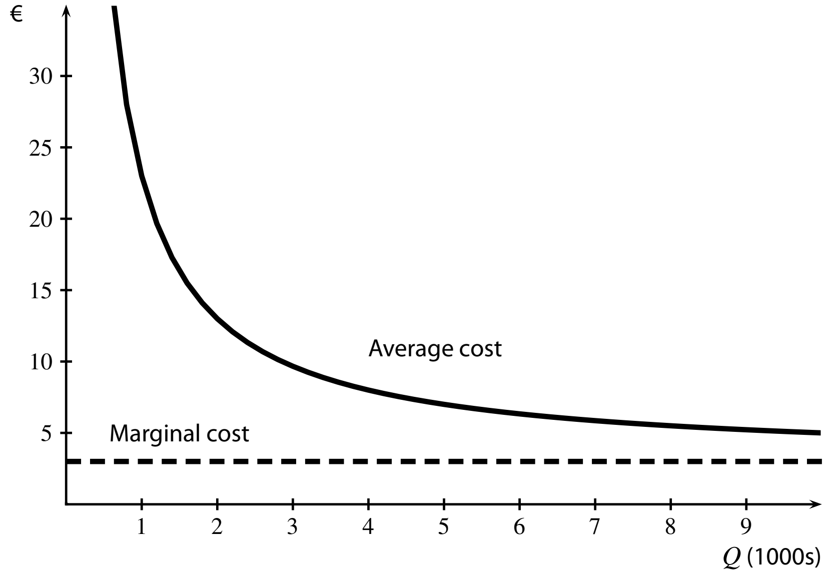


Figure 4.11



U-shaped average cost curve

We already saw a cost curve with U-shaped average cost in Section 4.5. In fact, we said that this is the realistic shape of most cost curves.

The simplest case of U-shaped average cost is when (a) $FC > 0$ and (b) marginal cost is increasing for all output levels; this is the case we use in subsequent applications.

For example:

$$\begin{aligned} c(Q) &= 6000 + 2.5Q + 0.0005Q^2\,, \\ FC &= 6000\,, \\ mc(Q) &= 2.5 + 0.001Q\,, \\ ac(Q) &= 6000/Q + 2.5 + 0.0005Q\,. \end{aligned}$$

The corresponding curves are shown in Figures 4.12 and 4.13. Figure 4.12 shows the total cost curve. We see that there is a fixed cost of €6000 and that the slope (marginal cost) is increasing. Figure 4.13 shows (a) the marginal cost curve, which is increasing, and (b) the average cost curve, which is initially decreasing and then increasing.

The efficient scale of production Q^u, where the average cost reaches its minimum, is approximately 3500 in Figure 4.13. The efficient scale of production is where the marginal cost and average cost curves intersect (have the same value), so this can be used as a shortcut to calculate the minimum average cost: one simply solves $ac(Q) = mc(Q)$ for Q. In the example, this yields

$$\begin{aligned} 6000/Q + 2.5 + 0.0005Q &= 2.5 + 0.001Q\,, \\ 6000/Q &= 0.0005Q\,, \\ 120{,}000{,}000 &= Q^2\,, \\ Q &\approx 3464\,. \end{aligned}$$

To then find the minimum average cost AC^u, we calculate the average cost of $Q = 3464$. That is, $AC^u = ac(Q^u) = ac(3464) \approx 5.96$.

Figure 4.12

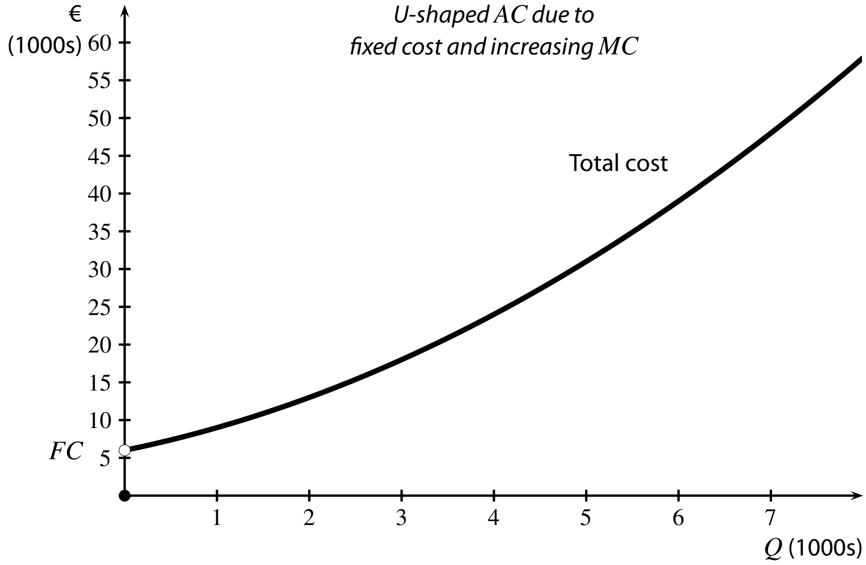
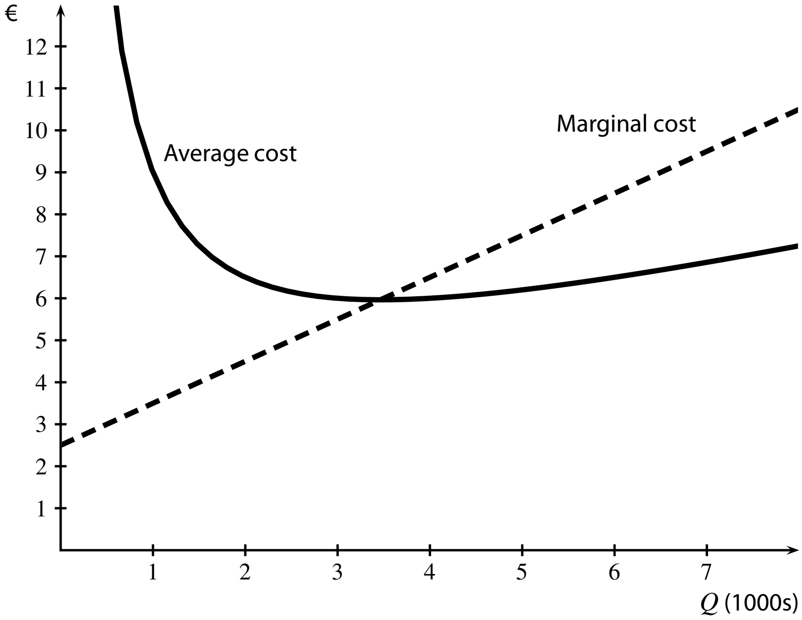


Figure 4.13



Exercises 4.4, 4.5, and 4.6 give you some practice. These exercises illustrate a distinction made in the Preliminaries chapter between high-data numerical examples and the low-data qualitative analysis that is the true purpose of this book. Exercise 4.4 is a numerical example that helps you understand the concepts. However, you are likely to lack such precise data in practice and you will instead have to perform analyses like those in Exercises 4.5 and 4.6. How much can you say with the small amount of soft information that you have available?

Exercise 4.4. Consider the following cost function:

$$c(Q) = 100 + 10Q + Q^2.$$

a. What are the formulas for the fixed cost, variable cost, average cost, and marginal cost?

b. At what output level Q^u is average cost lowest?

c. What is the minimum average cost AC^u?

Exercise 4.5. Suppose a firm has a fixed cost as well as increasing marginal cost and thus has a U-shaped average cost curve. Suppose that its marginal cost increases by a fixed amount ΔMC at all output levels—for example, a per-unit tax is imposed on the firm's output. Based only on this information, what can you say about how Q^u and AC^u change? A graph can help you figure out the answer and should be used to illustrate your it.

Exercise 4.6. Suppose a firm has a fixed cost as well as increasing marginal cost and thus has a U-shaped average cost curve. Suppose that its fixed cost increases by ΔFC—for example, the government imposes a yearly license fee to operate in the market or there is an increase in the firm's R&D cost. Based only on this information, what can you say about how Q^u and AC^u change? A graph can help you figure out the answer and should be used to illustrate it.

4.9 Wrap-up

Changes to production and to the use of inputs take time to implement. Some inputs, such as machinery and buildings, take a long time to adjust. Labor can take several months or longer to adjust, depending on restrictions on firing and on search costs in hiring. Utilities and consumable inputs can be adjusted quickly. Thus, starting from a given production process and input mix and shifting to a new output level, the minimum cost of production for the new output level depends on the time one has to adjust the inputs. A simple way to capture this is to consider two horizons: In the short run, some important inputs cannot be adjusted at all; in the long run, either a firm can adjust all inputs or it is starting from scratch and hence is not bound by previous production decisions. This chapter is about long-run costs.

Decisions should be guided by measures of cost that differ from the accounting values.

1. The *opportunity cost* of a resource owned by the firm is the return the resource could earn if used outside the firm. By including opportunity costs in the measurement of cost, we restore the heuristic "do a project if and only if revenue exceeds cost" and we avoid the mistake of ignoring the value of the alternative use of resources.
2. Costs are *sunk* if they cannot be changed by any decisions under consideration. Since sunk costs affect all options equally, they should not affect decisions. Ignoring sunk costs helps us to avoid letting them influence decisions.

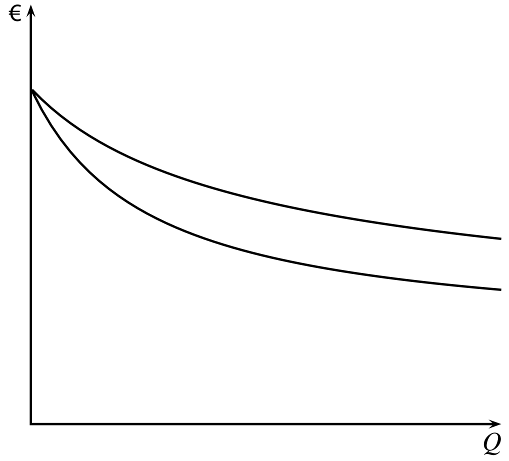
We separated out other components and transformations of the cost curve $c(Q)$, such as the average cost curve $ac(Q)$. If the average cost is decreasing, then a merger reduces the total cost of production; we then say that there are economies of scale. If $ac(Q)$ is constant (no economies of scale), then the total cost of output does not change following a merger. If $ac(Q)$ is increasing (diseconomies of scale), then the total cost of output goes up after the merger.

In applications in subsequent chapters, the easiest case is always when there is no fixed cost and when marginal cost is either constant or increasing (yielding constant or increasing average cost). Therefore, unless stated otherwise, we assume that the cost curve has these properties. When there is a fixed cost (and hence economies of scale over some range), marginal conditions may not be sufficient. This case is given special treatment.

Additional exercises

Exercise 4.7. Figure E4.1 shows the AC and the MC curves of a manufacturing firm. Without any further information, can you tell which one is which?

Figure E4.1



Exercise 4.8. Consider the following cost function:

$$c(Q) = 144 + 3Q + Q^2.$$

a. What are the formulas for the fixed cost, variable cost, average cost, and marginal cost?

b. At what output level Q^u is average cost lowest?

c. What is the minimum average cost AC^u?

Chapter 5

Competitive Supply and Market Price

5.1 Motives and objectives

Broadly

We revisit the notion of competitive markets, first taken up in Chapter 2. In that chapter, we studied a stylized market in which each seller has a single unit of a good to sell and each buyer wants to purchase at most one unit. We now have the tools in place to study more realistic and interesting markets—markets in which the suppliers are firms that make complex production decisions and the buyers are customers who may purchase any desired quantity of the good.

Perfect competition is a metaphor that approximately describes outcomes of a wide range of markets in which each trader has many substitute trading partners. The model of perfect competition does not describe the details by which transactions take place. It merely assumes the following:

1. there is a single market price known to all traders; and
2. each trader believes that she can sell or buy arbitrary amounts at this price but not at a higher price (if a seller) or a lower price (if a buyer).

These assumptions are often collectively stated as “each agent is a *price taker*”.

Each trader then has no flexibility about the price at which transactions take place; her only decision is how much to trade. A competitive market is in equilibrium when the price is such that supply equals demand; the resulting price is called the equilibrium price. If instead demand exceeds supply (for example), then some unmodeled process raises the price, thereby reducing demand and increasing supply until balance is reached.

There are no markets in which the competitive assumptions are exactly satisfied, but there are many markets in which they are approximately satisfied. Consider, for example, commodity markets such as that for wheat. This is a well-organized market in which traders are aware of the price at which transactions take place. Collectively the farmers’ production decisions affect the market price, but each individual farmer’s production is a small part of the total and so has little effect on the price of wheat. Stock markets are also competitive.

The competitive model is a useful approximation even for markets that are not nearly as competitive as commodity markets and stock markets. Although a firm in an industry should exploit any market power it has and should seek out every possible strategic advantage against its rivals, at the end of the day the difference between the outcome of this strategic interaction and the outcome of perfect competition may be small from the viewpoint of an outsider (such as a manager who purchases inputs from this market). Studying markets as if they were perfectly competitive is simpler than studying the detailed interaction between many agents who make complicated pricing decisions.

As a manager, you may feel that you want to avoid managing a firm in a competitive market because “there are no profits to be made”. Although there is a grain of truth to this statement, you can still earn a return on your expertise. Furthermore, market equilibrium is a static concept. In reality, markets are changing in response to changing input prices, customer preferences, and technologies. The equilibrium is where the market is headed, even if the environment changes yet again before the market gets there. One way to state the goal of this chapter is to help managers understand where markets are headed so that they can anticipate these changes. By being better and faster at predicting the paths that markets take, managers can increase the profits of their firms—even if there are no profits to be made in the hypothetical static equilibrium.

More specifically

The material in this chapter is divided into three parts.

Equilibrium and surplus. We first consider equilibrium, revisiting the main concepts of Chapter 2. We want to see if markets are efficient ways to trade and how the gains from trade are divided among buyers and suppliers.

Supply decisions. We then take a more careful look at firms’ supply decisions and see how a firm’s supply curve is related to its cost curve. We eventually consider three cases: increasing, constant, and U-shaped average cost curves. We skip the case of decreasing average cost (economies of scale) because such industries are natural monopolies.

Profits. Firms can earn profits in a competitive equilibrium if they enjoy privileged cost structures that cannot be replicated by potential entrants. However, if the firms have constant marginal cost or if any entrant has access to the technologies of incumbent firms, then competition pushes profits down to zero.

Furthermore, a theme running through the chapter are the entry/exit decisions of firms and their implications for market equilibrium.

5.2 Supply and equilibrium

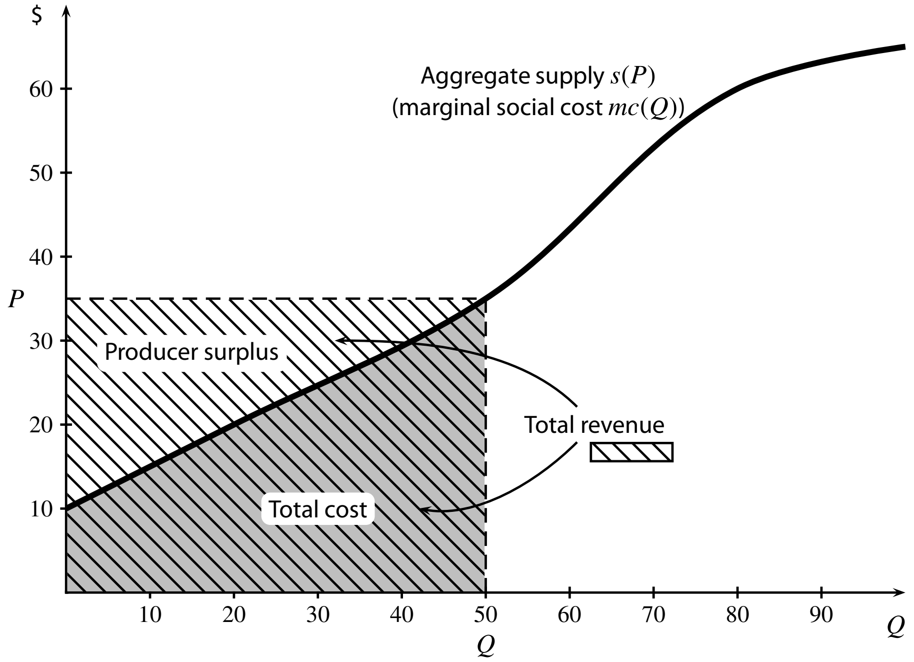
Chapters 1 and 2 also presented a model of a competitive market. What is new in this chapter is that firms may produce any amount—not just zero or one units—so that their supply decisions and entry and exit decisions are more complicated (and interesting). Also, as already introduced in Chapter 3, consumers may buy more than one unit.

We begin with some "bottom line" conclusions from Chapters 1 and 2 that remain true independently of the details of the firms' supply decisions.

Aggregate supply, cost, and producer surplus

At each market price P, each firm (including any potential entrant) decides whether and how much to produce, presuming that its decision does not affect the market price. We denote the total supply by $s(P)$. The bottom-line conclusions about the supply curve are illustrated in Figure 5.1.

Figure 5.1



Using the same method by which we illustrated demand, valuation, and consumer surplus, this figure illustrates supply, cost, and producer surplus. The supply curve is the inverse of the marginal cost curve. As with demand curves, we plot the supply curve with price on the vertical axis so that the supply curve and the marginal cost curve are graphed with the same orientation. At the price $P = 35$, supply is $Q = 50$. The total cost of producing Q is the area under the supply curve, that is, the area of the shaded region in Figure 5.1. The total revenue is the area of the striped region. The producer surplus is the difference between them.

These conclusions hold for individual and aggregate supply curves and for a range of circumstances regarding the firms' costs, the possibility of entry, and the possibility that input prices rise as the industry expands its output and hence its demand for the inputs.

Here is some intuition for why the supply curve is the inverse of the marginal cost curve (and vice versa). Suppose that, at a current output level, the price exceeded the marginal social cost. Then someone would figure out how to increase output by a single unit (e.g., by expanding output of an existing firm or entering the market), thereby making a profit equal to the difference between the price and the marginal social cost. Such expansion of supply would continue until the price equaled the marginal social cost of production.

Elasticity of supply

We measure the responsiveness of supply to changes in price by the *elasticity of supply*. Because elasticity of supply is mathematically similar to elasticity of demand, we can limit ourselves to a brief treatment.

Denote the elasticity of supply by E_s. It is, roughly, the percentage change in quantity divided by the percentage change in price:

$$E_s = \frac{\%\text{ change in } Q}{\%\text{ change in } P}.$$

Because a higher price causes a higher supply, elasticity as defined here is positive.

For discrete changes, we use the formula for arc elasticity:

$$E_s = \frac{Q_2 - Q_1}{\frac{1}{2}(Q_1 + Q_2)} \Bigg/ \frac{P_2 - P_1}{\frac{1}{2}(P_1 + P_2)}.$$

If the supply curve is smooth, then we define point elasticity by

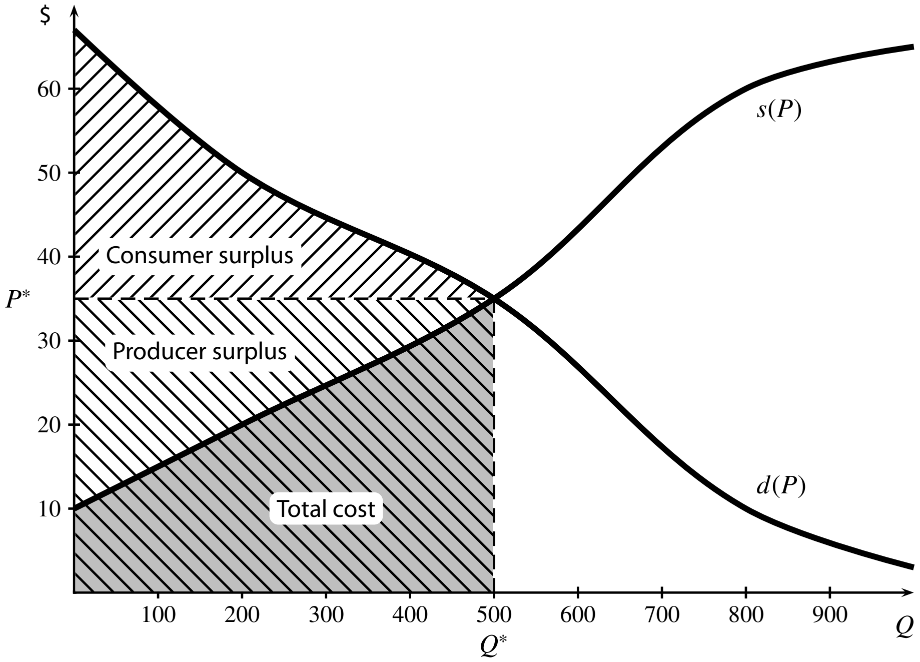
$$E_s = \frac{dQ}{dP}\frac{P}{Q}.$$

In the limiting case, as supply becomes very inelastic, the supply curve is a vertical line and is said to be *perfectly inelastic*. At the other extreme, as supply becomes very elastic, the supply curve is a horizontal line and is said to be *perfectly elastic*.

Equilibrium and efficiency

Denote the demand curve for the good by $d(P)$. The equilibrium price is P^* such that supply equals demand: $s(P^*) = d(P^*)$. Let Q^* denote the equilibrium trade. We illustrate an equilibrium in Figure 5.2. In this example, the equilibrium price is $P^* = 35$, where supply and demand both equal $Q^* = 500$.

Figure 5.2



The consumer surplus is the area above the horizontal line at P^* and below the demand curve (total valuation minus total expenditure). The producer surplus is the area below the horizontal line at P^* and above the supply curve (total revenue minus total cost).

In equilibrium, the marginal costs of the producers and the marginal valuations of the customers are equal to P and hence to each other. Therefore the marginal conditions for maximizing total surplus are satisfied and so the market is efficient.

5.3 An individual firm's supply decision

The firm's revenue curve and output decision

An individual firm's decision is just an extreme case of the profit-maximization problems studied in Chapter 7. Suppose the market price is P. Then the firm's revenue for any level of output Q is PQ. That is, its revenue curve is $r(Q) = PQ$. Marginal revenue—the extra revenue from selling one more unit—is the price P. The marginal condition $mc(Q) = mr(Q)$ is thus $mc(Q) = P$. The firm produces up the point where its marginal cost is P.

Supply with constant or decreasing average cost

If there are economies of scale (decreasing average cost) for any level of production, then we expect an industry to be a natural monopoly. Therefore, the competitive model is not relevant. (If we attempted to study this case with the competitive model, we would end up with nonsense.)

The competitive model is relevant with constant average cost, but this is an extreme case that we consider later.

Supply with increasing average cost

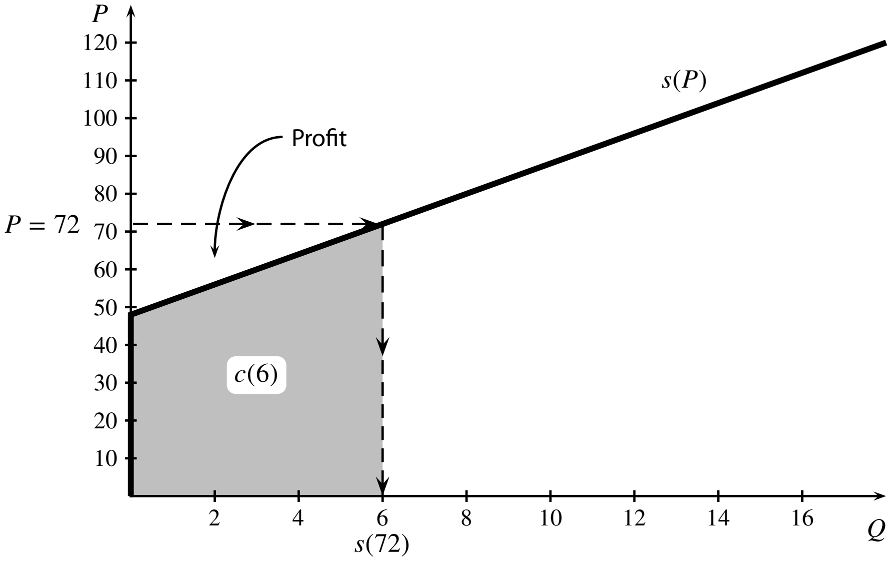
The leading case of diseconomies of scale involves (a) no fixed cost and (b) increasing marginal cost. Marginal conditions are sufficient. From the marginal condition, we know that $s(P) = Q$ if and only if $mc(Q) = P$. Thus, the firm's supply curve $s(P)$ is the inverse of its marginal cost curve $mc(Q)$. Total cost $c(Q)$ is the area under the marginal cost curve—hence under the supply curve—up to Q.

Suppose the firm's cost curve is $c(Q) = 48Q + 2Q^2$. Then its marginal cost is $mc(Q) = 48 + 4Q$. We find the supply curve by solving $mc(Q) = P$ for Q:

$$\begin{aligned} 48 + 4Q &= P\,, \\ 4Q &= -48 + P\,, \\ Q &= -12 + (1/4)P\,. \end{aligned}$$

To graph the supply curve with price on the vertical axis, we just graph the marginal cost curve *except* that we extend the graph along the vertical axis to show that supply is zero when the price is less than 48. This is shown in Figure 5.3.

Figure 5.3



Exercise 5.1. Suppose that a firm has no fixed cost and that its marginal cost equals 10+2Q. (Its cost curve is $c(Q) = 10Q + Q^2$.)

a. Write the equation for the firm's supply curve. Graph the supply curve with price on the vertical axis.

b. Calculate the firm's output and profit when $P = 20$ and when $P = 30$. For $P = 30$, illustrate the output decision, the cost, and the profit on the graph of the supply curve.

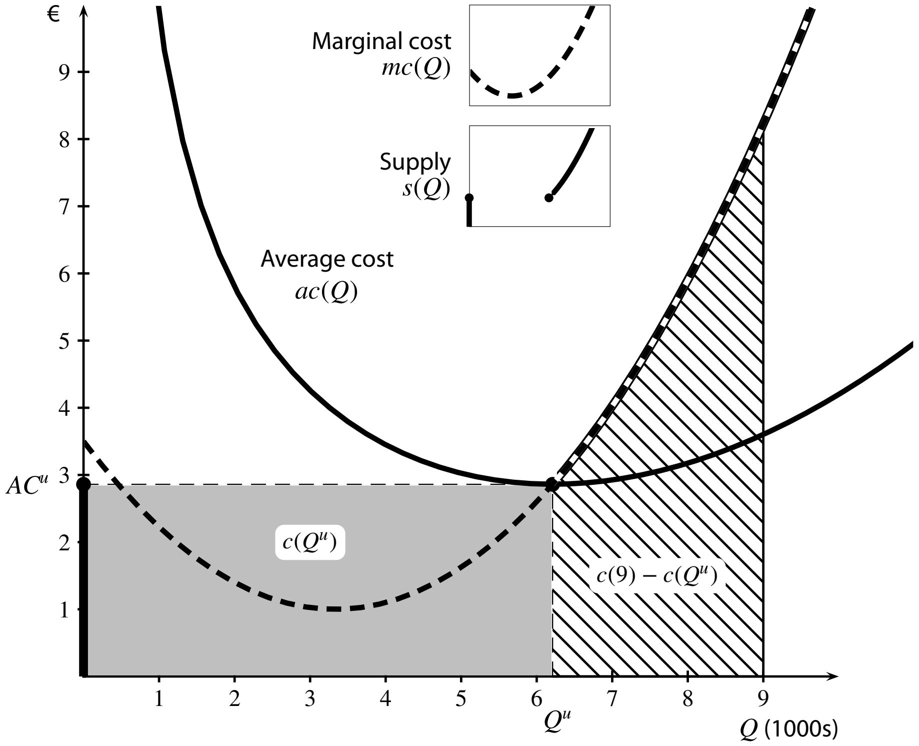
Supply with U-shaped average cost

A firm's average cost curve is U-shaped under these circumstance:

1. $FC > 0$ and MC is increasing;
2. $FC \geq 0$ and MC is first decreasing and then increasing.

Case 1 is simplest and is used later in the book, but we present an example of case 2 in Figure 5.4. It shows the average cost and marginal cost curves for the production function that was first analyzed in Figures 4.1 and 4.2.

Figure 5.4



Let AC^u be the value of this minimum average cost and let Q^u be the quantity that achieves it. Marginal conditions are "almost" sufficient with one caveat: the firm should check the shutdown option.

1. The firm should shut down (produce 0) whenever $P < AC^u$, for in that case it is impossible for the firm to break even.
2. When $P > AC^u$, the firm can make a profit. The supply curve is given by the marginal condition, that is, by the upward-sloping part of the marginal cost curve.

3. When $P = AC^u$, the firm is indifferent between shutting down or producing Q^u. Either way, it earns zero profit. This is the price at which entry or exit takes place and at which the supply curve jumps from the vertical axis to the marginal cost curve.

Thus, we obtain the supply curve shown in Figure 5.4.[1]

Exercise 5.2. Consider the firm in Exercise 5.1 (its variable cost is $10Q + Q^2$ and its marginal cost is $10 + 2Q$), but now suppose it has a fixed cost $FC = 100$ that can be eliminated by shutting down. Thus, its cost curve is $c(Q) = 100 + 10Q + Q^2$, the same as in Exercise 4.4.

a. If the firm does not shut down, how much does it produce?

b. In Exercise 4.4, you calculated the quantity that minimizes the average cost and determined the minimum average cost. Write these numbers again. For what prices should the firm shut down?

c. Graph the average cost curve and the marginal cost curve for quantities between 0 and 20. (using e.g. Excel or simply by hand). Draw in the supply curve.

5.4 Aggregate supply

Exit and entry

The aggregate supply curve shows the total amount supplied by all firms as a function of the market price. A change in the price changes the total amount supplied both because active firms adjust their output and because firms enter or exit the market.

Some markets have *barriers to entry* that keep new firms from entering. These barriers may be legal restrictions, such as licensing requirements. For example, in New York City there is a fixed number of permits for taxis, called “taxi medallions”. The barriers may instead be threats of retaliation by an incumbent monopolist. When there are no such barriers, then firms can enter the market.

Yet when economists say *free entry*, they mean much more than just the absence of such barriers. They mean that anyone can set up a firm with the same cost structure as any existing firm. This means that entrants have access to the same technology and organizational goodwill as incumbents. This is quite an extreme assumption that holds approximately for

[1] Although the supply curve has a jump in it, the total cost of $Q \geq Q^u$ units is equal to the area under the supply curve up to Q^u provided we draw an imaginary line to connect the supply curve. In Figure 5.4, the shaded rectangle is the total cost of Q^u units, since its area is $ac(Q^u) \times Q^u$. The striped region is the additional cost of producing 9 rather than Q^u units. Hence, the total cost of 9 units is the area of the shaded region plus the area of the striped region.

industries with simple technologies but not for those that require complex organizations.

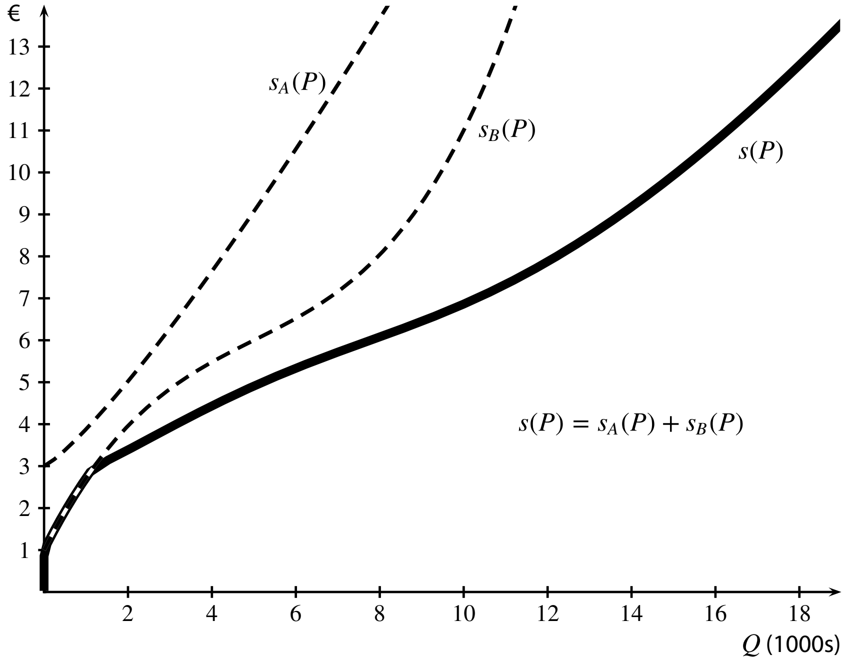
Our model of perfect competition does not rely on any form of entry. However the model is a good approximation of actual markets only if there are enough firms active in the market. A rule of thumb is that "enough" means at least four or five firms. For example, if New York taxi rates were not set by a regulatory agency then we could model the market for New York taxi services as being competitive.

In this section, we consider aggregate supply and equilibrium without free entry (though there may be entry and exit). We take up free entry in Section 5.5.

Aggregate supply and equilibrium

The aggregate supply curve is the sum of the supply curves of the individual firms. For example, Figure 5.5 shows the supply curves $s_A(P)$ and $s_B(P)$ of two firms with increasing average cost, along with the aggregate supply curve $s(P) = s_A(P) + s_B(P)$. Since price is on the vertical axis and quantity is on the horizontal axis, we sum the supply curves by adding up the quantities horizontally. As the price rises, it follows that (a) firms already in the market expand production and (b) new firms enter the market when the price exceeds their minimum average costs.

Figure 5.5



If firms have U-shaped average costs, then the aggregate supply curve jumps at prices where entry takes place. However, as long as each firm's scale of production at entry is small compared to the total size of the market, these jumps are not significant.

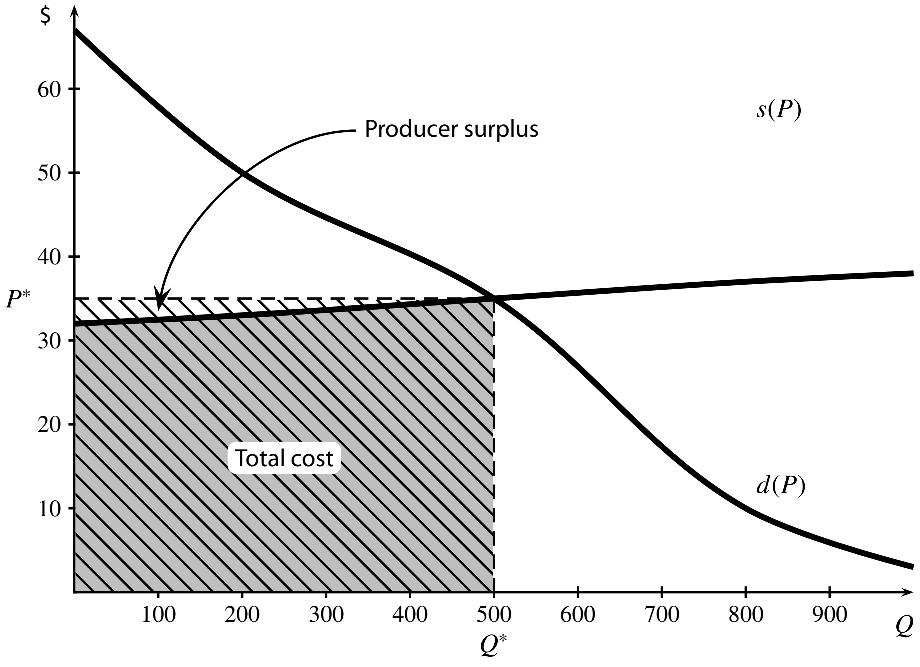
5.5 Equilibrium profits and very competitive markets

Equilibrium profits

As long as the supply curves are upward sloping, active firms earn positive profit in equilibrium. This was shown in Figure 5.2 and can be understood as follows. Each active firm chooses an output level where the price equals its marginal cost. Its supply curve is increasing when marginal cost is increasing and hence marginal cost exceeds average cost. Thus, the price exceeds the firm's average cost and so the firm earns a profit.

However, if we keep fixed the equilibrium price and quantity, then profits are lower the flatter or more elastic is the supply curve. Compare Figure 5.2 with Figure 5.6.

Figure 5.6



Aggregate supply can be very elastic for two reasons. First, this occurs if each firm's individual supply curve is very elastic, which means that its marginal cost curve is very flat. Competition is intense even if the number of firms is low[2] and even if there is no free entry, because each firm reacts to price increases by drastically increasing its own output level, thereby keeping prices from rising far above average cost. We obtain a simple model by taking this case to its limit: firms have identical constant average costs.

Aggregate supply is also very elastic if there are many potential firms with approximately the same cost curve. A small price increase causes a large increase in supply because many firms enter the market when the price goes above their minimum average costs, thereby keeping the price from rising far above the firms' average costs. We obtain a sim-

[2] But large enough for the firms to be price takers.

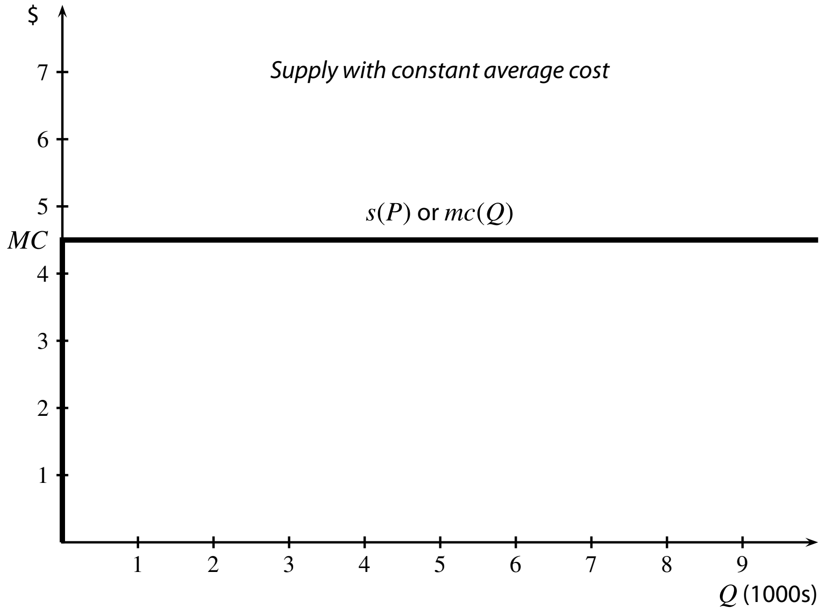
ple model by taking this case to its limit: free entry, in which any firm and any number of potential firms have access to the same technology.

We now develop each of these polar cases.

Constant average cost

Consider again a firm with no fixed cost and increasing marginal cost. Since the supply curve is the inverse of the marginal cost curve, supply is more elastic the more slowly marginal cost increases. In the limit, as marginal cost becomes constant, the supply curve becomes perfectly elastic. It is a horizontal line at the firm's marginal cost, as illustrated in Figure 5.7.

Figure 5.7



When the price is below $MC = 4.5$, the firm produces nothing because it can never break even. When the price is above MC, the firm gets a per-unit profit of $P - MC$ and therefore would like to produce an infinite amount. When the price is equal to MC, the firm is willing to produce any amount since it exactly breaks even regardless of its output.

If the industry consists of firms such as this one with constant average cost, then the aggregate supply curve is a horizontal line at the lowest average cost of the firms. The equilibrium price is therefore equal to this value. Firms that do not have the lowest cost are never active. Each active firm earns zero profit.

Free entry: The dynamics of entry and exit

The profits that accrue to each firm for a given number of firms that are in the market can be found from the model without free entry. When there *is* free entry, entry takes place until the profit is driven down to zero—that is, until one more firm could not enter the market and still make money. You are asked to develop this idea in Exercise 5.3.

Exercise 5.3. Consider a competitive market with N identical firms. Each firm has the cost curve given in Exercises 4.4 and 5.2:

$$c(Q) = 100 + 10Q + Q^2.$$

The demand curve is

$$d(P) = 1600 - 20P.$$

The following steps show you how to find the equilibrium price and equilibrium profit per firm as a function of N, and then determine how many firms would enter if there were free entry.

a. In order to calculate the equilibrium price when there are N firms, we must find the aggregate supply curve. Since the firms are identical, aggregate supply is equal to N times the supply of an individual firm. The fixed cost affects only entry or exit decisions, which we initially take as given. Thus, the individual supply depends only on marginal cost. In fact, you already found the individual supply curve for this marginal cost in Exercise 5.1. Take your answer from that exercise, which we denote by $s_i(P)$, and multiply it by N to get the aggregate supply curve: $s(P) = N \times s_i(P)$.

b. Solve $s(P) = d(P)$ for P to derive the equilibrium price as a function of N.

c. Use a spreadsheet to complete the exercise. You should create the following columns:

1. N, which ranges from 1 to 150.
2. P, calculated from N using the formula in the part b.
3. Q_i, the output per firm; this equals $s_i(P)$.
4. R_i, an individual firm's revenue; this equals PQ_i.
5. C_i, an individual firm's cost including the fixed cost; this equals $c(Q_i)$.
6. Π_i, an individual firm's profit; this equals $R_i - C_i$.

Scan down the last column. As long as the profit is positive, more firms would enter. If it is negative, firms would exit. Find the point where the profit is 0 or where it switches from positive to negative. This is the equilibrium number of firms when there is free entry. What is the price? How much does each firm produce?

Free entry: A shortcut

There is a useful shortcut for finding the equilibrium with free entry.

Aggregate supply and equilibrium with free entry is simple. All firms use the technology and organizational goodwill that yield the lowest cost. Let AC^u be the minimum average cost. When $P < AC^u$, firms in the market lose money and so choose to exit. When $P > AC^u$, entrants can earn a profit and so choose to enter. When $P = AC^u$, firms are indifferent between being in the market and being out; either way, their profit is zero. Thus, the aggregate supply curve is a flat line $P = AC^u$.

1. The only possible equilibrium price P^* is therefore AC^u, no matter what the demand curve is:
$$P^* = AC^u. \tag{5.1}$$
2. At this price, the output Q_i^* of each individual firm in the market must be the quantity Q^u that achieves the average cost AC^u; otherwise, it would lose money:
$$Q_i^* = Q^u. \tag{5.2}$$
3. Once we know the market price P, total output Q^* is determined by the demand curve:
$$Q^* = d(P^*). \tag{5.3}$$
4. Total output also equals $N^* Q_i^*$, where N^* is the number of firms in the market. From $Q^* = N^* Q_i^*$, we can find N^*:
$$N^* = Q^*/Q_i^*. \tag{5.4}$$

Equations (5.1)–(5.4) provide a recipe for calculating an equilibrium. As data, you need the demand curve $d(Q)$ of the market and the cost curve $c(Q)$ of a typical firm. A preliminary step is to calculate Q^u and AC^u for the cost curve, as described in Chapter 4. You can then plug these values into the four equations.

(This description of equilibrium is subject to one qualification, which may or may not be minor: The demand $d(AC^u)$ might not be a multiple of Q^u, in which case we must round down $d(AC^u)/Q^u$ to obtain the number of firms. The equilibrium price is then a little above AC^u. Each firm in the market earns a profit, but a potential entrant realizes that by entering it will push the market price below AC^u and hence such entry is not profitable. Our first description is a good approximation as long as the efficient scale Q^u of production is small compared to the size of the market. If it is large, then the profits of the firms in the market may be significant; there may be so few firms in the market that it is not competitive. For the exercises in this book, you need not concern yourself with this caveat. Either the number of firms works out evenly or you can use the noninteger solution as an approximation.)

Exercise 5.4 provides a numerical example to which you can apply the recipe. The cost curve is the same as in Exercise 4.4, so you can use your answer from that exercise as the preliminary step.

Exercise 5.4. Consider a competitive market with free entry. Each potential firm has the cost curve given in Exercises 4.4, 5.2, and 5.3:

$$c(Q) = 100 + 10Q + Q^2.$$

The demand curve is

$$d(P) = 1600 - 20P.$$

Use the values of Q^u and AC^u that you calculated for Exercise 4.4 as your starting point.

a. Find the equilibrium price.

b. How much does each firm produce?

c. What is the total output?

d. How many firms are in the market?

Exercises 5.5 and 5.6 are low-data qualitative exercises. They are related to Exercises 4.5 and 4.6, in which you had to determine how a per-unit tax or license fee would change Q^u and AC^u. Take your conclusions from those exercises and see how, via our recipe, the tax or fee affects P^*, Q_i^*, Q^*, and N^*.

Exercise 5.5. Consider a competitive industry with free entry and a U-shaped average cost curve. (You may assume a fixed cost and increasing marginal cost.) Suppose the government imposes a per-unit tax. What happens to the following?

a. The price of the product.

b. The output of each firm that stays in the market.

c. Total output.

d. The number of firms in the industry.

Exercise 5.6. Consider a competitive industry with free entry and a U-shaped average cost curve. (You may assume a fixed cost and increasing marginal cost.) Suppose the government imposes a yearly license fee on any firm in the market (the same for all firms, and independent of a firm's level of output). What happens to the following?

a. The price of the product.

b. The output of each firm that stays in the market.

c. Total output.

d. The number of firms in the industry.

Remarks on the zero-profit condition

Does perfect competition equal zero profit? Some textbooks treat free entry, and hence the zero-profit condition, as a component of the model of perfect competition. However, this approach has two disadvantages. First, it underestimates the usefulness of perfect competition as a model; recall that the model can also be a good approximation for markets with barriers to entry. Second, this approach exaggerates the prevalence of free entry. Such extreme free entry is a realistic assumption in markets served by small businesses such as taxis (when entry is not regulated), real-estate agencies (when a trade association does not act as a cartel), and hair salons. However, the cost structures of large organizations depend on complex "corporate culture", production processes, and managerial habits that are difficult to replicate. Hence, there is not free entry and firms may earn positive profits.

To restore the "zero profit" outcome, economists sometimes say that the goodwill and other non-replicable components of the organization should be treated as an asset and the profit should be treated as a return to this asset, so that actual "economic profit" is zero. This is just a change in terminology that some people find helpful, though it does abuse the notion of economic profit and opportunity cost because the "asset" cannot be transferred to another firm and has no opportunity cost.

That said, it is more dangerous to err by exaggerating economic profit—either by measuring it incorrectly (e.g., interpreting accounting profit as economic profit without factoring in the opportunity cost of the return on capital or on an entrepreneur's expertise) or by forecasting it incorrectly (e.g., not foreseeing the effect that free entry will have on the profitability of a market).

Is zero profit such a bad thing for a firm? A firm with zero *accounting* profit is a pretty sorry case. The shareholders are getting no return on their capital.

However, a firm with zero *economic* profit is earning a return on its assets that is as high as what these assets could obtain in any other use. Hence, these assets are not poorly invested.

Can there still be a return to being clever in a zero-profit market? The zero-profit condition is a long-term property of the equilibrium. However, the world is always changing and the long-term is never reached—it simply tells us where the world is heading. Those who understand markets well are better than others at forecasting changes in market prices and making entry, exit, and investment decisions—these people make an economic profit (which you may call a return to their expertise, if you like). Becoming more skilled at such forecasting is the largest managerial payoff from this chapter on perfect competition and from Chapter 6 on short-run decisions in competitive markets. You will thereby earn a return to your expertise that comes from creating value—by better investing the economy's resources.

In the "very competitive" models, who produces? How much? Each model of "very competitive markets" pins down the market price and the total quantity produced but leaves some things unspecified. These are given below for each model.

- *Constant AC.* In equilibrium, each firm is indifferent about its level of production—it earns zero profit no matter how much it produces. The model does not pin down how much any particular firm produces.
- *U-shaped AC and free entry.* In equilibrium, each firm is indifferent about whether it is in the market—either way, it earns zero profit. The model does not pin down which of the many potential firms are actually operating.

In the real world of very elastic but not perfectly elastic supply, as described at the beginning of this section, all these things are pinned down. Slightly increasing marginal cost determines how much each firm produces. Slight differences in cost curves determine which firms are in the market and which are not. This is comforting to know but not necessarily worth the trouble to model explicitly. The robust conclusions are the ones that come from the extreme models of constant average cost or free entry. What each firm or potential firm does is sensitive to small changes in cost structures and would require extremely accurate data to determine quantitatively.

5.6 Variable input prices

Except for the price of the good whose market we have studied, we have fixed all other prices—including the input prices.

If a firm is so large that its output decisions have significant effects on its input prices, then it is not competitive. At the aggregate level, however, an expansion of output in a large industry may drive up the prices of inputs that are special to that industry. When we take into account such endogenous determination of input prices, our conclusions from the previous sections change in the following ways.

1. Increases in input prices limit an industry's output expansion when the market price rises, so aggregate supply is less elastic than when input prices are fixed.
2. The area under the aggregate supply curve up to a quantity Q is still the total cost of producing Q, but this cost now includes the increasing marginal cost of inputs.
3. Producer surplus is therefore not the same as the profit of the firms in the market. Producer surplus equals profit plus the surplus that goes to the suppliers of inputs.

For an extreme comparison, consider an industry with constant average cost or with free entry. With constant input prices, the aggregate supply curve is horizontal and perfectly elastic. With variable input prices, the supply curve slopes upward. Firms earn zero profit in equilibrium, and all producer surplus goes to the suppliers of inputs.

5.7 Wrap-up

A competitive market is an idealized representation of a market in which there are many firms producing close substitutes, so that no firm has a strong influence on the market price. In the model, the firms produce perfect substitutes and choose their level of output taking the market price as given. The relationship between the market price and a firm's output decision is its supply curve. The sum of the firms' supply curves gives the aggregate supply curve, which associates to each price the total output supplied by all firms. Equilibrium is where supply equals demand—at the intersection of the supply and demand curves.

The nature of a firm's supply curve depends on the shape of its cost curve.

1. With no fixed cost and increasing marginal cost (increasing average cost or diseconomies of scale), marginal conditions are sufficient and the supply curve is the inverse of the marginal cost curve.
2. Hence, the more slowly marginal cost increases, the more elastic is the supply curve. In the limit, when average cost is constant, the supply curve is horizontal and perfectly elastic.
3. With a U-shaped average cost curve, a firm enters the market when the price reaches the firm's minimum average cost. As the price rises further, the firm sells the amount at which its marginal cost is increasing and equal to the price.

The profits that firms earn in an industry also depend on the shapes of their cost curves:

1. With increasing or U-shaped average costs, firms can earn positive profit if there is no entry or if potential entrants have heterogeneous costs.
2. When marginal cost increases slowly and hence firms' supply curves are highly elastic, competition is intense regardless of whether there is free entry. In the limit, with constant average costs, firms bid the price down to the lowest marginal cost and earn zero profit.
3. With free entry (anyone can set up a firm with the same cost structure as any other firm), entry pushes the price down to the minimum average cost. Firms again earn zero profit.

If we take into account that input prices increase as their demand rises, then an industry's aggregate supply curve slopes upward even when there is free entry or constant average cost. Producer surplus represents profit plus the surplus that accrues to the suppliers of the primary factors of production.

Additional exercises

Exercise 5.7. (Competitive supply) Evaluate: "In a competitive market, a firm sets its price equal to its marginal cost."

Exercise 5.8. (Profit and diseconomies of scale) Evaluate: "A firm in a competitive market without free entry is better off having diseconomies of scale, because only then can it earn a positive profit in equilibrium."

Exercise 5.9. (The need for patent protection) Suppose that each pharmaceutical company in a competitive drug market knows that, with $100 million of R&D, it can develop a cure for hay fever. The medicine will cost $20 per dosage to manufacture. However, there is no patent protection and so, once the drug is developed, any firm can also produce it at $20 per dosage. What will happen?

Exercise 5.10. (Supply with U-shaped average cost) Assume that your firm operates in a perfectly competitive market. Your total cost function is $c(Q) = 100 + Q^2$ and hence your marginal cost is $mc(Q) = 2Q$. If the market price is 60, then how much should you produce and what is your profit?

Exercise 5.11. (Free entry) Consider a perfectly competitive market with free entry. (There are firms and potential firms with access to the same technology and hence with the same cost curve.) The AC and MC curves for the common technology are given by

$$\begin{aligned} ac(Q) &= \frac{50}{Q} + Q\,, \\ mc(Q) &= 3Q\,. \end{aligned}$$

Find the (approximate) equilibrium price and the output of each firm that is active in the market.

Exercise 5.12. (Integrating different cost structures) Consider a competitive market in which the firms are grouped into two sectors, which use different technologies. The technologies cannot be replicated, so firms in each sector cannot adopt the technology of the other sector and entry possible. (However, there are many firms in each sector and so each firm behaves competitively).

The two production sectors and the demand for the good have the following properties:

Production Sector A. Each firm in Sector A has the same cost curve, which has a constant marginal cost of 100.

Production Sector B. Each firm in Sector B has the same cost function, which exhibits diseconomies of scale. The aggregate supply curve for this sector is as follows (we use M to denote "million"):

Price	50	75	100	125	150	175	200
Supply	1M	2M	3M	4M	5M	6M	7M

Demand. The demand curve for this market has these values:

Price	50	75	100	125	150	175	200
Demand	10M	9M	8M	7M	5M	4M	3M

The following questions ask you to determine the competitive equilibrium under various assumptions regarding (i) whether only one of the sectors or both the sectors serve the market and (ii) the presence of taxes.

a. Suppose the market is served only by Sector A. (Sector B does not exist.) What is the equilibrium price? How much is traded at that price? Do the firms earn a profit? Explain.

b. Suppose the market is served only by Sector B. (Sector A does not exist.) What is the equilibrium price? How much is traded at that price? Do the firms earn a profit?

c. Suppose the market is served by both sectors. What is the equilibrium? How much is produced by each sector? Do any of the firms earn a profit?

d. Suppose that the market is served only by Sector A and that a $25 (per-unit) sales tax is imposed. What is the new equilibrium price charged by the firms? How much is produced/consumed at that price? Who "bears the burden of the tax" (i.e., who pays more than the equilibrium price without the tax)?

e. Suppose that the market is served only by Sector B and that a $50 (per-unit) sales tax is imposed. What is the new equilibrium price charged by the firms? How much is produced/consumed at that price? Who bears the burden of the tax?

f. Briefly compare your answers about the tax burden in the previous two parts and relate the difference to the elasticities of supply.

Chapter 6

Short-Run Costs and Prices

6.1 Motives and objectives

Broadly

Production is a dynamic process. As market conditions change, the firm adjusts production, perhaps expanding output when demand rises and reducing output when demand falls. The firm must change the inputs used for production, but some inputs are fixed in the short run. This constraint causes costs to be higher in the short run than in the long run; in particular, the cost reduction by decreasing output is smaller and the extra cost from increasing output is higher. This leads the firm to respond less aggressively to changing market conditions in the short run than in the long run.

So far, we have studied only long-run production decisions, when all outputs can be varied and the firm can shut down. In this chapter, we consider short-run production decisions when some inputs are fixed and their cost cannot be eliminated even by shutting down. We reconsider competitive supply decisions in the short-run horizon and then compare these with long-run decisions.

More specifically

We have the following objectives concerning short-run production and cost:

1. to compare short-run and long-run cost by examining the production function;
2. to see how short-run cost depends on the status-quo input mix;
3. to define the short-run fixed cost as the cost that cannot be eliminated in the short run even by shutting down;
4. to state and provide intuition for the law of decreasing marginal return, and to see that it results in increasing marginal cost in the short run.

While we develop those ideas about cost and paint the big picture about the firm's planning horizons, we will consider the decisions of both a perfectly competitive firm (as described in the long run in Chapter 5) and a firm with market power (as described in the long run in the Preliminaries chapter and in Chapter 7).

Then we make the following detailed comparisons between short-run and long-run decisions by perfectly competitive firms.

1. Given a change in the market price, a firm's output decision is less responsive in the short run than in the long run.
2. After a shift in demand, the market price fluctuates more in the short run than in the long run.
3. A perfectly competitive firm has a lower profit if it does not anticipate that the market price fluctuates more in the short run than in the long run, and market volatility is higher if many firms make this mistake.

