APPENDIX

Introduction to Basic Economics Concepts

This appendix serves as a very brief overview of some of the main economics concepts used throughout this book. If the reader has had an introductory or intermediate economics course before this (and the book aims at such a student), this material should serve as a quick reminder of the basic concepts. For the student who has never encountered economic thinking before, please remember that entire books and one- to two-semester courses usually cover the material that follows, and judge the adequacy of the coverage of this material found here with mercy.

THE CONCEPT OF UTILITY AND DEMAND CURVES

Economists begin their study of human behavior by making what some would consider a rash assumption, namely that every consumer has a stable set of preferences that allows comparison of different “bundles” of goods (taken here very broadly to mean goods and services produced in the market or activities that might take place out of the market). These preferences are normally represented as a “utility function” showing how additional amounts of each good increase total “well-being” or (in the usual term) “utility.” Economists normally presume that “goods” continue to add utility as more are consumed, so sometimes the “good” becomes the removal of a noxious substance. For example, garbage removal may be the good rather than garbage. The list of possible goods is very large. Think about how many bar codes exist for all the stores in the world; each one describes a different good and this still
misses items such as “watching a sunset” for which no bar code (yet) exists. To simplify the discussion, we usually abstract from the specific names of these goods and simply call them $X_1, X_2, \ldots$ etc. Thus, the utility function becomes $\text{Utility} = U(X_1, X_2, \ldots X_N)$ where $N$ is the number of goods that one might possibly wish to consider. For reasons of simplicity—including our ability to diagram these concepts readily—economists often collapse the discussion to two goods (which can then be pictured on a two-dimensional graph, as you will see shortly). This actually does not cause a great loss of precision in thinking because these two goods can be one specific good (such as popcorn or peanuts) and the other a “composite bundle” of all other goods and services.

Economic thinking rests fundamentally on the premise that consumers make choices to maximize their utility, limited by the amounts of goods that the consumer’s income can purchase. In more elaborate versions of this problem, income itself represents choices made by the consumer in the current period (how much to work versus how much to enjoy leisure) as well as the cumulative effects of many past decisions (how much education to seek, how hard to study in classes of health economics, how much to save for the future). We assume that consumers not only know their preferences but also how to act to maximize their own utility. Although few economists believe that this literally occurs in every human choice, a very fruitful study of human behavior emerges from this simple yet powerful paradigm.1

These types of decisions can be portrayed easily either in graphs or by using the branch of mathematics known as calculus. The basic “problem” of the consumer is to maximize utility subject to a budget constraint,2 and this can be easily portrayed (in a simplified version) in a two-dimensional graph. In this graph, we focus on a specific good (say, $X_1$) and choices between that good and all other goods. Figure A.1 shows two key features about how goods affect utility. First, for any specified amount of $X_{\text{other}}$, adding more $X_1$ adds to total utility but at a decreasing rate. This conforms to the simple idea that we eventually become “sated” in our consumption of a specific good. The formal statement of this says that we have positive but diminishing marginal (or incremental) utility for any specific good such as $X_1$. (The same is said

1A classic discussion of these issues is found in Milton Friedman’s Essays in Positive Economics (1966).

2This type of problem commonly uses an approach devised by mathematician Joseph-Louis Lagrange, known (not by coincidence) as Lagrange multipliers. Lagrange is usually considered to be French, but he was born in Turin, Italy, and wrote his famous essay on finding maxima and minima in 1755 in Italian at age 19, probably younger than almost every reader of this textbook! (What had you done by age 19?)
symmetrically for \( X_{\text{other}} \) because it represents a bundle of all other goods and services. Indeed, any one of those could be pulled out and treated similarly to \( X_1 \). The other key feature of Figure A.1(a) shows that similarly shaped curves exist for each possible level of \( X_{\text{other}} \), but utility is lower as the amount of \( X_{\text{other}} \) falls so that these curves “nest” one under another as the amount of \( X_{\text{other}} \) falls.

Now comes an important step: We take the information from the curves in Figure A.1(a) and transform the same information into a different graph, one using \( X_1 \) and \( X_{\text{other}} \) on the axes. (This is why we want to work only with two goods. Otherwise this becomes an \( n \)-dimensional graph that’s difficult if not impossible to draw for \( n \geq 3 \).) To see how this works, pick a specific level of utility in Figure A.1(a) (say, \( U_1 \)), and think about the combinations of \( X_1 \) and \( X_{\text{other}} \) that could create exactly \( U_1 \). Figure A.1(a) shows three possible levels of \( X_{\text{other}} \), arbitrarily designated as 8, 9, and 10 (but they could be anything because we haven’t said in what units \( X_{\text{other}} \) is measured). Points A, B, and C all have exactly the same level of utility \( (U_1) \), but each has a different combination of \( X_1 \) and \( X_{\text{other}} \). The graph of those points (and all others like them if we’d drawn more curves in Figure A.1(a) for different amounts of \( X_{\text{other}} \)) creates an isoutility curve at utility level \( U_1 \) in Figure A.1(b). More commonly, economists call these indifference curves because the consumer is literally indifferent to which of any combination of goods on each curve is

\[ X_1^a, X_1^b, X_1^c \]

\[ X_1^d, X_1^e, X_1^f \]

\[ X_{\text{other}}^a, X_{\text{other}}^b, X_{\text{other}}^c \]

\[ X_{\text{other}}^d, X_{\text{other}}^e, X_{\text{other}}^f \]

\[ X_1, X_{\text{other}} \]

\[ U_1, U_2 \]

\[^3\text{These come from the Greek word } \textit{isos} \text{ meaning “equal.”}\]
best to consume. Points D, E, and F from Figure A.1(a) translate into a separate indifference curve $U_2$ in Figure A.1(b). We could create an extremely large number of such curves in Figure A.1(b) (an infinite number of them, actually) by picking specific levels of utility in Figure A.1(a) and graphing the relevant points over to Figure A.1(b). We usually draw only a few of such curves to symbolize the overall idea, and (as you will see in a moment), only one of them eventually matters—the highest one achievable with the budget available to the consumer.

