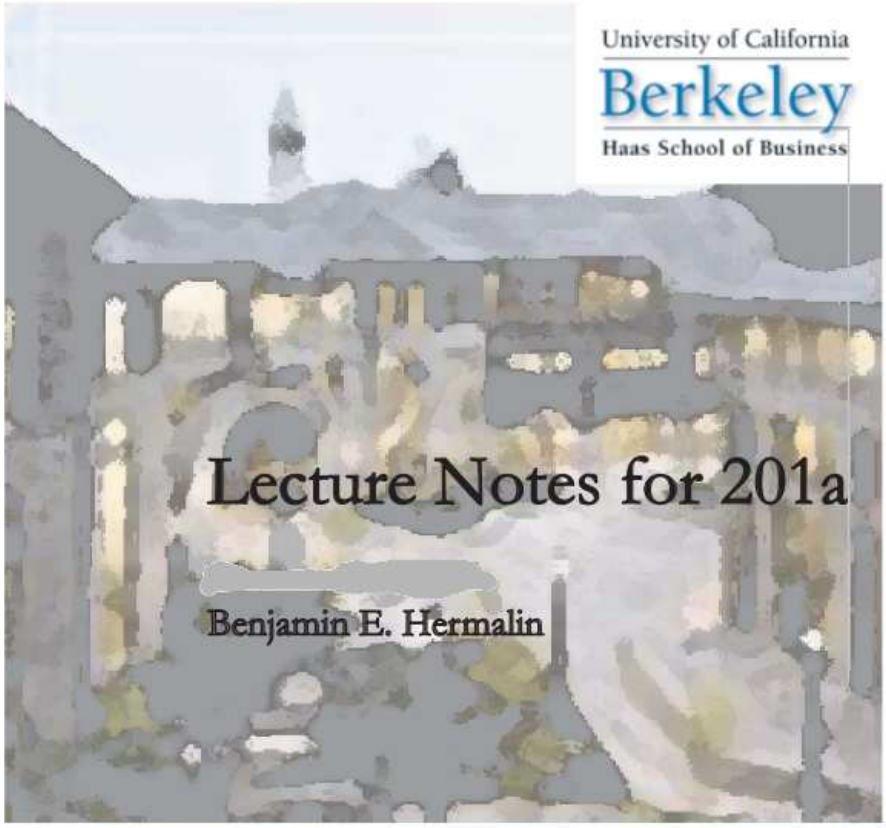

Contents

List of Tables

List of Figures

List of Examples

Preface

These lecture notes are intended to supplement the lectures and other materials in the Microeconomics for Business Decision Making course at the Haas School of Business.

A Word on Notation

Various typographic conventions are used to help guide you through this text.

Text that *looks like this* is an important definition. On the screen or printed using a color printer, such definitions should appear blue.

Notes in margins: *These denote important "takeaways."*

The symbol in the margin denotes a paragraph that may be hard to follow and, thus, requires particularly close attention (not that you should read any of this text without paying close attention). On rare occasions, a paragraph might be so difficult as to warrant a .

The OPT symbol at the beginning of certain footnotes indicates that the reading the footnote is *optional*.

The symbol $\int dx$ in the margin denotes a section that uses calculus. Observe that such sections are also indented relative to the rest of the text. Some technical footnotes are also prefaced with $\int dx$. It is understood that not everyone is comfortable with calculus and, for those who aren't, I simply ask that you skim those sections and make note of the main conclusions. You may also wish to read through Appendix A3. There is no expectation that every student will fully grasp the details of the analysis in those sections that employ calculus. When using calculus, prime indicate the derivatives of functions. Hence, for instance, the derivative of $f(\cdot)$ would be represented as $f'(\cdot)$. Occasionally, the dy/dx style of notation will be used for derivatives.

A Word on Currency

As a rule, I will refer to monetary amounts as dollars. Unless I'm citing actual data, there is nothing in these notes specific to dollars. The analysis is the same whether the monetary amounts are dollars, euros, pesos, pounds, yen, or what have you.

For those who feel I'm being too American-centric in using dollars, note I didn't specify which dollars. For all you know, I have Australian, Canadian, or Hong Kong dollars in mind.

A Word on the Appendices

There are two sets of appendices. The "A" appendices review basic algebra, solving equations, and calculus. The "B" appendices consider basic probability. Reading them is optional. Even though reading them is optional, this doesn't mean that using the mathematics they contain is necessarily optional. If, therefore, you feel rusty with respect to your math or probability skills, you should probably read through them. To the best of my knowledge, no harm will come to you if you do.

Fundamentals of Problem Solving & Decision Theory | 1

Many years ago, Greg Wolfson, a former student, and his wife were in the Caribbean as a hurrican. Understandably, they were nervous about the possibility of being stuck on the Turks and Caicos Islands during what looked like one of the worst storms of the century. Should they stay on the Islands or should they try to make it to Miami on route back home? If they stayed and the hurricane hit the Islands, then they faced having their vacation ruined or worse. If they left, then they gave up the rest of their vacation, incurred additional costs of getting last-minute plane tickets, and ran the risk of being caught in the hurricane while in Miami. Fortunately, Greg had studied *decision theory*. Decision theory helped Greg and his wife to think *systematically* through their decision problem—stay or flee—and reach their best decision. This turned out to be "stay," and a good thing too: While Miami was being battered by Hurricane Andrew, Greg and his wife were on a Hobie Cat, sailing the lovely turquoise waters off the Turks and Caicos Islandsanother MBA 201A success story!

Decision theory is a set of tools for deciding "which." For example, Greg and his wife were deciding which would be the better course of action for them, stay or flee. These tools help by *formalizing* your decision making. They help you recognize the alternatives available to you; they help you see what additional information would be useful in reaching a decision; and they make you aware of the assumptions or conditions that are critical to the decision you make.

A word of caution: As powerful as these tools are, they are *not* a substitute for your own thinking. Rather, they are aids to your thinking. Put another way, they are not magic formulæthat can make your decisions for you (which is just as well, since otherwise someone would program a computer with them, which would likely do you out of a job).

The First Rule: Know Your Problem | 1.1

Before you can solve a problem, you have to know what it is. This may seem so obvious that it hardly warrants mention. Obvious though it may be, the truth is that people aren't always that good at identifying what the problem is. This may surprise you—after all, you've been solving problems in and out of school for as long as you can remember. However, most of the problems we solve in school (and life) have been *given to us*. We're asked to solve problem

that someone else has posed. But part of good management is identifying the relevant problems.

An example may help to illustrate the issue.

> **Example 1 [Sterling Chemicals]:** Sterling Chemicals, Inc. was founded in 1986 in a $213 million leveraged buyout of Monsanto Corporation's Texas City plant.[1] The plant is located on Galveston Bay and manufactures seven commodity chemicals and their coproducts. The plant has the world's largest styrene monomer unit. It is the only domestic producer of synthetic lactic acid and tertiary-butylamine. In 1987, its first year of operation, Sterling employed 950 people and had sales of $413 million.
>
> At Sterling's Texas City plant, setting up scaffolding is the first task in most repair and maintenance jobs. If the required scaffolding is not available, the job falls behind schedule and workers end up waiting rather than working. Currently, scaffolding is available for only 43% of scheduled jobs.
>
> A carpenter with fourteen years of experience described the problem as follows:
>
> > Carpenters always complained about not being able to find enough scaffolding. The shortages were so bad that we were spending more time trying to find scaffolding than we spent erecting it. The necessary scaffolding was never at the scaffolding storage racks near the project sites, so we usually had to check storage racks throughout the plant. We calculated that $500,000 worth of labor was being spent each year looking for scaffolding.
>
> A study found that, for 57% of all maintenance projects, there was not enough scaffolding available at the scaffolding storage area nearest the project site. This required carpenters to search other, nearby racks for the necessary scaffolding. In 24% of the cases, they had to ask the truck department to search the plant for the scaffolding needed. In short, considerable time and effort were devoted to scrounging the necessary scaffolding.
>
> Some think the solution is obvious—the plant doesn't have enough scaffolding. One estimate is that Sterling needs $100,000 worth of additional scaffolding.
>
> Management is, however, wary about spending money on improvements unless it is absolutely necessary, so management wants further analysis and thought.

Problems: *A problem is a question.*

What was the problem at Sterling Chemical? The common answer, both at Sterling and when this problem is posed to first-year MBA students, is "Sterling Chemical has too little scaffolding." But that is wrong. The problem is "Why isn't scaffolding readily available?" A possible answer (*i.e., cause* of the problem) is that Sterling has too little scaffolding, but that is not the *problem* itself. A problem is something to be solved; it is a question, typically one that begins with "Why . . . "

[1] This case is drawn from "Science, Specific Knowledge, and Total Quality Management" by Karen H. Wruck and Michael C. Jensen, which appeared in the *Journal of Accounting and Economics*, Vol. 18 (1994), pp. 247–287.

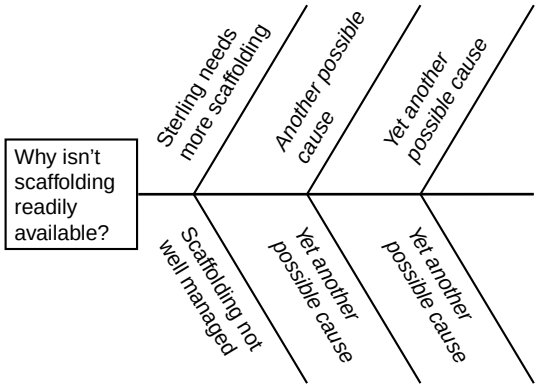

Figure 1.1: ***Fishbone Analysis.*** A problem (*e.g.*, "why isn't scaffolding readily available?") is put in a box ("the head") and possible causes are suggested ("the ribs"), with the ribs closest to the head denoting the most likely causes.

1.2 Fishbone Analysis

Once we've identified the problem we wish to solve, we need to solve it. There are many methods of solving problems and it is beyond these notes to cover them all. Instead, the focus will be on two methods. One, which we will take up later, is *decision analysis*, which is useful for solving decisions problems (this was the type of analysis employed by Greg Wolfson). The second, which is better suited to more open-ended problems, is *fishbone analysis*.

Figure 1.1 illustrates what fishbone analysis is all about. The problem—*e.g.*, "Why isn't scaffolding readily available?"—is put in a box at the front of the diagram. The possible causes for the problem are listed. Here, two have been given: "Sterling needs more scaffolding" and "scaffolding is not well managed." Room has been left for other possible causes. If you have some imagination, you can see why this is called fishbone analysis—the diagram resembles a fish's skeleton. Observe the problem to be solved is the "head" of the fish and the possible causes are the "ribs." Typically, possible causes that are most likely to be the true cause are put closest to the head. Less plausible causes are put further from the head.

Once your fish is drawn, the next step is to investigate each of the possible causes, starting with the ribs closest to the head. Following that course of action, consider the first rib: "Sterling needs more scaffolding." How do we know if this is the cause? One answer is to inventory the scaffolding and check. This is, in fact, what Sterling did after drawing its fish.

When Sterling inventoried its scaffolding, what it found "… was that we had more than enough scaffolding on site, but that it was frequently in the

wrong place at the wrong time."[2] In fact Sterling had 133 *extra* units of scaffolding. Clearly, the proposed cause, "Sterling needs more scaffolding," is wrong.

Moreover, the evidence from the scaffolding's inventorying supports the proposed cause "Scaffolding is not well managed." So Sterling investigated this. What it found was that various teams were hoarding scaffolding. As one carpenter put it

> We all knew that there were guys out there who hoarded scaffolding. If you ever needed a cross brace, you knew that Charley would have some. And if you needed a ladder section, you knew that Bob was a 'specialist' in those. They hoarded what they used frequently so that they wouldn't have to go scavenging. But this caused shortages at the other storage racks.[3]

As a consequence, Sterling adopted changes to its scaffolding management that all but eliminated shortages (necessary scaffolding was immediately available 97% of the time). Observe that by deploying fishbone analysis, Sterling avoided jumping to the "obvious"—but false—solution of buying more scaffolding. At the very least, this analysis kept Sterling from wasting $100,000. Furthermore, to the extent additional scaffolding wouldn't have fixed the availability problem, it potentially saved Sterling even more.

1.3 Introduction to Decision Analysis

The first rule of decision making: *Know your goals (objectives).*

In many cases, solving your problem involves choosing among *alternatives*. Your objective is to choose the alternative that is best, where "best" depends on what your goals are. Indeed, the first rule of decision making is to know what your goals are. For example, if your decision problem is which movie to see at the multiplex, then "best" means "most entertaining" (assuming being entertained is your goal).

Although the first rule of decision making may strike you as obvious, you would be surprised how often people start making decisions without thinking through what their goals are. For instance, obeying the first rule can often be a problem when a committee makes a decision, because committee members can have different goals. Sometimes the committee members recognize their differences in advance, but sometimes they are unspoken. Occasionally committee members believe they are in agreement with respect to their goals when, in fact, they are not (you've likely had conversations that began "it just didnt occur to me that you wanted . . . "). Psychology also plays a role here. You may not, for example, want to admit to yourself what your true goals are—perhaps because they are socially unacceptable—so you convince yourself that your goals are something else. Unfortunately, it is beyond these notes to make sure

[2] Wruck and Jensen, cite *supra*, page 256.

[3] Wruck and Jensen, cite *supra*, page 256.

that you obey the first rule. All they can do, as they have just done, is point out that obeying the first rule is not as easy as it may at first seem.

Having identified your goals, you next have to identify your alternatives. For some decision-making problems, your alternatives are obvious. For instance, if you are deciding which movie at the multiplex to see, then your alternatives are the movies playing plus, possibly, not seeing any movie at all. For other problems, however, identifying your alternatives is more difficult. For instance, if you are deciding which personal computer to buy, then it can be quite difficult to identify *all* your alternatives (*e.g.*, you may not know all the companies that make computers or all the optional configurations available). Fortunately, there are ways to overcome, at least partially, such difficulties, as we will see later. We will even study the *decision* of whether you should expend resources expanding your list of alternatives later in these notes.

Your choice of alternative will lead to some consequence. Depending on the decision-making problem you face, the consequence of choosing a given alternative will be either known or uncertain. If you are driving in your neighborhood, then you know where you will end up if you turn left at a given intersection. If you are investing in the stock market, then you are uncertain about what returns you will earn. Typically, we will suppose that even if you are uncertain about which particular consequence will occur, you know the set of possible consequences. For instance, although you don't know what your stock price will be a year from now, you do know that it will be some non-negative number. Moreover, you likely know something about which stock prices are more or less likely. For example, you may believe that it is more likely that your stock's price will change by 20% or less than it will change by 21% or more.

Sometimes, however, you may not know what all the possible consequences are. That is, some possible consequences could be *unforeseen*. To give an example, the author once met a vineyard owner who was proud of his "green" farming techniques. Unlike many of his fellow vintners, he used pesticides that killed only the "bad" bugs, leaving the "good" bugs—those that ate the bad bugs—alive. A consequence of this, which was unforeseen by the vintner, was that if he successfully killed the bad bugs, then the good bugs would be left with nothing to eat and would starve.

By their very nature, unforeseen consequences are difficult to identify prior to making your decisions. And for the same reason, it is difficult to predict which consequences will be unforeseen by others. As a practical manner, one way to *help* identify unforeseen consequences in your own decision making is to think about what your "un-goals" are; that is, the consequences you would like not to happen. For instance, an un-goal of the vintner was to kill the good bugs. Another way to identify unforeseen consequences is to reframe your way of thinking about your goals. For instance, instead of thinking about not killing the good bugs, think instead of helping the good bugs to survive. Reframed in this way, the adverse consequence of killing the good bugs' food supply might be more apparent. As discussed in any good organizational behavior class, how we think about a problem is very much tied to how the problem is

Un-goals: *A good manager tries to identify unintended consequences.*

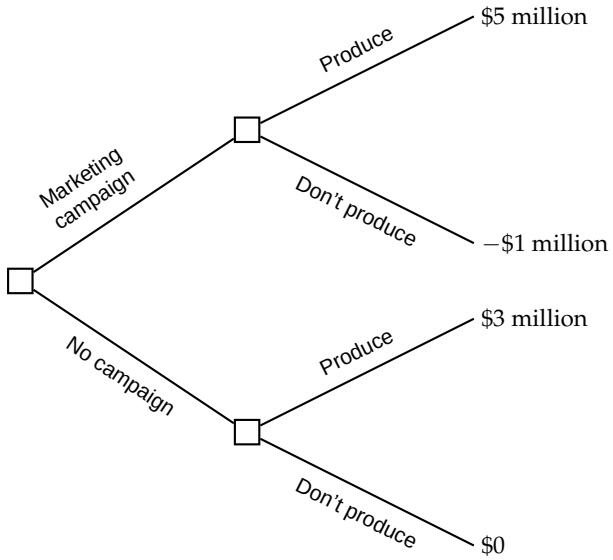

Figure 1.2: A decision tree for a firm that must first decide whether to launch a marketing campaign for a new product. The firm's subsequent decision is whether to actually produce the product. Square nodes are *decision nodes*. From them stem *branches*. At the end of the tree are *payoffs*.

framed. It is beyond the scope of these notes to discuss *framing effects* fully, but it is worth pointing out they exist.

1.4 Decision Making Under Certainty

Decision tree: *A graphical representation of a decision problem as a series of branching alternatives.*

To represent, in a schematic way, a decision problem, we draw a *decision tree*. An example of such a tree is shown in Figure 1.2. It represents the following problem: A firm is considering producing a new product. Prior to producing, the firm can conduct a marketing campaign (*e.g.*, advertise heavily) or not. Suppose that a marketing campaign costs $1 million. Suppose that if the product is marketed and produced, it will generate revenues of $8 million while costing $2 million to produce; so profit will be $5 million (= $(8 − 2 − 1) million). If the product is marketed, but not produced, it will generate no revenue and no production cost; so profit will be $1 million (*i.e.*, just the cost of the marketing campaign). If the product is not marketed, but produced, it will generate a revenue of $4 million but cost $1 million to produce; so profit will be $3 million. Finally, if the product is neither marketed, nor produced, then profit will be $0.

Each square in Figure 1.2 is a *decision node*. A decision node indicates

that the decision maker (in this case the firm) has a decision to make. The alternatives available to him or her at that decision node are represented by the *branches* that stem from the right-side of the decision node. For instance, the alternatives available to the firm at the left-most—first—decision node are "marketing campaign" and "no campaign." Note that the tree is read from left to right; moreover, going from left to right is meant to represent the sequence of decisions. For example, the firm must first decide whether to have a marketing campaign before it decides whether to produce the product. At the end of the tree are the consequences or *payoffs* from the sequence of decisions. For instance, if the firm chose "no campaign" and, then, "produce," it would make a profit of $3 million. Payoffs are expressed in terms relevant to the decision maker's goals. Here, the goal is to make money, so they are represented in monetary terms.

Having represented a decision problem by a tree, the next step is to solve it. Solving a tree means determining which decisions will best accomplish the decision makers goals. If, as in Figure 1.2, the goal is to make money, then this means determining the decisions that will lead to the most money. Trees are solved by working backwards, a procedure known as *backwards induction*: Start at the rightmost decision nodes and select the branches that give the decision maker the largest payoffs. For the tree in Figure 1.2, this means choosing "produce" at the top rightmost decision node—since a $5 million gain is better than a $1 million loss—and choosing "produce" at the bottom rightmost decision node—since a $3 million gain is better than $0. Next move left to the preceding decision nodes. Again, select the branches that give the decision maker the largest payoffs *taking into account, if necessary, the future decisions that will be made*. In Figure 1.2, this means choosing "marketing campaign" at the first node because doing so ultimately leads to a payoff of $5 million, whereas "no marketing campaign" ultimately leads to a payoff of $3 million. Were there any decision nodes to the left of the first node (*i.e.*, were there decisions to be taken prior to the decision of whether to market), then you would choose the alternatives at those nodes taking into account your decision to have a marketing campaign at the "marketing campaign/no marketing campaign" node.

Backwards induction: *Trees are solved by working backwards*

The tree in Figure 1.2 is straightforward, so solving it is straightforward as well. For more complicated trees, however, it is important to keep track of where you are as you work backwards. Two devices for keeping track are *arrowing* the correct decision and *valuing* the intermediate decision nodes (*i.e.*, the decision nodes other than the first). Arrowing means putting a little arrow (or other mark) on the correct decision. When youre done arrowing, the arrows will guide you through the tree. For instance, in Figure 1.2, you would put an arrow on the "marketing campaign" branch, because that is the correct alternative to choose. Valuing a decision node means writing the payoff from making the correct decision at that decision node. Figure 1.3 shows Figure 1.2 with arrows and values.

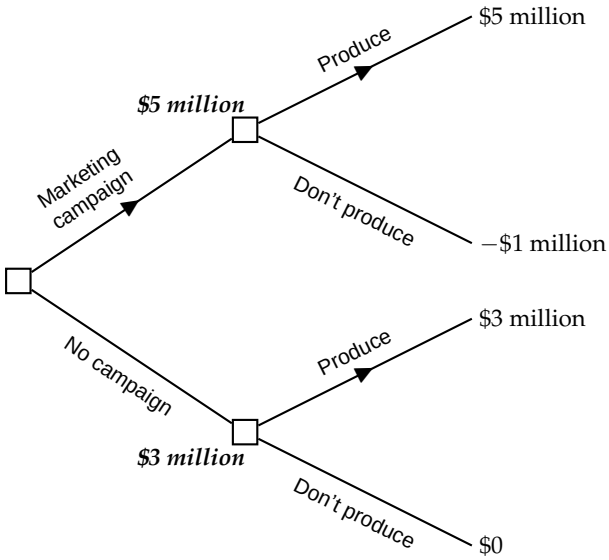

Figure 1.3: The correct branches to follow are noted with arrows (►). The values of the intermediate nodes are also noted.

1.5 Decision Making Under Uncertainty

[Before reading this section, you may wish to review the material in Appendices B1 and B3.]

Many decisions are made in situations of uncertainty. To represent such decision problems, we use a second kind of node: a *chance node*. A chance node, drawn as a circle, indicates that what follows is uncertain. Each branch stemming from the chance node shows a possible outcome of the random process that the chance node represents. For instance, Figure 1.4 represents the decision tree associated with launching a new product.

In Figure 1.4, the firm can either launch or not launch a new product. This is a decision of the firm, so represented by a decision node (square). If it doesnt launch, then its payoff is $0. If it does launch, its payoff is uncertain. Hence, the "launch" branch leads to a chance node (circle). There are three possible outcomes if the firm launches. The new product can be very successful—a blockbuster; or it do acceptably well; or it can fail. The profits associated with each outcome are indicated at the ends of the respective branches (*e.g.*, acceptable performance yields $3 million in profit). The probabilities of each of the possible outcomes are shown in square brackets. Thus, for instance, the probability of failure is .05. Note, summing across the branches that emanate from a chance node, the probabilities must sum to 1 (*e.g.*, $.55 + .40 + .05 = 1$).

Recall that the *expected value of a gamble* when there are N possible out-

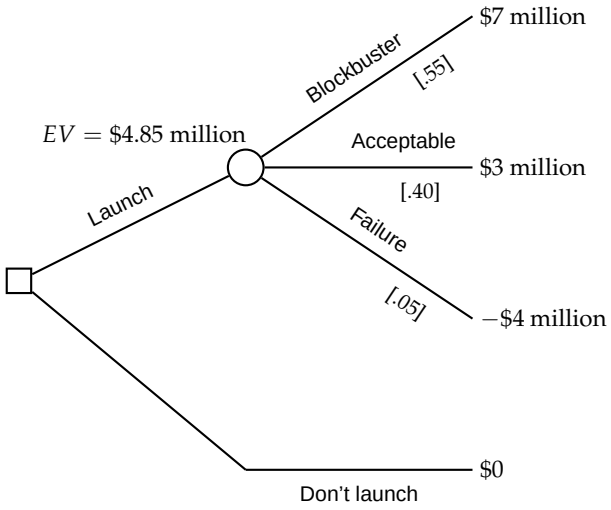

Figure 1.4: A firm faces the decision of launching a new product or not launching it. There are three possible outcomes if it launches: blockbuster, acceptable, and failure. The payoff from each is shown at the end of the tree and the probability of each is indicated in brackets.

comes, denoted EV , is given by the formula

$$EV = p_1 x_1 + \cdots p_N x_N \,,$$

where p_n is the probability of the nth outcome and x_n is the payoff should the nth outcome occur (see expression (B3.1) on page 171). For example, the expected value of the gamble shown in Figure 1.4 is, in millions,

$$\$4.85 = .55 \times \$7 + .4 \times \$3 + .05 \times (-\$4) \,.$$

It might strike you as odd that \$4.85 million is called the "expected" value of that gamble: How could \$4.85 million be "expected" if the only possible values are \$7 million, \$3 million, and \$4 million? There are two justifications for the term "expected." First, suppose that we repeated the gamble many times, say 10,000 times. Your average winnings—that is, the total of your winnings for the 10,000 repetitions divided by 10,000—would very high probability be close to \$4.85 million.[4] In other words, we would *expect* your average winnings to be \$4.85 million.

As a second justification, suppose that you ran a life insurance company. Then, for each insured, you are essentially gambling on when he or she will die (*i.e.*, x would be age at death). The expected value would, thus, refer to the expected age of death of an insured. If you insure a large enough population, then the average age at death among your insureds will be close to the expected age of death of a single insured.

Expected-value maximizer: *A decision maker who chooses, from among her alternatives, the one the yields the greatest expected value.*

Expected value is useful for decision theory because many decision makers are *expected-value maximizers*, which is to say that they choose among their alternatives the one that yields the greatest expected value.

To better understand an expected-value maximizers behavior, consider the decision problem shown in Figure 1.4. As noted above, the expected value of launching is \$4.85 million. Because \$4.85 million is greater than \$0 (the firms payoff if it doesnt launch), the firm would choose to launch if it is an expected-value maximizer.

Note that we solve the tree in Figure 1.4 in the same way we solve all trees—by working backwards: We start at the rightmost node, in this case a chance node, calculate the value of that node (*i.e.*, its expected value), move left to the preceding node (*i.e.*, the launch/no launch decision node), and choose the alternative that yields the greater expected value (here, launch).

To extend this example, suppose that, prior to making the launch/no launch decision, the firm could commission a survey that would tell it how successful the product will be. Of course, whether to conduct the survey is itself a decision, so it must be added to the tree. Figure 1.5 revises the tree in Figure 1.4 to reflect these changes.

[4] For instance, there is approximately an 80% probability that your average winnings would fall between \$4.75 and \$4.95 million and a 95% probability that your average winnings would fall between \$4.7 and \$5 million. If you have had an advanced course in probability, you may recognize that this is nothing more than the law of large numbers.

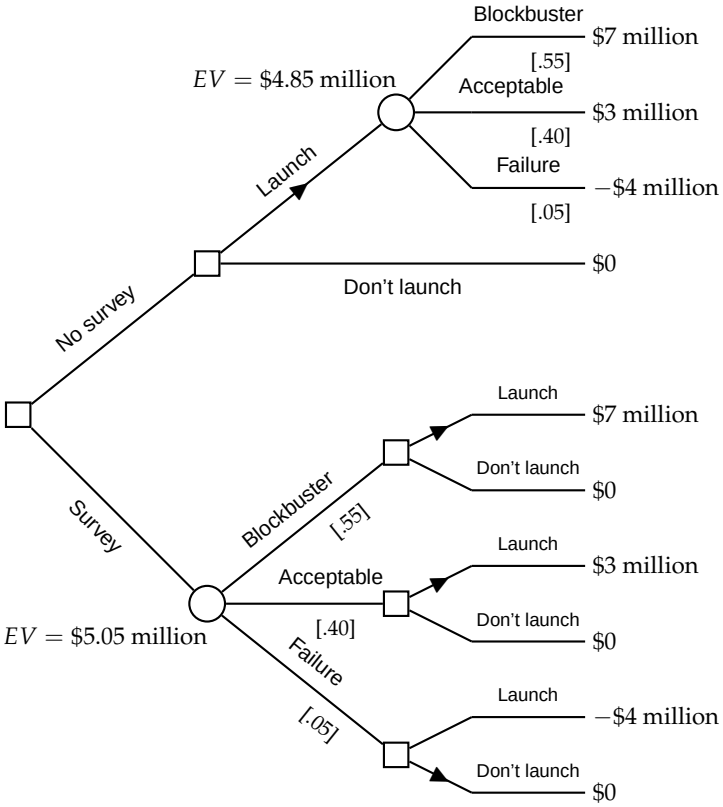

Figure 1.5: Now, the firm can conduct a survey, if it wishes, prior to making its launch decision.

As always, we solve the tree by working backwards. The top of tree is the same as Figure 1.4, so we know from our previous analysis that the firm would choose to launch. Its expected payoff is \$4.85 million. At the bottom of the tree, the rightmost nodes are all decision nodes. Note that they follow the chance node because, if the firm does a survey, it will know how successful its new product will be at the time it decides whether to launch. At the top two of these decision nodes, the firm would launch—positive amounts of money beat nothing. At bottom-most node it would not launch—\$0 beats suffering a loss. Note that the values of these three decision nodes have been appropriately labeled. When the firm decides to take a survey, it doesnt know what it will learn, so what it will learn is uncertain. This is reflected by the chance node that precedes the decision nodes. The expected value at that node is

$$\$5.05 \text{ million} = .55 \times \$7 \text{ million} + .4 \times \$3 \text{ million} + .05 \times \$0\,.$$

Comparing \$5.05 million to \$4.85 million, it follows that the firm would prefer to undertake the survey than make a decision without it.

This last example also illustrates how we can calculate the value of this survey: The value of the survey is the difference in the expected payoff with the survey, \$5.05 million, and the expected payoff without the survey, \$4.85 million; that is, the survey is worth \$200,000. The firm would pay up to \$200,000 to have this survey conducted.

To summarize:

Solving decision trees for expected-value maximizers:

1. For each of the rightmost nodes proceed as follows:
   (a) If the node is a decision node, determine the best alternative to take. The payoff from this alternative is the *value of this decision node*. Arrow the best alternative.
   (b) If the node is a chance node, calculate the expected value. This expected value is the *value of this chance node*.
2. For the nodes one to the left proceed as follows:
   (a) If the node is a decision node, determine the best alternative to take, using, as needed, the values of future nodes (nodes to the right) as payoffs. The payoff from this alternative is the value of this decision node. Arrow the best alternative.
   (b) If the node is a chance node, calculated the expected value, using, as needed, the values of future nodes (nodes to the right) as payoffs. The expected value is the value of this chance node.
3. Repeat Step 2 as needed, until the leftmost node is reached. Following the arrows from *left* to *right* gives the sequence of appropriate decisions to take.

1.6 Information

Often when we make a decision under uncertainty we would like more information about the uncertainty we face. You, for example, might read the prospectus for a security that you are considering purchasing, or you may ask a friend or a stock broker for advice about that security. We have, in fact, already seen a situation of information gathering. Recall Figure 1.5. In that tree, a firm was deciding whether to launch a new product or not. Prior to making this decision, the firm could, if it wished conduct a survey that would *perfectly* reveal how successful the product would be. Without conducting a survey, the firm would have to make its launch decision "in the dark." This is an example of a decision maker (*e.g.*, the firm) deciding whether to acquire additional information before making a decision. Note that whether to acquire additional information is, itself, a decision; one that we will study in this section.

Perfect information: *Information that completely reveals what will happen; that is, information that once learned means there is no longer uncertainty about a given event.*

Information is divided into two classes: perfect and imperfect. *Perfect information*, like that in Figure 1.5, is information that completely reveals in advance the outcome of some future uncertain event. In that figure, for instance, the survey completely revealed how successful the launch would be.

In other circumstances, we might expect to receive *imperfect information*. Imperfect information helps us to have a better idea of what will happen in the future, but it does not completely reveal what will happen.

Imperfect information: *Information that improves a decision maker's ability to predict the outcome of a future event, but which does not completely reveal what will happen.*

As an example of imperfect information, consider Figure 1.6, which illustrates the following situation: A firm uses sheet metal in making its product (*e.g.*, car parts). It is concerned with whether a recent shipment of sheet metal is up to its standards. It can use the sheet metal and do an entire production run, return the sheet metal to the supplier, or do a test run and then decide whether to do a production run or return the sheet metal. Unfortunately, a test run is not definitive as to whether the metal is up to the firms standards, although it gives indications. Let those indications be summarized as "likely okay" and "likely not okay." If the test run comes back "likely okay," then the probability that the metal is high quality is p and, hence the probability that the metal is low quality is $1 - p$. For reasons discussed below, we can assume $p > 1/2$. If the test run comes back "likely not okay," then the probability that the metal is high quality is $1 - p$ and the probability that it is low quality is p. Absent a test run, the probability that the metal is high quality is $1/2$ and, hence, the probability that it is low quality is $1/2$.

For our model of information to be consistent, it must be that that the expected probability of high quality prior to doing a test run is $1/2$ (taking an action can't change the underlying probabilities). Given our assumption about what the indicators mean, this consistency can be achieved only if the probability of getting the "likely okay" indicator is $1/2$ and, thus, the probability of the "likely not okay" indicator is $1/2$ (remember the probabilities at any chance node must sum to one).[5]

[5] If you want to see this demonstrated formally, let α be the probability of the "likely okay"

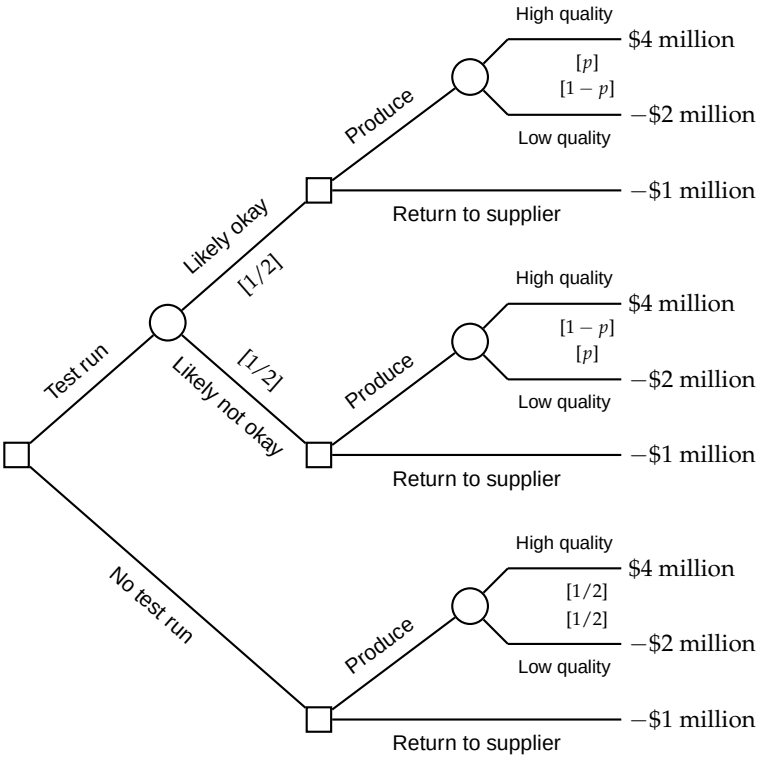

Figure 1.6: An example of *imperfect* information. By doing a test run, the firm gains information about how likely it is that the sheet metal is high quality.