6.2 Short-run vs. long-run cost

Now would be a good time to re-read Section 4.2, where we explained how the cost of production depends on the planning horizon. We also gave an example of an oil refinery that can adjust some inputs only slowly. Hence, following a change in the output level, the cost soon after the change is higher than the cost after some time has passed and the firm has been able to adjust all its inputs to the optimal mix.

To understand further the distinction between short-run and long-run costs, we need to peek inside the firm's technology and input-mix decisions. We include multiple inputs in the model so that some can be varied in the short run and others cannot. For simplicity, suppose there is (i) a single *variable* input whose quantity is L and price is P_L and (ii) a single *fixed* input whose quantity is M and price is P_M. Let's call the variable input "labor" and the fixed input "machinery". The qualifiers "variable" and "fixed" refer to the short run; in the long run both inputs can be varied freely. Let $f(L,M)$ be the output when the firm uses L units of labor and M units of machinery.

Consider first the long-run cost $c(Q)$ of producing Q units; it is the lowest cost of the various input mixes that yield Q units. Suppose the input prices are P_L and P_M, respectively. Then $c(Q)$ is the value of the following minimization problem:

$$\begin{aligned} \min_{L,M} \quad & P_L L + P_M M \\ \text{subject to:} \quad & f(L,M) = Q\,. \end{aligned}$$

For simplicity, assume there is no long-run fixed cost (this means that the production function $f(L,M)$ goes up smoothly as both inputs are increased from zero).

Given the long-run cost curve $c(Q)$ and given a stable demand curve for its own product (if it has market power) or given a stable market price (if it is perfectly competitive), the firm chooses a profit-maximizing level of output as described in the Chapter 7 or Chapter 5. Using the subscript "current" to denote *current* or status-quo levels, let Q_{current} be the level of output the firm produces and let L_{current} and M_{current} be the amounts of the inputs the firm uses.

Consider what happens after an unanticipated increase in demand. The firm sees the market price rise, so it should respond by changing its level of output. The firm makes this decision for two planning horizons.

Long-term planning. On the one hand, it tries to predict what the long-run market price will be. Then, using its long-run cost function, it chooses the optimal level of output by balancing revenue and cost. The firm thereby formulates a production plan that it will implement in the long run. This may entail adding more machinery; the firm obtains capital financing, seeks zoning permits for an enlargement to the factory, contracts out the construction of the new factory, and places orders for new machinery. These inputs may finally come into place in, say, one year.

Short-term adjustments. In the meantime, the firm may want to immediately change its output level in response to the new demand curve or market price. However, only its use of the variable input can be changed. Although the firm still chooses the output level using the principles studied in Chapter 5, the measurement of its cost curve (i.e., of the relationship between its short-run output level and its cost of production) is different from the measurement of its long-run cost. Because the firm is constrained to use exactly M_{current} units of machinery, its short-run cost $c_s(Q)$ of Q units of output is the value of the following minimization problem:

$$\begin{aligned} \min_{L} \quad & P_L L + P_M M_{\text{current}} \\ \text{subject to:} \quad & f(L, M_{\text{current}}) = Q\,. \end{aligned}$$

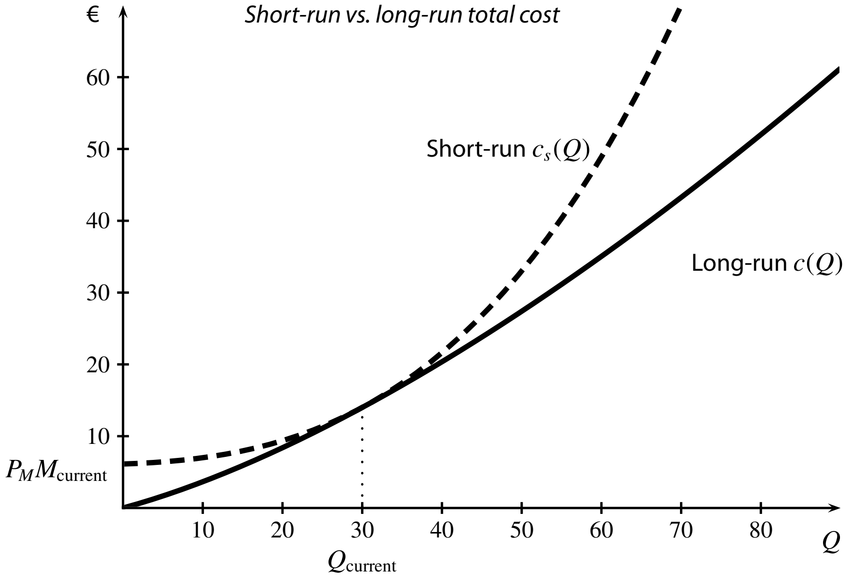
The constraint on the input mix makes production less efficient in the short run than in the long run. Whether the firm lowers or raises its output, it has a higher cost of production than what it can achieve a year from now—after it has had time to adjust its use of machinery. Furthermore, even if the firm shuts down, it incurs the cost $P_M M_{\text{current}}$. Only at the output level Q_{current} does the short-run cost equal the long-run cost, because then the firm can continue to use its current efficient production process.

In summary, if we draw the long-run and short-run cost curves on the same graph, then:

1. the short-run cost curve lies above the long-run cost curve except at Q_{current}; and
2. the short-run cost curve starts at $P_M M_{\text{current}}$, whereas the long-run cost curve starts at 0 (because we assumed for clarity that this product line has no long-run fixed cost).

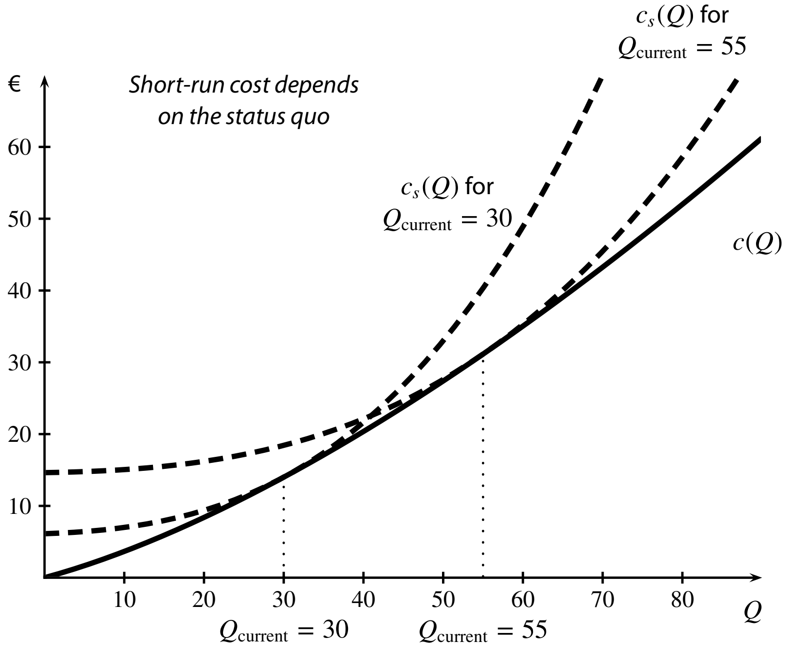
For example, if $Q_{\text{current}} = 30$ then the two curves could have the form shown in Figure 6.1.

Figure 6.1



The short-run cost curve depends on the current levels of the fixed inputs and hence on the status-quo output level. Figure 6.2 shows the short-run cost curve when the initial output level is 55 rather than 30. With the greater investment in fixed inputs, the short-run fixed cost is higher but the short-run variable cost is lower than with the smaller investment.

Figure 6.2



6.3 Law of diminishing return

Whereas long-run cost curves can have varied shapes and may exhibit economies or diseconomies of scale, a rule of thumb for short-run cost curves is that they exhibit increasing marginal cost. The reason is that, because some key inputs are fixed in the short run, the marginal product (the extra output per unit of additional input) of other inputs is decreasing. This is called the *law of diminishing return*.

6.4 Short-run vs. long-run price and output decisions

Sufficiency of marginal conditions

Consider the output decision by a firm, as outlined in Chapter 5 for a perfectly competitive firm or in the Preliminaries chapter and in Chapter 7 for a firm with market power. Suppose that the demand curve or market price shifts, leading to a new revenue curve $r_{\text{new}}(Q)$, and suppose that this shift is expected to be long-term. The firm adjusts its output level to Q_{short} in the short term in order to maximize $r_{\text{new}}(Q) - c_s(Q)$; it adjusts its output level to Q_{long} in the long term in order to maximize $r_{\text{new}}(Q) - c(Q)$. The purpose of this section is to make further comparisons between these two decisions.

Assume that marginal revenue is constant or decreasing. As a general rule, marginal conditions are sufficient in the short run, given the following considerations.

1. The short-run cost curve exhibits increasing marginal cost.
2. Although the short-run cost curve has a fixed cost, this fixed cost cannot be avoided by shutting down and hence is not relevant to the short-run output decision (it is "sunk" in the short run even though it can be modified in the long run).

For the sake of comparing the short-run and long-run decisions, assume that the long-run cost curve has no fixed cost and has increasing marginal cost, so that marginal conditions are also sufficient in the long run.

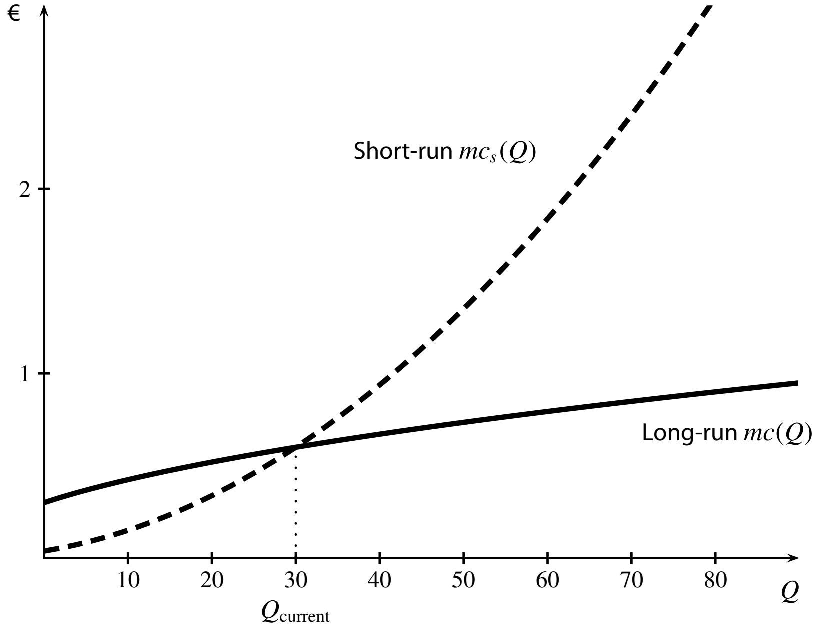
Comparison between short-run and long-run marginal cost

Because total cost is higher in the short run than in the long run except at Q_{current}, we have the following comparisons between short-run and long-run marginal costs.

1. For $Q > Q_{\text{current}}$, $mc_s(Q) > mc(Q)$. Because the firm is locked into a low level of the fixed input in the short run, it cannot increase output using the efficient input mix.
2. For $Q < Q_{\text{current}}$, $mc_s(Q) < mc(Q)$. The marginal cost of increasing output is also the marginal savings when reducing output. When the firm reduces output below Q_{current}, it saves less in the short run than in the long run because it cannot adjust all inputs.

Drawn on the same graph, the long-run and short-run marginal cost curves cross at Q_{current}. Below Q_{current}, the short-run curve lies below the long-run curve, whereas the opposite happens above Q_{current}. This can be seen in Figure 6.3, which shows the marginal cost curves that correspond to the total cost curves in Figure 6.1.[1]

Figure 6.3

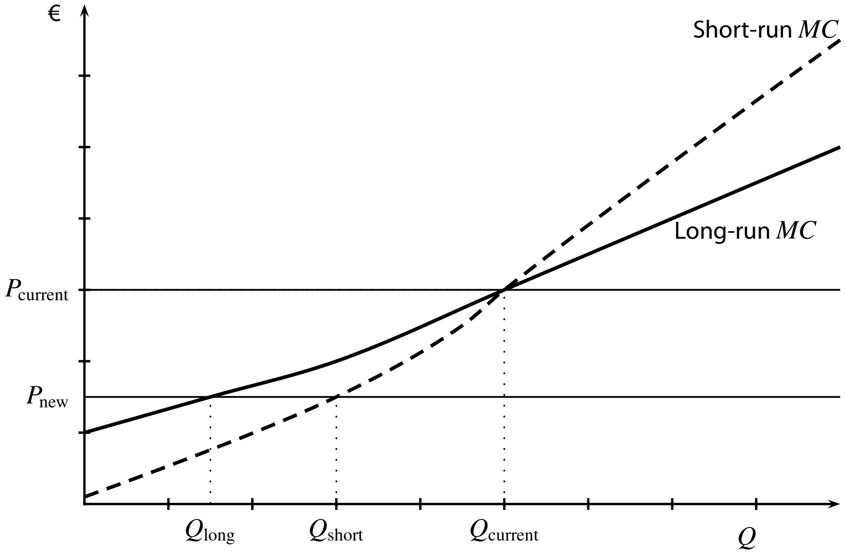


[1] These curves happen to intersect also at the origin, but this is not important.

Implications for short-run and long-run decisions

Because moving away from Q_{current} is more costly in the short run, *output is less responsive to price changes in the short run than in the long run.* We illustrate this in Figure 6.4.

Figure 6.4



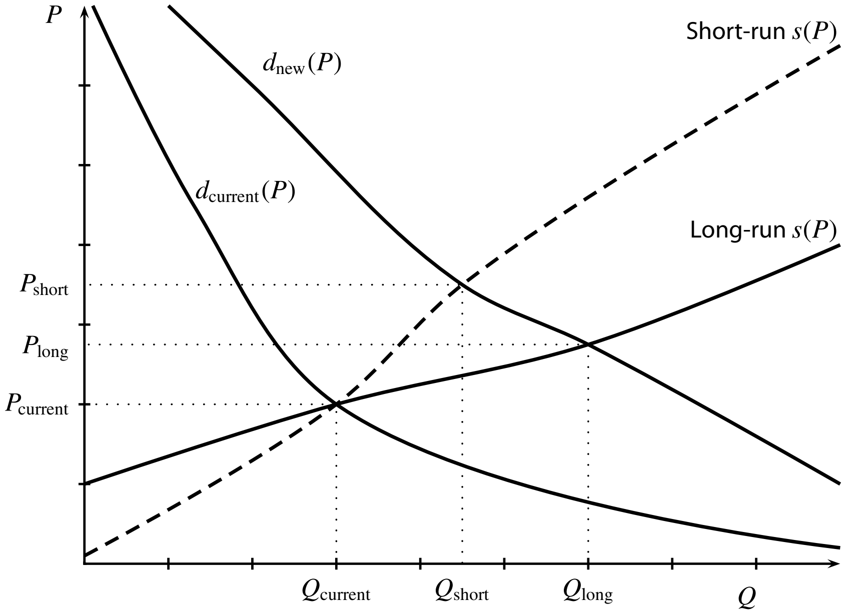
The short-run and long-run marginal cost curves intersect at Q_{current}. The short-run marginal cost curve is steeper than the long-run marginal cost curve. When the price falls from P_{current} to P_{new}, initially output drops only to Q_{short}, but in the long run it drops to Q_{long}.

Recall that the supply curve is the inverse of the marginal cost curve. We can thus think of the two upward-sloping marginal cost curves in Figure 6.4 as the short-run and the long-run supply curves. At their point of intersection, the short-run supply curve is steeper—and hence less price sensitive—than the long-run supply curve. Specifically, supply is less elastic in the short run than in the long run.

6.5 Short-run volatility in competitive industries

Just as a firm's individual supply is less elastic in the short run than in the long run, so is the aggregate supply in a competitive industry. Figure 6.5 provides an illustration of this fact and its consequences.

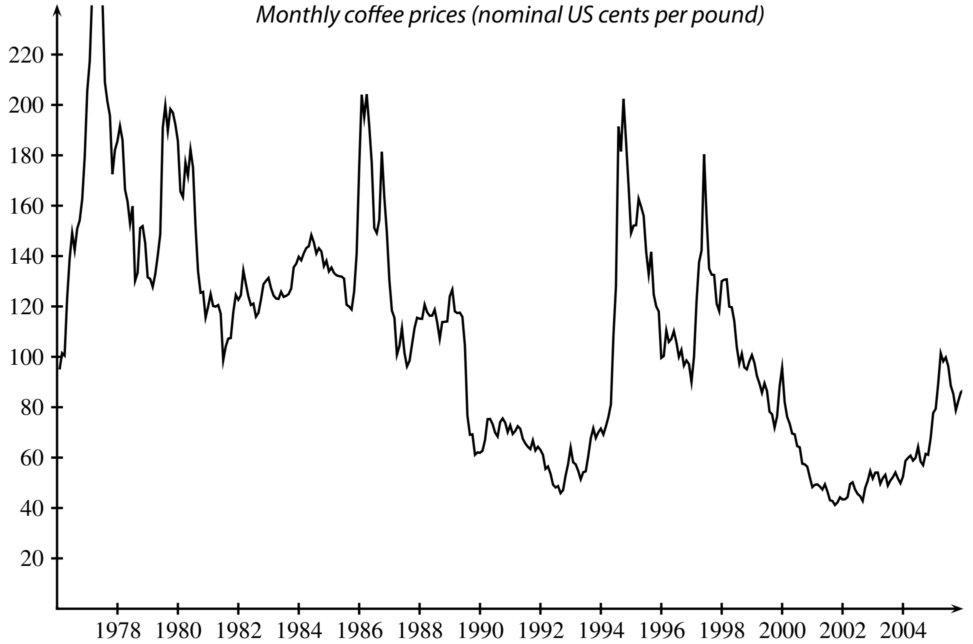
Figure 6.5



The graph shows the long-run supply curve and the current equilibrium price P_{current} and quantity Q_{current} given the current demand curve $d_{\text{current}}(P)$. The short-run supply curve, which reflects the current output levels of all the firms, is also shown. The demand curve then makes a long-term shift to $d_{\text{new}}(P)$. The short-run supply is not very elastic, so this shift has a large effect on price. The price rises to P_{short} and the supply increases only to Q_{short}. In the long run, however, as the firms adjust all the inputs, output increases to Q_{long} and the equilibrium price falls to P_{long}.

We thus see that market prices are more volatile in the short run than in the long run. This difference is particularly pronounced for industries in which there are crucial inputs that take a long time to adjust. For example, a coffee tree does not bear coffee beans until about five years after it is planted. In the meantime, yields can be increased only modestly through the use of more labor and chemicals. Short-run supply of coffee is therefore highly inelastic, and we see in Figure 6.6 that the price of coffee is quite volatile.

Figure 6.6



6.6 Overshooting

One of the apparent marvels of perfectly competitive markets is that firms and consumers need only know the market price, and then the "invisible hand" of market equilibrium is enough to coordinate production and consumption decisions efficiently. Firms need not know the market demand curve or the supply curves of the other firms.

However, this presumes that somehow an equilibrium is reached and then the market stays there. In fact, perpetual change is the rule; the static competitive equilibrium model is meant only to capture certain forces that operate in the long term. In a changing world, a firm must make plans for the future that require more information than current prices. At the very least the firm must forecast future prices, which requires predicting consumer demand as well as the supply decisions of other firms.

Consider again the scenario illustrated in Figure 6.5. Following the shift in demand, the market price increases to P_{short} as the firms adjust their use of variable inputs. In the meantime, the firms also must make long-term production plans and must adjust their use of fixed inputs. To forecast correctly that the long-term equilibrium price will be P_{long}, the firms must know the new demand curve and the long-run market supply curve.

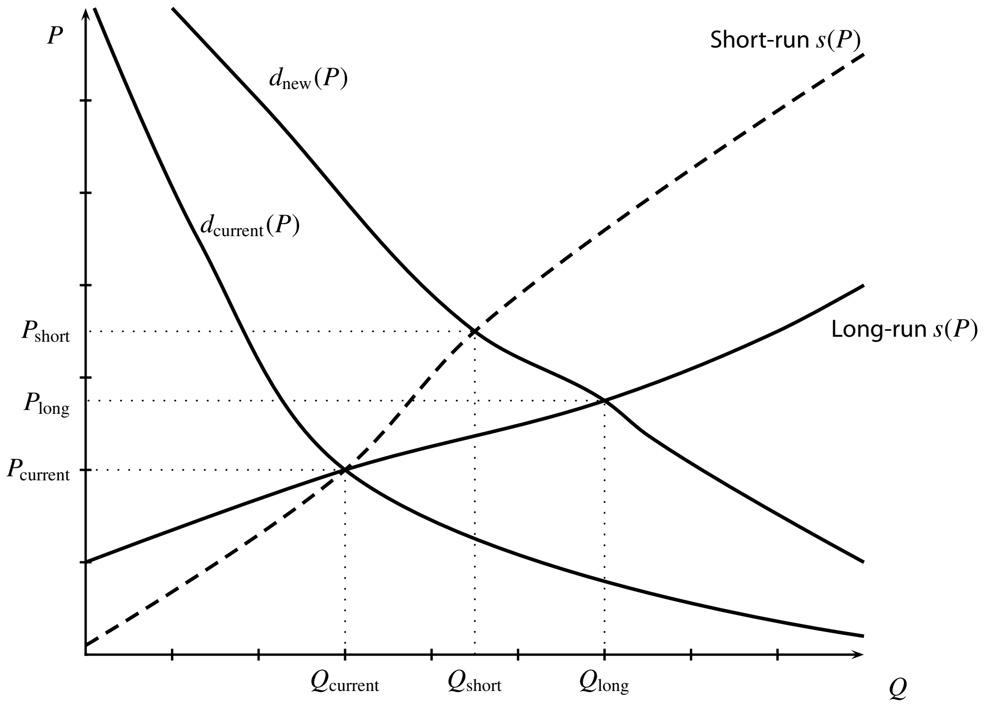
One mistake a firm could make is to assume that the price P_{short} that follows the shift in demand from $d_{\text{current}}(P)$ to $d_{\text{new}}(P)$ will persist indefinitely. Such a firm will expand output more than it should and regret its investment decision when the price ends up at P_{long} instead of P_{short}. The firm has forgotten that all the other firms also face fixed input constraints in the short run but will expand output further in the long run.

If all the firms make this mistake, then industry volatility is much worse. An initial increase in demand that pushes the price up in the short run can generate so much overinvestment that the price then falls to below its initial level. This is called “overshooting”.

Exercise 6.1. This exercise asks you to provide a graphical illustration of overshooting. It will test your understanding of the distinction between short-run and long-run supply.

Figure E6.1 shows the same initial scenario as Figure 6.5. The current equilibrium quantity and price are Q_{current} and P_{current}. Then the demand curve shifts from $d_{\text{current}}(P)$ to $d_{\text{new}}(P)$, causing the price to rise in the short run to P_{short}.

Figure E6.1



a. Suppose that all the firms believe the market price is going to stay at P_{short} for the long term, so they initiate investments and adjustments to production processes and inputs in order to maximize profit given P_{short}. Mark on the graph the intended long-run total output of the firms.

b. After all these changes and investments are in place, the firms have new short-run cost curves that determine the new short-run supply curve. Draw in a plausible short-run supply curve.

c. The new short-run equilibrium must be at the intersection of the new short-run supply curve and the new demand curve. Mark on the graph the short-run output and price levels.

d. Is the market now in a long-run equilibrium? Specifically, compare the resulting price to that which obtains in the long run when the firms do not overshoot. Do the firms regret their investments?

6.7 A capacity-constraint model

Overview

Here is a stark but simple and intuitive model of the distinction between the short run and the long run.

The model is motivated by industries in which an input, fixed in the short run, determines a capacity beyond which it is nearly impossible to produce. For example, once a supertanker is filled to capacity, it is difficult to transport more oil without buying another or a larger supertanker. Given a fixed supertanker capacity, the marginal variable costs (mainly, pumping costs plus fuel costs that vary with the weight of the cargo) are flat until the capacity is reached, and then they rise very steeply (the only way to increase output is to run the vessel more quickly). Other examples include paper mills, trains, coffee growing, most assembly-line industries, theaters, and stadiums.

To develop a simple model of such a technology, we assume that (a) the marginal cost of capacity is constant; and (b) the marginal variable cost is constant up to capacity, and it is impossible to exceed capacity.

One of the nice things about this capacity constraint model is that it is easy to obtain the required data. It is hard to know marginal cost, but it is easy to measure average cost (at year's end, take your total cost and divide by your total output). When marginal cost is constant, it equals average cost and hence is also easy to measure.

Cost structure of a single firm

Let's fix a numerical example to work with. Consider a movie theater. The short-run fixed costs of capacity are those that do not change no matter how many people actually show up at the theater. These costs are mainly the land, building, and seats in the theater as well as a share of the heating, lighting, and air conditioning. Suppose these costs sum to €3 per seat. The short-run variable costs are those that scale up as more people enter the theater. These are mainly the costs of labor and supplies for selling tickets and cleaning up. Suppose these costs sum to €1 per ticket sold.

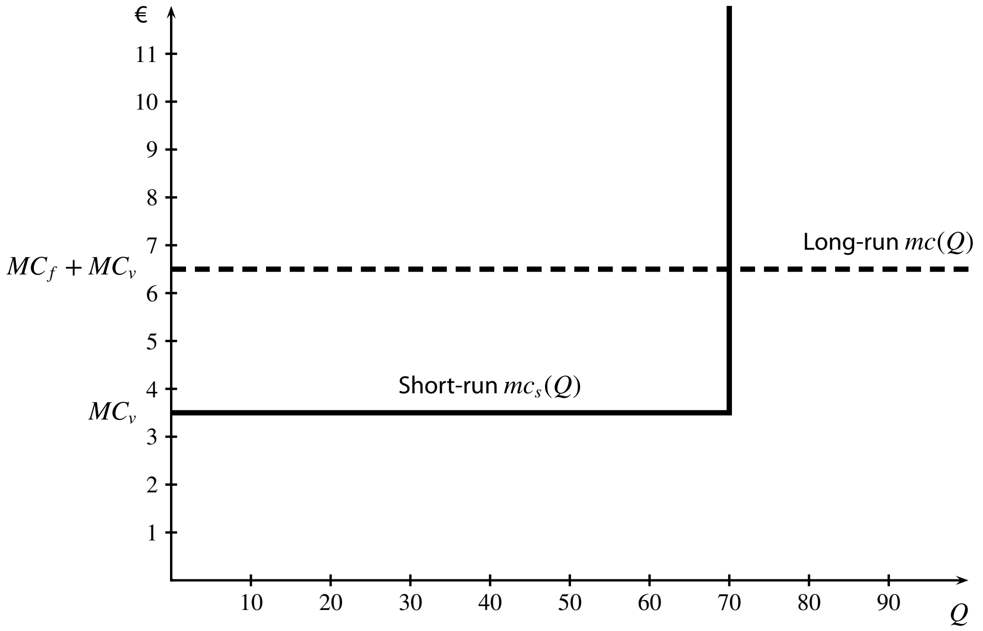
We are still missing an important cost: that of the movies themselves. Is this a capacity cost or a short-run variable cost? The answer depends on the type of contract between the theater and the movie distributor. In one form of contract, the cost is proportional to the capacity of the theater; then the cost is a capacity cost and is fixed in the short run. However, it is now standard practice that the cost is proportional to the number of tickets sold; then it is a variable cost in the short run. Suppose that this is the case and that it equals €2.50 per ticket.

In sum, the cost of capacity is €3 per seat; denote this by MC_f. The short-run variable cost is €3.50 per ticket; denote this by MC_v.

A movie theater faces demand that fluctuates during the week and across times of day and that is uncertain from one day to the next. Thus, even if the market remains otherwise perfectly stable, the firm does not sell out every showing every day. Modeling this kind of stochastic demand is beyond the scope of this book, so let's assume that demand is the same each day and at every showing *except* that occasionally the demand curve may shift. Therefore, when the market is in long-run equilibrium, the firm operates at capacity.

The long-run marginal cost per seat or ticket (given that the theater operates at capacity) is the sum $MC_f + MC_v = 6.50$. There is no long-run fixed cost. Thus, the supply curve of the firm is perfectly elastic at the price $P = MC_f + MC_v$. Figure 6.7 shows this long-run supply curve (the long-run marginal cost curve) as the dashed flat line at $P = €6.50$.

Figure 6.7



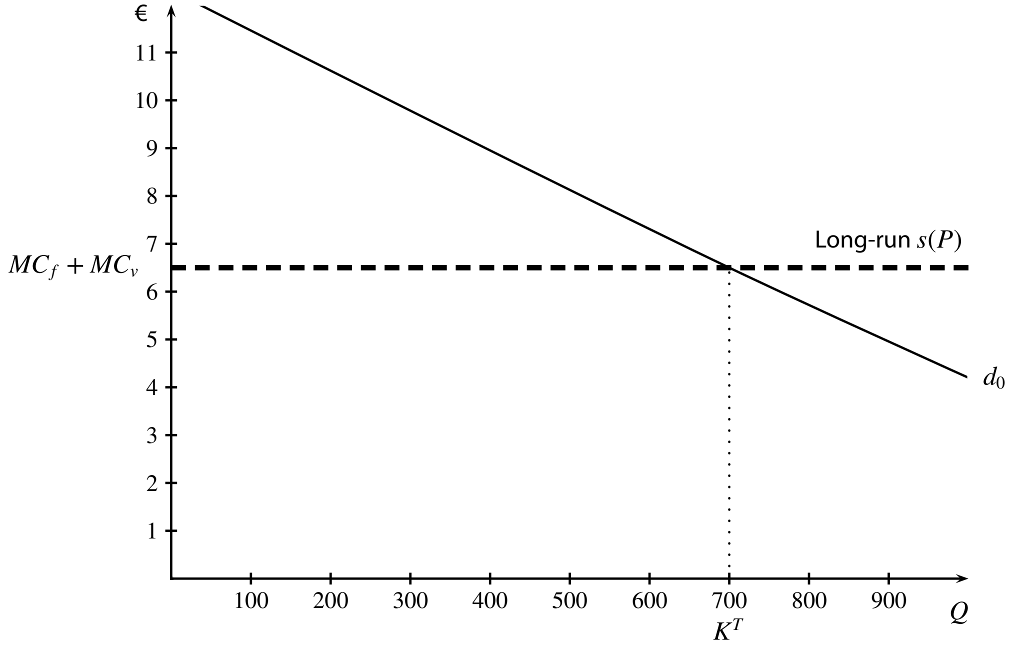
Suppose that the firm has set up a capacity of $K = 70$. For quantities between 0 and 70, the firm's short-run marginal cost is $MC_v = 3.50$. This is shown as the horizontal segment of the solid line in Figure 6.7. It is impossible to produce beyond 70 in the short run; we represent this graphically by extending the short-run marginal cost curve as a vertical line going off to infinity at $Q = 70$.

This short-run marginal cost curve is also the firm's short-run supply curve, as follows. Provided the price exceeds the short-run marginal cost, the firm chooses to fill the theater (i.e., to produce at capacity). Thus, for any $P > 3.50$, its short-run supply is 70; this is the vertical part of the short-run supply curve. When $P < 3.50$, the price does not even cover the short-run variable cost of each person who comes through the door and hence the theater shuts down. When $P = 3.50$, the price just covers the variable costs and hence the firm is indifferent between how many tickets it sells (between 0 and 70); this is the horizontal part of the short-run supply curve.

Market supply and equilibrium

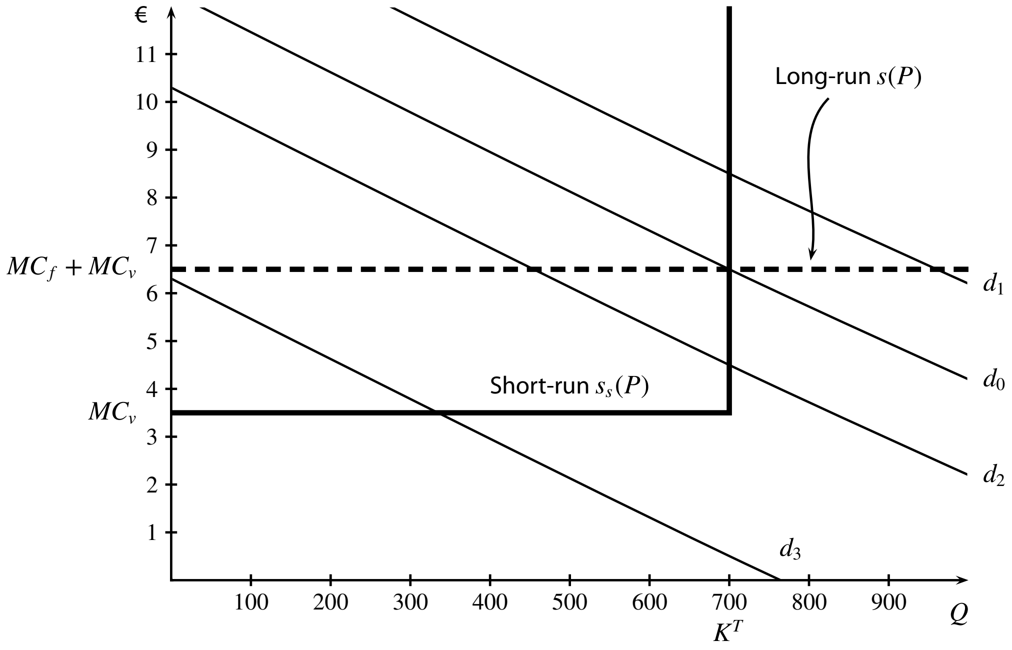
Suppose that all the firms in the market have the same cost structure. Then we know that long-run supply is perfectly elastic at the long-run marginal cost $MC_f + MC_v$, so this marginal cost must be the equilibrium price. The demand curve then determines how much is produced and traded at this price. Figure 6.8 shows the long-run supply curve and a demand curve d_0. The long-run equilibrium quantity—that is, the total capacity of the firms—is $K^T = 700$.

Figure 6.8



Let see what happens in the short run if the demand curve shifts. Figure 6.9 shows various other demand curves: d_1, d_2, and d_3. The total installed capacity is $K^T = 700$. Until the price falls below $MC_v = 3.50$, then all theaters operate at capacity; supply is perfectly inelastic at $Q = 700$. If the price falls below $MC_v = 3.50$, all firms would shut down because they would lose money on every patron. If the price is MC_v, theaters exactly cover their short-run variable costs and thus do not care how many tickets they sell; collectively, they are willing to supply any amount between 0 and 700. Hence, the short-run supply curve is the solid line in Figure 6.9.

Figure 6.9



To avoid cluttering the figure, we have not marked the short-run and long-run equilibria on the graph, but you should be able to easily visualize them as the intersection of the new demand curve with the short-run and long-run supply curves, respectively.

If demand shifts up to d_1, theaters cannot offer more tickets because they are already at capacity. All that happens is that the price rises until demand falls back to the capacity of the theaters and again equals the available supply.

If demand shifts down to d_2, the price falls but not below MC_v. All the theaters still want to operate at capacity because the price is above MC_v. The price falls enough that demand again equals total capacity.

Suppose demand shifts down to d_3. Even when the price falls to MC_v, demand is less than the available capacity. Now theaters begin to operate below capacity; some may close. But the price does not go below MC_v, for if it did then all the theaters would close and demand would exceed supply.

Exercise 6.2. Consider the numerical example in this section. Suppose that the market starts out in long-run equilibrium with the demand curve d_0, so that capacity is 700. Then a tax of τ = €2 per unit is imposed.

a. First model the new long-run equilibrium, treating the tax as a shift in the demand curve. What is the new long-run equilibrium price? Draw a graph showing the long-run supply curve, the initial and shifted demand curves, and the current and new long-run equilibria.

b. Next show what happens in the short run. What is the short-run price and quantity? Draw a graph showing the short-run supply curve and illustrate the equilibrium before and after the tax.

Short-run supply with heterogeneous firms

Some markets have many heterogeneous firms or plants, each of which approximately fits the capacity-constraint model. Because excess capacity can be difficult to dismantle or depreciate, the "short run" can last for a long time when there is excess capacity. As the price varies, different plants shut down and come back on line—when the price crosses the threshold of their short-run variable costs. This determines an upward-sloping short-run supply curve that can be derived from easily obtained data.

For example, suppose that there are 7 plants that the capacities and short-run variable costs listed in Table 6.1. (We have ordered the plants from lowest to highest variable cost.)

Table 6.1

Plant	Capacity	MC_v
1	1200	35
2	800	38
3	1400	43
4	1000	45
5	2200	49
6	1300	52
7	600	59

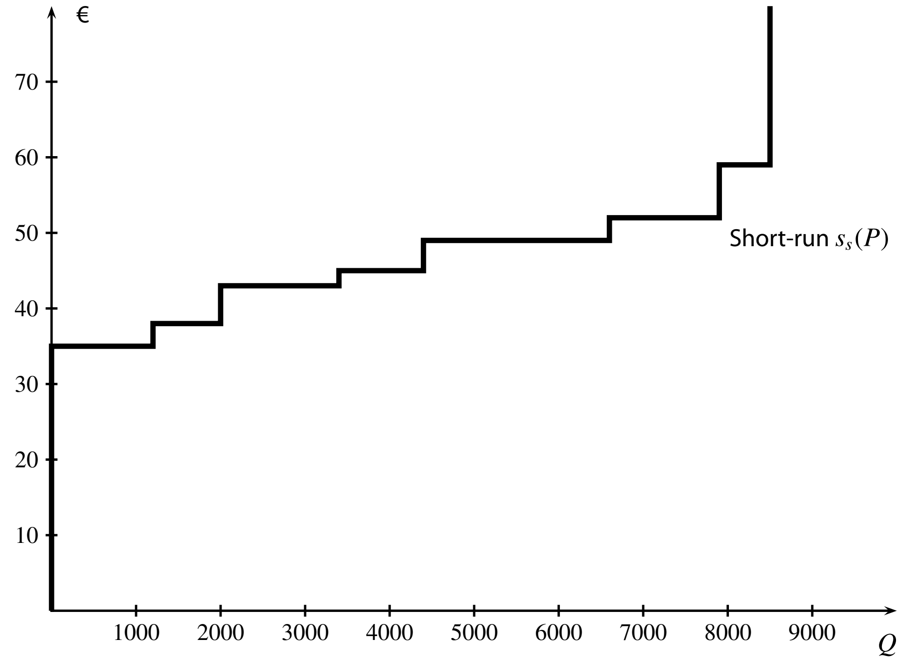
Consider the short-run supply decisions. If the price falls below 35 then no plant operates; supply is zero. When the price rises above 35, plant 1 comes on line; the supply is 1200. When the price rises further and passes 38, plant 2 starts up; total supply is 2000. Table 6.2 lists the threshold prices and the total capacity that comes on line at these prices; it describes the short-run supply curve.

Table 6.2

Thresholds	Capacity of plant that comes on line	Total supply
35	1200	1200
38	800	2000
43	1400	3400
45	1000	4400
49	2200	6600
52	1300	7900
59	600	8500

That short-run supply curve is graphed in Figure 6.10.

Figure 6.10



6.8 Wrap-up

In the short run, a firm cannot adjust certain inputs. Therefore, its short-run cost for a change in output is higher than its long-run cost. The short-run cost curve is characterized by (a) a fixed cost of the fixed inputs (which cannot be eliminated even by shutting down and hence is irrelevant to short-run decisions) and (b) increasing marginal cost.

Furthermore, the short-run marginal cost curve is steeper than the long-run marginal cost curve. As a consequence, after a change in market conditions, output is less responsive (but price is more volatile) in the short run than in the long run. A competitive firm that does not anticipate this difference in price volatility is likely to make incorrect investment decisions.

Additional exercises

Exercise 6.3. This is a numerical example of the comparison between short-run and long-run costs. The production function is $f(L, M) = L^{1/2}M^{1/2}$, where "M" = machines and "L" = labor. Suppose that $P_L = P_M = 1$.

a. The cost-minimizing long-run input mix given this production function and these prices is to have equal amounts of L and M. Derive from this information the cost function. (See how many units of L and M you would need to produce Q units; then $c(Q)$ is the cost of these inputs at the prices $P_L = P_M = 1$.)

b. What are the *AC* and *MC* curves? Are there (dis)economies of scale?

c. Suppose that initially $Q_{\text{current}} = 4$; then there is a shift in demand (or some other change that makes the firm want to adjust its production). In the short run, M is a fixed input and L is a variable input. Determine how many units of M are employed initially. What is the short-run *FC*?

d. Given the fixed amount of M just determined, derive output as a function of L.

e. Invert this function in order to find the amount of L needed to produce Q units.

f. The variable cost is the cost of the labor. Derive variable cost as a function of Q. (This is so trivial you may wonder if you got the right answer.)

g. Add the short-run fixed cost and the variable cost to obtain the short-run total cost curve.

h. Graph the short-run and long-run cost curves on the same axis.

i. What is the short-run marginal cost curve?

j. Graph the short-run and long-run marginal cost curves on the same axis.

k. Suppose, for example, that output is expanded from $Q = 4$ to $Q = 8$. What is the total cost in the short run? What is the total cost in the long run?

Chapter 7

Pricing with Market Power

7.1 Motives and objectives

Broadly

Much of this book is about how firms should decide what prices to charge and how much to sell of their goods. Firms do not make these decisions in isolation. The valuations of one firm's customers for that firm's goods depend on the prices that other firms charge for complementary and substitute goods. In this book, we examine pricing decisions with several different perspectives on the interaction between firms.

Monopoly (Chapters 7, 8, 17, 9, 10) The monopoly model takes as given the pricing decisions of other firms, and hence it has low emphasis on interaction between firms.

Perfect competition (Chapters 5 and 6) Perfect competition is an approximation for markets in which there are several or many firms producing close substitutes, so that their pricing decisions are tightly related. However, each firm is a small player in the market and hence does not unilaterally have a large effect on other firms or on the market price.

Oligopoly (Chapters 12–15) Oligopoly models consider the strategic interaction between a small number of firms. The emphasis is on how firms react to and anticipate the reactions of other firms.

The common labels "monopoly", "perfect competition", and "oligopoly" misleadingly suggest that only one model is relevant for each firm, depending on its degree of market dominance. In fact, all three perspectives are relevant to all firms. What distinguishes each model are the aspects of markets and pricing that the model focuses on.

"Monopoly" is a particularly restrictive term that suggests the topic is relevant only to the Microsofts and local cable television monopolies of the world. However, the model merely takes as given the pricing decisions of other firms and assumes that the firm has some market power. Market power means that the products of other firms are not perfect substitutes. For example, a cereal maker produces flavors and types of cereal that are slightly different from those of other manufacturers; each car model has special features, mechanical quality, and styling that differentiate it from other models; a gas station is closer to certain customers than is any other gas station; and even memory chip producers have different reputations for quality. Thus, market power is the rule, not the exception. However slight it may be, successful firms know how to exploit it. Thus, "pricing with market power" is a more accurate title for this topic. We also call it "pricing" for short.

More specifically

The chapters on pricing with market power differ with respect to the kinds of pricing strategies firms use. We begin, in Chapters 7 and 8, with the classic case of uniform pricing. The firm faces a standard demand curve of the kind we studied in Chapter 3. The firm charges the same price to all customers and allows them to purchase any amounts at the posted price. We show how to calculate the profit-maximizing quantity and price and by solving marginal conditions. We give formulas for the firm's profit-maximizing price for two special cases: linear demand and exponential demand (and constant marginal cost in both cases).

As a benchmark, we characterize the quantity of sales that maximizes the total gains from trade. To sell this efficient quantity, the firm must charge a price that leaves the customers with a substantial share of these gains. The firm faces the following trade-off. By decreasing sales and increasing price, the firm reduces the total available gains from trade but also appropriates some of the consumer surplus. This trade-off leads the firm to choose an output level that is below the efficient one, thereby causing a loss of total surplus known as a deadweight loss.

Then we revisit both the profit-maximizing and the efficient solutions when there is a (long-run) fixed cost. The fixed cost affects only the firm's entry and exit decisions.

7.2 From cost and demand to revenue and profit

In this chapter, you are the manager of a firm that produces a single good. Your cost function is $c(Q)$ and the demand curve of your branded product is $d(P)$. You charge a price P and sell Q units, choosing P and Q to maximize your profit: revenue minus cost.

Price and quantity are not independent decision variables; rather, they are tied to each other by your demand curve. There are two equivalent ways to view this decision problem. You can choose price P and then let the quantity you sell be given by $Q = d(P)$. Most managers frame their decision problems this way. However, you can instead choose Q and let your price be whatever you must charge in order to sell that amount. This price is given by the inverse of your demand curve, $P = p(Q)$.

We adopt the second perspective, with quantity as the choice variable, because it has several analytic advantages.

1. We get a nice decomposition between revenue and cost. We can write profit as $\pi(Q) = r(Q)-c(Q)$, where revenue $r(Q) = p(Q)\times Q$ depends only on the properties of demand (ask your marketing manager for this information) and your cost $c(Q)$ depends only on your cost of production (ask your production manager for this information).[1]

[1] In contrast, cost as a function of price depends in a complicated way on both your demand and your cost

2. A competitive firm is a price taker and chooses only the quantity to sell. By treating quantity as the choice variable when the firm has market power, we can better see the relationship between these two models.
3. We can also better see the relationship between the profit-maximizing solution and the efficient benchmark. Efficiency is a question of the quantity of output, not of prices, which only determine the distribution of surplus rather than how much surplus is generated. The fact that the inverse demand curve and the marginal valuation curve are the same facilitates the comparison.

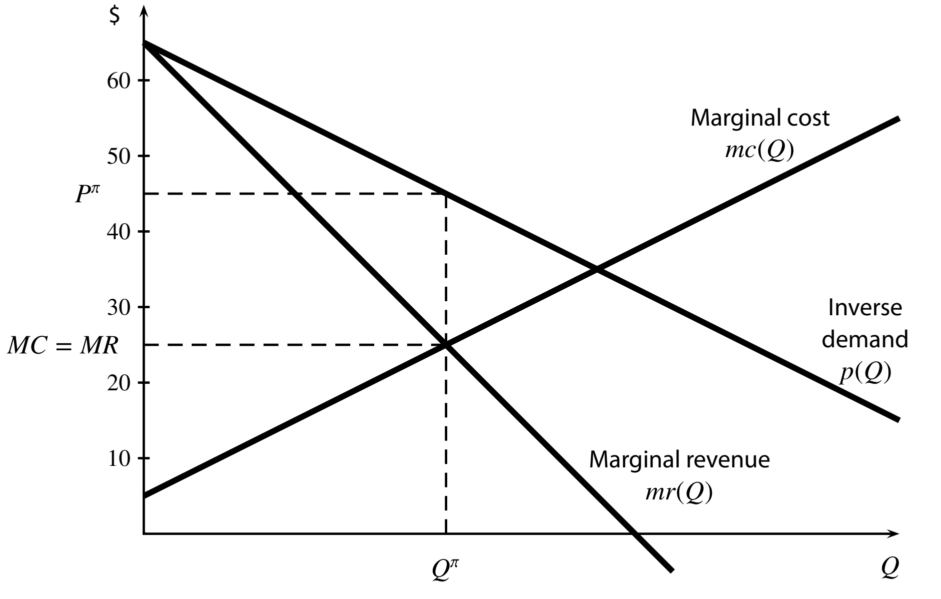
7.3 Profit-maximizing output level

Marginal conditions for profit maximization

The marginal condition for maximizing $\pi(Q) = r(Q) - c(Q)$ is that $mr(Q) = mc(Q)$. We can illustrate the solution as the intersection between the marginal revenue and marginal cost curves.

Figure 7.1 shows a demand curve, the marginal revenue curve, and a marginal cost curve. The profit-maximizing quantity, at the intersection of the marginal revenue and marginal cost curves (i.e., where marginal revenue equals marginal cost) is $Q^\pi = 40$. We look to the inverse demand curve to find the price that generates demand equal to 40. This is $p(40) = 45$, as shown in the graph.

Figure 7.1



curves; you first have to determine demand as a function of price and then find the cost of producing that level of output.

A numerical example

We now calculate the profit-maximizing solution for the data in Figure 7.1:

$$\begin{aligned} d(P) &= 130 - 2P\,, \\ c(Q) &= 5Q + Q^2/4\,. \end{aligned}$$

We follow a six-step recipe.

1. *Find the inverse demand curve.* We solve $Q = 130 - 2P$ for Q, which yields
$$P = 65 - Q/2\,.$$
2. *Find the revenue curve.* Revenue is
$$\begin{aligned} r(Q) &= p(Q) \times Q\,, \\ &= (65 - Q/2)Q\,, \\ &= 65Q - Q^2/2\,. \end{aligned}$$
3. *Find the marginal revenue curve.* It is the derivative of the revenue curve:
$$mr(Q) = 65 - Q\,.$$
4. *Find the marginal cost curve.* It is the derivative of the cost curve:
$$mc(Q) = 5 + Q/2\,.$$
5. *Solve the marginal condition for the profit-maximizing quantity.* We solve
$$\begin{aligned} mr(Q) &= mc(Q)\,, \\ 65 - Q &= 5 + Q/2\,, \\ 60 &= 3Q/2\,, \\ 40 &= Q\,. \end{aligned}$$
Thus, the profit-maximizing output is $Q^\pi = 40$.
6. *Find the price from the inverse demand curve.* Given output Q^π, the firm should charge $p(Q^\pi)$, that is,
$$P^\pi = p(40) = 65 - (40/2) = 45\,.$$

Pricing with linear demand and constant marginal cost

For linear demand $d(P) = A - BP$ and constant MC, the profit-maximizing price is

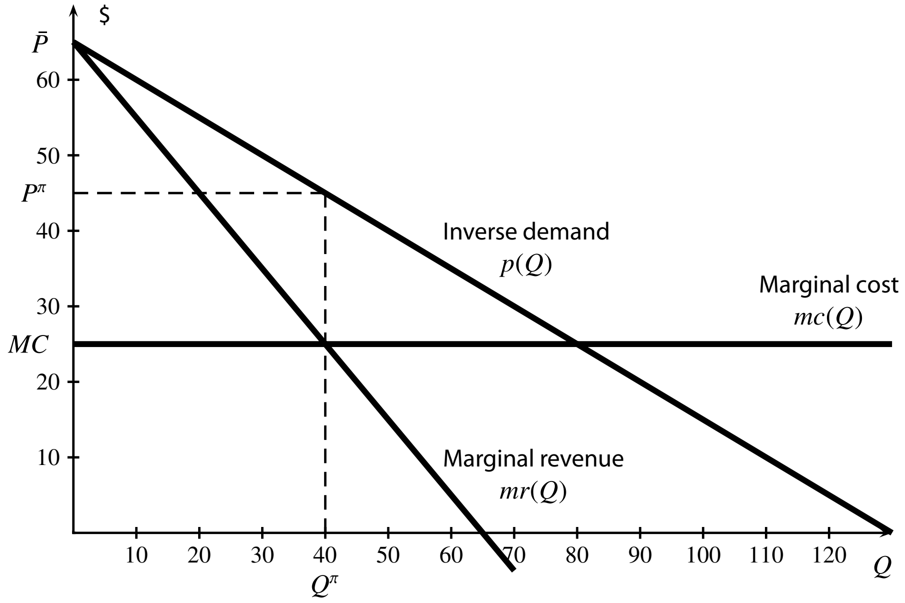
$$P = \frac{MC + \bar{P}}{2}, \tag{7.1}$$

where $\bar{P} = A/B$ is the choke price. This is the *midpoint pricing rule*: the price is the midpoint between the marginal cost and the choke price. You can derive this formula by applying the recipe in the numerical example to the general form of the demand curve.

Suppose that the demand curve is $d(P) = 130 - 2P$, and so the choke price is $\bar{P} = 65$. Suppose that $MC = 25$. Then the midpoint pricing rule yields $P = (25 + 65)/2 = 45$.

Let's check the formula by calculating the optimal price the long way. Recall that $mr(Q) = 65 - Q$ for this demand curve. The $MR = MC$ condition is $65 - Q = 25$, or $Q = 40$. The price is $p(40) = 45$—we get the same answer. See Figure 7.2.

Figure 7.2



Although linear demand and constant marginal cost are idealized conditions, the midpoint pricing rule can be used as a rough approximation even when they do not hold. Furthermore, the rule will come in handy in this book for numerical examples that illustrate our qualitative conclusions, since it saves us from repetitive $MR = MC$ calculations for each example. But watch out: *only apply the midpoint pricing rule when the demand curve is linear and marginal cost is constant*; otherwise, the formula is not valid.

Exercise 7.1. Suppose you produce minivans at a constant marginal cost of \$15K and your demand curve is $d(P) = 16 - 0.6P$. (Price is measured in 1000s of dollars and quantity is measured in 100,000s of minivans.) Find your optimal price and quantity.

Pricing with exponential demand and constant marginal cost

There is also a simple pricing formula for exponential demand with constant MC. Write the demand curve in the form $Q = AP^{-B}$. Then the profit-maximizing price is

$$P = \frac{B}{B-1}MC.$$

This formula is only valid if $B > 1$. We will derive it in Chapter 8; in the meantime, the next exercise puts it to use.

Exercise 7.2. What happens if a per-unit tax is imposed on a monopolist's product? One answer might be: "Since the monopolist can charge whatever it wants, it will pass the tax on to the consumer." But this does not say much, since any firm can charge whatever price it wants. The question is: *What price does it want to charge?* Let's see if we can provide a more informative answer.

When faced with a question like this, one strategy is to start by working out a few simple examples. This at least illustrates some possibilities. Find out how a per-unit tax changes the price of (a) a firm with constant marginal cost and linear demand and (b) a firm with constant marginal cost and exponential demand. You can use the formulas we just presented. You should treat the tax like a per-unit cost borne by the firm. If the firm's marginal cost is MC and the tax is τ, then the firm's marginal cost after the tax is $MC + \tau$. For the case of exponential demand, use a particular value of $B > 1$ such as 2 or 3.

7.4 Profit maximization versus social efficiency

Motivation

Let's review why even the manager of a profit-maximizing firm should be interested in understanding the socially efficient solution.

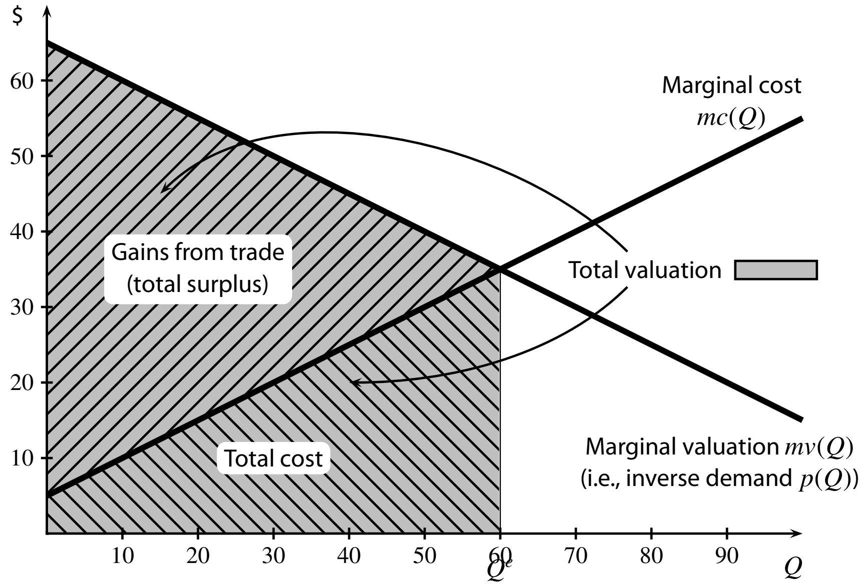
1. It helps her understand the profit-maximizing solution.
2. If the profit-maximizing solution is not efficient (as will happen in this case), then there are lost gains from trade that a more clever approach might generate and extract for the firm.
3. The manager will also be on the buying side of the market, whether on the job (since her firm buys inputs) or off the job (as a consumer).
4. It is important to understand the motives and arguments of the regulatory authorities that intervene in markets to make them more efficient.