These curves are really no different than many maps you’ve often read. Weather maps, for example, either with curves or colors, show areas of the country with the same temperature. Meteorologists call these isothermal maps, meaning “same temperature.” Topographical maps show lines of constant altitude, and anybody who’s ever used these maps soon becomes able to look at them and figure out what the terrain that the map depicts looks like.

The next step allows us to move from the ephemeral world of utility (which we can’t really see or measure) to the concrete world of prices and quantities. Begin with (and slightly generalize) the graph in Figure A.1(b) and then think about how to add information about the consumer’s buying power, the “budget” available to buy goods and services. In the simplest model, the consumer’s income ($I$) is fixed at a specific level (say, $I_1$). In Figure A.2(a), this income appears as a downward-sloping straight line with a slope determined by the relative prices of $X_1$ and $X_{\text{other}}$. The trick is to find the highest possible level of utility (remember that the consumer is trying to maximize utility within the constraint of the budget available), and it’s easy

---

4 A personal note: The word *indifferent* has a specific meaning to economists not universally shared by others. It refers to combinations of things that create the same utility. Early in my marriage, my wife would sometimes ask me, “Do you want peas or corn with the meat loaf tonight?” Because I’d just learned the concept of indifference in an economics class, I would respond “I’m indifferent,” meaning “either would be fine.” Before I explained this concept clearly, she heard instead “I don’t want either” as my message, and this came close to ruining our marriage until I explained the underlying economic idea to her. I can now say “I’m indifferent” without having to call a marriage counselor.

5 To show why the budget line is a straight and downward-sloping line in this graph, begin with the notion that all available income is spent on goods $X_1$ and $X_{\text{other}}$ purchased at prices of $p_1$ (for $X_1$) and (arbitrarily) a price of 1 for other goods. Thus $I = p_1X_1 + X_{\text{other}}$. Now we ask what combinations of goods could exactly consume the total budget $I$. To do this, increase $X_1$ by a little bit and ask how much $X_{\text{other}}$ has to fall to keep within the budget. Adding one unit of $X_1$ causes spending to increase by $p_1$, so the amount of $X_{\text{other}}$ has to fall by exactly $1/p_1$ to keep total spending the same. The constant-budget line in the graph with $X_1$ and $X_{\text{other}}$ on the axes thus has a slope of $-1/p_1$. More generally, in a graph of $X_1$ versus $X_2$, the budget line will have a slope of $-p_{12}/p_1$. 

to see that one does this by finding the highest possible indifference curve that just touches the budget constraint at one point. This is called a tangent point, and it’s easy to show that the consumer can’t do any better than to pick the bundle of goods where the budget line is exactly tangent to one of the indifference curves. (To see this, think about trying to achieve any higher level of utility; it would require using more income than is in the budget. Going in the other direction, any feasible point on the budget constraint except the tangent must lie on a lower indifference curve, and, hence, must create lower utility.)

Figure A.2(a), on the following page, shows two incomes ($I_1$ and $I_2$) that were picked because they are just tangent to the indifference curves $U_1$ and $U_2$. The optimal consumption for the consumer with these tastes (that is to say, with this utility function) and a budget of $I_1$ is the combination $(X_1^*, X_{\text{other}}^*)$. If the consumer’s income were to rise to $I_2$, the optimal combination would shift to $(X_1^{**}, X_{\text{other}}^{**})$.

Now comes the final step of creating demand curves from the indifference curve map. The “thought experiment” here is to ask what happens if income remains the same and the price of $X_{\text{other}}$ remains at 1, but the price $p_1$ goes up or down. Because the slope of the income line is the ratio $1/p_1$, raising $p_1$ causes the budget line to flatten out and vice versa. Figure A.2(b) shows three different prices for $p_1$, labeled $p_1^*, p_1^{**}$, and $p_1^{***}$, but it should be clear that we can pick as many different prices as we wish and just add complexity to the graph.

When $p_1$ is lowest among these three values (at $p_1^*$), the buying power is highest, so the top of these three income lines is the one that matters. The optimal choice leads to consumption of the pair $(X_1^*, X_{\text{other}}^*)$ in Figure A.2(a). If the price goes up to $p_1^{**}$, the budget line rotates downward around the lower right intersection (it must flatten out, as described before, but if all the budget were spent on $X_{\text{other}}$, the same amount would be available no matter what the price $p_1$, which is why it rotates around that specific point). For a price $p_1^{***}$, the optimal consumption is the pair $(X_1^{***}, X_{\text{other}}^{***})$. Similarly, if the price rises again to $p_1^{****}$, we get a new tangency at quantity $X_1^{****}$ and so on. We could make as many such points as we wanted by making very small changes in the price $p_1$ and filling in lots and lots of budget lines in Figure A.2(a) (but it would get very messy). Figure A.2(c) simply takes these tangency points and graphs them in a new way, showing combinations of $p_1$ and $X_1$ that follow directly from the consumer’s utility-maximizing choices for the given budget $I$.

---

6This comes from the Latin word tangens, meaning “touching.”
All of the discussion thus far really centers on a single individual, but the transition from the individual to a larger group (society) is quite simple. The aggregate demand curve simply totals at each price the quantities on each individual demand curve for every member of the society. This is called horizontal aggregation because of the common penchant for drawing demand curves with price on the vertical axis and quantity on the horizontal axis of the diagram. Adding things “horizontally” in such a diagram produces the correct aggregate demand curve by showing the quantity consumed at any price by all of the consumers in the market.
such an aggregation for a three-person society, but it should be clear that the process can continue for as many members of society as are relevant with the demand curve for each of the three persons shown and labeled as \(D_1\), \(D_2\), and \(D_3\). The aggregate demand curve [Figure A.3(b)] \(D_{\text{total}}\) adds up, at each possible price, the total quantities demanded. \(D_{\text{total}}\) coincides with \(D_1\) at higher prices because only person 1 has any positive demand for \(X_1\) at prices above \(p_2\). There is a kink in \(D_{\text{total}}\) at \(p_2\) where person 2’s demands begin to add in, and again at \(p_1\), where person 3’s demands begin to add in.