Why can we assume $p > 1/2$? If the test run is informative, which we assume it is, then a result of "likely okay" must cause the firm's updated probability that the metal is high quality to be greater than if it had received no information. Since the probability of high quality absent information is $1/2$, the probability having seen the metal is "likely okay" must be greater than $1/2$; that is, $p > 1/2$. Similarly, if the test run comes back "likely not okay," then the firms updated probability that the metal is high quality is lower than if it had received no information; that is, $1 - p < 1/2$.

The variable p is, therefore, a measure of how informative the test run is. The closer p is to $1/2$, the less informative the test run is. Indeed, were $p = 1/2$, then there would be no information in a test run, because the outcome of a test run would not change the probability that metal is high quality. Conversely, the farther p is from $1/2$ (equivalently, the closer it is to 1), the more informative the test run is. Indeed, were $p = 1$, then we would have *perfect* information.

We solve the tree in Figure 1.6 like any other tree—starting at the right and moving left. The top rightmost chance node yields an expected value of

$$EV_{\text{top}} = p \times \$4 \text{ million} + (1 - p) \times (-\$2) \text{ million}.$$

Because $p > 1/2$, it is readily seen that $EV_{\text{top}} > \$1$ million. Consequently, at the top right decision node, the decision would be to produce. The value of that decision node is EV_{top}.

Going down to the bottom rightmost chance node, we see it has an expected value of

$$\frac{1}{2} \times \$4 \text{ million} + \frac{1}{2} \times (-\$2) \text{ million} = \$1 \text{ million}.$$

Given that a gain of \$1 million beats a loss of \$1 million, the correct decision at the preceding decision node is to produce; hence, the value of this decision node is \$1 million.

Now consider the middle rightmost chance node. It has an expected value of

$$\begin{aligned} EV_{\text{middle}} &= (1 - p) \times \$4 \text{ million} + p \times (-\$2) \text{ million} \\ &= \$4 \text{ million} - p \times \$6 \text{ million}. \end{aligned}$$

---

indicator and, thus, $1 - \alpha$ is the probability of "likely not okay." The ultimate probability that the metal is high quality prior to any testing is $\alpha p + (1 - \alpha)(1 - p)$ (this is the law of total probability, Proposition 48 on page 166). So

$$\alpha p + (1 - \alpha)(1 - p) = \frac{1}{2};$$

or, manipulating the expression algebraically,

$$2\left(p - \frac{1}{2}\right)\alpha - p + 1 = \frac{1}{2}.$$

Adding $p - 1$ to both sides yields

$$2\left(p - \frac{1}{2}\right)\alpha = p - \frac{1}{2},$$

which entails that $\alpha = 1/2$ as required.

The decision to take at the preceding decision node depends on the value of p: If $EV_{\text{middle}} \geq -\$1$ million, then the firm should produce. If $EV_{\text{middle}} < -\$1$ million, then the firm should not produce. When is $EV_{\text{middle}} \geq -\$1$ million? Answer: when

$$\$4 \text{ million} - p \times \$6 \text{ million} \geq -\$1 \text{ million};$$

or, dividing both sides by $-\$1$ million (because this is a negative quantity, it reverses the inequality sign), when

$$6 \times p - 4 \leq 1\,.$$

This further simplifies to $p \leq 5/6$. So when $p \leq 5/6$—6it is not super likely the metal is low quality even conditional on an indication of "likely not okay"—then the firm should produce even though the result of the test run is not promising. On the other hand, if $p > 5/6$—6it is very likely the metal is low quality conditional on the test run returning "likely not okay"—then the firm should not produce if the test run is not promising.

Working back to the first chance node, we find that the expected value of that node is

$$\frac{1}{2} \times EV_{\text{top}} + \frac{1}{2} \times \begin{cases} EV_{\text{middle}}\,, \text{ if } p \leq \frac{5}{6} \\ -\$1 \text{ million}\,, \text{ if } p > \frac{5}{6} \end{cases}\,.$$

Finally, we can decide what to do at the first decision node; that is, we can determine whether it is worthwhile to do a test run or not. Suppose first, that a test run is not especially informative, which, here, means $p \leq 5/6$. The value of choosing the "test run" branch is, then,

$$\begin{aligned} \frac{1}{2} \times EV_{\text{top}} + \frac{1}{2} \times EV_{\text{middle}} &= \frac{1}{2} \times \big(p \times \$4 \text{ million} + (1-p) \times (-\$2) \text{ million}\big) \\ &\quad + \frac{1}{2} \times \big((1-p) \times \$4 \text{ million} + p \times (-\$2) \text{ million}\big) \\ &= \frac{1}{2} \times \$4 \text{ million} + \frac{1}{2} \times (-\$2) \text{ million} \\ &= \$1 \text{ million}. \end{aligned}$$

Observe this is exactly the same as the expected payoff from choosing the "no test run" branch. In this case, then, *doing the test run creates* no *additional value*.

Suppose, in contrast, that a test run is very informative, which, here, means

$p > 5/6$. The value of doing the test run is, then,

$$\begin{aligned}\frac{1}{2} \times EV_{\text{top}} + \frac{1}{2} \times (-\$1) \text{ million} &= \frac{1}{2} \times \big(p \times \$4 \text{ million} + (1-p) \times (-\$2) \text{ million}\big) \\ &\quad + \frac{1}{2}(-\$1) \text{ million} \\ &= p \times \$3 \text{ million} - \$1.5 \text{ million} \\ &> \$1 \text{ million (because } p > \frac{5}{6}) \,.\end{aligned}$$

In this case, *doing the test run does create value*.

Why were the two cases so different? That is, why was there value to doing a test run when $p > 5/6$ but not otherwise? The answer has to do with the difference in how the firm responds to "likely not okay" in the two cases. When $p \leq 5/6$, the firm produces even if the result from the test run is "likely not okay." This is also the action it takes absent any information (*i.e.,* at the bottom decision node) and if the result is "likely okay." That is, when $p \leq 5/6$, the information learned from the test run has no potential to affect the firms action—when $p \leq 5/6$, the firm always produces. In contrast, when $p > 5/6$, then the firm does not produce if the result from the test run is "likely not okay." That is, when $p > 5/6$, the information learned from the test run *does have* the potential to affect the firms action—the firm takes *different* actions depending on the result of the test run. This reflects a general result about the value of information:

Conclusion (The fundamental rule of information)**.** *Information has value* only if *it has the potential to affect a decision maker's choice of action.*

When $p \leq 5/6$, the information has no potential to affect the firm's decision, but when $p > 5/6$, the information has the potential to affect the firm's decision. This is why the information is valueless when $p \leq 5/6$, but valuable when $p > 5/6$.

How valuable is the information when it is valuable? To find out, we subtract the expected value of not doing the test run from the expected value of doing the test run:

$$\begin{aligned}&\underbrace{p \times \$3 \text{ million} - \$1.5 \text{ million}}_{EV \text{ if do test run}} - \underbrace{\$1 \text{ million}}_{EV \text{ don't do test run}} \\ &\qquad = p \times \$3 \text{ million} - \$2.5 \text{ million}\end{aligned}$$

(for $p > 5/6$). Figure 1.7 plots the value of information for p between $1/2$ and 1. Note that the value is zero for p between $1/2$ and $5/6$ and is increasing (upward sloping) for p between $5/6$ and 1. The information is most valuable when $p = 1$, which makes sense: We would expected perfect information (which is what $p = 1$ represents) to be more valuable than imperfect information. This, too, is a general result:

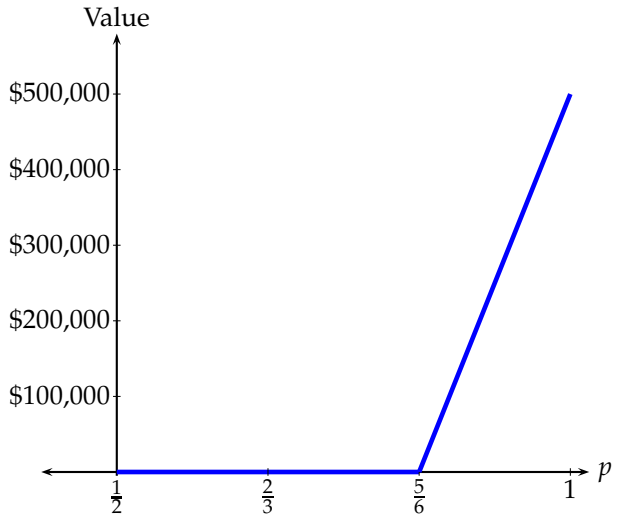

Figure 1.7: A plot of the value of information. Information is valueless for $p \leq 5/6$. It is valuable for $p > 5/6$. Perfect information (*i.e.*, $p = 1$) maximizes the value of the information.

Conclusion. *Perfect information is always at least as valuable as imperfect information.*

Figure 1.8 repeats Figure 1.6, except now the payoff if the metal is returned to the supplier is left as a variable, z, and the value of p is fixed at $5/6$. What we want to do now is understand how our conclusions change with z. In particular, we want to see how the value of information changes as z changes. Note that the expected values have been written in for each of the rightmost chance nodes. From this information, we see that we want to divide our analysis into four regions:

1. $z \leq -\$1$ million;
2. $-\$1$ million $< z \leq \$1$ million;
3. $\$1$ million $< z \leq \$3$ million; and
4. $\$3$ million $< z$.

In region 1, the decision to make at each of the three rightmost decision nodes is "produce." Moreover, this is the decision regardless of the information learned. Since the information, therefore, has no potential to change the firm's action, the information must be worthless (this is just the Fundamental Rule of Information).

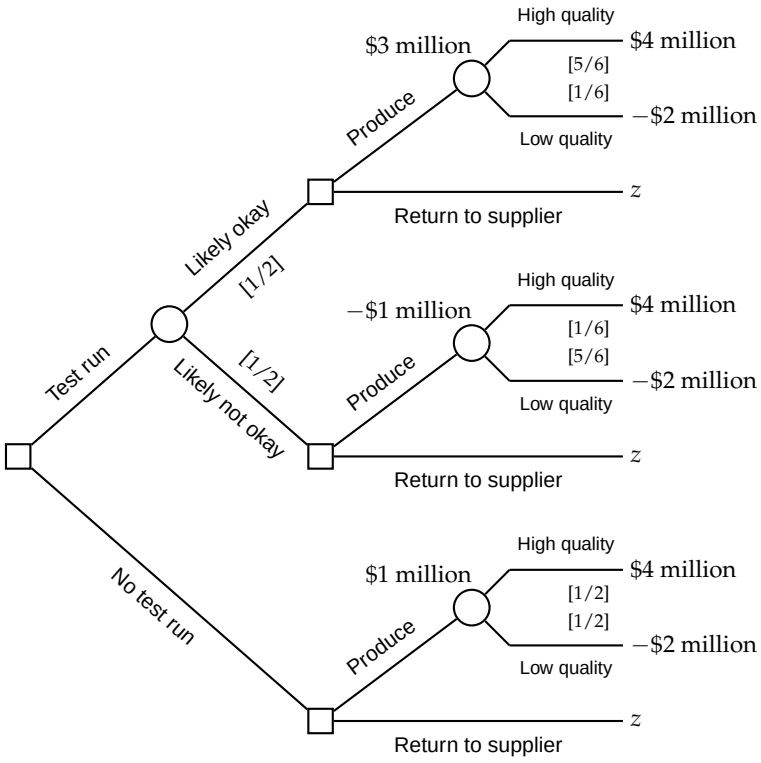

Figure 1.8: More on the value of information

In region 2, the firm produces at the top and bottom decision node, but returns the metal to the supplier at the middle node. The information has, therefore, the potential to change the firm's action, so we know it has value. Specifically,

$$\begin{aligned}\text{Value of information in region 2} &= \frac{1}{2} \times \$3 \text{ mil.} + \frac{1}{2} \times z - \$1 \text{ mil.} \\ &= \frac{1}{2} \times z + \$500,000\,.\end{aligned}$$

Note that, in region 2, the value of information is increasing in z.

In region 3, the firm produces only at the top decision node, but returns the metal to the supplier at the bottom two nodes. Again, the information has the potential to change the firms action, so we know it has value. Specifically,

$$\begin{aligned}\text{Value of information in region 3} &= \frac{1}{2} \times \$3 \text{ mil.} + \frac{1}{2} \times z - z \\ &= \$1,500,000 - \frac{1}{2}z\,.\end{aligned}$$

Note that, in region 3, the value of information is decreasing in z.

Finally, in region 4, the firm always does best to return the metal to the supplier. Moreover, this is the decision regardless of the information learned. Since the information has, therefore, no potential to change the firm's action, the information must be worthless.

Figure 1.9 plots the value of the information as a function of z. From the figure, it is clear that the information is most valuable when $z = \$1$ million. What's significant about \$1 million? It's the value of z that makes the firm indifferent between producing and returning the metal when it has *no* information. This is not a fluke, but an illustration of a general result:

Conclusion. *The value of information is maximized when the decision maker would,* absent the information*, view alternative actions as equally attractive.*

1.7 Real Options

In our discussion of information, we have so far treated information as something actively sought by the decision maker (*e.g.*, by doing a survey or a test run). Another way a decision maker can often acquire information is by waiting; that is, the passage of time resolves some of the uncertainty. The ability to use information learned over time can create certain options for the decision maker. To distinguish these options from financial options, they are often referred to as *real options*.

The main idea is that before you commit to an irreversible action, you have the option to commit or not to commit (just as with an in-the-money call option, where you have option to exercise or wait).

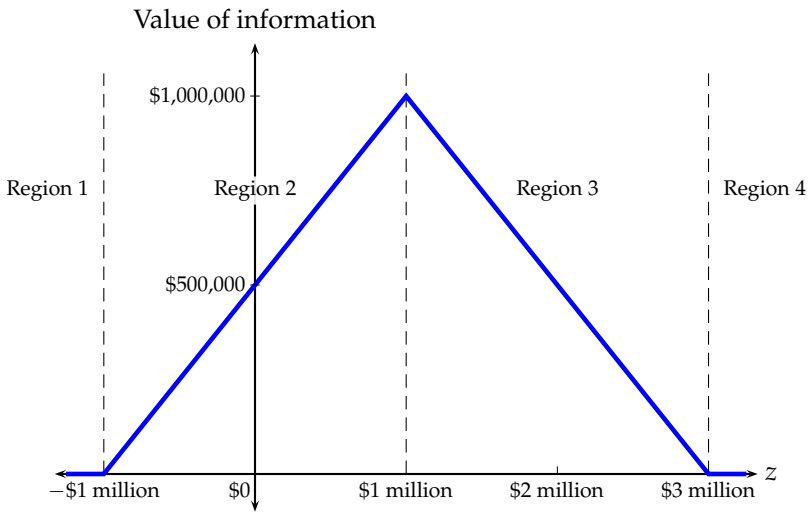

Figure 1.9: A plot of the value of information against the value of returning the metal to the supplier, z. (Note horizontal and vertical scales are different.)

Figure 1.10 illustrates a real option (note its resemblance to Figure 1.5). A firm can delay its launch of a new product or not delay. The success of the product depends on the growth rate of the economy. If the economy enjoys a high rate of growth, then the firm stands to earn $20 million. If the rate of growth is moderate, then the firm makes $5 million. Finally if there is low or no growth, then firm will lose $5 million. Observe from the figure that delay allows the firm to learn the state of the economy before making its launch decision. Delay is not without cost, however. If it delays, then the payoffs should it launch are all reduced by D, where $D \geq 0$.

From Figure 1.10, we see that the firm should delay if

$$\$6 \text{ million} - \frac{3}{5}D \geq \$4 \text{ million};$$

or, solving this last expression for D, if

$$\$3\frac{1}{3} \text{ million} \geq D.$$

Observe there is a difference between information gained by waiting and information bought prior to a decision (*e.g.*, a survey). Suppose that rather than the delay/no delay decision, the initial decision was forecast/no forecast, where forecast is an econometric forecast of the economy's future state. Let F be the cost of the forecast. Because F is spent regardless of whether the firm ultimately launches or not, the expected value of doing the survey is \$6 million$F$.

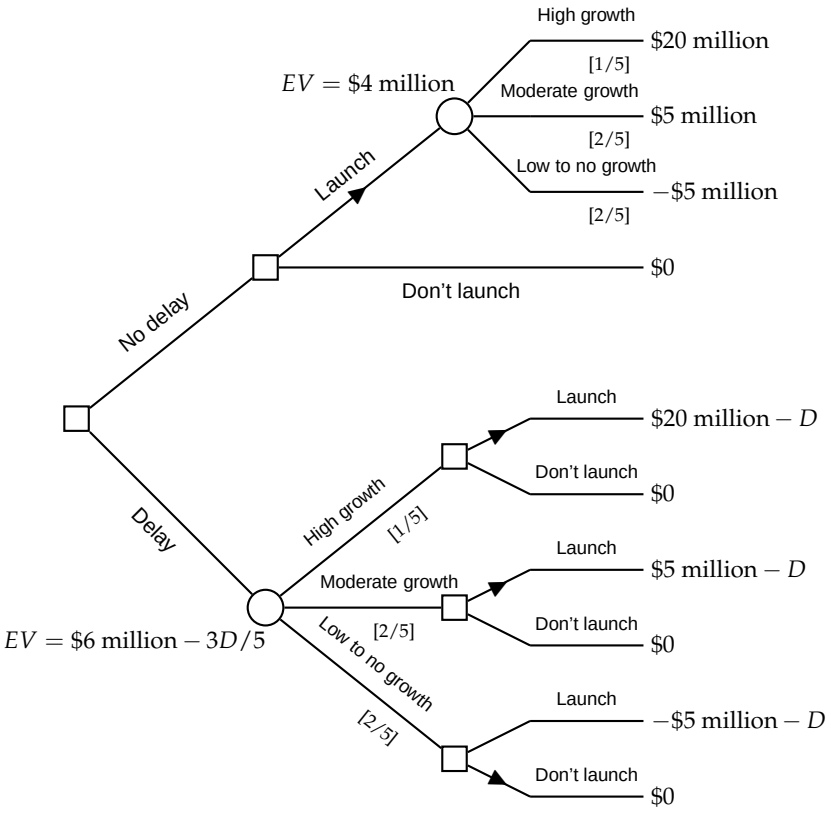

Figure 1.10: A firm has the *option* of delay.

So it pays to do the forecast only if it costs no more than $2 million. In other words, because the cost of delay is not always borne (*e.g.,* it is not borne if the firm decides not to launch), whereas the cost of *ex ante* information (*e.g.*, a forecast) is always borne, the cost of delay that a decision maker will accept is greater than the cost he or she will accept to obtain the information in advance (*e.g.,* through a survey or forecast).

1.8 Attitudes Towards Risk

Consider the following situation. You own a "lottery" ticket with the following properties. Tomorrow, a fair coin will be flipped. If it lands heads up, you will receive $1 million. If it lands tails up, you will receive nothing. The ticket is transferable; that is, you can give or sell it to another person, in which case this other person is entitled to the winnings from the lottery ticket (if any). Between today and tomorrow, people may approach you about buying your ticket. What is the smallest price that you would accept in exchange for your ticket?

A possible answer is $500,000 because that is the expected value of this lottery. Many people, however, would be willing to accept less than $500,000. You, for instance, might be willing to sell the ticket for as little as $400,000, under the view that a certain $400,000 was worth the same as a half chance at $1 million. Moreover, even if you are unwilling to sell your ticket for $400,000, many people in your situation would be.

Selling your ticket for less than $500,000 is, however, inconsistent with being an expected-value maximizer, because you wouldnt be choosing the alternative that yielded you the greatest expected value. So, if you would, in fact, accept $400,000, then you are not an expected-value maximizer. Moreover, even if *you* are (*i.e.,* you wouldnt sell for less than $500,000), others definitely aren't. As this discussion makes clear, we need some way to model decision makers who arent expected-value maximizers. That is the task of this section.

One reason that someone is not an expected-value maximizer is that he is concerned with the riskiness of the gambles he faces. For instance, a critical difference between an expected payoff of $500,000 and a certain payoff of $400,000 is that there is considerable risk with the former—you might win $1 million, but you also might end up with nothing—but there is no risk with the latter. Most people don't like risk, and they are willing, in fact, to pay to avoid it. Such people are called *risk averse*. By accepting less than $500,000 for the ticket, you are effectively paying to avoid risk; that is, you are behaving in a risk-averse fashion. When you buy insurance, thereby reducing or eliminating your risk of loss, you are behaving in a risk-averse fashion. When you put some of your wealth in low-risk assets, such as government-insured deposit accounts, rather than investing it all in high-risk stocks with greater expected returns, you are behaving in a risk-averse fashion.

Risk aversion: *A distaste for risk.*

To define risk aversion more formally, we begin with the concept of a *certainty-*

equivalent value:

Definition. *The* certainty-equivalent value *of a gamble is the minimum payment a decision maker would accept, if paid with certainty, rather than face a gamble. The certainty-equivalent value is often abbreviated CE.*

For example, if \$400,000 is the smallest amount that you would accept in exchange for the lottery ticket discussed previously, then your certainty-equivalent value for the gamble is \$400,000 (i.e., *i.e.*, $CE = \$400,000$).

We can now define risk aversion formally:

Definition. *A decision maker is* risk averse *if his or her certainty value for any given gamble is less than the expected value of that gamble. That is, we say an individual is risk averse if $CE \leq EV$ for all gambles and $CE < EV$for at least some gambles.*

Risk neutral: *Unaffected by risk.*

In contrast, an expected-value maximizer is risk neutral—his or her decisions are unaffected by risk. Formally,

Definition. *A decision maker is* risk neutral *if the certainty-equivalent value for any gamble is equal to the expected value of that gamble; that is, he or she is risk neutral if $CE = EV$ for all gambles.*

In rare instances, a decision maker likes risk; that is, he or she would be willing to pay to take on more risk (or, equivalently, require compensation to part with risk). For instance, if someone's certainty-equivalent value for the aforementioned lottery ticket were \$600,000 (*i.e.*, she had to be compensated for giving up the risk represent by the ticket), then she would be called *risk loving*. It is worth emphasizing that risk-loving behavior is fairly rare.[6]

At this point you might ask: When is it appropriate (*i.e.*, reasonably accurate) to assume a decision maker is risk neutral and when is it appropriate to assume he is risk averse? Some answers:

SMALL STAKES VERSUS LARGE STAKES: If the amounts of money involved in the gamble are small *relative* to the decision maker's wealth or income, then his behavior will tend to be approximately risk neutral. For example, for gambles involving sums less than \$10, most people's behavior is approximately risk neutral. On the other hand, if the amounts of money involved are large *relative* to the decision maker's wealth or income, then his behavior will tend to be risk averse. For example, for gambles involving sums of more than \$10,000, most people's behavior exhibits risk aversion.

SMALL RISKS VERSUS LARGE RISKS: If the possible payoffs (or at least the most likely to be realized payoffs) are close to the expected value, then the risk

[6] Although it is true that a number of people like to pay to gamble on occasion (*e.g.*, they visit casinos or play state lotteries), their more typical behavior can be described as risk neutral or risk averse (*e.g.*, even people who visit casinos typically purchase homeowner's insurance). Moreover, it is not clear that people gamble because they love the risk *per se*; they may simply like the excitement of the casino or like to dream about what they would do if they won the next Powerball drawing.

is small and the decision maker's behavior will be approximately risk neutral. For instance, if the gamble is heads you win \$500,001, but tails you win \$499,999, then your behavior will be close to risk neutral since both payoffs are close to the expected value (*i.e.*, \$500,000). On the other hand, if the possible payoffs are far from the expected value, then the risk is greater and the decision maker's behavior will tend to be risk averse. For instance, we saw that we should expect risk-averse behavior when the gamble was heads you win \$1 million, but tails you win \$0.

DIVERSIFICATION: So far the question of whether someone takes a gamble has been presented as an all-or-nothing proposition. In many instances, however, a person purchases a portion of a gamble. For example, investing in General Motors (or any other company) is a gamble, but you don't have to buy all of General Motors to participate in that gamble. Moreover, at the same time you buy stock in General Motors, you can purchase other securities, giving you a portfolio of investments. If you choose your portfolio wisely, you can diversify away much of the risk that is unique to a given company. That is, the risk that is unique to a given company in your portfolio no longer concerns you—you are risk neutral with respect to it. Consequently, you would like your firm to act as an expected-value maximizer. As we discuss in next, diversified decision makers are risk neutral (or approximately so), while undiversified decision makers are more likely to be risk averse.

Diversification

To clarify the issue of diversification, consider the following example. There are two companies in which you can invest. One sells ice cream. The other sells umbrellas. Ice cream sales are greater on sunny days than on rainy days, while umbrella sales are greater on rainy days than on sunny days. Suppose that, on average, one out of four days is rainy; that is, the probability of rain is 1/4. On a rainy day, the umbrella company makes a profit of \$100 and the ice cream company makes a profit of \$0. On a sunny day, the umbrella company makes a profit of \$0 and the ice cream factory makes a profit of \$200. Suppose you invest in the umbrella company only; specifically, suppose you own all of it. Then you face a gamble: on rainy days you receive \$100 and on sunny days you receive nothing. Your expected value is

$$\$25 = \frac{1}{4} \times \$100 + \frac{3}{4} \times \$0\,.$$

Suppose, in contrast, that you sell three quarters of your holdings in the umbrella company and use some of the proceeds to buy one eighth of the ice cream factory. Now on rainy days you receive \$25 from the umbrella company (since you can claim one quarter of the \$100 profit), but nothing from the ice cream company (since there are no profits). On sunny days you receive \$25 from the ice cream company (since you can claim one eighth of the \$200 profit), but nothing from the umbrella company (since there are no profits).

That is, rain or shine, you receive \$25—your risk has disappeared! Your expected value, however, has remained the same (*i.e.*, \$25). This is the magic of diversification.

Moreover, once you can diversify, you want your companies to make expected-value- maximizing decisions. Suppose, for instance, that the umbrella company could change its strategy so that it made a profit of \$150 on rainy days, but lost \$10 on sunny days. This would increase its daily expected profit by \$5—the new EV calculation is

$$\frac{1}{4} \times \$150 + \frac{3}{4}(-\$10) = \$30\,.$$

It would also, arguably, increase the riskiness of its profits by changing its strategy in this way. Suppose, for convenience, that 100% of a company trades on the stock exchange for 100 times its expected daily earnings.[7]

The entire ice cream company would, then, be worth \$15,000 ($= 100 \times (1/4 \times \$0 + 3/4 \times \$200)$) and the entire umbrella company would, then, be worth \$3000. To return to your position of complete diversification and earning \$25 a day, you would have to reduce your position in the umbrella company to hold one sixth of the company and you would have to increase your holdings of the ice cream company to 2/15th of the company:

$$\text{Earnings on a rainy day} : \frac{1}{6} \times \$150 + \frac{2}{15} \times \$0 = \$25\,;$$

and

$$\text{Earnings on a sunny day} : \frac{1}{6} \times (-\$10) + \frac{2}{15} \times \$200 = \$25\,.$$

Going from holding one fourth of the umbrella company to owning one sixth of the umbrella company means selling 1/12th of the umbrella company,[8] which would yield you \$250 ($= 1/12 \times \$3000$). Going from holding one eighth of the ice cream company to owning 2/15ths means buying an additional 1/120th of the ice cream company,[9] which would cost you \$125 ($= 1/120 \times \$15,000$). Your

[7] The price-to-earnings ratio is 100 here, but the value of the price-to-earnings ratio does not matter for the conclusions reached here. If the ratio were r, then decreasing your holdings of the umbrella company to 1/6th of the company and increasing your holdings of the ice cream company to 2/15th would yield a trading profit of

$$\frac{30r}{12} - \frac{150r}{120} = \frac{5}{4}r > 0\,.$$

Now you might wonder whether it is appropriate to use the same price-to-earnings ratio for both firms. In this case it is, at least if you believe that the stock price is driven by fundamentals (that is, future profits).

[8] Because $\frac{1}{4} - \frac{1}{6} = \frac{3}{12} - \frac{2}{12} = \frac{1}{12}$.

[9] Because $\frac{2}{15} - \frac{1}{8} = \frac{16}{120} - \frac{15}{120} = \frac{1}{120}$.

profit from these stock market trades would be $125. Moreover, you would still receive a riskless $25 per day. So because you can diversify, you benefit by having your umbrella company do something that increases its expected value, *even if it is riskier.*

Decision Making by Risk-Averse Decision Makers

Analyzing the behavior of risk-averse decision makers fully is beyond the scope of these notes. This is, however, a topic of importance, especially in finance, insurance, and the setting of incentive compensation. Finance courses, among others, provide tools for understanding the behavior of risk-averse decision makers.

1.9 Summary

This chapter began with a discussion of a general problem-solving method, *fishbone analysis*.

We then turned to decision making under certainty. Such decision problems could be represented by *decision trees* consisting of decision nodes (squares), branches for the alternatives, and payoffs. These trees, *like all trees*, were solved by working from right to left (although the tree is read left to right).

Next we added the possibility of uncertainty. We indicated uncertainty in our trees by using chance nodes (circles). The branches stemming from a chance node are the possible outcomes of the random event that the chance node represents. We practiced solving trees with uncertainty under the assumption that the decision makers were *expected-value maximizers*.

We next analyzed the value of information in decision-making problems. Three points were made: (1) Whether or not to seek more information is, itself, a decision that needs to be considered as part of the overall decision making process; (2) information is valuable only if it has the potential to change decisions; (3) information is maximally valuable when, absent the information, the decision maker is indifferent among his options.

We observed that delay is often a way to obtain information. In many ways it is similar to obtaining information ex ante (consider, *e.g.,* the similarities between Figures 1.5 and 1.10). However, because the cost of delay is not always borne, decision makers will generally be willing to pay a higher cost in terms of delay than they will to obtain the information *ex ante*.

Finally, we considered the issues of attitudes toward risk. While many decision makers are *risk averse*, we also saw that there were conditions under which *risk neutrality* could be assumed. One condition was when the decision maker was *diversified*. We saw that, once diversified, an individual wants the firms in which he or she invests to behave as expected-value maximizers.

Costs | 2

What is the cost of an activity (*e.g.*, providing a good or service)? The obvious answer might seem "the money spent on that activity." The purpose of this chapter is to convince you that, like so many "obvious" answers, this is not (necessarily) the correct answer.

Opportunity Cost | 2.1

The correct way to think about cost from the perspective of making good business decisions is to consider the cost of some activity (alternatively, good, decision, service, etc.) to be the value of the most highly valued forgone activity (*i.e.*, the value of the best alternative decision). Economists describe this way of viewing cost as considering the *opportunity cost* of an activity or decision. In other words, the cost of something is the value of the best opportunity given up.

Opportunity cost: *The value of the most highly valued forgone activity or use of a good.*

In many circumstances, the opportunity cost notion coincides with what might be termed the naïve view of cost, namely it's your expenditure on the activity in question. If, for instance, you purchase three tons of grapes for your wine-making business, then the cost of the grapes is the amount you pay for those grapes.

This naïve view, focusing on expenditures, can, however, lead one to make bad decisions. For instance, suppose you purchased 1000 liters of a chemical that is used in your production process. Suppose you paid a price of $10 per liter. Suppose, however, that after purchase, but before use, the price of this chemical jumps to $11 on the open market. The cost if you use this chemical is not your expenditure, $10,000, but rather $11,000, the current value of the 1000 liters. Why? Because your next best use of the chemical is to sell it on the open market and the value of this next-best activity is $11,000. You can see this by supposing that the products you would make from the 1000 liters would sell for a total of $10,500. If you produce, youll have $10,500 in the bank and you might, incorrectly, suppose yourself to have made a $500 profit. In fact, youve suffered a $500 loss: Had you sold the chemical, you would have had $11,000 in the bank and $11,000 beats $10,500. From this, we see both the danger of the naïve view and the benefit of adopting the opportunity-cost view.

This scenario also illustrates two concepts that derive from opportunity cost, *imputed cost* and *sunk expenditure*.

Imputed cost: *The imputed value of a forgone opportunity.*

Sunk expenditure: *An expenditure is sunk if it cannot be recovered or avoided over the relevant decision-making horizon. A sunk expenditure is* not *a cost.*

An imputed cost (*e.g.*, the \$11,000 in the above scenario) is a cost not associated with an expenditure; that is, you dont have to pay anyone \$11,000 to use the chemical, but it's a cost of your using the chemical nonetheless.

A sunk expenditure is an expenditure that has been made—sunk—in the past or an expenditure that one will make regardless of which of the relevant alternatives is chosen (*e.g.*, because of a previously made commitment).[1]

In particular, what you paid for machinery, property, services, inputs, or any other item in the past is not a cost todayit is sunk. Moreover, a past expenditure is irrelevant to any decision you are making now. As the old saying goes, "there is no point crying over spilled milk." Or, in terms of decision trees, one can view a past expenditure as a constant amount subtracted from *all* the payoffs. Given the discussion in the previous chapter, you should readily be able to convince yourself that decisions don't change if one subtracts the same amount from all payoffs.

Note, however, an expense doesn't have to have been made in the past to be sunk. Any expense, even one not yet paid, is sunk if it cannot be avoided given the relevant set of actions from which the decision maker must choose.

Some examples will further illustrate the benefits of adopting the opportunity-cost view.

> **Example 2:** Your store is situated in rented space. Your lease expires in six months time and you cannot break it prior to then. Monthly rent is \$5000. Your *other* monthly expenditures (*e.g.*, inventory, sales help, etc.) are \$3000, for total monthly expenditures of \$8000. Your monthly revenue (sales) total \$6000. You are considering whether to shutdown effective immediately. This might seem an attractive course of action: After all it would seem youre "losing" \$2000 a month. This, however, is an incorrect assessment of the situation. What is the opportunity cost of staying open? Well, if you stay open, your expenditures will be \$8000 per month. If you shutdown, your expenditures will be \$5000 a month (remember you can't break your lease). Thus, the true (opportunity) cost of staying open is \$3000 ($= \$8000 - \$5000$): Relative to your best alternative (shutting down), you are only forgoing \$3000 by staying open. Given that your true costs are half of your monthly revenues, you would choose to remain open over the next six months. To see it another way, note that your bank account will be \$30,000 poorer at the end of six months if you shutdown ($6 \times (\$5000)$); whereas, if you remain open, your bank account will be only \$12,000 poorer at the end of six months ($6 \times (\$6000 - \$5000 - \$3000)$).