Socially efficient solution

So let's suppose that—rather than maximizing your own profit—you benevolently maximize the total surplus of your customers and your firm. Let $v(Q)$ be the total valuation of the customers when Q units are efficiently distributed among them, and let $mv(Q)$ be the marginal valuation curve.

The total gains from trade are $v(Q) - c(Q)$, just as when we studied efficiency in Chapter 1. The difference is that there may be many customers, each of whom may buy more than one unit, and there is one firm that sells more than one unit. Yet the analysis is otherwise the same. The marginal condition for maximizing the gains from trade is $mv(Q) = mc(Q)$: you produce up to the point where an extra unit would cost the firm more than the customers are willing to pay for it. Thus, the efficient quantity Q^e is where the marginal valuation and marginal cost curves intersect. This is illustrated in Figure 7.3.

Figure 7.3



We recall the following facts.

1. The marginal valuation curve is the inverse of the demand curve: $mv(Q) = p(Q)$.
2. $v(Q)$ is the area under the marginal valuation curve—and hence under the demand curve—up through Q. Thus, $v(Q^e)$ is the area of the shaded region in Figure 7.3.
3. With no fixed cost, $c(Q)$ is the area under the marginal cost curve up to Q, so that $c(Q^e)$ is the area of the region marked "total cost".
4. Total gains from trade are $v(Q) - c(Q)$ and thus equal the area of the region labeled "gains from trade".

A numerical example

In Figure 7.3, the demand curve is $d(P) = 130 - 2P$ and the cost curve is $c(Q) = 5Q + Q^2/4$. These are the same as in our numerical example of the profit-maximizing quantity. For comparison, we calculate the efficient quantity in three steps.

1. *Find the marginal valuation curve.* The *MV* curve is the inverse demand curve:

$$mv(Q) = 65 - Q/2\,.$$

2. *Find the marginal cost curve.* The *MC* curve is the derivative of the cost curve:

$$mc(Q) = 5 + Q/2\,.$$

3. *Solve the marginal condition for Q.* We solve

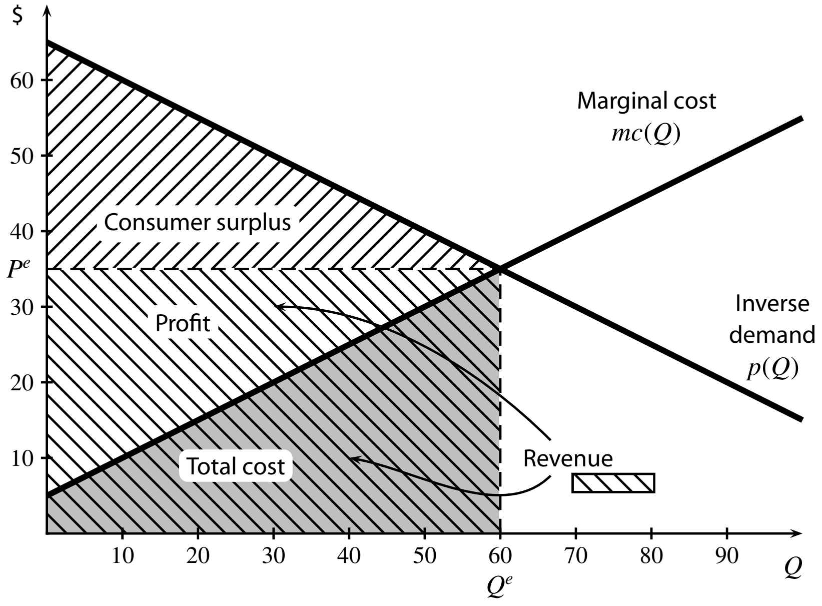
$$\begin{aligned} mv(Q) &= mc(Q)\,, \\ 65 - Q/2 &= 5 + Q/2\,, \\ 60 &= Q\,. \end{aligned}$$

Thus, the efficient quantity is $Q^e = 60$, as illustrated in Figure 7.3.

Implementing the efficient solution by setting a price

To implement the efficient solution, the firm could calculate Q^e and then charge the price $p(Q^e)$ such that the consumers demand Q^e units. In this case, the firm would charge $p(60) = 65 - (60/2) = 35$. The division of the total gains from trade between consumer surplus and profit is illustrated in Figure 7.4.

Figure 7.4



A shortcut when marginal cost is constant

This price the firm charges to implement the efficient solution is equal to its marginal cost at that solution:

1. it charges $p(Q^e)$;
2. since the inverse demand and marginal valuations are the same, $p(Q^e) = mv(Q^e)$;
3. the marginal condition for efficiency is $mv(Q^e) = mc(Q^e)$.

Thus, economists refer to the efficient solution as "marginal cost pricing".

If marginal cost is constant and equal to MC, then we know—even without calculating the efficient quantity—that the price should be MC. Therefore, the efficient quantity is whatever consumers demand at that price: $Q^e = d(MC)$. We do not need to follow the steps listed in the numerical example.

Take the demand curve $Q = 130 - 2P$. If marginal cost is constant and equal to 25, then the efficient quantity is

$$d(25) = 130 - (2 \times 25) = 80.$$

The profit-maximizing quantity is lower than the efficient quantity

Comparing Figures 7.3 and 7.1, we see that the profit-maximizing quantity is lower than the efficient quantity. Graphically, the marginal revenue curve lies *below* the inverse demand (marginal valuation) curve, so the marginal cost curve intersects the marginal revenue curve at a lower quantity than where it intersects the marginal valuation curve.

Here is why the marginal revenue curve lies below the inverse demand curve—that is, why, *at any point on the demand curve, marginal revenue is less than the price*. If you could sell any quantity you wanted at a price P, then each unit sold would increase your revenue by P. However, in practice you must lower your price in order to sell more. The extra revenue from selling one more unit is therefore less than the price P at which this unit is sold: the drop in price reduces your revenue on all the other units.

If the good is perfectly divisible, then from $r(Q) = p(Q) \times Q$ we can derive[2]

$$MR = P + \frac{dP}{dQ}\,Q. \tag{7.2}$$

Here dP/dQ is the amount by which the price changes per unit increase in output; hence $(dP/dQ) \times Q$ is the effect on revenue of the decrease in price. Because $dP/dQ < 0$, this effect is negative, and so again we obtain $MR < P$.

[2] Equation (7.2) sometimes causes confusion because of the roles of P and Q. Note that MR always means "extra revenue per unit increase in Q". Equation (7.2) expresses marginal revenue as a function of the initial point (P, Q) on the demand curve.

The trade-off between total gains from trade and consumer surplus

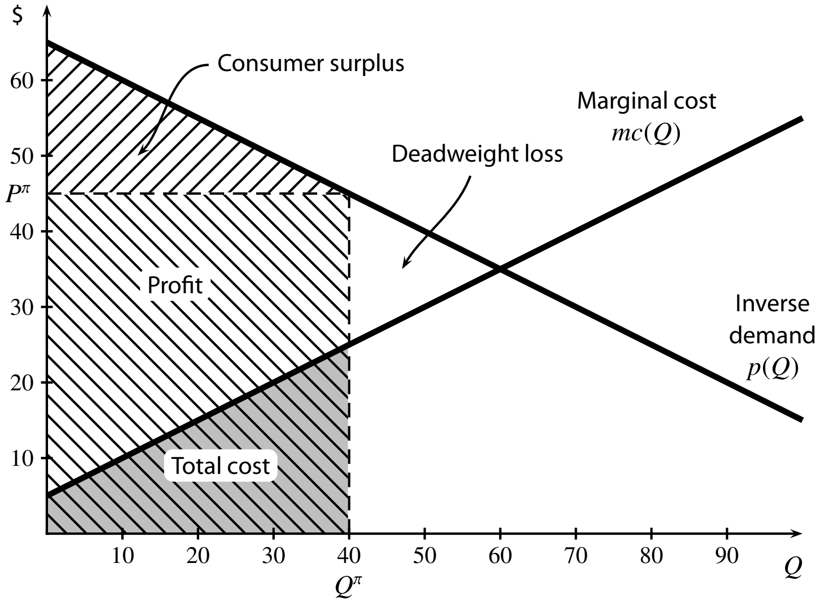
The firm chooses an inefficiently low quantity because it does not appropriate all the gains from trade.

Compare the distribution of gains from trade with efficient marginal-cost pricing, as in Figure 7.4, with the gains from trade and their distribution when you reduce output to 40 and increase the price to 45, as in Figure 7.5. Such a price increase involves the following trade-off:

1. *Total gains from trade go down.* Because the output level is below the efficient level, some potential gains from trade are not realized. The size of the loss in total surplus is the area of the triangle marked "Deadweight loss".
2. *Consumer surplus goes down.* The customers purchase less and at a higher price.

Profit is equal to total surplus minus consumer surplus. Thus, the decrease in gains from trade is bad for your firm yet the decrease in consumer surplus is good for your firm.

Figure 7.5



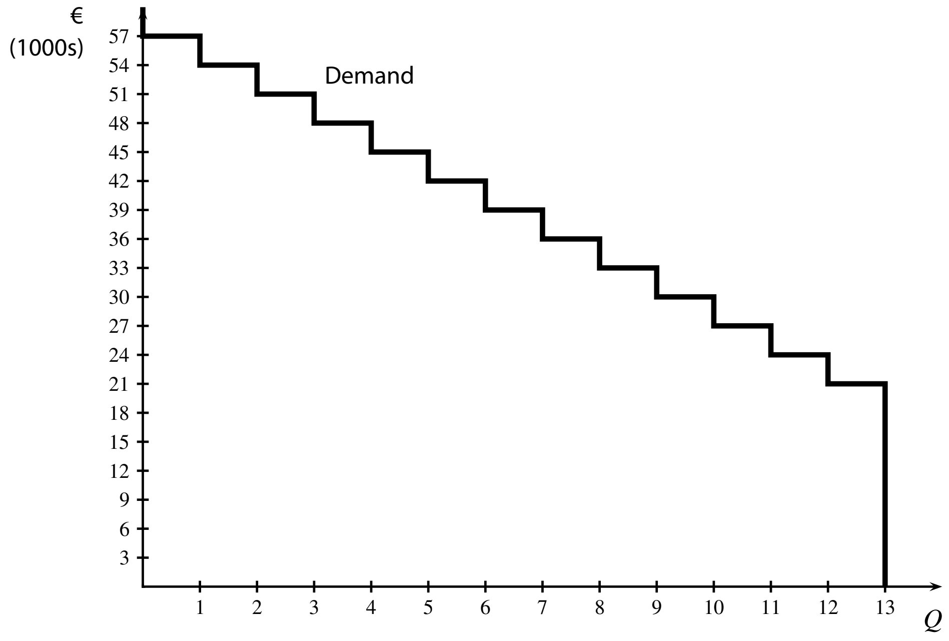
Starting at the efficient quantity, the decrease in total gains from trade is initially dominated by the reduction in consumer surplus, so that the firm gains by reducing output and increasing price. This is another way to see that the profit-maximizing quantity is lower than the efficient quantity.

Exercise 7.3. Suppose your firm produces a water purification system that you sell to small businesses. The demand for this indivisible good is shown in the first two columns of Table E7.1 (money values are in €1000s). Revenue is calculated for you in the third column. The fourth column shows your cost, and the profit is calculated in the fifth column. Figure E7.1 is a graph of the demand curve.

Table E7.1

Output	Price	Revenue	Cost	Profit
0	60	0	0	0
1	57	57	25	32
2	54	108	50	58
3	51	153	75	78
4	48	192	100	92
5	45	225	125	100
6	42	252	150	102
7	39	273	175	98
8	36	288	200	88
9	33	297	225	72
10	30	300	250	50
11	27	297	275	22
12	24	288	300	−12
13	21	273	325	−52

Figure E7.1



a. What would your price and quantity be if you maximized total surplus? (Choose the quantity that roughly equates marginal valuation to marginal cost, remembering that price at a given quantity is the marginal valuation.) How much is the total surplus?

b. What is the profit-maximizing price? What is your profit? How much is the consumer surplus? How much is the deadweight loss? Illustrate the profit, consumer surplus, and deadweight loss in Figure E7.1. (Your diagram should look similar to Figure 7.5.)

7.5 The effect of a long-run fixed cost

General procedure

A fixed cost is the same for all levels of output other than zero, and hence it does not affect the socially optimal nonzero quantity or the profit-maximizing nonzero quantity. The fixed cost only affects the decision of whether to operate at all (entry and exit decisions). For example, the R&D cost of developing a new drug affects whether or not you make the investment but not how much you charge if you actually do make the investment.

A fixed cost thus leads to the following three-step analysis, whether calculating the efficient solution or the profit-maximizing solution.

1. Calculate the solution ignoring the fixed cost (using marginal analysis).
2. Calculate the “variable profit” or “variable social surplus” ignoring the fixed cost.
3. Compare the variable profit or surplus to the fixed cost. If the fixed cost is greater, shut down or do not start up; otherwise, follow the production plan derived in step 1.

Profit-maximizing solution

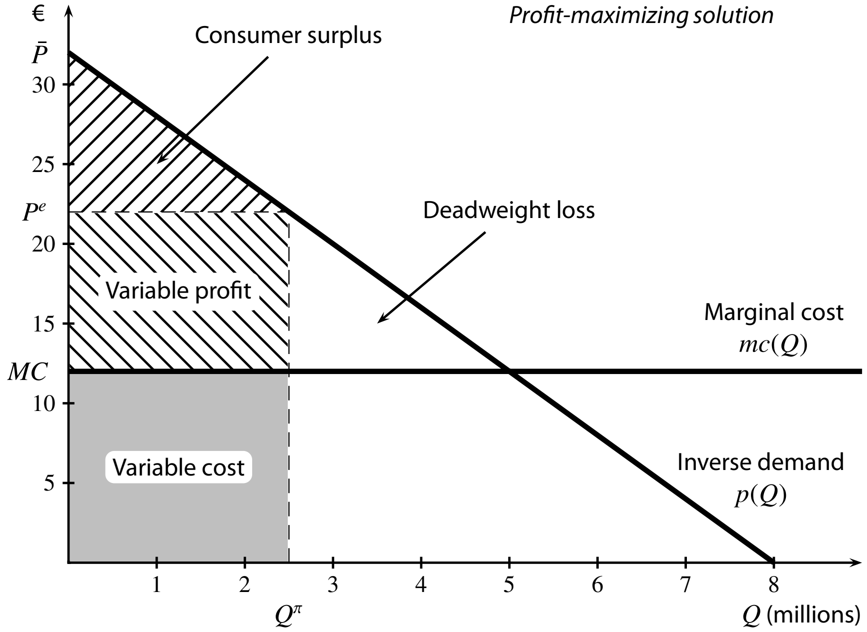
For example, suppose that a baldness medication, once developed, can be produced at a constant marginal cost of €12 per dose. Suppose the demand curve is $Q = 8 - P/4$, where quantity is measured in millions of doses of the medicine.

1. *Solution ignoring the fixed cost.* The choke price is €32, so the midpoint pricing rule tells us that the profit-maximizing price is $(12 + 32)/2 = 22$. This corresponds to output of $Q^\pi = 2.5\text{M}$.
2. *Variable profit.* The variable profit, shown in Figure 7.6, page 149, is

$$(P - MC)Q = (22 - 12) \times 2.5\text{M} = 25\text{M}.$$

3. *Entry/exit decision.* Hence, the firm should develop the medicine only if the R&D cost is lower than €25M.

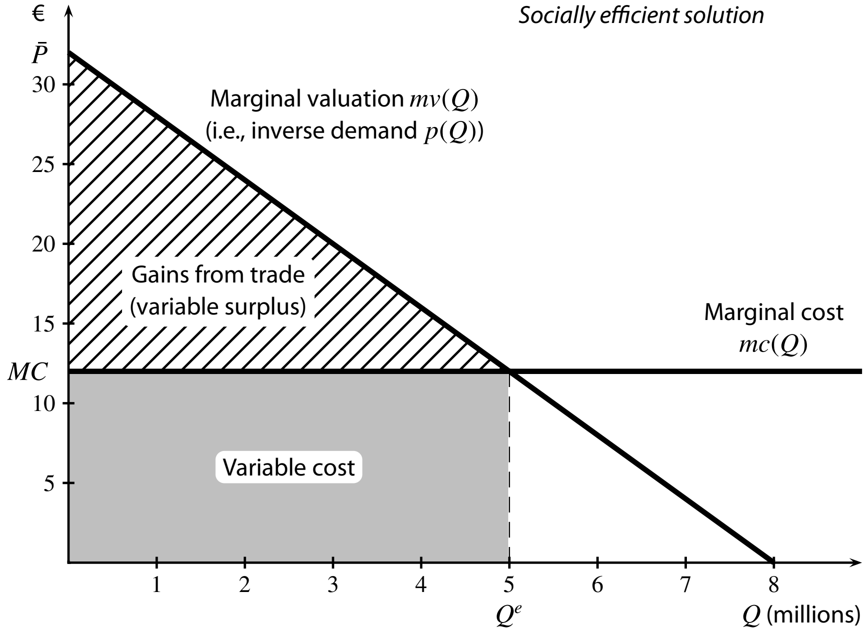
Figure 7.6



Socially efficient benchmark

Let's calculate the efficient solution for the same numerical example. Recall that the marginal cost is €12. The marginal valuation curve is the inverse demand curve: $mv(Q) = 32 - 4Q$. The marginal cost and marginal valuation curves are shown in Figure 7.7.

Figure 7.7



Ignoring the fixed cost, the efficient output level is $Q^e = 5$M and the gains from trade are €50M. (The base of the triangle is 5 and the height is 20, so the area is $(5 \times 20)/2 = 50$.) It is therefore efficient to develop the medication if the R&D cost is less than €50M.

Patents

Not only does the profit-maximizing firm produce too little (compared to the socially efficient level) once it decides to develop the medicine, it also makes a socially inefficient decision about whether to invest in the R&D. This is because the firm bears the entire R&D cost but cannot appropriate all the gains from trade that this investment can generate.

In our example, the deadweight loss if the firm decides to develop the medicine is €12.5M (the area of the white triangle in Figure 7.6). Suppose, however, that the R&D cost is €30M. The firm decides not to develop the medicine (its variable profit is €25M) even though it is socially efficient to do so (the variable surplus is €50M). No surplus is generated, whereas the socially efficient outcome would generate a surplus of €50M − €30M = €20M. Thus, all gains from trade are lost and the deadweight loss is €20M. Such underinvestment can lead to substantial inefficiencies.

What does this tell us about how well a patent system works for stimulating research?

Without a patent system (or some other means to keep other firms from copying innovations), no firm would invest in research and development: post-investment competition would drive the variable profit to zero, making it impossible to recover the R&D expense. A patent allows a firm to have market power and extract some of the gains from trade. Thus, it stimulates research and is better than having no patent system.

However, a patent system is an imperfect solution for two reasons.

1. Once a product is developed, the firm produces less than the socially efficient amount (charges too high a price).
2. The patent system still does not allow the firm to extract all the gains from trade generated by its investment, so investment remains below the socially efficient amount.

The alternative is for governments to subsidize research. If governments were perfect at deciding what research to finance and at conducting the research, this solution would be great. A government could fund the socially efficient research and then give away the knowledge so that competition leads to an efficient level of production. However, governments are not so perfect. (Furthermore, the taxation required to finance the research causes a deadweight loss.) Thus, most countries use a mixed approach, combining a patent system with government subsidies for research.

7.6 Wrap-up

We studied the price–quantity decision of a firm that has market power. The firm prices uniformly, charging the same price to all customers and allowing them to purchase any amount at the posted price.

The efficient quantity equates marginal valuation and marginal cost; it occurs at the intersection of the marginal cost curve and the demand curve. The profit-maximizing firm produces less than the efficient quantity, thereby causing a loss of social surplus called a deadweight loss. In choosing the optimal price and quantity, the firm faces the following trade-off: lower quantity reduces the total gains from trade but also reduces the surplus left to customers.

Additional exercises

The following exercises are linked and review all the calculations from this chapter.

Exercise 7.4. You manage a firm and must make the following decision. There is a good that could be developed with an R&D investment of 250, after which a patent would be obtained and the good could be produced at a constant marginal cost of 10. The potential market for this good has the following demand curve:

$$d(P) = 25 - (1/2)P\,.$$

The purpose of this exercise is to formulate a business plan before making the investment.

a. If you were to develop the product, what price would you charge and how much would you sell?

b. Calculate your "variable profit"—that is, your profit ignoring the up-front R&D cost.

c. Should you make the investment?

d. Suppose instead the R&D cost is only 100. How would your answers change?

e. Suppose you initially estimate the R&D cost to be 100 and that you go ahead and develop the product. However, the R&D cost ends up being 250. How does this affect your pricing when you launch the product or your decision of whether to actually launch the product?

Exercise 7.5. For the demand curve in Exercise 7.4, the total valuation curve for the consumers is

$$v(Q) = 50Q - Q^2\,.$$

a. If the R&D expense is 250, what is your profit? What is the consumer surplus? What are the total gains from trade?

b. If the R&D expense is 100, what is your profit? What is the consumer surplus? What are the total gains from trade?

Exercise 7.6. You are the CEO of Benevolent Dictators Ltd. There is a good that could be developed with an R&D investment of 250, after which the good could be produced at a constant marginal cost of 10. The potential market for this good has the following demand curve:

$$d(P) = 25 - (1/2)P\,.$$

This implies the following total valuation curve for the consumers:

$$v(Q) = 50Q - Q^2\,.$$

The purpose of this exercise is to formulate a plan before making the investment, given your objective to maximize total surplus (whether or not you can break even doing so).

a. If you were to develop the product, how much would you produce?

b. Calculate the "variable surplus"—that is, the surplus not taking into account the fixed cost.

c. Should you make the investment?

d. Suppose instead the R&D cost is only 100. How would your answers change?

Exercise 7.7. This problem refers back to Exercise 7.6.

a. What are the total gains from trade if the R&D expense is 250?

b. What are the total gains from trade if the R&D expense is 100?

Exercise 7.8. Combine the results from Exercises 7.4–7.7 to fill in the following table for the demand curve and marginal cost in those exercises.

FC	Possible surplus	Actual surplus	Deadweight loss
250			
100			

What is the difference between the sources of the deadweight loss in the two cases?

Chapter 8

How Pricing Depends on the Demand Curve

8.1 Motives and objectives

Broadly

A firm's demand curve may shift when (for example) a rival firm changes its price, there is a news report about the firm's product, or the firm engages in an advertising campaign. How does or should a firm adjust its price following such a shift?

If we have all the data—complete demand curves and cost curves—needed to calculate optimal prices before and after the shift, then the answer to this question just gives the "before" and "after" prices. However, we would like to be able to say something even when we lack such data. For example, under what easily verifiable conditions can we say that the firm should raise its price following a shift in its demand curve?

We will show, for example, that a firm should raise its price in response to an increase in the price of a rival firm's substitute good. However, it should lower its price if the price of a complementary good goes up. The effect of an advertising campaign on price depends on the nature of the campaign. However, typically it will be raise demand for the good and differentiate the good so that consumers are less price sensitive. Then the firm should raise its price following the campaign.

More specifically

A shift in the demand curve has two effects on pricing.

- *Volume effect.* Suppose that a firm has increasing marginal cost. If the demand curve shifts outward (so that, at any price, the firm would sell more) then the extra volume of sales pushes the firm to a region of higher marginal cost and causes the firm to raise its price.
- *Price-sensitivity effect.* Even with constant marginal cost, a firm should raise its price if demand becomes less price sensitive.

These two effects are typically both present. To understand each effect, we will isolate them. We first isolate the price-sensitivity effect by studying a firm with constant marginal cost. The main step in our analysis is to measure price sensitivity correctly—as elasticity—

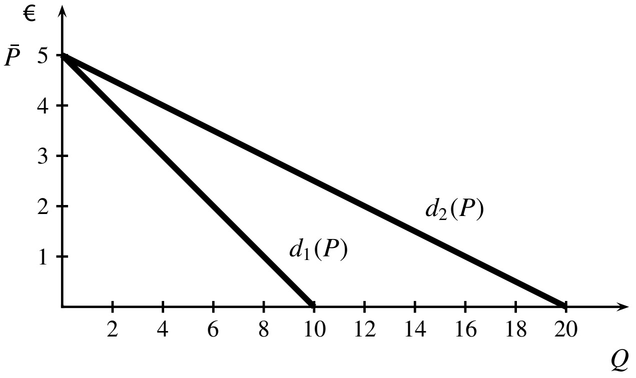
and to rephrase a firm's pricing decision in terms of elasticity.

To examine the volume effect by itself, we study a firm with increasing marginal cost and whose demand curve simply increases or decreases in volume without any change in price sensitivity. This means that the demand goes from $Q = d(P)$ to $Q = kd(P)$ for some $k > 0$. For example, if $k = 2$ then demand has doubled, as if the firm expanded into a new market of identical size and with the same properties as its original market.

8.2 A case where the price does not change

Consider a firm whose demand curve shifts, from $d_1(P)$ to $d_2(P)$, as described in the previous paragraph. In other words, there is a $k > 0$ such that $d_2(P) = kd_1(P)$ for every P. If $k = 0.5$ then demand has dropped in half, as if the firm abandoned half of its market. If $k = 2$ then demand has doubled after the shift. For example, suppose the initial demand curve is linear and equal to $d_1(P) = 10-2P$; then the new demand curve is $d_2(P) = 20-4P$. These two demand curves are shown in Figure 8.1.

Figure 8.1



Suppose also that the firm has constant marginal cost MC (the same before and after the shift). We will show that *the firm's profit-maximizing price does not change following the shift in the demand curve.*

We can see this for the example in Figure 8.1. Since both demand curves are linear and since marginal cost is constant, we can apply the midpoint pricing rule as a shortcut to calculate the optimal prices before and after the shift. Suppose that $MC = 1$. Note that both demand curves have the same choke price $\bar{P} = 5$. Thus, the midpoint pricing rule gives us the same answer for the two demand curves:

$$P = \frac{\bar{P} + MC}{2} = \frac{5+1}{2} = 3.$$

Hence, the shift in the demand curve has no effect on price!

This conclusion holds even if demand is not linear. That is, whenever the firm has constant marginal cost and its demand curve expands or contracts by a constant factor, the firm should not adjust its price. Here is one way to see this. With constant marginal cost, we can write profit as a function of price in the form of "mark-up times volume": $(P - MC) \times d(P)$. Say, for example, that demand doubles at each price. Then profit doubles at each price, becoming $2 \times (P - MC) \times d(P)$. Any price that maximizes $(P - MC) \times d(P)$ also maximizes twice that quantity. Thus, the profit-maximizing price does not change after the shift.

By assuming that marginal cost is constant, we have suppressed the volume effect. Apparently the two demand curves also have the same price sensitivity, since there is no price-sensitivity effect. Our measure of price sensitivity should be such that these two demand curves are equally price sensitive. This happens when we measure it by elasticity.

8.3 Marginal revenue and elasticity

Marginal revenue measures the change in revenue for a unit change in output. Marginal revenue varies along a demand curve (for different quantities or, equivalently, for different prices). Since marginal revenue depends on the slope of the demand curve and so does elasticity, it turns out that one can write marginal revenue for any point on the demand curve as a function of the price and the elasticity at that point on the demand curve:

$$MR = P\left(1 - \frac{1}{E}\right). \tag{8.1}$$

This formula has several uses.[1]

Elasticity and the sign of marginal revenue

We can see from equation (8.1) that the sign of marginal revenue depends on elasticity as shown in Table 8.1.

Table 8.1

If elasticity is	we say demand is	Then marginal revenue is	and an increase in output (a decrease in price) causes revenue to
< 1,	inelastic.	< 0,	fall.
$= 1$,	unit elastic.	$= 0$,	stay the same.
> 1,	elastic.	> 0,	rise.

That is, Table 8.1 classifies the effect that an increase in output has on revenue. In Section 3.4 we studied the effect that an increase in price has on expenditure. Since the consumers' expenditure equals the firm's revenue and since an increase in Q is the same as a decrease in P, Table 8.1 (derived from equation (8.1)) is just a restatement of Section 3.4

[1] Here is where this formula comes from. The formula for revenue is $r(Q) = p(Q) \times Q$. Marginal revenue is $r'(Q) = p(Q) + p'(Q)Q$. Replace $p(Q)$ by P and $p'(Q)$ by dP/dQ; this yields $MR = P + (dP/dQ)Q = P(1 + (dP/dQ)(Q/P))$. Note that $(dP/dQ)(Q/P) = -1/E$. Thus, $MR = P(1 - 1/E)$.

(derived from intuition).

Using the formula to check your pricing

Using equation (8.1), the marginal condition $MC = MR$ can be written

$$MC = P\left(1 - \frac{1}{E}\right). \tag{8.2}$$

We can rearrange this equation to read

$$P = \frac{E}{E-1}MC. \tag{8.3}$$

This looks like a nice formula for setting your price. You look up your marginal cost MC and your elasticity E; you then charge your marginal cost times a markup $E/(E-1)$, which is higher the less elastic demand is (the closer is E to 1).

Unfortunately, both elasticity and marginal cost depend on your price: elasticity because it varies along the demand curve, marginal cost because it varies with your level of output. Hence, P appears on both sides of equation (8.3). Only in the exceptional case in which both the elasticity of demand and the marginal cost are constant does equation (8.3) constitute a true pricing rule. When the constant-elasticity demand is written $Q = AP^{-B}$, where B is the elasticity, and the constant marginal cost is denoted MC, we obtain the pricing formula

$$P = \frac{B}{B-1}MC$$

that was stated in Section 7.3.

Nevertheless, even in the real world where elasticity and marginal cost are not constant, equation (8.3) is useful for checking whether your price is optimal and, if not, for giving the direction in which you should adjust your price. The procedure is the following.

Step 1. Given your current pricing decision, measure the elasticity of demand. Ask your marketing department, who will probably estimate a constant-elasticity demand function.

Step 2. Measure also your marginal cost. Ask your production department: though it would be difficult to determine an entire cost curve, it should be possible to estimate the change in cost for small changes in output from the current level.

Step 3. Use the two estimates to calculate the price according to equation (8.3).

Step 4. If the calculated price is close to your current price, then your price is approximately optimal; if the calculated price is higher than your current price then you should increase your price; if it is lower then you should decrease your price.

Step 5. The formula does not tell you by exactly how much to change your price. It exaggerates the price change, so try a price between your current price and the one given by the formula.

After operating a while at the new price, check again by repeating these steps.

Using the formula to understand market power

Equation (8.3) also gives some intuition about how price differs from marginal cost. Recall from Chapter 7 that a firm's pricing decision would be efficient if the firm charged marginal cost. We see from equation (8.3) that, the less elastic the firm's demand is, the more its price differs from marginal cost. When the firm is charging above marginal cost, it knows that there are unrealized gains from trade. It would like to generate these gains by lowering its price if it could extract some of the extra surplus . However, if demand is not very elastic then it is not worth lowering the price; the extra gains from trade that are generated will be small compared to the extra surplus the firm gives to its customers. Thus, price diverges further from marginal cost the less elastic is demand.

Therefore, regulatory authorities elasticity use elasticity as one measure of market power (where higher elasticity means less market power). The Lerner index measures market power as $(P - MC)/P$. If we replace MC in this formula by the right-hand side of equation (8.2), we see that the Lerner index is just $1/E$. It varies from 0 (infinitely elastic demand and no market power) to 1 (unit-elastic demand and high market power).[2]

8.4 The effect of an increase in marginal cost

Raise your price if your marginal cost goes up

The focus of this chapter is on how pricing should respond to a shift in the demand curve. However, let's see how pricing responds instead to a shift in the *cost* curve.

If your marginal cost increases, your profit-maximizing quantity falls (hence your profit-maximizing price rises). This unsurprising fact does not rely on marginal analysis but is easily demonstrated by it.

Suppose that your marginal revenue curve is decreasing.[3] You initially choose an output level that satisfies the marginal condition $MR = MC$. Then your marginal cost increases: perhaps one of your inputs has become more expensive or a per-unit tax has been imposed on your product. If you do not adjust your output, then marginal revenue does not change but your marginal cost is higher. Hence, $MR < MC$. To restore the marginal condition, you should lower your output; this causes marginal revenue to rise and marginal cost to fall or stay the same.

[2] Other measures of market power—namely, market share and the Herfindahl–Hirschman Index—are now more widely used by courts than the Lerner index.

[3] This is true, for example, if demand becomes more elastic at higher prices, which is an empirical regularity.

Output should be lower than that which maximizes revenue

A corollary is that the revenue-maximizing quantity Q^r is greater than the profit-maximizing quantity Q^π if a firm's marginal cost is not zero. The reason this is a corollary is that Q^r maximizes "profit" given a hypothetical cost curve with zero marginal cost, whereas Q^π maximizes profit given a higher marginal cost.

Never choose a price at which demand is inelastic

Let Q^r be the quantity that maximizes revenue. Then $mr(Q^r) = 0$. From equation (8.1), $E = 1$. Thus, a firm with zero marginal cost—and hence that maximizes revenue—should choose a point on the demand curve at which demand has unit elasticity.

Let Q^π be the profit-maximizing quantity for a firm with positive marginal cost. Then $mr(Q^\pi) > 0$. Therefore, at Q^π, $E > 1$: a firm with positive marginal cost should choose a point on the demand curve at which demand is elastic.

In April 1998, AOL increased its monthly membership billing rates by $2 per subscriber. The company expected lower subscriber additions and flat revenue as a consequence of the transition to this new plan. However, its revenue in the second quarter of 1998 rose to $667.5 million, up nearly 16% over the previous quarter. Thus, the company had unwittingly let its pricing drift into the inelastic portion of the demand curve, and the price increase was a step toward restoring the price to its profit-maximizing level. CEO Steve Case later commented: "I was surprised by the strength of the member growth this quarter. When we announced the [monthly unlimited access rate] increase, my prediction was that we would see no member growth in the June quarter as we transitioned to the new pricing plan. The fact that we had our best fourth quarter ever, despite the price increase and lower marketing expenditures, was a surprise and a delight to me." The firm had overestimated demand elasticity. It was pleasantly surprised to be wrong, but if the firm had better estimated demand it would have increased its price earlier and made even higher profit.

An example of an unpleasant surprise was when Sony slashed the price of its Playstation2 gaming console from $299 to $199 in May 2002. This price cut would necessarily increase sales and hence Sony's costs; it could only make sense if Sony expected sales to increase so much that revenue would rise even more than costs. However, Sony had overestimated the elasticity of demand. Overall gaming division revenues fell 5%, which was mainly due to this price decrease (according to Sony's 2003 annual report). To make matters worse, Sony's price cut was matched by Microsoft on its X-Box console, softening sales even further. It is estimated that console revenues in 2003 fell by nearly $2 billion.

Most state-owned or state-regulated firms are not supposed to maximize profit; instead they are typically expected to charge average or marginal cost. As a consequence, they may end up pricing where demand is inelastic. For example, US postage rates were hiked in January 2001, including an increase in the price of a first-class domestic stamp from 33 to 34 cents. The one-cent hike in the price of a first-class stamp brought the Postal Service about $1 billion per year, showing us that demand was inelastic. If the Postal Service were a for-profit firm, it would want to continue raising rates at least until any further increases would reduce revenue.

Exercise 8.1. The following is a nice application of these simple conclusions. It is difficult because it requires several steps of reasoning.

Suppose the author of a book has a contract with a book publisher, which pays him \$100,000 plus 7% of the wholesale value of all sales of his book. The publisher has a fixed production cost of \$200,000 (including the \$100,000 royalty) for the book, plus an additional cost of \$5 per copy for printing and distribution. Both parties (the author and the publisher) care only about maximizing their own profit. Compare the price that the publisher will set (the one that maximizes the publisher's profit) with the price the author would like the publisher to set (the one that maximizes the author's profit). Although you do not have enough information to determine these prices, you can say which one is higher. Explain your reasoning carefully.

8.5 The price-sensitivity effect

Main idea

We finally have the tools to return to one of the problems posed in Section 8.1: How does a change in the price sensitivity of a demand curve affect a firm's optimal price?

We hinted that we would compare the price sensitivity of two demand curves by comparing their elasticities. But what does this mean, given that elasticity varies along a demand curve? For example, elasticity varies from zero to infinity along any linear demand curve. How can we say that one demand curve is more elastic than another?

We rank the elasticities of two demand curves d_1 and d_2 by comparing their elasticities *at each price*. If the elasticity of d_2 is higher, at each price, than the elasticity of d_1, then we say that d_2 is *more elastic* than d_1.[4]

We then have the following important conclusion.

> Price-sensitivity effect: *Suppose a firm has constant marginal cost. If its demand curve shifts from d_1 to d_2, where d_2 is less elastic than d_1, then the firm should charge a higher price after the shift.*

Note: We can also use this conclusion to compare the pricing decisions of two firms that have the same constant marginal cost but face different demand curves. The one with less elastic demand should charge a higher price.

[4] Not all pairs of demand curves can be ranked unambiguously. Curve d_2 may be more elastic than d_1 at some prices but less elastic at other prices.

We can gain some intuition for this result from our formula for marginal revenue:

$$MR = P\left(1 - \frac{1}{E}\right).$$

At any price, lower elasticity implies lower marginal revenue and hence a greater incentive to decrease output and raise price.

Comparing elasticity of linear demand

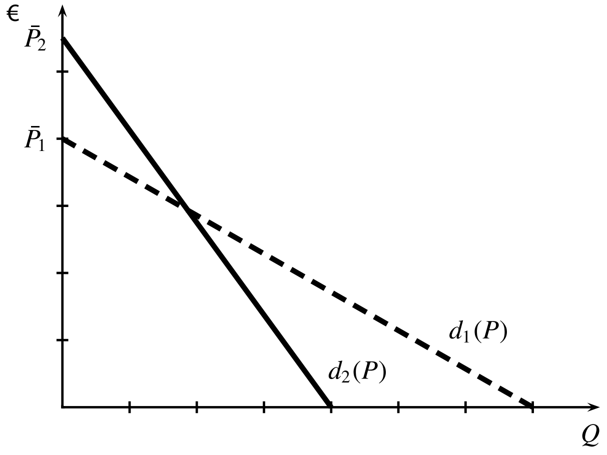
For linear demand $d(P) = A - BP$, recall that

$$E = \frac{P}{\bar{P} - P},$$

where $\bar{P} = A/B$ is the choke price. Demand is more elastic when $\bar{P}$ is lower. Thus, *when comparing two linear demand curves, the one with the lower choke price is more elastic.*

For example, the two demand curves in Figure 8.1 have the same choke price and hence are equally elastic. In Figure 8.2, d_2 has a higher choke price and hence is less elastic. A firm with constant marginal cost should increase its price if its demand curve shifts from d_1 to d_2.

Figure 8.2



Determinants of elasticity

Even if we do not have the data to measure the elasticities of two different demand curves, we can guess which one is more elastic by looking at certain qualitative data. Most of the intuitive statements we could make about price sensitivity are true when we measure price sensitivity by elasticity. In particular, demand for a good tends to be less elastic:

1. when it has fewer close substitutes;
2. when the buyers of the good have higher income;
3. following an advertising campaign.

Although an advertising campaign often makes demand less elastic by differentiating the brand from potential substitutes, that is not necessarily an objective or consequence of advertising. The payoff from the advertising may instead be an increase in volume. For example, we have seen already that mere expansion into a new market that has the same demand characteristics has no effect on elasticity. Furthermore, repositioning a product toward the mass market good could increase elasticity and lead to a decrease in price (presumably made up for by an increase in volume).

8.6 The volume effect

We say that demand curve d_2 has higher volume than d_1 if $d_2(P) > d_1(P)$ for all P. Consider now the volume effect:

> Volume effect: *Suppose that marginal cost is increasing and the demand curve shifts from d_1 to d_2, where d_2 has higher volume than d_1 but the two curves are equally elastic. Then the firm should charge a higher price following the shift.*

(The assumption about the demand curves simply means there is a $k > 1$ such that $d_2(P) = kd_1(P)$ for all P.)

We can use marginal analysis to give the following intuition behind the volume effect. The firm initially chooses an optimal price P_1 given demand curve d_1, so that marginal revenue equals marginal cost. Now suppose the demand curve shifts to d_2. If the firm does not change its price, then its new marginal revenue after satisfying the increased demand also does not change because elasticity has not changed. However, because demand is higher and because the firm has increasing marginal cost, the marginal cost is higher than before. Hence, marginal cost exceeds marginal revenue and the firm can increase its profit by reducing output and increasing its price.

8.7 Combining the price-sensitivity and volume effects

Main result

We can also predict the change in price when both the price-sensitivity and volume effects are present and go in the same direction.

> Combined price-sensitivity and volume effects: *Suppose that marginal cost is increasing and that the demand curve shifts from d_1 to d_2, where (a) d_2 is less elastic than d_1 and (b) d_2 has higher volume than d_1. Then the firm should charge a higher price following the shift.*

The shift following an advertising campaign

An advertising campaign will often increase volume and simultaneously make demand less price sensitive. Then, with either constant or increasing marginal cost, the firm should raise its prices.

The shift in demand when a competitor raises its price

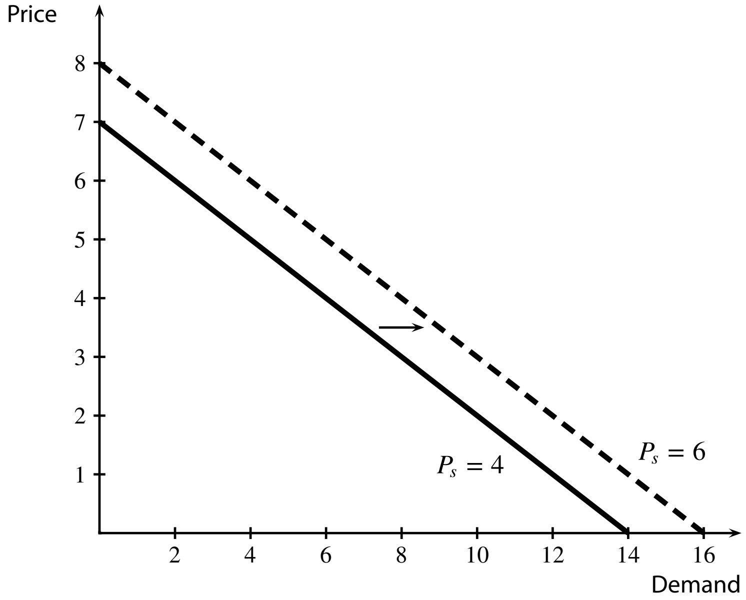
If the price of a substitute good goes up, then a firm's demand curve shifts. Such a price increase has two effects on demand.

1. An increase in the price of a substitute good increases demand for the firm's good. (This property defines a substitute good.) Thus, the volume effect is positive.
2. An increase in the price of a substitute good nearly always makes demand for the firm's good less elastic. (This is not necessarily true, but it is an empirical regularity; we will assume it is true in the rest of this book.) Thus, the price-sensitivity effect is also positive.

Therefore, the firm should raise its price in response to the competitor's price increase.

For example, suppose that demand as a function of own price P and the price P_s of a substitute good is linear: $Q = 10 - 2P + P_s$. Suppose initially that $P_s = 4$ and that P_s later increases to 6. The demand curves before and after the increase in P_s are $Q = 14 - 2P$ and $Q = 16 - 2P$, respectively; these are shown in Figure 8.3.

Figure 8.3



The choke price is initially 7 and then goes up to 8. Thus, demand becomes less elastic. Both the volume and price-sensitivity effects are present. The firm will increase its price whether it has constant or increasing marginal cost.

The shift in demand when the price of a complementary good goes up

An increase in the price of a complementary good also causes a firm's demand curve to shift. However, the effects are the opposite of those caused by an increase in the price of a substitute good.

1. An increase in the price of a complementary good lowers demand for the firm's good. (This property defines a complementary good.) Thus, the volume effect is negative.
2. An increase in the price of a complementary good nearly always makes demand for the firm's good more elastic. (This is not necessarily true, but it is an empirical regularity; we will assume it is true in the rest of this book.) Thus, the price-sensitivity effect is also negative.

Therefore, the firm should lower its price in response to the increase in price of the complementary good.

For example: with linear demand, an increase in the price of a complementary good causes the demand curve to shift to the left without changing in slope. Thus, the choke price falls and demand becomes more elastic; this was illustrated in Figure 3.2. Both the volume and price-sensitivity effects are present. Whether the firm has constant or increasing marginal cost, it should decrease its price in response.

The shift in demand when the income of consumers goes up

Suppose that the consumers' income rises. This causes demand for a normal good to rise. Demand typically becomes less elastic as well. Thus, the firm should raise its price.

For example: with linear demand, an increase in income causes the demand curve for a normal good to shift to the right without changing its slope. Thus, the choke price rises and demand becomes less elastic; volume goes up at as well. Whether the firm has constant or increasing marginal cost, it should increase its price.

8.8 Wrap-up

A shift in a firm's demand curve has two effects on price: a volume effect and a price-sensitivity effect. The volume effect is positive if demand increases at each price and if marginal cost is increasing. The price-sensitivity effect is positive if demand becomes less elastic.

We introduced elasticity of demand with the particular goal of measuring the price-sensitivity effect. However, elasticity has other uses and it is one property of demand that can be robustly measured. We relate elasticity to marginal revenue and show that a firm should never price in the inelastic part of its demand curve. We also show how a firm can check and adjust its price by measuring marginal cost and elasticity at its current price.

Additional exercises

Exercise 8.2. Evaluate: "After an advertising campaign, the cost of the advertising is sunk; hence, the advertising campaign should have no effect on the firm's pricing strategy."

Exercise 8.3. Suppose that two monopolists, F and S, sell the same product but in different markets. Firm F sells in France and and firm S sells in Switzerland. The demand in these two countries has the following characteristics:

1. at any given price, the quantity demanded in France is greater than the quantity demanded in Switzerland;
2. at any given price, the elasticity of demand is identical in both countries; and
3. demand becomes less elastic as one moves down the demand curve (toward lower price and higher quantity).

The two monopolists have identical cost functions with no fixed cost and *increasing* marginal cost.

Use this information to compare the firms' profit-maximizing prices. At play here is the *volume effect*. State the implication of the volume effect for this example. Then use marginal analysis to give a careful but succinct explanation of why this is true.

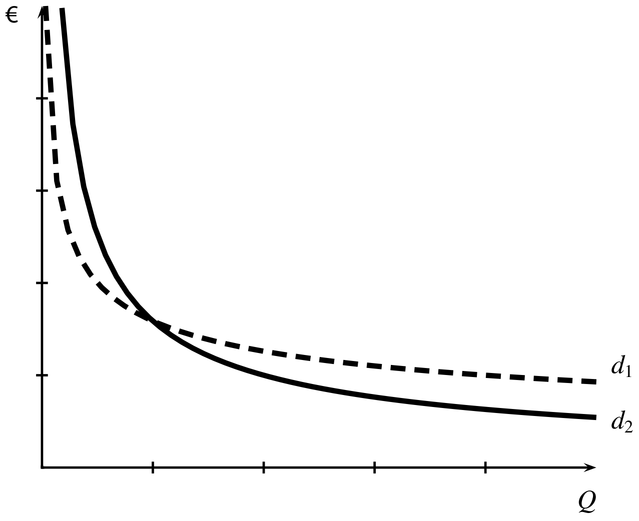
Exercise 8.4. Though elasticity is not the same as slope, slope does determine the relative elasticities of two demand curves at a point where they intersect. In the equation

$$E = -\frac{dQ}{dP}\frac{P}{Q},$$

P and Q are the same for the two demand curves at their intersection point. Hence, the curve whose slope dQ/dP is greater in magnitude has higher elasticity. Since we graph demand curves with price on the vertical axis, the curve with higher dQ/dP in magnitude is less steep. The curve with higher slope in magnitude—less steep, given the way we usually graph demand curves—is thus the more elastic one (at that point).

Use this information to answer the following question. You are told that one of the two demand curves in Figure E8.1 is less elastic than the other. Which one is less elastic? How did you determine this?

Figure E8.1



Chapter 9

Explicit Market Segmentation

9.1 Motivation and objectives

Broadly

Figure 9.1

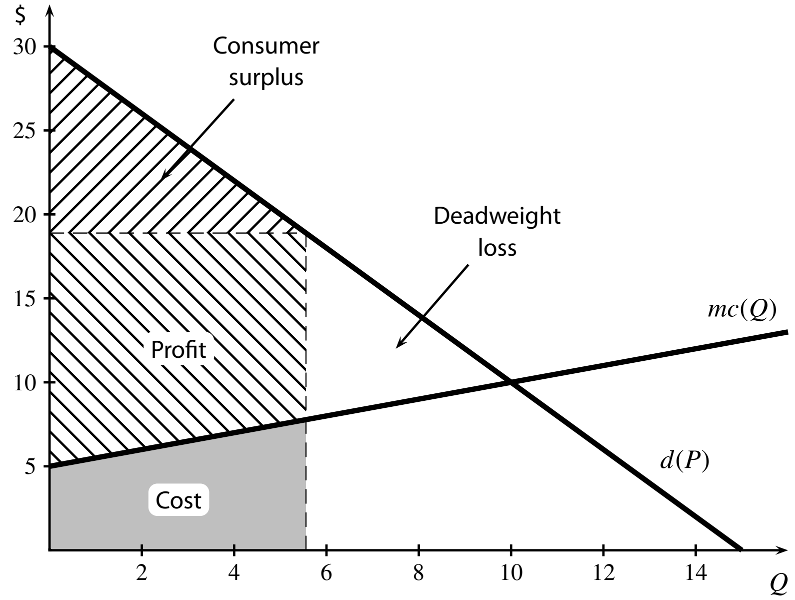


Figure 9.1 shows our familiar illustration of the deadweight loss and of the distribution of surplus for a firm with market power. The firm's manager is happy to see the large region labeled "profit", but she feels disappointed for two reasons. First, the customers are getting away with some surplus. Second, some potential gains from trade (the deadweight loss) are not realized and hence go to no one. Ideally, she would like to generate all possible gains from trade and extract these gains for the firm.

Her problem is not a lack of bargaining power. Her firm can make a take-it-or-leave-it price offer that allows for no haggling. Yet changing the price trades off consumer surplus for deadweight loss and so cannot eliminate both of them.

Chapters 9, 10, and 11 examine more sophisticated pricing strategies that increase profits by expanding gains from trade and reducing consumer surplus.

More specifically

The pricing we studied in the previous two chapters is so standard that it needed neither a special term nor a definition—until now, as we move toward more sophisticated strategies. Henceforth, we refer to the pricing from those chapters as *uniform pricing*. The other pricing strategies that we consider in Chapters 9, 10, and 11 are broadly classified as *price discrimination*. For now, it would be premature to define the distinction between uniform pricing and price discrimination. We will do so at the end of Chapter 11, when we can reflect back on the various pricing strategies.

A property of uniform pricing that we move away from in the current chapter is *equal treatment of customers*—that the firm offers all customers the same options for trade. Customers have heterogeneous characteristics, so a firm can raise its profit by charging a higher price to customers with higher valuations. Such differential treatment is called *explicit market segmentation*.

This chapter studies a common and simple form of explicit market segmentation in which the firm charges different prices to two or more market segments:

- pharmaceutical companies charge different prices in different countries;
- movie theaters and ski resorts have discounts for senior citizens;
- utility companies have different rates for residential and business customers;
- some countries that rely on income from tourism enforce a system of differential pricing for foreign visitors and (less wealthy) local residents.

What differences between the demand curves of two market segments leads the firm to charge a higher price to one segment than to the other? This is similar to the question we asked (and answered) in Chapter 8 *except* that, in Chapter 8, the two demand curves were before and after a shift in demand. The bottom-line conclusion remains nearly the same:

> *Charge a higher price to the market segment with less elastic demand.*

The conclusion is even simpler here because there is no mention of volume and of the shape of the marginal cost curve. (In particular, there is no volume effect.) The marginal cost of selling one more unit is the same for both market segments because they are served by the same firm and from the same production process.

9.2 Requirements for explicit market segmentation

Examples

A market segment means a subgroup of customers. It is sometimes the case that a firm can charge a different price to different market segments. If the demand curves in the different segments have sufficiently different properties, then such explicit market segmentation can raise the firm's profit.

Here are some examples.

1. A firm can offer student or senior-citizen discounts, enforced by requiring a student ID or proof of age.
2. A utility company may charge different rates to business and residential customers.
3. A pharmaceutical company can charge different prices in different countries for prescription medicine.
4. A pharmaceutical company can charge different prices for the same medicine depending on what it is prescribed for.
5. A software company can give an academic discount to people who have an academic affiliation.

Observability and no-arbitrage

In each of the five examples, customers in a market segment that is charged a higher price may try to circumvent the price discrimination by either:

- pretending they belong to the other segment, which is called "masquerading"; or
- buying the product through customers in the other segment, which is called "arbitrage".

The following would be examples of masquerading: a non-student uses a fake student ID in order to receive the student discount at a theater; a person running a business from his own home does not disclose this fact in order to obtain residential rates from a utility company; a patient with one ailment buys a medicine using a prescription for a different ailment.

The following would be examples of arbitrage: a person buys prescription medicine in India and resells it in the United States; a professor buys a software package for a friend in order to get the academic discount.

The easier are masquerading and arbitrage, the less effective is explicit market segmentation. In each of the five examples, masquerading and arbitrage are difficult or at least inconvenient. Speaking in more absolute terms, we would say that the no-masquerading (or "observability") and no-arbitrage conditions are satisfied. On the other hand, it is nearly impossible to distinguish the income level of concert goers in order to charge wealthier customers higher prices for the same tickets; the no-masquerading or observability condition is violated. If a book retailer tried to charge different prices to men and to women, it would be very easy for (say) a man to get a woman to buy a book for him; the no-arbitrage condition is violated.

The observability and no-arbitrage conditions are stringent. Being able to explicitly segment a market is the exception rather than the rule. Explicit market segmentation is nearly impossible if a product is sold through a retailer and the producing firm never has contact with the customer. A retailer of art supplies might offer a discount to local art students—thus having different mark-ups for different customers, which is to say different prices for its retailing services. However, the manufacturer of the art supplies would find it difficult to offer different wholesale prices that depend on the final customer who buys the product. In order to avoid masquerading, Apple offers an academic discount only at its own stores and through a limited number of authorized retailers, which must provide documentation to Apple each time they sell a computer with the discount. In order to avoid arbitrage, Apple limits the number of computers an individual can purchase with the discount.

Avoiding masquerading and arbitrage may be possible yet require costly systems to track customers and products. Furthermore, merely charging different prices carries a complexity cost. The firm must weigh these transaction costs against the benefits of explicitly segmenting the market.

There is a technology-based arms race between firms and customers that is constantly shifting the practicality of explicit market segmentation. On the one hand, information technology and internet-based transactions makes it easier to track customer characteristics and tailor offers to these characteristics. For example, Amazon has (controversially) used customers' browsing and purchasing histories to adjust the prices that it charges different customers for the same book. On the other hand, this same technology makes trade more anonymous (facilitating masquerading) and makes it easier for customers to arbitrage price differences (for example, it has become easier to purchase prescription medicine from a foreign country by internet even though national laws restrict such trade).

Exercise 9.1. Explicit market segmentation tends to be more common in the sale of services (e.g., discrimination by income for universities and by age for air transportation services) than in the sale of manufactured goods. Why do you think this is so?

A false example

If you had to give another example of explicit market segmentation, you might mention business and leisure travelers who pay different prices for air travel. However, this would be wrong. The defining characteristic of explicit market segmentation is that customers have different trading opportunities. For example, a 40-year-old cannot get the children's discount at an amusement park. Is this true of the airlines' pricing example?