**USING DEMAND CURVES TO MEASURE VALUE**

The previous discussion speaks as if the quantity of medical care that people demand is affected (for example) by price. Demand curves show this relationship, describing quantity as a function of price. They can also be “inverted” to describe the incremental (marginal) value consumers attach to additional consumption of medical care at any level of consumption observed. This approach uses the “willingness-to-pay” interpretation of a demand curve. The only thing that is done here is to read the demand curve in a direction other than is normally done. Rather than saying that the quantity demanded depends on price (the interpretation given above), we can say equally well that the incremental value of consuming more \(X_1\) is equivalent to the consumer’s willingness to pay for a bit more \(X_1\). Just as the quantity demanded falls as the price increases (the first interpretation of the demand curve), we can also see that the marginal value to consumers (incremental willingness to pay) falls as the amount consumed rises. We call these curves inverse demand curves or value curves.
Inverse demand curves (willingness to pay curves) slope downward generically because consumers have declining marginal utility for any good (see Figure A.1), expressing the more common idea of satiation. Ten wonderful restaurant meals a year is terrific. Ten a month would be fine too, but in most cities, one would tend to have worn out the novelty of restaurant visits. Thus, the tenth restaurant visit per month creates less added (marginal) value than the tenth per year. Ten restaurant meals a day would be downright burdensome not to mention unhealthy and would probably have negative marginal value for most people (the value curve would drop below the horizontal axis). No intelligent consumer would go to this extreme, of course, unless bribed to do so (negative prices for restaurant visits).

We can extend the same idea one step further by noting that the total value of using a certain amount consumed to consumers is the area under the demand curve out to the quantity consumed. If we subtract the cost of acquiring the good (or service), we have the unusual but important concept of “consumer surplus,” the amount of value received above and beyond the amount paid to acquire the good. As an example (using discrete quantities of consumption), suppose the first restaurant visit each month created $100 in value to the consumer. If the meal cost $30, the consumer would get $70 in consumer surplus out of that visit. A second meal per month might create an additional $75 in value, again costing $30, and creating $45 in consumer surplus. A third restaurant meal per month might create a marginal value of $35 and a consumer surplus of only $5. A fourth meal would create only $20 in marginal value and would cost $30. Table A.1 summarizes these data.

Two ideas appear in this discussion. First, demand curves can predict quantities consumed: Intelligent decision making will continue to expand the amount consumed until the marginal value received just equals the marginal cost of the service. (In the case of the restaurant meals, we don’t quite achieve “equality” because we described the incremental value of restaurant meals as a lumpy step function, dropping from $100 to $75 to $35 to $20, and the cost was described as $30.)

The second concept is that of total consumer surplus to the consumer from having a specific number of restaurant meals each month. As noted, intelligent planning would stop after the third meal each month. Total consumer surplus sums up all of the extra value to consumers (above the costs paid) for each restaurant meal consumed.

\(^8\)Note that we have to carefully specify the unit of time in demand curves. The quantity consumed is described as a rate per unit of time. Thus, the same ideas as discussed here on a “per month” basis would apply to 12 meals per year, 24 meals per year, and so on because the incremental utility of 1 meal a month ought to equate exactly to 12 meals a year.
TABLE A.1 CONSUMER SURPLUS IN RESTAURANT MEAL EXAMPLE

<table>
<thead>
<tr>
<th>Meal</th>
<th>Value Created ($)</th>
<th>Price Paid ($)</th>
<th>Consumer Surplus ($)</th>
<th>Total Consumer Surplus ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>30</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>2</td>
<td>75</td>
<td>30</td>
<td>45</td>
<td>115</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>30</td>
<td>5</td>
<td>120</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>30</td>
<td>-10</td>
<td>110</td>
</tr>
</tbody>
</table>

It is easy to prove that consumer surplus is maximized by a simple rule: Expand the amount consumed until the marginal benefit has just fallen to match the marginal cost. In this case, at the third restaurant meal per year, the marginal benefit has fallen to $35, the marginal cost is $30 per visit, and stopping at that rate of meals per year maximizes consumer surplus as the last column of Table A.1 shows. Of course, the type of “incremental value” data shown in Table A.1 is just a lumpy version of a demand curve. One could easily create a bar graph of the data in Table A.1 and then draw a smooth line through the midpoints of the tops of each of the bars in and call it a demand curve; adding up such curves across many individuals would smooth things out even more.

The concept of “consumer surplus” is one of the most powerful used by economists, often phrased as “consumer welfare” rather than “consumer surplus.” Many problems studied by economists seek ways to increase (or, best of all, maximize) consumer surplus for individuals or an entire society. These concepts appear in Chapter 4, in which you study some novel problems arising as the step is taken from the general ideas in this appendix to the specific problems associated with demand for medical care.

THE PROBLEM OF SUPPLY OF GOODS

Having gone through the concepts of consumer demand, we can now turn to the problem of how goods and services are supplied to the market. The concepts are actually quite similar. The idea of a “utility function” has a close

---

9Aggregating consumer surplus from individuals to the level of an entire society requires a decision about how one person’s welfare adds to another’s. One way to do it (the most common in economics) says “the welfare created by $1 is the same, no matter whom you give it to.” See, for example, Harberger (1971). Others would use some societal “weighting function” in which the value of improving a person’s well-being differs from person to person. For a detailed analysis of this concept, see Bator (1957) in an article with a very misleading title (it says “simple” but it’s not!). The work of Rawls (1971) delves into these concepts of “economic justice” in great detail, and his work has spawned an enormous literature of commentary best found in your library. All of this analysis comes under the field of “welfare economics.”
match in a “production function” (showing how inputs combine to create the final product).

For a first pass at the concepts of production theory, we can step back for a moment and reflect on our own experiences in “producing” things: More effort leads to more output. In most productive processes, that relationship between “more in” and “more out” doesn’t follow a straight path but one that includes dips and hills. In a discussion of production functions, we can speak about areas in which returns to scale are different—increasing, constant, and decreasing.