By employing the concept of opportunity cost you saved yourself \$18,000 in the preceding example. This is because you recognized that the rent over the next six months was not a cost, but rather a sunk expenditure.

> **Example 3:** You run a theater. At the moment, the show that you're running sells on average 100 tickets per night (you are risk neutral). The ticket

[1] An expenditure that is sunk is sometimes called a *sunk cost*. This, however, is unfortunate terminology, because if an expenditure is sunk than it is not a cost with regard to the decision at hand.

price is $50 per ticket. The nightly expenses of the show are $3000 and these are essentially independent of the number of people who attend. A local civic group approaches you about purchasing all the tickets in the house (300) for a given evening. Because they are a charitable group, they ask if you might be willing to sell them the tickets at a discount of $25 per ticket.

What is your cost of doing this? First, recognize the $3000 is wholly irrelevant. Why? Well whether or not you sell to this group, the show will run. The expenses of the show are, thus, sunk, and, thus, not a cost. The cost, remember, is the value of what you give up. What you give up are the 100 tickets sold at $50 per ticket; that is, $5000. So the cost of your donation is $5000.

Observe that, in both examples, the answer is dependent on the decision problem you face. In Example 2, over the *relevant decision-making horizon*— the next six months—the monthly rent is sunk. Were you, however, deciding to renew the lease at the end of the six months, then the rent would be a cost because you could avoid it by shutting down (*i.e.,* by not *renewing* the lease). Likewise, in Example 3, you are already committed to running the show. Hence, any expenses associated with running the show are sunk with respect to other decisions, such as whether to sell to the civic group at a discount. Had, however, the decision been a different one, specifically whether or not to close the show, then the nightly expenses of running the show would be a cost because they could then be avoided. Both of these examples illustrate one test for whether an expense is a cost: Is there a branch (outcome) of the relevant decision tree in which you avoid the expense? If the answer is no, then the expense cannot be a cost.

Conclusion. *If, within the set of relevant decisions, an expense is unavoidable, then it is* not *a cost.*

Example 4: Your company owns a warehouse. Currently, the warehouse is in such poor condition that it cannot be used. Your company would have to spend $500,000 to make it usable. If made usable, the value to your company, if it uses it over its productive life, is $300,000. You are also aware that another company, in a different industry than you, is looking to purchase a warehouse and would certainly purchase your warehouse if it were usable (note, then, this other company would not buy your warehouse "as is"). In thinking about what sales bid to make to this other company, is the $500,000 a cost; that is, relevant to your decision? What would your answer be if the value to your company of a usable warehouse was $600,000?

Under the original conditions (*i.e.,* the value to you of repaired warehouse is $300,000), the $500,000 is most certainly a cost. If you sell your warehouse you will have to spend $500,000. If, instead, you pursue your next best alternative— which is do nothing with the warehouse—you don't spend the $500,000. On the other hand, if the value to your company of a repaired warehouse were $600,000, then the $500,000 is not a cost; you will incur the expense if you sell the warehouse or if you pursue your next best alternative, which, now, is to use the warehouse yourself.

Example 4 suggests another test to see whether an expense is a cost, namely the question of *cost causation*:

Conclusion (Cost causation)**.** *If a decision causes you to incur an expense that you wouldn't incur under the next best alternative, then that expense is a cost; it has been caused by your decision. Conversely, if an expense would have also been incurred under the next best alternative, then it is not a cost of your decision.*

As an additional example, consider the following.

> **Example 5:** You may have noticed or read that when oil prices jump up (say because of Mideast unrest), prices at your local gas station quickly follow suit. Yet, as newspaper articles are quick to point out, the gasoline in your local gas station's tanks was produced from oil purchased at the old, lower, prices. Is your local gas station engaged in price gouging, as some allege, or is there a sensible opportunity-cost explanation?
>
> It won't surprise you that the latter is the answer. One way to think about it is as follows. The wholesale price of new gasoline will be higher than the wholesale price previously paid.[2] "Old" and "new" gasoline are perfect substitutes, so your local gas station could, in theory, resell the gas in its tanks to other stations in the wholesale market. In other words, the situation is like the chemical example considered at the beginning of this section.
>
> Even if reselling the gasoline in the wholesale market isn't feasible, the opportunity cost of selling the gasoline already in the tanks has increased. A rise in oil prices must, ultimately, lead to an increase in gasoline prices. Gasoline today is a substitute for gasoline tomorrow. So an alternative available to your local station is to choose not to sell gasoline now and to wait until the price goes up, at which time it can sell the gasoline it holds at the higher price.

Example 5 illustrates an example of decision makers' thinking about costs correctly. Yet, there are many cases in which business people fail to do so. One situation is when business people consider historic prices, such as those paid for inputs, as relevant for current decision making; that is, failing to recognize them as sunk. For instance, a well-known computer manufacturer used the price it had paid for the memory chips in its inventory when it priced its computers, ignoring the fact the price of these memory chips had fallen considerably from the time they had been purchased. By using historic price, rather than the chips' current value, the manufacturer overestimated its true costs. This, in turn, caused it to overprice its computers. As a consequence, it lost a considerable amount of market share and profit. In another example, a well-known car manufacturer, seeking a competitive advantage, hired people to forecast metal prices. The idea was that it would buy and stockpile metals whose prices were forecast to increase. Then, later, it would have "cost advantage" over its rivals when it used that metal in manufacturing. While

[2]The wholesale price is what the gas station pays its supplier for the gasoline it purchases to sell to consumers.

speculating in commodity markets, such as metals, might be an okay investment strategy for the firm, it can't provide any cost advantage if the metal is ultimately used in manufacturing—the cost of the metal is its market value at the time it is used, not what was paid for it in the past.

As a final example of opportunity costs at work, consider the following.

> **Example 6:** A large conglomerate knows now that it will need to send 50 mid-level executives to a two-day convention in Miami in February.[3] That is high season and hotels are routinely filled to capacity. Any hotel suitable for these executives charges $250 a night per guest. Seeking to avoid paying $25,000, someone in the conglomerate notes that it happens to own a suitable hotel in Miami. Said person proposes that the 50 executives just stay at that hotel for free; further claiming that the cost will be only the cost of room maintenance, which is $50 per night; that is, cost will be only $5000.
>
> Whatever the advantages of a conglomerate, this is not one. This someone has made a fundamental error. Recall that hotels sell out. Hence, what the company is forgoing by lodging its executives at one of its hotels is not $50 per night (it would pay that whether an executive is lodged in a room or an outside guest is), but rather $250, the forgone revenue from renting the room to an outside guest. That is, the cost per night is still $250—no savings can be achieved by this proposal.

Example 6 illustrates a rather common mistake, assuming that because one owns an asset, its use is free. Typically, unless there is no alternative use of the asset, its use is not free. The cost of using it is the value of its next best use (in Example 6, for instance, renting the rooms to outside guests).

Cost: *Expenditures plus imputed costs less sunk expenditures.*

A formula that can help relate expenditures to costs is the following:

$$\text{Cost} = \text{Expenditures} - \text{Sunk Expenditures} + \text{Imputed Costs}\,.$$

2.2 Cost Concepts

Consider a firm that makes a single product. Suppose that every morning, prior to the start of production, the machines used for production must be maintained. Suppose this costs $100 per day. Suppose each unit of the product requires $5 worth of raw materials, $8 worth of labor, and will cost $2 to ship. What is the cost of selling a single unit of the product? The answer depends on the decision-making horizon: If it is mid-day and the firm has been producing all morning, then the cost of a unit is $15 ($= \$5 + \$8 + \$2$); if, however, it is the *first* unit of the morning, then the cost is $115.

To understand these answers, recognize that the cost incurred by maintaining the machines is different than the costs incurred by producing an additional unit. Maintenance costs are, here, costs that are incurred once and that do not

[3] This is example is based on an actual incident.

Overhead cost: *A cost incurred by operating that does not vary with the number of units produced.*

Variable cost: *A cost that varies with the number of units produced.*

Average cost: *The total cost of production divided by the number of units produced.*

depend on the number of units produced. That is, they are incurred when the firm decides to produce that day (go from the 0th unit to the 1st unit). A cost that is incurred when a firm decides to produce rather than shutdown over the relevant decision-making horizon (here, a day) and that does not vary with the total number of units produced is an *overhead cost*.[4]

In contrast, the expenditures on raw materials, labor, and shipping are, here, examples of *variable costs*; that is, costs that vary with each unit produced.

Now these distinctions might strike you as odd: Why not simply use the total amount spent on producing a day's output divided by a day's output as the cost per unit of output (this measure is called *average cost*)?[5] For example, if this firm produced 20 units today, why not simply treat the cost per unit as \$20 $(= (20 \times \$15 + \$100)/20)$? There are two reasons why this is *a bad idea*. First, it depends on the number actually produced. If the firm's output varied, would you really want to think that its unit costs are varying even though the technology and input prices (*e.g.*, the cost of raw materials) had remained unchanged? The second, and more important reason is that this approach violates the fundamental notion of cost: The cost of an activity (*e.g.*, producing the 20th unit) is the value of the next best alternative (*e.g.*, stopping at the 19th unit). Here, if you stopped at the 19th unit you would save just the additional \$15 that you would need to spend to produce a 20th unit. Over the relevant decision-making horizon—make or don't make a 20th unit given that 19 units have already been made—the daily maintenance cost is sunk and, thus, not a cost over that decision-making horizon. To make this more concrete, consider the following example.

> **Example 7:** Suppose the firm under consideration gets an order to produce 20 units for \$21 per unit. The firm will accept this order because
>
> $$\underbrace{20 \times \$21}_{\text{revenue}} - \underbrace{(20 \times \$15 + \$100)}_{\text{cost}} = \$20\,;$$
>
> that is, the firm will make a positive profit. Suppose, after the firm has started production, someone else calls in an order for 10 units at \$18 per units. Should the firm accept this second order? Well if it mistakenly used average cost, \$20, as its cost per unit, it would reject the offer because it would "lose" \$2 per unit. If, however, it used the correct cost per unit, namely \$15, it would accept the offer because it would make a \$3 profit per unit. To verify that this is, indeed, the correct cost concept—that is, leads to the correct decision—observe that
>
> $$\underbrace{20 \times \$21 + 10 \times \$18}_{\text{revenue}} - \underbrace{\big((20+10) \times \$15 + \$100\big)}_{\text{cost}} = \$50 > \$20\,.$$

[4] Overhead costs are sometimes referred to as *fixed costs*. The latter is, however, unfortunate terminology: If some expenditure were truly fixed (*i.e.*, immutable), then it could not be a cost (it would be sunk).

[5] Observe that average cost, abbreviated AC, is cost, C, divided by the number of units produced, x. That is, $AC = C/x$.

This last example shows the importance of using *marginal cost*—the additional cost incurred by producing one more unit—in deciding how much to produce. In Example 7, the marginal cost of the 20th unit is \$15 because, as previously argued, \$15 is the additional cost incurred by producing the 20th unit. Note that we can equivalently define marginal cost as

Marginal cost: *The marginal cost of the nth unit is the additional cost incurred by producing the nth unit.*

$$\text{Marginal cost of } n\text{th unit} = \text{Cost of producing all } n \text{ units} - \text{Cost of producing } n-1 \text{ units}.$$

Rather than writing out "marginal cost of" and "cost of producing all," we will use the notation $MC(n)$ to denote the marginal cost of the nth unit and $C(n)$ to denote the cost of producing all n units. We will also refer to $C(n)$ as the *total cost* of producing n units.

Given the importance of the marginal cost concept, it is worth considering two more examples.

Example 8: Suppose the firm under consideration had to maintain the machines after every fifty units; that is, before producing the 1st, 51st, 101st, etc., units. Consequently, the marginal cost of the 1st unit is \$115, as is the marginal cost of the 51st unit, the 101st unit, etc. The marginal cost of all other units remains just \$15.

Example 9: Return to the assumption that the machinery only needs to be maintained once a day, first thing in the morning. Suppose, now, that if the firm produces 250 or more units, it needs to pay time-and-a-half overtime to its employees; that is, the cost of labor per unit rises to \$12 per unit. Then the marginal cost of the first unit is \$115. The marginal cost of the nth unit, $1 < n < 250$, is \$15. But now the marginal cost of the mth unit, $m \geq 250$, is \$19 (= \$5 in raw materials + \$2 in shipping + \$12 in labor).

2.3 Relations Among Costs

In this section we explore the relations among total cost, average cost, and marginal cost from an algebraic perspective.

First, observe that the total cost of producing nothing (*i.e.*, $C(0)$) must be zero. Why? If you're producing nothing, then you aren't forgoing anything by producing, so the cost must be zero.

Conclusion. $C(0) = 0$.

Turning into notation the definition of marginal cost given in the last section, we have

$$MC(n) = C(n) - C(n-1)\,. \tag{2.1}$$

Using expression (2.1), observe that we can express total cost as follows.

$$\begin{aligned} C(n) &= C(n) - C(0) && \text{(recall } C(0) = 0\text{)} \\ &= C(n) - \underbrace{C(n-1) + C(n-1)}_{+0} - \cdots - \underbrace{C(1) + C(1)}_{+0} - C(0) \\ &= \underbrace{C(n) - C(n-1)}_{MC(n)} + \cdots + \underbrace{C(1) - C(0)}_{MC(1)} && \text{(associative rule)} \\ &= MC(n) + \cdots + MC(1) = \sum_{j=1}^{n} MC(j)\,. \end{aligned}$$

We can summarize this as

Proposition 1. *The total cost of producing n units is the sum of the marginal costs of producing the first n units; that is, $C(n) = \sum_{j=1}^{n} MC(j)$.*

Recall that average cost of producing n units is

$$AC(n) = \frac{C(n)}{n}\,;$$

that is, average cost is total cost divided by the number of units produced. Using Proposition 1, we can rewrite this last expression as

$$AC(n) = \frac{\sum_{j=1}^{n} MC(j)}{n}\,. \tag{2.2}$$

A baseball analogy may help to explain expression (2.2). Think of $MC(j)$ as being 1 if a batter gets a hit on his jth at bat and as being 0 if he does not (to simplify matters, assume no walks, no errors, no hit batters, etc.). A batter's average, recall, is just the number of hits he gets divided by his number of at bats. It follows that expression (2.2) is the formula for his batting average.

We can also use this baseball analogy to consider how AC changes with MC. If a batter gets a hit at his latest at bat, his average goes up (a hit is like batting 1.000, which is greater than his average to that point). If he doesn't get a hit, his average goes down (failing to hit is like batting .000, which is less than his average to that point). In other words, we have

$$\begin{aligned} &MC(n+1) > AC(n)\,, \text{ then } AC(n+1) > AC(n)\,; \text{ and} \\ &MC(n+1) < AC(n)\,, \text{ then } AC(n+1) < AC(n)\,. \end{aligned}$$

Another way to state this conclusion is

Proposition 2. *If average cost is declining, then marginal cost is less than average cost. If average cost is increasing, then marginal cost is greater than average cost.*

2.4 Costs in a Continuous Context

For some goods (*e.g.*, liquids) a unit is a fairly arbitrary notion. Not only could we have liters versus quarts, say, but we can also half liters, quarter liters, and so forth. Likewise we could have tons, half tons, quarter tons, and so on of coal or wheat. For such goods it is useful to treat costs as continuous functions of output. Even for goods for which a unit is clearly defined (*e.g.*, shirts), it can prove convenient to act as if a continuum of units can be produced.

In such contexts, we treat the cost function, $C(x)$, as being a continuous function of output, x.[6,7] That is, we can talk of the cost of producing one liter, $C(1)$, the cost of producing 2.5 liters, $C(2.5)$, and so forth.

We define average cost as before, namely total cost divided by units. That is, $AC(x) = C(x)/x$. The average cost of 2.5 liters, $AC(2.5)$, would, thus, be $C(2.5)/2.5$. Note that because $C(\cdot)$ is a continuous function, $AC(\cdot)$ is also a continuous function.

Marginal cost is a bit trickier. The incremental cost of producing h more units of a good is easy enough to calculate, it is $C(x+h)C(x)$. Marginal cost, however, is a statement about the incremental cost per unit. That is, marginal cost can be seen, roughly, as the average of the incremental cost. That is,

$$MC(x) \approx \frac{C(x+h) - C(x)}{h}, \tag{2.3}$$

where the symbol $\approx$ means "approximately equal to."

An analogy may help. Suppose you have driven t hours. If you drive one hour more, then it is easy to calculate your speed:

$$\text{speed} = \frac{\big(D(t+1) - D(t)\big)\ \text{km}}{1\ \text{hour}},$$

where $D(\cdot)$ is your distance from your starting place as a function of time. Note the division by 1 hour, reflecting our desire to have speed measured as km per (whole) hour. Of course, we could also determine your speed after you've driven just a half hour:

$$\text{speed} = \frac{\Big(D(t+\tfrac{1}{2}) - D(t)\Big)\ \text{km}}{\tfrac{1}{2}\ \text{hour}}.$$

We divide by 1/2 hour because we want to express speed in terms of km per hour, not km per half hour. This explains why we divide by h in expression

[6] The cost function is sometimes called the *cost schedule*. In general, the word "schedule" is a synonym for "function."

[7] $\int dx$ As a technical matter, we also take $C(\cdot)$ to be differentiable except, possibly, at 0. that is, $C'(x)$ is defined for all $x > 0$.

(2.3), we want to express the increment in cost per (whole) unit. Of course these speed calculations only tell us the average speed you travelled over the hour or half hour respectively. At some points, you might have gone faster and at some points you might have gone slower. For this reason, if we wanted to know your speed at a particular moment in time, we would ideally like to consider a very small time interval—indeed, the smaller the better. That is, if we want to know your speed at time t, the best measure would be

$$\text{speed} = \frac{\big(D(t+h) - D(t)\big)\ \text{km}}{h\ \text{hour}}\,,$$

where h is some small fraction of an hour (say, 1/3600th of an hour, that is, a second).

Likewise, we typically want to know not the average incremental cost over some interval, but the incremental cost (per whole unit) at a specific point. We can do this by calculating expression (2.3) with as small an h as possible. Indeed, we want to do it as h goes to zero. That is, for continuous cost functions we define MC by

$$MC(x) = \lim_{h\to 0} \frac{C(x+h) - C(x)}{h}\,, \tag{2.4}$$

where "lim" means the limit of that ratio as h goes toward zero.[8] For instance, suppose that $C(x) = 4x$. Then

$$\frac{C(x+h) - C(x)}{h} = \frac{4\times(x+h) - 4x}{h} = \frac{4h}{h} = 4\,.$$

Clearly, as the expression doesn't depend on h, the limit as h goes to zero is 4; that is, if $C(x) = 4x$, then $MC(x) = 4$. In fact, replacing 4 with any constant γ, we have, if $C(x) = \gamma \times x$

$$\frac{C(x+h) - C(x)}{h} = \frac{\gamma\times(x+h) - \gamma\times x}{h} = \frac{\gamma\times h}{h} = \gamma\,.$$

Hence, if $C(x) = \gamma \times x$, then $MC(x) = \gamma$.

---

[8] For example, suppose $C(x) = x^2$ and we want to know $MC(1)$. If $h = 1$, then our approximation is

$$\frac{2^2 - 1^2}{1} = 3\,.$$

If $h = 1/2$, then our approximation is

$$\frac{1.5^2 - 1^2}{.5} = 2.5\,.$$

If $h = 1/4$, then the approximation is 2.25. If $h = 1/8$, then 2.125. If $h = .0001$, then 2.001. The limit is the number that this sequence approaches as h goes to zero, which looks to be 2. Indeed, as shown later (see Proposition 3), it is 2.

As a second example, suppose $C(x) = \alpha x^2 + \beta x$, where α and β are constants. Then

$$\frac{C(x+h) - C(x)}{h} = \frac{\alpha(x^2 + 2xh + h^2) + \beta(x+h) - \alpha x^2 - \beta x}{h}$$
$$= \frac{2\alpha xh + \alpha h^2 + \beta h}{h} = 2\alpha x + \alpha h + \beta\,.$$

As h goes to zero, this last expression goes to $2\alpha x + \beta$; that is, if $C(x) = \alpha x^2 + \beta x$, then we've shown that $MC(x) = 2\alpha x + \beta$.

Sometimes, there is an overhead cost associated with production. For instance, a chemical plant might have the following cost schedule:

$$C(x) = \begin{cases} 0\,, \text{ if } x = 0 \\ c(x) + F\,, \text{ if } x > 0 \end{cases};$$

that is, if the firm decides to produce at all, then it incurs an overhead cost of F, $F > 0$, as well as a variable cost of $c(x)$. This cost function is discontinuous at $x = 0$ and, thus, we cannot define $MC(0)$. Nevertheless, we can define $MC(x)$ for $x > 0$ using expression (2.4)–the F doesnt matter if $x > 0$:

$$MC(x) = \lim_{h\to 0} \frac{\big(c(x+h) + F\big) - \big(c(x) + F\big)}{h} = \lim_{h\to 0} \frac{c(x+h) - c(x)}{h}\,.$$

We can summarize our analysis of specific functional forms as follows:

Proposition 3. *Let $C(x) = \alpha x^2 + \beta x + F$, where α, β, and F are non-negative constants (i.e., each is greater than or equal to zero). Then $MC(x) = 2\alpha x + \beta$ for $x > 0$. If $F = 0$ (i.e., there is no overhead cost), then $MC(0)$ is defined and is equal to β.*

$\int dx$ Marginal Cost as Derivative

If you've had calculus, you no doubt recognize expression (2.4) as the definition of a derivative. That is, we have

$$MC(x) = \frac{d}{dx}C(x) = C'(x)\,.$$

2.5 A Graphical Analysis of Costs

We can also understand the relations among $C(x)$, $MC(x)$, and $AC(x)$ graphically.

For instance, because $AC(x) = C(x)/x$, it follows that

$$C(x) = xAC(x)\,. \tag{2.5}$$

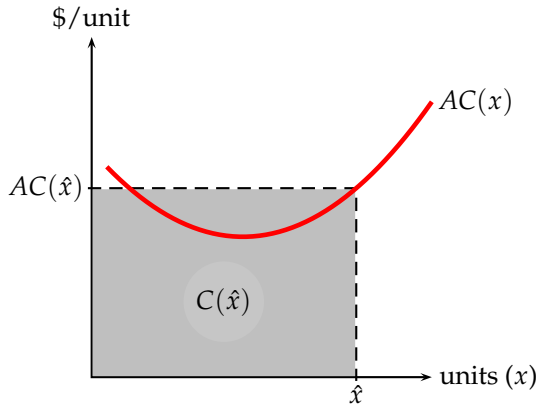

Figure 2.1: Total cost of $\hat{x}$ units is the area of the rectangle with width $\hat{x}$ and height $AC\hat{x}$.

We can illustrate this graphically. See Figure 2.1. In this figure, I have plotted AC (in red) as a function of units, x. Because AC is in dollars per unit, the vertical axis in Figure 2.1 is in terms of \$/unit. Consider a specific number of units, $\hat{x}$. Suppose $\hat{x}$ is the width of a rectangle whose height is $AC(\hat{x})$. The area of this rectangle (shown in gray in Figure 2.1) is its width times height, or $\hat{x}AC(\hat{x})$, which, from expression (2.5), is $C(\hat{x})$. That is, the total cost of $\hat{x}$ units can read off the graph of average cost as the area of the rectangle formed by $\hat{x}$ and $AC(\hat{x})$.

We can also relate MC and AC. Consider Figure 2.2. The average cost schedule is the same as in Figure 2.1. To this, I have added the marginal cost schedule shown in black. Consistent with Proposition 2, we see that $MC(x) < AC(x)$ for x at which $AC(\cdot)$ is decreasing (headed downward) and that $MC(x) > AC(x)$ for x at which $AC(\cdot)$ is increasing (headed upward). It follows, therefore, that at the point at which $AC(x)$ is at a minimum, x^*, MC and AC must coincide; that is, $MC(x^*) = AC(x^*)$.

Proposition 4. *The marginal cost schedule intersects the average cost schedule at the*

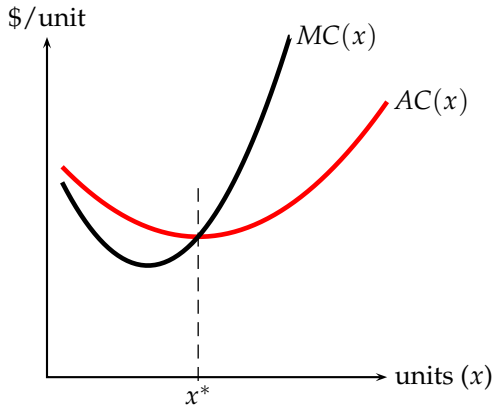

Figure 2.2: Average cost is falling whenever it exceeds marginal cost and it is increasing whenever it is less than marginal cost. The minimum of average cost occurs where marginal cost cuts average cost from below.

minimum of average cost.[9]

Finally, we can relate *MC* and total cost graphically. Figure 2.3 shows a plot of the marginal cost schedule. Because *MC* is expressed in terms of dollars per unit, the vertical axis is labeled $/unit. The marginal cost schedule that is plotted in this figure corresponds to a good that is produced in discrete units. That is,

$$MC(n) = C(n) - C(n-1)\,.$$

Observe the gray area beneath the *MC* schedule has been divided into rectangles. Each rectangle has width 1. The height of the $n$the rectangle is $MC(n)$. Recall that the area of a rectangle is height times width, so the area of the nth rectangle is $MC(n)$. But this means that the area under the *MC* schedule from

[9] $\boxed{\int dx}$ Observe that, because $AC(x) = C(x)/x$, it follows that

$$AC'(x) = \frac{xC'(x) - C(x)}{x^2} = \frac{xMC(x) - C(x)}{x^2}\,.$$

When *AC* is at a minimum, $AC'(x) = 0$. This means that the numerator of the last expression equals zero: $xMC(x) - C(x) = 0$. Dividing both sides by x and rearranging yields

$$MC(x) = \frac{C(x)}{x} = AC(x)\,.$$

Observe, then, this argument proves that at every local minima of $AC(\cdot)$, $AC = MC$.

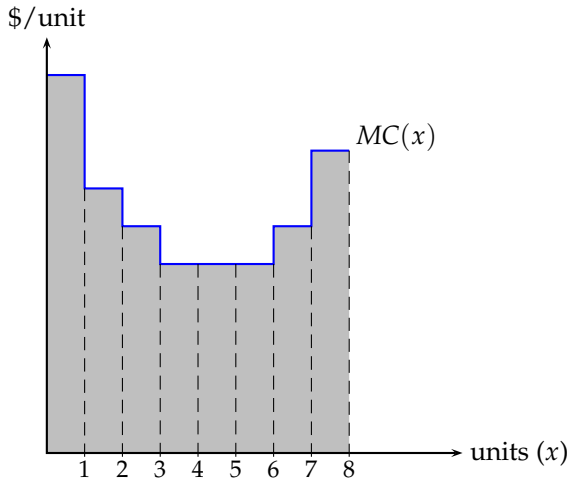

Figure 2.3: A marginal cost schedule has been plotted. The area beneath the marginal cost schedule between two points corresponds to the difference in total costs between those two points.

n to m, $m > n$, units is

$$\begin{aligned}\text{Area under } MC \text{ from } n \text{ to } m &= MC(n+1) + \cdots + MC(m) \\ &= \sum_{k=n+1}^{m} MC(k)\,.\end{aligned}$$

Recall that

$$C(n) = \sum_{k=1}^{n} MC(k)\,.$$

Hence,

$$\begin{aligned}C(m) - C(n) &= \sum_{k=n+1}^{m} MC(k) \\ &= \text{Area under } MC \text{ from } n \text{ to } m.\end{aligned}$$

This is a general result:

Proposition 5. *The area beneath the marginal cost curve between two points x_1 and x_2, $0 < x_1 < x_2$, equals $C(x_2) - C(x_1)$.*

Note, although we derived Proposition 5 using the case of a discrete good (*e.g.*, cars), the result applies equally well to the case of a continuous good (*e.g.*, a

chemical). Thus, for example, the area under the marginal cost schedule of a chemical between 4.8 and 5.3 metric tons equals $C(5.3) - C(4.8)$.

Moreover, it should be noted that Proposition 5 reflects a general result about the relation between marginal and total functions:

Proposition 6. *The area beneath any marginal curve between two points, x_1 and x_2, $x_1 < x_2$, equals $G(x_2) - G(x_1)$, where $G(\cdot)$ is the corresponding* total *schedule.*[10]

In Proposition 5, we required that $x_1 > 0$. Why? Because, if there is an overhead cost and the good in question is continuous, then $MC(0)$ is not well defined. Fortunately, we can consider, in that case, the *MC* schedule starting arbitrarily close to zero. The area to the left of our arbitrarily close starting point is negligible and can, thus, be ignored. Hence, the area under the marginal cost curve starting almost from zero to some point x must be the total *variable* cost of x units. Total cost is the sum of variable and overhead cost. Hence, $C(x)$ is the area under the marginal cost schedule starting almost at 0 to x plus the overhead cost.

Proposition 7. *If there is an overhead cost, $F > 0$, and units are continuous, then*

$$C(x) = \textit{Area under MC from almost 0 to } x + F\,.$$

$\int dx$ Total Costs as an Integral

Write

$$C(x) = \begin{cases} 0\,, \text{ if } x = 0 \\ c(x) + F\,, \text{ if } x > 0 \end{cases}\,,$$

where $F \geq 0$ is overhead cost and $c(\cdot)$ is the *variable* cost function. Assume $c(\cdot)$ is differentiable. By the definition of a variable cost, it must be that $c(0) = 0$.

Observe that for $x > 0$, $MC(x) = c'(x)$. Because a single point cannot affect the value of an integral, there is no loss in our treating $MC(0)$ as equaling $c'(0)$; that is, it is immaterial how we define $MC(0)$ with respect to integration, so we're free to define it as $c'(0)$.

By the fundamental theorem of calculus,

$$c(x_2) - c(x_1) = \int_{x_1}^{x_2} c'(x)dx = \int_{x_1}^{x_2} MC(x)dx\,. \tag{2.6}$$

Hence, the area under the marginal cost schedule between x_1 and x_2 is the difference in variable cost from producing x_2 units as opposed to x_1 units. For $x_2 > x_1 > 0$,

$$C(x_2) - C(x_1) = \big(c(x_2) + F\big) - \big(c(x_1) + F\big) = c(x_2) - c(x_1)\,.$$

[10] $\int dx$ This is just the fundamental theorem of calculus.

Therefore, we have

$$C(x_2) - C(x_1) = \int_{x_1}^{x_2} MC(x)dx \qquad (x_2 > x_1 > 0)$$

This is just Proposition 5.

If we let $x_1 = 0$ in expression (2.6), we see

$$c(x_2) = \int_0^{x_2} MC(x)dx$$

(recall $c(0) = 0$). Substituting into the definition of $C(\cdot)$, we arrive at

$$C(x_2) = c(x_2) + F = \int_0^{x_2} MC(x)dx + F\,.$$

2.6 The Cost of Capital

As we saw in Section 2.1, whether an expenditure is a cost or not (*i.e.,* sunk) can depend on the decision-making horizon. Recall, for instance, Example 2. Once locked into a lease, rent payments are not a cost. But, at the time of lease renewal, they are a cost.

One *might* view capital expenditures (*e.g.,* machinery, vehicles, plant and facilities, etc.) similarly. That is, once purchased, the purchase price is a sunk expenditure. Prior to purchase, the purchase price is a cost. Yet, while it is definitely true that the price paid is sunk after purchase, this does not entail there is no cost to using the capital. It would be costless only if, as with the lease, there was nothing else that could be done. Typically, however, there is at least one thing else that can be done with capital assets: resale.

The ability to resell capital goods creates two sources of imputed cost. To identify them, consider a relevant decision-making horizon; namely, the time between points at which you can resell the capital good or asset. This might, for instance, be a day, a month, or longer depending on circumstances. Let V_0 be the value—the resale price—for this asset at the beginning, time 0, of the time period and let V_1 be its value—resale price—at the end of the period, time 1. Let r be the amount that the firm can earn per dollar during this time period on money optimally invested (this could just be the interest rate on funds banked). If the firm sold the asset at the beginning of the period, time 0, then, at time 1, it would have $(1+r)V_0$. If it sells it at the end of the period, time 1, then it would have V_1. The difference, $(1+r)V_0V_1$, is what is forgone by using the asset over the time interval; that is, it is the cost of using the asset.

Decompose this cost as follows:

$$\text{capital cost} = \underbrace{rV_0}_{\text{forgone return}} + \underbrace{V_0 - V_1}_{\text{depreciation}}\,. \tag{2.7}$$

Depreciation is the amount by which the resale value of an asset falls over time. Depreciation is driven by two factors. First, using an asset often means wear and tear, reducing the asset's productive life and, hence, its resale value. Second, changes that take place in technology over time can also lower resale value (consider, *e.g.,* that a computer with a pentium chip, even if never used, is worth far less today than when it was first sold).

Divide the cost of capital, expression (2.7), by V_0. This yields a capital-cost rate equal to $r + \delta$, where

$$\delta = \frac{V_0 - V_1}{V_0}$$

is the *rate of depreciation*. From our earlier discussion, observe that δ is greater the greater is the rate of wear and tear and the greater is the rate of technology change.

> **Example 10:** Suppose you are in the business of renting residential housing. In a stable housing market, the resale value of a house today is essentially equal to its resale value next month (at least assuming normal use). Ignoring tax considerations and the "joys of home ownership," the value of a house to a rational consumer is equal to the discounted value of the rent payments she would have to make for equivalent housing; that is,[11]
>
> $$V = \sum_{t=1}^{\infty} \frac{\rho}{(1+r)^t} = \frac{\rho}{r},$$
>
> where ρ is the rental price. If you and your tenants face the same rate of return (*i.e.,* interest rate), then, in a stable housing market, it will be the case that rental prices and house prices are related by the condition $\rho = rV$. In other words, rental prices should equal the cost of the capital that the house represents.
>
> Let's continue this example by supposing that house prices appreciate suddenly. Then rents will go up, because $V_1 > V_0$, but those tenants with leases will find themselves paying below market rates. This explains why, in an appreciating housing market, average house prices are seen to be rising faster than average rents.
>
> Further consider a housing market with a forecast appreciation rate of α, driven, perhaps, by in migration due to a strong labor market. The rate α is like a negative δ, so the cost of renting a house over the relevant period is $(r\alpha)V$. Consequently, you, as a landlord, make a profit for any rent $\rho(r\alpha)V$.