A business traveler and leisure traveler are sitting side by side in coach class on their way from Amsterdam to Barcelona. They converse and reveal that the business traveler paid €700 for her ticket whereas the leisure traveler paid €250 for his. Could the business traveler have bought the same ticket as the leisure traveler? Certainly. The leisure traveler purchased his ticket in advance, took a Saturday-night stay, and accepted penalties and nonrefundability in case of changes. The business traveler could have done the same but chose not to.

Airlines cannot explicitly segment the market owing to the lack of observability: they cannot tell whether a customer is a business or leisure traveler. A business traveler does not have a "B" branded on her forehead, and there is no leisure-traveler's ID card that the airlines could demand before giving a leisure-traveler's discount.

However, it is clear that airlines' complex pricing of restricted and non-restricted tickets is closely related to explicit market segmentation. It is an example of *implicit market segmentation*, which is a way of indirectly and imperfectly pursuing the goals of explicit market segmentation when the latter is not possible. We will study this in Chapter 10. For now, as you read through the current chapter, it is still useful to ask yourself what airlines *would* do if they *could* explicitly segment the market. That is the first step toward understanding how they should handle implicit market segmentation.

9.3 Different prices for different segments

As we have noted, the observability and no-arbitrage conditions need only be approximately satisfied in order for a firm to explicitly segment a market. To simplify our model, we assume that the two conditions are perfectly satisfied and we ignore any transaction costs.

Cost-based price differences that we do not study

A firm may charge a higher price to one segment than another simply because it is more costly to serve customers in the first segment (as a result, e.g., of transportation costs). More commonly, such cost-based price differences are entangled with (but do not invalidate) the demand-based price differences that we study. Sometimes the two effects are difficult to untangle empirically. Does a laundry service charge more for a woman's blouse than for a man's shirt because the former are, on average, more difficult to iron? Or is it because of differences in the elasticity of demand for the market segments? Probably both factors are present.

On the other hand, if a Japanese car manufacturer charges a lower price in the United States than in Japan for a car produced in Japan—in spite of the transportation cost—then differences in demand are driving the price difference.

At an analytic level we can untangle the two effects by studying one effect at a time, just as we untangled the volume and price-sensitivity effects in Chapter 8. In this chapter, we devote all our time to the more-interesting demand-based effects. We neutralize the cost-based effects by assuming that the goods or services sold to the two segments are identical—hence have the same production cost—and that there are no transportation costs.

Bottom line

There is no point in explicitly segmenting your market unless the demand curves of the segments are sufficiently different. What differences should there be? When we lack the data needed to calculate the profit-maximizing prices of the two segments, can we still use qualitative information about the segments to determine which segment should be charged a higher price?

The bottom line is that the price-sensitivity effect studied in Chapter 8 arises here as well, but is even more robust because there is no volume effect:

> *Your firm should charge a* higher price *to the segment with* less elastic *demand.*

Marginal conditions for revenue maximization

This conclusion again comes from studying marginal conditions.

Suppose you have two market segments, which we label 1 and 2. Once again, though we want to understand price, we attack the problem by framing it in terms of quantities: your must choose the amounts Q_1 and Q_2 to sell to the two segments.

Recall that we assume the output comes from a common production process without transportation costs. Then total cost is a function of the total amount supplied to all segments, not of how it is divided among the segments. The marginal cost of producing another unit is the same no matter which segment it goes to.

Let's think about the decomposition of your problem between production and marketing.

- *Production.* On the production side, we simply have a cost curve $c(Q)$ measuring cost as a function of total output $Q = Q_1 + Q_2$; our production managers need not concern themselves with market segmentation.
- *Marketing.* It is the task of the market managers to divide the output among the market segments in order to maximize revenue. Let $r_1(Q_1)$ and $r_2(Q_2)$ be the revenue from the two segments. Given total output Q, the marketing department's problem is to choose Q_1 and Q_2 (such that $Q_1+Q_2 = Q$) in order to maximize total revenue $r_1(Q_1)+r_2(Q_2)$. The total amount of revenue that can be obtained depends on Q; denote this by $r(Q)$.
- *Strategic planning.* The central strategy group chooses Q to maximize $r(Q) - c(Q)$.

If the firm is to maximize profit, each group must do its job right. For the marketing department, this means that they divide the output among the segments to obtain the highest possible revenue from what has been produced. The marginal condition is $MR_1 = MR_2$; otherwise, revenue can be increased by shifting output from the segment with low marginal revenue to the segment with high marginal revenue. For example, if $MR_1 > MR_2$, then shifting one unit of output from segment 2 to segment 1 increases total revenue by $MR_1 - MR_2$.

As in Chapter 8, the equality $MR_1 = MR_2$ allows us to conclude that the price is higher for the less elastic demand curve. Recall the formula

$$MR = P\left(1 - \frac{1}{E}\right).$$

Suppose demand is more elastic in segment 1 than in segment 2. As a benchmark, let output be divided so that the same price P holds in both market segments. Then $E_1 > E_2$ at the common price and

$$MR_1 = P\left(1 - \frac{1}{E_1}\right) > P\left(1 - \frac{1}{E_2}\right) = MR_2.$$

Because $MR_1 > MR_2$, revenue can be increased by shifting output from segment 2 to segment 1, causing the price to fall in segment 1 and to rise in segment 2. Therefore, the optimal division of output is such that the price in segment 1 is lower than in segment 2. (This argument assumes that marginal conditions are sufficient, but there is an alternative proof of the same result without such an assumption.)

For example, suppose that

$$d_1(P) = 20 - P \quad \text{and} \quad d_2(P) = 60 - 2P\,. \tag{9.1}$$

The choke prices are $\bar{P}_1 = 20$ and $\bar{P}_2 = 30$, so the demand in segment 1 is more elastic demand than that in segment 2. Therefore, you should charge less to segment 1 than to segment 2.

Determining the optimal level of production

This price comparison is the main point of this section. However, let's complete the analysis by including the strategic planning problem in order to see how quantities and prices are calculated. There are three cases.

Constant marginal cost. Each market segment represents an independent uniform pricing problem. You solve the two separate equations:

$$\begin{aligned} mr_1(Q_1) &= MC\,, \text{and} \\ mr_2(Q_2) &= MC\,. \end{aligned}$$

For example, suppose you have constant marginal cost of 10 and that the demand curves are those in equation (9.1). Using the midpoint pricing rule, your optimal prices are

$$\begin{aligned} P_1 &= (\bar{P}_1 + MC)/2 = (20 + 10)/2 = 15\,, \text{and} \\ P_2 &= (\bar{P}_1 + MC)/2 = (30 + 10)/2 = 20\,. \end{aligned}$$

Observe that $P_1 < P_2$ as predicted.

Increasing marginal cost. If marginal cost is not constant, then pricing across different segments is interrelated because the marginal cost is a function of the total amount produced. The optimal quantities Q_1 and Q_2 should solve the system of equations

$$\begin{aligned} mr_1(Q_1) &= mc(Q_1 + Q_2)\,, \\ mr_2(Q_2) &= mc(Q_1 + Q_2)\,. \end{aligned}$$

For the demand curves in equation (9.1), $mr_1(Q_1) = 20 - 2Q_1$ and $mr_2(Q_2) = 30 - Q_2$. Suppose that $c(Q) = Q + Q^2/4$ and hence $mc(Q) = 1 + Q/2$. The equations are then

$$\begin{aligned} 20 - 2Q_1 &= 1 + (Q_1 + Q_2)/2\,, \\ 30 - Q_2 &= 1 + (Q_1 + Q_2)/2\,. \end{aligned}$$

One can verify that the solution is $Q_1 = 4$ and $Q_2 = 18$, which corresponds to prices $P_1 = 16$ and $P_2 = 24$. Again, $P_1 < P_2$ as predicted.

Fixed capacity. Suppose you have a fixed capacity Q. Then you split this capacity to equate marginal revenue in the segments, solving

$$\begin{aligned} &mr_1(Q_1) = mr_2(Q_2)\,, \\ &Q_1 + Q_2 = Q. \end{aligned}$$

Sticking to the same demand curves and assuming $Q = 34$, these equations are

$$\begin{aligned} &20 - 2Q_1 = 30 - Q_2\,, \\ &Q_1 + Q_2 = 34\,. \end{aligned}$$

One can verify that the solution is $Q_1 = 8$ and $Q_2 = 26$, which correspond to prices $P_1 = 12$ and $P_2 = 17$. Once again, $P_1 < P_2$.

9.4 Wrap-up

By charging different prices to different market segments—assuming that the firm can observe some characteristic that differentiates demand—the firm can increase total surplus and appropriate more of this surplus. The main conclusion is that the firm charges a higher price in the market segment whose demand is less elastic.

Chapter 10

Implicit Market Segmentation (Screening)

10.1 Motives and objectives

Broadly

The explicit market segmentation studied in Chapter 9 is only partially successful. It is often impossible to explicitly segment a market. Even when possible, the segmentation will be crude. A software company many offer different prices for students and non-students, yet customers within each group are still quite heterogeneous. Owing to the lack of observability, the company cannot further customize prices to match the customers' valuations.

With explicit market segmentation, the firm imposes different trades on customers based on observable differences. For nonobservable differences, the firm can try another strategy: let the customers adapt the trades to their differences through the choices that they make. Such a strategy is called *implicit market segmentation* or *screening*.[1] What the customers do—adapting trades to their characteristics by their own choices—is called *self-selection*.

More specifically

This chapter studies screening with unit demand. Chapter 11 studies screening via nonlinear pricing with multi-unit demand. (We will define nonlinear pricing in that chapter.)

Paradoxically, in order to understand what firms should do given constraints on explicit market segmentation, it helps to first study what they would do when faced with no such constraints. Thus, we begin each of these two chapters (unit demand, then multi-unit demand) by studying the benchmark model of *perfect price discrimination*.

This model is only a hypothetical benchmark because it makes an impossible assumption about the firm's information: that the firm knows each buyer's valuations. With this information, the deal the firm offers each customer can be fully customized to his preferences, and there is no point in allowing the customer a choice of different deals.

But a firm that lacks such perfect information should allow its customers some discre-

[1] "Screening" is the standard term in economics. It is less descriptive than "implicit market segmentation", but it has only 2 syllables instead of 9. Do not confuse this topic with other distinct meanings of the English word "screening", such as screening job candidates by acquiring information about them through tests and interviews.

tion in what trades to execute, so that the trades are adapted to the customers' preferences via the choices they make. The selection of deals that the firm offers is called a *menu*. The set of possible menus is enormous and limited only by our imagination. Hence, we cannot give a simple answer to the question: "What is the optimal menu for each market?" Instead, we first describe the basic principles and then illustrate them by exploring several types of menus and working through examples. We study screening via product differentiation and bundling in this chapter, and then screening via nonlinear pricing in Chapter 11.

10.2 Perfect price discrimination with unit demand

Main idea

Let's take the idea of explicit market segmentation to an absurd limit. Suppose that (a) each customer has unit demand, (b) you know each customer's valuation (perfect observability), (c) arbitrage is not possible, and (d) charging each customer a different price involves no transaction cost. Your profit-maximizing strategy is then simple.

- *Surplus extraction.* If you trade with a customer, you charge him his valuation. Therefore, you get all the gains from trade that are generated.
- *Surplus generation.* It is then in your interest to maximize the gains from trade, so you sell the good to any customer whose valuation is at least as high as your marginal cost.

You have reached a profit-maximizing utopia: you *generate* and *extract* all gains from trade. This is called *perfect price discrimination* (PPD).

Perfect price discrimination yields an efficient outcome even when there is a fixed cost of production: because a firm receives all the variable surplus generated by a movie or invention, the firm pays the fixed cost of making the movie or developing the invention whenever such surplus exceeds the fixed cost. (The customers, however, end up with no surplus.)

An example

Suppose your firm produces a water purification system that you sell to small businesses. You have a constant marginal cost of €25,000. You have 13 potential customers with the following valuations (measured in €1000s).

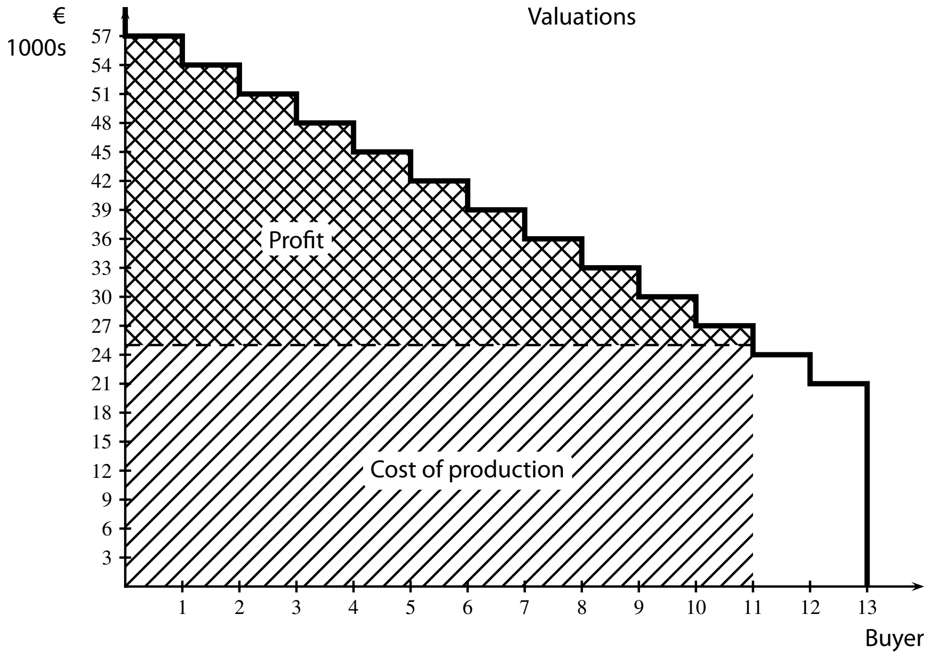
Buyer	1	2	3	4	5	6	7	8	9	10	11	12	13
Valuation	57	54	51	48	45	42	39	36	33	30	27	24	21

(We deliberately use the data from Exercise 2.1 and Exercise 7.3 to help you see the links between the associated topics.)

If you charged a uniform price, then your demand curve would be the same as the one given in Exercise 7.3. Our goal is to compare perfect price discrimination with the uniform-price outcome calculated in that exercise.

Since you extract any gains from trade, you should sell to the 11 customers whose valuation are at least 25. You should charge each one his valuation. Trade is efficient. You get all the gains from trade and the customers get no surplus.

Figure 10.1



10.3 Screening: basic framework

If you recognize the heterogeneity of the customers but lack the information needed for explicit segmentation, you can still try to *implicitly* segment the market. You propose options to your customers and let the trades adapt themselves to the customers' characteristics through the choices that *they* make.

Let's suppose that you have ten customers. You have a set of potential deals you can offer them. For example, a deal might be to trade a particular quantity of your single product for a sum of money. You might also have several different products you can offer, in which case the set of potential deals is more varied.

In the case of perfect price discrimination, you can identify each customer and you know his preferences over deals. You offer him the deal that gives you the highest profit, subject to the constraint that he be willing to accept the deal (when his alternative is to not trade with you at all). There is no point in offering a customer more than one deal. If we number the customers from 1 to 10, we could list the deals that you offer them by $D_1, \ldots, D_{10}$.

In the case of screening, you know the ten different preferences that your customers have but you cannot identify which customer has which preferences. For example, suppose you manage a gas station. Then you know how many customers live in each geographical location around your gas station and you also know that a customer's location determines the premium he is willing to pay in order to buy gasoline from you instead of from one of your competitors. However, when a customer drives up to your gas station, you cannot tell where he lives and hence you cannot customize your price to this information. Similarly, suppose you manage an airline. Through marketing studies you can learn how many of your customers are business travelers and how many are leisure travelers. However, when a customer calls to purchase a ticket, you cannot tell what kind he is. Hence, you must offer all customers the same set of options.

From your information about the distribution of preferences among your customers, you can determine which option each *type* of customer picks from your menu. (As when we studied bargaining in Chapter 2, we assume that, when a customer is indifferent between two options, he selects the one that maximizes your profit; you can then tweak the menu so that he has a strict preference for that option.) We can list the deals that the different types of customers pick by $D_1, \ldots, D_{10}$. There is no point in offering deals that are not picked, so this list might as well also be the menu you offer. Your problem is to design this list in a way that maximizes your profit.

Compared to the case of perfect price discrimination, you face an additional constraint: since customers themselves pick from the menu, they must prefer the deal intended for them over all other deals on the menu.

We have thus framed your decision problems in the two cases of perfect price discrimination and screening in a way that allows a nice comparison between them. In both cases:

- you pick a deal for each type of customer;
- each deal gives you a certain profit;
- you want to pick the deals that give you the highest combined profit.

However, you have an extra constraint in the case of screening.

- In both cases, each type of customer must weakly prefer his intended deal over not trading at all. This is called the *participation constraint*.
- With screening but not with perfect price discrimination, each type of customer must weakly prefer his intended deal over the other deals that you offer. This is called the *self-selection constraint*.

Typically, each deal has two components: fixed quantities of the good(s) and service(s) that you sell, which we call a *bundle*, and the amount you charge for the bundle, which we call a *tariff*.[2] The same bundle should not appear twice on a menu with two different

[2] "Price" is the amount charged per unit of a good, whereas "tariff" is the total amount charged for a bundle. For example, if a firm charges a uniform price of $5, then the tariff for a bundle of 8 units is $40. If the bundle contains a single unit of a single good, then price and tariff are identical.

tariffs, since any customer will ignore the higher-tariff option. The design problem is thus to choose which bundles appear on the menu and what tariff to charge for each bundle.

Therefore, in order to study screening, we must allow for multiple bundles. There are several ways in which bundles can be differentiated.

1. *By varying characteristics of differentiated substitute goods.* If it is possible to produce differentiated varieties of the good or service and if each customer buys at most one of these, then the bundles can differ by the versions they contain. In the software industry, in which a firm may offer a "standard" and "professional" version of a product, this is called *versioning*; more generally, it is called *screening via product differentiation*. We study this in Section 10.4.
2. *By varying combinations of complementary goods.* If the firm sells multiple products of which each customer may buy more than one, then the bundles can contain different combinations of the goods. For example, office suites contain several software programs, and their price is not the sum of the prices of the individual components. This is called *bundling* and is studied in Section 10.8.
3. *By varying quantities of a single good.* If each customer may buy different quantities of a single good (multi-unit demand), then the bundles can differ by the quantities they contain. This is called *nonlinear pricing* and is studied in Chapter 11.

The self-selection constraint means that the deal you can offer one customer depends on the deals offered to all the other customers. This makes the screening problem more complicated than perfect price discrimination, particularly if there are many types of customers and many potential bundles. However, we can carve out some simple problems by restricting the types of customers and the potential bundles. In this chapter, we assume there are only two, three, or four possible bundles and only two or three types of customers.

10.4 Differentiated products as screening

Overview

In screening via product differentiation, differentiated versions of what is essentially the same product are implicitly targeted to different market segments. The versions are designed with the screening objective in mind.

For example, consider an airline that offers restricted and unrestricted tickets. The airline knows it has heterogeneous customers—business travelers with inelastic demand and leisure travelers with elastic demand—and would like to segment the market. However, it cannot distinguish between the travelers at the time of purchase. If it sells a single product, then the airline must choose between: (a) charging a high price, thereby extracting surplus from the business travelers but losing gains from trade with the leisure travelers; or (b) charging a low price, thereby generating gains from trade with all travelers but leaving the business travelers with more of these gains. To overcome this trade-off, the airline looks for a product characteristic—convenience and flexibility—that business travelers value more than leisure travelers; it then sells a discounted "damaged" product line in which this characteristic is absent. The airline may thus sell the undamaged product at a high price to the business travelers yet still extract gains from the leisure travelers.

The airline's implicit market segmentation is not as effective as perfect price discrimination. The airline must be conscious of the self-selection constraints. Even though the restricted tickets are not convenient, the gap between the restricted and unrestricted fares cannot be set arbitrarily high without prompting business travelers to switch to the restricted tickets. Hence, the airline may have to keep business fares lower than it would like and leisure fares higher than it would like. Furthermore, offering damaged goods to the leisure travelers reduces the gains from trade that can be extracted.

Product differentiation to satisfy heterogeneous tastes

Screening is not the only reason that firms offer different versions of products. They also try to match their customers' heterogeneous tastes. For example, the Coca-Cola company offers regular Coke and Diet Coke simply because some customers prefer a low-calorie product.

It is hard to define formally what distinguishes product differentiation as screening (e.g., the airlines example) from product differentiation as a means to satisfy heterogeneous tastes (e.g., different types of soda). However, we can say that product differentiation is motivated by screening whenever it would disappear or be much more limited if the firm had the ability to explicitly segment its market. The Coca-Cola company would continue to produce Coke, Fanta Orange, and Sprite to satisfy the different tastes of its customers even if it could observe each customer's tastes and charge different prices according to those tastes. In contrast, the airlines would not impose Saturday-night-stay requirements if customers entered ticket agencies with "business traveler" or "leisure traveler" tattooed on their foreheads.

Most screening via product differentiation involves quality differentiation such that the differences in quality are exaggerated compared with those resulting from perfect information and such that the differences in prices exceed the differences in production costs for the various qualities. This is how the French economist Jules Dupuit recognized screening by passenger railroads. In 1849 he wrote in the *Annales de Ponts et Chaussées*:

> It is not because a few thousand francs which would have to be spent to put a roof over the third-class carriages or to upholster the third-class seats that some company or other has open carriages with wooden benches.... What the company is trying to do is to prevent the passengers who can pay the second-class fare from traveling third-class; it hits the poor, not because it wants to hurt them, but to frighten the rich.[3]

[3] "On Tolls and Transport Charges," translation in *International Economic Papers*. London: MacMillan, 1962.

In the same spirit, Leon Walras made the following observation about French railway companies in 1875:

> In reality, the companies consider, rightly or wrongly, the average price of 7.66c … to be the profit-maximizing price [of passenger service]; but they do not want to miss the chance of taking more from passengers willing to pay more, nor to turn away passengers not willing to pay as much. This is why there are three separate classes, and great efforts made to accentuate on the one hand the advantages of the first class and on the other hand the disadvantages of the third class. When some time ago there was an outcry that third class coaches should have windows fitted as laid down in the regulations for 1857–8, and now when heating is demanded of them in the winter, people complain about the meanness of the companies without understanding its true cause. If the third class coaches were comfortable enough for many first and second class passengers to go in them, total net product would fall. That is all there is to it.[4]

10.5 Example: Software versioning

Scenario

Suppose that your company produces a graphics arts software package called ArtWorks. You might broadly distinguish between two very different types of customers. Some of your customers—such as advertising agencies and other graphics art designers—use the package every day as part of high-value art services. Other customers are just folks using it in their spare time for fun and hobbies. For simplicity, suppose that you do have exactly two types of customers: all professional users have identical valuations (these are your high-valuation customers) and all hobbyists have identical valuations (these are your low-valuation customers).

The pro's valuation of the product is \$700; the hobbyist's valuation is \$220. Software is cheap to reproduce and distribute, so your marginal cost is only \$20.[5]

We are interested in how your firm's pricing decision depends on the number of high- and low-valuation customers. Let H denote the number of high-valuation customers and L the number of low-valuation customers. For various menus, we write out your profit as a function of H and L. For example, if you have a profit of \$30 per high-valuation customer and \$10 per low-valuation customer, then your total profit is (H × \$30) + (L × \$10).

[4] Translation by P. Holmes in *Journal of Public Economics*, 13:81–100, 1980.

[5] You might have a large fixed cost, but this is not relevant to this pricing problem. We ask what prices you should charge given that you are operating, not whether you should start up or shut down.

Uniform pricing

We first suppose that you use uniform pricing. Your demand curve has just two steps in it, so this is a particularly simple version of uniform pricing from Chapter 7. There are only two reasonable options:

1. Charge \$700 and sell only to the professionals. Your profit is $H \times \$680$.
2. Charge \$220 and sell to everyone. Your profit is $(H + L) \times \$200$.

If L is very low compared to H, then your profit is higher by charging the high price (margin matters); if L is very high compared to H then your profit is higher by charging the low price (volume matters).

Benchmark: Explicit market segmentation

What you wish you could do is perfectly price discriminate. Then you sell to each customer and charge each his respective valuation. That is, you charge the professionals \$700 and the hobbyists \$220. Since your unit cost is \$20, your profit is $(H \times \$680) + (L \times \$200)$. You get both margin and volume.

Screening

Suppose you cannot explicitly segment the market. How can you pursue the goals of explicit market segmentation and avoid the fallback strategy of uniform pricing?

Your best strategy is to create two versions of your software. ArtWorks has many features, so you can create a stripped-down version that is targeted to the hobbyists. This is truly a “damaged good”, since the stripped-down version is worse than the full version, yet the two versions cost the exact same amount to make. (Once you have developed the software, the reproduction and distribution cost is the same whether or not you load everything on the CD. In fact, there is even an extra fixed cost involved in created the stunted version.) You decide to brand the full version as ArtWorks Pro and the stripped-down version as ArtWorks Essentials.

You have many options when deciding what to omit from ArtWorks Essentials. This is a complex design problem that is beyond the scope of this chapter. However, you want to drop features that matter a lot to the professionals (so that they are willing to pay a large premium to have the Pro version rather than the Essentials version) and that matter little to the hobbyists (so that you do not reduce too much the value of the product to them).

Table 10.1

Type of customer	Valuation	
	ArtWorks Pro	ArtWorks Essentials
Professionals	\$700	\$300
Hobbyists	\$220	\$200

Suppose, after careful designing of the two versions, the valuations of your customers are as shown in Table 10.1. These numbers are fictitious but not unreasonable. The professionals really make a lot of use of this software. In fact, the only reason they are not willing to pay much more than \$700 for it is that there are competitors with similar products selling for under \$1000. Once you drop a few key features, such as the ability to calibrate Pantone colors or produce color-separated artwork to send to printers, the value of this product drops off quickly. Yet such features have almost no value to the hobbyists.

Now your task is to price the two versions, keeping in mind that the Pro version is targeted to professionals and the Essentials version is targeted to hobbyists.

A naïve solution would be to price the Pro version at the professional's valuation of \$700 and to price the Essentials version to the hobbyist's valuation of \$200, in order to extract all the surplus. Think for a moment what would happen.

Your professional customers would buy the Essentials version! This would give them a surplus of $300 - 200 = 100$ versus the surplus of \$0 when buying ArtWorks Pro. In fancier words, this menu would violate the pro customers' self-selection constraint because they prefer the deal you were targeting to the hobbyists over the deal you were targeting to them.

Loosely, your objective is to set the prices "as high as possible" subject to the self-selection and participation constraints. The binding parts of these constraints are as follows.

- *Low types' participation constraint.* You must not set the price of ArtWorks Essentials so high that the hobbyists prefer not to purchase the product at all.
- *High types' self-selection constraint.* You must not set the price of ArtWorks Pro so high that the high types prefer to buy ArtWorks Essentials.

With these two principles as a guide, we can determine the optimal prices. You should set the price of ArtWorks Essentials to the hobbyists' valuation of \$200. You should set the price of ArtWorks Pro to the highest amount such that the professionals weakly prefer to purchase it rather than ArtWorks Essentials. That is, for the Pro version, you charge the price of Essentials (\$200) plus the professionals' incremental willingness to pay for Pro over Essentials ($700 - 300 = 400$), for a total price of \$600.

Since your unit cost is \$20, your profit is now \$580 per high-valuation customer and \$180 per low-valuation customer. Your total profit is $(H \times \$580) + (L \times \$180)$.

Comparison of the menus

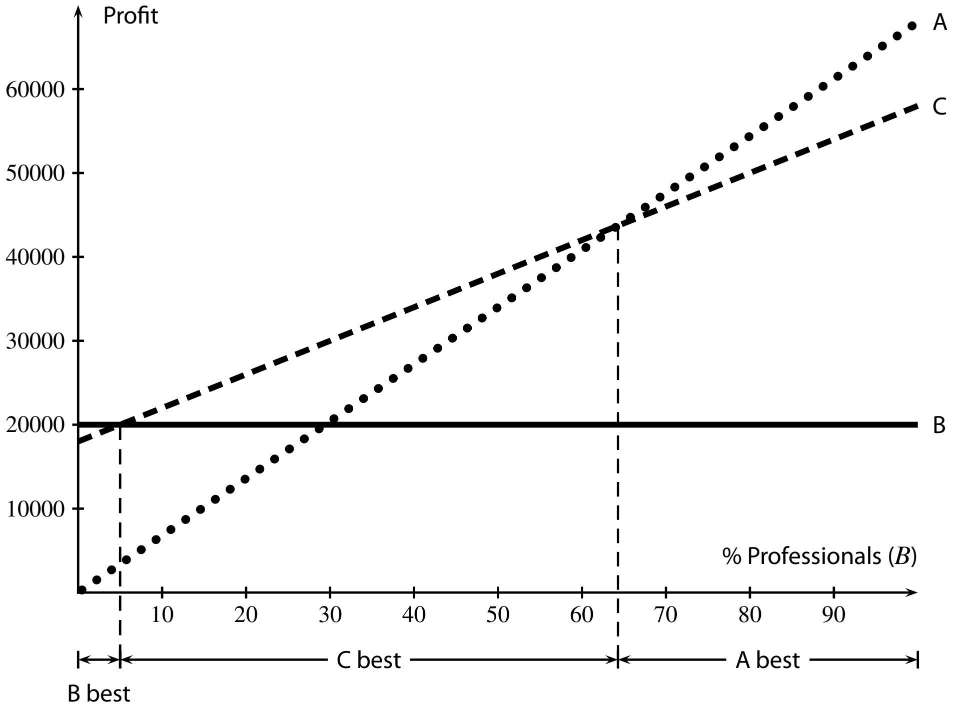
We have thus narrowed your options down to three potentially good ones:

A. sell ArtWorks Pro to professionals for \$700 [profit = $H \times \$680$];
B. sell ArtWorks Pro to all customers for \$220 [profit = $(H + L) \times \$200$];
C. sell ArtWorks Pro to professionals for \$600 and ArtWorks Essentials to hobbyists for \$200 [profit = $(H \times \$580) + (L \times \$180)$].

The ranking of these menus depends on how many customers of each type you have. Fix the total number of customers at 100, so that H is the percentage that are high-valuation customers and $L = 100 - H$. Figure 10.2 shows the profit of each menu as a function of H.

Figure 10.2

Option	Profit as a function of H	Shown in graph as	
A	$680H$	dotted line	● ● ● ● ●
B	$100 \times 200 = 20{,}000$	solid line	━━━
C	$H \times 580 + (100 - H) \times 180 = 18{,}000 + 400H$	dashed line	▬ ▬ ▬ ▬



We can now see the trade-off between these menus. Menus A and B do not involve true screening or separation of the types, because each customer gets the same deal (in fact, only one deal is offered). The trade-off between these two menus is the one we studied in Chapter 7 for a firm with market power. You can price low in order to generate extra gains from trade, but at the expense of giving the customers more of these gains. It is better to price high (menu A) when the high end of the market is large (above 64%) and to price low (menu B) when the low end of the market is large (above 95%).

Screening (menu C) is an attempt to "have your cake and eat it too". You sell to the high-valuation customers at a high price in order to extract as much surplus from them as possible. However, you reap extra gains from trade with the low-valuation customers by selling them an inefficient damaged good at a price that does not make it attractive for high-valuation customers. This menu is optimal when the market has a mix of high and low types.

10.6 Example: Airlines

Exercises 10.1–10.4 work through another example in which a “damaged good” is used to screen customers. The next paragraphs set up the example, which is framed in terms of an airline that uses restricted tickets to screen business travelers and leisure travelers. You are a manager of this airline.

One of the decisions you must make when using differentiated products to screen customers is the design of the product mix. For example, what restrictions should you, as an airline manager, impose on restricted fares? How many categories of service and fares should you have? The other decision is how much to charge for each product. These are not independent decisions; you cannot know the profit from a given product mix unless you solve the pricing problem for that mix.

To keep this exercise from being too complex, we simplify the product-mix decision by assuming that only two possible products are available: unrestricted tickets and restricted tickets. You must decide whether to offer one or both of these products and you must choose a price for each product you offer.

In practice, some restrictions (e.g., advance purchase requirements and change fees) reduce the cost of carrying a passenger by a small amount whereas others (e.g., Saturday-night-stay requirements) have no effect on cost. To emphasize that the motive for introducing a lower-quality product is not merely to lower cost, assume that an unrestricted ticket and a restricted ticket have exactly the same cost for your airline: €300 per ticket.

We assume that each customer buys either 0 or 1 ticket. The valuations for the two types of tickets and the two types of customers are shown in Table 10.2. (Note that the two types of tickets are substitutes; a traveler never buys both.)

Table 10.2

Type of	Valuation	
customer	Unrestricted	Restricted
Business	€1000	€600
Leisure	€600	€500

(*Airline’s cost is €300 per ticket.*)

Let B denote the number of business travelers and L the number of leisure travelers. For many of the questions that follow, you will be asked to write out a formula for your profit as a function of B and L.

Exercise 10.1. (Uniform pricing) Suppose you decide to offer only unrestricted tickets. There are only two prices you might charge (depending on the values of B and L). What are they? For each of the two prices, (i) describe who purchases the tickets and (ii) write your profit as a function of B and L.

Exercise 10.2. (Benchmark: Explicit market segmentation) Suppose your airline can perfectly price discriminate.

a. What products do you offer (unrestricted, restricted, both, or neither), and what price(s) do you charge each market segment?

b. What is your total profit as a function of B and L?

Exercise 10.3. (Screening) Suppose you offer both restricted and unrestricted tickets, with the intention that the business travelers buy the unrestricted ticket and the leisure travelers buy the restricted ticket. Note that the business travelers are willing to pay more than the leisure travelers for both types of tickets. The important fact is that the extra amount the business travelers are willing to pay for the unrestricted ticket is more than the extra amount the leisure travelers are willing to pay.

a. What would happen if you attempted to extract all the surplus by setting the price of the unrestricted ticket equal to the business travelers' valuation (€1000) and the price of the restricted ticket equal to the leisure travelers' valuation (€500)?

b. The binding parts of the participation and self-selection constraints are:

1. you must not set the price of restricted tickets so high that the leisure travelers prefer not to travel at all;
2. you must not set the price of unrestricted tickets so high that the business travelers prefer to buy restricted tickets.

Identify these constraints by name.

c. Calculate your optimal prices.

d. Calculate the profit (given the optimal prices) as a function of B and L.

Exercise 10.4. (Comparison) You have thus narrowed your options down to three potentially good ones:

A. sell only unrestricted tickets at a high price to business travelers only;
B. sell only unrestricted tickets at a lower price to all travelers;
C. sell unrestricted tickets to business travelers and restricted tickets to leisure travelers at the prices you obtained in Exercise 10.3.

The optimal prices you obtained for options A, B, and C do not depend on the numbers of business and leisure travelers. However, the ranking of the options does. The purpose of this exercise is to see this relationship.

a. Suppose there are 50 business travelers and 50 leisure travelers. Calculate your profit for each of the three options and find which option has the highest profit.

b. Suppose there are 99 business travelers and 1 leisure traveler. How do the profits of the three options compare?

c. Suppose there are 99 leisure travelers and 1 business traveler. How do the profits of the three options compare?

d. Fix the total number of customers at 100, so that B is the percentage that are business travelers and L is the percentage that are leisure travelers. On a single graph, plot the profit as a function of B (for B between 0 and 100) for each of the three options. (For each option, profit as a function of B is a line and hence is easy to draw by hand.) For each option, identify the region on the graph (i.e., the range of values of B) for which that option yields the highest profit.

10.7 Intertemporal product differentiation

The key to understanding how to price a physical product at different times and locations is to realize that *products are differentiated not just by their physical characteristics* (color, flavor, features, quality, ...) *but also by where and when they are sold.* On the consumption end, physically identical products at different locations or times are not perfect substitutes (e.g., consumers are willing to pay a premium to buy gasoline from the nearest gas station; commuters care a lot about the time at which they take a train.) On the production side, products at different locations and times have different costs (e.g., some locations involve higher transportation costs; the production cost for a new product may fall over time as a result of the learning curve; the marginal cost of supplying transportation or utility services is higher at peak times than at off-peak times).

Thus, we have positive interest rates because (a) people value, on the margin, consumption today more than consumption in the future and (b) it is not possible to costlessly shift output from the future to today. Furthermore, a chain of stores may have different prices at different locations because (a) the nearby customers differ from one store to another and (b) the price differences cannot be costlessly arbitraged owing to transportation costs.

Intertemporal product differentiation can be used to screen customers. We should first note, however, that many intertemporal price differences are not due to screening. This is true, for example, of much peak-load pricing. A firm's basic cost curve may be the same at peak and off-peak times, but in equilibrium the firm sells more and and has higher marginal cost at peak times than at off-peak times. Thus, even a public electric company that tries to maximize total surplus may charge different rates at different times of the day, because the marginal cost of increasing output is much higher at times that are already at capacity than at times below capacity. A profit-maximizing utility company whose customers were identical (making screening irrelevant) would further differentiate peak and off-peak prices because the customers' elasticity of demand would be different at different times of day.

Intertemporal production differentiation as screening is common in the publishing and entertainment industries. If a publisher of a new novel could perfectly price discriminate, it would sell the novel right away to all customers who were interested, with each customer paying his valuation. Because it cannot directly observe customers' valuations, the publisher looks for product features for which the high-valuation customers are willing to pay a higher premium than are low-valuation customers: (a) having a hardcover instead of paperback and (b) reading the book just after it is released. It thus first introduces a hardcover edition at a high price, which it sells mainly to the high-valuation customers. Later it charges a lower price for a paperback edition (and also discounts the hardcover edition), selling these mainly to the low-valuation customers. We need to invoke screening to understand this, because the delay in introducing the paperback edition does not reduce cost but merely damages the product line.

10.8 Bundling

What is bundling?

Bundling means taking two or more distinct goods and selling them as a bundle at a tariff that is not the sum of the prices of the individual components. For example, a symphony may offer a subscription to a concert series, a magazine publisher may offer a discount when someone subscribes to two magazines, and Microsoft includes some software packages in its operating systems rather than selling them separately.

Restrictions on bundling

There is a bundle discount (resp., premium) if the bundles costs less (resp., more) than the sum of the tariffs of the components. Arbitrage limits bundle discounts, because people could make a profit by buying the bundle and selling off the individual components. Masquerading limits bundle premia, because a customer could buy the individual components under multiple identities rather than buying the bundle.

Bundling as a way to smooth preferences

The inability to perfectly price discriminate has a greater impact on a firm's profit the more heterogeneous its customers' preferences are. If most customers have almost the same valuation, then the firm can generate and extract almost all the gains from trade even if it has to offer the same deal to all customers. One use of bundling is to make customers' preferences more homogeneous, because preferences over the bundle are often less heterogeneous than preferences over individual components.

Consider a software firm that makes a desktop publisher, a spreadsheet, and a calendar program. For simplicity, suppose each customer views these programs as neither substitutes nor complements; that is, his valuation of a program does not depend on which other program(s) he buys. Suppose the firm has three types of customers with valuations as shown in Table 10.3.

Table 10.3

Type of	Valuation ($)		
customer	Publisher	Spreadsheet	Calendar
A	90	30	30
B	30	90	20
C	20	20	90

For example, if a type-B customer buys the publisher and spreadsheet for \$80, then his total valuation for the two programs is \$30 + \$90 = \$120 and his surplus is \$120 − \$80 = \$40. Let the types be in equal proportions; we treat each type as a single customer.

The marginal cost of software is small; for simplicity suppose it is zero, so the publisher maximizes revenue. (The fixed development cost affects the decision of whether or not to develop the software but not the pricing once it has been developed.)

The efficient trade is to sell the three programs to all types of customers. The total gains from trade (which equal the sum of the valuations) are \$420.

Suppose you do not bundle the programs. For each program, you can either sell 1 unit for \$90, 2 units for \$30 each, or 3 units for \$20 each. Your revenue is highest when you charge \$90 for each program. Your total revenue (profit) is \$270. Customers get no surplus, but you have not generated all possible gains from trade. The deadweight loss is \$420 − \$270 = \$150.

Suppose that instead you bundle the three programs as an office suite. The customer valuations for the bundle are shown in Table 10.4.

Table 10.4

Type of	Valuation ($)
customer	Office suite
A	150
B	140
C	130

There is no longer much disparity between the customers' valuations. You can sell 1 unit for \$150, 2 for \$140, or 3 for \$130. Selling 3 units yields the highest revenue: \$390. There is no deadweight loss and you have extracted almost all the gains from trade.

Mixed bundling as screening

If bundling does not achieve enough homogeneity of valuations, then the firm can offer both the bundle and some of its components (or various types of bundles) as a screening mechanism to deal with the residual heterogeneity of valuations.

Suppose that, in our software example, the customers' valuations are as shown in Table 10.5.

Table 10.5

Type of	Valuation ($)			
customer	Publisher	Spreadsheet	Calendar	Office suite
A	90	50	70	210
B	60	90	20	170
C	10	10	90	110

If you price the three programs separately, then your optimal prices, sales, and revenues are as shown in Table 10.6.

Table 10.6

	Publisher	Spreadsheet	Calendar
Price	60	50	70
Sales	2	2	2
Revenue	120	100	140

Pricing without bundling
Total revenue = $360

If instead you offer only the office suite, then you should sell 2 units at $170 each for revenue of $340. This revenue is even lower than when you price individually because you do not make any sales to customer C. However, if you screen by offering the calendar program on its own for $90 (in addition to the bundle for $170), then customer C buys it, while customers A and B prefer to purchase the office suite. Your total revenue is now $170 + $170 + $90 = $430.

You can improve further on this menu. Customer A is keeping $40 ($210 − $170) of the surplus. If you raise the price on the office suite to extract this surplus, then customer B decides not to purchase anything, and so your profit goes down. The solution is to raise the price on the office suite to $210 and to offer another bundle containing only the publisher and the spreadsheet, "Office Suite—Small Business Edition", for $150 so that customer B buys it. Your total profit is then $210 + $150 + $90 = $450.

In summary, your optimal menu is shown in Table 10.7.

Table 10.7

	Bundle		
	Calendar	Publisher Spreadsheet	Publisher Spreadsheet Calendar
Price	90	150	210
Sold to	C	B	A

(A complete demonstration that this is the optimal menu involves comparing it with all possible combinations of bundles. There are seven possible bundles—three single products, three pairs of products, and the full suite—and 35 possible combinations of bundles. We have only compared four of these combinations.)

Exercise 10.5. You are in charge of sales for a symphony that has a mini-season of two concerts, one featuring music by Wagner and the other featuring new music by John Harbison. Some in the potential audience like old music much more than contemporary, others like both equally, and others like contemporary much more than old music. You must decide how to price the individual tickets and the series in order to maximize profit.

We make some simplifying assumptions: The symphony has a very large concert hall relative to its popularity and hence capacity constraints are not an issue; all seats are equally desirable; and the marginal cost of each concertgoer is zero. Hence, the symphony's goal is to maximize revenue.

Assume the market is highly segmented, with only three types of customers. There is one customer of each type (which is equivalent to assuming there are equal numbers of each type). The valuations of these customers for each of the two concerts are as shown in Table E10.1

Table E10.1

Type of	Valuation	
customer	Wagner	Harbison
A	50	5
B	40	40
C	5	50

A customer may go to one or both of the concerts. A customer's valuation of a bundle equals the sum of his valuations of the concerts in the bundle.

a. (Benchmark: Explicit market segmentation) Suppose, hypothetically, that you can perfectly price discriminate. How much should you charge each type of customer for each concert? What is the total revenue?

b. (No bundling) Return to the real situation in which you cannot perfectly price discriminate. Suppose you sell only individual tickets. What price should you charge for each concert? Who buys tickets? What is the total revenue?

c. (Pure bundling) Suppose you offer only the series and do not sell individual tickets. What price should you charge? Who buys? What is your total revenue?

d. (Mixed bundling) Suppose you sell individual tickets and also the series. What prices should you charge? What does each type of customer buy? What is the total revenue?

10.9 Wrap-up

Even when a firm cannot explicitly segment its market, it should pay attention to the composition of its market and not merely choose a price based on its aggregate demand curve. By offering a menu of trading options to customers, trades can be customized to individual preferences through the customers' own choices. By careful design of the menu, a firm can partially achieve the goals of generating all gains from trade and extracting those gains for itself.

Additional exercises

Exercise 10.6. (Perfect price discrimination) Suppose that you produce an indivisible good at a constant marginal cost of \$14. You have 12 customers, each of whom buys at most one unit of the good. They have the following valuations:

$$\$10, \$12, \$15, \$16, \$17, \$19, \$22, \$22, \$25, \$26, \$27, \$30.$$

a. If you cannot price discriminate, what price should you charge? What is your total profit? (A spreadsheet may be useful.)

b. Take as the status quo your decision in the previous part. The purpose of this part is to demonstrate that the outcome is not economically efficient. Show that there is a customer who is currently not buying the product and to whom you could sell the good at a price that would make both you and that customer better off (if you could identify the customer and if the transaction would not disturb your current sales to other customers).

c. Now suppose that you can perfectly price discriminate (charge each customer a different price). Which customers do you sell to, how much do you charge, and what is your total profit?

Exercise 10.7. (Perfect price discrimination) Consider a market in which each customer purchases at most one unit of an indivisible good. Consider a shift from a monopolist who cannot price discriminate to the same monopolist with perfect price discrimination. Which customers are better off, worse off, or indifferent?

Exercise 10.8. (Perfect price discrimination) Zahra, a profit-maximizing entrepreneur, sells an indivisible product of which each customer buys at most one unit. She produces the good at a constant marginal cost of $5. Zahra's market contains many customers, whose diverse valuations she knows.However, she is initially prohibited by law from price discrimination. She chooses to charge $10, which results in sales to 10,000 customers. She calculates that these customers obtain a total of $50,000 in consumer surplus. Suppose now that the prohibition is lifted and so Zahra can engage in perfect price discrimination. Based on this limited information, what can you say about (a) how many customers she will sell to, (b) what range of prices she will charge, and (c) by how much her profit will go up? Be as specific as possible and explain your answers. (*Note*: Do not assume linear demand, since that would not be in the "limited information" spirit of this question and since the data are not consistent with linear demand.)

Exercise 10.9. (Screening via differentiated products) The purpose of a numerical example like that of Exercises 10.1–10.4 is to obtain intuition about real-world pricing problems by working through a simple example that you can "touch and feel". Imagine that six months after reading this book you find yourself with a real-world pricing problem that is similar—in that you consider offering multiple quality levels and there are two broad market segments—although without the stark simplicity of our example. Write a brief (e.g., three-paragraph) summary of the intuition you have obtained from the numerical example. You should think of this summary as a brief memo or presentation whose purpose is to analyze the problem for your colleagues.

Exercise 10.10. (Bundling) You are a movie distributor with two movies to offer: an arts film called *Sorrow and Loneliness* and an action film called *Death Machine*. You distribute to three theaters: "Supermall Megaplex", "Downtown Deluxe", and "Ethereal Visions". You know that the amounts each theater would pay for these movies are as shown in Table E10.2.

Table E10.2

	Valuation	
Theater	*Sorrow*	*Death*
Supermall	$60	$100
Downtown	$90	$60
Ethereal	$120	$10

(The amount a theater is willing to pay for one of these movies does not depend on whether or not it chooses to buy the other movie.) Your marginal cost is zero and so your objective is to maximize revenue. In each of the following problems, you cannot explicitly segment your market.

a. Suppose that you do not bundle. What is the optimal price of *Sorrow* and what is the optimal price of *Death*? What is your total revenue?

b. Now suppose that you offer the two movies only as a bundle (pure bundling). What is the optimal price for the bundle? What is your revenue?

c. Suppose you can engage in mixed bundling. What is your optimal pricing strategy? What does each theater buy? What is your revenue?

Exercise 10.11. (Bundling) You are a monopolist selling two different types of concert tickets, for rock music and world music. You have zero marginal cost and hence your objective is to maximize revenue. You face three groups of potential customers, with an equal number of customers in each group. Table E10.3 summarizes the valuation of each group for each concert.

Table E10.3

	Valuation	
Type	Rock	World
A	5	60
B	35	65
C	40	70

a. (Pure bundling): What is the optimal price for the bundle of both tickets? What is your profit?

b. (Mixed bundling) Find one mixed bundle pricing strategy that gives you higher profit than your answer for pure bundling.

Chapter 11

Nonlinear Pricing

11.1 Motivation and objectives

In Chapter 10 we looked at sophisticated pricing strategies that implicitly differentiate among customers. The entire chapter examined models with unit demand, because there are additional things to consider in the case of multi-unit demand.

In particular, the pricing strategies we have studied always assume that a firm posts a price and allows each customer to buy as much or as little as he wants at that price. This is called *linear pricing*. Anything else is called nonlinear pricing. This chapter is about the advantages of nonlinear pricing when it is feasible.

The terms "linear" and "nonlinear" have to do with the graph of the price schedule. A price schedule is just a function $t(Q)$ that states the tariff (the total amount) a customer must pay to buy Q units. If the firm uses linear pricing with the price P then the price schedule is simply $t(Q) = PQ$; it is a linear function whose graph is a line that starts at the origin. Any other pricing schedule will be a nonlinear function. Nonlinear price schedules may have volume discounts (the average tariff per unit is decreasing in the number of units). Examples include "one for 59 cents or two for a dollar" and two-part tariffs (Section 11.5). Price schedules may also have quantity premia (e.g., "your first movie is offered to you at half price") as well as quantity restrictions (e.g., "the minimum order is 10 shirts").

11.2 Perfect price discrimination, revisited

We revisit perfect price discrimination but now for multi-unit demand.

With perfect price discrimination, we remove all constraints that would keep a firm from generating and extracting all gains from trade. Therefore:

1. to generate all gains from trade, what we sell to each customer must be efficient;
2. to extract all gains from trade, the amount we charge to each customer must equal his valuation for what we sell him.

One customer with multi-unit demand

Now suppose you have one customer who may be interested in buying any quantity of the good. The good may be divisible or indivisible. Suppose that you know the customer's entire valuation curve.

If you charge a certain price and let the customer decide how much to buy at this price, then you are stuck in the same situation as in Chapter 7. The customer purchases only up to the point where marginal valuation equals the price you charge. Because marginal valuation is decreasing, his average valuation exceeds his marginal valuation—and hence exceeds your per-unit price. Thus, he takes home consumer surplus. If you raise the price to reduce this surplus, then the customer cuts back on his demand, thereby reducing the total gains from trade.

The solution is to offer the customer only one option: to purchase the efficient quantity Q^e. To extract the gains from trade, you should charge the customer his total valuation for Q^e.

For example, suppose you have the following data:

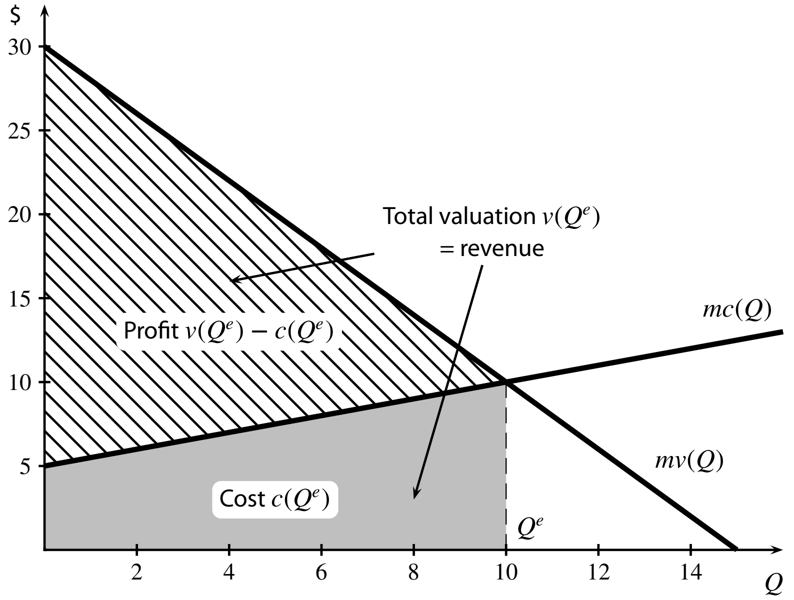
$$\begin{aligned} v(Q) &= 30Q - Q^2, \\ mv(Q) &= 30 - 2Q, \\ c(Q) &= 5Q + \tfrac{1}{4}Q^2, \\ mc(Q) &= 5 + \tfrac{1}{2}Q. \end{aligned}$$

To find the efficient quantity, you set $MV = MC$; that is, you solve $30-2Q = 5+\frac{1}{2}Q$, which yields $Q^e = 10$. The customer's total valuation for 10 units is $v(10) = (30\times 10)-10^2 = 200$. You say to the customer: "You can buy 10 units for \$200; if you don't like this deal then we won't trade at all." The consumer gets the same surplus (i.e., zero) whether or not he trades, so he accepts your proposal.[1]

Your cost is $c(10) = 50 + \frac{1}{4}10^2 = 75$, and your profit is $v(Q) - c(Q) = 125$. These values are illustrated in Figure 11.1.

[1] As usual, in practice you would lower the amount you charge by the smallest currency unit to make sure he strictly prefers accepting your deal over walking away.

Figure 11.1



Multiple customers

This scenario supposes that you have many customers and that you know each customer's valuations. You would like to deal with each customer separately, using one of the pricing strategies for a single customer that we just studied. You then generate all possible gains from trade and extract them for the firm.

Because you generate all gains from trade and extract them for the firm, the graphical illustration of the outcome at the market level is the same as in Figure 11.1—but with one difference of interpretation. In Figure 11.1, $mv(Q)$ is the single customer's marginal valuation curve; with many customers, $mv(Q)$ is the *market* marginal valuation curve (i.e., the market inverse demand curve $p(Q)$).

The calculations of the quantities you sell depend on whether your marginal cost is constant or variable. If your marginal cost is constant, then each single-customer pricing problem is independent from the others. You sell to each customer the quantity that equates his marginal valuation to your marginal cost, and you charge him his valuation for this quantity.

11.3 More on nonlinear pricing

Consumer choice

We assume that, *given a price schedule $t(Q)$, the consumer ranks the possible quantities by the consumer surplus $v(Q) - t(Q)$ that they generate.* Hence, when choosing between two quantities, the consumer chooses the higher quantity if and only if his extra valuation is as high as the extra cost.

The consumer's decision is analogous to a firm's output decision (as previewed in the Preliminaries chapter). Whereas the firm chooses how much to produce and sell based on trade-offs between revenue and cost, the consumer chooses how much to demand based on trade-offs between valuation and expenditure. In the smooth case, the marginal condition is that marginal valuation equals the marginal tariff. It is typically the case that marginal valuation is decreasing.

Limitations to nonlinear pricing

Nonlinear pricing incurs two administrative burdens:

1. the firm must inform buyers of its price schedule, which is more complex than a simple flat price;
2. the firm must keep track of how much each customer purchases.

It is easy to keep track of the size of individual transactions (e.g., sell packages of four bars of soap at a lower price than a single bar of soap). However, if the price schedule is to be applied to cumulative transactions then trading cannot be anonymous. For example, a retailer may make trading non-anonymous by requiring shoppers to use a "fidelity" card in order to obtain discounts or special gifts based on the quantity of purchases.

With nonlinear pricing, not everyone pays the same amount per unit of the good. Nonlinear pricing can therefore be impeded by customers' efforts to obtain the lowest per-unit tariff, as follows.

Arbitrage. If the firm offers volume discounts, then some customers may buy large quantities and resell small quantities to other customers. This is a form of arbitrage. The easier arbitrage is, the less steep volume discounts can be. Arbitrage may also be performed by intermediaries. For example, how does a manufacturer of soap sell a customer four bars at a cheaper per-unit price than one bar? It cannot simply put four bars in a box and sell these to the retailer at a lower price per bar than single bars, because the retailer could then stock his shelf with single bars by opening the four-bar boxes. Therefore, one observes four bars glued together so that they cannot be separated without damaging the individual packages.