Figure A.4 illustrates these concepts for one of several possible inputs (which we’ll call $A_1$). On the top portion, we see the production function graphed showing total output as input $A_1$ increases. Initially (in region $a$),
output rises faster than the input. This is the realm of “increasing returns to scale.” Then come realms b and c, where output rises with more use of input $A_1$, but at a decreasing rate. This is the realm of decreasing returns to scale. Finally, at the boundary between regions c and d, output reaches a maximum and then begins to fall. Region d is irrelevant for any rational productive process because it requires additional resources (more $A_1$) but yields less output $X$.

On the bottom of Figure A.4, the same production function is shown, but here the vertical axis scale has become the ratio of $X/A$ rather than the total amount of $X$ produced (as on the top). The upper and lower portions of Figure A.4 relate in the following way: If you draw a ray from the origin in the top portion to any point on the curve, it shows the ratio $X/A$. If you plot that ratio in the bottom (i.e., the slope of the ray in the top), you get the average product $(AP)$ of $A_1$ ($X/A_1$). The AP peaks at the point dividing regions b and c in the top (where the slope of the ray is as large as it can get).

The bottom portion of Figure A.4 also shows the marginal product of $A_1$, defined as the rate at which $X$ changes for a tiny change in the amount of $A_1$ used, holding constant all other inputs. The marginal product $MP$ rises faster and then falls faster than the average product $AP$.

Now comes a bit of calculus sleight of hand: It’s easy to prove that the optimal use of input $A_1$ comes at the point at which the marginal product of $A_1$ just equals the ratio of $w_1/P$, where $P$ is the price of the final product and $w_1$ is the price of $A_1$, or slightly recast, $P \times MP = w_1$. The idea makes intuitive sense: It says to expand the use of input $A_1$ just to the point at which the “value of the marginal product” (price times the $MP$) just equals the cost of adding another unit of $A_1$.

We can now see that rational production processes always operate in the realm in which the $MP$ is positive but declining (region c in Figure A.4). The optimum—the best that the firm can do—comes at one of the two points at which the $MP$ curve intersects the line $w_1/P$ in the bottom of Figure A.4. One of those intersections occurs in region a and the other in region c. But to stop in region a means not using a very productive input $A_1$ through the entire

---

10In the notation of calculus, the marginal product is defined as $\partial X/\partial A_1$.

11Economists commonly use the symbol $w$ for prices of inputs (sometimes called “factor prices”) because one of the inputs is always “labor” with a “wage” attached to it. Thus, the letter $w$ generally indicates the price in an input market. The letter $W$ took on a new meaning in the U.S. presidential election of 2000.

12Proof: The firm seeks to maximize profits, defined as $\Pi = PX - C(X)$ where $X$ is a function of $A_1, A_2$, etc. Also, the cost of production is $C = w_1A_1 + w_2A_2 + \cdots + w_NA_N$. Thus, to find the optimum, differentiate the profit function with respect to $A_1$, giving $P \partial X/\partial A_1 - w_1 = 0$. This solves for the expression $\partial X/\partial A_1 = w_1/P$. QED.
range at which the value of the marginal product exceeds the cost of adding more input. Thus, the second of the intersections is the relevant one—in region c—and shows the rule for optimizing the firm's profit. (This also works for not-for-profit firms that seek to minimize costs of a given amount of production, an issue that arises in Chapters 8 and 9 in considering not-for-profit hospitals.)

The same process can be shown in another (now familiar) way: We can show the production function (and how it relates to several inputs at once) using Figure A.5 (similar to the “indifference curve” maps showing consumer utility in Figure A.1).

To keep things notationally correct, we can talk about the production of a specific good (say, $X$), with inputs we'll call $A_1$, $A_2$, etc., available to the company producing $X$ at prices $w_1$, $w_2$, etc. The producer's production function is “given” the same way that the consumer’s utility function is “given” for the analysis. The inputs in the production function combine (as defined by the function itself) to produce the output $X$. We'll use the generic function $g(\cdot)$ to describe production functions here, so $X = g(A_1, A_2, \ldots, A_M)$ where there are $M$ inputs in the production function. We can readily graph the production function with the same techniques as we used to graph the utility function, using any two of the inputs ($A_1$ and $A_2$ if we wish) on the axes of the graph.

**FIGURE A.5**

\[ A_1 \]

\[ TC_{12} \]

\[ TC_{11} \]

\[ TC_{10} \]

\[ TC_9 \]

\[ TC_8 \]

\[ x = 12 \]

\[ x = 11 \]

\[ x = 10 \]

\[ x = 9 \]

\[ A_2 \]

\[ $/X$ \]

\[ MC \]

\[ AC \]

\[ x \]

13This simple introduction ignores the problem of optimal investment in R&D to improve the production function.
and now drawing “isoquants” showing combinations of the various inputs that produce the same level of output (some specific quantity of $X$ such as $X = 10, X = 11$, etc.). Figure A.5 shows such a figure.\footnote{Just as in the consumer demand case, we really are talking about rates of production for a given amount of time (e.g., a day, a month, or a year).}

Because the quantities of $X$ have specific (and observable) values (unlike the utilities of the previous discussion), we can say something specific about the spacing between the isoquants. Think about the space between the isoquants for $X = 10, X = 11$, and $X = 12$. If they are the same distance apart (being somewhat loose for the moment about what “the same distance” means), then we have a production that has “constant returns to scale” in the sense that (for example) adding $k\%$ to all of the inputs will add $k\%$ the output. If we get less than an $k\%$ increase in the output for an $k\%$ increase in the inputs, we have “decreasing returns to scale,” and in this case, the isoquant curves will get further and further apart as we move to higher levels of $X$. Conversely, if we get more than an $k\%$ increase in output when we add $k\%$ to all the inputs, we have “increasing returns to scale,” and the isoquants will be spaced closer and closer together.