It is important to observe that depreciation (or appreciation) is determined by the resale value of the asset, which is a function of the actual wear and tear on the asset, as well as trends in the resale market (*e.g.,* technological changes). This should be contrasted with accounting measures of depreciation. Accounting measures employ formulæ that has little to no relation to actual changes in

[11] We can treat the value of the house as a perpetuity because, even though any one owner won't live forever, when he or she sells it, he or she will capture its expected future value (including expected future resale value).

market value. For instance, the IRS states that cars and computers have a useful life of five years and allows them to be depreciated at the same rate. That is, the accounting procedure acts as if their market values fall at the same rate. Yet, experience tells us that the rates at which cars and computers actually lose market value are very different.

2.7 Introduction to Cost Accounting

This section introduces and explains some cost-accounting terminology. As we saw with depreciation, accounting terminology overlaps considerably with economic terminology. The problem is that, although the words are the same, their meanings are not. What are deemed costs by accountants may not be costs as weve defined them or they may not be properly allocated for the purposes of making sound managerial decisions. In addition, some true costs can fail to be accounted for by the accounting system.

On the other hand, accounting information is valuable informationit is not to be disregarded. To exploit this information, however, one must learn to read the information correctly and not be mislead by the terminology.

The principal goal of accounting

Although cost accounting serves many purposes, its principal purpose is to *account* for the consumption of resources by the firm. In essence, accounting seeks to determine where the money went. This is clearly a necessary activity in any firm.

The problem and an example

A problem is that, in addition to accounting for expenditures, most accounting methods also allocate expenditures. This allocation can be at odds with what economics tells us.

To consider a simple scenario, suppose that a firm makes lefthanded scissors and righthanded scissors on the same machine. The labor required for each pair of scissors (unit) made is $1. The raw inputs for each pair of scissors is $1. In addition, each time the machine is switched from one type of scissors to the other, a set-up cost of $200 is incurred. Suppose the machine is set-up once at the beginning of the day for righthanded scissors and, then, later in the day it is switched to lefthanded scissors. Finally, suppose 900 righthanded scissors and 100 lefthanded scissors are produced each day. Daily expenditure is, therefore, $2400. A traditional accounting method for allocating these expenditures would call the $2 per scissors incurred from labor and raw inputs a direct cost. Hence, righthanded scissors incur $1800 per day in direct costs, while lefthanded scissors incur $200 per day in direct costs. This traditional method would also treat the set-ups as overhead and allocate this overhead between the two types of scissors on the basis of units produced. Consequently, 90%

of the $400 in daily overhead would be allocated to righthanded scissors and 10% of the $400 would be allocated to lefthanded scissors. The total "cost" of righthanded scissors would be $2160 and the total "cost" of lefthanded scissors would be $240 (note these sum to $2400). The per-unit "cost" of righthanded scissors would, therefore, be $2.40 and the per-unit "cost" of lefthanded scissors would also be $2.40. This allocation of expenditures would make it seem that righthanded and lefthanded scissors were equally expensive to produce. This conclusion, however, is wrong!

To see why it's wrong, suppose the firm in question made only righthanded scissors and was considering adding lefthanded scissors. How would expenditures change? Making only righthanded scissors, there are no set-ups (except on the very first day of production). So adding lefthanded scissors means incurring an additional $400 per day in set-up costs plus $200 in direct costs—a total of $600. If the price at which scissors were sold was $5 per pair, then it would not be profitable to add lefthanded scissors: total additional revenue = $500 ¡ $600 = total additional cost. A conclusion that would have been missed if the traditional allocation had been used.[12]

What this example illustrates is the danger of assigning costs by any means other than cost causation.

Moral: *Assign costs by cost causation.*

Accounting terminology

Basis of allocation: How shared overhead (see below) is allocated among different products. In the example above, set-up, which was treated as shared overhead, was allocated on the basis of units produced. Other common bases of allocation are labor hours and machine hours. A synonym for basis is *overhead rate*.

Direct costs: expenditures that can be traced to a single product or activity. In the example above, the $2 in labor and raw inputs are direct costs. See, also, *variable costs* below.

Direct labor: the compensation paid to employees whose time and effort can be traced to the product in question. The $1 in labor of each unit in the example above would represent direct labor.

Direct materials: the raw inputs that can be traced directly to the product in question. The $1 in raw inputs for each unit in the example above would represent direct materials.

Fixed costs: expenditures that do not vary with the number of units produced or that are not always increasing with the number of units produced.

Manufacturing cost: the sum of *direct labor*, *direct materials*, *manufacturing overhead*, and work beginning in process inventory minus work ending in process inventory.

[12] To verify: Daily profits if the firm produces both types is $2600 (= $5000 − $2400). Daily profit if it produces only righthanded scissors is $2700 (= $4500 − $1800).

Manufacturing overhead: all manufacturing expenditures except for *direct costs*. Included in manufacturing overhead are indirect materials, indirect labor, maintenance, utilities, rent, insurance, depreciation, and taxes. Manufacturing overhead is a *product cost* (see below). Manufacturing overhead is sometimes called *direct overhead* or *factory overhead*.

Overhead: expenditures that are not directly attributed to a given unit of output. For instance, in the example above, set-up costs are considered overhead because they are not directly attributed to a given unit of output.[13]

Overhead pools: different accounts into which overhead may be divided if the basis of allocation for these different accounts differ. For instance, if someoverhead is to be allocated on the basis of labor hours and other overhead is to be allocated on the basis of machine hours, then the first set of overhead would be one pool and the second set of overhead would be another pool.

Period costs: expenditures incurred during a given period of time (usually the period that an accounting statement covers). They are not allocated to the production process itself. If, in the above example, the firm in question spent \$100 per day marketing scissors, then this \$100 would be a period cost (assuming, unrealistically, that a daily accounting statement was produced for the entire firm). Period costs are necessarily overhead.

Product costs: costs incurred during production that can be allocated to production itself. In the example above, all the expenditures would be treated as product costs.

Shared overhead: overhead common to more than one product. In the example above, the salary of the factory manager would be shared overhead.

Unit cost: the accounting cost of a product divided by the number of units produced. In the example above, the unit cost of a pair of scissors was \$2.40. Note that the unit cost is *not* the same as marginal cost. When accounting cost equals economic cost, a synonym for unit cost is average cost.

Variable costs: expenditures that vary with each unit produced. See direct costs.

> **Example 11 [Parable of Red Pens and Blue Pens]:** Once upon a time a little firm made two products, red pens and blue pens. Each pen used 15 cents worth of labor and raw materials. Each pen was run through a machine, the daily cost of which was \$1000 regardless of how many pens were run through or their colors. The firm could sell the first 5000 red pens

[13] From an economic standpoint, however, we could attribute the set-up costs to a given unit, namely the first unit produced after the machine is switched. That is, for example, the marginal cost schedule for lefthanded scissors is $MC(1) = \$202$ and $MC(2) = \cdots = MC(100) = \$2$.

it made each day at 30 cents each; additional red pens, however, were sold at 20 cents per pen. The firm could sell all the blue pens it wanted at 25 cents per pen. The firm, however, could make no more than 8000 pens a day and it could not expand over its relevant decision-making horizon. The little firm chose to manufacture 5000 red pens and 3000 blue pens a day for a daily profit of

$$\$50 = 5000 \times (.30 - .15) \text{ dollars} + 3000 \times (.25 - .15) \text{ dollars} - \$1000\,.$$

As this was greater than \$0, the little firm was happy to produce.

One day an evil accountant came along and said the little firm should adopt an accounting system that allocated shared overhead (*e.g.*, the \$1000 for the aforementioned machine). Being naïve, the little firm went along. The accountant chose to allocate the shared overhead on the basis of output—thus, the red-pen line was billed \$625, which is five eighths of \$1000,[14] and the blue-pen line was billed the remaining \$375. The new accounting is shown in Table 2.1.

	Pens	Revenue	Direct Cost	Shared Overhead	Total Expense	Profit
Red Pens	5000	\$1500	\$750	\$625	\$1375	\$125
Blue Pens	3000	\$750	\$450	\$375	\$825	−\$75
Total	8000	\$2250	\$1200	\$1000	\$2200	\$50

Table 2.1: The Evil Accountant's New Accounting System

Upon examining the accounting data, the evil accountant snickered, "Aha! Your blue-pen line is unprofitable—you should shut it down." Dutifully, the naïve little firm shut down its blue-pen line and switched over to producing nothing but red pens. Now the firms revenues were

$$\$2100 = 5000 \times \$0.30 + 3000 \times \$0.20\,.$$

Because the blue-pen line was shut, the \$1000 cost of the machine was fully allocated to the red-pen line. The firm's new accounting is shown in Table 2.2.

	Pens	Revenue	Direct Cost	Shared Overhead	Total Expense	Profit
Red Pens	8000	\$2100	\$1200	\$1000	\$2200	−\$100
Blue Pens	0	\$0	\$0	\$0	\$0	\$0
Total	8000	\$2100	\$1200	\$1000	\$2200	−\$100

Table 2.2: New Accounting After Blue-Pen Line Shut

[14] Allocating on the basis of output means taking the output of the red-pen line, 5000 pens, and dividing it by total output, 8000 pens, to get the red-pen line's share. Similarly, the blue-pen line's share would be 3000/8000 or 3/8.

> Upon examining the accounting data, the evil accountant chortled, "Aha! Your entire company is unprofitableyou should shut down completely." Dutifully, the naïve little firm did, shutting its doors forever. So thanks to the evil accountant, the little firm went from making a tidy profit of $50 a day to going out of business!

Shared overhead allocation: *Has nothing to do with good decision making.*

The moral of this parable is simple: Don't allocate shared overhead, it can only lead to dopey decisions.

In the lefthanded-righthanded-scissors example, the problem was that the accounting system missed that an overhead cost should properly be allocated fully to lefthanded scissors—this is a case where allocatable overhead is mistakenly treated as shared, with the mistake further compounded by allocating it on the basis of an *ad hoc* rule. In the red-pens-blue-pens example, the overhead cost shouldn't be allocated to either line—it is true shared overhead. In both examples, because the basis of allocation had nothing to do with cost causation, the accounting gave a misleading picture of what was going on.

Summary | 2.8

The key takeaways of this chapter are:

- To make the best business decisions, remember the cost of something is the value of the best forgone alternative. That is, use the notion of *opportunity cost*.
- Cost = Expenses − Sunk expenditures + Imputed costs.
- Sunk expenditures are *irrelevant* for decision making.
- Understand marginal cost (MC) and average cost (AC), as well as their relations between each other and with total cost (C).
- Understand the cost of capital.
- Cost causation is the right way to allocate costs. Allocation of overhead on any other basis will be misleading for decision making (recall lefthanded and righthanded scissors).
- Shared overhead should not be allocated for the basis of decision making (recall red pens and blue pens).

Introduction to Pricing | 3

This chapter describes how to figure out the optimal price and quantity if your firm is engaging in *simple pricing*; that is, pricing in which the firm sets a given price per unit, which is paid by all customers, each of whom is free to buy as many units at that price as he or she desires.[1] After defining "optimal," we will discuss the concept of marginal revenue at length. While there is some involved analysis required, the topic justifies the pain. Optimal pricing and quantity setting in this context means making as much money as possible. The important takeaways from this chapter are

- Marginal revenue equals marginal cost at the optimal quantity produced (this equality may be approximate in the case of discrete goods).
- Marginal revenue comes from an underlying demand curve. You should know how to derive a marginal revenue curve from a demand curve assuming simple pricing.
- Demand curves themselves come from consumer preferences and from the prices of all goods. You should be able to evaluate the effects of changes in preferences and these other prices on your best choices.

Simple Pricing Defined | 3.1

This note discusses optimal pricing of a product by a firm that must charge the same price per unit to all of its consumers.

Definition. *A firm engages in* simple pricing *for a particular product if that product is sold for the same price per unit no matter who the buyer is or how many units the buyer purchases.*

Observe that simple pricing is *nondiscriminatory pricing*.

In subsequent chapters, we will consider the possibility that a firm chooses to charge either different prices to different individuals or prices in such a way that a consumers expenditure is not a simple multiple of a per-unit price (*i.e.*, his or her expenditure is not proportional to quantity purchased). Simple pricing applies when the identity of the buyer cannot be observed or inferred at reasonable cost. It also applies when the seller cannot prevent *arbitrage* among

[1] This is sometimes called *linear pricing*.

buyers when buyers can purchase multiple units. Arbitrage means taking advantage of an ability to buy low and sell high. So, if one set of buyers faces a lower price than another set and the former can resell to the latter, then the consequent arbitrage opportunity will force a single price to hold. For example, if a chemical producer proposed to charge a different price for the same chemical depending on the buyers industry, then those buyers in the industry facing a low price could resell to those buyers in the industry facing the high price and reap a profit. The producer would, effectively, be limited to a single price, the lowest one it sets.

3.2 Profit Maximization

Note: The topics considered in this and following sections are quite general and pertain not only to simple pricing, but other forms of pricing as well.

Profit: *Is revenue minus cost*

A firm's *profit* is the revenue it takes in minus its cost. If we let $R(x)$ be the firm's revenue from selling x units, then its profit from selling x units, $\pi(x)$, is $R(x)C(x)$, where $C(x)$ is the total cost of x units.

If the firm sets a price of p per unit—engages in simple pricing—then $R(x) = px$.

In choosing the amount to produce and sell, the firm seeks to find the quantity, x, that maximizes profit, $\pi(x)$. Under the real-life conditions that govern revenue and cost, an x that maximizes $\pi(x)$ must exist. Let's use an asterisk to denote that quantity (*e.g.*, x^*, n^*, etc.).

Suppose, first, that we are considering a discrete good (*e.g.*, shirts) rather than a continuous good (*e.g.*, a liquid). Saying that n^* is the profit-maximizing amount is the same as saying that $\pi(n^*) \geq \pi(n)$ for all other n. In particular, consider the quantities $n^* - 1$ and $n^* + 1$. We know that

$$\pi(n^*) \geq \pi(n^* - 1) \text{ and} \tag{3.1}$$
$$\pi(n^* + 1) \leq \pi(n^*)\,. \tag{3.2}$$

Substituting $R(n) - C(n)$ for $\pi(n)$ yields

$$R(n^*) - C(n^*) \geq R(n^* - 1) - C(n^* - 1) \text{ and}$$
$$R(n^* + 1) - C(n^* + 1) \leq R(n^*) - C(n^*)\,;$$

or, further rearranging,

$$R(n^*) - R(n^* - 1) \geq \underbrace{C(n^*) - C(n^* - 1)}_{MC(n^*)} \text{ and}$$
$$R(n^* + 1) - R(n^*) \leq \underbrace{C(n^* + 1) - C(n^*)}_{MC(n^*+1)}\,.$$

If we define $MR(n) = R(n) - R(n-1)$, we can rewrite these last expressions as

$$MR(n^*) \geq MC(n^*) \text{ and} \tag{3.3}$$
$$MR(n^* + 1) \leq MC(n^* + 1)\,. \tag{3.4}$$

Marginal revenue: *The change in revenue from selling an additional unit.*

What is $MR(n)$? It is the change in revenue incurred by selling the nth unit rather than selling only $n-1$ units and it is called *marginal revenue*.

Expression (3.3) tells us that for n^* to be the profit-maximizing quantity, then the marginal revenue from the nth unit needs to be at least as great as the marginal cost of the nth unit—which makes sense; if it weren't true (*i.e.*, $MR(n^*) < MC(n^*)$), then, whatever the gain in revenue from the n^*th unit, it is outweighed by the additional cost of producing the n^*th unit. Hence, it wouldnt be sensible to produce n^* units ($n^* - 1$ units would be better). Expression (3.4) tells us that for n^* to be the profit-maximizing quantity, then the marginal revenue from the $n^* + 1$st unit cannot exceed the additional cost incurred by producing the $n^* + 1$st unit; that is, that $n^* + 1$ units cannot be better than n^* units.

Proposition 8. *A necessary condition for n^* to be the profit-maximizing output is that expressions* (3.3) *and* (3.4) *both hold true.*

3.3 The Continuous Case

As noted in the previous chapter, for some goods, a unit is a fairly arbitrary notion, and amounts of such goods can be treated as continuous. Moreover, even for discrete goods, it is simply easier sometimes to treat quantity as an approximately continuous variable.

As with cost, we need to define marginal revenue in the continuous context. To that end, by analogy to expressions (2.3) and (2.4) in Section 2.4, we define MR as

$$MR(x) = \lim_{h \to 0} \frac{R(x+h) - R(x)}{h}\,. \tag{3.5}$$

For example, if $R(x) = Ax - Bx^2$, where A and B are constants, $A > 0$ and

$B \geq 0$, then

$$\begin{aligned} MR(x) &= \lim_{h\to 0} \frac{\big(A(x+h) - B(x+h)^2\big) - (Ax - Bx^2)}{h} \\ &= \lim_{h\to 0} \frac{Ax + Ah - Bx^2 - 2Bxh - Bh^2 - Ax + Bx^2}{h} \\ &= \lim_{h\to 0} \frac{Ah - 2Bxh - Bh^2}{h} \\ &= \lim_{h\to 0} A - 2Bx - Bh \\ &= A - 2Bx\,. \end{aligned}$$

It will be useful to memorialize this example:

Proposition 9. *If $R(x) = Ax - Bx^2$, $A > 0$ and $B \geq 0$, then $MR(x) = A - 2Bx$.*

In the continuous case, we can rewrite the conditions for x^* to be the profit-maximizing quantity, expressions (3.1) and (3.2), as

$$\begin{aligned} \pi(x^*) &\geq \pi(x^* - h) \text{ and} \\ \pi(x^* + h) &\leq \pi(x^*)\,. \end{aligned}$$

Substituting $R(x) - C(x)$ for $\pi(x)$ and rearranging yields

$$\begin{aligned} R(x^*) - R(x^* - h) &\geq C(x^*) - C(x^* - h) \text{ and} \\ R(x^* + h) - R(n^*) &\leq C(x^* + h) - C(n^*)\,. \end{aligned}$$

Dividing both sides of those inequalities by h and taking limits as h goes to zero yields

$$\begin{aligned} MR(x^*) &\geq MC(x^*) \text{ and} \\ MR(x^*) &\leq MC(x^*)\,. \end{aligned}$$

The only way both inequalities can be satisfied is if

$$MR(x^*) = MC(x^*)\,. \tag{3.6}$$

We can conclude, therefore, as follows:

Proposition 10 (MR = MC rule)**.** *In the continuous case, a necessary condition for x^* to be the profit-maximizing output is that $MR(x^*) = MC(x^*)$.*

$\int dx$ Profit Maximization

If you've had calculus, you no doubt recognize expression (3.5) as the definition of a derivative. That is, we have

$$MR(x) = \frac{d}{dx}R(x) = R'(x)\,.$$

Consider maximizing profits. We know that a function (*e.g.*, $\pi(\cdot)$) is increasing wherever its derivative is positive and decreasing wherever its derivative is negative. It follows, therefore, that at an optimum, the derivative must be zero (*i.e.*, the top of a hill is flat). Hence, if x^* is the profit-maximizing quantity, then

$$\frac{d}{dx}\pi(x)_{|x=x^*} = 0\,;$$

that is, $\pi'(x^*) = 0$. Substituting $R(x) - C(x)$ for $\pi(x)$, this implies

$$R'(x^*) - C'(x^*) = MR(x^*) - MC(x^*) = 0\,.$$

From which we again see that a necessary condition for x^* to be the profit-maximizing quantity is that $MR(x^*) = MC(x^*)$.

3.4 Sufficiency and the Shutdown Rule

Propositions 8 and 10 are only necessary conditions; that is, they identify possible candidates for being the profit-maximizing quantity, but they do not guarantee that a given n or x that satisfies those conditions is the profit-maximizing quantity. This is similar to the fact that while a high GMAT score is necessary for being admitted to a top MBA, knowing someone has such a high score is not sufficient to know—does not guarantee—she has been admitted to a top MBA program. Fortunately, there is a condition that insures that, if the firm should be in business at all, the conditions stated in Propositions 8 and 10 are also sufficient (*i.e.*, identify the profit-maximizing quantity).

We will establish the sufficiency condition for the continuous case first. Let x^* be the candidate for the profit-maximizing quantity identified by expression (3.6). If $MR(x) > MC(x)$ for all $x < x^*$ and $MR(x) < MC(x)$ for all $x > x^*$, then x^* must be the profit-maximizing quantity (assuming the firm should operate at all). To see why, observe that marginal profit, $MR(x)MC(x)$ is positive for all $x < x^*$; that is, every additional unit in this region contributes positively to total profit. On the other hand, marginal profit is negative for all $x > x^*$; that is, every additional unit in this region reduces total profit. Graphically, we're climbing up the total profit "hill" in the region $x < x^*$ and we're descending the total profit "hill" in the region $x > x^*$. See Figure 3.1.

We've established:

Proposition 11. *If*

(i) $MR(x^*) = MC(x^*)$,

(ii) $MR(x) > MC(x)$ *for all* $x < x^*$, *and*

(iii) $MR(x) < MC(x)$ *for all* $x > x^*$,

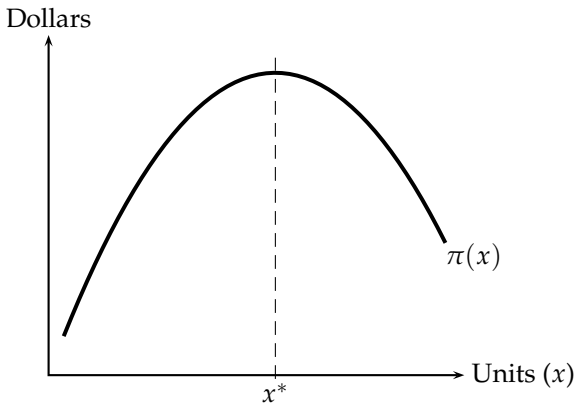

Figure 3.1: Profits are increasing with x if $x < x^*$. Profits are decreasing with x if $x > x^*$.

then x^* is the profit-maximizing quantity for the firm to produce (if it should be in business at all).

Another way to view Proposition 11 is that x^* is the profit-maximizing quantity if the marginal revenue schedule, $MR(\cdot)$, crosses the marginal cost schedule, $MC(\cdot)$, from above at x^*. See Figure 3.2 (note the two vertical scales—profits are in dollars, while marginal revenue and marginal cost are in dollars/unit). An equivalent way to state Proposition 11 is, thus,

Proposition 12. *If marginal revenue crosses marginal cost once at x^* and does so from above, then x^* is the profit-maximizing quantity (if the firm should be in business at all).*

Now consider the discrete case. Observe that we could plot marginal revenue and marginal cost against output, n. If we connected the dots, then we would have, essentially, marginal revenue and marginal cost curves respectively. By Proposition 12, if the marginal revenue curve crosses marginal cost once from above at a given point, then that point approximates the profit-maximizing quantity. Why approximate? Well, remember, this is the discrete case and the curves might cross at a non-integer point. But, then, the profit-maximizing quantity would be the highest integer to the left of the cross; that is, an n^* such that $MR(n^*)MC(n^*)$ and $MR(n^*+1) < MC(n^*+1)$.

We can translate this analysis into a statement like Proposition 11. However, before we do so, we might question the likelihood that the marginal revenue schedule would cross the marginal cost schedule only once and from above in the discrete case. Remember, in the discrete case, $MC(1)$ can be quite large because all the overhead costs are triggered by starting production (going from

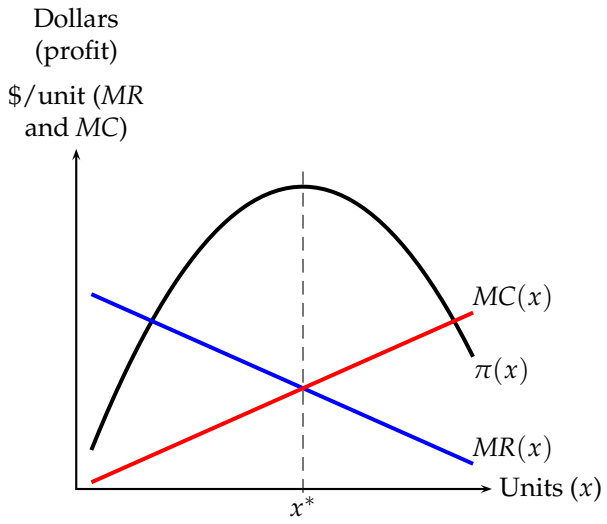

Figure 3.2: Marginal revenue (blue line) crosses marginal cost (red line) from above at the profit-maximizing quantity, x^*.

0 to 1 unit). Hence, $MC(1) \gg MR(1)$. Fortunately, we can typically ignore that first unit for reasons that we will take up in a moment. We have, therefore,

Proposition 13. *Consider the discrete case. If*

(i) $MR(n^*) \geq MC(n^*)$,

(ii) $MR(n) > MC(n)$ *for all* n, $2 \leq n < n^*$, *and*

(iii) $MR(n) < MC(n)$ *for all* $n > x^*$,

then n^* *is the profit-maximizing quantity for the firm to produce (if it should be in business at all).*

The caveat "if it should be in business at all" has been used a number of times in this section. What do we mean by it? There is a final test that n^* (discrete case) or x^* (continuous case) must satisfy before we can conclude they're the amounts the firm should produce. That test is *would the firm lose money at that level of output?* If the answer is no, then they're the amounts that should be produced. If the answer is yes, then the firm should shutdown.

To go into detail, recall that $C(0) = 0$—if you're not producing, then, by definition, you're not forgoing the best alternative, so the cost must be zero. Not surprisingly, if you don't produce, you don't earn any revenue. That is, $R(0) = 0$. Hence, if the firm shuts down (produces nothing), then its profit is zero (*i.e.*,$\pi(0) = R(0)C(0) = 0$). We know that if the firm produces, the best it can do is produce n^* or x^*. So the alternatives are, in the discrete case, produce

n^* or 0. The former is worse than the latter only if $\pi(n^*) < \pi(0) = 0$. Likewise, for the continuous case, the alternatives are produce x^* or 0. Again, the former is worse only if $\pi(x^*) < \pi(0) = 0$.

We can summarize this as

Proposition 14 (Shutdown rule)**.** *If profit at the quantity determined by Propositions 11 or 13, as applicable, is non-negative, then the firm should produce that quantity. Otherwise, it should shutdown.*

Another way to describe this is in terms of average revenue and average costs. Let q^* be x^* or n^* as appropriate. The shutdown rule states that the firm should operate if and only if

$$R(q^*) - C(q^*) \geq 0\,.$$

Dividing through by q^* and rearranging, this yields

$$AR(q^*) \geq AC(q^*)\,, \tag{3.7}$$

where $AR(q) = R(q)/q$ is *average revenue*. In other words, the firm should operate if and only if $AR(q^*) \geq AC(q^*)$.

Example 12: A firm has revenue, $R(x)$, given by

$$R(x) = 10x - \frac{1}{1000}x^2\,.$$

Its costs, $C(x)$, are given by

$$C(x) = \begin{cases} 0\,, \text{ if } x = 0 \\ 2x + F\,, \text{ if } x > 0 \end{cases}\,,$$

where F is a non-negative constant. From Proposition 9,

$$MR(x) = 10 - 2\frac{1}{1000}x = 10 - \frac{1}{500}x\,.$$

From Proposition 3, $MC(x) = 2$ for $x > 0$. Setting $MR(x) = MC(x)$, we have

$$10 - \frac{1}{500}x^* = 2\,,$$

which, solving for x^*, implies $x^* = 4000$. Observe that $MR(x) > MC(x)$ for all $x < x^*$ and $MR(x) < MC(x)$ for all $x > x^*$. So, therefore, if the firm should produce at all, it should produce 4000 units. Should it produce? Observe

$$AR(x) = 10 - \frac{1}{1000}x \text{ and } AC(x) = 2 + \frac{F}{x}\,.$$

We have $AR(x^*) \geq AC(x^*)$ if

$$10 - \frac{1}{1000}4000 \geq 2 + \frac{F}{4000}\,;$$

that is, if

$$6 \geq 2 + \frac{F}{4000},$$

or if $16,000 \geq F$. So, provided the overhead cost, F, does not exceed \$16,000, the firm maximizes its profit by producing 4000 units. If F does exceed \$16,000, then the firm should shutdown.

3.5 Demand

Where does the revenue function come from? The answer is it depends on how the firm prices and what its consumers demands are for its product. In this section, we explore consumer demand and related issues.

Individual demand

Consider an individual and a good (which may be a physical product or a service). If the consumer receives q units of this good, she enjoys some benefit, $b(q)$, from these q units. We can think of this benefit as being the monetary equivalent of the happiness she enjoys from the q units. That is, she would be indifferent between having the q units or having $b(q)$ dollars in cash.

This benefit function, $b(\cdot)$, derives from the preferences, likes and dislikes, of the individual in question. It can also depend, as we will see later, on her income and the prices of other goods that she buys.

We can also define a *marginal benefit* schedule, $mb(\cdot)$, for this individual. In the discrete case, we have

$$mb(n) = b(n) - b(n-1);$$

that is, as always with marginals, the marginal benefit is the increment in her total benefit from adding the nth unit.

In the continuous case, we have

$$mb(x) = \lim_{h \to 0} \frac{b(x+h) - b(x)}{h}.$$

Note the similarity of this definition to the definition of marginal cost and marginal revenue in the continuous case.

Regardless of the case, a property of marginal benefit schedules is the following.

Observation. *Individuals' marginal benefits schedules are decreasing functions. That is, if $q_1 > q_0$, then $mb(q_1) < mb(q_0)$.*

Slope of marginal benefit: *Marginal benefit is a decreasing function.*

Why is this? Well, it follows from two factors. First, the increase in most people's enjoyment or happiness from more of the same thing is diminishing in the amount they receive. For example, the additional enjoyment gained from

a second candy bar is typically smaller than the enjoyment the first candy bar provided. Or, for those of you on the Atkins diet, the additional enjoyment gained from the second steak is typically smaller than the enjoyment the first steak provided. This property is known as *diminishing marginal benefit* (or, sometimes, *diminishing marginal utility*).

The second factor is simple "crowding out." If you eat one thing, you might not have room for something else. If you spend time in one activity, you might not have time for another activity. Hence, consumption of one thing often means forgoing consumption of something else. So, even if the marginal gross benefit of consumption isn't decreasing (*i.e.*, the effect on neural responses and neuro-transmitter production is constant), the overall marginal benefit would be decreasing because we are forgoing increasingly valued alternatives. That is, the opportunity cost is rising. For example, you might schedule your first hour of video game playing when there is nothing to watch on TV. If you play a second hour, however, it could mean forgoing shows you somewhat like. A third hour, could mean missing some of your favorite shows, etc.

Although we could do the analysis both for the discrete and the continuous case, it is far easier to do it for the continuous case. Hence, we will consider that case only. As, however, should become clear, marginal benefit is a lot like marginal revenue; so the analysis of marginal revenue and profit maximization in the discrete case carries over to marginal benefit and surplus maximization in the discrete case.

Goods are not typically given to us for free—usually, we have to pay for them. Because the money we spend on them could be spent on other things we like, we are always forgoing other things when we buy things. Opportunity cost! Hence, to determine how well off we are from buying q units of a goodhow much we *profit*—we need to subtract the expenditure on the q units from our benefit, $b(q)$. Under simple pricing, the expenditure is pq, where p is the price per unit of the good. The individuals profit—called her *consumer surplus*—is

$$\begin{aligned} cs(x) &= b(x) - \text{expenditure on } x \\ &= b(x) - px \qquad \text{(under simple pricing)} \end{aligned} \tag{3.8}$$

Using Proposition 3, the consumers "marginal cost" is just p, the price. Note her marginal cost is a constant. If we assume $mb(0) \geq p$, then the $mb(\cdot)$ schedule will cross her "marginal cost" schedule (*i.e.*, the horizontal line at height p) once from above. We can then invoke Proposition 12 to conclude that the consumer maximizes her "profit"—that is, her consumer surplus—by equating marginal benefit with price. In other words, she maximizes her consumer surplus by purchasing the amount x^* that solves

$$mb(x^*) = p\,. \tag{3.9}$$

Figure 3.3 illustrates

If $mb(0) < p$, then the consumer does best not to purchase any units. As is true of all marginal and total schedules, total benefit, $b(x)$, is the area beneath the $mb(\cdot)$ schedule from 0 to x units. Because of diminishing marginal

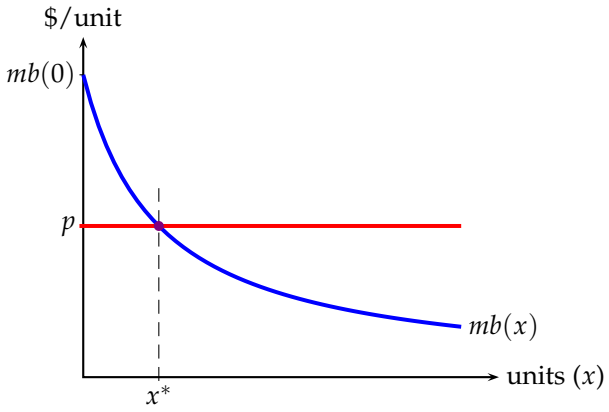

Figure 3.3: The marginal benefit schedule, $mb(\cdot)$, crosses the price line, p, once from above.

benefit, that area must be less than the rectangle whose area is $x \times mb(0)$ (see Figure 3.3). Hence, $b(x) < xmb(0)$. But if $mb(0) < p$, then $b(x) < px$ for all x and the consumer does better not to buy any amount.

Because $mb(\cdot)$ is decreasing, it is invertible (see Section A1.1). Hence, expression (3.9) can be read as defining x^* as a function of p. That is, if we vary p, we can see how the consumer's choice of optimal x varies using expression (3.9). Figure 3.4 illustrates. Note that we get a different x_t^* for each $p_t \leq mb(0)$. All $p_t > mb(0)$ map to the same x^*, namely 0.

(Individual) demand curve: *The relation between price and the amount an individual wishes to purchase.*

We can use the relation identified by Figure 3.4 to define a *demand curve* (alternatively, demand function or demand schedule) for the individual in question. Specifically, for $p > mb(0)$, we see that the quantity demanded (*i.e.*, that the consumer wishes to purchase) is 0. For $p \leq mb(0)$, we see that the quantity demanded (*i.e.*, that the consumer wishes to purchase) is the inverse of marginal benefit; that is, the amount demanded at p is $mb^{-1}(p)$. If we let $d(\cdot)$ denote the demand function (*i.e.*, the amount the consumer in question wishes to purchase at a given price), then we have

$$d(p) = \begin{cases} 0\,, \text{ if } p > mb(0) \\ mb^{-1}(p)\,, \text{ if } p \leq mb(0) \end{cases} . \tag{3.10}$$

Figure 3.5 illustrates.

Observe, from Figure 3.5, that we determine the amount demanded at a price, say p_0, by reading across from that price until we hit the demand curve, then down from where we hit the demand curve. Hence, as illustrated, $d(p_0)$—a point on the horizontal axis—is the amount demanded at price p_0—a point on the vertical axis.