Masquerading. If the price schedule is such that the average tariff is increasing (volume premia instead of volume discounts), then customers may try to place several small orders under different identities rather than placing all orders under the same identity. Such masquerading is nearly impossible for customers of a utility company but is easy for on-line shoppers. Therefore, whereas a Web merchant could require each customer to create and use an identity in order to receive a volume discount, it could not induce to customer to do so if he paid more than when shopping under multiple identities.

Whenever we study examples of nonlinear pricing, we implicitly assume that arbitrage and masquerading are impossible or at least difficult.

Nonlinear pricing due to a nonlinear cost structure

Nonlinear pricing can be due in part to a nonlinear cost of transacting with each customer. For example, a telephone company has a fixed cost of maintaining a household's phone number and billing the household each month, whether or not the household makes any phone calls. When a customer enters a nightclub, there is a fixed (opportunity) cost of the space the customer takes up, independent of the number of drinks he has. Per-unit delivery costs are usually lower for large deliveries than for small ones. Tax-return preparation involves client-specific learning by doing, so preparing a client's return four years in a row is not four times as expensive as preparing the return once. Such nonlinear cost structures lead to nonlinear pricing even in competitive markets. Our interest throughout this chapter is in the use of nonlinear pricing when transactions have no such nonlinear cost structure.

11.4 An example of nonlinear pricing as screening

Scenario

Suppose that you have two types of customers. Suppose further that each customer is interested in purchasing either 0, 1, or 2 units of the good. Since 0 represents "no trade" and requires no tariff, our problem is to pick the tariff we charge for 1 unit and the tariff we charge for 2 units.

Let the two types of customers be called "high valuation" and "low valuation", or just "high" and "low" for short. Their total valuations for the different quantities are shown in Table 11.1. Note that both the *total* and *marginal* valuations are higher for the high type than for the low type.

Table 11.1

Type of	Valuation	
customer	1 unit	2 units
High	\$45	\$75
Low	\$35	\$50

Assume that you have a constant marginal cost of production that is equal to \$10. Hence, trade is efficient if you sell 2 units to each customer.

We are interested in how your firm's pricing decision depends on the number of high- and low-valuation customers. Let H denote the number of high-valuation customers and L the number of low-valuation customers. For various menus, we write out your profit as a function of H and L. For example, if you have a profit of \$30 per high-valuation customer and \$10 per low-valuation customer, then your total profit is $(H \times \$30) + (L \times \$10)$.

Benchmark: Explicit market segmentation

Suppose first that your company can perfectly price discriminate. Then you sell 2 units to all customers and charge them their respective valuations. That is, you charge the high types \$75 and the low types \$50. Since your cost for 2 units is \$20, your profit is $(H \times \$55) + (L \times \$30)$.

Optimal pricing when each customer buys 2 units

Return to the scenario in which you cannot explicitly segment the market. Suppose you offer only bundles with 2 units. There are two potentially optimal tariffs. Either you charge the high types' valuation of \$75 and sell only to high types (for a profit of $H \times \$55$) or you charge the low types' valuation of \$50 and sell to both types (for a profit of $(H+L) \times \$30$). If L is very low compared to H then your profit is higher by charging the high price (margin matters); if L is very high compared to H then your profit is higher by charging the low price (volume matters).

Optimal pricing when customers buy different amounts

Suppose you offer a price schedule such that the high types buy 2 units and the low types buy 1 unit. The important fact is that *the extra amount the high types are willing to pay for a second unit is more than the extra amount the low types are willing to pay.*

Exercise 11.1. A naïve approach to this pricing problem is to set the tariff for 2 units equal to the high types' valuation (\$75) and to set the tariff of 1 unit equal to the low types' valuation (\$35) in order to extract all the surplus. Explain what would happen if you set prices this way.

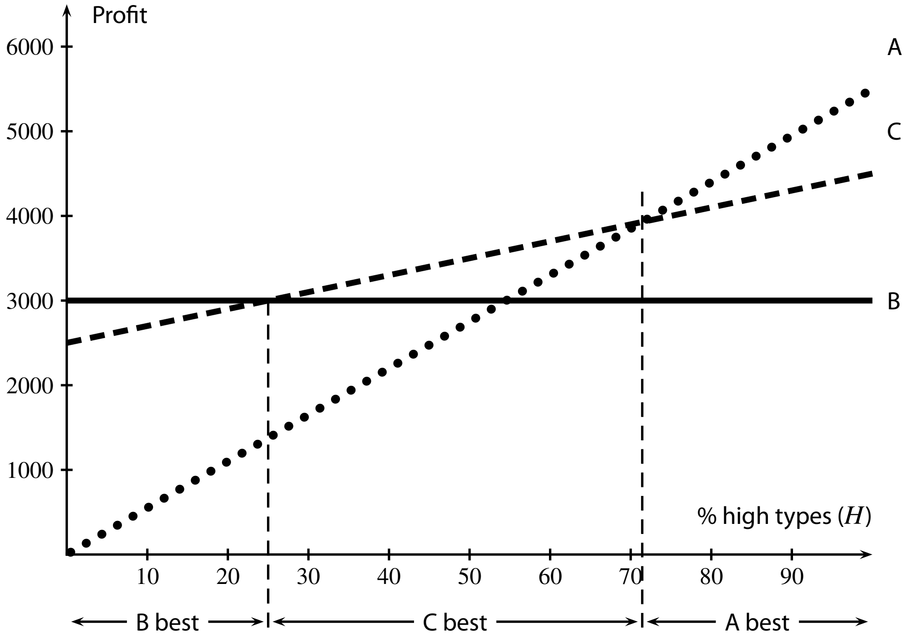
Loosely, your objective is to set the prices "as high as possible", subject to the self-selection and participation constraints. As in Section 10.5, the binding parts of these constraints are as follows.

- *Low types' participation constraint.* You must not set the price of 1 unit so high that the low types prefer not to purchase any units at all.
- *High types' self-selection constraint.* You must not set the price of 2 units so high that the high types prefer to buy just 1 unit.

With these two principles as a guide, we can determine the optimal prices. You should set the price of 1 unit equal to the low types' valuation of \$35. You should set the price of 2 units to the highest amount such that the high types weakly prefer to purchase 2 units; that

Figure 11.2

Option	Profit as a function of H	Shown in graph as	
A	$H \times \$55$	dotted line	• • • • •
B	$100 \times \$30 = \$3000$	solid line	━━━
C	$H \times \$45 + (100 - H) \times \$25$ $= 2500 + 20H$	dashed line	▬ ▬ ▬ ▬



is, for 2 units you charge the one-unit price ($35) plus the high types' marginal valuation for the second unit ($30), which sum to $65.

Your profit is then $45 per high-valuation customer and $25 per low-valuation customer. Your total profit is $(H \times \$45) + (L \times \$25)$.

Comparison of the menus

We have thus narrowed your options down to three potentially good ones:

A. sell 2 units at a price of $75 to high-valuation customers [profit = $H \times \$55$];
B. sell 2 units at a price of $50 to all customers [profit = $(H + L) \times \$30$];
C. sell 2 units to high-valuation customers for $65 and 1 unit to low-valuation customers for $35 [profit = $(H \times \$45) + (L \times \$25)$].

The ranking of these menus depends on how many customers of each type you have. Fix the total number of customers at 100, so that H is the percentage that are high-valuation customers and $L = 100 - H$. Figure 11.2 shows the profit of each menu as a function of H.

We can now see the trade-off between these menus. Menus A and B do not involve true screening or separation of the types, because each customer gets the same deal (in fact, only one deal is offered). The trade-off between these two menus is the one we studied in Chapter 7 for a firm with market power. You can price low in order to generate extra gains from trade, but at the expense of giving the customers more gains from trade. In this case, it is better to price high (menu A) when the high end of the market is large and to price low (menu B) when the low end of the market is large.

Screening (menu C) is an attempt to "have your cake and eat it too". You sell to the high-valuation customers at a high price in order to extract as much surplus from them as possible. However, you reap extra gains from trade with the low-valuation customers by selling them an inefficient amount at a price that does not make it attractive for high-valuation customers. It is optimal when the market has a mix of high and low types.

The optimal screening menu in the above example has a quantity discount. One unit costs \$35; two units cost \$65. However, such screening can involve a quantity premium, depending on the valuations. Exercise 11.2 asks you to recalculate the optima screening menu after an adjustment to the valuations that results in a quantity premium.

Quantity discounts are more common than quantity premia. Quantity premia are difficult to enforce. Consumers can show up under different identities to buy smaller quantities several times in order to obtain a smaller per-unit price.

Exercise 11.2. Find the optimal screening menu for the following valuations.

Table E11.1

Type of	Valuation	
customer	1 unit	2 units
High	\$40	\$80
Low	\$35	\$50

11.5 Two-part tariffs

Overview

The complexity of solving for the exact solution to a screening problem grows quickly as we increase the number of types of customers and the number of potential options to put in the menu. Imagine, if you can, trying to solve a nonlinear pricing problem with 10,000 heterogeneous customers. The complexity of the optimal menu—first in figuring out what it is and then in implementing it—has a cost that does not appear in our model. When this cost is taken into account, it is better to sacrifice the profitability of the menu in return for simplicity. For example, although the millions of airline customers each have slightly different preferences, the airlines offer only a limited number of fare classes and design these classes by grouping customers into broad categories. The principles of screening that we have studied in this chapter still apply, but the firms attempt to find merely a good screening mechanism rather than an "optimal" one.

Firms that engage in nonlinear pricing with a large number of heterogeneous customers often restrict their attention to a simple type of nonlinear price schedule called the *two-part tariff*. This is like a linear price P, but you add on a "fixed cost" T for doing any business at all, and so the tariff for Q units is $T + PQ$. The fixed cost T is called the *entry fee* and P is called the *usage fee*. Here are some examples:

1. local telephone companies charge a connection fee plus a usage fee that depends on the number of phone calls;
2. warehouse clubs charge a membership fee that is independent of quantities purchased;
3. bars have cover charges, which are independent of the amount of alcohol consumed.

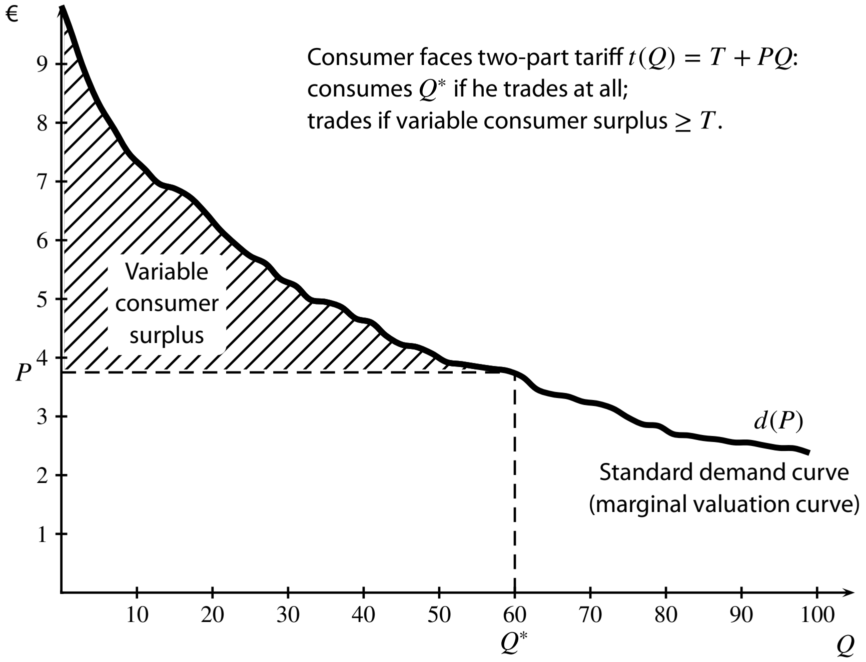
A two-part tariff has a volume discount: the average tariff, $(T/Q) + P$, is decreasing in Q.

Because a two-part tariff $t(Q) = T + PQ$ has only two parameters, T and P, solving for the optimal two-part tariff is just a two-variable optimization problem, which is only slightly more difficult than choosing a single linear price. With data about the composition of your market, you can estimate demand—and hence profit—for each two-part tariff. You then search for the one that maximizes your profit.

Consumer choice under a two-part tariff

Suppose a customer faces a two-part tariff $t(Q) = T + PQ$. The entry fee T is a fixed cost for the customer that affects the customer's decision about whether to trade at all but does not affect his decision about how much to trade. Hence, if the customer's demand curve is $d(P)$ and his valuation curve is $v(Q)$, then he purchases $Q^* = d(P)$ if he decides to trade. His "variable consumer surplus", not taking into account the entry fee, is $v(Q^*) - PQ^*$. He chooses to trade as long as $v(Q^*) - PQ^* \geq T$, that is, as long as his total valuation $v(Q^*)$ is as high as the total cost $T + PQ^*$ of the Q^* units. This is illustrated in Figure 11.3.

Figure 11.3



Using two-part tariffs in perfect price discrimination

In perfect price discrimination, the firm makes a take-it-or-leave-it offer of a single quantity at a single tariff to each customer. You can think of such an offer as an extreme form of nonlinear pricing.

There is a two-part tariff that leads to the same outcome while appearing to give the customer more choice. Since the firm gets all the gains from trade and hence wants the customer to purchase the socially efficient amount of the good, the usage fee *P* should be set to the firm's marginal cost. The customer then purchases up to the point where his marginal valuation equals the firm's marginal cost. Then the firm sets the entry fee in order to extract all the gains from trade, so that the customer is indifferent between entering the market and staying out. Thus, the firm sets T equal to the consumer surplus the customer would receive if he faced the linear price *P*.

That perfect price discrimination can be achieved with a two-part tariff helps us understand both perfect price discrimination and two-part tariffs. However, it is not an explanation of why two-part tariffs are used. First, recall that perfect price discrimination is only a hypothetical benchmark. Second, a two-part tariff is a more complicated scheme than the "single quantity at a single tariff".

Exercise 11.3. Suppose the demand curve in Exercise 7.3 is that of a single customer, and let your marginal cost be 25. Suppose you implement perfect price discrimination with a two-part tariff. What is your usage fee? What is the entry fee?

Two-part tariffs as screening mechanisms

There is a counterexample to nearly any general statement about two-part tariffs when used as screening. However, a few general tendencies can be stated.

The entry fee should usually be strictly positive. Starting from a zero entry fee, an increase in the entry fee only causes those customers who are getting the least surplus to drop out of the market. These customers typically purchase very little and the loss of such sales is negligible. In contrast, the additional surplus extracted from those who remain in the market is significant.

The usage fee is usually strictly higher than marginal cost. Starting with a usage fee equal to marginal cost and with an entry fee set to its optimal level, suppose the usage fee is increased and the entry fee is decreased in order to keep all current customers in the market. The higher usage fee causes a deadweight loss among the current customers, but this effect is initially negligible. The higher usage fee also increases the extraction of surplus from those who have high valuations, and this effect is significant.

11.6 "Degrees" of price discrimination

We say that a firm price discriminates if it does not sell all units of a good at the same per-unit cost. Price discrimination occurs when the firm explicitly segments its market, uses nonlinear pricing, or uses mixed bundling. (Screening via differentiated products is a gray area; each version of the product is sold at only one price, but different versions are sold at different prices even if they cost the same to the firm.)

There is a categorization of price discrimination into *first*, *second*, and *third degrees*. We have so far avoided these terms because they suggest a ranking and a comprehensive categorization that do not exist. Now that we understand price discrimination, we can see how these common terms apply. They are defined according to market segmentation and nonlinear pricing as shown in Table 11.2.

Table 11.2

Market	Pricing	
segmentation	Linear	Nonlinear
Perfect	3rd-degree	1st-degree
Imperfect	3rd-degree	(no name)
None	None	2nd-degree

Degrees of price discrimination

Here is how what we have studied fits in.

- *Perfect price discrimination* is the same as *first-degree price discrimination*.

- *Explicit market segmentation with linear pricing*, studied in Section 9.3, is the same as *third-degree price discrimination*.
- *Nonlinear pricing as screening*, studied in Section 11.4, is the same as *second-degree price discrimination*.
- Bundling and screening via production differentiation do not fall into any of these categories.

Additional exercises

Exercise 11.4. (Nonlinear pricing) You sell a divisible good to two customers, A and B. Their demand curves are

$$\begin{aligned} d_A(P) &= 3\tfrac{3}{5} - \tfrac{2}{5}P\,, \text{ and} \\ d_B(P) &= 12 - 2P\,. \end{aligned}$$

One can show that their total valuation curves are therefore

$$\begin{aligned} v_A(Q) &= 9Q - \tfrac{5}{4}Q^2\,, \text{ and} \\ v_B(Q) &= 6Q - \tfrac{1}{4}Q^2\,, \end{aligned}$$

respectively. You have a constant marginal cost of 4.

You cannot explicitly segment the market, but nonlinear pricing is possible. Find the optimal menu, using the following hint: you can achieve efficient trade *and* extract all the gains from trade. Be sure to check that your menu satisfies all constraints. Could you achieve the same outcome using a two-part tariff (i.e., the same two-part tariff for both customers)?

Chapter 12

Static Games and Nash Equilibrium

12.1 Motives and objectives

The science of business is a social science. It is about people who interact and who make decisions with some information about each other's motives and decisions. The key features of such a social situation are the decision makers who are involved (individuals or organizations), the order in which decisions are made, the actions available to each decision maker each time she makes a decision, and each decision maker's motives or preferences over the outcome. A *game*, as a model in *game theory*, is a careful formal representation of these key features. The game has more structure than exists in the real world, and captures only some of the important features, so that it can then be subjected to careful analysis. However, the basic analysis that goes into the formal model is the same that we use, or should use, when approaching any real-world situation. Such analysis brings strategic reasoning to the forefront.

This is our first chapter on game theory and its applications. This chapter introduces static games, which are games in which each player makes one decision and the decisions are made at the same time. We first focus in the individual decision maker in a game and then look at the mutual interaction between them.

12.2 Players, actions, and timing

In a game, the decision makers are called *players*. In this chapter we consider only games in which each player makes just one decision. This is obviously just a "snapshot" of real-world social situations, which nearly always involve some ongoing interaction, Furthermore, this chapter considers only games in which each player makes her decision without first observing the decision of the other players. Such games are called *static games* or *simultaneous-move games*.

For example, consider the following competition in a market. There currently is one firm in the market, acting as a monopolist; we call this firm the Incumbent. Another firm is contemplating entry; we call this firm the Entrant. The Incumbent's decision is whether to advertise or not. The Entrant's decision is whether to enter the market or not. These decisions are made “simultaneously”. This could happen if each of these decisions takes some time to be executed, so that (a) the Incumbent must develop and pay for an advertising campaign before knowing whether the Entrant will be in the market and (b) the Entrant must invest in a product to get into the market before seeing whether there is an advertising campaign.

In this example, each player has only two possible actions. Any game, like this one, in which each player has only a few (or finitely many) possible actions, is called a *game with discrete actions*. In many applications, however, a player’s action can have many values. Here are some examples:

1. Firms with competing products choose what prices to charge.
2. Fishing vessels operating in the same waters choose how intensively to fish.
3. Listeners of a private but nonprofit (and listener-supported) radio station choose how much to contribute to the station.
4. Firms in an R&D race choose how much to invest in R&D.
5. The members of a team decide how much effort to devote to the team project.
6. Firms competing in the same market choose how much to spend on advertising.

We call these *games with numerical actions*. The basic concepts are the same whether a game has discrete or numerical actions, but some tools differ.

12.3 Payoffs

Entrant–Incumbent game

Let’s suppose that you are the Entrant in the Entrant–Incumbent game. How do you approach the situation and decide what action to take? You have already made the first step: identifying the key players with whom you must interact, you are interacting with, the decisions each of you faces, and the timing of those decisions.

The next step is figuring out your objectives—that is, how you rank the possible outcomes, where an outcome means the action that each player takes.

We will use numerical values to measure your ranking of outcomes. For the Entrant–Incumbent story, it is natural to think of these numbers as profits. But if you had other motives, such as the thrill of entering a new market or compassion for the manager of the other firm (either way, don’t tell your shareholders!), then the numerical values can reflect these things as well. In general, we call these values your *payoffs*.

For a discrete game like this one, we can depict the payoffs in a table or matrix, as shown in Table 12.1.

Table 12.1

		Entrant	
		Enter	Not enter
Incumbent	Advertise	−1 □	0 □
	Not advertise	1 □	0 □

The rows correspond to the actions of one player—in this case, the Incumbent. The columns correspond to the actions of the other player—in this case, the Entrant. Thus, each cell corresponds to a pair of actions, that is, to an outcome of the game. In each cell, we put the two players' payoffs for that outcome. We have positioned the Entrant's payoffs in the upper right corner of each cell, to leave room for the Incumbent's payoffs in the bottom left corner. Because we do not yet make use of the Incumbent's payoffs, they are denoted by □.

The outcome (*Not advertise*, *Enter*) gives you (the Entrant) the highest payoff. However, you cannot just pick your preferred outcome, because you can control only your own action. If you are sure that the Incumbent is going to advertise, then the relevant comparison is that you prefer the outcome (*Advertise*, *Not enter*) over the outcome (*Advertise*, *Enter*), and hence you prefer the action *Enter* for yourself.

Partnership game

For a game with numerical actions, we cannot show payoffs using a matrix. Instead, we write down the payoffs with a function. For example, if there are two players 1 and 2 whose actions are denoted by A_1 and A_2, respectively, then we write the payoff of player 1 as $u_1(A_1, A_2)$. (We also use $\pi_1(A_1, A_2)$ when it is natural to think of the payoffs as profits.)

Our example is of a "partnership game", which refers to any kind of team production in which players devote effort or resources to a common project whose fruits are then shared. Examples include (a) partners who own a small company, (b) two firms that engage in a joint venture, (c) students working together on a group project, and (d) co-workers on a project team.

Consider a partnership game with two players, 1 and 2. Let's view them as partners of a small firm and let A_i be the effort of partner i. Suppose that the firm's accounting profit depends on the level of effort that each of the partners puts in, such that $\pi(A_1, A_2) = 16A_1 + 16A_2 + 2A_1A_2$. The partners split the profit evenly and so each partner gets $\frac{1}{2}\pi(A_1, A_2)$. Each partner's surplus is the difference between her share of the profit and a disutility or cost of effort, which is $c_i(A_i) = A_i^2$ for partner i.

As we work through this example, you will assume the role of player 1. Your payoff function is

$$u_1(A_1,A_2) = \tfrac{1}{2}\,\pi(A_1,A_2) - c_1(A_1) = 8A_1 + 8A_2 + A_1A_2 - A_1^2\,.$$

12.4 Dominant strategies

Your next step is to forecast the action of the other player, right? Not so fast. This is not necessary if you have an action that is best no matter what the other player does. That action is then called a *dominant strategy*. For example, if you are firm B in the game shown in Table 12.2, then *Rebate* is a dominant strategy for you.

Table 12.2

		Firm B	
		Rebate	No rebate
Firm A	Rebate	60 □	10 □
	No rebate	70 □	0 □

The term “dominant strategy” applies even if the strategy sometimes “ties” with one of your other strategies. In none of the other games that we have considered do you have a dominant strategy, even in this weaker sense. However, here is a famous game in which you do have a dominant strategy.

A sealed-bid second-price auction is an auction with the following rules. Each bidder submits a bid “in a sealed envelope” without being able to observe the bids of the other players (hence, this is a static game). Then the envelopes are opened and the bidder with the highest bid gets the object that is being auctioned off. Here is why it is called a “second-price” auction: the winning bidder pays the value of the second-highest bid rather than the value of his own bid.

Exercise 12.1. What is your dominant strategy in a sealed-bid second-price auction in which you know your valuation of the object? Explain.

12.5 Best responses

If you do not have a dominant strategy, then your best response depends on what you think the other player will do. Our next step is to identify what you should do if you know the other player’s action (for each possible action of the other player).

Discrete games

For example, consider the Entrant–Incumbent game in which your payoffs (as the Entrant) are as shown in Table 12.3.

Table 12.3

		Entrant	
		Enter	Not enter
Incumbent	Advertise	−1 □	0 □
	Not advertise	1 □	0 □

You can see that your best response to *Advertise* is *Not Enter*: you are comparing your payoffs for your two actions that are in the row for the action *Advertise*. Similarly, your best response to *Not Advertise* is *Enter*. This is illustrated in Table 12.4 by circling the highest payoff in each row, corresponding to the action by the Entrant that is the best response to the action of the Incumbent.

Table 12.4

		Entrant	
		Enter	Not enter
Incumbent	Advertise	−1 □	[✓ 0] □
	Not advertise	[✓ 1] □	0 □

Exercise 12.2. Consider a game between players A and B; you are player B. The actions and your payoffs are shown in Table E12.1.

Table E12.1

		Player B			
		I	II	III	IV
Player A	W	0 □	0 □	2 □	3 □
	X	1 □	3 □	2 □	1 □
	Y	3 □	2 □	5 □	1 □
	Z	2 □	1 □	3 □	2 □

The purpose of using this large game with no story is to drive home how simple and mechanical an exercise it is to find best responses in a discrete game. For example, to find your best response to Y, you just compare the four numbers in the row for action Y. For each action by player A, circle the payoff for your action that is the best response.

The partnership game

In a game with numerical actions, we cannot simply put circles on a table to show best responses. Instead, we summarize the best response of player 1 (for example) by using a function $A_1 = b_1(A_2)$. This is called player 1's *best-response curve* or *reaction curve* (even though player 1 is anticipating rather than reacting to player 1's action).

Most examples of games with numerical actions use calculus to find the best responses. This is true of our partnership game. However, we leave the calculus in footnotes because it is not the point of this example.

Suppose that you are player 1 in the partnership game. Recall that your payoff function is

$$u_1(A_1,A_2) = 8A_1 + 8A_2 + A_1A_2 - A_1^2 .$$

Suppose you expected that player 2 would choose $A_2 = 10$ and you wanted to calculate your best response. Replacing A_2 by 10 in your payoff function shows you how your payoff depends on your own action A_1:

$$\begin{aligned} u_1(A_1, 10) &= 8A_1 + (8 \times 10) + (A_1 \times 10) - A_1^2 \\ &= 80 + 18A_1 - A_1^2 . \end{aligned}$$

You want to find the value of A_1 that gives you the highest payoff.

One way to solve this problem is to solve the marginal condition that your marginal payoff (the first derivative of your payoff function with respect to your own action) be equal to zero.[1] The solution is $A_1 = 9$.

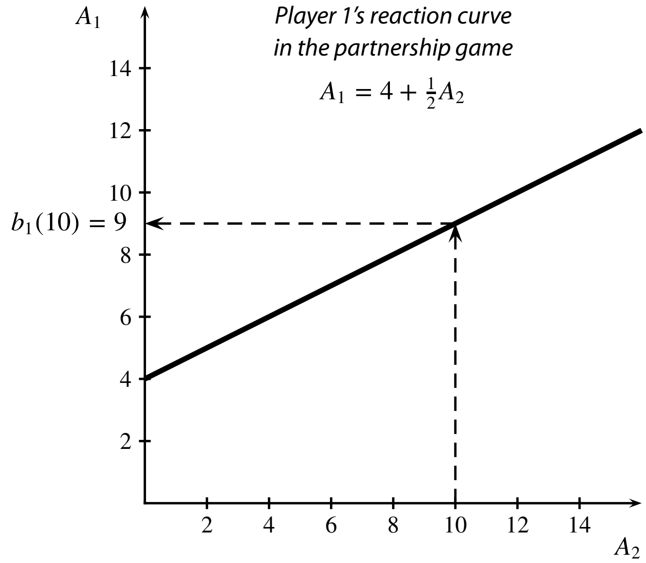
In order to find the general functional form for your best-response curve, we can perform the same exercise while leaving A_2 unspecified.[2] The solution is $A_1 = 4 + \frac{1}{2}A_2$. We can also write that the best-response curve is $b_1(A_2) = 4 + \frac{1}{2}A_2$. Note that $b_1(10) = 9$, which means that your best response to $A_2 = 10$ is $A_1 = 9$. This is what we calculated previously.

We can graph the reaction curve to illustrate the best responses. We put the independent variable, A_2, on the horizontal axis and the dependent variable, A_1, on the vertical axis; see Figure 12.1.

[1] This first derivative is $\partial u_1(A_1, 10)/\partial A_1 = 18-2A_1$. Setting it equal to zero and solving for A_1 yields $18-2A_1 = 0$ or $A_1 = 9$.

[2] This first derivative is $\partial u_1(A_1,A_2)/\partial A_1 = 8 + A_2 - 2A_1$. Setting it equal to zero and solving for A_1 yields $8 + A_2 - 2A_1 = 0$ or $A_1 = 4 + \frac{1}{2}A_2$.

Figure 12.1



Strategic complements and substitutes

The word “complements” captures the idea that two variables reinforce each other; the word “substitutes” captures the idea that two variables replace each other. For a consumer, two goods are complements if consuming more of one good makes him want to consume more of the other; if the opposite happens, the goods are substitutes. In a production process, inputs are complements if an increase in one input raises the marginal product of the other good and hence leads the firm to use more of the other input; if the opposite occurs, the inputs are substitutes.

We can also use these terms to characterize, for a game with numerical actions, the interaction between the action of one player and the best response of the other player. Suppose the players are 1 and 2. If an increase in player 2’s action causes player 1’s best response to rise, then the actions are *strategic complements* for player 1; if the opposite occurs, the actions are *strategic substitutes*. In other words, the actions are strategic complements for player 1 if her best-response curve is increasing; they are strategic substitutes if it is decreasing.

In the partnership game, player 1’s best response curve is increasing. This can be seen either by inspecting the curve’s formula $A_1 = 4 + \frac{1}{2}A_2$ or by inspecting its graph in Figure 12.1.

Strategic complements and substitutes are qualitative concepts, which means that they characterize whether best responses go up or down but not by how much. They play an important role in Chapter 15. As usual, the value of such qualitative concepts is that we can often apply them even when we do not have the data necessary for the kind of numerical calculations in our examples. Let’s revisit the partnership game to understand qualitatively why the actions are strategic complements.

In that game, the effort levels are complementary inputs in the production process. Thus, an increase in effort by player 2 raises the marginal return to effort by player 1, causing player 1 to increase her own effort. This occurs here because of the term A_1A_2 in the team production function. When does it happen in real life?

Consider two partners who own a firm, where one is a production type and the other is a marketing type. These two skill sets complement each other. If the production manager works hard and thereby produces a better product at a lower price, then there is a greater return to the marketing manager's effort and so the marketing manager is motivated to work hard as well.

Contrast this with a partnership setting that is similar but in which the partners are members of a study group who must hand in a group project for the class. Suppose that the group members bring the same skills to the group (their efforts are substitute inputs) and that each group member's concern is mainly to get an above-average grade. The more one member of the group works, the less the other members feel compelled to work in order to avoid a bad grade. Then the effort levels are strategic substitutes.

Here is how we might classify two other examples.

1. In the game between housemates who control their own stereos, the volume levels are probably strategic complements for the most part. If one housemate turns up his stereo, the other does the same in order to hear his own music. (However, unless they really hate each other's music, if one turns his stereo up loud enough then the other may just give up and turn his off rather than trying to drown out his housemate.)
2. In the game between competing fishing vessels, the fishing intensities are probably strategic substitutes. The more one boat fishes and hence depletes the stock of fish, the lower is the return to fishing for the other boat.

12.6 Equilibrium

By "equilibrium" we mean a possible outcome of the game, taking into account both players' decisions. We thus shift away from the perspective of a single player's decision problem to take more of a bird's-eye view. Yet whatever outcomes we predict can also serve to guide the individual players.

Equilibrium in dominant strategies

The simplest and most convincing prediction is made when both players have a dominant strategy. Then we can safely predict that both players will choose their dominant strategies. This is called a dominant-strategy equilibrium.

Consider the following famous game, called the "Prisoners' Dilemma". Two suspects (Akbar and Jeff) to a crime in Texas are interrogated in separate rooms. If both suspects resist confessing, the evidence is enough to convict them of second-degree murder. If a single prisoner confesses and testifies against the other one, the charge against the confessor is reduced to manslaughter whereas the other prisoner will be convicted of first-degree murder and then executed. If both prisoners confess, they are both convicted of first-degree murder but their sentences are reduced to life in prison in return for their guilty pleas.

The payoffs (representing merely preferences over the outcomes) are written in Table 12.5 for both players. We put the payoff of Jeff (the "row" player) in the bottom left of each cell; we put the payoff of the Akbar (the "column" player) in the top right of each cell.

Table 12.5

		Akbar		
		Resist	Confess	
Jeff	Resist	2 2	3 0	Akbar's payoff Jeff's payoff
	Confess	0 3	1 1	

For each player, *Confess* is a dominant strategy. (Verify this by examining Table 12.5 carefully. Now that we have both players' payoffs in the matrix, you need to look at the correct payoffs when examining the behavior of one of the players.)

When a game has a dominant-strategy equilibrium, there is no strategic interaction. Neither player needs to think about the other player's thinking and actions in order to decide what to do. It sounds pretty boring as a subject of long discussion. What makes the Prisoner's' Dilemma such a famous and oft-discussed game is that it illustrates, quite starkly, that the equilibrium play of a game need not lead to a collectively desirable outcome. Each player acts out of self-interest and without considering the impact of his or her actions on the other player. *In the Prisoners' Dilemma, it is in the individual interest of each player to confess, yet they would both be better off if neither confessed.* This tension, and ways to resolve it, are the subject of Chapter 14.

(The term "Prisoners' Dilemma" is used more broadly for the entire class of 2 × 2 games—with two players, each of whom has two actions—that have the two properties just highlighted: (a) each player has a dominant strategy; and (b) the outcome when each chooses his dominant strategy is Pareto inferior to the outcome when each chooses his other strategy.)

Exercise 12.3. Cigarette manufacturers often claim that the purpose of their advertising is to attract smokers from other brands, not to attract new smokers or to increase the amount that each consumer smokes. Let's suppose this claim is correct and then construct a model of the advertising decisions of the firms. To simplify the situation, suppose that: (a) there are only two firms, RJR and Philip Morris; and (b) each firm has only two options with regard t advertising, *Advertise* and *Not advertise*.

Define "gross profit" to mean profit that does not taking into account the cost of advertising and define “net profit” to be the gross profit minus the advertising expenditure. Let the cost of advertising be 30. We take the statement that advertising does not encourage smoking to mean that the total gross profit of the industry is the same regardless of the amount of advertising; advertising affects only the distribution of the industry gross profit among the firms. Assume that when a firm chooses to advertise, its gross profit increases by 40 and the gross profit of its competitor falls by 40 (keeping fixed the advertising decision of the competitor). Let the gross profit for each firm when neither firm advertises be 160.

a. Construct a 2 × 2 matrix showing the net profit (payoff) of each firm for each combination of advertising decisions.

b. What outcome do you predict for this game?

c. Assuming that the claim of the cigarette manufacturers (stated in the first sentence of this problem) is true, should the two companies lobby for or against a law that prohibits all cigarette advertising?

Recall the sealed-bid, second-price auction from Section 12.4. Each player has dominant strategy. Therefore, such play is a dominant-strategy equilibrium.

Here is another example. In a takeover bid, a two-tier tender offer is one in which the suitor offers to buy a certain percentage of the shares (usually a majority) at one price and then the remaining shares at a lower price. In its most coercive form, the remaining shares are forcibly “bought out” at below-market price. This may occur by merging the takeover target with the suitor firm on unfavorable terms (after obtaining majority control of the target).

Here is how this might work. Suppose the per-share market value of a firm is €100. The suitor offers to buy 51% of the shares at a price of €105 and then, through some means, can acquire the remaining share for a value of €90 each. Shareholders have some deadline for tendering their shares. If more than 51% of the shares are tendered, then only a fraction of each shareholder’s tendered shares are purchased at the price of €105, so that the total amount purchased at this price is still 51%. The remainder is purchased at the unfavorable value of €90. (This is called “blending”.)

Clearly, this takeover is not in the collective interest of the shareholders. If successful, the average value of their shares is $(105 + 90)/2 = 97.50$, which is below the market value. However, for each individual shareholder, it is a dominant strategy to tender all shares. For example, suppose no else else tenders their shares, so that the takeover is not successful. If you tender your shares, you get €105 per share rather than the market price of €100. Suppose instead enough shareholders tender their shares for the takeover to be is successful. If you tender your shares then you get some blend of €105 and €90; otherwise you get only €90 per share.

You can see why many countries have regulations that restrict such coercive two-tier offers.

Nash equilibrium

When a game does not have a dominant-strategy equilibrium, each player's action depends on what she expects the other player to do. We thus turn to the notion of *Nash equilibrium*, which requires that players' expectations be consistent with each other but does not explain how this consistency is achieved. For example, a profile of actions (A_1, A_2) in a two-player game is a Nash equilibrium if:

1. player 1, expecting 2 to play A_2, finds that her best response is A_1; and
2. player 2, expecting 1 to play A_1, finds that his best response is A_2.

A game can have one or more Nash equilibria or it can have no Nash equilibrium at all.

For example, consider two firms that develop and sell highly complementary products. For a concrete story, suppose there is a single manufacturer of televisions, called Thompson, and a single broadcaster, called TF1. These firms are each thinking of investing in HDTV, and each firm's available actions are *Invest* and *Not invest*. Of course, no consumers will watch HDTV broadcasts if they do not have HDTV sets, and without HDTV broadcasts, no consumers will buy HDTV sets. Though it may be profitable if both firms develop the new technology, TF1 would simply lose its investment if it developed HDTV broadcasts and Thompson did not develop HDTV sets. The payoff matrix could be as shown in Table 12.6.

Table 12.6

		Thompson	
		Invest	Not invest
TF1	Invest	100 100	20 −50
	Not invest	−50 20	20 20

Let's now check each pair of actions (listing TF1's action first) to see which are Nash equilibria

- *Invest, Invest.* This is a Nash equilibrium. For each firm, it is best to invest if the other firm invests.
- *Not invest, Not invest.* This is also a Nash equilibrium. For each firm, it is best not to invest if the other firm does not invest.
- *Not invest, Invest.* This is not a Nash equilibrium. For example, TF1 prefers to invest if it thinks that Thompson will invest.
- *Invest, Not invest.* This is not a Nash equilibrium either.

This game, with two Nash equilibria, illustrates the possibility of coordination failure. Both players are better off if they invest, but it is also a Nash equilibrium that neither player invests: neither expects the other to invest, therefore neither wants to invest, therefore neither does invest, therefore the players are correct in their expectations.

Let's use our example from Exercise 12.2 to focus on the mechanics of Nash equilibrium. In that exercise, you found the best responses for Player B. Table 12.7 shows the payoffs of both players and the best responses for player A.

Table 12.7

		Player B			
		I	II	III	IV
Player A	W	0 1	0 0	2 3	3 2
	X	1 4	3 2	2 3	1 0
	Y	3 3	2 2	5 4	1 1
	Z	2 0	1 3	3 1	2 1

We now show the best responses for both players in the same matrix in Table 12.8. Any cell that corresponds to a best response for *both* players is a Nash equilibrium. There are two: (Y, III) and (X, IV).

Table 12.8

		Player B			
		I	II	III	IV
Player A	W	0 1	0 0	2 3	3 2
	X	1 4	3 2	2 3	1 0
	Y	3 3	2 2	5 4	1 1
	Z	2 0	1 3	3 1	2 1

Exercise 12.4. Table E12.2 shows another game without telling you the story that lies behind it.

Table E12.2

		Player B	
		Y	Z
Player A	W	−20 −20	0 100
	X	100 0	0 0

a. Find the Nash equilibrium (equilibria) of the game.

b. Tell two stories for which this game could plausibly be a formal model. One story should be of two firms; the other story should be from everyday life and involve two individuals.

Exercise 12.5. Table E12.3 shows the profits for a pair of duopolists who can choose among three output strategies: *Zero*, *Small*, or *Large*.

Table E12.3

			Firm B	
		Zero	Small	Large
	Zero	0 0	1500 0	2000 0
Firm A	Small	0 1500	1300 1300	1400 800
	Large	0 2000	800 1400	500 500

Mark the best responses for both players on this matrix. What are the Nash equilibria?

Nash equilibrium with numerical actions

For a game with numerical actions, we find a Nash equilibrium as follows.

Consider a typical game with two players, 1 and 2, and with numerical actions, denoted A_1 and A_2, respectively. Let $A_1 = b_1(A_2)$ be player 1's reaction curve and let $A_2 = b_2(A_1)$ be player 2's reaction curve. The Nash equilibrium conditions can be translated into a system of equations involving the reaction curves, as laid out in Table 12.9.

Table 12.9

Conditions for (A_1^*, A_2^*) to be a Nash equilibrium

In words …	As an equation …
If player 1 expects 2 to choose A_2^*, then 1 chooses A_1^*	$A_1^* = b_1(A_2^*)$
If player 2 expects 1 to choose A_1^*, then 2 chooses A_2^*	$A_2^* = b_2(A_1^*)$

We can find the equilibrium by solving these two equations. We can illustrate the equilibrium by drawing the two reaction curves on the same axes: the Nash equilibrium is at their intersection.[3]

Consider the partnership game. We already derived player 1's reaction curve; now we need to find player 2's reaction curve. There is a shortcut for this game because it is *symmetric*, meaning loosely that the game looks the same from either player's perspective. More specifically, a game is symmetric if the players have the same available actions and the same payoffs, except that the roles of the two actions are reversed. For example, the payoffs of players 1 and 2 in the partnership game can be written

$$u_1(A_1,A_2) = 8A_1 + 8A_2 + A_1A_2 - A_1^2\,,$$

$$u_2(A_2,A_1) = 8A_2 + 8A_1 + A_2A_1 - A_2^2\,.$$

This game would not be symmetric if (a) the players received different shares of the profit; (b) they had different disutilities of effort; or (c) their efforts entered asymmetrically into the production function.

The world is not symmetric. The point of studying and recognizing symmetric games is that they are simpler. In particular, the two players' reaction curves will have the same form. Since player 1's reaction curve is $b_1(A_2) = 4 + \frac{1}{2}A_2$, it follows (in a symmetric game) that player 2's reaction curve must be $b_2(A_1) = 4 + \frac{1}{2}A_1$. We thus save ourselves the trouble of deriving it from scratch.

A Nash equilibrium is thus a solution to the equations

$$A_1 = 4 + \tfrac{1}{2}A_2\,, \tag{12.1}$$

$$A_2 = 4 + \tfrac{1}{2}A_1\,. \tag{12.2}$$

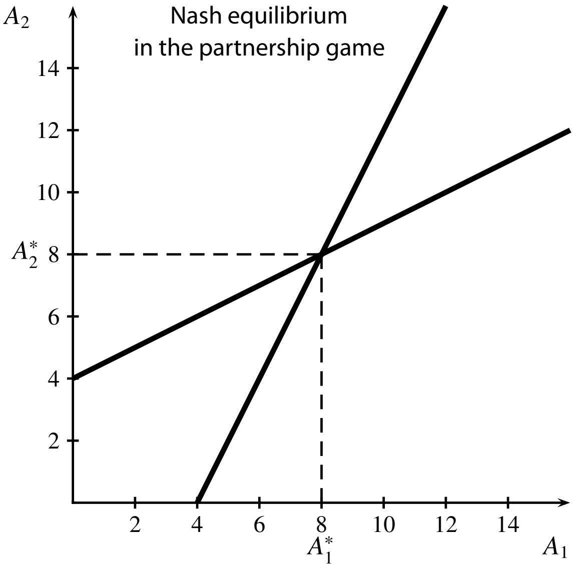
[3] If the point (A_1,A_2) is at an intersection of the reaction curves, then: (a) A_1 is the best response by player 1 to A_2, because the point is on 1's reaction curve; and (b) A_2 is a best response by player 2 to A_1, because the point is on 2's reaction curve. Hence, the point is a Nash equilibrium.

This is a pretty simple system of linear equations, but there is also a shortcut for finding an equilibrium. A symmetric game typically has a symmetric Nash equilibrium, in which each player chooses the same action. If A is the common action in a symmetric equilibrium, then equations (12.1) and (12.2) become

$$A = 4 + \tfrac{1}{2}A\,,$$
$$A = 4 + \tfrac{1}{2}A\,.$$

Observe that these equations are identical. Thus, rather than having two equations and two unknowns, we have only one equation and one unknown.[4] The solution to this equation is $A = 8$. Therefore, in the symmetric Nash equilibrium, each player's effort level is 8.

Figure 12.2



The two reaction curves are drawn on the same axes in Figure 12.2. Player 2's reaction curve is drawn with the usual orientation—independent variable A_1 on the horizontal axis. Player 1's reaction curve is drawn with the reverse orientation—independent variable A_2 on the vertical axis—and so it should be read as mapping points on the vertical axis to best responses on the horizontal axis. We can see in Figure 12.2 that the Nash equilibrium, where the two reaction curves intersect, is $A_1^* = A_2^* = 8$, as we calculated previously.

Interpretation of a Nash equilibrium

In general, if there is some means by which players' expectations about each other's behavior are coordinated, then such coordination can only settle on a Nash equilibrium. How such coordination actually occurs is not spelled out by our model. Here are three means by

[4] This trick can work for games with many players. Suppose the game has 20 players. Finding a Nash equilibrium would normally mean solving a system of 20 equations (one for each player's best-response condition) and 20 unknowns. However, if the game is symmetric, finding a symmetric equilibrium still involves just one equation and one unknown.

which this may occur, that is, three interpretations of a Nash equilibrium.

1. *Self-enforcing agreement.* By some unmodeled pre-play communication, an agreement is reached. An agreement is self-enforcing—meaning that each player will voluntarily follow through with her agreed-upon action if she believes the other player will also do so—if and only if it is a Nash equilibrium.

TF1 and Thompson, involved in the game in Table 12.6, get together to agree on a strategy for developing HDTV. They know that it will be hard to monitor and enforce an agreement in court, so they do not even bother writing a contract. They both agree to make the investment. When the representatives return to their companies, each company finds it in its interest to follow through with the agreement, because it expects the other company to do the same.

2. *Social norm or convention.* In this interpretation, a similar game is played frequently in a society but the players involved change each time. The Nash equilibrium is a social convention or norm that has evolved historically by some unmodeled process. The social norm is stable because players have no incentive to deviate from it.

In a corporation, a social norm could develop that meetings always start five minutes late. Each person has no incentive to show up on time for a meeting because he (correctly) expects that no one else will show up on time.

3. *Steady state of repeated interaction.* In this interpretation, the same players play the same game repeatedly. They do not use sophisticated reasoning to predict how their actions today will influence the actions of others in the future; instead, they form beliefs about what the other players will do today based on their play in the recent past. The players may initially try different out-of-equilibrium strategies, but they stop revising their beliefs and changing their actions once an equilibrium is reached. A Nash equilibrium thus represents a steady state of a dynamic process.

TF1 and Thompson are not the only firms involved in HDTV monitors and broadcasting; we made it a two-firm game merely to have a simple example. The many firms actually involved can find themselves in a classic chicken-and-egg situation. No one has an incentive to invest because no one else is investing. This is a stable situation—a Nash equilibrium—that must be overcome every time a costly new technology, with complementary components produced by different firms, hits the market.

12.7 Wrap-up

We have studied static games, in which players choose their actions at the same time. Players cannot base their decisions on observed actions but only on expectations about what the other players will do.

We focus first on the decision problem of a particular player. The player has to identify the elements of the game and determine the payoffs. She should look for a dominant strategy. She should consider how she would act if she knew the actions of the author players.

Then we looked at the interrelationship between the players' actions. We want to identify outcomes where all players are choosing in their own interest and are conscious of the decisions that other players are making. We described two kinds of equilibria. An equilibrium in dominant strategies is robust since each player has a best action without regard to what the other players are doing. We then considered Nash equilibria, in which players at least have consistent expectations about what each other is doing. Specifically, a Nash equilibrium is an action profile in which each player chooses his equilibrium action if he expects the other players to do so as well.

Chapter 13

Imperfect Competition

13.1 Motivation and objectives

In Chapters 7–10, we considered the pricing decisions of a single firm in isolation, treating its demand curve as fixed. However, a firm's demand curve depends on prices set by other firms. This leads to interesting interaction.

We studied an extreme form of such interaction in the model of perfect competition (Chapters 5 and 6). In perfect competition, the firms produce identical products and there are enough firms so that no single firm influences the market price.

In this chapter (and continuing in the next two chapters), we use the tools of game theory to study an intermediate form of strategic interaction in which only a few firms have a significant effect on each other's demand. Such models are called *oligopoly* or *imperfect competition*. The firms may produce perfect substitutes or they may produce differentiated products. In either case, the price-taking assumption of perfect competition is no longer reasonable. Instead, each firm has residual market power.

We introduce two models of strategic competition: (a) price competition between firms producing substitute goods; and (b) competition in quantities (or investments in capacity) between firms producing perfect substitutes. We study the Nash equilibria of these games.

13.2 Price competition

Scenario

We consider price competition between a small number of firms that produce imperfect substitutes, such as the manufacturers of mobile phone handsets, airlines competing on the same route, automobile manufacturers, car rental companies in the same local market, conference hotels in the same city, and brand-name jean designers. We restrict attention to just two firms, although the insights from this model are valid when there are more than two. We could think of competition between Avianca and American Airlines, which are the only two carriers for the route from Miami to Bogotá. Their products (tickets on this route) are substitutes, but they are not perfect substitutes because they have different schedules, they provide slightly different service, and customers have different loyalties. (This example has additional relevant details, such as price regulation on international routes and multiple fare classes, that we will not consider.) Another example is competition between Coca-Cola and Pepsi in the cola market.

The big picture, before the details

Numerical examples of price competition games necessarily get heavy. Before we see too many symbols and parentheses, let's outline the big picture, including (in italics) what it takes to fully specify such a game and solve for the Nash equilibrium.

1. Each firm has a demand function (Chapter 3) that shows how demand for its product depends not only on the firm's own price but also on the prices set by other firms.

We must first identify the firms whose interaction we care to model. Then we need to know their demand functions—taking into account the dependencies on each other's price—and also the individual firms' cost curves.

2. Each firm's individual problem is the one studied in Chapters 7 and 8. Given prices charged by other firms (either it anticipates these prices or just sees them being charged), it collapses its demand function to a demand curve and then solves the pricing problem for this demand curve.

When we do this step, we should treat the other firms' prices as variables rather than specific numbers. Then we end up deriving the firm's reaction curve: its optimal price as a function of the other firms' prices.

3. However, the firms' decisions are interrelated, something we did not take into account in Chapters 7 and 8. A Nash equilibrium consists of prices charged by all the firms such that each firm's price maximizes its profit given the prices of the other firms. That is, there is consistency between the prices that firms use to derive their demand curves and the prices they set from such calculations.

We solve for the Nash equilibrium, either graphically or numerically, as the solution to the system of equations given by firms' reaction curves.

The details

Suppose there are two firms, 1 and 2. Each must set a price, understanding that demand for its product depends also on the price charged by the other firm. We define also the notation listed in Table 13.1.

Table 13.1

	Price	Sales	Cost curve	Demand function
Firm 1	P_1	Q_1	$c_1(Q_1)$	$d_1(P_1, P_2)$
Firm 2	P_2	Q_2	$c_2(Q_2)$	$d_2(P_2, P_1)$

When the firms choose the prices P_1 and P_2, firm 1 sells $d_1(P_1, P_2)$; hence, its revenue is $P_1 \times d_1(P_1, P_2)$ and its cost is $c_1(\,d_1(P_1, P_2)\,)$. Firm 2's revenue and cost are analogous. Thus, the firms' profits, as functions of the prices the two firms charge, are

$$\begin{aligned} \pi_1(P_1, P_2) &= P_1 \times d_1(P_1, P_2) - c_1(\,d_1(P_1, P_2)\,)\,, \\ \pi_2(P_2, P_1) &= P_2 \times d_2(P_2, P_1) - c_2(\,d_2(P_2, P_1)\,)\,. \end{aligned}$$

Example 13.1 (Part 1) Suppose that the two firm's demand functions are

$$\begin{aligned} Q_1 &= 48 - 3P_1 + 2P_2\,, \\ Q_2 &= 80 - 4P_2 + 3P_1\,. \end{aligned}$$

Suppose also that each firm has constant marginal cost, with $MC_1 = 8$ and $MC_2 = 13$.

Given constant marginal cost, a firm's profit is equal to its per-unit markup multiplied by sales: $(P - MC) \times Q$. Hence, the firms' profit functions are

$$\pi_1(P_1, P_2) = (P_1 - 8)\,(48 - 3P_1 + 2P_2)\,, \qquad (13.1)$$

$$\pi_2(P_2, P_1) = (P_2 - 13)\,(80 - 4P_2 + 3P_1)\,. \qquad (13.2)$$

Residual demand, best responses, and reaction curves

The strategic interaction between such firms is typically ongoing. We consider Nash equilibrium as a possible steady state of this interaction. Each firm chooses its price to maximize its profit given the price that the other firm charges.

The first step in the analysis is to determine each firm's reaction curve. A direct approach is to solve the marginal condition that marginal profit be zero. However, we work in terms of demand curves (as outlined previously) because doing so is more insightful.

Firm 2's price affects firm 1 by shifting firm 1's demand curve. Once firm 1 determines the resulting demand curve, it should apply the uniform pricing methods we studied in Chapters 7 and 8, equating marginal revenue and marginal cost. This approach emphasizes that the interaction is through the demand side and it relates the strategic pricing model to the market power model. In fact, this analysis should look very familiar—we did it already in Section 8.7.

Example 13.1 (Part 2) Suppose that, in Example 13.1, $P_2 = 21$. Then firm 1's demand curve—which we call its *residual demand curve* to emphasize its dependence on firm 2's price—is

$$d_1^r(P_1) = 48 - 3P_1 + (2 \times 21) = 90 - 3P_1\,.$$

The choke price for this demand curve is $90/3 = 30$. Firm 1 should charge the midpoint between its marginal cost and its choke price, or $P_1 = (8 + 30)/2 = 19$.

If instead firm 2 charges $P_2 = 30$, then firm 1's residual demand curve is

$$d_1^r(P_1) = 48 - 3P_1 + (2 \times 30) = 108 - 3P_1\,.$$

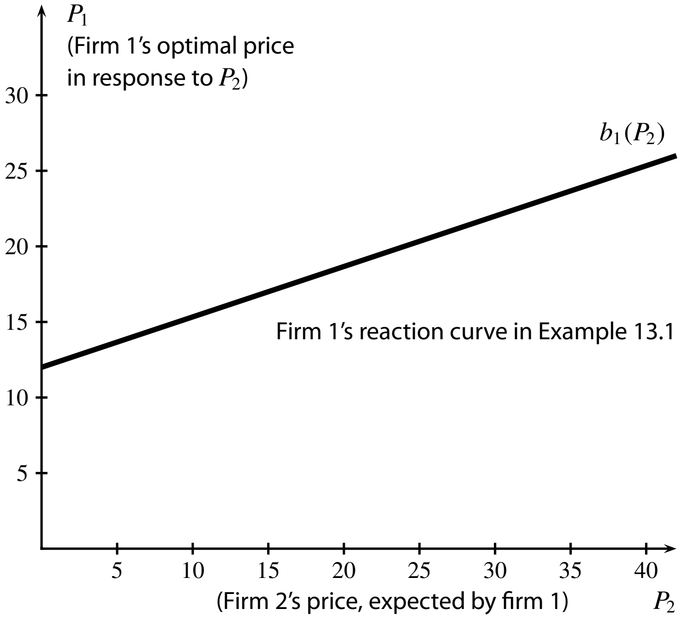
Firm 1's choke price is 36, so it should charge $P_1 = (8 + 36)/2 = 22$.