Just as in the problem of consumer demand, we can talk about the total costs of the firm as the adding up of all the costs of its inputs. Thus, total cost ($TC$) is defined as $TC = w_1A_1 + w_2A_2 + \cdots + w_NA_N$. Again to keep the figures manageable, we can (without loss of generality) restrict ourselves to two inputs, so $TC = w_1A_1 + w_2A_2$ (where $A_2$ can be “other inputs besides $A_1$”). The $TC$ line has a downward slope for the same reasons that the budget line of the individual consumer has a downward slope, and the slope of that line is given by the ratio of the relative prices of the two inputs (in this case, the slope is $-w_2/w_1$ because we have graphed $A_1$ on the vertical axis and $A_2$ on the horizontal axis. The firm wishes to find the lowest possible total cost for any level of output (say, $X = 10$). It does so by finding the $TC$ line just tangent to the isoquant for an output of $X = 10$. Any lower $TC$ line would bring a combination of inputs not sufficient to produce an output of $X = 10$, and any higher cost would, by definition, be higher than necessary. The lowest possible cost occurs at the tangency of an isoquant and a $TC$ line. Figure A.5(a) shows such tangent lines for output quantities $X = 9, X = 10$, and so forth, labeled as $TC_9, TC_{10}$, and so on.

Two more concepts—those of marginal cost and average cost—become useful in a bit. Marginal cost asks the following simple question: “If I move from $X = 10$ to $X = 11$ in output, how much does the $TC$ change?” Of course, the amount it changes can differ as you move from $X = 1$ to $X = 2$ and then again from $X = 10$ to $X = 11$, and similarly for $X = 21$ to $X = 22$, and so on. Thus, the marginal cost ($MC$) in Figure A.5(b) will
vary depending on the level of production, so we need to think about the marginal cost at \( X = 10, X = 11 \), \ldots and so forth, which we could describe as \( MC(X = 10) \), \( MC(X = 11) \), and so on, or in more general form, \( MC(X) \).

The average cost (\( AC \)) is an easier concept: It indicates what the total cost is per unit of output, found by simple division of \( TC/X \). Of course, \( AC \) can also vary with the level of output, so we really need to think about the \( AC \) also varying with the level of output, hence, \( AC(X = 10) \), \( AC(X = 11) \), and so on, or more generally, \( AC(X) \).

We can readily graph \( TC(X) \) versus \( X \) as well as \( MC(X) \) versus \( X \) and \( AC(X) \) versus \( X \).\(^{15} \) Figure A.5(b) converts the information in Figure A.5(a) to a new graph showing quantity produced (\( X \)) on the horizontal axis and both \( MC(X) \) and \( AC(X) \) on the vertical axis. This shift really just shows the spacing between the \( TC \) curves that are tangent to each isoquant. To calculate the \( AC \), we just compute the ratio of the \( TC \) line associated with a given \( X \) by the amount of \( X \) itself. To calculate the \( MC \) for a given value of \( X \), we take the difference between \( TC \) for that amount of \( X \) and for the next smaller amount of \( X \).\(^{16} \)

Returning for the moment to the ideas of increasing, constant, or decreasing returns to scale, it can easily be shown (play with the figures if you wish) that when we have constant returns to scale, we have an \( AC(X) \) curve that’s a straight, flat line. When we have increasing returns to scale, we have a downward-sloping \( AC \) curve. Conversely, when we have decreasing returns to scale, we have an upward-sloping \( AC \) curve. Most production functions exhibit initially increasing returns to scale and then as output expands, decreasing returns to scale. This turns into the “typical” \( AC \) curve with a U shape as characteristically shown in economics texts and here.

The \( MC \) curve has typically more of a J-shaped appearance, turning up faster than the \( AC \) curve. “Standard” production functions also have the characteristic that \( MC = AC \) (i.e., the \( MC \) and \( AC \) curves intersect) at the point at which \( AC \) is at the lowest possible point. The “traditional” shapes of \( MC \) and \( AC \) curves follow directly from the same assumptions about “returns to scale” shown in Figure A.4, with initially increasing, then decreasing returns to scale, corresponding directly to the concepts of declining and then increasing \( AC \).

\(^{15}\)The Appendix to Chapter 3 develops these ideas more fully for the specific case of medical care.

\(^{16}\)If we think of units of \( X \) as being discrete (such as tons of steel), then the \( MC \) curve will be a little kinked, but if we were to think about the \( MC \) curve for pounds, ounces, or grams of steel, it would get much smoother. Calculus simply works out the problem assuming that we can make very, very small changes in \( X \) and then calculate the change in \( TC \) associated with that change. In other words, we get a completely smooth curve by assuming that \( X \) can be produced in continuously divisible units.
We are now in a position to describe the supply curve for a competitive firm if we assume that it wishes to maximize profits (the normal assumption for normal firms, but one we diverged from in studying not-for-profit hospitals in later chapters). It is easy to prove (but we won’t do so here) that the profit-maximizing competitive firm will expand output up to the point—we’ll call it $X^*$—at which the price in the market (a “given” for a competitive firm) just matches the $MC(X^*)$. Thus, as the price increases, the competitive firm will expand output by “marching up” its $MC$ curve. In other words, the $MC$ curve is the supply curve of a competitive firm. Again, this makes good intuitive sense: The profit-maximizing firm will decide to increase output by one more unit if (and only if) the price it receives from selling that one added unit of output at least covers the added costs ($MC$) of producing it. Limiting production to any amount below the point at which means that the firm was forgoing some potential profits, and expanding output above that point means that the firm spent more to produce the added units of production than it received in revenue when selling them. Thus, the profit-maximizing rule says “produce up to the point at which $MC = \text{price}$.”

This rule requires another caveat: If the price is below $AC$ of production, the firm will stop producing because it would lose money compared to quitting business. Thus, the relevant supply curve of a competitive firm is the segment of the $MC$ curve lying at or above the $AC$ curve. Figure A.5(b) shows the relevant segment of the $MC$ curve as a heavier line.

Finally, we can add all of the supply curves of all the firms willing to participate in the market (i.e., those that can at least cover their average costs), adding “horizontally” the same way we can add demand curves of consumers. This gives us the market supply curve in aggregate. This aggregation takes place in exactly the same way that consumers’ demand curves are aggregated to form a market-level demand curve (horizontal aggregation).