Observe, too, that because a demand curve is derived from the marginal

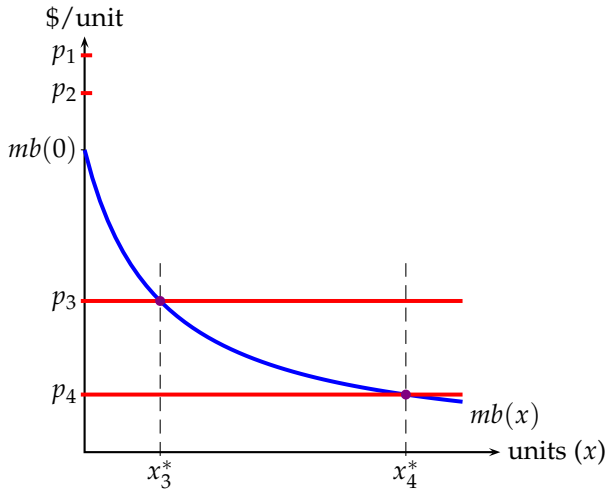

Figure 3.4: By varying the price, the marginal benefit curve can be used to determine the quantity the consumer wants (demands). Note that at high prices (*e.g.*, p_1 and p_2), demand is zero (*i.e.*, $x_1^* = x_2^* = 0$).

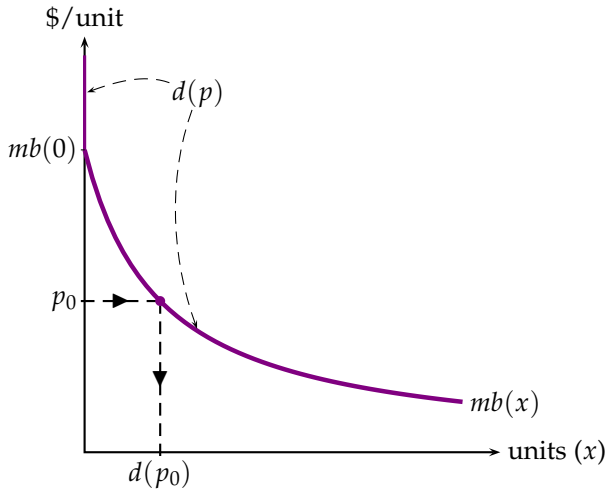

Figure 3.5: The demand curve, shown as a violet curve, corresponds to marginal benefit (thicker curve) for $p \leq mb(0)$ and corresponds to the vertical axis (thinner line) for $p > mb(0)$.

benefit schedule, which is decreasing, the amount demanded is less the greater is the price. That is, if p and p' are two prices, $p' > p$, then

$$d(p') < d(p) \text{ if } d(p) > 0\,; \text{ otherwise}$$
$$d(p') = d(p) \text{ if } d(p) = 0\,.$$

Consistent with Figure 3.5, we describe this by saying that demand curves slope down.[2]

Note: *Demand curves slope down.*

Example 13: Suppose that an individual's benefit schedule for a particular good is $10x - x^2$; that is, $b(x) = 10x - x^2$. Using Proposition 9, we see that

$$mb(x) = 10 - 2x\,.$$

Observe $mb(0) = 10$. Let's derive this individual's demand curve. For $p > 10$, he purchases nothing; that is, $d(p) = 0$ for all $p > 10$. For $p \leq 10$, we need to invert the marginal benefit schedule. Observe that if

$$p = mb(x) = 10 - 2x\,,$$

then

$$mb^{-1}(p) = \frac{10 - p}{2} = 5 - \frac{p}{2}\,.$$

We can conclude that

$$d(p) = \begin{cases} 5 - p/2\,, \text{ if } p \leq 10 \\ 0\,, \text{ if } p > 10 \end{cases}\,.$$

Properties of demand: Complements and substitutes

Our derivation of the individuals demand curve has held constant a number of factors. In particular, it has assumed that the prices of other goods has remained fixed. But what happens to demand for one good if the price of another good changes? The answer depends on whether the two goods are complements or substitutes.[3]

Two goods are *complements* if they are goods that tend to be consumed Examples of such pairs would be chips and salsa, PCs and operating systems, and whips and chains. Some pairs of goods are complements for some people (*e.g.*, orange juice and vodka), but not for other people (*e.g.*, vodka tonic drinkers).

Complements: *Goods consumed together.*

Two goods are *substitutes* if they are goods that tend to be seen as alternatives. Examples would include DVDs and in-theater movies, beer and wine, and email and telephone calls.

Substitutes: *Goods that are alternatives.*

[2] If you've taken economics before and heard of Giffen goods, forget about them—they are a theoretical possibility that have never been shown to exist in reality. If you dont know what a Giffen good is, consider yourself lucky.

[3] Point about spelling: Note that "complement" has no "i." Compliment with an "i" is a different word (as a verb it means to praise or give for free; as a noun it means a statement of praise).

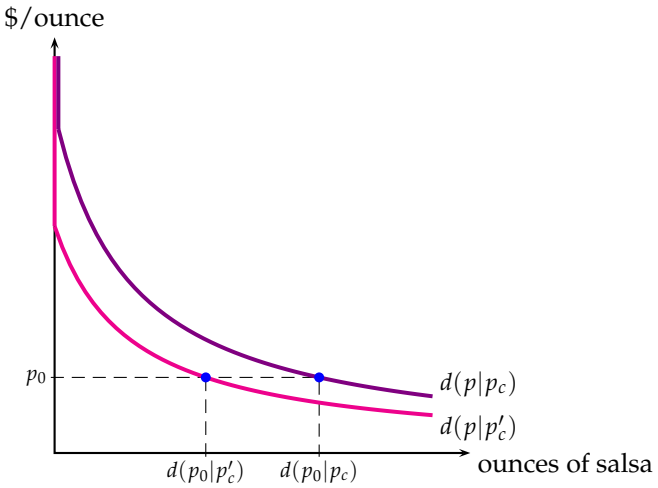

Figure 3.6: As the price of chips, the complementary good, increases from p_c to p'_c, the demand curve for salsa shifts in (goes from being the violet curve, $d(\cdot|p_c)$, to being the magenta curve, $d(\cdot|p'_c)$.

If two goods, say chips and salsa, are complements and the price of one, say chips, goes up, then the demand for the other at a given price will tend to be less. That is, if p_c and p'_c are two prices for chips, $p_c < p'_c$, then

$$d(p_s|p'_c) \leq d(p_s|p_c)\,,$$

where $d(p_s|\hat{p}_c)$ means the amount of salsa demanded as a function of the price of salsa, p_s, given that the price of chips is $\hat{p}_c$. As illustrated in Figure 3.6, we can describe this as saying that an increase in the price of the complementary good causes the demand curve for the other good to shift in (equivalently, shift to the left). The reason for this is that the benefit of consuming salsa depends on the number of chips one consumes. If, because the price of chips has increased, one will consume fewer chips, then one has less demand for salsa. Another way to put this is that the marginal benefit of an ounce more salsa is less the fewer chips on which it can go.

If two goods, say beer and wine, are substitutes and the price of one, say beer, goes up, then the demand for the other at a given price will tend to be greater. That is, if p_b and p'_b are two prices for beer, $p_b < p'_b$, then

$$d(p_w|p'_b) \geq d(p_w|p_b)\,,$$

where $d(p_w|\hat{p}_b)$ means the amount of wine demanded as a function of the price of wine, p_w, given that the price of beer is $\hat{p}_b$. Thinking of Figure 3.6 "in reverse," we can describe this as saying that an increase in the price of the substitute good causes the demand curve for the other good to shift out (equivalently,

shift to the right). The reason for this is that the benefit of consuming wine depends on the amount of other alcohol one can consume. If one is buying less beer because the price of beer has gone up, then the demand for other alcohol (*e.g.*, wine) will go up.

> **Example 14 [Fish and Bushmeat: Economics meets Ecology]:** Because of various European Union (EU) policies,[4] there was more fishing by European fleets off the west African coast. This led to worse harvests for West African fishermen, including Ghanaian fishermen. Hence, the price of fish in Ghanaian markets went up. Bushmeat (hunting wild animals) is a substitute for fish (*i.e.*, is an alternative source of protein). Hence, the outcome of the EU's policy was a greater slaughter of African wildlife.

Properties of demand: Other shifters

Changes in the prices of complements and substitutes are not the only factors that can shift demand. For example, when a hurricane is coming, the benefit of having bottled water, plywood, and canned foods increases. Because benefit is the area under the marginal benefit schedule (recall Proposition 6 on page 43), a shift up in benefit must correspond to a shift up in marginal benefit. But, from Figure 3.6, it is clear that a shift up in marginal benefit is equivalent to a shift out in demand. That is why, for instance, we see more bottled water, plywood, and canned foods being demanded when a hurricane is coming.

Other changes, such as changes in technology, can also shift demand curves. The advent of the car, for instance, greatly lessened the benefit of buggy whips, which meant marginal benefit fell, which was equivalent to the demand for buggy whips shifting in. Likewise, the advent of USB-port jump (flash) drives has caused the demand for diskettes (floppies) to shift in (indeed, disappear).

Some technologically driven demand shifts, such as that caused by jump drives, can also be seen as examples of substitutes at work. If a product doesn't exist, thats equivalent to its price being infinity. The introduction of jump drives can be seen as a (dramatic) fall in the price of jump drives. Jump drives and diskettes are substitutes. Hence, not surprisingly, the advent of jump drives caused demand for diskettes to shift in.

Properties of demand: Income effects

Suppose your income doubled. One response might be that you eat more meals out than you did before. Of course, if you eat more meals out, then you are eating fewer meals at home, which means you're buying fewer items to cook at home (*e.g.*, frozen dinners). In terms of your demand curves, a doubling of your income causes your demand curve for restaurant meals to shift out and your demand curve for frozen dinners to shift in.

---

[4] This example is taken from a study by a Berkeley faculty member, Justin Brashares: Brashares, J.S., P. Arcese, M.K. Sam, P.B. Coppolillo, A.R.E. Sinclair, and A. Balmford. "Bushmeat Hunting, Wildlife Declines and Fish Supply in West Africa," *Science* Vol. 306 (2004), pp. 1180–1183.

A good for which the demand curve shifts out as income rises is called a *normal good*. A good for which the demand curve shifts in as income rises is called an *inferior good*. Inferior goods are those goods that are seen as less desirable than certain substitutes for them (*e.g.*, a meal at Rivoli is better than a Swanson frozen dinner). The problem is that, at a given level of income, one can afford only so many restaurant dinners. At a higher level of income, one can afford more, which crowds out the frozen dinners.[5]

This discussion of income effects highlights an omission in our discussion of demand to this point. What about the ability of an individual to afford things? That is, for example, if an individual's marginal benefit schedule intersects the price line at 100 units and a price of $10, but she only has $900 to spend, then it wouldn't make sense to say her demand is 100—she can't afford that many. The goods news is that there is a way to accommodate the issue of affordability into the analysis. The bad news is that it is rather complicated (see the following subsection). Fortunately, for most goods the affordability issue is generally not that important. The number of candy bars, for instance, that you buy varies little or not at all with your income. Another way to say this is that, for most goods, income effects are negligible. In this text, we will, therefore, ignore income effects.

$\int dx$ Utility Maximization

OPT We think of consumers as deriving utility—that is to say happiness—from the consumption of different goods. A general way to write utility for a potential customer is to index all the different products, such as Cinnamon Apple Cheerios, sports cars, broccoli, Warriors tickets, etc., by i and write utility as

$$U(q_1, q_2, \ldots, q_i, \ldots q_N)\,,$$

where N is the total number of different goods. Consumers also have *budget constraints*. That is, a consumer's total expenditure cannot exceed his income, I. We write this as

$$p_1 q_1 + \cdots + p_N q_N \leq I\,,$$

where p_i is the price of a unit of the ith good. Note the lefthand side could be written as $\sum_{i=1}^{N} p_i q_i$.

The consumer wishes to maximize his wellbeing—his utility— subject to his budget constraint. This can be done by Lagrange maximization: Define λ to be the Lagrange multiplier or shadow value of income. That is, λ is the value of the marginal dollar in terms of

[5] Alternatively, one might stay home, but have more steaks and fewer frozen dinners as income rises

utility (*i.e.,* the increase in utility if income were increased by a dollar). Then this constrained maximization program can be written as

$$\max_{\{q_1,\ldots,q_N\}} U(q_1,\ldots,q_N) + \lambda\left(I - \sum_{i=1}^{N} p_i q_i\right).$$

Although we could solve this expression (the Lagrangean) directly, it is easier to use some intuition. If someone is going to pay a dollar to consume more of one good, it cannot be the case that spending that dollar on another good will yield more utility. The amount an additional dollar buys of good i is $1/p_i$. If MU_i is the marginal utility from the ith good (i.e., $MU_i = \partial U/\partial q_i$), then $MU_i \times 1/p_i$ (marginal utility times quantity) is the amount of additional utility derived from a dollar spent on the ith good. If the consumer has maximized his utility, it cannot be that moving the marginal dollar from one good to another can increase his utility. That is, we have for any two goods i and j,

$$MU_i \times \frac{1}{p_i} \not> MU_j \times \frac{1}{p_j} \text{ and}$$

$$MU_j \times \frac{1}{p_j} \not> MU_i \times \frac{1}{p_i}.$$

But this just says that for any pair of good it must be that

$$\frac{MU_i}{p_i} = \frac{MU_j}{p_j} \tag{3.11}$$

if the individual is maximizing his utility subject to his budget constraint.[6] Equation (3.11) says that the benefit of consuming one good divided by the cost (sometimes called the "bang for the buck") must be equal to the same ratio (bang for the buck) for any other good.

Your bang for the buck is the value, in terms of utility, of the marginal dollar of income. That is, it is the shadow value of income. We've thus derived:

$$\frac{MU_i}{p_i} = \lambda,$$

or, rewriting,

$$MU_i = \lambda p_i. \tag{3.12}$$

Observe that we have a function of quantity only on the lefthand side and a function of price on the righthand side. If, as we are

[6] To be precise, this holds only for goods for which the consumer actually purchases positive amounts.

maintaining, the consumer gets a diminishing marginal utility (or benefit) from each good, then MU_i is an invertible function of q_i. If we invert it, we get

$$q_i = MU_i^{-1}(\lambda p_i)\,. \tag{3.13}$$

This relates the amount the consumer wants to the price of the good. So, holding everything else constant, this is a demand curve.

To square this analysis with our earlier derivation of individual demand from marginal benefit schedules, we need to assume that the utility function has the form

$$U(q_1,\ldots,q_N) = u(q_1,\ldots,q_{N-1}) + q_N\,.$$

There is no further loss of generality in normalizing the units of the Nth good so that the price per unit is \$1; that is, so $p_N = 1$.[7] Clearly, $MU_N = 1$, hence, from expression (3.12) for $i = N$, we have $1 = \lambda \times 1$; that is, the shadow value of income, λ, is 1. Observe, therefore, that λ does not depend on income, I. Holding constant the amount of the goods other than i in expression (3.12), we have an expression that relates marginal benefit to price; that is, condition (3.9). Inverting— employing expression (3.13)—gives us

$$q_i = MU_i^{-1}(p_i) = d_i(p_i)$$

(recall $\lambda = 1$). Observe this means that there are no income effects, therefore, in the demand for good i.

From individual to aggregate demand

For simple pricing, what a firm cares about is not each individual's demand but *aggregate demand*, the amount that all individuals want at a given price. Aggregating all the individual demand curves gives the *aggregate demand curve*.

Aggregating demand is straightforward. If Rose and Noah are the only two consumers and Rose wants 3 units at a given price and Noah wants 5 units at a given price, then aggregate demand at that price is 8. Similarly, if Rose's demand curve is $d_R(p)$ and Noah's is $d_N(p)$, then the aggregate demand curve, $D(p)$, is given by

$$D(p) = d_R(p) + d_N(p)\,.$$

In general, if we have N consumers, indexed by n, then the aggregate demand curve is given by

$$D(p) = d_1(p) + \cdots + d_N(p) = \sum_{n=1}^{N} d_n(p)\,. \tag{3.14}$$

[7] The Nth good is the *numéraire* good.

Example 15: Suppose that there are two types of consumers. There are 1000 people of type A, and each of these people have a demand for your product given by

$$d_A(p) = \begin{cases} 150 - 3p\,, \text{ if } p \leq 50 \\ 0\,, \text{ if } p > 50 \end{cases}\,.$$

There are an additional 2000 people of type B, each of whom has a demand for your product given by

$$d_B(p) = \begin{cases} 1000 - 2p\,, \text{ if } p \leq 500 \\ 0\,, \text{ if } p > 500 \end{cases}\,.$$

The aggregate demand of type-A consumers is

$$D_A(p) = \sum_{n=1}^{1} 000 d_A(p) = 1000 d_A(p) = \begin{cases} 150,000 - 3000p\,, \text{ if } p \leq 50 \\ 0\,, \text{ if } p > 50 \end{cases}\,.$$

The aggregate demand of type-B consumers is

$$D_B(p) = \sum_{n=1}^{2} 000 d_B(p) = 2000 d_B(p) = \begin{cases} 2,000,000 - 4000p\,, \text{ if } p \leq 500 \\ 0\,, \text{ if } p > 500 \end{cases}\,.$$

So overall aggregate demand is

$$D(p) = D_A(p) + D_B(p)$$

$$\begin{cases} 150,000 - 3000p + 2,000,000 - 4000p = 2,150,000 - 7000p\,, \text{ if } p \leq 50 \\ 0 + 2,000,000 - 4000p = 2,000,000 - 4000p\,, \text{ if } 50 < p \leq 500 \\ 0 + 0 = 0\,, \text{ if } 500 < p \end{cases}\,.$$

3.6 Demand Elasticity

For reasons that will become clear later, it is useful to define a quantity known as the *elasticity of demand*. The elasticity of demand, ϵ_D, is defined to be

$$\epsilon_D = -1 \times \text{slope of demand at } p \times \frac{p}{D(p)}\,. \tag{3.15}$$

Because demand curves slope down, we multiply by -1 to make $\epsilon_D > 0$.

What ϵ_D is telling us is the percentage change in quantity demanded per percentage change in price. That is, it can be shown that,

$$\epsilon_D - 1 \times \left(\frac{\Delta D(p)}{D(p)} \times 100\%\right) \div \left(\frac{\Delta p}{p} \times 100\%\right)\,,$$

where Δ denotes "change in."[8]

[8] $\int dx$ Recall that the slope of $D(\cdot)$ is denoted as $dD(p)/dp$. Hence, equation (3.15) can be rewritten as $-dD(p)/dp \times p/D(p)$. Rearranging, we get the calculus analog of (3.6): $\epsilon_D = -dD(p)/dp \div dp/p$.

3.7 Revenue and Marginal Revenue under Simple Pricing

We can now go from demand to a revenue function for a firm that engages in simple pricing.

Derivation of revenue

Consider a firm that faces aggregate demand $D(p)$. What this says is that if it charges a price of p, then it will sell $D(p)$ units. Its revenue is price times units sold, or $pD(p)$. Observe that, if it were to raise its price, two things would happen. It would earn more per unit (that's the good news). But it would sell fewer units—demand curves, recall, slope down (that's the bad news).

Our analysis of profit maximization was done in terms of quantity, whereas we have just derived an expression for revenue in terms of p. To utilize our earlier analysis, we need to convert $pD(p)$ into an expression that is expressed in terms of quantity. Observe that $D(p)$ is a quantity, call it q. Because demand curves slope down, $D(\cdot)$ is invertible. We can, thus, find a unique value of p that solves the equation

$$q = D(p)\,.$$

Let $P(q)$ be the value of p that solves that equation. Because $P(\cdot)$ is the inverse of the demand function, we call it the *inverse demand function* (alternatively, inverse demand curve or inverse demand schedule). So, substituting P(q) for p and q for D(p), we can rewrite revenue in terms of quantity:

$$R(q) = P(q)q\,.$$

Derivation of marginal revenue

To find the profit-maximizing quantity, we want to use the $MR = MC$ rule. But this requires us to know MR.

If the firm wishes to increase q, it will have to lower price. Put another way, because demand curves slope down, $P(\cdot)$ is a decreasing function—if q goes up, it goes down. We therefore face the same "good-news-bad-news" situation noted above. The good news from increasing q is that we sell more units. The bad news is that we can only do so at a lower price on all units sold. This second effect is sometimes referred to as the *driving-down-the-price effect*.

Consider a very small positive increment in the quantity to be sold, dq. As just discussed, this will result in a small change in price, dp. Because demand curves slope down, $dp < 0$. Figure 3.7 illustrates. The small change in revenue, dR, is captured by the two shaded areas. Because the firm is selling dq more units, it adds $dq \times \big(P(q) + dp\big)$ to its revenue. This is the gray area labelled A. That's the good news. The bad news is the driving-down-the-price effect. Price is changed by dp. So, on all the q units it would have sold had it not tried to increase sales, it gets $-dp$ less (recall $dp < 0$). So it loses revenue equal to

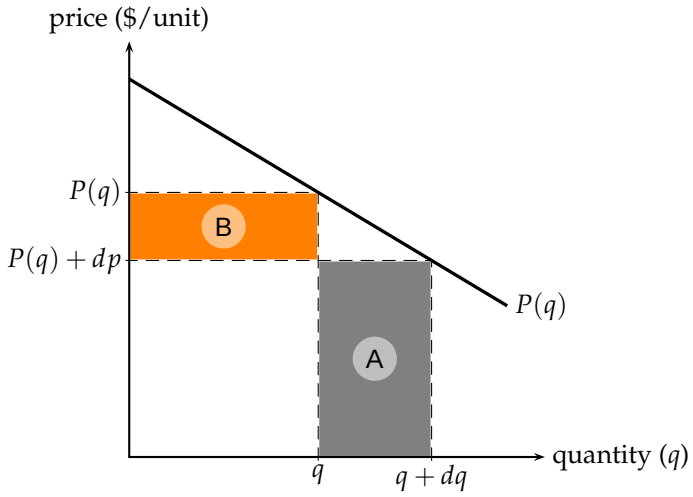

Figure 3.7: Increasing units sold by dq changes price by $dp < 0$. The firm gains, in revenue, the gray area, labeled A, but loses the orange area, labeled B. The latter area represents the driving-down-the-price effect.

the orange area labelled , which is $dp \times q$. Hence, the change in revenue, dR, is given by

$$dR = dq \times \big(P(q) + dp\big) - (-dp \times q) = dq \times \big(P(q) + dp\big)\,.$$

Dividing the end terms by dq, we get

$$\frac{dR}{dq} = P(q) + dp + \frac{dp}{dq}q\,.$$

Now a small change in revenue per unit change in quantity is just marginal revenue; that is, $dR/dq = MR(q)$. Similarly, a small change in the price per unit change in quantity is just the slope of the inverse demand curve. We thus have

$$\begin{aligned} MR(q) &= P(q) + dp + q \times \text{slope of } \textit{inverse} \text{ demand at } q \\ &= P(q) + dp + qP'(q)\,. \end{aligned}$$

Note the use of the notation $P'(q)$ to denote the slope of the inverse demand schedule at q. The quantity dp is very small, essentially zero, so it can be ignored. In fact, in the limit, as we consider an infinitesimally small change in quantity, it is zero. We've thus arrived at the formula for marginal revenue

***MR* under simple pricing:** $P(q) + qP'(q)$.

under simple pricing:[9]

Proposition 15. $MR(q) = P(q) + qP'(q)$.

The $qP'(q)$ term, which is negative—demand curves slope down—represents the driving-down-the-price effect.

Example 16: Suppose a firm faces demand given by

$$D(p) = 500,000 - 2500p\,.$$

To calculate marginal revenue, (i) we need to calculate *inverse* demand, then (ii) we need to employ Proposition 15.

Inverting demand, we have:

$$q = 500,000 - 2500 \times P(q)\,;\ \text{hence, } P(q) = 200 - \frac{q}{2500}\,.$$

Observe the second expression is a line in slope-intercept form (see Section A1.5) and the slope, $P'(q)$, is $-1/2500$ (it is negative of course because demand curves slope down). Using Proposition 15 yields:

$$MR(q) = \underbrace{\left(200 - \frac{q}{500}\right)}_{P(q)} + \underbrace{\left(-q\frac{1}{2500}\right)}_{qP'(q)}$$

$$= 200 - \frac{2q}{2500} = 200 - \frac{q}{1250}\,.$$

Observe the following, *all of which are general results*,

- If the inverse demand schedule is linear, then the marginal revenue schedule is also linear.
- The inverse demand schedule and the marginal revenue schedule share the same intercept (in this example, 200).
- If the inverse demand schedule is linear, then the slope of the marginal revenue schedule is twice the slope of the inverse demand schedule ($-1/1250$ versus $-1/2500$, respectively, in this example).

Marginal revenue and elasticity

The fact that marginal revenue under simple pricing has both a good-news term and a bad-news term raises the question of whether the bad-news term could dominate the good-news term. That is, could marginal revenue ever be negative? The answer is yes. Moreover, as we will show, the sign of marginal revenue is related to elasticity of demand, ϵ_D.

[9] $\boxed{\int dx}$ A quick derivation via calculus: $R(q) = qP(q)$. Differentiating, we have $MR(q) = P(q) + qP'(q)$; where the derivative of the righthand side follows by the product rule (see Proposition 32 on page 152).

Because price is positive, $MR(q)/P(q)$ has the same sign as $MR(q)$. Observe, from Proposition 15, that

$$\frac{MR(q)}{P(q)} = 1 + P'(q)\frac{q}{P(q)}. \tag{3.16}$$

Recall that $P'(q)$ is the slope of *inverse* demand, $q = D(p)$, and $P(q) = p$. Hence, we can rewrite expression (3.16) as

$$\begin{aligned}\frac{MR(q)}{p} &= 1 + \text{slope of } \textit{inverse} \text{ demand at } q \times \frac{D(p)}{p} \\ &= 1 + \frac{1}{\left(\text{slope of } \textit{inverse} \text{ demand at } p \times \frac{p}{D(p)}\right)}.\end{aligned}$$

The denominator in the last expression is -1 times the elasticity of demand (see expression (3.15) above). So we have

$$\frac{MR(q)}{p} = 1 - \frac{1}{\epsilon_D}. \tag{3.17}$$

Observe this last expression implies

Proposition 16. *If elasticity is greater than 1 ($\epsilon_D > 1$), then marginal revenue is positive; if it equals 1 ($\epsilon_D = 1$), then marginal revenue is zero; and if it is less than 1 ($\epsilon_D < 1$), then marginal revenue is negative.*

When $\epsilon_D > 1$, we say that we are on the *elastic* portion of demand. When $\epsilon_D = 1$, we say that we are on the *unitary elastic* of demand. When $\epsilon_D < 1$, we say that we are on the *inelastic* portion of demand.

When $\epsilon_D > 1$, a one-percentage-point change in price causes a more than one percent change in quantity, which is why demand is called elastic in this case. Because revenue is price times quantity, this tells us that, on the elastic portion of demand, revenue goes up if we lower price (increase quantity); and revenue goes down if we raise price (decrease quantity). When $\epsilon_D < 1$, a one-percentage-point change in price causes less than a one-percent change in quantity, which is why demand is called inelastic in this case. Moreover, this means that, on the inelastic portion of demand, revenue goes down if we lower price (increase quantity); and it goes up if we raise price (decrease quantity).

Observe that a firm engaged in simple pricing would never want to operate on the inelastic portion of demand. If it were on the inelastic portion, then it could raise price (equivalently, cutback output), which would raise revenue. Moreover, because it is producing less, it would also be reducing cost. Any move that both raises revenue and cuts cost is a winning move—so, at any point on the inelastic portion of demand, it would want to raise price, which means it would never operate on the inelastic portion.

Proposition 17. *A firm engaged in simple pricing does* not *choose its price or output so as to be on the inelastic portion of its demand curve.*

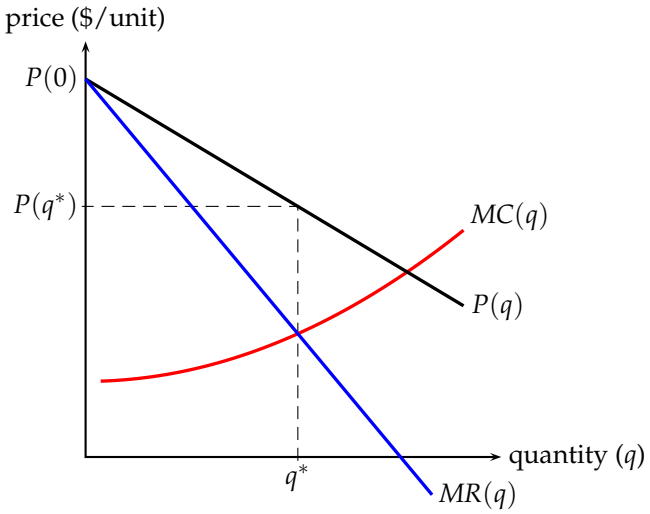

Figure 3.8: The profit-maximizing quantity, q^*, is determined by the intersection of the marginal-revenue schedule (in blue) and the marginal-cost schedule (in red). Note that *MR* crosses *MC* once and from above. The profit-maximizing price, $P(q^*)$, is then read off the inverse demand curve (in black) at the profit-maximizing quantity, q.

3.8 The Profit-Maximizing Price

We now have all the ingredients in place to determine the profit-maximizing price.

Under fairly general conditions, the marginal-revenue schedule under simple pricing is downward sloping. From Example 16, it will certainly be downward sloping if demand is linear.

Typically, we can expect the marginal-cost schedule to be either relatively flat or increasing. Either way, if the marginal-revenue schedule is decreasing, this means that, if the schedules cross at all, the marginal-revenue schedule crosses the marginal-cost schedule once and from above. Observe that, if $P(0) > MC(0)$, the curves will cross given their predicted slopes.[10]

Assuming, therefore, that the $MR = MC$ rule is sufficient, as well as necessary, for determining the profit-maximizing quantity, we need to translate that into a price. This is straightforward. If q^* solves the expression $MR(q^*) = MC(q^*)$, then the profit-maximizing price is $P(q^*)$. Figure 3.8 illustrates.

[10] $\boxed{\int dx}$ If, because of an overhead cost, $MC(0)$ is not defined, then this condition can be restated as $P(0) > \lim_{h \downarrow 0} MC(h)$.

The firm's profit is revenue minus cost. So the maximum profit is

$$\pi(q^*) = q^* P(q^*) - C(q^*)\,.$$

We do need, of course, to check the shutdown rule; that is, make sure the firm should operate at all. Recall that, under simple pricing, average revenue is price. Hence, the condition for operating at all, expression (3.7), becomes

$$P(q^*) \geq AC(q^*)\,.$$

Example 17 [From start to finish]: Consider a firm that faces demand

$$D(p) = 1,000,000 - 50,000p\,.$$

Suppose its costs are given by

$$C(q) = \begin{cases} 0\,, \text{ if } q = 0 \\ 6q + 1,400,000\,, \text{ if } q > 0 \end{cases}\,.$$

For $q > 0$, it is readily seen that $MC(q) = 6$ (recall Proposition 3 on page 39). To derive $MR(\cdot)$, we need, first, to calculate inverse demand. Then, we need to calculate marginal revenue using Proposition 15. Letting $q = D(p)$ and $p = P(q)$, we have

$$\begin{aligned} D(p) &= 1,000,000 - 50,000p\,; \text{ hence,} \\ q &= 1,000,000 - 50,000P(q)\,. \end{aligned}$$

So, solving for $P(q)$, we have

$$P(q) = 20 - \frac{q}{50,000}\,.$$

Observe the slope is $-/50,000$. Hence, Proposition 15 tells us that

$$MR(q) = \left(20 - \frac{q}{50,000}\right) + q\frac{-1}{50,000} = 20 - \frac{q}{25,000}\,.$$

Observe, as we knew had to be the case from Example 16, the MR schedule has the same intercept as the inverse demand curve and has twice its slope.

Because (i) $P(0) = 20 > 6 = MC$, (ii) MC is flat, and (iii) $MR(\cdot)$ downward sloping, we see that the MR schedule crosses the MC schedule once, from above. That is, the only candidate for the profit-maximizing quantity (unless the firm should shutdown) is the value of q that solves $MR(q) = MC(q)$. To find that q:

$$20 - \frac{q}{25,000} = 6\,;$$

hence, solving for q, we obtain $q = (20 - 6) \times 25,000 = 350,000$. That is, $q^* = 350,000$.

To get the profit-maximizing price, we insert that q^* into $P(\cdot)$:

$$P(q^*) = 20 - \frac{350,000}{50,000} = 13\,.$$

So, if it is optimal for it to operate, the profit-maximizing price is \$13.

Should the firm operate? To answer, we need to determine whether \$13 is greater than $AC(q^*)$. Observe

$$AC(q) = 6 + \frac{1,400,000}{q}\,.$$

Therefore,

$$AC(q^*) = 6 + \frac{1,400,000}{350,000} = 6 + 4 = 10\,.$$

Because \$13 > \$10, the firm should produce.

Finally, its profit is revenue minus cost

$$\pi(350,000) = \underbrace{13 \times 350,000}_{\text{revenue}} - \underbrace{(6 \times 350,000 + 1,400,000)}_{\text{cost}} = 1,050,000\,.$$

The firm's profit, if it prices correctly, is \$1,050,000.

3.9 The Lerner Markup Rule

We can relate the price markup over marginal cost to the elasticity of demand. The markup is $p^*MC(q^*)$, where we have used p^* to denote the profit-maximizing price (*i.e.*, $p^* = P(q^*)$). What we want to do is determine what proportion of price is markup; that is, we want to calculate

$$\frac{p^* - MC(q^*)}{p^*}\,.$$

Recall, earlier, that we established that

$$MR(q) = p \times \left(1 - \frac{1}{\epsilon_D}\right) \tag{3.18}$$

(see expression (3.17) above). We also know that $MR = MC$ at the profit-maximizing quantity. Hence, we have

$$\begin{aligned}
\frac{p^* - MC(q^*)}{p^*} &= \frac{p^* - MR(q^*)}{p^*} && (MR = MC \text{ at profit-max'ing quantity}) \\
&= \frac{p^* - p^* \times \left(1 - \frac{1}{\epsilon_D}\right)}{p^*} && (\text{expression (3.18)}) \\
&= \frac{1}{\epsilon_D}\,.
\end{aligned}$$

We have just established the Lerner markup rule:

Proposition 18 (Lerner markup rule)**.** *Under simple pricing, at the profit-maximizing price and quantity, the proportion of the price that is markup over marginal cost is* $1/\epsilon_D$*; that is,*

$$\frac{p^* - MC(q^*)}{p^*} = \frac{1}{\epsilon_D}.$$

The Lerner markup rule is useful for optimally adjusting price following small changes in cost.