We can easily derive the general functional relationship between firm 2's price and the price that firm 1 should charge. Given any P_2, firm 1's residual demand is $d_1^r(P_1) = (48+2P_2)-3P_1$. (Compared to the demand function, this residual demand curve emphasizes that the price P_2 is merely a parameter that determines firm 1's demand curve.) The choke price is $(48+2P_2)/3 = 16 + \frac{2}{3}P_2$, so firm 1 should charge

$$\begin{aligned} P_1 &= \frac{8 + \left(16 + \frac{2}{3}P_2\right)}{2} \\ &= 12 + \tfrac{1}{3}P_2 \,. \end{aligned}$$

The function $P_1 = 12 + \frac{1}{3}P_2$ is firm 1's reaction curve and is drawn in Figure 13.1.

Figure 13.1



We calculate firm 2's reaction curve the same way. When firm 1 charges P_1, firm 2 makes the following calculations.

$$\begin{aligned} \text{Residual demand curve:} \quad d_2^r(P_1) &= (80 + 3P_1) - 4P_2 \,; \\ \text{Choke price:} \quad \bar{P}_2^r &= 20 + \tfrac{3}{4}P_1 \,; \\ \text{Optimal price:} \quad P_2 &= \frac{13 + (20 + \frac{3}{4}P_1)}{2} = \frac{33}{2} + \frac{3}{8}P_1 \,; \\ \text{Reaction curve:} \quad b_2(P_1) &= \tfrac{33}{2} + \tfrac{3}{8}P_1 \,. \end{aligned}$$

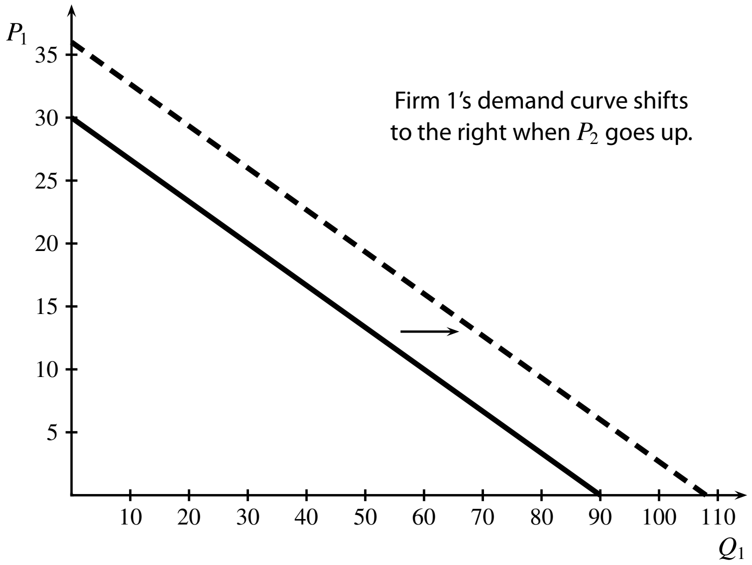
Strategic complements

Warning: We are about to use the English words "complements" and "substitutes" in two different, unrelated ways in the same passage. We will talk about goods being complements or substitutes, which is a property of consumer demand. We will also discuss whether firms' prices, in a price competition game, are strategic complements or substitutes; this is a property of the firms' reaction curves. One should not expect, a priori, any particular link between these two concepts merely because they make use of the same English words.

In Example 13.1 (price competition with substitute goods), firm 1's reaction curve is $P_1 = 12 + \frac{1}{3}P_2$. This curve is increasing, as seen by inspecting the function and as shown in Figure 13.1. Firm 2's reaction curve is similar; the actions are strategic complements.

Let's try to understand why the reaction curve in this example has strategic complements in order to see whether this is a general property of such price competition games. Recall that firm 1's demand function is $48 - 3P_1 + 2P_2$. When firm 2 chooses $P_2 = 21$, then firm 1's demand curve is $Q_1 = 90 - 3P_1$. If instead firm 2 chooses $P_2 = 30$, firm 1's demand curve shifts to $Q_1 = 108 - 3P_1$.

Figure 13.2



As discussed in Chapter 8, such a shift in demand has two effects, both of which lead firm 1 to charge a higher price. Demand is larger at any price; this volume effect leads the firm to increase its price if it has increasing marginal cost but has no effect if the firm has constant marginal cost (as in this example). The demand curve has also become less elastic (we see this because demand is linear and the choke price is higher); this price-sensitivity effect also leads to an increase in price, even with constant marginal cost.

Let's consider more generally what can happen in the case of price competition with substitute goods. Section 8.7 explains the following: (a) the volume effect is always positive when the price of a substitute good rises—by definition of a substitute good, an increase in its price causes demand of the other good to rise. (b) It is an empirical regularity that the price-sensitivity effect is positive (always true for linear demand and nearly always true in the real world). Thus, the volume and price-sensitivity effects move in the same direction. A higher price charged by the competitor leads the firm to raise its own price. We conclude that, for nearly all demand functions and whether marginal cost is constant or increasing, the prices in a price competition game with substitute goods are strategic complements.

Exercise 13.1. Suppose the two goods are complements rather than substitutes.

a. Write down an example of a linear demand function for good 1 such that good 2 is a complementary good. Write down and then graph the residual demand curve for two different prices of good 2.

b. Analyze the volume and price-sensitivity effects of an increase in the price of good 2.

c. In strategic competition with complementary goods, would you therefore conclude that the prices are strategic complements or strategic substitutes?

Finding the Nash equilibrium

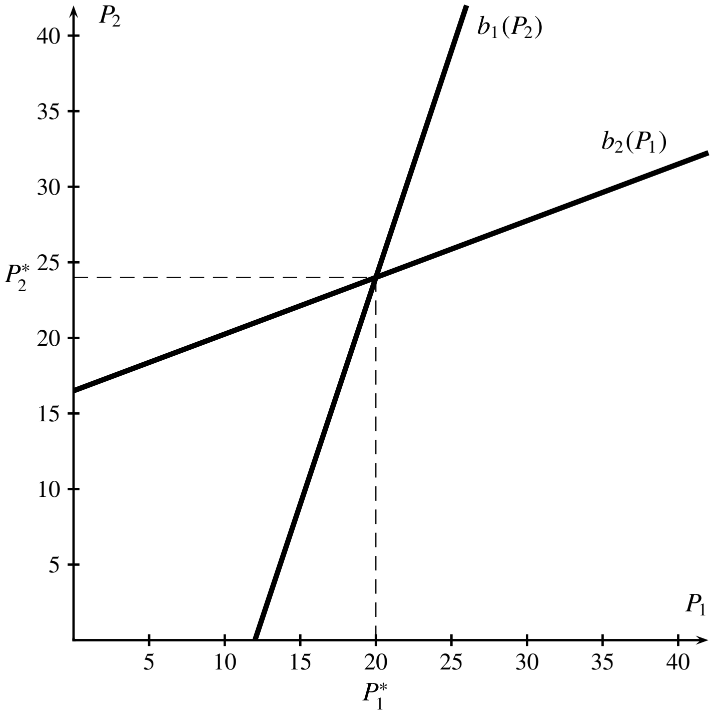
Recall that the Nash equilibrium conditions can be translated into the system of equations $P_1 = b_1(P_2)$ and $P_2 = b_2(P_1)$.

Example 13.1 (Part 3) As we saw in part 2 of this example, the reaction curves are

$$\begin{aligned} P_1 &= 12 + \tfrac{1}{3}P_2\,, \\ P_2 &= \tfrac{33}{2} + \tfrac{3}{8}P_1\,. \end{aligned}$$

The solution to these equations is $P_1^* = 20$ and $P_2^* = 24$. Figure 13.3 shows the two reaction curves on the same axes. We see that they intersect at the Nash equilibrium $P_1^* = 20$ and $P_2^* = 24$.

Figure 13.3



These equilibrium prices result in the following profits:

$$\begin{aligned}\pi_1(20,24) &= (20-8)\,(48-(3\times 20)+(2\times 24)) = 432\,;\\ \pi_2(24,20) &= (24-13)\,(80-(4\times 24)+(3\times 20)) = 484\,.\end{aligned}$$

Exercise 13.2. Consider a price competition model with two firms, 1 and 2, whose demand functions are as follows:

$$\begin{aligned}Q_1 &= 90-3P_1+2P_2\,;\\ Q_2 &= 90-3P_2+2P_1\,.\end{aligned}$$

Each firm's marginal cost is 6. This game is symmetric, and you can use this symmetry to simplify some calculations.

a. Write down firm 1's profit function $\pi_1(P_1, P_2)$.

b. For any P_2, write firm 1's residual demand curve in the linear form $Q_1 = A + BP_1$. Specifically, what are the values of A and B (which may include P_2 as parameters) and what is the choke price?

c. Using the midpoint pricing rule, what is firm 1's optimal price given P_2? That is, what is firm 1's reaction curve in this game?

d. Because the game is symmetric, you can obtain firm 2's reaction curve by changing the indices in firm 1's reaction curve. Graph the two curves on the same axes, with firm 1's price on the horizontal axis and firm 2's price on the vertical axis. Estimate from the graph the Nash equilibrium prices.

e. Because the game is symmetric, there is a symmetric equilibrium. Let P^* be the common price. Write down and solve the single equation that defines a symmetric equilibrium.

f. Calculate the profit of each firm in the Nash equilibrium.

13.3 Price competition with perfect substitutes

As the goods become better and better substitutes, each firm's residual demand becomes more and more elastic, providing greater incentives to undercut the price of the competitor in order to attract market share. As a consequence, the equilibrium prices are pushed down to the firms' marginal cost (if the firms have the same marginal cost), which is the competitive equilibrium price.

Let's consider the polar case in which the goods are perfect substitutes. Then all demand goes to the firm that charges the lowest price; if both firms charge the same price, they split the demand. Assume that the firms have the same marginal cost MC. In this model, having a perfectly divisible currency adds confusion, so assume that there is a smallest currency unit (which is nevertheless quite small). Let this unit be \$0.01.

This game is called the *Bertrand model*. Charging MC (or less) is a dominated strategy because it results in zero profit. However, the equilibrium price cannot be higher than $MC +$ \$0.01 either, because a firm that garners only half the market or less would have an incentive to undercut the other firm's price by a penny, thereby obtaining the entire market at nearly the same price.

As a close approximation, we can summarize the conclusions as follows: If the goods are perfect substitutes and the firms have constant marginal cost, then the price is bid down to the marginal cost and the firms earn zero profit. The price and output are the same as in the competitive model.

13.4 Cournot model: Quantity competition

One story: Price competition with capacity constraints

We consider an alternate model of competition between firms producing perfect substitutes. Rather than choosing individual prices and letting the market determine the demand for their branded products, they choose quantities and let the market determine a common price. This quantity-competition model is also called *Cournot competition*, and a Nash equilibrium of the game is also called a *Cournot equilibrium*.

The distinction between the Cournot and Bertrand models appears to be an arbitrary modeling decision. Yet the outcomes—as we shall see—are quite different. We pause to understand what lies behind this difference.

The Bertrand model has a surprisingly strong conclusion whose message is relevant not only to competition between firms. For example, suppose you are selling a house. Suppose that the house is worth £300K to you and is worth £400K to the only buyer who has expressed any interest. Suppose also that you and the buyer know each others' valuations. Then the two of you will negotiate a price that is likely to be halfway between your valuations. Suppose instead that two buyers have expressed interest and both have valuations of £400K. Then you can play the two buyers off each other, selling the house to the highest bidder. Each buyer has an incentive to make a higher offer than that of the other buyer, as long as the price does not exceed £400K. Hence, you are able to sell the house for nearly £400K and the buyers get almost no surplus.

Nevertheless, the conclusion of the Bertrand model is implausibly strong when applied to competition between two firms in a large market. Is it really true that two firms are enough to make a market perfectly competitive? Will two gas stations located next to each other undercut each other's prices until the price falls to their marginal cost? Perhaps, but there are several reasons why this may not happen. First, the owners may chat over the fence and agree to fix the prices above their marginal cost. Even if this is not possible (it is harder for CEOs of large firms to get away with such illegal collusion), the firms are engaged in a repeated game and each understands that price cuts may be met with retaliatory price cuts; we will see in Chapter 14 that this can lead to stable equilibria in which prices are higher than marginal cost.

There is a third reason, which we now explore in greater depth. Suppose that, if both gas stations charge their marginal cost, then demand at the gas stations exceeds their capacity (determined mainly by the number of pumps they have installed); hence lines form and some drivers go elsewhere. Then either gas station can increase its price above the marginal cost without reducing its sales. The two gas stations will bid down the price only to the point where both gas stations are at capacity. Hence, if the total demand for gasoline at their location is $d(P)$, if the inverse demand is $p(Q)$, and if the firms install capacities Q_1 and Q_2, then the price is bid down to $p(Q_1 + Q_2)$.

These capacities are not fixed forever, of course. If the current price is above the stations' long-run marginal cost, then each gas station appears to have an incentive to increase its capacity. However, the gas stations realize that, once the capacity is installed, the ensuing price competition will cause the price to fall. Hence, each takes into account the effect that its capacity has on the price when choosing how much capacity to invest in. This is different from the behavior of a competitive firm, which invests in capacity presuming it has no effect on the market price.

This capacity-investment model is just one of the stories that can underpin quantity competition. Cournot is also a quite plausible model of any commodity whose output is sold in a market that determines the market clearing price, which all producers receive. Oil and many other commodity markets have this feature. Producers realize that, by trying to sell more output in this market, they will push down the market price. The large oil-producing countries consider most carefully the effect that their supply decisions have on the market price for oil. Electricity producers in California, when supplying surplus electricity to electronic spot markets created in the late 1990s, have been rather conscious about the effects of their decisions on the spot-market price.

The big picture, before the details

Numerical examples of quantity competition games also get heavy, so it is once again useful to first outline the big picture, including (in italics) what it takes to fully specify such a game and solve for the Nash equilibrium.

1. You need to identify the firms in the market and their cost structures.
In numerical examples, we always use constant MC for simplicity.

2. The market has a demand curve for the homogeneous good, which we use to determine how the market price depends on output.

We use the inverse demand curve $p(Q)$. If you are given a demand curve $d(P)$, the first step is to take its inverse.

3. Each firm's individual problem is also like the one studied in Chapters 7 and 8. Given the output decisions of the other firms, there is a downward-sloping relationship between the firm's output and price—it perceives and exploits this market power when choosing how much to produce.

The firm chooses the output level that equates its marginal cost and its marginal revenue, given the influence that its output has on price, which in turn depends on the total output of the other firms. Solve for the firm's optimal quantity decision, treating total output of the other firms as a variable. You then obtain the firm's reaction curve: profit-maximizing output given the output decisions of the other firms.

4. The firms' output decisions are interrelated. A Nash equilibrium consists of quantities produced by all the firms such that each firm's quantity maximizes its profit given the total output of the other firms.

We solve for the Nash equilibrium, either graphically or numerically, as the solution to the system of equations given by firms' reaction curves.

The details

Suppose there are two firms, 1 and 2. They both have the same constant MC. (This is the only cost structure we consider; the game is then symmetric.)

Let $p(Q)$ be the market inverse demand curve. When the firms choose quantities Q_1 and Q_2, the market price is $p(Q_1 + Q_2)$ and hence each firm's margin is $p(Q_1 + Q_2) - MC$. Therefore, firm 1's and firm 2's profits are

$$\begin{aligned}\pi_1(Q_1, Q_2) &= \big(p(Q_1 + Q_2) - MC\big) \times Q_1\,,\\ \pi_2(Q_2, Q_1) &= \big(p(Q_1 + Q_2) - MC\big) \times Q_2\,.\end{aligned}$$

In particular, the firms' revenue curves are

$$\begin{aligned}r_1(Q_1, Q_2) &= p(Q_1 + Q_2) \times Q_1\,,\\ r_2(Q_2, Q_1) &= p(Q_1 + Q_2) \times Q_2\,.\end{aligned}$$

We can solve for player 1's reaction curve by solving the marginal condition $MR_1 = MC_1$. For example, suppose that the inverse demand curve is $p(Q) = 400 - Q$ and that each firm's marginal cost is 100. Firm 1's revenue curve is thus

$$r_1(Q_1, Q_2) = (400 - Q_1 - Q_2) \times Q_1\,.$$

The first derivative with respect to firm 1's quantity is

$$MR_1 = 400 - Q_2 - 2Q_1\,.$$

We then solve

$$
\begin{aligned}
MR_1 &= MC_1\,, \\
400 - Q_2 - 2Q_1 &= 100\,, \\
Q_1 &= 150 - \tfrac{1}{2}Q_2\,.
\end{aligned}
$$

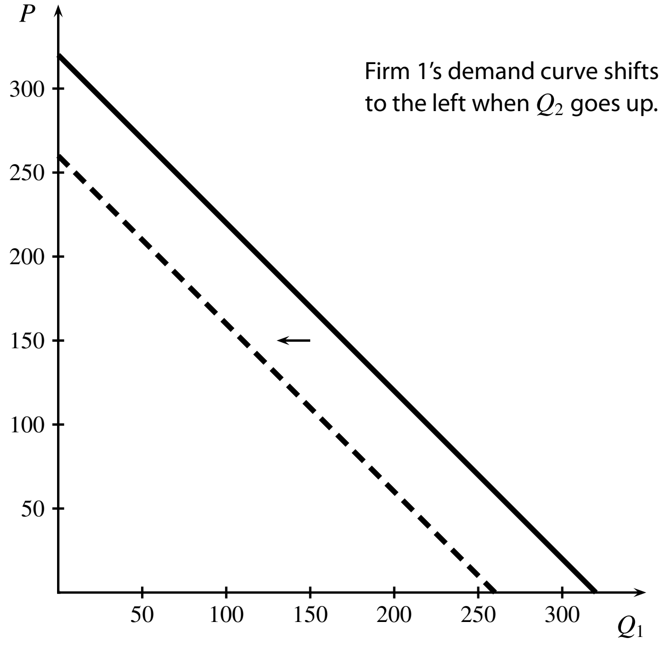
Hence, player 1's reaction curve is $b_1(Q_2) = 150 - \frac{1}{2}Q_2$.

Strategic substitutes

In our example of Cournot competition, firm 1's reaction curve is decreasing. Since firm 2's reaction curve has the same form, the capacities are strategic substitutes.

Let's see whether this is a general property of Cournot competition. We can also frame a firm's decision in that game as one of choosing a point on a residual demand curve. However, it is easiest to work with inverse demand curves. Recall that the inverse demand curve is $p(Q) = 400 - Q$. If firm 2 chooses output $Q_2 = 80$, then firm 1 perceives that the price, as a function of firm 1's output, is $p^r(Q_1) = 400 - (Q_1 + 80) = 320 - Q_1$. Firm 1's problem is to choose the optimal quantity given this inverse demand curve. If instead firm 2's output is $Q_2 = 140$, then firm 1 faces the inverse demand curve $p^r(Q_1) = 400 - (Q_1 + 140) = 260 - Q_1$. Graphically, it is shifted to the left by exactly the extra amount (60) that firm 2 produces. This is seen in Figure 13.4.

Figure 13.4

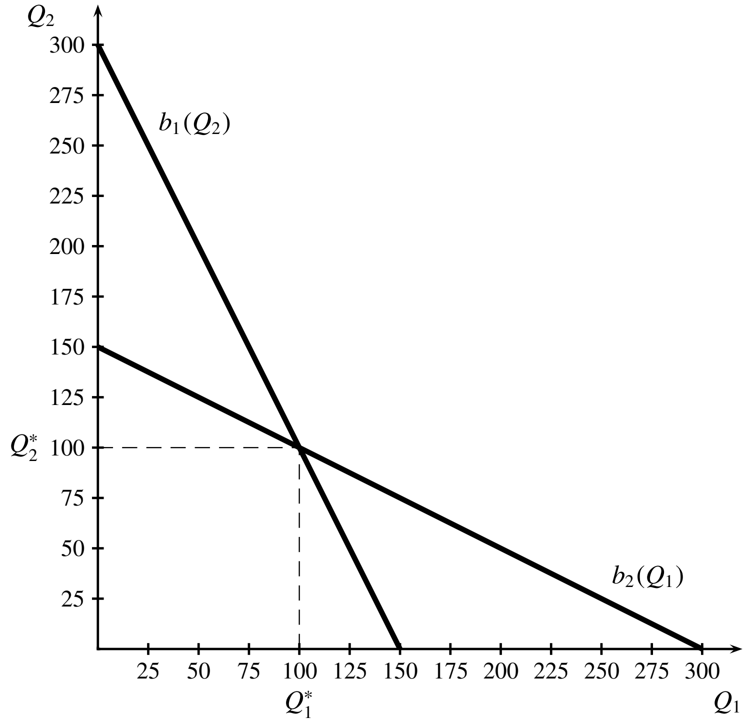


The optimal price for firm 1 is lower but so is the optimal quantity. The shift in the demand curve has reduced the marginal revenue for any extra unit produced because the price at which the extra unit would be sold is lower. This is always true with linear demand curves and is nearly always true for real-world demand curves. Thus, it is a general property that the quantities or capacities in a Cournot game are strategic substitutes.

Nash equilibrium

Because the game is symmetric, firm 2's reaction curve is $b_2(Q_1) = 150 - \frac{1}{2}Q_1$. The reaction curves are shown in Figure 13.5. The equilibrium is at their intersection: $Q_1^* = Q_2^* = 100$.

Figure 13.5



We can also find the equilibrium as a solution to the system of equations $Q_1 = b_1(Q_2)$ and $Q_2 = b_2(Q_1)$. Because the game is symmetric, we can find a symmetric equilibrium as a solution to $Q = b_1(Q)$, or $Q = 150 - \frac{1}{2}Q$. The solution is $Q^* = 100$. Each firm's profit in equilibrium is (400 – 200 – 100) × 100 = 10,000.

Exercise 13.3. Two identical firms produce 128-Mb memory chips in plants that are capacity constrained. The cost per unit of capacity (including short-run marginal cost) is 30, and the market's inverse demand curve is

$$P = 150 - Q\,.$$

You are to analyze the Cournot model of quantity competition.

a. Write down firm 1's revenue and profit as a function of the quantity levels for the two firms.

b. Find each firm's reaction curve by solving $MR = MC$.

c. Graph the reaction curves, with firm 1's quantity on the horizontal axis and firm 2's quantity on the vertical axis. Estimate the Nash equilibrium as the intersection of these two curves.

d. Calculate the Cournot equilibrium capacities.

e. Calculate the Cournot equilibrium market price and the profit for each firm.

f. Suppose the firms were to behave as in the model of perfect competition. What is the equilibrium price and total output? Compare these values with the Cournot equilibrium price and total output.

Cournot model becomes competitive as the number of firms increases

In the Bertrand model, the outcome is competitive even with two firms. A nice feature of the Cournot model is that there is a gradual shift from monopoly to perfect competition as the number of firms increases from one to many. The fewer firms there are, the greater is each firm's effect on the price and hence the more each firm shades its quantity decisions and the less competitive is the outcome. In the other direction, as the number of firms grows, the effect that each firm has on the price becomes negligible and hence the equilibrium prices and total output converge to their competitive levels. This supports the view that the competitive model is a good approximation when there are many firms.

13.5 Quantity vs. price competition

We have presented two models of imperfect competition:

- *Price competition.* With price competition, firms set prices simultaneously and then sell whatever is demanded at those prices. The limiting case in which the goods are perfect substitutes is called the Bertrand model.
- *Quantity competition.* With quantity competition, firms choose how much to sell and then the prices end up being set such that these quantities are demanded by the consumers. We studied only the limit case in which the goods are perfect substitutes, which is called the Cournot model.

When the goods are not very close substitutes, these two models yield similar predictions (though we did not see this). However, that the models are different is most clearly seen in the case of perfect substitutes. The Bertrand model predicts that prices are bid down to marginal cost; the Cournot model predicts that prices stay above marginal cost. This raises the following question: Which model is best?

The price competition model tends to be simpler. It is accurate and intuitive when the goods are not close substitutes. However, if the goods are nearly perfect substitutes, the Cournot model is more accurate and intuitive.

If you want to understand this bottom line, consider these three interpretations of the capacity competition model.

1. It is a reduced form of a two-stage game in which firms first choose long-term fixed investments and then engage in price competition based on short-run marginal cost.[1]
2. It captures an unmodeled dynamic aspect of price competition, namely that undercutting a price will likely lead to a quick price decrease by the competitor. Firms foresee that the outcome of such price competition is determined by the target quantities they intend to sell.
3. It applies to organized markets for commodity goods in which each supplier never sets a price for its own output but rather simply chooses how much to supply at the price for the commodity that is determined by the market. This is approximately how, for example, the market for oil works.

13.6 Imperfect competition with free entry

We have seen various strategic models in which firms earn positive profits. Such profits become smaller as more firms enter the market. We now augment these models by introducing a form of exit and entry. Entry can mean producing a perfect substitute of an existing product or introducing a product that is similar to yet differentiated from existing products. We assume that there is free entry; that is, any firm has access to the same varieties of products and can produce these with the same technology. There is a fixed cost to being in the market.

Firms are forward-looking in that, when deciding whether to enter, each potential firm anticipates what the equilibrium prices will be—given the post-entry market structure—according to one of the models of Sections 13.2–13.5. In particular, they understand how their entry decision will affect the prices of the other firms.

The entry–exit equilibrium conditions are (a) that no active firm takes a loss (taking into account its fixed cost) and (b) that no potential firm can enter and earn an economic profit. The following picture emerges when the firms have no fixed investments but produce differentiated products. When a firm enters, it chooses a product that is differentiated from existing ones in order to avoid excess competition, which would drive its price close to its marginal cost. Each firm thus earns a variable profit (not taking into account its fixed cost). However, as more firms enter, each firm has more competitors with similar products and its variable profit falls. Eventually, the variable profit just balances the fixed cost and there is no further entry or exit.

[1] For the way we studied Cournot competition, this interpretation is an exact fit if the technology is of the following type: fixed investments determine a capacity of production, with a constant marginal cost of capacity; and once capacity is installed, the short-run marginal cost of production is constant up to the capacity but exceeding the capacity is impossible. However, one can adapt the model for more general technologies.

When the fixed costs are high, the array of firms and products is sparse and each firm has considerable market power. When the fixed costs are low, the array of firms and products is rich and each firm has little market power. In the hypothetical limit, as the fixed costs fall to zero, the market becomes saturated and perfectly competitive.

A similar picture emerges when the firms produce perfect substitutes and compete in quantities. As more firms enter the market, the variable profit (the Cournot profit) falls. Eventually, the variable profit just balances the fixed cost and there is no more entry or exit.

13.7 Wrap-up

We studied two games of imperfect competition. In the first game, the firms' goods are related but they are not perfect substitutes, and the firms engage in price competition. A change in one firm's price shifts the other firm's demand curve and hence the other firm's optimal price. In the second game, firms produce perfect substitutes and decide how much to produce. An increase in capacity or output for one firm reduces the market price, thereby changing the optimal capacity or output for the other firm.

Chapter 14

Explicit and Implicit Cooperation

14.1 Motivation and objectives

In a Nash equilibrium, each player chooses her action thinking only about the effect on her own payoff. Therefore, as illustrated starkly in the Prisoners' Dilemma, the Nash equilibrium typically is not efficient from the point of view of the two players. That is, it typically does not maximize the total payoffs of the players and there are often other outcomes that both players would prefer.

The purpose of this chapter is twofold. First, we study in more detail the comparison between Nash equilibrium and what could be achieved through collective action. We introduce the concept of positive and negative externalities for this purpose.

Then we consider how, through repeated interaction, it can be possible to sustain cooperation between the players.

14.2 Positive and negative externalities

The possible inefficiency of Nash equilibrium occurs because each player does not take into account the effect that his or her actions have on the other players' payoffs. Such effects are called *externalities*.[1]

For games with numerical actions, we classify the externalities as positive or negative, as follows.

1. If a higher action of player 1 benefits player 2—meaning that $u_2(A_2, A_1)$ is an increasing function of A_1—then player 1's action has *positive externalities*.
2. If a higher action of player 1 hurts player 2—meaning that $u_2(A_2, A_1)$ is a decreasing function of A_1—then player 1's action has *negative externalities*.

The effort levels in the partnership game have positive externalities because higher effort by one player increases the profit and hence helps the other player. Table 14.1 shows other examples where the actions can be classified as having either positive or negative

[1] This is a rather broad definition of externalities. Classic narrower examples include the direct, nonprice effects that one person's consumption has on others (consumption externalities—e.g., using your cellphone on an elevator has a negative consumption externality) and that a firm's production has on others (production externalities—e.g., pollution is a negative production externality).

externalities.

Table 14.1

Example	Externalities
Two housemates with different music tastes set the volumes on their bedroom stereos.	Negative
Fishing vessels in the same waters choose how intensively to fish.	Negative
Listeners of a private but non-profit (and listener-supported) radio station choose how much to contribute to the station.	Positive
Firms in an R&D race choose how much to invest in R&D.	Negative
Members of a team decide how much effort to devote to the team project.	Positive
Competing firms choose advertising expenditures.	Negative

In price competition with substitute goods

In Example 13.1, which is a game of price competition with substitute goods, the firms' profit functions are

$$\pi_1(P_1, P_2) = (P_1 - 8)(48 - 3P_1 + 2P_2), \tag{14.1}$$

$$\pi_2(P_2, P_1) = (P_2 - 13)(80 - 4P_2 + 3P_1). \tag{14.2}$$

We can see that $\pi_1(P_1, P_2)$ is increasing in P_2 and $\pi_2(P_2, P_1)$ is increasing in P_1. Hence, this game also has positive externalities.

Without having such precise data, we know that any game of price competition with substitute goods has positive externalities: If one firm raises its price, then demand for the other firm's good rises (this is what we mean by substitute goods) and so the other firm's profit rises too.

(You would normally think of a firm as being "hurt" by its competitors. So why do we say that the prices have positive externalities? Because the firm is better off when its competitors set *higher* prices.)

In price competition with complementary goods

Consider price "competition" between two firms whose goods are *complements*. Perhaps firm 1 is Intel (selling microprocessors) and firm 2 is Microsoft (selling operating systems). An increase in one firm's price reduces demand for the other firm's good (this is the definition of complementary goods) and hence hurts the other firm. Therefore, the prices in such a game have negative externalities.

In Cournot competition

In the Cournot quantity competition game of Section 13.4 in Chapter 13, the profit functions are

$$\begin{aligned}\pi_1(Q_1, Q_2) &= (300 - Q_2)Q_1 - Q_1^2 ,\\ \pi_2(Q_2, Q_1) &= (300 - Q_1)Q_2 - Q_2^2 .\end{aligned}$$

We can see that $\pi_1(Q_1, Q_2)$ is decreasing in Q_2 and $\pi_2(Q_2, Q_1)$ is decreasing in Q_1. Thus, this game has *negative* externalities.

Furthermore, *any* Cournot game has negative externalities. No matter what the inverse demand curve is or what the firms' cost curves are, an increase in quantity by one firm pushes down the market price and thus hurts the other firm.

14.3 Individual vs. collective action

General idea

Collective action refers to what the players could achieve if they could sit down, hammer out an agreement on the actions they will take, and then sign a binding contract or otherwise be induced by law or some other mechanism to stick to the agreement. In a Nash equilibrium, on the other hand, players make individual choices and cannot commit to any actions.

The purpose of many laws is to shift behavior away from the Nash equilibrium and thereby benefit everyone. Communitywide collective action is best achieved through such laws. Collective action among a smaller number of parties can be achieved by signing an enforceable contract, by sustaining implicit cooperation through repeated interaction, or by otherwise sustaining trust (for which the field of organization behavior still has better models than does economics). For firms, collective action can also be achieved by forming cartels, by tacit collusion in games with repeated interaction, or by mergers.

If the actions can be classified as having positive or negative externalities, then we can systematically compare the actions that would result from collective action, in the mutual benefit of the two players, with the Nash equilibrium actions.

Positive externalities ⇒ Nash actions are too low. With positive externalities, players (by thinking only about their own payoffs rather than the collective good) do too little of the good thing in the Nash equilibrium—their actions are too low compared to the collective optimum.

Negative externalities ⇒ Nash actions are too high. With negative externalities, players do too much of a bad thing in the Nash equilibrium—their actions are too high.

To be more specific, the phrase "Nash actions are too low" means the following.

1. Starting at the Nash equilibrium, both players' payoffs would go up if the actions were

increased a bit (in the smooth case).

2. Any pair of actions that make both players better off—and hence that could result from bargaining among the players when the default outcome, should no agreement be reached, is to revert to Nash—are higher than the Nash equilibrium actions.
3. The reason a collective optimum is not self-enforcing (i.e., is not a Nash equilibrium) is that each player has an incentive to lower her action.

For example, the speed of driving has a negative externality on other drivers. However, in a Nash equilibrium, each driver disregards the effect his speed has on others. Hence, if the drivers could enforce a binding agreement, then they would collectively choose to drive more slowly. (For example, they may all vote for laws that restrict their speed.)

Partnership game

Because a partnership game has positive externalities, the players' collectively preferred effort levels are higher than their Nash effort levels. The effort levels are not easily observable by courts, but the collective effort levels might be sustained through tit-for-tat strategies by the players (each puts in high effort out of fear that the other player will respond to shirking by shirking himself).

Price competition with substitute goods

Because price competition with substitute goods has positive externalities, the collusive or cartel prices are higher than the Nash prices. Such cooperative action cannot be enforced in courts but can be sustained through repeated interaction. Each firm keeps its price high out of fear that the other firm will retaliate to a price cut by cutting its own price.

Furthermore, if the two firms merge then prices following the merger should rise. This is why firms producing substitute goods can have an incentive to merge even if there are no cost savings.

Either way, the higher prices benefit the firms but hurt the consumers by even more, and so the deadweight loss increases. This is why such mergers may be blocked by antitrust authorities.

Example 14.1 In Example 13.1, one can show that the prices that maximize total profit are $P_1 = 36.7$ and $P_2 = 37.4$. The resulting total profit is 1353. In comparison, we found that the Nash equilibrium prices are $P_1^* = 20$ and $P_2^* = 24$, resulting in a total profit of 916.

Exercise 14.1. Suppose the two firms in Exercise 13.2 merge, so that the two prices are chosen to maximize total profits. (Alternatively, the two firms form a cartel.) This is a two-variable maximization problem, but again we can use its symmetry to reduce it to a one-variable maximization problem because the two prices should be equal to each other. Let *P* be this common price.

a. Write total demand as a function of P. What is the choke price of this demand curve?

b. Determine the optimal post-merger common price using the midpoint pricing rule. What is the total profit? Compare the post-merger price and profit with the Nash equilibrium prices and profit.

There is a message here about setting up incentives in multidivisional firms. Suppose the firm has several product lines that are substitute goods. The prices should be chosen so as to maximize the total profit of the firm. However, if the firm is divided by product line and the manager of each division is told to maximize her division's profit (perhaps with a yearly bonus tied to that profit), then the division managers are engaged in a price competition game just as if they managed independent firms. They will end up choosing prices below those those that maximize total profit of the firm. (Such bonuses are useful to provide incentives for managers to work hard; one must simply be aware of this downside.)

Price competition with complementary goods

If firms produce complementary goods then their prices have negative externalities. In the Nash equilibrium, each firm sets its price without taking into account that a low price would stimulate demand for the other firms' complementary goods and hence benefit the other firms. Thus, if two such firms merge, their prices would fall. This benefits the two firms and also benefits consumers, and there is a drop in the deadweight loss. Such mergers are viewed favorably by antitrust authorities.

Cournot competition

Cournot competition has negative externalities. Thus, the collusive or cartel quantities are lower than the Nash equilibrium quantities. For example, OPEC regulates production among its members to keep output below that which would reign in a Nash equilibrium. (But each country has an incentive to cheat and increase its output, which makes such cartels unstable.) The collusive market price is thus higher, and so is the deadweight loss. Such collusion or mergers among such firms is thus viewed unfavorably by antitrust authorities.

Exercise 14.2. Suppose the firms in the Cournot model of Exercise 13.3 form a cartel that maximizes total profit, with each firm supplying half the market.

a. How much will each firm produce and what will its profit be?

b. Suppose that firm 1 abides by the cartel agreement you just calculated, but that firm 2 cheats by increasing its quantity. What is firm 2's optimal quantity if it assumes (perhaps erroneously) that firm 1 will not react by changing its own quantity? What is firm 2's profit (until firm 1 reacts)?

14.4 Achieving cooperation through repeated interaction

Overview

With repeated interaction and long-term relationships between players, tacit cooperation becomes possible. “Tacit” means that it is sustained by the players own’ reactions in the games, and not by binding contracts or some external enforcement. This is also called *implicit cooperation* or, when the cooperation represents collusion between firms operating in the same market, *tacit* or *implicit collusion*. This provides one way out of the Prisoners’ Dilemma.

Consider a Prisoners’ Dilemma game in which each player’s available actions are C (“cooperate”) and D (“defect”), and where the payoffs are as shown in Table 14.2. In this game, D is a dominant strategy for each player, but (C,C) yields a higher payoff for both players than does (D,D).

Table 14.2

		Player B	
		D	C
Player A	D	1 1	0 4
	C	4 0	3 3

If the game is played repeatedly then D is no longer a dominant strategy each round because what a player does in one round may influence the behavior of the other player in subsequent rounds. Perhaps this fact can be used to achieve tacit cooperation. For example, each player may be willing to cooperate in the short run to encourage the other player to do the same in subsequent rounds. Put another way, it may be possible to sustain a cooperative outcome by means of credible threats and promises.

In order to study such games, we define a class of games, called *repeated games*, in which the same static game—called the *stage game*—is played repeatedly. A strategy for a player is a plan of what action to take at each decision node, even at contingencies the player does not expect to reach. A *subgame perfect equilibrium* is a strategy for each player such that, at every decision node, if the node is reached then the player cannot achieve a higher payoff by unilaterally changing her strategy in the rest of the game. Requiring optimality even at decision nodes that are not reached is how we exclude noncredible threats.

Repeated Nash

It is important to understand that repetition does not automatically guarantee tacit cooperation. Specifically, it is always an equilibrium in the repeated game to play, in each round, a Nash equilibrium of the stage game.

To see this, consider an infinitely repeated Prisoners' Dilemma. Suppose that each player's strategy is to defect no matter what has happened before. Is there any reason for a player to deviate from her strategy? Each round, each player realizes that her action will have no effect on the future play of the game. Therefore, she should choose her current action to maximize the current period payoff. This means defecting.

Grim-trigger strategies

Suppose the Prisoners' Dilemma is played repeatedly and that you are player A. You might reason as follows:

> In the first round, I will cooperate. This way, I signal to player B that I can be trusted to cooperate in the future. If player B sends me a similar signal in the first round, and if my signal catches on so that player B also cooperates in subsequent rounds, then I will reciprocate by continuing to cooperate. However, once player B defects, I will lose confidence and defect for as long as the game is played.

This strategy—"cooperate until the other player defects, and then defect forever"—is called *grim trigger*. ("Trigger" because a defection by the other player triggers defection by you; "grim" because, once your defection is triggered, you defect forever and give up hope of reestablishing tacit cooperation.) This sounds like a reasonable strategy. It is based on a system of threats and promises of the sort: *I promise to cooperate as long as you cooperate, but I will punish you if you deviate from cooperation.*

We now examine whether grim-trigger strategies survive a formal argument. Specifically, are they a subgame perfect equilibrium?

Infinitely repeated games

Assume the game does not have a fixed number of rounds. Either the players expect to play the game forever or, more realistically, they are never sure when the game is going to end. When players trade off current payoffs for future payoffs, they discount the future payoffs either because they are impatient or because they know that the game might end before the future payoffs are obtained.

Consider the decision you face in each round, given the expectation that the other player is also following a grim-trigger strategy and assuming that you have both cooperated so far. If you follow the grim-trigger strategy, then (C, C) is played and your payoff is 3 each period for as long as the game lasts. Suppose instead you defect. As a short-term gain, your payoff is 4 instead of 3 in that round. However, in every round that follows, the other player defects and your payoff is at most 1. Thus, you are weighing the immediate *one-shot* gain of playing D (an extra $4 - 3 = 1$ in that period), against all the future benefits of playing C (an extra $3 - 1 = 2$ in every subsequent round). As long as you do not discount the future too much (you are not too impatient or it is not too likely that the game will end soon), you prefer to continue cooperating.

This argument shows how you are willing to cooperate because of the threat of being punished for defecting. The same reasoning applies to the other player. However, we still have to check that these threats are credible. That is, if a defection were to occur, would the players still follow the grim-trigger strategies? Do the continuation strategies themselves form an equilibrium? The answer is "yes", as follows. Following a defection, the strategies are that the players thereafter play a Nash equilibrium of the stage game. We already showed that such "repeated Nash" strategies are an equilibrium.

Exercise 14.3. Consider the infinitely repeated Prisoners' Dilemma in Table 14.2. Use tools for discounting flows of profits to determine for what discount rate the grim trigger strategy profile is an equilibrium.

Assume that the other player will stick to grim trigger. We check whether it is profitable for you to deviate in the first period. (The same calculations tell us whether it would be profitable to deviate any late date, if no one has yet defected.)

Denote the discount (interest) rate by r.

a. Calculate the present discounted value of your payoff if you continue cooperating forever.

b. Calculate the present discounted value of your payoff is you deviate today (and then continue deviating, since there is no longer any point in cooperating).

c. What is the threshold discount rate above which grim trigger is not an equilibrium?

Repeated games in the real world

There are other strategies than can achieve cooperation in infinitely repeated games. We can get rid of the "grimness" of grim-trigger. For example, suppose instead that, following a defection by player B, player A defects but will switch back to cooperation as long as player B cooperates for 10 periods. We can view these 10 periods as a phase during which player B tries to "make amends". You can check that such strategies also form an equilibrium if the players do not discount the future too much. These strategies have the advantage of being less fragile than grim trigger: if occasionally someone accidentally defects, it is possible to restore cooperation.

There are many other equilibria, including just playing Nash over and over again (no cooperation at all), and it is possible to achieve intermediate payoffs between those of "(C,C) forever" and "(D,D) forever". This "anything can happen" result is known as the "folk theorem for repeated games".[2]

However, as a rule of thumb, the equilibria that tend to arise as social norms are those that lead to high payoffs for the players. Also, if one is playing such a game in the absence

[2] It is so called because it was known to game theorists long before someone published a formal proof.

of a social norm, then a good strategy is to start cooperating and see whether the other player plays along. You risk getting a low payoff for a few periods, but this risk can be outweighed by the prospect of a higher payoff in the long term.

Such tacit cooperation breaks down when people discount the future more, such as when a long-term relationship seems likely to end soon. An example is that we often observe a collapse of social norms in a society during a civil war or other internal turmoil, which breaks up relationships and creates great uncertainty about the future benefits of long-term relationships.

An example of documented collusion sustained as the equilibrium of a repeated game was the NASDAQ collusion in the early 1990s. The NASDAQ traders could not sign contracts to enforce the collusion and they were too numerous to work out a self-enforcing agreement in a smoke-filled room. Yet a social norm developed: traders who did not "play along" would not have trades passed to them; if you were a trader who passed trades to someone who wasn't playing along, then trades would not be passed to you either. No one had an incentive to deviate from this social norm, even though short-term profits could be made by undercutting other traders.

14.5 Wrap-up

Because Nash equilibrium is based on the idea of individual action, there is no reason to expect that the equilibria of games will be efficient compared with what could be achieved by collective action. This chapter uses the ideas of negative and positive externalities to classify how the equilibrium actions compare to the collectively desirable actions. We then consider ways to achieve collective action, such as through repeated interaction.

Chapter 15

Strategic Commitment

15.1 Motives and objectives

Actions you take today influence the actions that others take later on. This chapter looks at how timing matters and how to look ahead when making decisions. Players need to anticipate the moves of others and also understand how players can anticipate one's own moves. Our focus is on how to use strategic commitment in order to shift other players' actions in a way that helps you.

Here are some examples:

1. You are plan to engage in price competition with a competitor. Beforehand, you have the opportunity to make an investment that reduces marginal cost. Beyond the trade-offs you would face as a monopolist, what are your strategic considerations?
2. You plan to engage in price competition with a competitor. Beforehand, you can invest in an advertising campaign that differentiates your product from that of your competitor. It does not make your product appear better, just different. What are your strategic considerations?
3. Your firm is currently a monopolist but faces entry from a potential entrant. You must now decide how irreversible to make your investments in production capacity. What are your strategic considerations?

In each case, your actions affect the payoffs, and hence the behavior, of all players (including yourself) in subsequent stages of the game. You want to use this influence to your advantage. To do so, you must answer two questions: (a) What do I want to achieve? (b) How do I achieve it?

15.2 Sequential games

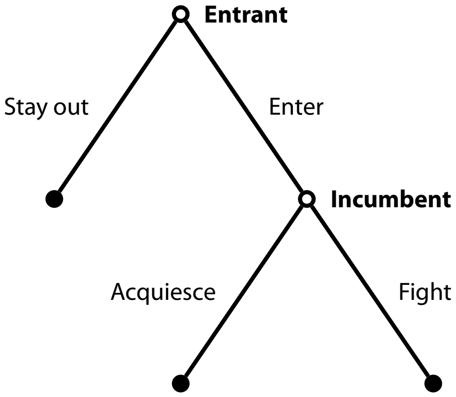
In a (purely) sequential game, players never choose actions simultaneously. Whenever a player chooses an action, she knows the preceding decisions that other players have made.

Game trees

Consider a situation in which a firm (the Incumbent) currently has a monopoly in a market and a competitor (the Entrant) is thinking of entering. The Entrant is not sure how aggressively the Incumbent will respond. To construct a game-theory model, we focus on the decisions that interest us the most, namely: (1) the Entrant must decide whether to enter the market or to stay out. (2) If the Entrant enters, the monopolist must decide whether to start a price war (*Fight*) or to share the market as duopolists (*Acquiesce*).

We can represent the decisions and their sequencing by a *game tree*. Each node of the tree represents a decision that one of the players must make; we label the node by the name of the player. The branches from the node represent the available actions. We label each branch by the name of the action. The sequence of decisions is from the *root* or initial node of the tree to the *leaves* or terminal nodes of the tree. The game tree for the situation previously described is shown in Figure 15.1.

Figure 15.1



An outcome of this game is an entire sequence of actions (path of play). This game has three possible outcomes: the Entrant stays out; the Entrant enters and then the Incumbent acquiesces; or the Entrant enters and then the Incumbent fights. Note that we can identify each outcome by the terminal node of the tree that is reached.

Missing so far is the data about how the players feel about the outcomes. We assume that the players are maximizing profit and that a player's payoff is her profit. Suppose we have the data about profits given in Table 15.1.

Table 15.1

Incumbent's monopoly profit =	12
Each firm's duopoly profit =	4
Each firm's profit if there is a price war =	1
Entrant's entry cost =	2
Entrant's profit if she never enters the market =	0

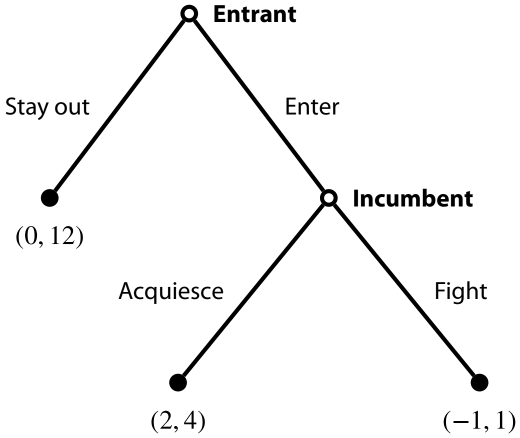
For example, the Entrant's overall payoff if she enters and then the Incumbent fights is -1. Then Table 15.2 shows the players' payoffs for each of the outcomes.

Table 15.2

Outcome	Entrant's payoff	Incumbent's payoff
Stay out	0	12
Enter, Acquiesce	2	4
Enter, Fight	−1	1

To summarize this information, we label each terminal node of the game tree with the payoffs of the players for that outcome. A common convention is to write the payoffs as a pair $(2, 4)$, where the first number is the payoff of the first mover (the Entrant, in this case) and the second number is the payoff of the second mover (the Incumbent, in this case). The game tree with the payoffs is shown as Figure 15.2.

Figure 15.2



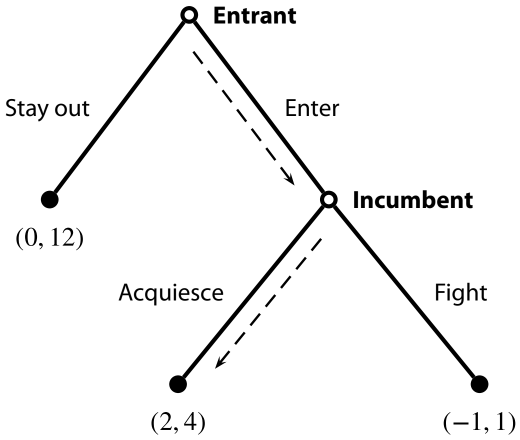
Strategic reasoning

When the Entrant decides whether or not to enter, she first predicts what the Incumbent will do. For example, If the Entrant believes that the Incumbent will fight, then expects that her own payoff for entering is -1. The Entrant therefore prefers to stay out (payoff equal to 0).

The Entrant should not form arbitrary beliefs about the Incumbent's behavior. Instead, a key element of the Entrant's strategic reasoning is to examine the Incumbent's strategic reasoning. Although the Incumbent might wish to threaten the Entrant with a price war in order to deter entry, this is an empty threat that he would willingly carry out. Once the Entrant enters, it is a *fait accompli*; the Incumbent's best response is to acquiesce (payoff of 4) rather than fight (payoff of 1). Thus, if the Entrant has confidence that she understands the payoffs of the Incumbent and that the Incumbent is a rational decision maker, then the Entrant should predict that the Incumbent will respond to entry by acquiescing. The Entrant concludes that her payoff from entering is 2, which is higher than her payoff from staying out. The outcome is that the Entrant enters and the Incumbent acquiesces.

We can represent this outcome on the game tree by drawing an arrow next to the edge for the action that is chosen at each node, as shown in Figure 15.3.

Figure 15.3



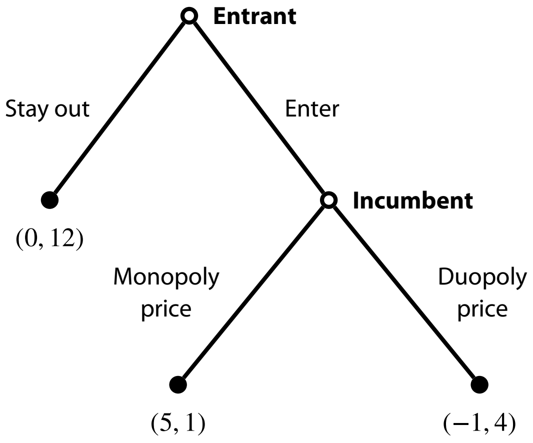
Let's work through another example. The story is similar to the previous one. This time the Incumbent has to decide whether to keep his monopoly price—which would allow the Entrant to gain a large market share by undercutting the monopolist—or to reduce the price to a level where the Incumbent and the Entrant share the market as duopolists. Also, the Entrant's entry cost is now higher than in the previous example. The basic profit structure is shown in Table 15.3.

Table 15.3

Incumbent's monopoly profit =	12
Each firm's duopoly profit =	4
Incumbent's profit if he does not reduce his price =	1
Entrant's profit if Incumbent does not reduce his price =	10
Entrant's entry cost =	5 .

The game tree is thus as shown in Figure 15.4.

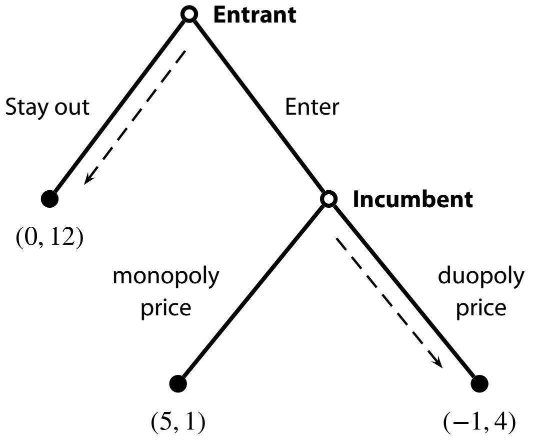
Figure 15.4



What outcome do you predict for this game?

The Entrant sees a market in which the monopolist is charging a high price. There seems to be an incentive for the Entrant to come into the market, undercut the monopolist, and make a sizable profit. However, if the Entrant does enter then the Incumbent should respond by lowering his price to the duopoly price, since this price will yield a payoff of 4 instead of 1. The Entrant anticipates this response by the Incumbent, thereby realizing that entering will yield a payoff of -1. Therefore, she decides to stay out. This outcome is depicted in Figure 15.5.

Figure 15.5



Note that Figure 15.5 shows an arrow indicating that the Incumbent would change its price to the duopoly price if the Entrant entered the market, even though this contingency is never reached. This contingent behavior is what leads the Entrant to stay out of the market, and hence it is important for understanding the outcome of the game. A plan for what the player will or would do at every node—even those nodes that the player expects will not be reached—is called a *strategy*.

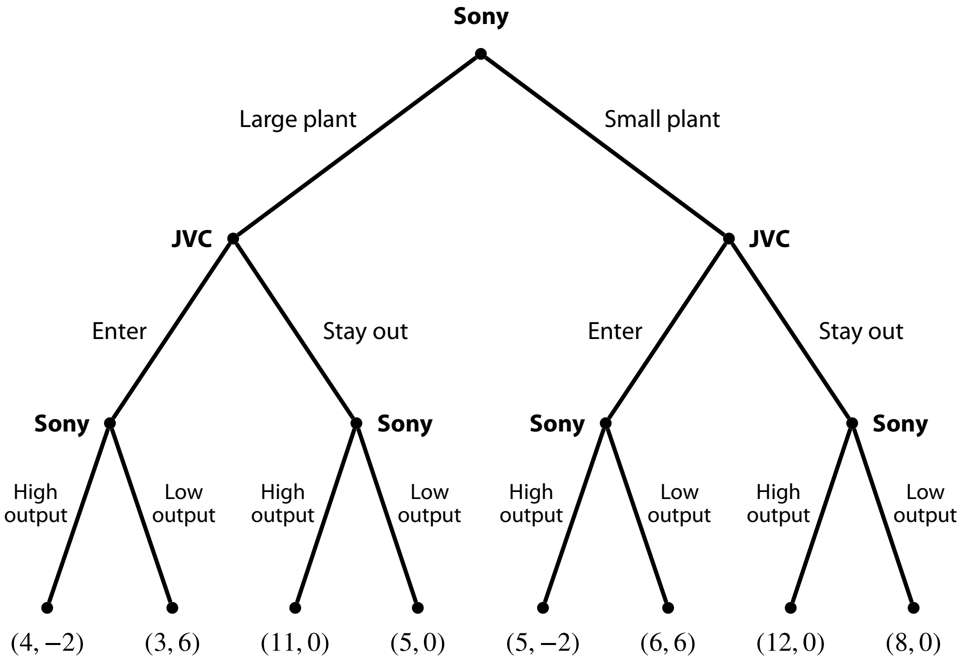
Backward induction

The analysis we used for predicting the outcome of these games is called *backward induction*. We first consider the decision nodes that are immediately before the end of the game, asking what the decision maker would do if that node were reached. This tells us the payoffs that all the players expect if that node is reached. Then we consider the next nodes up in the tree, and so on.[1] It is called "backward" because you start at the end of the game and reason backward in time. This procedure represents not only the way an observer might analyze the game but also the way a player decides what to do at each decision node. Exercise 15.1 asks you to apply backward induction to a more complex game.

[1] This is a multiperson version of dynamic programming.

Exercise 15.1. Consider the following strategic situation. Sony has a monopoly on Discman. It plans to construct a plant this spring and can build either a large plant or a small plant. JVC has a one-time opportunity to enter this market in the summer. It can choose either to enter or to stay out. Sony will then choose its output level; it can choose either high output or low output. This sequence of actions is depicted in Figure E15.1.