**MARKET EQUILIBRIUM**

The preceding two sections have shown how economists think about consumers’ demands for various products and how those add up to a market-level demand curve. Similarly, we can see how production theory leads us to using the $MC$ curve of each firm (in the segment above $AC$) as its supply curve, and we can add up those curves for each firm participating in the market to create the market-level supply curve. With the concepts of demand and supply in hand, we’re now in position to determine the equilibrium output and price in a competitive market. To do this, we simply place the market-level demand and supply curves into the same figure (Figure A.6) and find
the intersection point. That intersection determines the equilibrium market-level quantity \( X^* \) and price \( p^* \).

To see why the point \((X^*, p^*)\) determines the equilibrium, think about moving away from that point with quantity either rising or falling. If quantity “tried” to be higher (for example, if producers tried to push more output into the market), consumer’s willingness to pay would fall below the equilibrium price \( p^* \) for any output larger than \( X^* \), but the added (marginal) costs of producers would rise (because they would be marching up the upward-sloping portions of the \( MC \) curve), and all producers would lose money on every amount sold. This would hold true for any quantity where \( X > X^* \) in Figure A.6.

Similarly, if something tried to hold back the amount produced below \( X^* \), then consumers in aggregate (represented by the aggregate demand curve) would willingly pay a price above \( p^* \) and it would cost less than \( p^* \) to produce a bit more output (because the producers would be on the upward-sloping segment of their \( MC \) curves). Such a situation would produce an opportunity for profits enough to make any profit-maximizing entrepreneur drool—price higher than incremental cost!—and every producer would try to expand output. Producers could do so successfully, but the price would fall to induce consumers to buy the added production (as they march down the aggregate demand curve). This eventually leads back to the point at which the demand curve and the supply curve intersect. At that point, the quantity supplied by producers and the quantity consumed by consumers just equal each other, and the consumer’s willingness to pay (the height of the demand curve)
just matches the incremental costs of producers (the MC curve). These are the essential features of a competitive equilibrium.\(^\text{17}\)

This means that at the level of the individual firm, the competitive equilibrium story is really quite simple: Each firm faces a demand curve that is flat (horizontal), and in fact, the height of that demand curve must be exactly the point at which the AC curve reaches its minimum. The reason why this remarkable result occurs is because of search and entry. Consumers are presumed to search for lower prices and flee from any seller charging a price higher than can be found elsewhere in the market. Similarly, producers are presumed to be willing and able to enter the market when opportunities arise to make unusually large profits. Such profits occur whenever price exceeds average costs of production. Since producers all sell along their MC curves, the only point that meets all of these criteria is where \( P = MC = AC \), which occurs at the very point where AC reaches a minimum.\(^\text{18}\)

A separate branch of economic analysis studies the demand for various “input factors” in their own markets, creating demand curves for these inputs from the producers of the “final products” that can then be matched (in their own markets) with supply curves for those inputs. This creates (in a very similar way) market equilibria for inputs, showing the quantities demanded and the prices for those input factors (described earlier as \( w_1, w_2 \), etc.). These “factor demand curves” ultimately “derive” from the demands consumers have for the final products, as mediated by the production processes available to producers, hence the common name of derived demand curves for input factors.

The essential point here for this level of review of economic theory is that we

\(^\text{17}\)Another interesting feature relates to the concept of welfare economics: Total consumer surplus is maximized in a competitive market, and nothing—no social planner, no genius producer, no government—can improve on the competitive market equilibrium in terms of adding consumer surplus. The original proof of this concept came from Italian economist V. Pareto, who (ironically) spent most of his life as an economic planner for the Italian Nazi (state socialist) party trying to do as a planner something he’d proved impossible as an economist.

To be clear, in many situations, a competitive market will \textit{not} lead to the best social outcome. Most prominent among these settings are cases in which externalities of production or consumption exist. Air pollution from automobiles or smokestacks, traffic congestion on freeways, and the spread of infectious diseases are classic examples of externalities. Chapter 14 is devoted in its entirety to a discussion of externalities. Anybody interested in these topics must absolutely read (as a beginning step) “The Problem of Social Cost,” by Nobel Laureate Ronald Coase (Coase, 1960).

\(^\text{18}\)The proof is fairly simple, again using calculus. Define \( AC = \frac{TC(X)}{X} \) and find the point at which it reaches a minimum by taking the derivative and setting it to zero. Thus [recalling that \( MC = \frac{d(TC)}{dx} \), \( d(AC)/dx = \frac{dX}{dX} \)]

\[ MC = \frac{MC(X) - TC(X)}{X^2}. \]

Setting this equal to zero and rearranging terms gives the solution that \( MC = TC(X)/X \), and, hence, \( MC = AC \). This proves that when AC reaches a minimum, \( MC = AC \). This is why economists always draw MC and AC curves with the MC curve passing through the AC curve at its minimum.
can use the concepts of supply and demand equally well for final product markets and input markets, and the analysis of market behavior in either case.

**MONOPOLY PRICING**

All of the analysis to date presumes a “large” number of buyers and sellers in the market, each behaving independently (no collusion), with consumers seeking to maximize utility and producers seeking to maximize profits. This situation changes in some important ways when the number of sellers is small and (in the extreme case) when the number of sellers in a market is exactly one, we have a “monopoly.” If the monopolist has the same motives as other producers (maximizing profits), the market will function differently than in a competitive market.

The key distinction between “monopoly” markets and competitive markets is the concept of “price taker.” In a purely competitive market, enough buyers and sellers are participating in the market so that each one of these “economic agents” acts as if its own behavior did not affect the final market price. On the buyers’ side, it means that nobody’s purchases of $X$ have enough impact on the market to alter the equilibrium price. On the sellers’ side, it means that nobody’s output decisions have an effect on the equilibrium price.

One way to state this same idea is that the supply curve to any consumer is a flat line (in the usual price-quantity diagram of supply and demand curves), suggesting that—in effect—the market can produce all of the product the consumer wishes to buy at the same price. Similarly, it says that sellers individually confront a demand curve that is flat (a constant price received for their product) in the sense that if they tried to raise their price above that market-equilibrium price, all of their potential buyers would migrate to some other seller who kept the price at $p^*$. Thus, in either the case of buyers or sellers, a competitive market means that both face a constant price in all of their decision making; hence, we can view both buyers and sellers in a competitive market as price takers. In other words, they “take” the price as a given fixed amount in their own decision making.