Example 18: Due to a rise in the price of one of its raw materials, a company finds its marginal cost of production has increased from \$10 per unit to \$10.50 (a 5% increase). Data suggests that elasticity of demand at current price is 1.25. Assuming the company's old price was optimal, what should its new price be (approximately)?[11] Using the Lerner rule we have

$$\frac{p - 10.50}{p} = \frac{1}{1.25} = .8.$$

Solving for p, we have $p = \$52.50$; that is, the new price should be \$52.50. (Can you determine what the *old* price was?)

Example 19: Suppose that your firm's marginal cost increases by 2%. By what percentage (approximately) should you raise your price, assuming you were initially pricing optimally?

To determine this, let δ denote the proportion by which you should increase your price; that is, $\delta \times 100\%$ is the percentage by which you should increase your price. Let p and c denote your original price and marginal cost, respectively. The Lerner rule tells you that

$$\frac{p - c}{p} = \frac{1}{\epsilon_D}.$$

Your new price, which is $p + \delta p$, must also satisfy the Lerner markup rule when your marginal cost is $c + .02c = 1.02c$:

$$\frac{p + \delta p - 1.02c}{p + \delta p} = \frac{1}{\epsilon_D}.$$

If two things equal the same third thing, then the two things equal each other:

$$\begin{aligned}\frac{p - c}{p} &= \frac{p + \delta p - 1.02c}{p + \delta p} \\ &= \frac{(1 + \delta)p - 1.02c}{(1 + \delta)p}.\end{aligned}$$

[11] Why "approximately"? Because elasticity is typically not a constant. If we adjust the price, we're on a different part of the demand curve, so the elasticity will be different. For small movements in price, however, the change in elasticity is very small, so treating it as constant for small price changes yields a good approximation.

Observe that $1 + \delta = 1.02$ solves this last expression (since, then, the common 1.02 can be canceled from numerator and denominator of the last expression). It follows, therefore, that $\delta = .02$—you should raise your price 2%. This is general: For *small* increases in marginal cost, the *approximate* best response is to increase price by the same percentage.

Finally, if elasticity were 1.5, then by increasing your price 2%, you could expect an approximate reduction in units sold of 3%.

Summary | 3.10

The focus of this section was on simple pricing; that is, charging the same price for all units and to all consumers. Before we could determine the profit-maximizing price, however, we needed to develop a suitable toolkit.

First, we had to work out how a firm maximizes profits. This involves a number of steps:

1. Determine the marginal-revenue and marginal-cost schedules.
2. Solve for the quantity that equates marginal revenue and marginal cost. This quantity is a *candidate* for being the profit-maximizing quantity. If the marginal-revenue schedule crosses the marginal-cost schedule once and from above, then this quantity is the profit-maximizing quantity if the firm should produce at all.
3. Check whether the firm should shutdown. That is, check that overhead costs are not so great as to make shutting down the preferable strategy.

Next, we investigated how individual preferences translated into demand. By maximizing their consumer surplus, individuals' determine how much they demand at any given price. Aggregating individual demand curves yields the firm's aggregate demand curve.

Knowing demand, we could then derive marginal revenue under simple pricing. Under simple pricing, marginal revenue has two components: a "good-news" component that points toward increased revenue if more units are sold; and a "bad-news" component—the driving-down-the-price effect—that arises because the price received on *all* units must be lowered to sell more units.

Once marginal revenue is determined, we can employ the steps outlined above to determine the profit-maximizing price.

Lastly, we considered the elasticity of demand and how it related to the profit-maximizing price via the Lerner markup rule.

4 Advanced Pricing

In this chapter, we consider other forms of pricing. Although simple pricing (sometimes called *linear pricing* or *uniform pricing*) is quite prevalent, it is neither the only nor the most desirable form of pricing. On your trips to the supermarket, you have no doubt noticed that the liter bottle of soda does not cost twice as much as the half-liter bottle of soda. In your purchases of airplane tickets, you have no doubt noticed that how much you pay is a function of what restrictions you are willing to accept (*e.g.*, staying over a Saturday night). Some goods and services you buy may have entry fees (*e.g.*, club membership) and per-use charges (*e.g.*, greens fees). All of these are examples of *price discrimination* or *nonlinear pricing*.

4.1 What Simple Pricing Loses

To motivate other forms of pricing, it helps to know what simple pricing's deficiencies are. Basically, there are two. First, simple pricing leaves "money on the table," in a sense to be made formal below. Second, it leaves potential profit in the hands of consumers in the form of their consumer surplus.

What do we mean that simple pricing "leaves money on the table"? Consider Figure 4.1. It reproduces Figure 3.8. Now suppose the firm in question wishes to sell one more unit. Observe that if it could sell that additional unit, the $q^* + 1$st unit, with*out* having to lower the price on the other q^* units, then it would profit by doing so. Some consumer is willing to pay the firm $P(q^* + 1)$ for the $q^* + 1$st unit and the incremental cost to the firm of producing that $q^* + 1$st unit is $MC(q^* + 1)$. As Figure 5.1 shows, $MC(q^* + 1) < P(q^* + 1)$. In other words, *if* the firm could charge $P(q^* + 1)$ for the $q^* + 1$st unit, but still charge $P(q^*)$ for the other q^* units, it would gain $P(q^* + 1) - MC(q^* + 1)$ in profit. The problem is that, under simple pricing, it can*not* charge a price for the $q^* + 1$st unit that is different than the price it charges for the other q^* units. And the fact that it would, under simple pricing, have to lower the price on all q^* units to induce someone to buy the $q^* + 1$st unit, makes selling that $q^* + 1$st unit undesirable.

In essence, then, the firm is leaving $P(q^* + 1) - MC(q^* + 1)$ on the table; that is, it is a potential sale that would be profitable except for the driving-down- the-price effect of simple pricing. Observe, too, that we could repeat this analysis for the $q^* + 2$nd unit, the $q^* + 3$rd unit, and so forth until we reach the unit at which inverse demand falls below marginal cost (obviously, we would

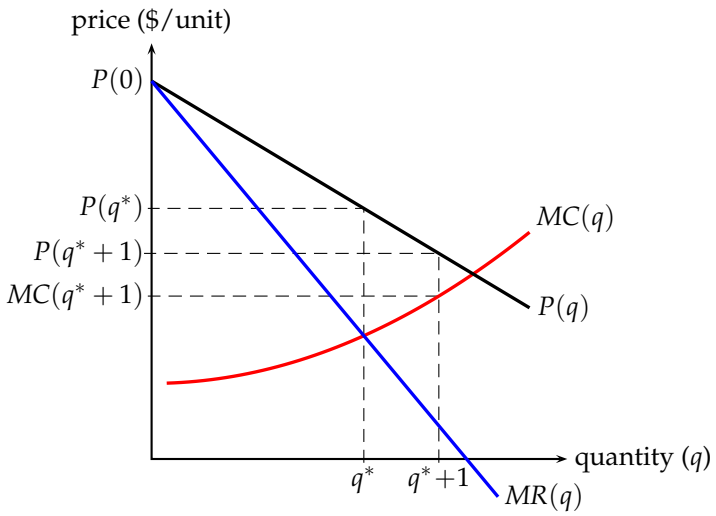

Figure 4.1: Were it not for the consequent driving-down-the-price effect, the firm would find it profitable to sell the q^*+1st unit at price $P(q^*+1)$.

never want to sell the qth unit if $P(q) < MC(q)$). Treating units as continuous, we have thus shown that the firm forgoes the area labeled deadweight loss in Figure 4.2. This region is called the *deadweight loss* because it corresponds to the potentially profitable sales that don't get made.

What these two figures show is that if a firm could devise a way to price that avoided the driving-down-the-price effect, then it could potentially increase its profits over those it can earn using simple pricing.

We also observed that simple pricing leaves profits, in the form of consumer surplus, in the hands of consumers. The firm would also like to capture some of this, as well. To understand this second "loss" from simple pricing, however, we will need digress and show how individual consumer surplus aggregates to overall consumer surplus.

4.2 Aggregate Consumer Surplus

Our objectives in this section are (i) to reconsider individual consumer surplus and (ii) show how it aggregates into overall or total consumer surplus.

Recall, from our derivation of individual demand (see pages 59–63), that an individuals consumer surplus is

$$cs(q) = b(q) - pq$$

when she pays p per unit. The function$b(\cdot)$, recall, is her benefit function and

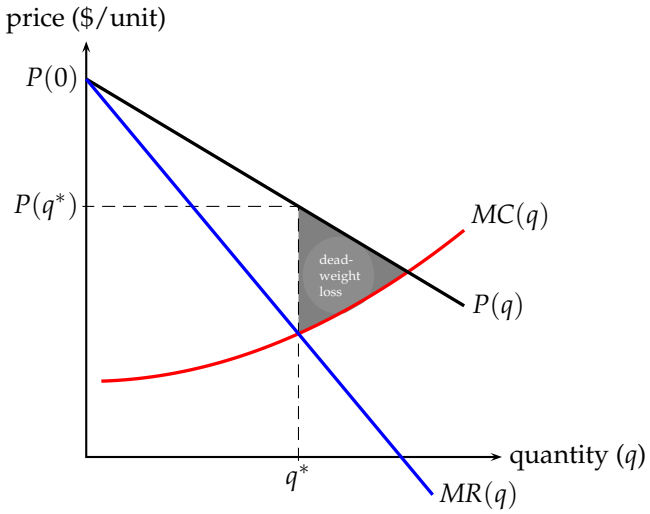

Figure 4.2: The shaded triangular region is the deadweight loss induced by simple pricing.

we interpret $b(q)$ to be her benefit, measured in dollars, of having q units of the good in question. Also remember that, for $q > 0$, her demand curve and her marginal benefit schedule are coincident (see Figure 3.5). For $q = 0$, her demand curve is coincident with the vertical (price) axis. Lastly, remember, Proposition 6 on page 43: The area under a marginal curve is the total function. Observe, therefore, that a consumers total benefit from q units, $b(q)$, is the area under her marginal benefit curve from 0 to q units.[1] Her marginal cost is just p, so her total cost of q units is the area under the price line (the horizontal line of height p) from 0 to q units; that is, pq. Graphically, if we subtract the area under the price line from the area under the marginal benefit curve, then we get the area under the marginal benefit curve but above the price line. In other words, the individual's consumer surplus is the area beneath her marginal benefit curve and above the price line from 0 to q units. Figure 4.3 illustrates.

One way to think about the individual's consumer surplus is the following. What is it worth to the consumer to be able to buy q units at a price of p per unit? Well her "profit," her consumer surplus, is $cs(q)$. If she had to pay more than $cs(q)$ for the ability to purchase q units at p per unit, she wouldn't do it—she would lose. If she had to pay less than $cs(q)$, she would do it because she'd still come out ahead. Therefore, we can consider $cs(q)$ to represent the most that the consumer would be willing to pay for the right to buy q units at a price of p.

Consumer surplus: *The consumers value of the right to buy a given number of units at a given price per unit.*

An area is an area. It doesn't matter, therefore, if we view the individual's

[1] In case you were wondering why I wrote $b(q)$ and not $b(q) - b(0)$, the answer is $b(0) = 0$.

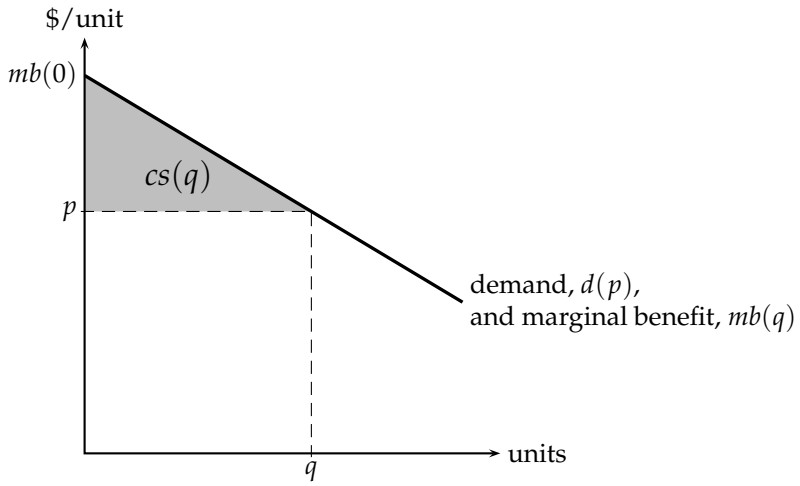

Figure 4.3: If we subtract the area of the rectangle whose lower left corner is the origin and whose upper right corner is (q, p) from the area beneath the marginal benefit curve, $mb(\cdot)$, from 0 to q units, we're left with the shaded triangle, which is the individual's consumer surplus, $cs(q)$, from q units purchased at p per unit.

consumer surplus as the area beneath the marginal benefit schedule, but above the price line, from 0 to q units or we view it as the area under her demand curve from p to $mb(0)$. Suppose, further, that we add to this area the area under her demand curve from $p = mb(0)$ to infinity. This doesn't change the amount of the original area because there is no area under the demand curve from $mb(0)$ to infinity (recall the demand line is coincident with the vertical axis in this interval). Hence, we don;t change the area shown in Figure 4.3 if we add on the area under her demand curve from $mb(0)$ to infinity (the reason for adding this zero will become clear later). We can, therefore, describe the individuals consumer surplus as follows.

Proposition 19. *At price p, an individual's consumer surplus is the area under her demand curve from p to infinity.*

In light of this result, in addition to writing her consumer surplus as $cs(q)$, we will also write it as $A_p^\infty(d)$, where this notation means the area under demand, $d(\cdot)$, from p to infinity.[2]

It makes sense to define aggregate consumer surplus as the sum of the individuals' consumer surpluses. Let CS denote aggregate consumer surplus. At a given price, p, we have, therefore, that

$$CS = A_p^\infty(d_1) + \cdots + A_p^\infty(d_N)\,,$$

where d_i denotes the demand schedule of the ith individual in a population of N consumers. The area operator, $A_p^\infty(\cdot)$, is linear, meaning that we can distribute it out of the summation. That is,

$$\begin{aligned} CS &= A_p^\infty(d_1) + \cdots + A_p^\infty(d_N) \\ &= A_p^\infty(d_1 + \cdots + d_N) && \text{(distributing out)} \\ &= A_p^\infty(D) && \text{(sum of individuals' demands is aggregate demand)} \end{aligned}$$

In other words, we've just established the following.[3]

Proposition 20. *Aggregate consumer surplus, CS, at price p is the area under the aggregate demand curve, $D(\cdot)$, between p and infinity.*

We also know that

[2] $\int \mathrm{d}x$ Observe $A_p^\infty(d) = \int_p^\infty d(z)dz$, where z is the dummy of integration.

[3] $\int \mathrm{d}x$ Observe,

$$CS = \sum_{i=1}^{N} \int_p^\infty d_i(z)dz = \int_p^\infty \left(\sum_{i=1}^{N} d_i(z) \right) dz = \int_p^\infty D(z)dz\,,$$

where the first equals follows because the summation operator can be passed through the integral (*i.e.*, integration is a linear operation).

$$\begin{aligned} CS &= cs_1(q_1) + \cdots + cs_N(q_N) \\ &= \big(b_1(q_1) - pq_1\big) + \cdots + \big(b_N(q_N) - pq_N\big) \\ &= \sum_{i=1}^{N} b_i(q_i) - \sum_{i=1}^{N} pq_i \\ &= \sum_{i=1}^{N} b_i(q_i) - pQ\,, \end{aligned}$$

where Q is the total amount demanded by all consumers at p; that is, $Q = D(p)$. Hence,

$$A_p^{\infty}(D) = CS = \sum_{i=1}^{N} b_i(q_i) - pQ\,. \tag{4.1}$$

Observe that we can also view $A_p^{\infty}(D) + pQ$ as the area under the aggregate *inverse* demand curve from 0 to Q units. From the last equation,

$$\text{Area under aggregate inverse demand} = \sum_{i=1}^{N} b_i(q_i) \equiv B(Q)\,,$$

where we define $B(Q)$, aggregate benefit, to be the sum of individual benefits. We have, thus, derived:

Proposition 21. *The area under the* inverse *demand curve, $P(\cdot)$, from 0 to Q units is the total or aggregate benefit, $B(Q)$, enjoyed by consumers from the Q units.*

If the area under the aggregate inverse demand curve is aggregate benefit, then it must be that the curve itself is marginal aggregate benefit (recall Proposition 6). In other words, if $MB(\cdot)$ is the marginal aggregate benefit schedule, then we've shown that

Proposition 22. *$P(q) = MB(q)$ for all $q > 0$.*

4.3 The Holy Grail of Pricing

Observe that the benefit generated by q units sold is $B(q)$. The cost to the firm of producing q units is $C(q)$. Hence, the total value created is $B(q)C(q)$. This difference is called total welfare and denoted by $W(q)$; that is,

$$W(q) = B(q) - C(q)\,.$$

If a firm sold q units, it could never hope to gain more profits that $W(q)$. Why? Because $W(q)$ is the total value created—the whole "pie" as it were—so if the firm were getting more than the whole pie, then consumers must be receiving less than zero. But consumers will opt not to buy rather than get less

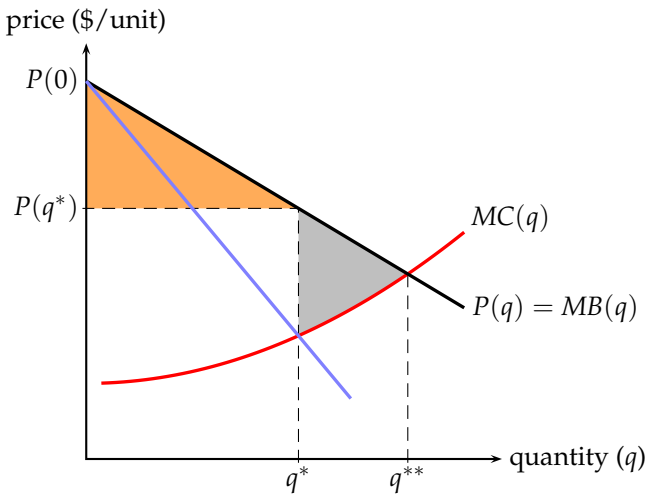

Figure 4.4: Welfare (total surplus) is maximized if q^{**} units trade.

than zero. Therefore, $W(q)$ represents the most that the firm could ever hope to capture.

Suppose the firm could capture all of $W(q)$. How much would it choose to produce? Looking at $B(\cdot)$ as being like $R(\cdot)$, our earlier analysis of simple pricing tells us that the candidate for the welfare-maximizing amount, q^{**}, is the one that solves

$$MB(q) = MC(q)\,.$$

From Proposition 22, this can be rewritten as

$$P(q^{**}) = MC(q^{**})\,.$$

We can summarize this as

Proposition 23. *If the firm can capture all the welfare generated from selling q units (i.e., $W(q)$), then the firm will want to produce to the point at which inverse demand intersects marginal cost.*

Figure 4.4, which replicates Figure 4.2, illustrates.

From Figure 4.4, we see that part of the gain the firm would enjoy, were it to capture maximized welfare, is the deadweight loss (the gray area). That is, were it able to capture all the welfare, it would be able to pick up the money left on the table by simple pricing. Furthermore, because the firm is capturing all the gains from trade (*i.e.*, all the value created), there can't be any left for the customers. That is the firm is also capturing all of the consumers' surplus (the orange area) as well.

As noted, producing q^{**} units and capturing all of welfare is the very best the firm could ever do. It is, therefore, the "Holy Grail" of pricing; it is what the firm would ideally like to do.

Because this outcome is so good, any form of pricing that achieves this Holy Grail is known as *perfect price discrimination*. For historic reasons, perfect price discrimination is also known as *first-degree price discrimination*.

4.4 Two-Part Tariffs

Can the firm ever obtain the Holy Grail? Generally, the answer is no. However, it can certainly get closer than it does by using simple pricing. In fact, if all consumers are the same in terms of their individual demand for the good, then the firm could actually obtain the Grail.

Two-part tariff: *Pricing with an entry fee and a per-unit charge.*

In this section, we explore a type of pricing called a *two-part tariff*. A two-part tariff can help get a firm closer to the Grail than can simple pricing. A two-part tariff is a pricing scheme (tariff) with, as the name indicates, two parts. One part is what's called the *entry fee*. It is the amount that the consumer must pay before she can buy any units at all. In this sense, it is like a consumer overhead charge. As a real-life example, consider it to be like the fee one pays to enter an amusement park.[4] The second part of a two-part tariff is the *per-unit charge*. It is the amount that the consumer must pay for each unit she chooses to purchase. At an amusement park, it would correspond to the price of each ride ticket. In some instances, as with some—but not all—amusement parks, the per-unit charge might even be set to zero.

Formally, consider a two-part tariff consisting of an entry fee, F, and a per-unit charge, p. A consumers expenditure, $T(q)$, if she buys q units is

$$T(q) = \begin{cases} 0\text{, if } q = 0 \\ pq + F \text{ if } q > 0 \end{cases},$$

The function $T(\cdot)$ is the tariff.

How much will a consumer buy under this tariff? If she pays the entry fee, then it is sunk; and her decision on how many to buy is the same as under simple pricing. That is, she chooses the q that equates her marginal benefit to p (*i.e.*, the q such that $mb(q) = p$). This is equivalent to saying that she buys $d(p)$ units, where $d(\cdot)$ is her individual demand schedule. We noted earlier that the value to a consumer of being *able* to buy q units at a price of p per unit was worth $cs(q) = cs\big(d(p)\big)$. Hence, when she decides whether or not to pay the entry fee, she asks herself whether F exceeds $cs\big(d(p)\big)$. If it does, then it isn't worth it to her to pay the fee; she's better off not buying at all. If it doesn't, then she should pay the fee. We can summarize the consumer's decision as

$$\text{units consumer buys} = \begin{cases} 0\text{, if } F > cs\big(d(p)\big) \\ d(p)\text{, if } F \leq cs\big(d(p)\big) \end{cases}. \tag{4.2}$$

[4] In fact, for this reason, two-part tariffs are sometimes called *"Disneyland" pricing*.

The question of whether F is less than or greater than $cs\big(d(p)\big)$ is known as a *participation constraint*. It determines whether or not the consumer is willing to participate (*i.e.*, buy) under the pricing scheme.

All consumers are identical

In this section, we consider the case in which there are N consumers, each of whom has the same demand for the firm's product or service. Using expression (4.2), we see that each consumer buys $d(p)$ units if $F \leq cs\big(d(p)\big)$ and 0 units otherwise. This means the firm's profits are

$$\pi = \begin{cases} 0\,, \text{ if } F > cs\big(d(p)\big) \\ N\big(F + pd(p)\big) - C\big(Nd(p)\big)\,, \text{ if } F \leq cs\big(d(p)\big) \end{cases}\,, \tag{4.3}$$

where the $Nd(p)$ term arises because if it sells $d(p)$ units to each customer, it sells $Nd(p)$ units in total.

We anticipate that the firm wants to operate, so it will set

$$F \leq cs\big(d(p)\big)\,.$$

From expression (4.3), the firm's profit is increasing in F, so it wants to make F as large as possible. But, as just seen, the largest possible F is $cs\big(d(p)\big)$. Therefore, the profit-maximizing entry fee is $F = cs\big(d(p)\big)$. Using this, we can substitute out F in expression (4.3) and write

$$\pi = N\big(cs\big(d(p)\big) + pd(p)\big) - C\big(Nd(p)\big)\,.$$

Observe that $Nd(p)$ is aggregate demand, $D(p)$. Let $Q = D(p)$ and observe that $p = P(Q)$, where $P(\cdot)$ is aggregate *inverse* demand. We can, therefore, rewrite the firm's profit as

$$\pi = Ncs\left(\frac{Q}{N}\right) + P(Q)Q - C(Q)\,.$$

Recall, from our discussion in Section 4.2, that aggregate consumer surplus, CS, is the sum of the individual consumer surpluses. Hence,

$$\pi = CS + P(Q)Q - C(Q)\,.$$

Observe, in this last expression, that profit is the sum of the consumers' surplus *and* the profits that the firm would have made under simple pricing. We can see already, therefore, that this two-part tariff yields greater profits than simple pricing would.

Finally, recall from expression (4.1) that CS is total benefit less expenditure, pQ. We have, therefore, that

$$CS = B(Q) - P(Q)Q\,;$$

hence,

$$\begin{aligned}\pi &= B(Q) - P(Q)Q + P(Q)Q - C(Q) \\ &= B(Q) - C(Q) = W(Q)\,. \end{aligned} \tag{4.4}$$

We've thus shown that, if all the consumers are identical (homogeneous), profits under a two-part tariff achieve the Holy Grail! That is, a two-part tariff allows the firm to capture all the gains to trade. In other words, a two-part tariff with identical consumers achieves perfect or first-degree price discrimination.

From Proposition 23, we know that a firm that captures all the gains to trade maximizes profit by producing the amount that equates inverse demand and marginal cost. We can, therefore, conclude as follows.

Proposition 24. *Suppose consumers have identical demands (are homogeneous). Under the profit-maximizing two-part tariff, the firm*

(i) produces q^{**} units, where $P(q^{**}) = MC(q^{**})$;

(ii) sets the per-unit charge, p, equal to $P(q^{**})$; and

(iii) sets the entry fee, F, to equal $cs\big(d(p)\big)$ (or, equivalently, sets it equal to CS/N).

Example 20 [An amusement park]: Consider an amusement park. Ignoring possible congestion costs, it seems reasonable to approximate the park's marginal cost of a seat on a ride as being 0. Suppose that all 50,000 patrons who come into the park have approximately the same demand for rides. Suppose that market research has revealed that demand to be

$$d(p) = \begin{cases} 50 - 12.5p\,, \text{ if } p \leq 4 \\ 0\,, \text{ if } p > 4 \end{cases}\,.$$

What is the optimal two-part tariff for this amusement park to use?

From Proposition 24, the per-unit charge—that is, the price per ride—should be set equal to marginal cost, which in this case is zero. So $p = 0$. What about the entry fee? It should be set to $cs\big(d(p)\big)$. We can calculate this quantity in two ways. First, we can calculate it as $A_0^{\infty}(d)$; that is, as the area under individual demand from 0 to infinity. Alternatively, we can determine the marginal-benefit schedule and, then, calculate consumer surplus as the area under the marginal-benefit schedule and above the price line (here, the horizontal axis) from 0 to $d(0)$ units. For purposes of illustration, we will do both.

At $p = 4$, $d(p) = 0$. Hence, the area under the demand curve from 0 to infinity is equal to the area under the demand curve from 0 to 4. If you plot the demand curve, you will see that it is a right triangle with height 4 and, because $d(0) = 50$, width 50. Its area is, thus, 100.[5] So the entry fee is \$100.

Alternatively, we know

$$q = 50 - 12.5mb(q)\,;$$

[5] Recall the area of a triangle is base $\times$ height $\div$ 2.

hence,

$$mb(q) = 4 - .08q\,.$$

Plotting this, we see it is a triangle with height 4 and width $d(0) = 50$. So its area is also 100 (of course, it had be to be the same—these are equivalent procedures). So the entry fee is \$100.

Example 21: A firm's cost is $C(q) = q^2/2000$. It faces 1000 identical customers, each of whom has demand

$$d(p) = \begin{cases} 10 - p\,, \text{ if } p \leq 10 \\ 0\,, \text{ if } p > 10 \end{cases}\,.$$

What is the optimal two-part tariff for it to employ? Unlike the amusement park example, here marginal cost is not a constant. Using Proposition 3 on page 39, marginal cost is $MC(q) = q/1000$.

We, next, need to calculate *inverse* aggregate demand. To do so, first we need to determine aggregate demand, $D(\cdot)$, then invert that to get *inverse* aggregate demand, $P(\cdot)$.

$$D(p) = 1000d(p) = \begin{cases} 10,000 - 1000p\,, \text{ if } p \leq 10 \\ 0\,, \text{ if } p > 10 \end{cases}\,.$$

When demand is positive, we have

$$Q = 10,000 - 1000P(Q)\,;$$

hence,

$$P(Q) = 10 - \frac{Q}{1000}\,.$$

Next, equate $P(Q)$ and $MC(Q)$:

$$10 - \frac{Q}{1000} = \frac{Q}{1000}\,.$$

So $Q^{**} = 5000$. $P(Q^{**}) = P(5000) = 5$. So the per-unit charge, p, is \$5.

What about the entry fee? The area under $d(\cdot)$ from 5 to infinity is, because $d(10) = 0$, the same as the area under $d(\cdot)$ from 5 to 10. Plotting, we see that corresponds to the area of the right triangle with vertices (0,5), (5,0), and (0,10). Its height is 5 and its width is 5, so its area is 12.5. Therefore, the entry fee is \$12.50.

Consumers are heterogenous

What if consumers don't all have the same demand curves? The firm can still profit from using a two-part tariff, but designing the optimal two-part tariff becomes much more complicated. Moreover, the elements of its design depend critically on a number of properties of the demand curves and how they vary across individuals (for example, the analysis is sensitive to whether individual

demand curves cross each other or not). Because it is so involved, an investigation of two-part tariff design with heterogeneous customers is beyond the scope of text such as this.[6]

One strategy, which we will explore in greater depth when we deal with third-degree price discrimination, would be to divide customers into homogeneous groups, then employ the optimal two-part tariff on each group. For example, an amusement park might conclude that seniors have very different demands for rides than others, so it has a different admission price for them.

Real-life use of two-part tariffs

Amusement parks have already been offered as a real-life example of firms that use two-part tariffs. Cover charges at bars, night clubs, and similar venues, are examples of entry fees, with the price per drink being the per-unit charge. Club stores (*e.g.*, Costco) that have membership fees are yet another example of two-part tariffs.

Because the per-unit charge could be zero, many two-part tariffs are hard to recognize initially. For example, service plans or tech-support plans are "disguised" two-part tariffs: The entry fee is the price of the plan and the per-unit charge is often zero. Of course, some plans like these have both a positive entry fee (plan price) and a positive per-unit charge (service-call charge).

The pricing of telephony is another example. For land-line phones, a common plan is one in which you pay an access fee per month plus some amount per call. Many calling plans are more complicated than this; they, for instance, charge a monthly fee, provide some minutes worth of calls at a very low fee (often zero), and then charge a different price per call for minutes beyond the provided minutes. These plans are examples of *multi-part tariffs*. Some of our analysis of second-degree price discrimination touches on multi-part tariffs, but a general analysis of multi-part tariffs is beyond the scope of this course..[7]

What limits two-part tariffs?

If two (or multi-) part tariffs are so great, why don't we see more of them? That is, why are amusement park rides and telephone minutes sold using such tariffs, while things such as cereal and soda are not (or at least appear not to be)?

The answer has, in part, to do with arbitrage.[8] Suppose a grocery store attempted to sell soda using a two-part tariff. It would pay someone to break

[6] For the motivated, a good book on nonlinear pricing is Robert B. Wilson's *Nonlinear Pricing*, Oxford: Oxford University Press, 1993. This book will only be accessible if you're very comfortable with calculus.

[7] Again, the motivated student may wish to consider Robert B. Wilson's *Nonlinear Pricing*, Oxford: Oxford University Press, 1993.

[8] Some other reasons for not using two-part tariffs are the costs of monitoring entry and heterogeneity among customers.

the scheme by paying the entry fee, then buying more soda than he wants in order to resell it to others. As long as his resale price exceeds his average cost, $p + F/q$, where q is the amount he is reselling, he is making a profit. Because they are avoiding the entry fee, his customers can find it cheaper to buy soda from him than from the grocery store. So we see that a grocery store attempting to sell soda using a two-part tariff would find its scheme collapsing due to arbitrage. In the limit, it would sell lots of soda to just one enterprising customer.

Unlike soda, which is easily resold, items like amusement park rides, telephone calls from your house or cell phone, service calls to your place of business, etc., are difficult, if not impossible, to resell. There is little danger of arbitrage. Hence, we expect to see two-part tariffs with items that are difficult to resell and not to see such tariffs with items that are readily resold.

Two-part tariffs: *Are more readily used when arbitrage is difficult.*

However, firms can attempt to defeat arbitrage. For instance, the grocery store could use a two-part tariff for some goods in the following way. Suppose that rather than being sold in packages, sugar were kept in a large bin and customers could take as much or as little as they like (*e.g.*, like fruit at most groceries or as is actually done at some stores like Berkeley Natural Grocery). Suppose that, under the optimal two-part tariff, each consumer would buy two pounds of sugar at a price p per pound, which would yield him or her surplus of cs, which would be captured by the store using an entry fee of $F = cs$. Of course, arbitrage makes this infeasible. So suppose, in response, that rather then let consumers buy as much or as little sugar as they want, the store packaged sugar in two-pound bags, for which it charged $2p + cs$ per bag. Each customer would face the decision of whether to have 0 pounds or 2 pounds. Each customer's total benefit from two pounds is $2p + cs$, so each would just be willing to pay $2p + cs$ for the package of sugar. Because the entry fee is paid on every two-pound bag, the store has devised an arbitrage-proof (albeit disguised) two-part tariff. In other words, *packaging*—taking away customers ability to buy as much or as little as they wish—can represent an arbitrage-proof way of employing a two-part tariff.

Packaging: *An arbitrage-proof way of implementing a two-part tariff.*

Metering

In addition to packaging and setting the per-unit price to zero, a third way in which two-part tariffs are disguised is via *metering*. Metering—also known as *tying*—has to do with situations in which the entry fee is the purchase price for a durable good (*e.g.*, an instant camera, a razor handle, a punch-card sorting machine, etc.) and the per-unit charge is the purchase price for a complementary good (*e.g.*, instant film, cartridge razor blades, punch cards, etc.). For example, one can view Gillette as using a two-part tariff to sell Mach 3 razor cartridges. The price of the handle (or a package with handle and some cartridges) represents the entry fee. The price of cartridge replacement packs is the per-unit charge.

Metering only works if the firm can keep others from making the complementary good. In the cases of Gillette or Polaroid, intellectual property rights

prevent others from making cartridges that fit Gillette handles or instant film that works in Polaroid cameras. For commodity products, like punch cards and paper, firms used to attempt to prevent others from providing the complementary good through contracts. That is, IBM would, through various contract terms, require or strongly encourage users of its punch-card sorting machines to buy IBM punch cards. Xerox would likewise require or strongly encourage users of its photocopiers to use Xerox paper (as well as Xerox toner, etc.). In the US, such requirements or pressure via contract have been deemed to violate the antitrust laws (these are forms of illegal tying).

Although we have introduced metering as a way to engage in a two-part tariff, metering can also be useful for other forms of price discrimination (*e.g.*, second-degree price discrimination through quantity discounts). Metering can also be useful as a way of providing customer credit—the camera price or the printer price is a down payment, the customer is "loaned" the rest of the camera's or printer's full price, and the film price or toner cartridge price contains a repayment (principal and interest) portion on the original loan taken out by buying the camera or printer. For instance, if the auto companies could compel you to buy gasoline from them only, then they could provide car financing through a low car price and a high gasoline price.