Figure E15.1



(*Note*: Sony's profit is listed first and JVC's profit is listed second.)

The profits of the firms are determined by the actions they take. For example, JVC gets a profit of zero if it does not enter the market and incurs a fixed cost if it does enter. The market price depends both on whether JVC is also in the market and on whether Sony chooses a high or low output level. Sony's fixed costs are higher and its variable costs are lower if it constructs a large plant than if it constructs a small plant. These factors lead to the profits that are labeled at the bottom of each terminal node.

a. Apply backward induction to this tree and thereby determine, at each of the decision nodes in the tree, what action the firm would take if it found itself at this node. (Mark the branch for each decision.)

b. What is the overall sequence of actions that will be taken, and what is each firm's profit?

c. How is this model related to "entry deterrence"?

15.3 Stackelberg games

Overview

For each two-player simultaneous-move game, we can imagine a sequential game in which one of the players (the "leader") makes her decision first and then the other player (the "follower") makes his decision after observing what the leader did. Actually, there are two such sequential games, depending on which player is the leader. These are called *Stackelberg games*. We can determine the outcome by applying backward induction. Note that we consider the Stackelberg games *different games* from the simultaneous-move game, even though they share some of the same components (players, actions, payoffs).

Example 15.1 Consider the following competition game. There is an Incumbent firm and a potential Entrant. The decision of the Incumbent is to advertise or not. The decision of the Entrant is to enter or not. The payoffs are as follows.

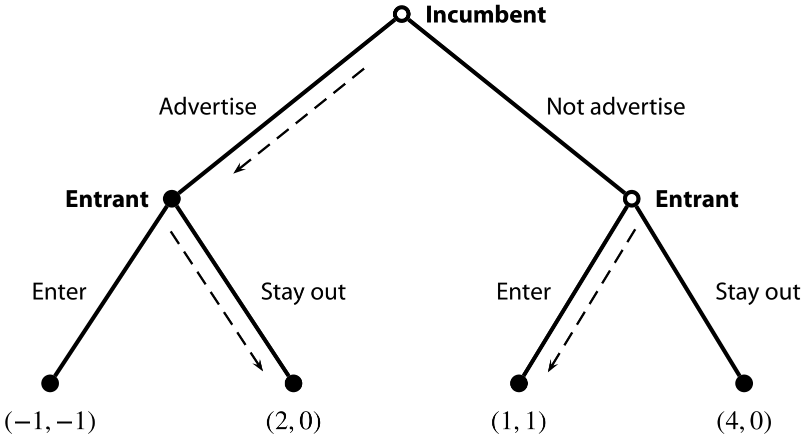
Table 15.4

		Entrant	
		Enter	Not enter
Incumbent	Advertise	−1 −1	0 2
	Not advertise	1 1	0 4

If the firms make their decisions simultaneously, then "Not advertise" is a dominant strategy for the Incumbent. The Entrant therefore enters. This yields the unique Nash equilibrium.

Suppose instead that the Incumbent's advertising decision is made first and is observed by the Entrant before the Entrant decides whether or not to enter. Then firms then play the sequential game depicted in Figure 15.6. Not advertising is no longer a dominant strategy, because the Incumbent's action influences the entry decision. The backward induction solution is revealed in the figure: the Incumbent advertises in order to deter entry.

Figure 15.6

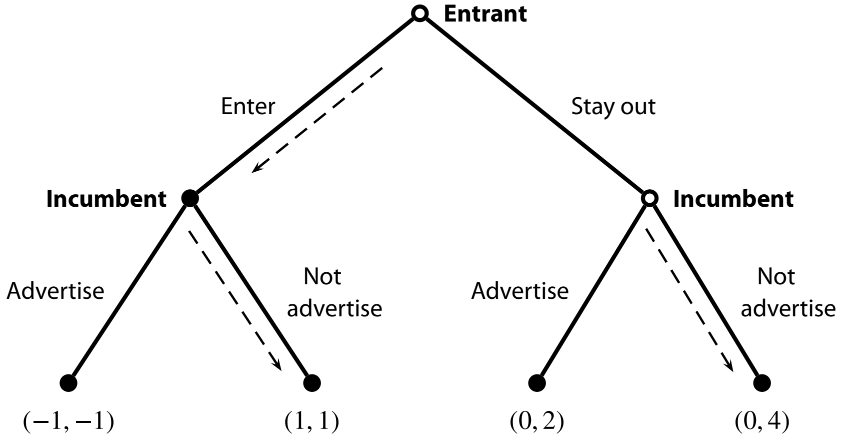


In Example 15.1, the Incumbent's payoff is higher as leader in the Stackelberg game than it is in the Nash equilibrium of the simultaneous-move game. This fact is not fortuitous: *in a Stackelberg game, the leader can cause any Nash equilibrium of the static game to be played.* She need only play her own Nash equilibrium action; the follower's best response is then to play his Nash equilibrium action. In Example 15.1, if the Incumbent chooses to not advertise then the Entrant reacts by entering, and this outcome is the Nash equilibrium of the static game. Therefore, *the Stackelberg leader's payoff is at least as high as her best Nash equilibrium payoff.* The leader cannot always do strictly better than her Nash equilibrium payoff, but often she can.

In Example 15.1, the Entrant gets a lower payoff as follower than in the Nash equilibrium. There are other games in which the follower instead gets a higher payoff than in the Nash equilibrium of the static game.

Suppose the timing in Example 15.1 is reversed, so that the Entrant is leader. Then the firms play the sequential game in Figure 15.7.

Figure 15.7



This time it is the Entrant who can secure at least her Nash equilibrium payoff, by choosing to enter so that the Incumbent responds by not advertising. This is the best she can do.

Comparing the two Stackelberg games in Figures 15.6 and 15.7, we observe that the Incumbent and the Entrant each get a higher payoff as leader than as follower. We therefore say that each has a *first-mover advantage*. If instead a player gets a lower payoff as leader than as follower, we say that she has *first-mover disadvantage*.

Timing and commitment

The purpose of comparing a static game (the two players move simultaneously) and the two corresponding Stackelberg games (one in which one player moves first, another in which the other player moves first) is to see how the timing of moves matters.

A Stackelberg game is meant to capture, in a simple way, the possibility that a player can commit to actions. Real-world commitment is usually more complicated than that captured in a Stackelberg game. For example it may be difficult for a player to truly lock in an advertising campaign; yet by (say) paying for part of it in advance, the player can make it in her interest to have the campaign. Similarly, it is difficult to tie one's hands from concerning prices, but one can commit to low prices by making large up-front investments that reduce the marginal cost of production. A full-blown game-theory model would incorporate the actions that allow one to achieve such commitment, but a shortcut is simply to model the Stackelberg game in order to understand (i) how one would exploit such commitment and (ii) what value it has.

How timing matters

Let's compare the behavior of the Entrant in Example 15.1. More generally, let's compare the behavior of the follower in the Stackelberg game with her behavior in the static game.

In the static game, the Entrant chooses a best response to what she *expects* the Incumbent to do. In the Stackelberg game, the Entrant chooses a best response to what she *has observed* the Incumbent do. In neither case can the Entrant influence the Incumbent. Thus, it is not the distinction between how the Entrant (follower) views the games that makes the two games so different.

In contrast, consider how the Incumbent (leader) views the games. In the Nash equilibrium of the static game, he also simply chooses a best response to what she expects the Entrant to do. However, in the backward-induction solution of the Stackelberg game, the Incumbent understands that *he can influence* the Entrant's action. Thus, to see how timing matters, we need to understand the following.

- How would the Incumbent (leader) like to influence the follower's action? This brings into play the externalities that the Entrant's action has on the Incumbent. Which shift in the Entrant's actions increase the Incumbent's payoff?
- How can the Incumbent achieve such a shift? This brings into play the strategic interaction, and the concepts of strategic complements and substitutes. How should the

Incumbent shift her own action in order to induce the Entrant to respond in the right way?

In Example 15.1, entry by the Entrant hurts the Incumbent. Thus, the Incumbent's goal when committing to an advertising strategy is to deter such entry. Advertising by the Incumbent makes entry less profitable for the Entrant, so the Incumbent deters such entry by advertising.

15.4 Stackelberg games with numerical actions

Backward induction and reaction curves

Consider a Stackelberg game with two players, 1 and 2, and numerical actions. Suppose player 1 is the leader. We solve this by backward induction, starting with the follower's move.

- We already have a tool for summarizing how the follower behaves in the Stackelberg game: After player 1 chooses her action A_1, player 2 responds by choosing $b_2(A_1)$. In this setting, the name "reaction curve" for b_2 makes sense, because it shows player 2's reaction to player 1's action.
- Moving to the next level of the game, player 1 realizes that her payoff from choosing an action A_1 is $u_1(A_1, b_2(A_1))$. She chooses her action to maximize this inferred payoff.

Once we have determined player 1's Stackelberg action A_1^s, we find player 2's Stackelberg action as $A_2^s = b_2(A_1^s)$. We can then calculate the Stackelberg payoffs of both players.

Exercise 15.2. Consider a price competition model involving two firms, firms 1 and 2. Firm 1 is Stackelberg leader and firm 2 is a follower. You are given the following information:

1. firm 1's demand function is $Q_1 = 90 - 3P_1 + P_2$;
2. firm 2's reaction curve is $P_2 = 5 + P_1/2$;
3. firm 1 has no fixed cost and a constant marginal cost of 10.

Write down firm 1's profit as a function of the price that it charges.

Three examples

The appendix to this chapter works through three such examples: a partnership game, a game of price competition with substitute goods, and a Cournot game. The numerical examples are in an appendix because they are a bit mechanical and calculation-intensive.

Table 15.5

Partnership Game

	Player 1		Player 2	
	A_1^*	U_1^*	A_2^*	U_2^*
Nash	8	128	8	128
Stackelberg: 1 leads	16	160	12	272
Stackelberg: 2 leads	12	272	16	160

Price Competition with Substitute Goods

	Firm 1		Firm 2	
	P_1^*	Π_1^*	P_2^*	Π_2^*
Nash	20	432	24	484
Stackelberg: 1 leads	22	441	$24\frac{3}{4}$	$552\frac{1}{4}$
Stackelberg: 2 leads	$20\frac{11}{18}$	477	$25\frac{5}{6}$	494

Cournot Competition

	Firm 1		Firm 2	
	Q_1^*	Π_1^*	Q_2^*	Π_2^*
Nash	100	10,000	100	10,000
Stackelberg: 1 leads	150	11,260	75	5,625
Stackelberg: 2 leads	75	5,625	150	11,260

However, (a) they drive home the way in which the leader should reason about her decision problem and (b) they provide grist for our mill when we draw more qualitative conclusions about strategic commitment.

The results are summarized in Table 15.5; see the appendix for the parameters of the examples.

15.5 Comparisons between Nash and Stackelberg

(Assume, for the rest of this chapter, that there is a unique Nash equilibrium.)

Some observations about the three examples

The following hold for both the partnership game and the price competition game.

1. The Stackelberg payoffs of both players are higher than the Nash payoffs (we know this is always true for the leader; in this game it is true also for the follower).

2. The Stackelberg actions of both players are higher than the Nash actions.
3. Each player has a higher payoff when follower than when leader. Hence, the Stackelberg game has a first-mover disadvantage.

In contrast, in the Cournot game, the following statements hold.

1. The follower's Stackelberg payoff is lower than its Nash payoff.
2. The Stackelberg action of the leader is higher than Nash, but the Stackelberg action of the follower is lower than Nash.
3. Each player has a higher payoff when leader than when follower. Hence, the Stackelberg game has a first-mover advantage.

Can we systematically understand these comparisons? What do the partnership game and the price competition game with substitute goods have in common? How are they different from the Cournot game?

Basic principles: Who thinks differently

To understand how the Stackelberg and Nash equilibria compare, we need to think about how the play of the games differs.

The follower's strategic thinking in the Stackelberg game is not much different from his thinking in the static game. In each case, he either expects or observes an action by the other player and then chooses a best response to it, without being able to influence the other player's action in any way.

The leader's thinking is what differs. In the Nash equilibrium of the static game, she chooses a best response to what she expects that the other player will do, not thinking that she has any influence over his behavior. However, a Stackelberg leader knows that the follower will adjust his action in response to the leader's action.

Comparing the follower's actions

So the first thing we should ask is, "how does player 1 want player 2 to adjust his action?" That is: What is the leader trying to achieve by moving away from her Nash action? The answer to this question tells us how the follower's action differs from his Nash action.

The answer depends on the direction of the externalities. If the game has positive externalities, then the leader wants the follower to increase her action. This is true in the partnership game and in the price competition game with substitute goods, which is why the follower's Stackelberg action is higher than his Nash action. If instead the game has negative externalities, then the leader wants the follower to decrease his action. This is true in the Cournot game, which is why the follower's Stackelberg action is lower than his Nash action.

Comparing the leader's actions

Now we can determine how the leader's Stackelberg action differs from Nash by seeing how the leader has to adjust her action in order to achieve the desired effect on the follower. If the actions are strategic complements, then the leader should adjust her action in the same direction that she wants the follower's action to move. This is true in both the partnership game and the price competition game with substitute goods. We previously concluded that, in both games, the leader wants the follower to raise his action (positive externalities); thus, the leader raises her own action. That is, in the partnership game, the leader's Stackelberg effort is higher than her Nash effort; in the price competition game, the leader's Stackelberg price is higher than her Nash price.

If the actions are strategic substitutes, then the leader should adjust her action in the opposite direction. This is true in the Cournot game. The actions in that game have negative externalities and so the leader wants the follower to lower its action. To achieve this, the leader raises its action. Thus, the leader's Stackelberg capacity is higher than her Nash capacity.

Comparing the follower's payoffs

The follower chooses a best response whether in the static game or the Stackelberg game. So the follower's payoff depends on what he is choosing a best response to. His Stackelberg payoff is higher than his Nash payoff if he likes the leader's Stackelberg action more than the leader's Nash action; it is lower if he likes the leader's Stackelberg action less. This depends on how the leader's Stackelberg action differs from her Nash action and on the direction of the externalities.

For example, in the partnership and price competition games, the leader's Stackelberg action is higher than her Nash action. Furthermore, the game has positive externalities. Thus, the increase in the leader's Stackelberg action benefits the follower, and so the follower's Stackelberg payoff is higher than his Nash payoff.

Consider instead the Cournot game. The leader's Stackelberg action is also higher than her Nash action. However, the game has negative externalities, so the increase in the leader's action hurts the follower. Therefore, the follower's Stackelberg profit is lower than his Nash profit.

We can deduce the following pattern for a game with positive or negative externalities.

1. The follower's Stackelberg payoff is higher than Nash if the actions are strategic complements.
2. The follower's Stackelberg payoff is lower than Nash if the actions are strategic substitutes.

The reasoning is as follows. The leader shifts her action away from Nash in order to induce the follower to help her. If the actions are strategic complements, then the leader shifts her action in the same direction that she wants the follower's action to move, and so she also helps the follower. In instead the actions are strategic substitutes, then—to induce the follower to help her—she hurts the follower.

First-mover advantage or disadvantage

We can deduce the first-mover advantage or disadvantage as follows.

1. Whenever the follower's Stackelberg payoff is lower than his Nash payoff, there is a first-mover advantage: the follower would obtain more than his Nash payoff if he were instead the leader. (See Example 15.4.)
2. When the follower's payoff is higher than his Nash payoff, there is a first-mover disadvantage. This is because the leader shifts her action in a beneficial way more than the follower does. The follower has the advantage of simply choosing a best response to the leader's action. For example, in the partnership game, partner 1 increases her effort more than partner 2 does.

Thus, there is a first-mover advantage if the actions are strategic substitutes and there is a first-mover disadvantage if the actions are strategic complements.

15.6 Examples using the basic principles

Three examples, revisited

We have broken the comparison into three questions. For the partnership game, it is summarized in Table 15.6.

Table 15.6

Question: Compared to Nash …	Answer	Reason	Implication
What does the leader want the follower to do?	Raise A_2	Positive externalities	$A_2^s > A_2^n$
How does the leader achieve this?	Raises A_1	Strategic complements	$A_1^s > A_1^n$
How does the leader's shift affect the follower?	Raises U_2	Positive externalities	$U_2^s > U_2^n$

For the price competition game with substitute goods, the analysis is summarized in Table 15.7. The comparisons and the analysis are the same as for the partnership game (Example 15.2) because, in both that game and in this price competition game, the actions have positive externalities and are strategic complements.

Table 15.7

Question: Compared to Nash …	Answer	Reason	Implication
What does the leader want the follower to do?	Raise P_2	Positive externalities	$P_2^s > P_2^n$
How does the leader achieve this?	Raises P_1	Strategic complements	$P_1^s > P_1^n$
How does the leader's shift affect the follower?	Raises Π_2	Positive externalities	$\Pi_2^s > \Pi_2^n$

The Cournot game is different, because the actions have negative externalities and are strategic substitutes. See Table 15.8.

Table 15.8

Question: Compared to Nash …	Answer	Reason	Implication
What does the leader want the follower to do?	Lower Q_2	Negative externalities	$Q_2^s < Q_2^n$
How does the leader achieve this?	Raises Q_1	Strategic substitutes	$Q_1^s > Q_1^n$
How does the leader's shift affect the follower?	Lowers Π_2	Negative externalities	$\Pi_2^s < \Pi_2^n$

Price competition with complementary goods

Consider a price competition game with complementary goods. The prices have negative externalities and are strategic substitutes. Therefore, the analysis is the same as for the Cournot game. See Table 15.9.

Table 15.9

Question: Compared to Nash …	Answer	Reason	Implication
What does the leader want the follower to do?	Lower P_2	Negative externalities	$P_2^s < P_2^n$
How does the leader achieve this?	Raise P_1	Strategic substitutes	$P_1^s > P_1^n$
How does the leader's shift affect the follower?	Lower Π_2	Negative externalities	$\Pi_2^s < \Pi_2^n$

Your turn

Once you have absorbed the three questions and understand how to identify the direction of the externalities and the strategic interaction, it is easy to analyze a new situation. The following two exercises give you an opportunity to do so.

Exercise 15.3. Suppose your firm is one of two large multinationals with critical mining operations in a small country. There is a risk that a socialist party will come to power with a program to nationalize the mines, so the two firms make financial contributions to the incumbent party to help keep it in power. The higher is the other firm's contribution, the better off your firm is because it increases the odds that the incumbents will retain power. Hence, the actions have positive externalities. The higher is one firm's contribution, the less the other firm feels it needs to contribute in order to keep the incumbents in power. Hence, the actions are strategic substitutes.

Let's say each firm would contribute $40M if the contributions were simultaneous. However, suppose that your firm has the opportunity to make a contribution before the other firm does and can thereafter commit to not making more contributions. Analyze how the outcome here differs from when the case where the contributions are made simultaneously.

Exercise 15.4. Consider two firms that produce brand-differentiated substitute goods. Suppose that actions are advertising expenditures, where the advertising promotes one brand at the expense of the other. The advertising expenditures have negative externalities. Furthermore, it is likely that the expenditures are strategic complements; if one firm increases its advertising, the other firm does the same in order to retain market share. Compare (a) the outcome when you can commit to your advertising expenditure before the other player does with (b) the outcome when the firms choose their advertising levels simultaneously.

15.7 Strategic commitment

A Stackelberg game captures the ability of a player to make a strategic commitment. The insights apply not only to games in which each player takes a single decision before the game ends. First, even in ongoing relationships, there may be significant events that occur infrequently (such as the investments in capacity for production of a new product) and the timing comparison we have made captures the differences between when the decisions are made simultaneously and when one player is able to make—and commit to—its decision first.

Second, in repeated games, sustaining cooperation is usually better for both players than any Stackelberg outcome, yet this may be difficult to achieve. Perhaps there are too many small players or the duration of the interaction is too uncertain. Then it may be to a player's advantage to stake out a position and let the other players react to it, effectively becoming a Stackelberg leader. For example, when a market contains one large firm and many small firms, the large firm may become such a Stackelberg leader.

Third, the qualitative insights that come from studying Stackelberg games apply to other kinds of strategic commitment. The idea is roughly as follows. Firms may choose actions that affect their payoffs and therefore affect their incentives to choose actions. When considering how to do this, the logic of the Stackelberg games provides good guidance. Suppose that a firm, engaged in price competition with substitute goods, invests in technology that has an upfront cost but then reduces the marginal cost of production. Making a large investment is like committing to more aggressive pricing. But we know that this will lead the other firm to price aggressively as well. Therefore, the firm prefers to make a small investment in order to soften the competition. On the other hand, suppose instead that the action of the other firm is to enter or stay out. Then making a large investment can deter entry.

15.8 Wrap-up

We can systematically compare the Stackelberg payoffs and actions with the Nash payoffs and actions according to whether the actions have positive or negative externalities and are strategic complements or substitutes.

By repeating the analysis for the various combinations of positive vs. negative externalities and strategic complements vs. substitutes, we obtain Table 15.10.

Table 15.10

Positive/ negative externalities?	Strategic complements/ substitutes?	(Stackelberg value vs. Nash value)				First-mover advantage/ disadvantage?
		Leader		Follower		
		Payoff	Action	Payoff	Action	
Positive	Complements	Higher	Higher	Higher	Higher	Disadvantage
Positive	Substitutes	Higher	Lower	Lower	Higher	Advantage
Negative	Complements	Higher	Lower	Higher	Lower	Disadvantage
Negative	Substitutes	Higher	Higher	Lower	Lower	Advantage

Appendix: Three numerical examples of Stackelberg games

Partnership game

Example 15.2 Consider the partnership game from Chapter 12. Player 1's payoff function is

$$u_1(A_1,A_2) = 8A_1 + 8A_2 + A_1A_2 - A_1^2 .$$

Player 2's reaction curve is $A_2 = 4 + \frac{1}{2}A_1$. Hence, to calculate player 1's inferred payoff when she is the leader in the Stackelberg game, we replace A_2 by $4 + \frac{1}{2}A_1$ in player 1's payoff function. This yields

$$u_1(A_1, b_2(A_1)) = 8A_1 + 8\big(4 + \tfrac{1}{2}A_1\big) + A_1\big(4 + \tfrac{1}{2}A_1\big) - A_1^2 = 32 + 16A_1 - \tfrac{1}{2}A_1^2 .$$

The marginal condition for maximizing player 1's payoff is $16 - A_1 = 0$, or $A_1^s = 16$. Therefore, player 2's Stackelberg action is $A_2^s = b_2(16) = 4 + \frac{1}{2}16 = 12$. The players' payoffs U_1^s and U_2^s are

$$U_1^s = u_1(A_1^s, A_2^s) = 8 \times 16 + 8 \times 12 + 16 \times 12 - 16^2 = 160 ,$$
$$U_2^s = u_2(A_2^s, A_1^s) = 8 \times 12 + 8 \times 16 + 12 \times 16 - 12^2 = 272 .$$

Suppose instead that player 2 is the leader. Because this game is symmetric, we get the same answers for the leader's and follower's actions and payoffs. Thus, we have $A_2^s = 16$, $A_1^s = 12$, $U_2^s = 160$, and $U_1^s = 272$.

Stackelberg games of price competition

Consider the Stackelberg game of price competition with goods that are substitutes.

If firm 1 charges P_1, then firm 2 reacts by charging $P_2 = b_2(P_1)$ and firm 1's demand is $d_1(P_1, b_2(P_1))$. Firm 1 chooses a price that maximizes profit assuming it faced this inferred demand curve.

Example 15.3 In Example 13.1, firm 1's demand curve is $Q_1 = 48 - 3P_1 + 2P_2$ and firm 2's reaction curve is $P_2 = \frac{33}{2} + \frac{3}{8}P_1$. Firm 1's inferred demand is therefore

$$d_1(P_1, b_2(P_1)) = \big(48 - 3P_1 + 2\big(\tfrac{33}{2} + \tfrac{3}{8}P_1\big)\big) = \big(81 - \tfrac{9}{4}P_1\big) .$$

The choke price of this demand curve is $\frac{4}{9} \times 81 = 36$. Since firm 1's marginal cost is 8, firm 1's optimal price according to the midpoint price rule is $P_1 = \frac{1}{2}(8 + 36) = 22$. Now that we know firm 1's optimal price, we can determine firm 2's response and the resulting profits. Firm 2 chooses $P_2 = \frac{33}{2} + \frac{3}{8}(22) = 24\frac{3}{4}$. Firm 1's profit is then

$$(22 - 8)\big(48 - 3 \times 22 + 2 \times 24\tfrac{3}{4}\big) = 441$$

and firm 2's profit is

$$\big(24\tfrac{3}{4} - 13\big)\big(80 - 4 \times 24\tfrac{3}{4} + 3 \times 22\big) = 552\tfrac{1}{4} .$$

Suppose instead that firm 2 is the leader. Firm 2's demand curve is $Q_2 = 80 - 4P_2 + 3P_1$, and firm 1's reaction curve is $P_1 = 12 + \frac{1}{3}P_2$. Firm 2's inferred demand is therefore

$$d_2(P_2, b_1(P_2)) = \left(80 - 4P_2 + 3\left(12 + \tfrac{1}{3}P_2\right)\right) = 116 - 3P_2\,.$$

Using similar calculations, we can show that the Stackelberg actions are $A_2^s = 25\frac{5}{6}$ and $A_1^s = 20\frac{11}{18}$ and that the Stackelberg profits are $\Pi_2^s \approx 494$ and $\Pi_1^s \approx 477$.

Exercise 15.5. Recall the price competition model in Exercise 13.2. Consider the corresponding sequential game in which firm 1 is the leader and firm 2 is the follower.

a. Using firm 2's reaction curve, determine firm 1's demand as a function of the price it charges. Write firm 1's profit as a function of the price it charges.

b. What is the choke price of firm 1's inferred demand curve? What is firm 1's optimal price according to the midpoint pricing rule?

c. Calculate the price that firm 2 charges in equilibrium and the profits for the two firms.

d. Does this Stackelberg game have a first-mover advantage or disadvantage?

Stackelberg game of capacity competition

Example 15.4 Consider a Stackelberg version of the quantity competition (Cournot) model from Chapter 13. Because we are are studying strategic commitment, we view quantity decisions as irreversible investments in installed capacity; thus, we also refer to this as capacity competition.

Let firm 1 be the leader. The inverse demand curve in that example is $p(Q) = 400 - Q$, and each firm's marginal cost is constant and equal to 100. Firm 2's reaction curve is $Q_2 = 150 - \frac{1}{2}Q_1$. Hence, firm 1's profit, as a function of Q_1, is

$$(p(Q_1 + b_2(Q_1)) - MC)Q_1 = (400 - (Q_1 + (150 - \tfrac{1}{2}]Q_1)) - 100)Q_1 = \left(150 - \tfrac{1}{2}Q_1\right)Q_1\,.$$

The first derivative with respect to firm 1's price is

$$\frac{d\pi_1}{dQ_1} = 150 - Q_1\,.$$

Setting this to zero (the marginal condition that marginal profit equals zero) and then solving for Q_1, we obtain $Q_1 = 150$. Firm 2's response is therefore $b_2(150) = 150 - \frac{1}{2}150 = 75$. The market price is $p(150 + 75) = 400 - 225 = 175$, and the firms' profits are

$$\begin{aligned}\Pi_1^s &= (175 - 100)150 = 11{,}250\,,\\ \Pi_2^s &= (175 - 100)75 = 5625\,.\end{aligned}$$

(Since the game is symmetric, when firm 2 is the leader the calculations are the same but the roles of firms 1 and 2 are reversed.)

Exercise 15.6. Consider the quantity competition model of Exercise 13.3. Suppose that one firm can act as a Stackelberg leader. Calculate the leader's optimal capacity. How much will the other firm produce? What is the market price? What is each firm's profit?

Chapter 16

Games with Incomplete Information

16.1 Motives and objectives

We have so far studied games under the simplifying assumption that the players know the payoffs (both their own and the payoffs of the other players). We now relax this assumption. This makes the prediction of other players' actions more complicated than in a Nash equilibrium or backward induction solution. Furthermore, players realize that their actions influence the beliefs of other players.

We are interested in situations where players do not merely face uncertainty. There can be uncertainty even in single-person decision problems, such as when you must decide whether to fix up the guest room, not knowing how many people will actually visit you. Instead, we are interested in cases in which the players have different information. We call this *asymmetric* or *incomplete information*. The information that one player has, but that other players do not, is called her *private information*.

For example, the seller of a used car has information about the reliability of the car that is not available to the buyer; a firm's customers have information about their valuations that is not observed by the firm; an employee has better information about his ability and activities on the job than his employer does; and traders in the stock market do not have the same information about the companies whose stocks they are trading.

Private or asymmetric information is the rule, not the exception. Private information has so far come up in this book mainly in the pricing model. Except in the benchmark of perfect price discrimination, consumers have private information about their valuations. Otherwise, we have assumed in this book that the parties in economic transactions have the same information (in fact, that all parties are perfectly informed). We have thereby focused on those forces in economic situations—also important and pervasive—that are not driven by private information. Private information adds many other fascinating and important features to economic situations in general and to managerial decision making in particular.

In this chapter, we study auctions as leading examples of static games with incomplete information. We describe various auction forms and consider both the case of private values, where each person knows his own valuation of the object being sold, and the case of common values, where all bidders have the same valuation of the object but possess different information about what this valuation is.

We also study dynamic games with incomplete information (in Section 16.4). There our emphasis is on reputation and how a small amount of private information can have a large effect when the horizon of the game is long enough.

16.2 Private values

In any game of incomplete information (static or dynamic), it is possible that each player knows her own payoffs but is uncertain about the payoffs of others. Then we say that there are *private values*. This asymmetry of information is important because each player finds it more difficult to predict the actions of other players. Here are some examples with approximately private values: (a) in competitive bidding for a construction project, each bidder knows its own cost structure but not that of the other bidders; and (b) a concertgoer and a ticket scalper may bargain over the price of a ticket, each knowing how much the ticket is worth to himself but not to the other party.

If you are in a game without private values, asymmetry of information creates an additional strategic consideration. Other players may then have information that would help you determine your own payoffs. For example, (a) if a seller of a used car is eager to trade, you may infer that the car is not in good condition; (b) if a company's stock price falls even though there is no new public information, you may infer that an insider has learned some bad news and has traded based on the information; and (c) if a company repurchases its own stock, you may infer that the management believes the stock is undervalued.

16.3 Auctions

Motivation

Consider the following two situations.

1. Suppose you own a construction company and are going to bid on a project to create an underground tunnel for a busy highway as it passes in front of a business school. The technical details of this project are well known to all the bidders, but the companies' costs for this project differ and each company knows only its own cost. Your firm's cost is €100M. The rules of the bidding are as follows. Each firm submits a sealed bid. The firm with the lowest bid wins the contract, but receives, as payment for the project, the value of the second-lowest bid.
2. Suppose you are a wine broker. A large collection of wine is to be auctioned off in one large lot via an English auction (the price rises until only one bidder is left, who then wins the auction and pays the final bid). Because of the amount of wine involved, the only bidders are wine brokers like you, who are interested in the wine for resale

rather than for personal consumption. You and the other wine brokers are equally good at reselling wine, but you are uncertain about the market value of the wine and have different estimates. You estimate the value to be 100,000.

How much should you bid in each auction? How are the two situations alike? How are they different?

Types of auctions

Auctions are organized markets for selling goods that involve some procedure for eliciting bids from the buyers as well as rules for determining—based on the bids—who gets the goods and how much the bidders pay. In the simplest auctions, a single good or bundle is sold (e.g., works of art, collectors' items, confiscated goods, and houses are often sold by auctions). Auctions can also be used to sell multiple units of goods (e.g., government debt such as U.S. Treasury bonds) or to sell multiple interrelated goods (e.g., communications airwaves). These auctions are more complex and we will not have time to consider them.

We can interchange the roles of the buyers and seller in auctions. That is, auctions can also be used by a buyer to purchase a good from one of several sellers. Governments frequently use auctions ("competitive bidding") to procure goods such as roadways or military equipment. The analytics of these auctions are the mirror image of the analytics for auctions used to sell goods, so we can restrict attention to auctions for selling goods without any loss of generality.

One broad classification of auctions is between *oral* auctions and *sealed-bid* auctions. In sealed-bid auctions, each bidder submits a single bid without knowing the bids of the other bidders. In an oral auction, bidders revise their bids based on the history of bidding. There are two main types of oral auctions.

- *Dutch auction.* The price moves down (starting from an initial high level) until one of the bidder announces that he wants the object at the current price.[1]
- *English auction.* The price moves up (starting from an initial low level) until no bidder is willing to go above the current price. The last bidder in the auction wins the object and pays the final price.[2]

There are two common forms of sealed-bid auctions.

- *First-price.* The highest bid wins and the winner pays the value of her bid.
- *Second-price.* The highest bid wins but the winner pays the value of the *second-highest* bid.[3]

[1] It is called a "Dutch auction" because it has been used for a long time in the Netherlands to sell flowers. It is also called a "descending" auction.

[2] The word "auction" has its roots in this common and ancient trading mechanisms. It comes from the Latin word *augere*, which means "to increase". English auctions are also called "ascending" auctions.

[3] A sealed-bid, second-price auction is also called a *Vickrey* auction, named after the economist who proposed this auction form 40 years ago and who since won the Nobel prize in economics.

Oral and sealed-bid auctions are not as different as they appear. The first-price sealed-bid auction and the Dutch oral auction are equivalent, as follows:

1. In a first-price sealed-bid auction, each bidder chooses a bid. The bidder with the highest bid wins and pays the value of this bid.
2. In a Dutch auction, each bidder decides a threshold at which she will claim the object. The bidder with the highest threshold wins, and pays the value of this threshold.

The second-price sealed-bid auction and the English oral auction are roughly equivalent, as follows:

1. In a second-price sealed-bid auction, each bidder chooses a bid. The bidder with the highest bid wins and pays the value of the second-highest bid.
2. In an English auction, each bidder chooses a threshold at which she will drop out of the bidding. The bidder with the highest threshold wins, and she pays the going price when the second-to-last bidder dropped out—this is the value of the second-highest threshold.[4]

We therefore restrict attention to sealed-bid auctions and refer to them merely as “auctions”.

Private-value vs. common-value auctions

A sealed-bid auction can be modeled as a static game. The players are the bidders and each bidder’s action is a bid. We let V_i be bidder i’s valuation of the object. If bidder i wins the auction and pays P, then her payoff is $V_i - P$. If she loses the auction and hence neither receives the good nor pays anything, then her payoff is 0.

There is usually incomplete information, meaning that valuations of the bidders are not common knowledge. We said that a game of incomplete information has private values if each player knows her own payoffs. In an auction, this means that bidder i knows V_i but may not know the other bidders’ valuations. Our earlier example of the construction contract is likely to have (approximately) private values: the bidders might have different costs, but each knows its own cost of building the tunnel.

What the private-value assumption excludes is the possibility that a bidder has information that would help the other bidders determine their valuations. For example, suppose that art collectors are buying a painting not merely to hang in their living rooms and enjoy but also as an investment. Each bidder brings to the auction some information that helps predict the future value of the painting and hence how much each bidder would be willing to pay for the painting today. We then say that the valuations are correlated.

The opposite extreme from private values is when each bidder has the exact same valuation for the good but the bidders come with different information about this valuation. We then say the bidders have *common values*. Here are some examples with approximately

[4] There is a small difference between the English and second-price sealed-bid auctions: in the English auction, bidders can observe at what price each bidder drops out of the bidding.

common values.

1. If all the bidders at an art auction are dealers who intend to immediately resell the painting, then the painting has the same value (equal to its resale value) to all the bidders, but they may have different estimates of this value (based on different information about the art market).
2. In the wine auction example, the value is equal to the revenue that will be generated net of distribution and marketing costs, and this is approximately the same for all bidders. Yet different bidders have different estimates of these revenues.
3. Suppose the auction is for the right to exploit an oil field and the bidders are oil companies with identical cost structures. Each bidder has its geologists perform exploratory studies, which provide partial information about the value of the oil field. However, at the end of the day, the oil field will generate the same profit no matter who owns it.

Private-value auctions

In a private-value auction, as in general private-value games, the bidders care about the payoffs of the other bidders because they want to predict their strategies. However, this is irrelevant to a bidder (or player in any private-value game) if she has a dominant strategy. Each bidder has a dominant strategy in a second-price auction with private values: to be her true valuation (see Exercise 12.1). The second-price auction is appealing because each player has a simple, truthful strategy and because it is efficient, in that the player with the highest valuation always wins the auction.

Consider now a first-price auction with private values. Bidding one's valuation is clearly not optimal because it guarantees a payoff of 0. Instead, each bidder shades her bid below her valuation, realizing that the lower is the bid, the higher is the risk of losing the auction but the lower is the price to be paid when winning the auction. How much lower than her valuation should bidder i bid? This will depend on her beliefs about others' bidding behavior.

Specifically, given beliefs about the other players' valuations and about their bidding strategies, she can derive the cumulative distribution function $F(B^\circ)$ of B°, the highest bid of the other players. Remember that, for example, $F(10)$ is the probability that $B^\circ \leq 10$. If she bids B_i, then she wins if $B^\circ \leq B_i$; this occurs with probability $F(B_i)$.[5] When she wins, her payoff is $V_i - B_i$. Therefore, her expected payoff is $(V_i - B_i)F(B_i)$. As B_i goes up, the probability $F(B_i)$ of winning goes up but the margin $V_i - B_i$ goes down.

Calculating an equilibrium is complicated and beyond the scope of this book. Nevertheless, using the following result (which is also interesting on its own right) we will be able to gain some further insight into the first-price sealed-bid auction without having to calculate explicitly the Nash equilibrium.

[5] We are ignoring what happens when there are ties, which are honestly not worth worrying about.

The Revenue Equivalence Theorem

Which auction, first-price or second-price, gives the seller a higher expected revenue? At first instance, you might think that the first-price auction does because the winner pays her own bid instead of paying the second-highest bid. However, the bids are lower in a first-price auction. It turns out that these two considerations balanced each other out, so that *the first-price and second-price auction have the same expected revenue*. This result is called the *Revenue Equivalence Theorem*.

From this fact we can determine, for example, what happens to the bids in a first-price auction as there are more bidders. We know that bidders bid their valuations in the second-price auction. The more bidders there are, the lower is the gap (in expectation) between the highest and second-highest valuations. Therefore, the lower is the surplus (in expectation) that the winner obtains. The same must happen in the first-price auction if it is to generate the same revenue. We can therefore conclude that, as the number of bidders increases in a first-price auction, the competition among them becomes fiercer and hence the bids in a first-price auction are closer to the bidders' true valuations.

Common-value auctions

When there are common values, you would like to use the other bidders' private information in order to determine how much the object is worth. Those bidders will not reveal that information to you directly, but you can infer something from their actions. For example, if the other bidders bid high, you may infer that they have favorable information about the value of the object and so you may revise your own estimate upward.

You might think that such inference is hopeless in a sealed-bid auction because you must submit your own bid without observing any of the other bids. We will illustrate that this is not the case. Suppose each bidder adopts this reasoning and therefore concludes that, in a second-price auction, her best strategy is to bid her estimate of the object's value based on her own information. Suppose that your own estimate is €100,000, that you win the auction with this bid, and that the second-highest bid is €95,000. How do you feel? You have just learned that you had the highest estimate of all the bidders. In particular, you have learned that one other bidder had private information that caused him to have an estimate of €95,000 and also that all the other bidders had information that led to estimates below €95,000. This should make you revise your own estimate downward. In fact, your revised estimate may be below the €95,000 that you have to pay, and you regret winning the auction. This is called the *winner's curse*.

The same can occur if you bid naïvely in a first-price auction. For example, assume that you enter the wine auction without having any information on other bidders' valuations. Further, assume that the wine will be awarded by means of a first-price sealed-bid auction. Suppose you shade your bid by 10% based on your analysis of the private-value auction and hence bid €90,000 . To make matters simple, assume that all other bidders follow a similar reasoning and thus bid 90 percent of their respective estimates of the value. Once again, by winning the auction you learn that all the other bidders' information led to lower estimates, which makes you revise your estimate downward, perhaps below €90,000. In particular, the greater the number of bidders, the greater the amount by which you have overestimated the true value.

The winner's curse was first documented in 1971 by oil engineers, who observed that oil companies that won auctions for drilling rights ended up discovering far fewer oil reserves than they originally hoped for.

To conclude, in both the second-price and first-price auctions, you should shade your bid below what you would bid if there were private values in order to avoid the winner's curse. Furthermore, the more bidders there are and the better informed you think they are, the more conservatively you should bid.

Wrap-up on auctions

The main points to be taken from our discussion on auctions are the following.

- Auctions are games of incomplete information that can be analyzed using the tools of game theory.
- Depending on the nature of the uncertainty, we may classify actions in terms of common versus private values.
- Bidding strategies depend on whether the values are common or private, on the type of auction that is used (first-price versus second-price sealed), and on the number of participants.
- In a second-price auction with private values, it is a dominant strategy for each bidder to bid his true valuation.
- In a first-price auction with private values, there is no equilibrium in dominant strategies. In a Nash equilibrium, each bidder bids less than his valuation. The more bidders there are, the closer are the bids to the bidders' true valuations.
- With private values, all auction forms yield the same expected revenue for the seller.
- When bidding in a common-value auction, one should be aware of the winner's curse and therefore bid more conservatively than in a private-value auction.

Let's return to and compare our motivating examples: the construction bid and the wine auction.

1. In the first example, competitive bidding is used by the state to *buy* something; in the second example, an auction is used to *sell* something. However, we noted that this is not a important analytic distinction.
2. The first example is a second-price sealed-bid auction; the second example is an English auction. We noted that such auctions are equivalent.
3. According to the construction procurement story, there are private values. Each firm knows its own cost but not that of the other firms. Therefore, each firm should bid its true cost (since this is a second-price auction).

4. According to the wine auction story, there are common values. In the end, the wine collection is worth the same amount no matter which broker wins it, but the brokers come into the auction with different information about this value. Although the English auction is like a second-price sealed-bid auction, the brokers should not bid their estimates of the value of the wine. Instead, each broker should bid below his own estimate to avoid the winner's curse. The better is the information of competing bidders, the more a broker should shade his own bid.

16.4 Reputation and grains of doubt

Backward induction is a compelling principle that people apply intuitively in strategic situations. You can improve your strategic ability by a conscious application of this principle. However, applying backward induction with the simplifying assumption of complete information can lead to misleading conclusions when the horizon is long, because a small amount of incomplete information in such games can have a large effect.

Let's start with some examples in which we end up with unrealistic conclusions; we shall then see how incomplete information can resolve the paradoxes.

Finitely repeated games

Consider a repeated Prisoners' Dilemma game that has a known, fixed horizon. For example, suppose everyone knows that the stage game is played exactly ten times. Then we have a finitely repeated game. In a grim-trigger strategy, players cooperate because they fear triggering future retaliation. In the last round of the game, such fear is not an issue and hence no one has an incentive to cooperate. Thus, in the last round of any repeated game, the players will play a Nash equilibrium.

Hence, in round 10 of this finitely repeated Prisoners' Dilemma, the players will both defect no matter what has happened before. In round 9, players know that (D,D) will be played in round 10 no matter what either player does in round 9. Hence, once again there is no fear of retaliation (or promise of cooperation) that will induce players to cooperate in round 9, and instead they end up playing (D,D) no matter what happened before. Continuing with this backward induction, cooperation unravels completely and the equilibrium is the Nash equilibrium (D,D) in each round.

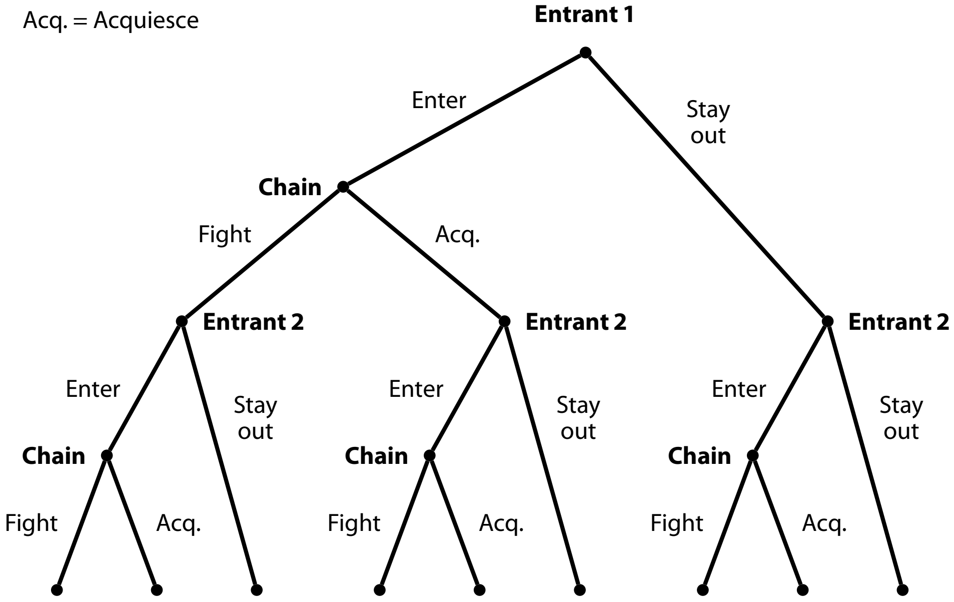
Thus, no matter how long the horizon is (as long as it is finite), the only equilibrium is for the players to defect in each period, even though we would suspect that some cooperation can be sustained. For example, why would a player not cooperate in initial rounds in order to develop a reputation as someone who can be trusted, so that the other player cooperates as well?

Chain-store paradox

Here is another example, called the "chain-store paradox".[6] Consider a chain store with many outlets, each of which faces the threat of entry in its local market. Suppose that entry would take place at different times for different stores. Each time there is an entrant, the chain must decide whether to "fight" or "acquiesce". To fight means to spend money on an advertising campaign and to engage in a costly price war, in order to drive the entrant out of the market or at least make it regret entering, even though such retaliation hurts the chain as well. The game tree, not showing payoffs, is shown in Figure 16.1 for a chain with two stores.

Figure 16.1

Acq. = Acquiesce



Though the game tree would become unwieldy to draw, you can imagine what this game would be like if the chain had, say, 100 stores and so might have to fend off entry from 100 potential entrants.

Let's solve the game by backward induction. The final stage is the potential entry in the 100th market and the reaction by the chain. According to the story we told, this stage has the same kind of payoffs as in Figure 15.3. Following entry, it is in the chain's interest to acquiesce rather than to fight the entry, because the price war and advertising campaign are so costly. Therefore the entrant chooses to enter. All this takes place no matter what has happened before in the other 99 markets.

Moving one step back, consider the possible entry and reaction in the 99th market. The entrant and the chain, having applied backward induction themselves, both realize that their actions in this round have no bearing on the rest of the game, and hence they concern

[6] This example was presented by the Nobel laureate Reinhard Selten to motivate a theory of dynamic games with incomplete information.

themselves only with the profits in this market. By the same logic as in the 100th market, the chain acquiesces following entry and the entrant chooses to enter.

Continuing backward in time, we reach the same conclusion for each market. Hence, the outcome of the game is that, in each market, the entrant enters and the chain acquiesces.

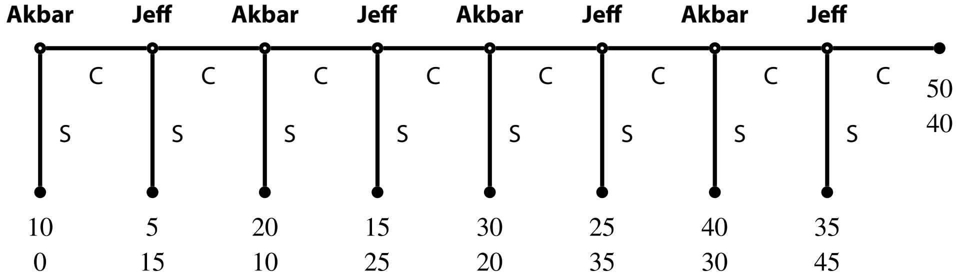
This is counterintuitive. Why wouldn't the chain fight the first entry, in order to develop a "tough" reputation that might deter future entry?

Exercise 16.1. Another striking example is the *centipede game*. It is not drawn from real life, though it has been used in laboratory experiments.

Here is the story behind the centipede game. There are two players, Akbar and Jeff, who take turns deciding whether to continue or stop the game. There is a prize that is initially worth $10. Each time one of the players says "Continue", the prize grows by $10 and it is the next player's turn. As soon as one of the player says "Stop", the game ends. The player who says "Stop" receives half of the prize plus $5; the other player gets the rest of the prize (half the prize minus $5). For example, if Akbar says "Stop" when the prize is $70, then he gets $40 and Jeff gets $30. The game continues for a fixed number of rounds. In the last round, if the player says "Continue" then the game also ends, but the payoffs are as if the game had continued one more round and the next player had said "Stop".

Suppose that Akbar goes first and the play continues for 8 rounds. Then the game tree is as shown in Figure E16.1.

Figure E16.1



Play starts at the upper-left node and continues to the right or down. At each terminal node, Akbar's payoff is written above Jeff's. This is called the "centipede game" because the game tree resembles a centipede (use your imagination!).

a. Apply backward induction to this centipede game and thereby determine, at each of the decision nodes in the tree, what action the player would take if he finds himself at this node. (Mark the branch for each decision.) What is the outcome of the game?

b. What would you actually do in the centipede game if you were Akbar? How would your answer change if the game lasted only 2 rounds? If it lasted 50 rounds?

Reputation

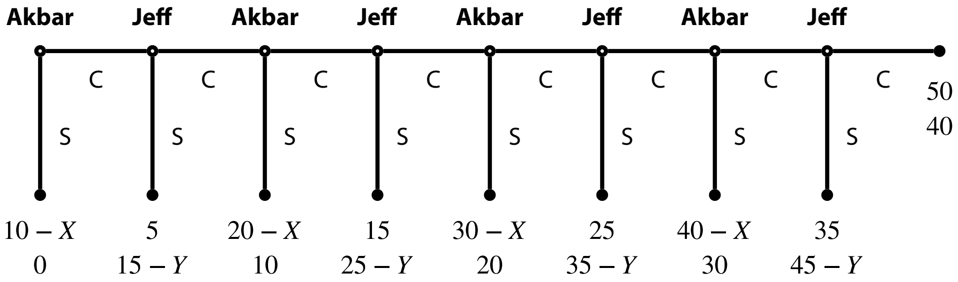
In the finitely repeated Prisoners' Dilemma, the chain-store paradox, and the centipede game, the intuition not captured by the formal model is that some players would choose certain actions early in the game in order to develop a reputation and thereby influence the actions that other players choose.

Developing a reputation means to influence other players' beliefs about what kind of player you are. However, because we started out assuming that there is complete information, the payoffs of these games are common knowledge and the players know everything about each other. To capture the intuition of "reputation", we need to introduce incomplete information.

Our main point will be: a little bit of incomplete information can go a long way. Because actually calculating the equilibria of these long-horizon games is tedious, we proceed informally.

Consider first the centipede game, but now we introduce some uncertainty. Perhaps the payoffs shown in Figure E16.1 are the monetary payoffs set up by the experimenter, yet the players may care about more than just the monetary payoffs. For example, perhaps each player may feel bad about being the one who says "Stop" and so reducing the total prize money that is shared. Let X and Y be the "psychic cost" of such guilt for Akbar and Jeff, respectively, so that the actual game is

Figure 16.2



This is a game with private values. Akbar's psychic cost X is known to him but not to Jeff, whereas Jeff's psychic cost Y is known to him but not to Akbar. We refer to X and Y as Akbar's type and Jeff's type, respectively. Each player believes that the other player's type ranges between 0 and some number greater than 5.

If a player's type is above 5, then that player's dominant strategy is to always say "Continue". We say that the player is a "nice guy". If the other player knows this, then he should also say "Continue" except perhaps in the last round. However, our interest is in the possibility that players say "continue" at the beginning of the game even if initially each puts small probability on the other's being a nice guy.

Let's suppose that the initial beliefs are that X and Y are distributed uniformly on $[0, 6]$. This means that each player believes the other is a nice guy with probability $1/6$. We will show that the following is the equilibrium: Each player always says C ("continue") until the second-to-last round. Then Akbar says C if $X \geq 2.5$; in the final round, Jeff says C if $Y \geq 5$.

It is clear that, in the final round, Jeff says C if and only if he is a nice guy. Consider the second-to-last round. Akbar says C not only if he is a nice guy, but also if his value of X is less than 5 because there is chance that Jeff is a nice guy and hence says C in the next round. Specifically, Akbar makes the following trade-off. If he says S ("stop"), then he gets $40 - X$ for sure. If he says C, then he gets 35 if $Y < 5$ or 50 if $Y \geq 5$. Because Jeff has so far always said C, Akbar still believes that Y is uniform on $[0, 6]$. Therefore, the expected payoff from saying C is $(5/6)(35) + (1/6)50 = 37.5$. He says C if $37.5 \geq 40 - X$, or $X \geq 2.5$.

In the third-to-last round, Jeff realizes that Akbar will say C not only if he is a nice guy, but also for other values of X (because Akbar's believes that Jeff might be a nice guy). Specifically, Jeff realizes that Akbar will say C with probability $3.5/6$. Jeff's payoff from S is $35 - Y$, which is at most 35. Jeff's payoff from C is at least what he would get if he says S in the final round, which is $(2.5/6)(30) + (3.5/6)(45 - Y)$. Therefore, he says C if

$$\begin{aligned}(2.5/6)(30) + (3.5/6)(45 - Y) &\geq 35 - Y\,,\\ Y &\geq -9\,.\end{aligned}$$

Since $Y \geq 0$, he always says C in this round.

In preceding rounds, each player says C because he knows the other is going to say C in the next round.

As we decrease the probability that the players are good guys, so that the game becomes increasingly one of complete information, the players still play C in the initial part of the game; however, this pattern of behavior breaks down more quickly. For example, suppose that X and Y are believed to be distributed uniformly on $[0, 5.02]$. Each player thinks the other is a nice guy with probability $0.02/5.02 \approx 0.004$. That is a pretty small grain of doubt that the other player might be a nice guy! Yet one can show that the players continue until the last three rounds. Then Jeff says C if $Y \geq 3.7$, after which Akbar says C if $X \geq 4.7$, after which Jeff says C if $Y \geq 5$. This is also how the game would be played if there were 100 or 1000 rounds. Thus, just a "grain of doubt" about each other's payoffs (or simply about each other's actions, perhaps because of doubts about whether the players are applying backward induction or are otherwise reasoning rationally) leads to C being played for most rounds, if the game is long enough.

In the finitely repeated Prisoners' Dilemma, there may be doubts about the rationality of the players or about whether the players have considerations such as feeling guilty about defecting when the other cooperates. We will observe the following equilibrium behavior. Once one player defects, both players defect thereafter because at least one is sure that the other is not a "nice guy". Each player cooperates in the beginning of the game, not only because she might be nice but also because (a) she wants to preserve her reputation as possibly being nice, and (b) she knows that the other player will do the same. Cooperation breaks down toward the end of the game.

In the chain-store game, there may be doubts about whether the chain is rational or about whether the chain actually has very low costs that make fighting a dominant strategy. The equilibrium is as follows. Once there is entry and the chain does not fight back, the chain is known not to have low costs and entrants come into the other markets. The chain, unless it has very high costs, fights any entry early in the game, in order to preserve its reputation as a chain that fights entry and may have low cost. Potential entrants realize that the chain is likely to fight, not merely because the chain might actually have low costs but also in order to preserve its reputation. Therefore, only the occasional low-cost entrant tests the situation by entering.