In monopoly markets, just the reverse situation occurs: The seller must recognize that when the amount produced changes, the price will change accordingly because the seller confronts a demand curve that matches the market demand curve. Thus, if the seller wishes to expand output, the price in the market must fall (along the demand curve) to induce consumers to actually buy the additional amount produced. So long as the seller can’t set different prices for different buyers, this means that the monopoly seller’s profit-maximizing decision must account for consumer preferences for their product (as expressed in the market-level demand curve).
Figure A.7 shows the essential features of the monopolist’s problem. The new and essential feature in this diagram is the “marginal revenue” curve, a curve showing how much new total revenue the producer will receive by expanding output by one unit. Think about a situation in which the producer sold 100 units at a price of $4 per unit. Now suppose that to expand to 101 units of output, the price would have to fall to $3.99 per unit to increase total buying by consumers to 101 units. (Somebody has to decide to increase the quantity demanded by 1 unit, and in this example, the reduction from $4 to $3.99 causes just one person to increase his or her demand by 1 unit.) Revenue at 100 units is $4 \times 100 = \$400$. Revenue at 101 units is $3.99 \times 101 = \$402.99$. Thus, the marginal revenue at 101 units is $2.99—\text{not} the \$3.99 received for the 101st unit of sales. The difference is that the producer had to give up $0.01 on 100 other units of sale ($1.00 loss) to induce consumers collectively to buy one more unit. Conveniently, when the demand curve is drawn as a straight line (as in Figure A.7(a)) the marginal revenue (MR) curve is also a straight line, beginning at the same point on the vertical (price) axis of the figure, but dropping at twice the rate of the demand curve.19

The monopolist’s goal of maximizing profits has a simple logic: Keep expanding output until the $MR$ received just matches the $MC$ of adding the next unit of output and then stop. In other words, expand production until $MR = MC$. In Figure A.7, this occurs at an output level of $X_m$. Then the monopolist sets the price (using the demand curve) to clear the market by

19Thus, for example, if the demand curve is of the form $p = a - bX$, then the $MR$ curve has the form $MR = a - 2bX$. 

**FIGURE A.7**
setting price at \( p_m \). At that price, consumers wish to buy exactly \( X_m \) units, thus “clearing the market.” The profits captured by the monopolist can’t be any larger at any other output.\(^{20}\)

Several important features distinguish monopoly from competitive markets. First, and most obviously, the price is higher and the total amount produced (and hence consumed) is smaller in the monopoly market than a corresponding competitive market would produce with the same demand curves of consumers and similar production characteristics on the part of suppliers. The other point—and the reason economists usually express a distaste for monopolies—is that the monopoly market produces less consumer surplus than a competitive market would produce. (More formally, it produces less economic surplus defined as the sum of consumer surplus and an equivalent concept of producer surplus that we’ll not delve into here.) To see this briefly, look at Figure A.7(b), reproducing Figure A.7(a) but adding the equivalent price and quantity (\( p_c \) and \( X_c \), respectively) that a competitive market would produce if the same \( MC \) curve were present in a competitive market. If somehow the competitive market outcome could be achieved, output (and consumption) would expand from \( X_m \) to \( X_c \), and the price would fall from \( p_m \) to \( p_c \). Consumer surplus would rise because consumers would get to both consume more and pay less for what they consumed. The rectangle \( A \) in Figure A.7(a) represents the transfer of profits gained through monopoly pricing from the monopolist to the consumers. The triangle \( B \) represents the gain in consumer surplus created by expanding output to \( X_c \). Those who own part of the company achieving the monopoly profits shown in rectangle \( A \) probably feel differently about the virtues of shifting to a competitive price than consumers of the product. But triangle \( B \) represents a pure social waste because consumers gain the consumer surplus at nobody’s expense (in the sense that the monopolist isn’t losing triangle \( B \)).

The other feature about monopolies is that they tend to attract entrants into their (unusually profitable) markets. Abnormally large profits by existing producers imply profit opportunities for potential rivals. Thus, unless something stands in the way of entry, economists usually expect to see monopoly profits erode.

\(^{20}\)The proof requires a bit of calculus: Define profits as \( \Pi = p(X)X - TC(X) \), where \( p(X) \) is the demand curve (expressed with price as a function of quantity) and \( TC(X) \) is the total cost of producing \( X \). To find the maximum profit, take the derivative with respect to \( X \), set it to zero, and solve the resulting equation. Thus, \( d\Pi/dX = X(dp/dX) + p - d\left[TC(X)\right]/dX \)

\( = X(dp/dX) + p - MC(X) = 0 \). Reshuffling these terms slightly gives \( MC = p + X(dp/dX) \).

The latter term is just the marginal revenue function—the price (from the demand curve) adjusted by the amount that price falls as you change \( X(dp/dX) \). Because the demand curve is downward sloping, \( dp/dX \) is negative, and \( MR \) is less than price.
What types of things can stand in the way of entry? Often the answer is government interference in the market. Many countries have long had industrial policies designed to protect the profits of incumbent producers. (Another distinct literature studies the reasons for this, known as political economy.) Governments requiring licenses for producers is a common example, and in the United States, governments require licenses not only for physicians, nurses, and pharmacists but also for barbers, lawyers, taxicabs, ice cream vendors, and an amazingly large additional array of classes of sellers.

Sometimes the nature of production creates a “natural monopoly.” For example, if the market is so small that a single firm can produce all of the desired output while still experiencing declining average costs, then the market usually sustains only one seller. Classic examples of such markets are the distribution of natural gas and electricity and (formerly) telephone service. But as commonly happens in a world with technical progress, new techniques can evolve to allow competition in what was once a “natural monopoly.” The entrance of cellular telephones to compete with land-line telephone systems provides a useful case study of the effects of new technologies on the way a market functions. Often the government steps in when the market has a true natural monopoly to control the prices charged; examples include gas, electricity, and (formerly) telephone distribution.