4.5 Third-Degree Price Discrimination

Third-degree price discrimination: *Charging different prices on the basis of observed group membership.*

In this section, we consider third-degree price discrimination.[9] *Third-degree price discrimination* means charging different tariffs to different consumers on the basis of identifiable characteristics. An example of third-degree price discrimination is a movie theater that charges a different price for children, a different price for seniors, and a different price for other adults.[10] The characteristics on which the theater distinguishes among patrons are readily identifiable, either by cashier observation or presentation of ID.

Senior-citizen discounts, child discounts, student discounts, and family discounts are all familiar forms of third-degree price discrimination. Ladies' nights at bars and similar venues are another form of third-degree price discrimination (albeit one also motivated by network externalities).[11] Sometimes third-degree price discrimination is on the basis of membership in certain organizations (*e.g.*, AAA discounts or discounts to members of the local public radio station).

Geography-based third-degree price discrimination is also quite prevalent. For example, a firm may quote different prices to buyers in one area than it

[9] Yes, I can count. Regardless of the order implied by their names, it makes more sense pedagogically to consider third-degree price discrimination before second-degree price discrimination.

[10] If the over-60 set are senior citizens, does that make the rest of us "junior citizens"?

[11] Network externalities refer to situations in which one person's demand (*e.g.*, a guy looking to meet women) is a function of the demand of others (*e.g.*, how many women are at the bar).

does to buyers in another area. This happens, for instance, in air travel, where a round-trip ticket going city A to city B back to city A has a different price than one going B to A back to B (this is especially true if A and B are in different countries). A third-degree pricing example that occasionally engenders newspaper articles and editorials is the different prices pharmaceutical companies charge for their drugs in different countries.

The simple case

The most basic situation of third-degree price discrimination is when a firm engages in simple pricing to two distinct groups when the firm faces no capacity constraints.

Let $P_i(\cdot)$ be the *inverse* aggregate demand of two populations indexed by i, $i = 1$ or 2. For instance, $P_1(\cdot)$ could be the inverse aggregate demand of students for a concert and $P_2(\cdot)$ could be the inverse aggregate demand of non-students. Let $C(\cdot)$ be the firm's cost function. Let q_i be the amount sold to population i. The firm's profit is its revenue from each population less its costs:

$$\pi(q_1, q_2) = q_1 P_1(q_1) + q_2 P_1(q_2) - C(q_1 + q_2)\,.$$

Using Proposition 15 on page 72, we can write

$$MR_i(q_i) = P_i(q_i) + q_i P'(q_i)\,.$$

We denote marginal cost in the usual way, $MC(q_1 + q_2)$.

How much should the firm sell to each population? From our earlier analysis, a good guess would be the amounts that equate the marginal revenues to marginal cost. This guess is, in fact, correct. Let's see why. Clearly, it cannot be optimal to produce so that

$$MR_i(q_i) < MC(q_1 + q_2)$$

because the firm could increase its profit by

$$MC(q_1 + q_2) - MR_i(q_i)$$

if it sold one less unit to population i. What if

$$MR_i(q_i) > MC(q_1 + q_2)\,? \tag{4.5}$$

Then the firm could increase its profit by

$$MR_i(q_i) - MC(q_1 + q_2)$$

if it sold one more unit to population i. So it cannot be optimal to produce so that expression (4.5) holds. By process of elimination, this leaves

$$MR_1(q_1^*) = MR_2(q_2^*) = MC(q_1^* + q_2^*) \tag{4.6}$$

at the profit-maximizing quantities, q_1^* and q_2^*. This analysis generalizes to N populations:

Proposition 25. *A firm* unconstrained *in the amount it can produce and that can engage in simple pricing to N distinct populations maximizes its profit by selling q_i^* to population i, where*

$$MR_1(q_1^*) = \cdots = MR_N(q_N^*) = MC\left(\sum_{i=1}^{N} q_i^*\right).$$

Observe this expression could also be written as

$$MR_1(q_1^*) = MC\left(\sum_{i=1}^{N} q_i^*\right)$$
$$\vdots$$
$$MR_N(q_N^*) = MC\left(\sum_{i=1}^{N} q_i^*\right).$$

As with simple pricing, the price charged population i is $P_i(q_i^*)$.

Because a possible result with this type of third-degree price discrimination is that $P_1(q_1^*) = \cdots = P_N(q_N^*)$, optimal nondiscriminatory simple pricing (*i.e.*,treating all populations the same) is a possible outcome of third-degree price discrimination. From this, we see that third-degree price discrimination can never do worse than nondiscriminatory simple pricing. Moreover, if we find $P_i(q_i^*) \neq P_j(q_j^*)$ for two populations i and j, then it must be that third-degree price discrimination is generating strictly greater profits.

Example 22: Consider a firm (*e.g.*, a pharmaceutical company) that has the cost function $C(Q) = Q^2/1000$. Observe, using Proposition 3, that its marginal cost is $Q/500$. Suppose it faces two populations (*e.g.*, Americans and Canadians), $i = 1, 2$. Let their aggregate demands be

$$D_1(p) = \begin{cases} 100,000 - 1000p\,, \text{ if } p \leq 100 \\ 0\,, \text{ if } p > 100 \end{cases} \text{ and}$$

$$D_2(p) = \begin{cases} 150,000 - 2000p\,, \text{ if } p \leq 75 \\ 0\,, \text{ if } p > 75 \end{cases}.$$

Inverting these demands over the range of positive demand:

$$q_1 = 100,000 - 1000P_1(q_1) \text{ and}$$
$$q_2 = 150,000 - 2000P_2(q_2)\,.$$

Therefore,

$$P_1(q_1) = 100 - \frac{q_1}{1000} \text{ and}$$
$$P_2(q_2) = 75 - \frac{q_2}{2000}\,.$$

Using Proposition 15, we have

$$MR_1(q_1) = 100 - \frac{q_1}{1000} + q_1\left(\frac{-1}{1000}\right) = 100 - \frac{q_1}{500} \text{ and}$$

$$MR_2(q_2) = 75 - \frac{q_2}{2000} + q_2\left(\frac{-1}{2000}\right) = 75 - \frac{q_2}{1000}\,.$$

We can now employ Proposition 25 to solve for q_1^* and q_2^*:

$$100 - \frac{q_1}{500} = \frac{q_1 + q_2}{500} \text{ and}$$

$$75 - \frac{q_2}{1000} = \frac{q_1 + q_2}{500}\,.$$

Using the methods for two equations in two unknowns set forth in Appendix A2, we obtain: $q_1^* = 18,750$ and $q_2^* = 12,500$. This yields prices:

$$P_1(18,750) = \$81.25 \text{ and } P_2(12,500) = \$68.75\,.$$

Observe the firm's profit is

$$18,750 \times \$81.25 + 12,500 \times \$68.75 - \frac{31,250^2}{1000} \text{ dollars} = \$1,406,250\,.$$

Because the prices to the two population are not the same, we know the firm's profit is greater than if it were forced to charge the same price to both populations.[12]

The capacity-constrained firm

Consider a concert hall that is considering different prices to students and non-students for a particular show. Assume this company's cost function is

$$C(q) = \begin{cases} 0\,, \text{ if } q = 0 \\ q + 20,000\,, \text{ if } q > 0 \end{cases}\,,$$

where q is the number of seats sold. Observe the marginal cost per seat is small, just \$1, while the overhead cost of the show is relatively high.

Unlike the situation in Example 22, the company is constrained: It cannot sell more tickets than it has seats. This will pose a problem if the optimal third-degree pricing scheme indicates q_S^* tickets should be sold to students and q_N^* tickets should be sold to non-students, but

$$q_S^* + q_N^* > K\,,$$

where K is the number of seats in the concert hall (the capacity).

[12]It can be shown—indeed, as an exercise you may wish to do the calculations—that the firm's profit if it had to charge the same price in both markets would be \$1,302,080.

Let's suppose that expression () is true. Then the firm's problem in designing the optimal third-degree price discrimination scheme is to choose the q_S and q_N that maximize profit,

$$q_S P_S(q_S) + q_N P_N(q_N) - \underbrace{(q_S + q_N + 20,000)}_{C(q_S+q_N)}$$

subject to the constraint $q_S + q_N \leq K$. By assumption, this constraint is binding; that is, we know $q_S + q_N = K$ (if it weren't binding, then we could design the scheme ignoring it, but then we would get a solution, q_S^* and q_N^*, that violated the constraint). Substituting into our expression for profit, we can write the profit as

$$q_S P_S(q_S) + q_N P_N(q_N) - \underbrace{(K + 20,000)}_{C(q_S+q_N)} .$$

Note that expenses, $K + 20,000$, are a constant; that is, our problem is independent of them. This makes sense from an opportunity-cost point of view: We've already decided we're selling all K seats—that decision is sunk—what we're deciding, now, is how to allocate those seats.

Thinking further along opportunity cost lines, we can solve for the optimal constrained allocation. What is the opportunity cost of selling the marginal seat to a student? It is the forgone value of selling that seat to a non-student. What's that value? It's the marginal revenue that would be realized by selling that seat to a non-student. In other words, we've just seen that the marginal cost of selling a seat to a student is the marginal revenue that seat would yield if sold to a non-student. Formally, we have

$$MC_S(q_S) = MR_N(q_N) = P(q_N) + q_N P'(q_N) ,$$

where the second inequality follows from the usual formula for marginal revenue under simple pricing and $MC_S(\cdot)$ is the marginal cost schedule for seats sold to students.

Profit-maximization requires equating marginal revenue to marginal cost. So we know that

$$MR_S(q_S) = MR_S(q_S) .$$

at the optimal number of seats to sell to students. Substituting for $MC_S(q_S)$, we have

$$MR_S(q_S) = MR_N(q_N) = P(q_N) + q_N P'(q_N) . \tag{4.7}$$

That is, given constrained capacity, the optimal quantities for the two populations will equate their marginal revenues. Expression (4.7) is one equation in two unknowns, q_S and q_N. Fortunately, we have a second equation, namely the constraint:

$$q_S + q_N = K . \tag{4.8}$$

We can conclude:

Proposition 26. *If a firm faces a binding constraint of K units and it faces two populations, N and S, then the optimal quantities, $\hat{q}_N$ and $\hat{q}_S$, to be sold to the two populations equate the two populations' marginal revenues and they sum to K; that is, the quantities solve expressions* (4.7) *and* (4.8)*.*

The prices to be charged to the two populations are $P_S(\hat{q}_S)$ and $P_N(\hat{q}_N)$ for students and non-students, respectively.

Example 23: Let's put some more numbers to our analysis of this concert hall's pricing problem. Specifically, suppose

$$P_S(q_S) = 41 - q_S/10 \text{ and}$$
$$P_N(q_N) = 101 - q_N/10\,.$$

Suppose that capacity, K, is 500 seats.

If the number of seats in the hall *were not* constrained, the firm would determine its optimal allocation from $MR_S(q_S) = MC(q_S + q_N)$ and $MR_S(q_S) = MC(q_S + q_N)$ (see Example 22). Here, those expressions are

$$41 - q_S/5 = 1 \text{ and}$$
$$101 - q_N/5 = 1,,$$

respectively. The solutions are $q_S^* = 200$ and $q_N^* = 500$. Because $200 + 500 > 500$, this "solution" doesn't work given the concert hall's seating capacity.

Now employ Proposition 26: Equating the marginal revenues, we have

$$41 - q_S/5 = 101 - q_N/5\,.$$

We also have the constraint

$$q_S + q_N = 500\,.$$

Solving these two equations yields $\hat{q}_S = 100$ and $\hat{q}_N = 400$. The ticket prices are $P_S(100) = \$31$ and $P_N(400) = \$61$ for students and non-students, respectively. Profit is

$$100 \times \$31 + 400 \times \$61 - \$20,500 = \$7000\,.$$

Note, a naïve approach might have been the following: Seeing that $q_N^* = 500 = K$, one might have concluded that one should sell to non-students only. Selling 500 seats to non-students would mean a price of $P_N(500) = \$51$. Observe this price is *greater* than the price any student would be willing to pay, so charging \$51 per seat would, indeed, result in only non-students attending. Yet, the profit from this naïve approach is less than the optimal amount: $500 \times \$51 - \$20,500 = \$5000$. This naïve approach would cost the concert hall \$2000 in profit.

Third-degree price discrimination and other forms of discrimination

We have seen that segmenting the population on the basis of identifiable groupings and, then, treating each group separately for the purpose of simple pricing yields a firm greater profit than setting a single price for everyone.

Sometimes a firm can do even better if it combines segmentation on the basis of groupings and other forms of discrimination. For example, a local bar with a live band once advertised: "No cover charge for ladies." That is, the bar was using a two-part tariff for men—an entry fee and a per-unit (*i.e.*, per drink) charge—but a single-part tariff for women.

When, for instance, a firm sells a package for one price in one country and a different price in another, then it is combining third-degree price discrimination (pricing on the basis of group) and a two-part tariff (remember, a package is one way to implement a two-part tariff).

Going through all the ways in which different forms of price discrimination can be combined is beyond the scope of this text—but as you think about pricing in your careers, keep in mind that these different forms can be combined.

Arbitrage

Because, under third-degree price discrimination, different populations face different tariffs, there is the opportunity for arbitrage. A student, for instance, could resell his ticket to a non-student. Hence, the ability to employ third-degree price discrimination effectively is dependent on an ability to prevent arbitrage (this is true, actually, of all discriminatory pricing).

In many instances, the ability to distinguish members of the different populations forecloses arbitrage. A non-senior with a senior ticket can be denied admission, for example. In other situations, especially when geography is used to identify different groups, it is much harder (*e.g.*, witness Americans going to Canada to buy prescription medications). In some cases, transportation costs deter geographic arbitrage. This is why, for instance, a firm can charge very different prices in West Virginia and California, but needs to charge fairly similar prices in San Francisco and San Jose.

Second-Degree Price Discrimination | 4.6

As we've seen, third-degree price discrimination can yield greater profits than simple pricing. But it requires being able to identify the members of different populations (*e.g.*, students from non-students). What if you can't readily identify who's in what group (market segment)? For instance, we know that business travelers are willing to pay more for plane tickets than vacationers. But mere inspection wont tell you which would-be flier is a business woman and

which is a woman going on vacation.[13] Fortunately, in some circumstances, it is possible to induce customers to reveal the group to which they belong. Price discrimination via induced revelation of preferences is known as *second-degree price discrimination*.

Second-degree price discrimination: *Price discrimination via* induced *revelation of preferences.*

A well-known solution to the problem of not being able to distinguish business from leisure travelers by inspection is for the airlines to offer different kinds of tickets. For instance, because business travelers don't wish to stay over the weekend or often can't book much in advance, the airlines charge more for round-trip tickets that don't involve a Saturday-night stayover or that are purchased within a few days of the flight (*i.e.*, in the latter situation, there is a discount for advance purchase). Observe an airline still can't observe which type of traveler is which, but by offering different kinds of service it hopes to induce revelation of which type is which.

Restricted tickets are one example of second-degree price discrimination, specifically of second-degree price discrimination via *quality distortions*. Because a restricted ticket is less useful than an unrestricted ticket, a restricted ticket can be viewed as being lower quality than an unrestricted ticket. Other examples include:

- Different classes of service (*e.g.*, first and second-class carriages on trains). The classic example here is the French railroads in the 19th century, which removed the roofs from second-class carriages to create third-class carriages.

- Hobbling a product. This is popular in high-tech, where, for instance, Intel once produced two versions of a chip by "brain-damaging" the state-of-the-art chip. Another example is software, where "regular" and "pro" versions (or "home" and "business" versions) of the same product are often sold.

- Restrictions. Saturday-night stayovers and advance-ticketing requirements are a classic example. Another example is limited versus full memberships at health clubs.

The other common form of second-degree price discrimination is via *quantity discounts*. This is why, for instance, the liter bottle of soda is typically less than twice as expensive as the half-liter bottle. Quantity discounts can often be operationalized through multi-part tariffs, so many multi-part tariffs are examples of price discrimination via quantity discounts (*e.g.*, choices in calling plans between say a low monthly fee, few "free" minutes, and a high per-minute charge thereafter versus a high monthly fee, more "free" minutes, and a lower per-minute charge thereafter).

[13] Although the two might dress differently, if airlines started charging different prices on the basis of what their passengers wore, then their passengers would all show up wearing whatever clothes get them the cheapest fare.

Quality distortions

Suppose there are two types of users of a given kind of software. Casual users value the core capabilities of the software at $50 and the ancillary capabilities at $25 (think of the core capabilities as, *e.g.*, basic wordprocessing and the ancillary capabilities as, *e.g.*, the ability to do mail-merges). Business users value the core capabilities at $80 and the ancillary capabilities at $60. Assume the marginal cost of producing a unit of the software is zero and that the seller of the software faces irrelevantly small costs to produce a version without the ancillary capabilities.

Let r denote the ratio of business to casual users. What we wish to explore is, as a function of r, what version or versions does the software manufacturer produce and what price or prices does it charge.

A naïve answer might be that, regardless, of r, the manufacturer offers a home version with the core capabilities at a price $50 and a pro/business version with the ancillary features at $\$80 + \$60 = \$140$. But, were it to do so, then no one would buy the business version. This is obvious for the casual (home) user: The value of the business version to him is $75, which is less than $140. But it is also true of the business user. To see this, recall that any consumer seeks to maximize his or her consumer surplus—that is, the difference between the benefit he or she derives and the amount he or she pays. The consumer surplus enjoyed by a business user who purchases the business version is $\$0 (= \$140 - \$140)$. The consumer surplus the business user who purchases the home version is $\$30 (= \$80 - \$50)$. The business user would, thus, opt to buy the home version.

Our consideration of the naïve answer reveals what is critical to the answer: If the manufacturer is to induce a business user to reveal she is a business user by getting her to buy the business version, then the manufacturer must leave her at least as much surplus as she would get were she to buy the home version. That is, if both products are sold, with the home version being priced at $50, then the difference between the price of the business version and a business user's value for the business version must be at least $30. This $30 has a name: It is the business user's *information rent*. The reason for the name is that the information rent is what the manufacturer must "pay" the business user (more generally, a high-value consumer) for the information that she is a business user. In other words, the information rent is the cost of inducing revelation.

Information rent: *The surplus that a high-valuation user captures under 2nd-degree price discrimination.*

Hence, if the software manufacturer offers both version, with the home version aimed at the casual user and the business version aimed at the business user, then the profit-maximizing prices for the two are $50 and $110.

The other option available to the manufacturer is simply to offer one version. If it does, which version? Well, given there is no additional cost to producing the business version (core plus ancillary capabilities) over the home version (core capabilities only) and *both* types of potential buyer value the business version more, it should sell just the business version. If it chooses to sell the business version only, what price should it charge? Well either $75 and sell to everyone or $140 and sell to just business users.

To summarize to this point. We've identified three options for the software manufacturer as potentially being profit maximizing:

1. Sell a business version for \$110 and a home version for \$50.
2. Sell a business version only and do so at a price of \$75.
3. Sell a business version only and do so at a price of \$140.

Which option is best? The answer depends on r, the ratio of business to casual users. Observe that the *total* number of potential buyers is proportional to $1+r$, the number of casual users proportional to 1, and the number of business users proportional to r. So the profits under the three option are proportional to

1. $\$50 + \$110r$.
2. $\$75(1+r)$.
3. $\$140r$.

The first option beats the second if $\$35r \geq \$25$; that is, if $r \geq 5/7$. The first option beats the third if $50 \geq \$30r$; that is, if $r \leq 5/3$. The second beats the third if $\$75 \geq 65r$; that is, if $r \leq 15/13$. Putting all this together, we can conclude:

Conclusion. *If $r < 5/7$ (i.e., if there are fewer than 5 business users for each 7 casual users), then the manufacturer should choose the second option, the business version at a price of \$75. If $5/7 \leq r \leq 5/3$, then the manufacturer should choose the first option, sell both versions at prices \$110 for the business version and \$50 for the home version. Finally, if $5/3 < r$, then the manufacturer should choose the third option, the business version at a price of \$140.*

Although the discussion here has been tied to a specific example, the basic principles should be clear:

- Discrimination via quality distortion means introducing a product that is *worse* than the product that would offered were you not discriminating.
- If you offer multiple products, then high-valuation consumers—the ones for whom the better products are intended—must get information rents reflecting that the they need inducement not to buy the inferior products that are priced so low-valuation consumers will buy them.

Quantity discounts

The fact that different size containers of the same good often cost different amounts on a per-unit basis is well known. Typically, the larger the package, the less it costs per-unit; that is, for example, the liter bottle of Pepsi typically costs less than twice the price of the half-liter bottle of Pepsi. In this section we consider why such quantity discounts can be an effective form of second-degree price discrimination.

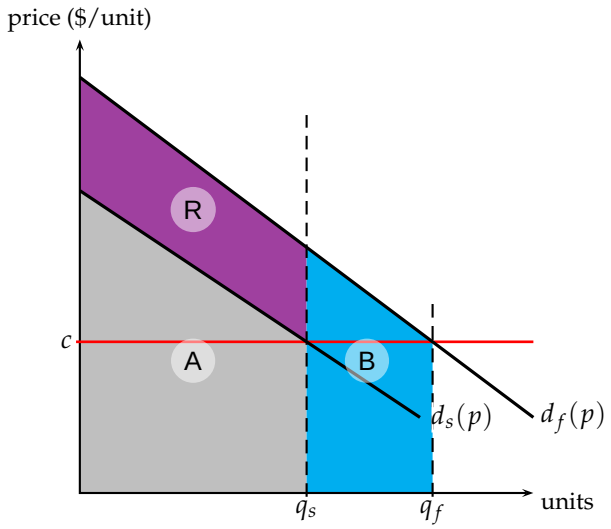

Figure 4.5: The individual demands of the two types of consumers (family and single), $d_f(\cdot)$ and $d_s(\cdot)$, respectively, are shown.

A fully general treatment can get quite complicated, so here we will restrict attention to a situation in which the firm has constant marginal cost, c.

Suppose the population of potential buyers is divided into families (indexed by f) and single people (indexed by s). Let $d_f(\cdot)$ denote the demand of an *individual* family and let $d_s(\cdot)$ denote the demand of an *individual* single. Figure 4.5 shows the two demands. Note that, at any price, a family's demand exceeds a single's demand.

The ideal, from the firm's perspective, would be the following. Suppose it could freely identify singles from families. It would then offer two different two-part tariffs (packages) to the two populations. It would make the per-unit charge c and the entry fee the respective consumer surpluses. Equivalently—and more practically—consider packaging. The package for singles would have q_s units and sell for a single's total benefit, $b_s(q_s)$. This is the area labeled A in Figure 4.5. Similarly, the family package would have q_f units and sell for a family's total benefit of $b_f(q_f)$. This is the sum of the three labeled areas in Figure 4.5.

The ideal is not, however, achievable. The firm can*not* freely distinguish singles from families. It must *induce* revelation; that is, it must devise a *second*-degree scheme. Observe that the ideal scheme won't work as a second-degree scheme. Although a single would still purchase a package of q_s units at $b_s(q_s)$, a family would not purchase a package of q_f units at $b_f(q_f)$. Why? Well, were the family to purchase the latter package it would, by design, earn no consumer surplus. Suppose, instead, it purchased the package intended for singles. Its

total benefit from doing so is the sum of areas A and R in Figure 4.5. It pays $b_s(q_s)$, which is just area A, so it would enjoy a surplus equal to area R. In other words, the family would deviate from the intended package, with q_f units, which yields it no surplus, to the unintended package, with q_s units, which yields it a positive surplus equal to area R.

Observe that the firm could induce revelation—that is, get the family to buy the intended package—if it cut the price of the q_f-unit package. Specifically, if it reduced the price to the sum of areas A and B, then a family would enjoy a surplus equal to area R whether it purchased the q_s-unit package (at price = area A) or it purchased the intended q_f-unit package (at price = area A + area B). Area R is a family's information rent.

Although that scheme induces revelation, it is not necessarily the *profit-maximizing* scheme. To see why, consider Figure 4.6. Suppose that the firm reduced the size of the package intended for singles. Specifically, suppose it reduced it to $\hat{q}_s$ units, where $\hat{q}_s = q_s - h$. Given that it has shrunk the package, it would need to reduce the price it charges for it. The benefit that a single would derive from $\hat{q}_s$ units is the area beneath its inverse demand curve between 0 and $\hat{q}_s$ units; that is, the area labeled A′. Note that the firm is forgoing revenues equal to area J by doing this. But the surplus that a family could get by purchasing a $\hat{q}_s$-unit package is also smaller; it is now the area labeled R′. This means that the firm could raise the price of the q_f-unit package by the area labeled H. Regardless of which package it purchases, a family can only obtain surplus equal to area R′. In other words, by reducing the quantity sold to the "low type" (a single), the firm reduces the information rent captured by the "high type" (a family).

Is it worthwhile for the firm to trade area J for area H? Observe that the *profit* represented by area J is rather modest: While selling the additional h units to a single adds area J in revenue it also adds ch in cost. As drawn, the profit from the additional h units is the small triangle at the top of area J. In contrast, area H represents pure profit—regardless of how many units it intends to sell to singles, the firm is selling q_f units to each family (*i.e.*, cq_f is a sunk expenditure with respect to the decision of how many units to sell each single). So, as drawn, this looks like a very worthwhile trade for the firm to make.

A caution, though: The figure only compares a single family against a single single. What if there were lots of singles relative to families? Observe that the total profit loss from reducing the package intended for singles by h is

$$(\text{area J} - ch) \times N_s\,,$$

where N_s is the number of singles in the population. The profit gain from reducing that package is

$$\text{area H} \times N_f\,,$$

where N_f is the number of families. If N_s is much larger than N_f, then this reduction in package size is not worthwhile. On the other hand if the two

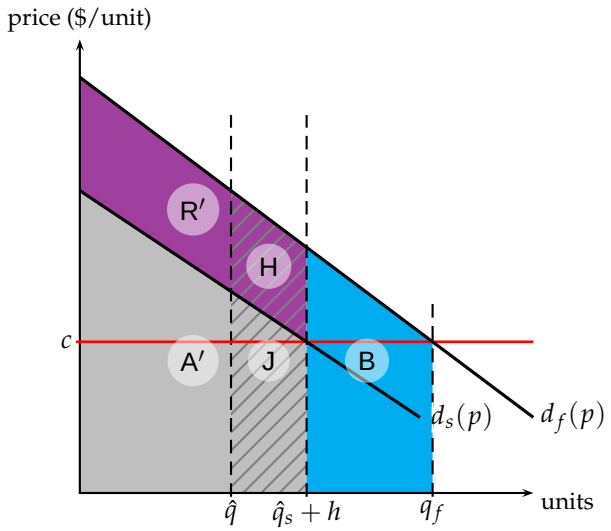

Figure 4.6: By reducing the quantity in the package intended for singles, the firm loses revenue equal to area J, but gains revenue equal to area H.

populations are roughly equal in size or N_f is larger, then reducing the package for singles by more than h could be optimal.

How do we determine the amount by which to reduce the package intended for singles (*i.e.*, the smaller package)? That is, how do we figure out what h should be? As usual, the answer is that we fall back on our $MR = MC$ rule. Consider a small expansion of the smaller package from $\hat{q}_s$. Because we are using an implicit two-part tariff (packaging) on the singles, the change in revenue—that is, marginal revenue—is the change in a single's benefit (*i.e.*, $mb_s(\hat{q}_s)$) times the number of singles. That is,

$$MR(\hat{q}_s) = N_s mb_s(\hat{q}_s)\,.$$

Recall that the marginal benefit schedule is inverse demand. So if we let $\rho_s(\cdot)$ denote the inverse *individual* demand of a single, then we can write

$$MR(\hat{q}_s) = N_s \rho_s(\hat{q}_s)\,. \tag{4.9}$$

What about MC? Well, if we increase the amount in the smaller package we incur costs from two sources. First, each additional unit raises production costs by c. Second, we increase each family's information rent (*i.e.*, area H shrinks). Observe that area H is the area between the two demand curves (thus, between the two *inverse* demand curves) between $\hat{q}_s$ and $\hat{q}_s + h$. This means that the marginal reduction in area H is

$$\rho_f(\hat{q}_s) - \rho_s(\hat{q}_s)\,,$$

where $\rho_f(\cdot)$ is the inverse demand of a family. Scaling by the appropriate population sizes and adding them together, we have

$$MC(\hat{q}_s) = N_s c + N_f\big(\rho_f(\hat{q}_s) - \rho_s(\hat{q}_s)\big)\,. \tag{4.10}$$

Some observations:

1. Observe that if we evaluate expressions (4.9) and (4.10) at the q_s shown in Figure 4.5, we have

$$\begin{aligned} MR(q_s) &= N_s\rho_s(q_s) \text{ and} \\ MC(q_s) &= N_s c + N_f\big(\rho_f(q_s) - \rho_s(q_s)\big)\,. \end{aligned}$$

   Subtract the second equation from the first:

$$\begin{aligned} MR(q_s) - MC(q_s) &= N_s(\rho_s(q_s) - c) - N_f\big(\rho_f(q_s) - \rho_s(q_s)\big) \\ &= -N_f\big(\rho_f(q_s) - \rho_s(q_s)\big) \\ &< 0\,, \end{aligned}$$

   where the second equality follows because, as seen in Figure 4.5, $\rho_s(q_s) = c$ (*i.e.*, q_s is the quantity that equates inverse demand and marginal cost). Hence, provided $N_f > 0$, we see that the profit-maximizing second-degree pricing scheme sells the low type (*e.g.*, singles) less than the welfare-maximizing quantity (*i.e.*, there is a deadweight loss of area $\mathsf{J} - ch$). In other words, as we saw previously, there is *distortion at the bottom*.

2. How do we know we want the family package to have q_f units? Well, clearly we wouldn't want it to have more—the marginal benefit we could capture would be less than our marginal cost. If we reduced the package size, we would be creating deadweight loss. Furthermore, because we don't have to worry about singles' buying packages intended for families (that incentive compatibility constraint is slack) we can't gain by creating such a deadweight loss (unlike with the smaller package, where the deadweight loss is offset by the reduction in the information rent enjoyed by families). We can summarize this as there being no distortion at the top.

3. Do we know that the profit-maximizing $\hat{q}_s$ is positive? That is, do we know that a solution to $MR = MC$ exists in this situation? The answer is no. It is possible, especially if there are a lot of families relative to singles, that it might be profit-maximizing to set $\hat{q}_s = 0$; that is, sell only one package, the q_f-unit package, which only families buy. This will be the case if $MR(0) \leq MC(0)$. In other words, if

$$N_s(\rho_s(0) - c) - N_f\big(\rho_f(0) - \rho_s(0)\big) \leq 0 \tag{4.11}$$

4. On the other hand, it will often be the case that the profit-maximizing $\hat{q}_s$ is positive, in which case it will be determined by equating expressions (4.9) and (4.10).

Example 24: Consider a cell-phone-service provider. It faces two types of customers, those who seldom have someone to talk to (indexed by s) and those who frequently have someone to talked to (indexed by f). Within each type, customers are homogeneous. The marginal cost of providing connection to a cell phone is 5 cents a minute (for convenience, all currency units are cents). A member of the s-population has demand:

$$x_s(p) = \begin{cases} 450 - 10p\,, \text{ if } p \leq 45 \\ 0\,, \text{ if } p > 45 \end{cases}.$$

A member of the f-population has demand:

$$x_f(p) = \begin{cases} 650 - 10p\,, \text{ if } p \leq 65 \\ 0\,, \text{ if } p > 65 \end{cases}.$$

There are 1,000,000 f-type consumers. There are N_s s-type consumers.

What is the profit-maximizing second-degree pricing scheme to use? How many minutes are in each package? What are the prices?

It is clear that the f types are the high types (like families in our previous analysis). There is no distortion at the top, so we know we sell an f type the number of minutes that equates demand and marginal cost; that is,

$$q_f^* = x_f(c) = 600\,.$$

We need to find q_s^*. To do this, we need to employ expressions (4.9) and (4.10). They, in turn, require us to know $\rho_s(\cdot)$ and $\rho_f(\cdot)$. Considering the regions of positive demand, we have:

$$\begin{aligned} q_s &= 450 - 10\rho_s(q_s) \text{ and} \\ q_f &= 650 - 10\rho_f(q_f)\,; \end{aligned}$$

hence,

$$\begin{aligned} \rho_s(q_s) &= 45 - \frac{q_s}{10} \text{ and} \\ \rho_f(q_f) &= 65 - \frac{q_f}{10}\,. \end{aligned}$$

Using expression (4.9), marginal revenue from q_s is, therefore,

$$MR(q_s) = N_s\rho_s(q_s) = N_s \times \left(45 - \frac{q_s}{10}\right).$$

Marginal cost of q_s (including forgone surplus extraction from the f type) is

$$\begin{aligned} MC(q_s) &= N_s c + N_f\big(\rho_f(q_s) - \rho_s(q_s)\big) \\ &= 5N_s + 1,000,000\left(65 - \frac{q_f}{10} - 45 + \frac{q_f}{10}\right) \\ &= 5N_s + 20,000,000\,. \end{aligned}$$

Do we want to shut out the s-types altogether? Employing expression (4.11), the answer is yes if

$$40N_s - 20,000,000 < 0\,;$$

that is, if $N_s < 500,000$. When the s-types are shut out, the price for 600 minutes (*i.e.*, q_f^*) is $b_f(600)$, which is

$$\begin{aligned} b_f(600) &= \text{Area under } \rho_f(\cdot) \text{ from 0 to 600} \\ &= \int_0^{600} \rho_f(z)dz \\ &= \int_0^{600} \left(65 - \frac{z}{10}\right) dz \\ &= 21,000 \end{aligned}$$

cents or \$210.

Suppose that $N_s \geq 500,000$. Then, equating MR and MC, we have

$$N_s \times \left(45 - \frac{q_s}{10}\right) = 5N_s + 20,000,000\,;$$

hence,

$$q_s^* = 400 - \frac{200,000,000}{N_s}\,.$$

The low type retains no surplus, so the price for q_s^* minutes is $b_s(q_s^*)$, which equals the area under $\rho_s(\cdot)$ from 0 to q_s^*. This can be shown (see derivation of $b_f(600)$ above) to be

$$\begin{aligned} b_s(q_s^*) &= \text{Area under } \rho_s(\cdot) \text{ from 0 to } q_s^* \\ &= \int_0^{q_s^*} \left(45 - \frac{q}{10}\right) dq \\ &= 45q_s^* - \frac{q_s^{*2}}{20}\,. \end{aligned}$$

The price charged the f types for their 600 minutes is $b_f(600)$ less their information rent, which is the equivalent of area R' in Figure 4.6.

$$\text{Area } \mathsf{R}' = \int_0^{q_s^*} \left(65 - \frac{q}{45} - 45 + \frac{q}{45}\right) dq = 20q_s^*\,.$$

So the price charged for 600 minutes is $21,000 - 20q_s^*$ cents (\$210 $- q_s^*/5$).