Wrap-up on reputation

From these examples, we have the following advice for playing long-horizon games. The intuition that comes from the infinite-horizon version is applicable when the horizon is finite. You should think about what kind of behavior would yield a high payoff if you could convince the other player that this is how you will behave. In the centipede game, you want to commit to always saying *C*; in the repeated Prisoners' Dilemma, you want to commit to cooperating as long as the other player cooperates (or playing a tit-for-tat strategy); in the chain-store game, you want to commit to always fighting entry. If the payoff from such commitment is high compared with the full-information equilibrium, adopt the behavior in the hope of convincing the other player that this is what you are going to do throughout the game. For example, in the centipede game, if you say *C* the first time but the other player follows with *S*, you lose €5 compared with just saying *S* the first time; but if the other player follows along and the game continues for several rounds, you can gain much more than that. So give *C* a try. In the repeated Prisoners' Dilemma, if you cooperate the first round but the other player defects, then you lose a little in that round compared to if you had also defected. However, if instead your strategy works and cooperation is sustained, your gain is higher. In the chain store, you lose some money by fighting several entrants, but you gain much more if you can thereby deter entry.

Chapter 17

Network Externalities

17.1 Motives and objectives

Broadly

Consider the following scenarios:

1. Sun Microsystems produces an operating system called Solaris. When a consumer decides whether to purchase it, he cares about how many other people use this operating system because being compatible with other people's system makes file sharing, communication, and technical support easier. The more people who also use the system, the more valuable it is to this consumer.
2. AquaBoulevard is a waterpark on the edge of Paris. When a consumer decides whether to go, he cares about how many other people are also going because the waterpark is less pleasant when it is crowded. The more people who also go to the waterpark, the less valuable it is to this consumer.

These goods do not fit our model of consumer demand because each consumer's valuation depends on how many other people buy the same good. We say that such a good has *network externalities*. Network externalities can be positive or negative:

Positive. The network externality is positive if the consumer's valuation is higher when other consumers use the good. Examples: the usefulness of a computer operating system such as Solaris or of a software package is higher when other consumers also use it, because of the benefit of sharing files and computers; the usefulness of trading in a market such as Ebay is higher when there are many other traders.

Negative. The network externality is negative if the consumer's valuation is lower when other consumers use the good. Examples: each consumer's valuation of waterpark such as AquaBoulevard or of a theme park is lower the more crowded the park.

Our focus in this chapter is on the implications of network externalities for the firms producing the goods, but the starting point is to understand the implications for consumers.

Externalities effect. Each potential consumer helps (if the externalities are positive) or hurts (if the externalities are negative) other consumers by using the product, but he only thinks about his own valuation when deciding whether to buy the product. As a consequence, equilibrium consumption is typically lower than socially optimal when the network externalities are positive; it is typically higher than socially optimal when the network externalities are negative.

Strategic effect. Consumers' decisions are interrelated. Each consumer's decision to buy depends on how many other consumers decide to buy. When the network externalities are positive, an increase in the number of consumers who buy or in the amount each consumer buys causes other consumers to buy the good or to buy more of it; hence, the purchase decisions are strategic complements. The opposite is true when the network externalities are negative.

Firms must understand a fundamental idea: because of the strategic effect, customers' purchasing decisions should be modeled as a game. Otherwise, firms may make the following mistakes:

1. with *positive* network externalities, firms may *underestimate* the elasticity of demand;
2. with *negative* network externalities, firms may *overestimate* the elasticity of demand.

Consider first an example of negative network externalities or congestion. Suppose AquaBoulevard wants to determine how much demand would rise if it lowered its admission price by 5% (say, one euro). It surveys a random sample of current and potential customers, asking the following questions: "How often do you come to the waterpark now? How often would you come if we lowered the price by one euro?" Each customer is likely to answer this last question just thinking about how much he or she currently values coming to AquaBoulevard, given the usual crowdedness of the water park. Based on the survey, AquaBoulevard might predict that demand would rise by 10%.

Suppose then that AquaBoulevard goes ahead with the price cut. The immediate *price* effect is that more customers come each day because of the reduced cost. However, as the park becomes more crowded, it also becomes less appealing to AquaBoulevard's customers; this *congestion* effect dampens demand. Since the price effect of the cut is positive (more demand) but the congestion effect is negative (less demand), the net effect is that customers come less frequently than what they reported in the survey. Demand might rise by only 6%.

Consider next an example of positive network externalities. Suppose that Sun Microsystems wants to determine how much demand would increase if it lowered the price of its Solaris operating systems by 5% (say, $25). To make the discussion simple, suppose that consumers pay a yearly license fee to have and use the latest version of Solaris. Sun surveys a random sample of its current and potential customers, asking the following questions: "How many licenses of Solaris do you currently lease? How many would you lease if the price fell by $25?" Each customer is likely to answer this last question just thinking about how much he or she values Solaris given, the number of other people who currently use it. Based on the survey, Sun might predict that demand would increase by 6%.

Suppose then that Sun goes ahead with the price cut. The immediate *price* effect is that some customers purchase more licenses because they are less expensive. However, as the Solaris user base increases, the operating system becomes more useful to the customers, which causes a further increase in demand that we call the *bandwagon* effect. The price and bandwagon effects of the price decrease are both positive; the total effect is that customers purchase more licenses that they said they would in the survey. Demand might go up by 10% rather than only 6%.

Network externalities affect industry concentration if they are specific to each brand's differentiated product. For example, a user's valuation of Solaris increases when other people also use Solaris but *not* when others use a different company's (incompatible) operating system. Such brand-specific positive network externalities can be a strong force toward industry concentration because customers flock to the same brand. If nearly every workstation is running Solaris, it is difficult for Hewlett-Packard or IBM to draw customers to their competing products. On the other hand, brand-specific congestion externalities reduce industry concentration because the congestion keeps customers from flocking to the same firm. If AquaBoulevard is very crowded, it is easy for another firm to set up a competing waterpark and draw customers away from AquaBoulevard.

More specifically

For the rest of this chapter, we focus entirely on positive network externalities. Compared to negative network externalities, they are more complicated and more important for managerial decisions because of their implications for dynamic pricing strategies and industry concentration. From now on, we will just say "network externality" with the understanding that we mean the case of positive network externalities. (This is consistent with common discourse. Anyone who says or writes "network externalities" without a qualifier is probably referring to positive network externalities. Negative network externalities are usually called "congestion".)

We first extend our model of consumer valuation and demand in order to incorporate network externalities. We then consider the implications of network externalities for pricing and competitive strategy. In the formal analysis, we consider only the case of a monopoly firm, but this analysis is still relevant to competing firms. In particular, our models will help us to understand the dynamic process by which a firm can position itself as an industry leader and then reap the rewards of this status.

17.2 A two-consumer example

Although we are interested in and will ultimately study markets with many consumers, we can illustrate a few basic ideas with a "toy" model that features only two consumers.

Nash equilibrium

Suppose there are two friends, Akbar and Jeff, who do not live together but might go to the same local movie theater each Friday. Akbar values a movie at €3 if he goes alone and at

€8 if Jeff also goes. Jeff likes Akbar's company equally well, but he places higher intrinsic value on movies. Hence, his valuation is €4 if he goes alone and €9 if Akbar also goes.

Let P be the price of a movie. We can model the two friends' decisions as a game in which each friend's actions are "Go" and "Stay home" and in which each player's payoff is his consumer surplus, as shown in Table 17.1(a). The payoff matrix and hence the Nash equilibria of this game depend on P.

1. Suppose $P < 4$. (Table 17.1(b) shows the game for $P = 3.5$.) Jeff's dominant strategy is to go to the movie. Given that Jeff goes, Akbar also wants to go. Hence, the unique Nash equilibrium is that both friends go.
2. Suppose $P > 8$. (Table 17.1(c) shows the game for $P = 8.5$.) Then Akbar's dominant strategy is to stay home. Given that Akbar stays home, Jeff also wants to stay home. Hence, the unique Nash equilibrium is that both friends stay home.
3. Suppose $4 < P < 8$. (Table 17.1(d) shows the game for $P = 6$.) Then there are two Nash equilibria: (a) both friends go and (b) both friends stay home. The first, "high demand", equilibrium is preferred by both friends to the other equilibrium.

Dynamic pricing strategies

If a game has several Nash equilibria, as happens in the example when $4 < P < 8$, basic game theory and the definition of Nash equilibrium provide no criteria for deciding which is more robust and probable. One typically must extend the model or look outside any formal model to decide this.

If Akbar and Jeff were literally the only two consumers, they could simply meet and agree to go to the movies, since they prefer this equilibrium to the one in which they stay home. If they cannot meet but they can observe each other's actions, then Jeff could choose to go, knowing that Akbar will then find it in his interest to follow.

However, we are really interested in markets with thousands or millions of consumers; we cannot presume that equilibria are selected by a grand meeting or that one consumer's decision will by itself greatly influence the rest of the market. We therefore turn to a dynamic interpretation of Nash equilibrium. We presume, as in this story, that the interaction is ongoing. Each Friday, each customer expects the number of other moviegoers to be the same as the previous week and then decides whether or not to go to the movies based on this expectation and based on the current ticket price. When the price is kept constant, customers stop adjusting their decisions when they reach a Nash equilibrium (that is, the Nash equilibria are the steady states of this dynamic adjustment process).

Which equilibrium is reached is tied to the historical circumstances. For example, suppose $P = 6$. If initially Akbar and Jeff are going to the movies, then they will continue to do so; if initially they both stay home, then they continue to stay home.

However, the movie theater can control this dynamic pricing through its pricing strategy. Assume that Akbar and Jeff are the only customers and that the cost of serving a customer is €5. The best outcome for the theater is to charge nearly €8 and have both friends go.

Table 17.1

a. General game form

		Akbar	
		Go	Stay home
Jeff	Go	$8 - P$ $9 - P$	0 $4 - P$
	Stay home	$3 - P$ 0	0 0

b. Game when $P = 3.5$

		Akbar	
		Go	Stay home
Jeff	Go	[✓ 4.5 5.5]	0 0.5
	Stay home	−0.5 0	0 0

c. Game when $P = 8.5$

		Akbar	
		Go	Stay home
Jeff	Go	−0.5 0.5	0 −4.5
	Stay home	−5.5 0	[✓ 0 0]

d. Game when $P = 6$

		Akbar	
		Go	Stay home
Jeff	Go	[✓ 2 3]	0 −2
	Stay home	−3 0	[✓ 0 0]

However, if initially the friends are not going to the theater, then neither one has an incentive to start going. The theater can try to break out of this vicious cycle via an advertising and publicity campaign meant to create self-fulfilling expectations ("buzz") that consumers will start going to the movie. However, such expectations are difficult to control. Instead, since the equilibria of this game depend on the price the movie theater charges, the theater can use price to move the behavior of the (here, two) consumers toward the high-demand equilibrium. All the theater needs to do is drop its price to below €4 for two weeks. The first week, Jeff attends because it is a dominant strategy to do so (this is the price effect). The second week Akbar, goes to the movies because he expects Jeff to do the same (this is the bandwagon effect). The third week, both friends expect the other to go to the movies, so the movie theater can raise its price to €8 and still attract both customers. Although the theater sells at below cost for a short period, it can recover this loss by long-term sales at €8.

Recap

Although the small number of customers in this example strains our credulity (a deficiency we will soon remedy), it successfully illustrates the following points.

1. With network externalities, the customers' purchasing decisions are a game.
2. There can be multiple Nash equilibria, of which the high-demand equilibrium is unanimously preferred by the customers and the firm.
3. A price cut has two effects on demand: first, it brings in extra demand simply because demand is price-sensitive (the price effect). Second, the extra demand increases the value of the product and hence increases demand further (the bandwagon effect).
4. To promote a new product or an old product that is stuck in a low-demand equilibrium, a firm may initially decrease its price to attract the first few customers. It can then raise its price after the network externality has increased customers' valuations.

17.3 Large markets: A linear example

We next consider an example with a large market. This example does not have multiple equilibria, but it does illustrate other properties of markets with network externalities.

The data

We consider a market with a fixed number of potential customers, each of whom might purchase 0 or 1 unit of the good. We let Q be the fraction of customers who purchase the good. That is, Q is our measure of demand.

Each customer's decision depends not only on the price but also on how many other customers purchase the product. To determine the equilibrium demand at each price, we need to know the following: if the price were P and each customer *expected* fraction Q^e of all potential customers to purchase the good, what fraction Q of the customers would actually purchase the good? We denote this data by the function $Q = d(P, Q^e)$. For example, suppose $d(P, Q^e) = 1 + \frac{1}{2}Q^e - P$. If $P = 3/4$ and each customer expects that $1/4$ of all customers will purchase the good, then actually $3/8$ of the customers do so. The middle term $\frac{1}{2}Q^e$ is the network effect.

If one does not know the function $d(P, Q^e)$, then one can estimate it by asking a random sample of customers such questions as "Would you buy the product if the price were 10 and you expected 3/4 of the other customers to buy the product also?" You may take a collection of such data points and then use them to estimate the parameters of a simple functional form for $d(P, Q^e)$.

Nash equilibrium

For our numerical example in this section, we assume $d(P, Q^e) = 1+\frac{1}{2}Q^e-P$. We just noted that if $P = 3/4$ and each customer expects $1/4$ of all customers to purchase the good, then actually $3/8$ of the customers do so. The customers are wrong in their expectations, so this is not an equilibrium. The market is in a Nash equilibrium when the actual demand equals the expected demand. For example, if $P = 3/4$ and everyone expects 1/2 of all customers to buy, then in fact 1/2 of the customers do purchase the good.

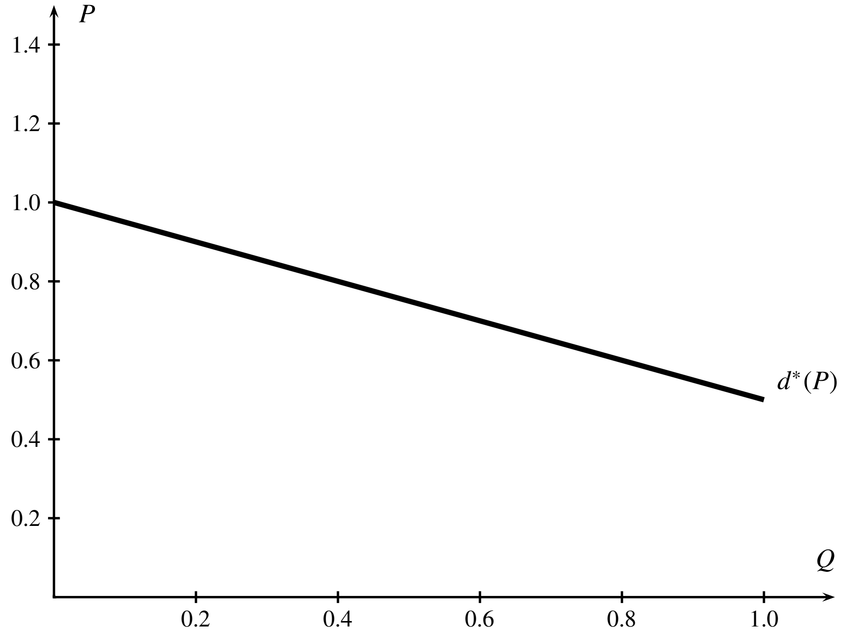
More generally, for a given price P, the demand Q^* is a Nash equilibrium if $Q^* = d(P, Q^*)$. We can solve this equation for Q^* as a function of P to determine the Nash equilibria for each price P:

$$\begin{aligned} Q^* &= 1 + \tfrac{1}{2}Q^* - P, \\ \tfrac{1}{2}Q^* &= 1 - P, \\ Q^* &= 2 - 2P. \end{aligned}$$

For example, if $P = 3/4$ then $Q^* = 1/2$, as we calculated previously. If the price goes down to $P = 2/3$, then the Nash equilibrium demand increases to $Q^* = 2/3$.

We call the curve $Q^* = 2 - 2P$ the *equilibrium demand curve* and denote it by $d^*(P) = 2 - 2P$. Figure 17.1 shows its graph, drawn with price on the vertical axis (as usual for demand curves).

Figure 17.1



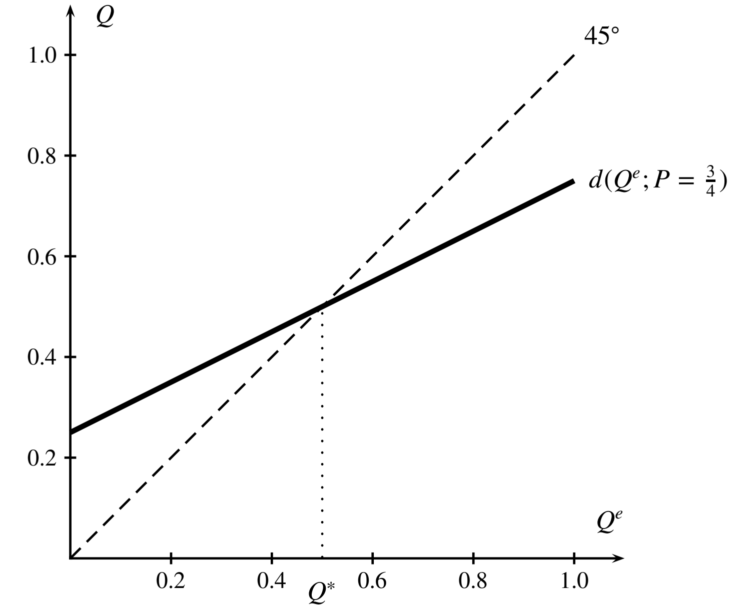
Unlike in our two-consumer example, there is only one equilibrium demand for each price and so the equilibrium demand curve $d^*(P)$ looks like a conventional demand curve. The pricing problem of a firm with market power is then no different from the one studied in Chapter 7 (or, if this market is actually competitive, the analysis is the same as in Chapter 5). What is new is the following: to estimate the demand curve $d^*(P)$ correctly, we must gather the data $d(P, Q^e)$ and then model demand at each price P as the equilibrium of a game, played by consumers, in which P is a parameter.

Illustrating Nash equilibrium using the aggregate reaction curve

Before moving to a more complicated scenario in which there are multiple equilibrium demands, it is worth illustrating further properties of this model.

In Chapters 12 and 13, we illustrated Nash equilibrium as the intersection of the two players' reaction curves. A related graphical illustration is possible here even though there are many consumers. Fix P and hence the game the consumers play. Let's denote the demand data by $d(Q^e; P)$ to emphasize that we are keeping P fixed and that our focus is on how actual demand depends on expected demand. We call the curve $Q = d(Q^e; P)$ an *aggregate reaction curve*. When $P = 3/4$, this curve is $Q = \frac{1}{4} + \frac{1}{2}Q^e$. It is shown in Figure 17.2.

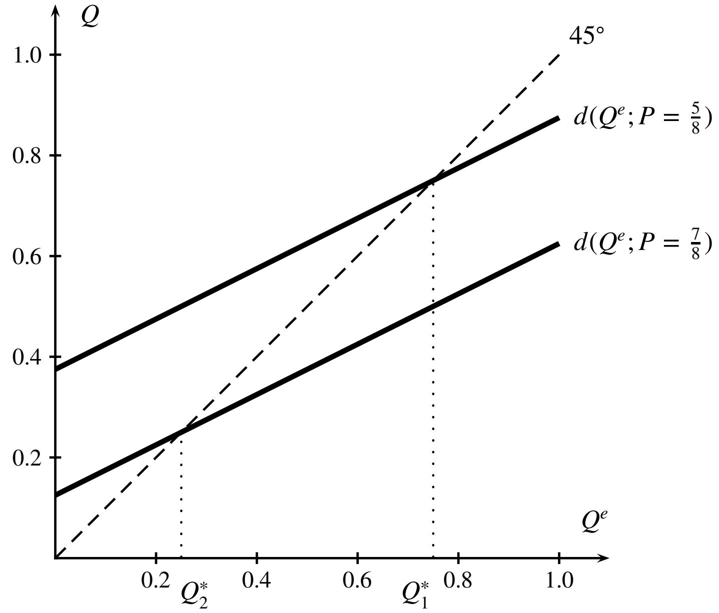
Figure 17.2



The 45° line is a visual aid for finding the equilibrium. This line is the set of points for which the equilibrium condition $Q = Q^e$ holds. Thus, a Nash equilibrium occurs where the aggregate reaction curve $d(Q^e; P)$ intersects the 45° line.

A change in the price will shift the aggregate reaction curve and hence the equilibrium. For example, Figure 17.3 shows $d(Q^e; P = \frac{5}{8})$ and $d(Q^e; P = \frac{7}{8})$.

Figure 17.3



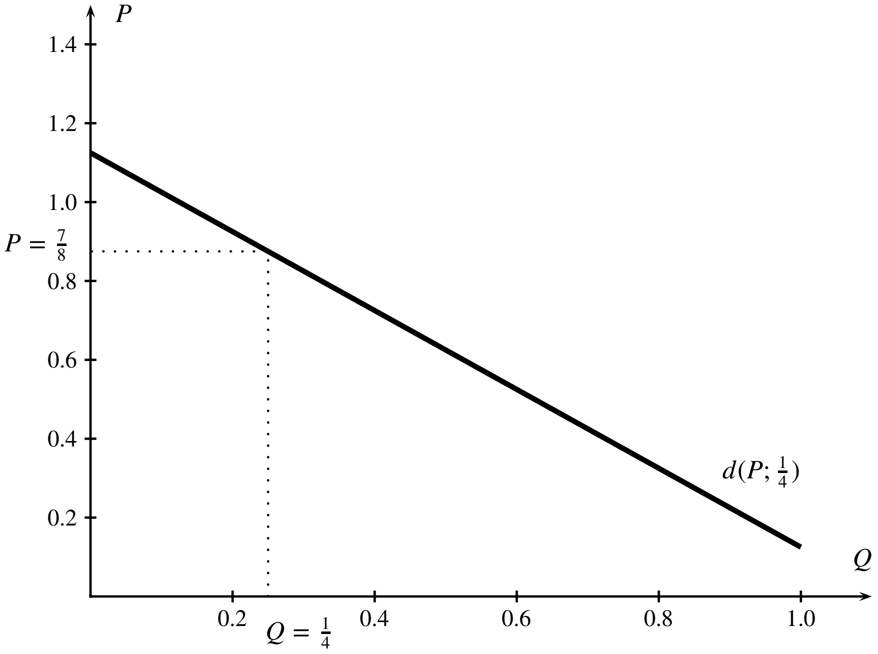
Observe that an increase in P (from $\frac{5}{8}$ to $\frac{7}{8}$) decreases actual demand for any given level of expected demand, and hence it shifts the aggregate reaction curve down. As a consequence, the equilibrium demand decreases (from Q_1^* to Q_2^*).

Price effects and bandwagon effects

Here is a second graphical perspective, in which we instead fix expectations Q^e and focus on how demand depends on price. For this purpose, we denote the demand data by $d(P; Q^e)$; we call this curve a *constant-expectations demand curve*.

For example, suppose we fix $Q^e = 1/4$. Then demand as a function of price is $Q = d(P; \frac{1}{4})$, or $Q = \frac{9}{8} - P$. This curve is shown in Figure 17.4, with price on the vertical axis.

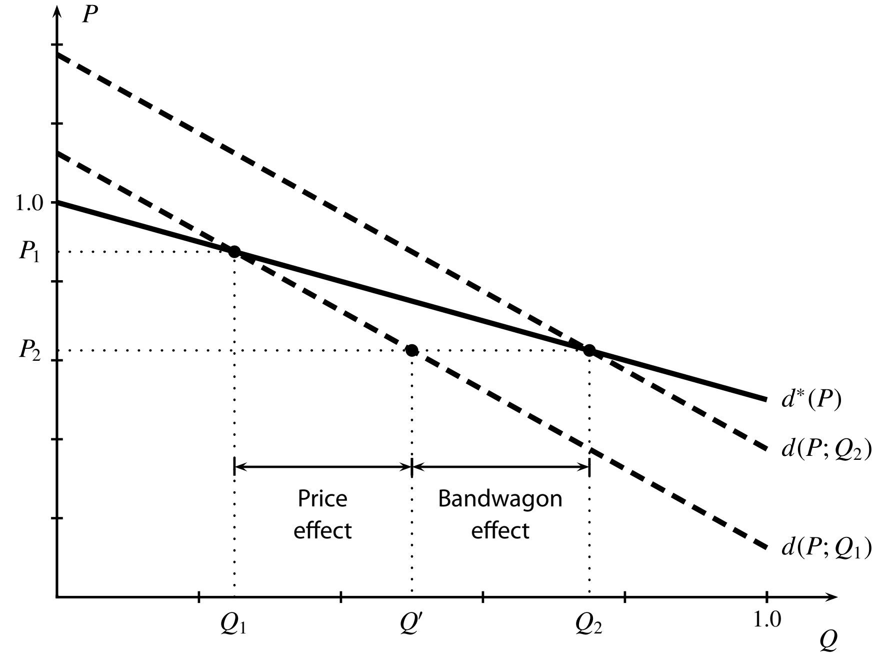
Figure 17.4



To find out the price at which $\frac{1}{4}$ is an equilibrium, we just need to look at the price that corresponds to $Q = \frac{1}{4}$ on the graph. Observe that this price is $P = \frac{7}{8}$.

We can use constant-expectations demand curves to illustrate the price effects and bandwagon effects of a change in price. Figure 17.5 shows the equilibrium demand curve along with the constant-expectations demand curves for $Q^e = Q_1 = 0.25$ and $Q^e = Q_2 = 0.75$.

Figure 17.5

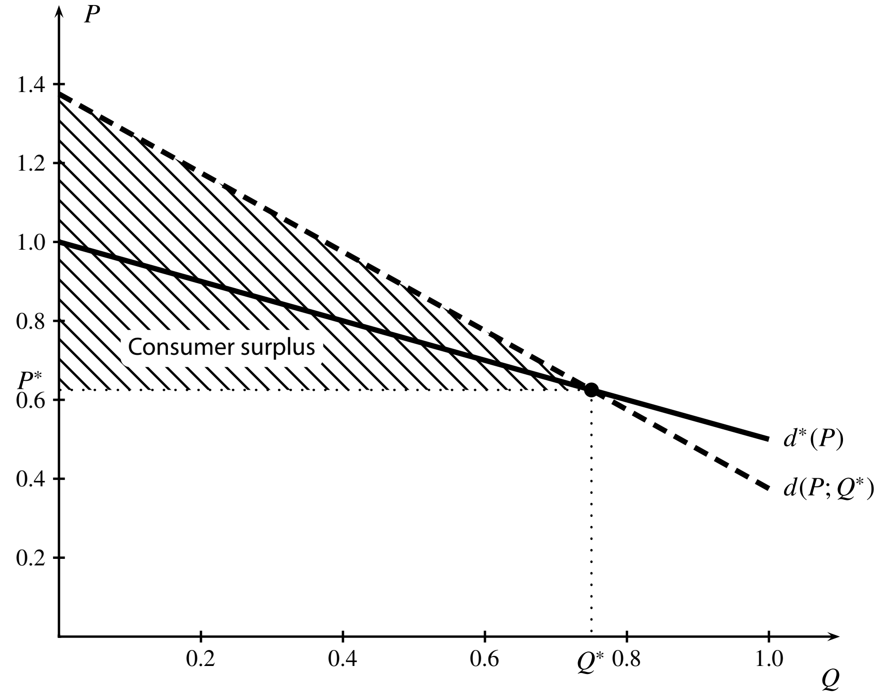


This graph illustrates the following. Suppose initially demand is Q_1 and the price is P_1. When the price falls to P_2, demand increases in part because, given consumers' current valuations of the good as determined by their current expectations $Q^e = Q_1$, more are willing to buy at the lower price. If this were the only story (i.e., if network externalities did not play a role), then the market would move along the constant-expectations demand curve $d(P; Q_1)$ and demand would increase to Q'. This increase in demand, $Q' - Q_1$, is the *price effect* on demand. However, via the network externality, the extra demand increases the valuations of all customers and thus brings of them into the market; the influx increases the valuations further and attracts yet more customers, and so on until a new equilibrium demand is reached at Q_2. The additional increase in demand, $Q_2 - Q'$, is called the *bandwagon effect*. Observe that, owing to this bandwagon effect, equilibrium demand is more elastic than the constant-expectations demand.

Consumer surplus

We can also use constant-expectations demand curves to illustrate consumer surplus. Figure 17.6 shows the equilibrium demand curve along with the constant-expectations demand curve for $Q^e = Q^* = 0.75$.

Figure 17.6



Suppose that the price is P^* and hence equilibrium demand is Q^*. For each consumer who buys the good, his surplus is his valuation *given that total demand is* Q^* minus the price P^*. If we graph these valuations from highest to lowest in order to form the marginal valuation curve, we get the inverse of the constant-expectations demand curve rather than of the equilibrium demand curve. Here is why. Suppose, for concreteness, that there are 100 consumers and so $Q^* = 0.75$ means that equilibrium demand is 75. What is the 30th-highest valuation given that 75 people buy the good? It is given by the inverse of the constant-expectations demand curve at $Q = 0.30$ because that is the price at which 30 people are willing to buy the good when they expect 75 people to buy the good. In contrast, the inverse of the equilibrium demand curve at $Q = 0.30$ gives the price at which 30 people are willing to buy the good when they expect 30 people to buy the good ($Q^e = 0.30$); this is the 30th highest valuation given that 30 people buy the good. Hence, total consumer valuation is the area under the constant-expectations demand curve $d(P; Q^*)$ up to Q^*; consumer surplus is the area between this constant-expectations demand curve and the line at P^*. This is shown in Figure 17.6; observe that the consumer surplus correctly calculated is larger than when calculated using the equilibrium demand curve.

Dynamic adjustment

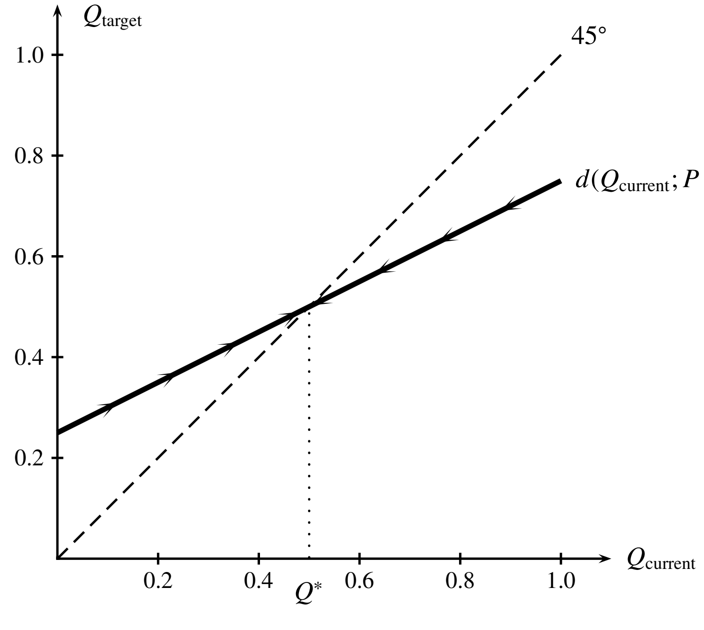
In the two-consumer example of Section 17.2, we focused on the interpretation of Nash equilibrium as a steady state in a game that is played repeatedly, and we considered how demand adjusts toward equilibrium and how a firm's dynamic pricing strategy can control this adjustment. Now that we have a model with many consumers, we can explore these topics with greater realism and richness.

In the dynamic model, the consumers' expectations about demand are formed by observing current demand. To emphasize this, we use the symbol Q_{current} to replace Q^e. If current demand is Q_{current}, then the market is in equilibrium if $d(P, Q_{\text{current}}) = Q_{\text{current}}$.

We postulate that, when the market is not in equilibrium, demand adjusts gradually because of inertia in consumers' purchasing decisions. That is, if $d(P, Q_{\text{current}}) \neq Q_{\text{current}}$, then demand adjusts only gradually toward $d(P, Q_{\text{current}})$ rather than instantaneously. We can think of $d(P, Q_{\text{current}})$ as the consumers' target demand (the demand the consumers would like, given the price P and given the network externality provided by the current level of demand); to emphasize this we use the symbol Q_{target}.

Figure 17.7 illustrates this adjustment for fixed P. The graph shows the aggregate reaction curve $d(Q_{\text{current}}; P)$.

Figure 17.7



Observe the following.

1. Where the graph of $d(Q_{\text{current}}; P)$ lies above the 45° line, $Q_{\text{target}} > Q_{\text{current}}$. Therefore, demand increases.
2. Where the graph of $d(Q_{\text{current}}; P)$ lies below the 45° line, $Q_{\text{target}} < Q_{\text{current}}$. Therefore, demand decreases.

In either case, demand moves toward Q^*, so we say that Q^* is *dynamically stable*.

Recap

In this section, we have studied an example with many consumers, in which there is a unique equilibrium demand at each price. This example has illustrated the following points, which remain valid when there are multiple equilibrium demands.

1. To estimate demand correctly, we must gather the data $Q = d(P, Q^e)$ and then model demand at each price P as the equilibrium of a game played by consumers.
2. A price cut increases demand for the conventional reason, but there is a further bandwagon effect as additional consumers find the good more valuable when they see others using it. Demand is more elastic when we take into account this network effect than when we do not.
3. Consumers' marginal valuation is measured by the inverse of the constant-expectations demand curve. Hence, consumer surplus is larger than if mistakenly measured by the area between the equilibrium demand curve and the price line.

We also explored dynamic adjustment toward equilibrium. In this linear example, there is a unique equilibrium and demand adjusts toward it, regardless of the initial value.

17.4 A nonlinear example with multiple equilibria

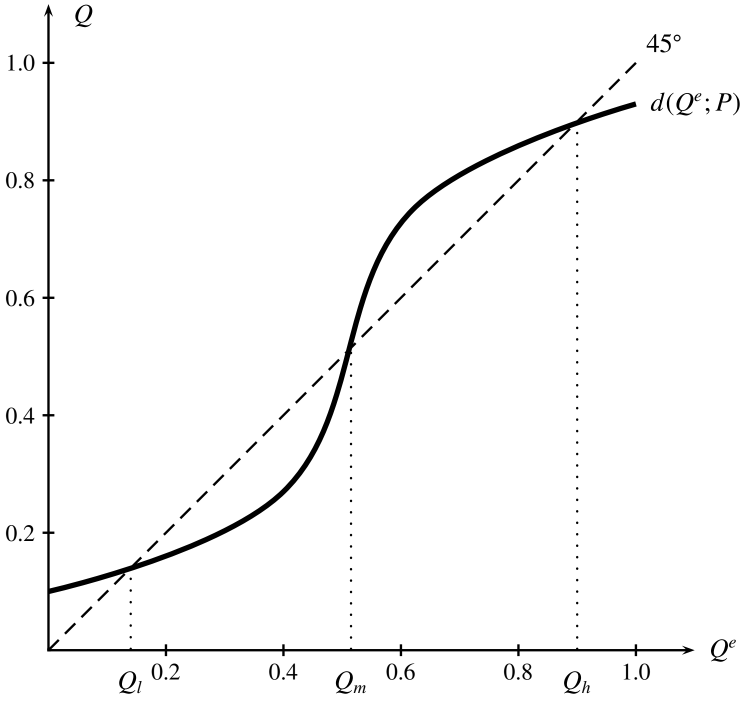
We continue with the basic model of a large market as presented in Section 17.3. The demand data is again summarized by $Q = d(P, Q^e)$. For a fixed price P, we can graph the aggregate reaction curve $Q = d(Q^e; P)$, which shows actual demand as a function of expected demand. For a fixed expected demand Q^e, we can graph the constant-expectations demand curve $Q = d(P; Q^e)$, which shows actual demand as a function of price when expectations are fixed.

The innovation of this section is that we study an example in which the demand data are not described by a linear function. Our goal is to understand general qualitative properties of such a situation rather than the numerical calculations, so the latter are not shown.

Aggregate reaction curve and equilibrium

Figure 17.8 shows a possible nonlinear aggregate reaction curve.

Figure 17.8



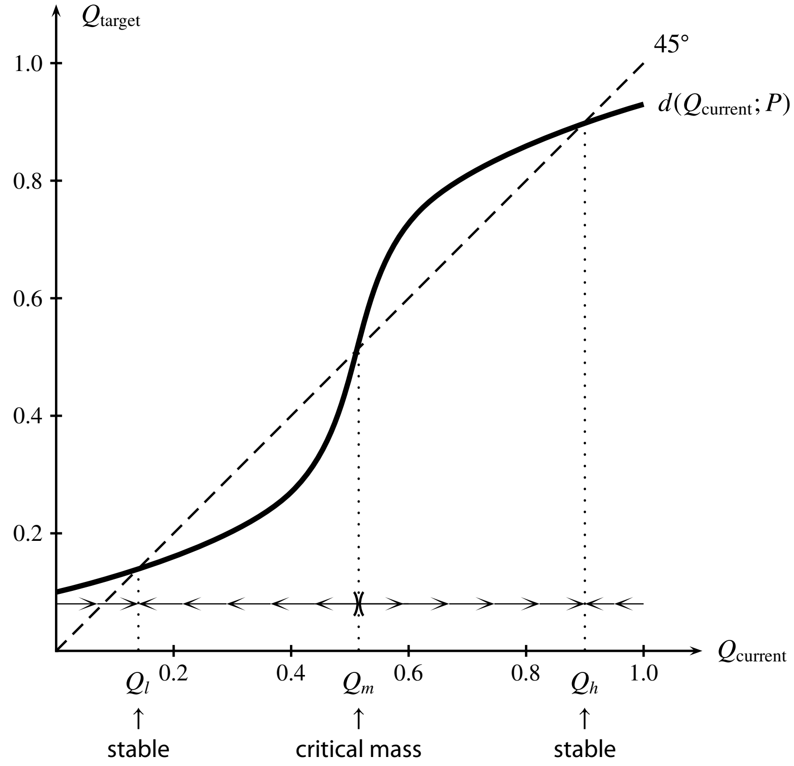
Its qualitative properties have a plausible interpretation. The curve starts above 0 (i.e., $d(0; P) > 0$) because there are some people who are willing to purchase the good even if no one else does. Initially, actual demand is not very responsive to the expected demand, indicating that the network externality does not become important until there are a significant number of users. Then, after 40% of the potential customers have adopted the good, the value of using the good increases quickly with the number of users, so the aggregate reaction curve is steep. Once many users have adopted, the remaining non-adopters are not very responsive to the network externality, so the aggregate reaction curve is again flat. The curve ends up below 1 (i.e., $d(1; P) < 1$) because there are some users who are not going to buy the good even if everybody else does.

We see that there are three equilibria: Q_l, Q_m, Q_h. Consumers prefer a higher equilibrium demand over a lower one. On the one hand, the consumers who purchase in the low-demand equilibrium are better off in the high-demand equilibrium because of the network externality (which leads to higher valuations and hence higher surplus in the high-demand equilibrium); on the other hand, there are other consumers who do not buy in the low-demand equilibrium and hence get zero surplus but who do buy in the high-demand equilibrium and hence get positive surplus.

Dynamic adjustment toward equilibrium

Figure 17.9 shows the same aggregate reaction curve as Figure 17.8, but with Q^e labeled as Q_{current} and Q labeled as Q_{target} to emphasize the dynamic process.

Figure 17.9



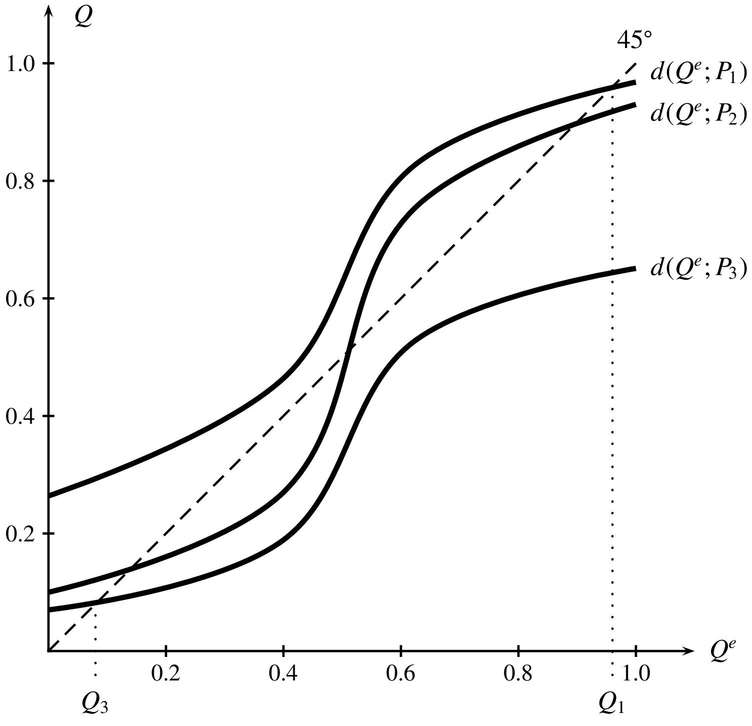
Recall that demand is falling when the graph of $d(Q_{\text{current}}; P)$ is below the 45° line, whereas demand is rising when the graph of $d(Q_{\text{current}}; P)$ is above the 45° line. Arrows in the graph show the direction in which this dynamic adjustment moves Q_{current}, depending on its current value. We see that the equilibria Q_l and Q_h are stable: when demand is initially close enough to either of these values, it moves toward it until equilibrium is reached. However, the equilibrium Q_m is unstable: any small movement away from it causes demand to move farther away until it reaches one of the other equilibria. Therefore, although Q_m is a Nash equilibrium, it requires an impossible balancing act that we cannot expect to occur. The value Q_m is of interest mainly because it divides the region in which demand converges toward Q_l from the region in which demand converges toward Q_h. We call Q_m the *critical mass* for the high-demand equilibrium; once demand is higher than Q_m, it converges naturally to Q_h.

The equilibrium demand correspondence

In this example, there can be multiple equilibria for any given price. Hence, we can no longer speak of an equilibrium demand curve (which specifies the unique equilibrium demand for each price). Instead, we have an *equilibrium demand correspondence*, which shows the set of equilibrium demands for each price.

Consider what relationship between price and equilibria would be plausible in this example. Figure 17.10 shows the same aggregate reaction curve that appeared in Figure 17.8 (the price is denoted P_2) along with aggregate reaction curves for a lower price P_1 and a higher price P_3.

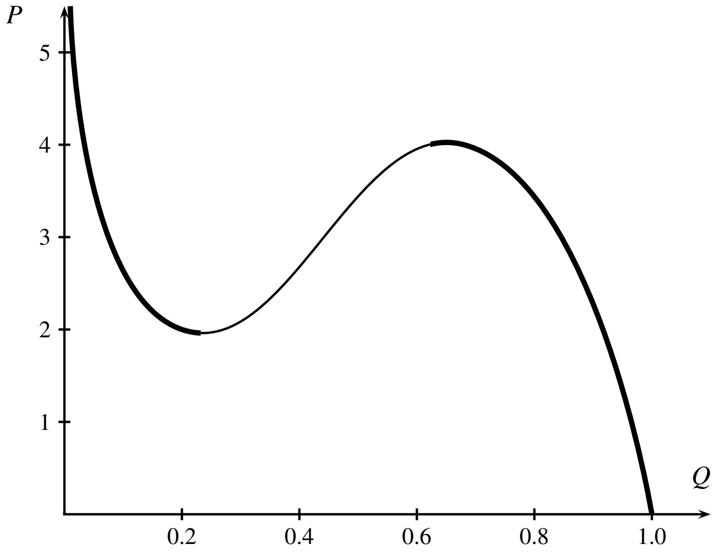
Figure 17.10



Observe that at the higher price P_3, there is a unique equilibrium; it has low demand (Q_3) and is stable. At the lower price P_1, there is also a unique equilibrium; it has high demand (Q_1) and is stable. The middle price P_2 is the one with two stable equilibria separated by an unstable equilibrium or critical-mass value.

The equilibrium demand correspondence could thus be as shown in Figure 17.11.

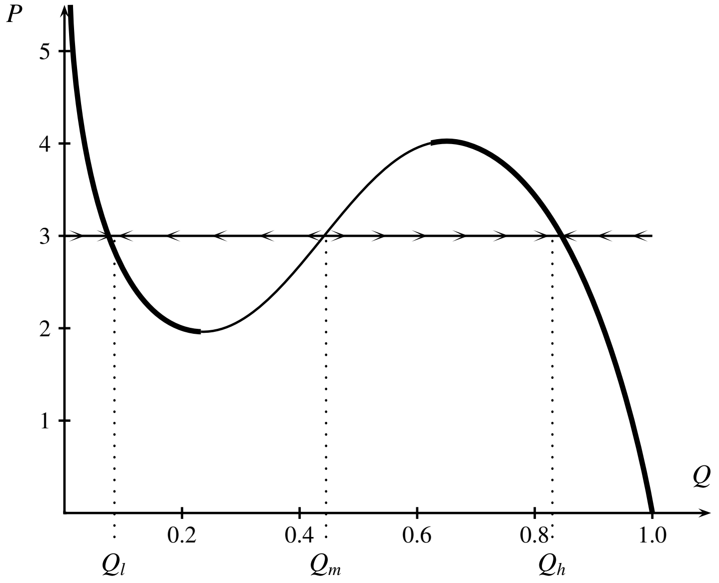
Figure 17.11



The thin portion of the curve denotes the unstable equilibria or critical-mass values. For high prices, there is a single low-demand equilibrium that is stable. For low prices, there is a single high-demand equilibrium that is stable. For intermediate prices, there are three equilibria: a stable low-demand equilibrium; a stable high-demand equilibrium; and an unstable equilibrium in the middle, which marks the critical mass needed for demand to converge to the high-demand equilibrium.

Figure 17.12 shows the same demand correspondence together with the dynamic adjustment direction when $P = 3$.

Figure 17.12



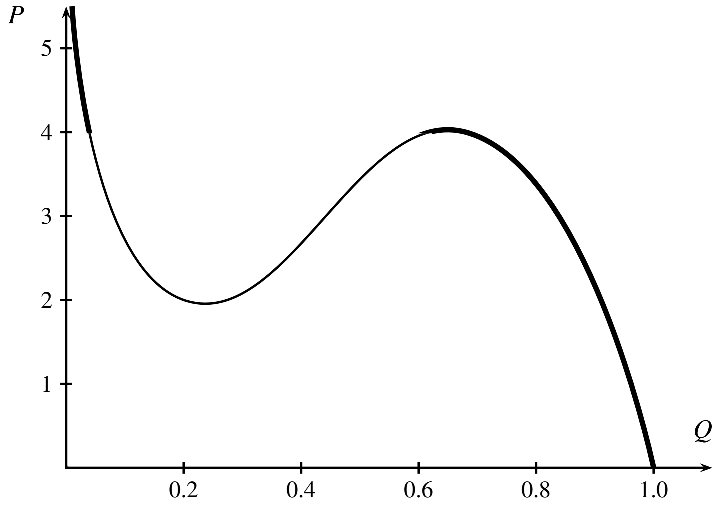
When demand is initially below Q_m, it moves toward the low-demand equilibrium Q_l; when demand is initially above Q_m, it moves toward the high-demand equilibrium Q_h.

Pricing when the firm can control consumer expectations

If a firm has determined that its equilibrium demand correspondence is as shown in Figure 17.11, how should it set the price of its product?

The first way to approach this problem is to assume that the firm can always induce the consumers to coordinate on the high-demand equilibrium (because this is what the consumers want). Then the firm's demand curve is the thicker part shown in Figure 17.13.

Figure 17.13



This demand curve slopes downward even though it has a discontinuous jump. The implication of this jump for the pricing problem is that marginal conditions are not sufficient. There may be a price above 4 that is locally optimal (because the demand curve there is quite inelastic), yet lowering the price to below 4 yields the firm a huge jump in demand that leads to higher profit.

Consumer surplus

It is worth illustrating the dramatic surplus that consumers may obtain but that the firm cannot extract even if it can coordinate consumer expectations around the high-demand equilibria. Suppose that the firm's profit-maximizing price and demand are P^* and Q^*, as shown in Figure 17.14.

Figure 17.14

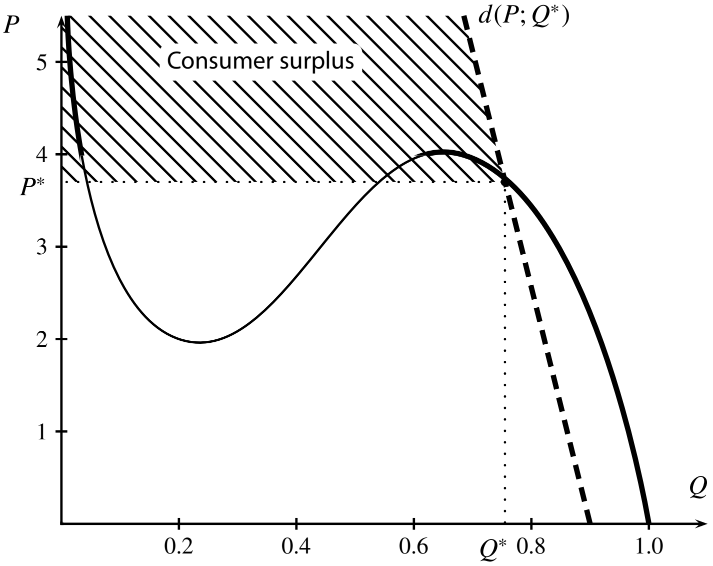


Figure 17.14 also shows the constant-expectations demand curve given Q^* and the consumer surplus. Consumers place high value on being in the market because there are so many other consumers in the market. However, if the firm tries to extract this surplus by increasing the price, it faces the following problem. As the first few customers get off the bandwagon in response to the price change, others follow because of the (reverse) bandwagon effect and there is a significant drop in demand. In fact, if the price rises to above 4 then demand collapses.

Dynamic pricing strategies

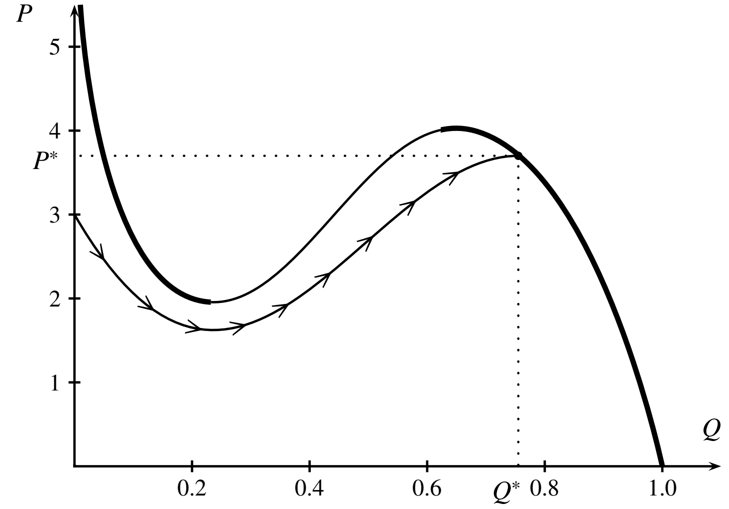
Suppose that the firm has just entered the market and hence starts with low demand, even though its ultimate objective is to reach the point (Q^*, P^*) shown in Figure 17.14. How can it get there?

It can attempt to use advertising and fanfare to create a "buzz" for the product, so that consumers' expectations that others will buy induces them to buy also. Such strategies are important but should not be relied upon. Instead, by initially setting the price low, the firm can bring consumers into the market until a critical mass is reached, after which it can increase the price.

This is a dynamic control problem. We would need information about the adjustment speed of demand and the interest rate at which the firm discounts its profit flow in order to calculate the exact solution. However, the following trade-off is involved. The greater is the discrepancy between $d(P, Q_{\text{current}})$ and Q_{current}, the faster demand adjusts. If the price is set very low at the beginning, then demand adjusts quickly and the firm soon can raise the price to P^*. However, the firm has given away product at a very low price; this is especially problematic if the good is durable and consumers who purchase at the low price may not purchase again for some time. Alternatively, the firm can price higher at the beginning, but then demand adjusts more slowly and so it takes longer before the firm can raise its price to P^*.

Figure 17.15 gives an example of a dynamic pricing strategy that is fairly conservative; adjustment is slow but the firm keeps the price relatively high throughout.

Figure 17.15



This path represents the price and demand combinations that evolve over time,[1] as controlled by the firm. The firm sets a fairly high price initially, but demand increases as it is pulled toward the low-demand equilibrium. However, the firm does not want to get stuck at such an equilibrium, so it gradually lowers the price. When the price is below 2, demand is pulled toward the high-demand equilibrium. However, the firm does not merely stick to such a low price and wait for this adjustment to happen. Instead, as demand increases it adjusts its price upward, taking care that the price-demand combination is always to the right of the thin line showing the critical-mass points. As long as this is true, demand continues increase toward the high-demand equilibrium.

[1] In no way should this curve be called a “demand curve”.

Recap

This example with multiple equilibria has illustrated the following points beyond those of Section 17.3.

1. There can be multiple equilibria, some of which are dynamically unstable.
2. There can be a critical mass (corresponding to an unstable Nash equilibrium), above which demand increases naturally to a high-demand equilibrium and below which demand decreases to a low-demand equilibrium.
3. A firm can use pricing strategies over time to coax demand to the target equilibrium. It may want to offer an initial discount in order to build demand up to the critical mass of its target equilibrium and then recover the lost profit when the target is reached.

Exercise 17.1. Consider a market with network externalities. Suppose that actual demand (measured as a fraction of the total number of consumers) as a function of price P and expected demand Q^e can be written $d(P, Q^e) = 1 - P + 2Q^e$, with the following caveat: when this function exceeds 1, demand is 1; when this function is negative, demand is 0.

a. Graph the aggregate reaction curve when $P = 3/2$. Identify the Nash equilibria and identify which of them are stable.

b. If you are introducing this good for the first time and so the demand is initially 0, outline a plausible dynamic pricing strategy.

17.5 Wrap-up

We have studied how to model demand for a single firm's product when that product has network externalities. We considered the dynamic adjustment toward equilibrium and how a firm should control that process through dynamic pricing strategies.

Without introducing further formal methods, consider what intuition these models can provide about strategies when multiple firms are competing for customers and each firm's product has network externalities. We have in mind, for example, competition between software packages: each consumer would buy only one of the packages and places higher value (controlling for quality and individual preferences) on the package that garners more users.

Consider the case of two firms. Let's start with the equilibrium. An outcome in which the two firms share the market is not stable. Even if the firms kept their prices fixed, shifts by a few consumers from one firm to the other would cause other consumers to follow, until one firm dominates the market. The interesting question is: What dynamic process determines which firm takes over the market?

Suppose the firms enter the market sequentially. The market leader should follow a strategy like the ones outlined in this chapter—with the caveat that speed is important because market penetration should be achieved before the second firm enters.

Now consider the entrant's strategy, assuming first that its product is not superior to the market leader's product. If the market leader is charging its profit-maximizing monopoly price and if (hypothetically) it did not change this price after the entrant came into the market, then the entrant could take over the market by following a dynamic pricing strategy like the one outlined previously. However, the market leader should (and surely will) react, lowering its own price to retain its market share. A rule of thumb for the outcome of such a price war—in which ultimately only one firm can remain in the market—is that the winner is the firm that has some cost or pricing advantage. In this case, if the incumbent tries simply to retain market share then it can do so at a price that is much higher than the entrant's price, owing to network externalities. Therefore, the entrant would lose the price war and would only incur losses in the process. Entering is a bad idea.

If the entrant's product is sufficiently superior to the market leader's product, then it can displace the leader from the market. A rule of thumb is that the entrant can displace the leader if consumers would shift toward the entrant's product (in spite of the network externality) whenever the two products are offered with the same markup over marginal cost. This rule of thumb is conservative; entry may succeed under less favorable conditions.

If two firms enter the market at exactly the same time, they will price low to become the market leader. The ultimate winner will be determined by chance events. The equilibrium of such a game is that the firms end up dissipating (as initial losses) the expected profit to be earned as monopolist—the prize of the game. Hence, although ultimately one of the firms will win and may earn a large monopoly profit in the long run, each firm's overall expected profit is zero.

Thus, markets with such network externalities are ones in which there is a huge gain to being first. There are no better examples than Microsoft and Intel, which continue to earn enormous monopoly profits and have had little trouble fending off entry from competing operating systems and microprocessors.