Something else that can get in the way of competitive pricing is simple collusion between sellers. If sellers get together and agree on output, they can act “as one” and hence achieve monopoly prices even when many firms participate in the market. OPEC is one prominent international organization that deliberately tries to control the output of crude oil in world markets. Within individual countries, government policy can either aid and abet collusion (a common outcome in many countries), or it can vigorously work to oppose collusion. In the United States, major laws passed early in the twentieth century (most notable the Sherman Anti-Trust Act and the Robinson-Patman Act) make collusion illegal, and the U.S. Department of Justice and the Federal Trade Commission act as major “watchdogs” to prevent collusion in U.S. markets. In other countries (and sometimes in the United States in the past and at present), the government actively works to support collusion among producers. One way to support such activities comes when the government restricts entry into a market (such as with licenses for taxicabs).

MONOPOLISTIC COMPETITION

Some markets behave neither as competitive markets nor monopolistic markets. A huge volume of literature has evolved to attempt to understand how markets work when they fit neither the competitive nor the monopolistic
paradigm, a branch of economic analysis known as *industrial organization* (or more succinctly, *IO*). This literature seeks to understand how producers interact with one another in their behavior in these “in-between” situations. Possible outcomes in price and quantity commonly range between the competitive and monopoly equilibria, but the market’s actual characteristics and the nature of interaction between producers determine just where the equilibrium will end up. The analysis of all of these potential markets lies far beyond the scope of this appendix and, indeed, for the analysis in the main text in general. However, one particular type of market equilibrium appears useful in studying many health care markets—the outcome known as *monopolistic competition*—so we will provide a brief overview of that market here.21

Monopolistic competition (as the phrase suggests) combines elements of both monopoly and competitive markets. Most usefully, we can think about them as emerging in situations in which one of the key elements of a competitive market—consumer search—breaks down. Thus, we can add one more item to the list of things at the end of the previous section that leads to monopoly pricing: failure of consumer search.

Think for a moment about a market with many buyers and a large handful of sellers, none of whom colludes on output or price, and into which market entry is perfectly free. If entry is truly unrestricted, then unusually large profits will induce entry, ultimately to the point of eroding away those extra profit opportunities. In monopolistic competition, we presume that such free entry can occur readily.

Now continue for a moment with the thought experiment of how the market would behave if *nobody* looked around for a low price when going out to purchase some good or service, and suppose that all of the sellers knew that all of the consumers behaved in this way. Each seller would then logically behave as a monopolist, knowing that whatever fraction of all buyers decided to buy from each seller would represent an opportunity for price setting just as a true monopolist would have. Each seller would, in effect, have a mini-monopoly, his or her own share of the total market, without fear that charging a higher price would cause consumers to flee to other sellers. (They wouldn’t flee, of course, if they never shopped for a better price than the first one they found.)

In this extreme-form case (absolutely no comparison shopping), the market equilibrium would look like a series of monopoly markets. Now let’s

---

21 The discussion that follows paraphrases work by Schwartz and Wilde (1979), Schwartz and Wilde (1982a,b), and Sadanand and Wilde (1982).
alter the story a bit, having some but not all consumers in the market engaging in a search for lower prices. Now each seller faces a different pricing problem: If each one keeps its price high, each will make more money from the nonshoppers but runs the risk of losing the sale to the shoppers (who might find a lower price from another seller in their search for a better price). Thus, each seller will set prices in a way that trades off higher profits from nonshoppers in return for attracting more business from shoppers.

It doesn’t take a lot of mathematics to understand the way such markets will work: The higher the percentage of consumers who shop (and the more they shop before buying), the closer the market will move toward the competitive equilibrium. The less people shop, the more it will move toward the monopoly equilibrium. Thus, the nature of the monopolistically competitive equilibrium will hinge greatly on the costs of search that consumers encounter.

Another way to phrase this story is based on the nature of the demand curve facing each seller. In a purely competitive case, the demand curve is completely flat in the usual price-quantity diagram, and we describe each seller as a price taker. Each seller has no control over the price it receives. In the other extreme—the monopoly market—the demand curve facing the seller is downward-sloped, and in fact is the same (by definition) as the market demand curve (derived by adding all of the demand curves of all of the consumers in the market). In the monopolistic competition model, the demand curve facing each seller is a mix of those two cases, the mix being determined by the extent of consumer shopping taking place. With much shopping occurring, the mix mostly contains the competitive model (a flat demand curve). With little shopping occurring, the mix mostly reflects the monopoly model (a downward-sloping demand curve reflecting the market demand curve). Thus, the slope of the demand curve facing each firm in a monopolistically competitive market must lie somewhere between the slope of the market demand curve and a slope of zero (a horizontal, flat demand curve).

Figure A.8 shows the final equilibrium picture for each firm in the market. The demand curve still slopes downward (because incomplete shopping mixes in some of the slope of the market demand curve with the flat demand curve of pure competition). However, profits must fall to zero because of the presumption of completely free (unfettered) entry. The latter condition occurs when the average costs of production just match the price received by the firm. Thus, to understand the monopolistically competitive equilibrium, we need to know the average cost curve of the firm (but not its marginal cost as in the case of purely competitive markets or monopoly markets). Entry by competitors spreads the available customers around to different firms, which we portray by showing that as entry occurs, the demand curve for the firm shifts inward (while still retaining a nonzero slope because of incomplete shopping). The
shift inward can and must continue only to the point at which revenue from sales just covers costs of production. This takes place when the demand curve for each firm has shifted in to the point at which it just touches the AC curve of the firm at one point (on the left-hand branch of the AC curve). This defines the monopolistically competitive equilibrium: Each firm's demand curve is tangent to its AC curve, so each firm makes no unusually large profits, and each firm faces a downward sloping demand curve.

The equilibrium price charged by this firm will be $p_{mc}$ and its output will be $X_{mc}$. Obviously, the price will be higher than the competitive case [where $P_c = \min(AC)$], and each firm's output will be lower than would occur in a competitive market. As a final observation, we might note that if the extent of comparison shopping increased sufficiently, the demand curves facing each firm would rotate more and more toward the flat demand curves facing a competitive firm, and when enough shopping takes place (it turns out not to require 100% of the buyers to carry out comparison shopping), the market “collapses” to the competitive equilibrium.