To conclude: If $N_s < 500,000$, then the firm sells only a package with 600 minutes for \$210. In this case, only f types buy. If $N_s \geq 500,000$, then the firm sells a package with 600 minutes, purchased by the f types, for $210 - q_s^*/5$ dollars; and it also sells a package with q_s^* minutes for a price of $b_s(q_s^*)$ dollars. For example, if $N_s = 5,000,000$, then the two plans are (i) 600 minutes for \$138; and (ii) 360 minutes for \$97.20.

4.7 Bundling

Often we see goods sold in packages. For instance, a CD often contains many different songs. A restaurant may offer a prix fixe menu that combines an appetizer, main course, and dessert. Theater companies, symphonies, and operas may sell season tickets for a variety of different shows. Such packages are called *bundles* and the practice of selling such packages is called *bundling*.

In some instances, the goods are available only in the bundle (*e.g.*, it may be impossible to buy songs individually). Sometimes the goods are also available individually (*e.g.*, the restaurant permits you to order *à la* carte). The former case is called *pure bundling*, the latter case is called *mixed bundling*.

Why bundle? One answer is it can be a useful competitive strategy; for instance, it is claimed that the advent of Microsoft Office, which bundled a word-processor, spreadsheet program, database program, presentation program, etc., helped Microsoft "kill off" strong competitor products that werent bundled (*e.g.*, WordPerfect, Lotus 123, Harvard Graphics, etc.).

Another answer, and one relevant to this chapter, is that it can help price discriminate. To see this, suppose a Shakespeare company will produce two plays, a comedy and a tragedy, during a season. Type-C consumers tend to prefer comedies and, thus, value the comedy at \$40 and the tragedy at \$30. Hence, a type-C consumer will pay \$70 for a season ticket (*i.e.*, access to both shows). Type-D consumers tend to prefer dramas and, thus, value the comedy at \$25 and the tragedy at \$45. Hence, a type-D consumer will pay \$70 for a season ticket. Assume no capacity constraint (*i.e.*, the shows dont sell out) and a constant marginal cost, which we can take to be negligible; that is, $MC = 0$. Let N_t denote the number of type-t theater goers.

If the company sold the shows separately, then its profit is

$$\underbrace{\max\{25(N_C+N_D), 40N_C\}}_{\text{profit from comedy}} + \underbrace{\max\{30(N_C+N_D), 45N_D\}}_{\text{profit from tragedy}} < 70(N_C + N_D)\,.$$

But if the theater company sold season tickets, it would get \$70 from both types and this, as just shown, would yield greater profit. This is an example in which pure bundling does better than selling the goods separately.

For an example where mixed bundling is profit maximizing, change the assumptions so that type-D consumers are now willing to pay only \$20 for the comedy. If $N_C > 4N_D$ (*i.e.*, type-C consumers are more than 80% of the market), then the profit-maximizing solution is to sell season tickets for \$70, but now make the tragedy available separately for \$45.

Observe how the negative correlation between preferences for comedy versus tragedy helps the theater company price discriminate. Effectively, this negative correlation can be exploited by the company to induce the two types to reveal who they are for the purpose of price discrimination. It follows that bundling is related to the forms of second-degree price discrimination considered earlier.

Summary 4.8

In this chapter we explored price discrimination. Ideally, a firm would like to capture all the gains to trade (welfare), which means leaving no profitable trades unmade (no deadweight loss) and capturing all of the consumers surplus. This ideal, which we called the "Holy Grail" of pricing, is known as perfect or first-degree price discrimination.

If all consumers are identical, then perfect price discrimination can be achieved by a two-part tariff. One part, the per-unit charge, is set so as to equate inverse demand and marginal cost. The second part, the entry fee, is set equal to the consumer surplus that each consumer would realize were he or she able to buy as many units as he or she desired at that per-unit charge.

When consumers are not identical, then it is typically not possible to achieve perfect discrimination.

We saw that two-part tariffs are often disguised. Sometimes they are disguised because the per-unit charge is zero. Sometimes they are disguised through packaging; consumers have a choice between buying nothing or a package of fixed size, the price of which is set equal to a consumers total benefit from the number of units in the package. Finally, metering is a way to execute a two-part tariff (although there are also other reasons to use metering).

When consumers are heterogeneous it is sometimes possible to divide them into different populations that are more homogeneous. If the firm can freely identify the population to which a consumer belongs (his or her "type"), then the firm can engage in third-degree price discrimination; that is, offer different tariffs to different populations. In other circumstances, the firm cannot freely identify consumers' types. It can, however, induce them to reveal their types. This form of price discrimination is known as second-degree price discrimination. Two prevalent forms of second-degree price discrimination are using quality distortions and quantity discounts. In any second-degree price discrimination scheme, the consumer type who values the good more (the "high type") is the one who must be induced to reveal his or her type. Because this knowledge is valuable, the high-type consumer retains some of its value in the form of an information rent. On the other hand, there is no distortion in what is sold the high type (*e.g.*, she gets the unrestricted ticket or the welfare-maximizing quantity). The other type (the "low type") doesn't possess valuable information, so earns no information rent. Moreover, to reduce the rent earned by the high type, what is sold to the low type is distorted; either he gets a good of reduced quality or he gets less than the welfare-maximizing quantity (so-called distortion at the bottom). In some cases, it pays to shut out the low type altogether.

A final method of price discrimination explored was bundling. Bundling allows the firm to take advantage of the correlations that exist between consumers' preferences for different products.

All discriminatory pricing is, in theory, vulnerable to arbitrage—the advan-

taged reselling to the disadvantaged. We considered when arbitrage might prevent the use of price discrimination and how firms might, in turn, seek to deter or reduce arbitrage.

5 Game Theory

A tool that is of great use in analyzing strategic situations is *game theory*. Like the tools considered in the previous chapter, game theory is not a substitute for your thinking, but an aid to it.

What is game theory? At a technical level, game theory is a body of mathematical knowledge that has arisen to analyze strategic situations or "games." Its history dates back to the beginning of the last century, although aspects of it can be glimpsed in work of the 19th century as well. As a field, game theory can be a course in itself and there are many thick textbooks dedicated to teaching it. Given this, we cannot expect to do complete justice to the field in a single chapter; we will, instead, consider a few ideas that are particularly useful in business strategy.

5.1 Introduction to Game Theory

Game theory seeks to help people make predictions about how people will behave in strategic situations. Because the earliest strategic situations analyzed were games, such as chess, the field became known as game theory and the strategic situations analyzed are called games. Hence, in a business context, one might speak of *entry games*; that is, strategic decisions in which one or more firms is considering entering a market and one or more incumbent firms is considering how to deter entry. Another "game" might be a *pricing game*—different firms in an industry face the strategic situation of what prices to charge. The actors in such games (*e.g.*, the firms in these two examples) are called the *players* of the game. What the players receive at the end of the game are known as their *payoffs*.

Let's consider a game. Suppose there are two firms, Row Inc. and Column Co.—Row and Column for short. Each firm is contemplating whether to advertise on local television. If neither firm advertises, then each firm's monthly profit—it's payoff—will be $100,000. If one firm advertises, but the other doesn't, then the advertising firm will have profits of $125,000 once the cost of advertising is taken into account; the non-advertising firm will have profits of $50,000. If both firms advertise, then the advertising of one partially cancels out the advertising of the other and *vice versa*; the profits of each, once the costs of advertising are accounted for, will be $75,000. Figure 5.1 illustrates

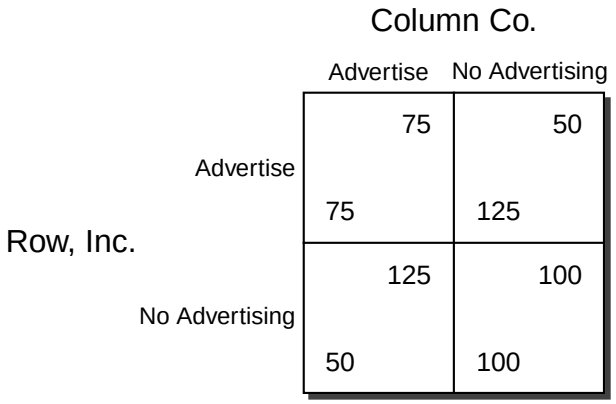

Figure 5.1: A game between Row, Inc. and Column Co. Payoffs are in hundreds of thousands of dollars. In each cell, the payoff to Row, Inc. is the number in the lower left-hand corner and the payoff to Column Co. is the number in the upper right-hand corner. Row, Inc.'s possible strategies are the row headings and Column Co.'s are the column headings.

in the form of a *payoff matrix*.[1]

Having specified the game, we wish to determine what will happen; what strategies will the firms actually play? This is known as *solving the game*. To do so, consider the situation from Row's perspective. If Column will advertise (*i.e.,* will choose the left column of the matrix), then Row's payoff will be \$75,000 if it also advertises, but only \$50,000 if doesn't. So, if Row believes Column will advertise, Row does better to advertise as well. If Column won't advertise (*i.e.,* will choose the right column of the matrix), then Row's payoff will be \$125,000 if it advertises, but only \$100,000 if it doesn't. So, if Row believes Column won't advertise, Row does better to advertise. Observe, therefore, that no matter what Row believes Column will do, Row does better to advertise than not. Hence, that must be what Row does. In the terminology of game theory, advertise is a *dominant strategy* for Row. More generally, a strategy is dominant for a player if it is best for that player regardless of what other players will do. What about Column? A similar analysis reveals that advertise is a dominant strategy for Column ($75 > 50$ and $125 > 100$). That is, Column does better to advertise regardless of what it believes Row will do. Hence, Column will advertise. So the solution of the game—the predicted outcome of how these firms will play—is they will both choose to advertise and each earn monthly profits of \$75,000.

Dominant Strategy: *A strategy that is best to play regardless of the strategies pursued by other players.*

[1]The matrix in Figure 5.1 is technically the representation of the game in *normal form*.

Observe that the outcome of this advertising game—both firms choose to advertise—is not as desirable an outcome for the firms as if neither advertised. In the latter outcome, both firms earn monthly profits of $100,000. But, as just seen, competition leads them to both advertise. This illustrates an important point: Competition between firms can lead to outcomes that the firms wouldn't see as desirable. The game in Figure 5.1 is an example of a common form of game called the *Prisoners' Dilemma*.[2] Other examples of Prisoners' Dilemmas are arms races between two nations (advertise = build up arms) and "rat-racing" between employees seeking to be the one promoted (advertise = work many hours overtime).

Prisoners' Dilemma: *A game in which playing their dominant strategies results in an outcome the players find undesirable.*

Nash Equilibrium

We have seen that if a player has a dominant strategy, then he/she/it will play it. Consequently, it is a straightforward matter to solve games in which all the players have dominant strategies—the solution is they all play their dominant strategies. But what if one or more players doesn't have a dominant strategy? To answer that, we need a more general method of solving games. Fortunately, one exists, it is known as *Nash equilibrium*.[3]

To understand Nash equilibrium, we need first to understand the notion of a best response. A *best response* is the best strategy for you to play against a *particular* set of strategies of your opponents *that you think they will play*. We would say that strategy is the *best response to that set of strategies*. Note that the name "best response" is a bit of a misnomer. You are *not* responding to what other players *actually do*; instead, it is the best response to what you *believe* they will do. To help illustrate this concept, consider Figure 5.2 a somewhat arbitrary game between two players, Row and Column.

Observe, first, that neither player has a dominant strategy. If Row believes Column will play Right, then Row does best to play either Up or Down; but if Row believes Column will play Center, then Row does best to play Middle. So there is no *one* strategy for Row that is best regardless of what Column will do. Similarly for Column: if Column believes Row will play up, then Column does best to play Left; but if Column believes Row will play Middle, then Column does best to play Center. So there is no *one* strategy for Column that is best regardless of what Row will do.

Now let's find best responses. What is Row's *best response* if it believes Column will play Left? The answer is Middle ($9 > 6$ and $9 > 8$). If it believes Column will play Center? Answer: Middle ($10 > 9$). If it believes Column will play Right? Answer: Up or Down ($12 > 11$). Now consider Column. What is

[2]The reason for the name is that one of the earliest illustrations was in the fictional context of two criminals who had been arrested. The prosecutor wants them to confess (the equivalent of advertising in Figure 5.1). The players—the prisoners—do better not confessing (the equivalent of not advertising). However, the prosecutor structures the players' payoffs (the amount of time by which their jail terms are reduced) so that confessing is a dominant strategy for both.

[3]Although some earlier work hinted at the concept of Nash equilibrium, this solution concept is largely due to John Nash, hence the name.

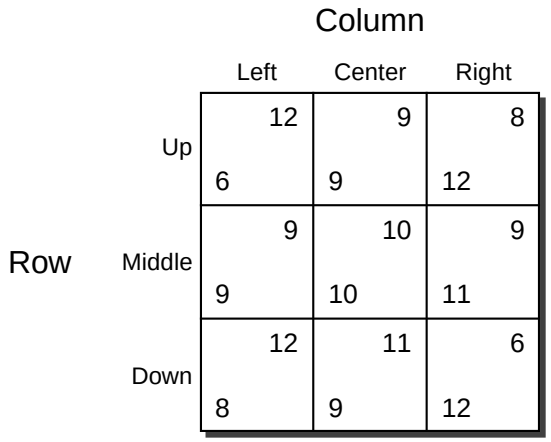

Figure 5.2: A game between Row and Column. In each cell, the payoff to Row is the number in the lower left-hand corner and the payoff to Column is the number in the upper right-hand corner. Row's possible strategies are the row headings and Column's are the column headings.

Column's best response if it believes Row will play Up? Answer: Left ($12 > 9$ and $12 > 8$). If it believes Row will play Middle? Answer: Center ($10 > 9$). If it believes Row will play Down? Answer: Left ($12 > 11$ and $12 > 6$).

We are now in position to solve the game and define a Nash equilibrium. A Nash equilibrium is a situation in which all players are playing best responses to the strategies they believe their opponents are playing and these beliefs are all correct. That is, all players are correctly anticipating what their opponents will do and all are playing accordingly. Observe, therefore, that to find a Nash equilibrium of a game we need to find a set of strategies, one for each player, such that these strategies are all best responses to each other; that is, a situation of *mutual best responses.* What then is a Nash equilibrium of the game in Figure 5.2? We need to find mutual best responses. From the previous paragraph we know:

Nash Equilibrium: *A situation in which all players are playing best responses to each others' strategies.*

Row's Strategy	Column's Best Response	Row's Best Response to that Best Response
Up	Left	Middle
Middle	Center	Middle
Down	Left	Middle

Because we are looking for mutual best responses, we know that a Nash equilibrium will correspond to a row of the preceding table in which the first and last entry of the row are the same. Consequently, we can conclude that the Nash equilibrium of the game is that Row plays Middle and Column plays

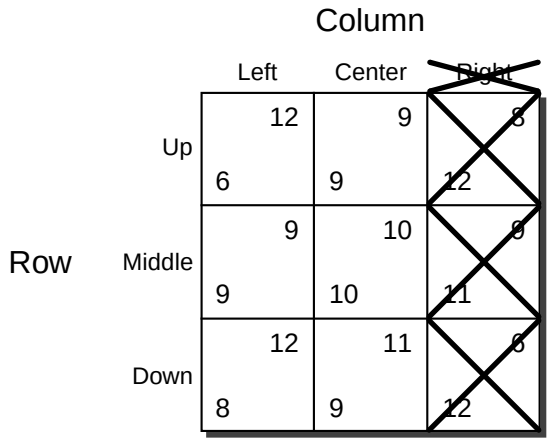

Figure 5.3: The game of Figure 5.2 with the dominated strategy Right "removed."

Center. In other words, we would predict that Middle-Center will be the outcome of the game.

As a second example, recall the game of Figure 5.1. If a strategy is a dominant strategy, then it is best against any particular strategies of the opponents; that is, a best response to all strategies of the opponents. It follows, therefore, that the Nash equilibrium of the Figure 5.1 game is for both firms to advertise. This is a general result: *A solution of game in which all players have dominant strategies is a Nash equilibrium.*

How do Nash equilibria come to be played? That is, how might players facing the Figure 5.2 game arrive at the Nash equilibrium? Presumably, they engage in *strategic thinking*. Specifically, they put themselves in their opponent's shoes and ask what they would do were they their opponent. For instance, Row, putting herself in Column's shoes, would recognize that no matter what Column thought she, Row, would do, Column would never wish to play Right: His payoff from playing Right never exceeds what he would get by playing Left or Center and is always less than playing one or the other or both. In the terminology of game theory, Right is a *dominated strategy*. Players don't play dominated strategies. Hence, Row would reason that Column would never play Right. Hence, the game is effectively the game shown in Figure 5.3. In this "pruned" game, Row would realize she has a dominant strategy, namely Middle. So Row would conclude that she should always play Middle. Column, likewise, would go through the same thinking and conclude that Row would, thus, play Middle. Column's best response to Middle is, as shown before, Center; hence, Column would conclude he should play Center.

5.2 The Bertrand Trap

Game-theoretic analysis and the notion of Nash equilibrium can be extended to strategic situations in which the players make choices over essentially continuous variables, such as price. In particular, it is worth exploring a particular pricing game known as the Bertrand model; or, as I like to refer to it, the Bertrand trap.[4]

The Six Conditions Underlying the Bertrand Model

The *Bertrand model* considers price competition among two or more firms under the following six conditions:

Homogeneity: The good or service produced by the firms is identical in quality, image, and function across the firms; that is, customers perceive no difference between the product of one firm and that of another.

Knowledge of price: Customers know the prices being charged by all the firms prior to making purchasing decision.

No lock-in: Customers are not locked into any firm. They incur no costs if they switch firms and they exhibit no firm loyalty.

All Sales to Lowest-Price Firm: *Customers buy from the firm charging the lowest price under Bertrand competition.*

Observe that these first three conditions imply that, because the only possible differentiation between the firms are the prices they charge, *customers will buy from the firm charging the lowest price.*

The last three conditions are:

Constant unit cost: All firms have the same constant unit cost; that is, no firm has a cost advantage over any other and there are no economies of scale.

No capacity constraints: Any one firm can meet all demand; that is, no firm is constrained by its capacity.

Myopic play: Firms do not consider their future interactions in making pricing decisions today.

Not many markets are described by all six of these conditions, although a few, such as commodity markets (*e.g.*, the market for hard red winter wheat) or those that involve bidding for large contracts (*e.g.*, bidding to supply standard parts to the military) come close. This, however, is okay, because the Bertrand trap is largely a cautionary tale—a warning of what could happen or what could go wrong if you don't make good strategic decisions today.

[4] The Bertrand model was developed by the French mathematician Joseph Bertrand.

Analysis of the Bertrand Model

To analyze the Bertrand model—solve the game—we need some notation. Let c denote the constant unit cost; that is, c is some number of dollars per unit. Let $D(p)$ denote market demand (*i.e.*, how much, in total, consumers wish to purchase) if the price for a unit of the product is p. Note p is some number of dollars per unit.

To make the analysis as straightforward as possible, let's consider a duopoly. Let the two firms in the duopoly be called A and B. Let $d(p_A)$ denote the demand faced by A if its price is p_A. Similarly, let $d(p_B)$ be B's demand if its price is p_B. For the moment, let's suppose A's price is no greater than B's. Recalling that customers all go to the firm charging the lower price, if B's price exceeds A's, then B's demand is zero if its price exceeds A's. In other words, if $p_B > p_A$, then B makes no sales. Moreover, if $p_B > p_A$, then A makes all the sales. In this case, note $d(p_A) = D(p_A)$.

What if $p_B = p_A$? In this case, we may suppose that the firms divide the demand at that common price between them, with neither firm getting all the sales.[5]

What are the firms' payoffs? Answer: their profits. Firm A's profits are $d(p_A) \times (p_A - c)$ and B's are $d(p_B) \times (p_B - c)$. For example, suppose $D(p) = 100 - p$, $p_A = 2$, $p_B = 3$, and $c = 1$, then A's profit would be

$$d(p_A) \times (p_A - c) = (100 - 2) \times (2 - 1) = 98\,,$$

where the final number is in dollars. B's profit would be

$$d(p_B) \times (p_B - c) = 0 \times (3 - 1) = 0\,.$$

To find the Nash equilibrium of this game, we need to find a pair of mutual best responses. This entails determining what each firm's best response is to the possible pricing of the other. Given the symmetry between the firms, we can look at just one firm's best response to the other's pricing. Let A be the one firm. There are three cases to consider:

$p_B > c$: If A prices above p_B, it has no sales and, thus, zero profit. If it prices below p_B it captures the entire market and earns profit $D(p_A) \times (p_A - c)$. Hence $p_A > p_B$ cannot be a best response and among all the $p_A < p_B$, the best is $p_A = p_B - \varepsilon$, where $\varepsilon > 0$ but exceedingly small. If it matches B's price, its profit is *less than* $D(p_B)(p_B - c)$. By setting $p_A = p_B - \varepsilon$, A captures the entire market with profits per sale approximately equal to $p_B - c$, which is better than it does by matching B's price. Conclusion: A's best response is to undercut B's price very slightly and capture the entire market.

[5] To be precise, this assumption is not necessary. Because, however, it ensures symmetry between the firms, it helps to keep the exposition as straightforward as possible.

$p_B < c$: If A prices above p_B, it has no sales and, thus, zero profit. If it prices at or below p_B it will lose money because price is below cost. Conclusion: A's best responses are all prices greater than B's.

$p_B = c$: If A prices at or above p_B, it makes zero profit. If it prices below B it captures the market, but loses money on every sale. Conclusion: A's best responses are all prices greater than or equal to B's.

What then is a situation of *mutual* best responses? Answer $p_A = p_B = c$. To see this, recall that if $p_B = c$, then among A's best responses is $p_A = c$. By symmetry, if $p_A = c$, then among B's best responses is $p_B = c$. Moreover, there are no other Nash equilibria: Given the analysis above, one or the other or both firms are *not* playing a best response if it is (they are) pricing below cost. Nor can it be an equilibrium for one of them to price above cost; either that firm should drop price to undercut its rival or its rival should raise price to be just below it.

Bertrand Trap: *Given the conditions of the Bertrand model, firms are* trapped *earning zero profits.*

Observe, then, that in the equilibrium of the Bertrand model both firms are pricing at unit cost and, therefore, both firms are earning zero profits. Earning zero profits is an undesirable outcome, which is why I refer to this as the *Bertrand trap*. Although this analysis was done for a duopoly, it readily applies to an industry of any size (except, of course, a monopoly). In a market in which the entire market can be captured by undercutting your rivals, the only equilibrium price can be pricing at cost.

5.3 Avoiding the Bertrand Trap

The conclusion of the last section, namely that the outcome of Bertrand competition is zero profits, is unavoidable if you find yourself engaged in Bertrand competition. The only way to avoid the dire outcome of the Bertrand trap is to avoid playing the Bertrand game in the first place. That is, like the Ghost of Christmas Future in Dickens's *A Christmas Carol*, the Bertrand model is a warning of what will happen to you unless you take steps *today* to change things. What things should you change? Well, you need to make sure that one or more of the conditions that underlie the Bertrand model aren't met. This section considers ways to change those conditions.

Differentiate Your Product

One factor that makes competition less fierce is the degree to which rivals' products are differentiated. This suggests that one condition of the Bertrand model that a firm might try to relax is the homogeneity of the products.

To see how product differentiation can help, assume customers are uniformly distributed in their inherent preferences for the products of two competing firms, named 0 and 1. Specifically, each consumer has a preference x, where $0 \leq x \leq 1$. Consumers also have a common intensity of preference,

$t \geq 0$, which can be interpreted as the degree of product differentiation (higher t means greater differentiation). A consumer whose inherent preference is x loses value tx if he buys from firm 0 and he loses value $t(1-x)$ if he buys from firm 1. Hence a consumer whose $x = 0$ is very partial to the product produced by firm 0 and a consumer whose $x = 1$ is very partial to the product produced by firm 1. A consumer whose $x = 1/2$ likes the two products equally well. If there are N total customers, then there are $N \times X$ with a value of $x \leq X$ and $N \times (1 - X)$ with a value of $x \geq X$.

Each consumer buys at most one unit of the good in question. His value for the good is $v - tx - p$ if he buys it from firm 0 at price p (he doesn't buy, of course, if that quantity is negative) and $v - t(1-x) - p$ if he buys it from firm 1 (again, he doesn't buy if that quantity is negative).

What is each firm's demand? To answer, we'll limit attention to the case in which v is relatively large. Let p_0 be the price charged by firm 0 and p_1 be the price charged by firm 1. A consumer will buy from firm 0 if his payoff from doing so exceeds his payoff from buying from firm 1; that is, he buys from firm 0 if

$$v - tx - p_0 > v - t(1-x) - p_1 .$$

He buys from firm 1 if the opposite holds; that is, he buys from firm 1 if

$$v - tx - p_0 < v - t(1-x) - p_1 .$$

(Because we've limited attention to v large, we can ignore the case in which he prefers to buy from neither.) Firm 0's demand will be the number of consumers for whom the first inequality holds and firm 1's demand will be the number of consumers for whom the second inequality holds. To determine these two demands, we begin with consumers who are just indifferent between buying from 0 or from 1; these are the consumers for whom

$$v - tx - p_0 = v - t(1-x) - p_1 . \tag{5.1}$$

Solving (5.1) for x, we find that consumers with x less than

$$\frac{t + p_1 - p_0}{2t}$$

buy from firm 0 and those with an x greater than that amount buy from firm 1. If we call that amount X, then, from above, there are $N \times X$ customers with x less than that amount and $N \times (1 - X)$ with x above that amount. Consequently, firm 0's demand, $d_0(p_0)$, is

$$d_0(p_0) = N \times \frac{t + p_1 - p_0}{2t} = \frac{Nt + Np_1}{2t} - \frac{N}{2t} p_0 .$$

Observe that demand is linear. Firm 1's demand, $d_1(p_1)$, is

$$d_1(p_1) = N \times \left(1 - \frac{t + p_1 - p_0}{2t}\right) = \frac{Nt + Np_0}{2t} - \frac{N}{2t} p_1 .$$

Observe it, too, is linear.

We need to solve this variation of the Bertrand game; that is, find the Nash equilibrium. To do so, recall the discussion of pricing in Chapter 3. We know that the profit-maximizing price—the best response given the price of the rival firm—is the average of the choke price and unit cost, c. For firm 0, the choke price is $t + p_1$, so its best response to p_1 is

$$p_0 = \frac{t + p_1 + c}{2}.$$

For firm 1, the choke price is $t + p_0$, so its best response to p_0 is

$$p_1 = \frac{t + p_0 + c}{2}.$$

Because we're looking for a situation of mutual best responses, both these equations must hold in equilibrium. Algebra will reveal that these equations are both satisfied if

$$p_0 = p_1 = t + c. \tag{5.2}$$

Product Differentiation: *If products are differentiated, then firms avoid the Bertrand trap.*

Observe, critically, that as long as there is differentiation (*i.e.*, $t > 0$), price will be greater than cost; differentiation has allowed the firms to avoid the Bertrand trap!

This last point can also be made by calculating each firm's profit in equilibrium. Substituting the equilibrium prices (those given by expression (5.2)) into the expressions for demand, we find that $d_0(p_0) = d_1(p_1) = N/2$. Each firm's profit is demand times the difference between price and unit cost; hence,

$$\text{profit} = \frac{N}{2} \times \big((t + c) - c\big) = \frac{Nt}{2}.$$

A firm's profit is zero if there is no differentiation (*i.e.*, $t = 0$) and positive if there is. Moreover, for relevant values of the parameters, *profit is increasing in the degree of differentiation* (*i.e.*, increasing in t).

This analysis underscores our discussion of product differentiation in the previous chapter. In particular, it shows why firms often work hard to differentiate their products (either in reality or in terms of image). If everyone thought all colas were the same, then price competition between Coke and Pepsi would be fiercer than it is.

Suppressing Price Information

Another key condition of Bertrand competition is that customers know the prices being charged by all firms. Suppose, instead, that customers had limited knowledge of prices.

To be concrete, assume there are two firms A and B. Half the potential customers know the price being charged by A and half the price being charged by B. At a cost of $k > 0$, a customer who knows the price being charged by one firm can learn the price being charged by the other. The cost k can be

considered the opportunity cost of the time required to call the other firm or hunt up its web site, etc.

Consider a customer who knows the price at firm A, p_A, but not the price at firm B, p_B. Although she doesn't know the price at B, she holds some idea of what it might be. In particular, she either believes it is sufficiently less than p_A to make incurring cost k to learn p_B worthwhile (*i.e.*, she believes $p_A - p_B > k$) or she believes it's not sufficiently less (*i.e.*, she believes $p_A - p_B \leq k$). If she holds the former belief, she will learn B's price and purchase from whichever firm has the lower price. If she holds the latter belief, she will not learn B's price and she will purchase from A. A requirement of any equilibrium is that, in equilibrium, her beliefs be correct.[6]

Our first result is that there is a Nash equilibrium in which both firms charge the *monopoly* price. To see this, observe that if customers believe the two firms are charging the same price, then they will never incur cost k to search out prices. Consequently, those customers who know A's price will buy from A (if they buy at all) and those who know B's price will buy from B (if they buy at all). Consequently, each firm is "the only game in town" for the customers who know its price. In other words, it is as if each firm is a monopolist *vis-à-vis* half the customers. So each firm wishes to maximize

$$\frac{1}{2}D(p) \times (p - c)\,, \tag{5.3}$$

given its demand at price p is $\frac{1}{2}D(p)$ (half the market). Observe that the $\frac{1}{2}$ in expression (5.3) doesn't affect what the profit-maximizing price is; in particular, the profit-maximizing price will be the same as the price that maximizes

$$D(p) \times (p - c)\,.$$

But that price is the price a monopolist would charge.

We need to verify that charging the monopoly price is the best response to your rival charging the monopoly price in this game. If you charge more, then your profits cannot be greater—either you keep all your customers, but are charging too much, or you lose your customers. If you charge less, then, unlike Bertrand competition, you don't attract any of your rival's customers—they don't know you've cut your price and, because they believe you won't, they don't bother to check. So you have the same customers, but you're charging them less than the profit-maximizing price. We can conclude that, indeed, your best response to your rival's charging the monopoly price is for you, too, to charge the monopoly price. *The firms have avoided the Bertrand trap!*

Price Ignorance: *If consumers find it costly to acquire knowledge of prices, then firms avoid the Bertrand trap.*

Moreover, this is the only possible equilibrium. To see this, note, first, that there cannot be an equilibrium in which consumers learn price. If consumers were to learn all prices, then the logic of Bertrand competition would drive the

[6] **OPT** This requirement arises because a rational consumer understands the game being played between the firms and anticipates what the equilibrium will be. Her beliefs about the difference in prices must, then, accord with the equilibrium.

firms to charge unit cost. But if they are charging the same price in "equilibrium," then the consumers will rationally not incur cost k. So the only equilibria must be those in which the consumers don't learn prices. But then, as just seen, each firm is a *de facto* monopoly and should, thus, charge the monopoly price.

This analysis helps to explain why many retailers seek to make comparison shopping difficult. The less consumers know about prices, the less fierce competition will be.

Lock-in Customers

As discussed in the previous chapter, one way to reduce the ferocity of competition is to lock-in customers. Although a complete analysis of lock-in requires techniques beyond the scope of this text, an intuitive understanding of why lock-in is effective can be seen from our earlier analysis of the Bertrand model.

Consider again two firms, A and B. Suppose half the potential customers incur switching costs going from A to B (are locked in to A) and, likewise, half incur switching costs going from B to A (are locked into B). Let the cost to a consumer of switching firms be $s > 0$. This means that a customer locked into A will only switch to B if B's price, p_B, is sufficiently less than A's, p_A, that it is worth incurring the switching cost; that is, such a customer switches if and only if $p_A - p_B > s$.

Recall the logic of Bertrand competition: In Bertrand competition, B can steal all of A's customers by just undercutting A's price (*i.e.*, by charging $p_A - \varepsilon$, where, recall, ε denotes an arbitrarily small positive amount). When, however, there is a switching cost, B can't just undercut A's price by a little, it has, instead, to undercut by at least s in order to steal customers from A. Consider a situation in which the two firms are expected to charge the same price (greater than unit cost). If B undercuts A by enough to steal any customers, it is also lowering the price to those customers locked into it by s. Hence, on the one hand, it gains, but on the other, it loses. Correspondingly, B will be less tempted to undercut A when customers incur a switching cost than when they don't.

Given the limited amount of game theory covered in this text, it is not possible to derive or even clearly state what the equilibrium is for this game. We can, though, verify that whatever the equilibrium is, it entails firms pricing *above* unit cost. To see this, we need merely show that pricing at unit cost is *not* a best response to your rival's pricing at unit cost; that is, show that $p_A = p_B = c$ is *not* a situation of mutual best responses (c recall denotes the unit cost). What is the best response to your rival's charging unit cost? Observe that you can charge up to s more than your rival and not lose any customers. Because positive profit beat zero profit, your best response is, thus, some price p between c and $s + c$ (*i.e.*, $c < p < s + c$). Observe it is *not* charging unit cost.

This analysis helps to explain why firms seek to raise consumers' switching costs. As discussed in the previous chapter, methods of doing so include the use of loyalty programs (*e.g.*, frequent-flier programs); making your product incompatible with that of your rivals (*e.g.*, only your printer cartridges fit into

your printers); and causing consumers to incur a cost if they switch (*e.g.*, as was once the case when you switched mobile-phone service provider and had to give up your existing phone number; or as would currently be the case if you switched broadband service provider and had an email account through your existing provider).

Obtain a Cost Advantage

Recall that the Bertrand model posited that the firms had identical unit costs. What if one firm had a lower unit cost than others? As before consider two firms, A and B, and suppose that A's unit cost, c_A, is less than $B's$, c_B. From our earlier analysis of the Bertrand model, we know that there cannot be an equilibrium in which either or both firms are charging a price above c_B. We also know that it is a dominated strategy for any firm to charge a price below its cost—there is no way to make a profit if you're pricing below cost. This tells us that firm B cannot be charging above c_B in equilibrium nor below it. The conclusion is, thus, that it is charging c_B. What about firm A? Observe it can undercut B slightly, capture the entire market, and still make a profit because $c_A < c_B$. The equilibrium is that B charges c_B and A charges $c_B - \varepsilon$ (where, again, ε is an arbitrarily small positive amount). We can conclude that *having a cost advantage over your rivals allows you to avoid the Bertrand trap.*

Limit Capacity

One of the factors that makes a firm in the Bertrand model so eager to undercut their rivals is that all demand comes to it and, moreover, it can handle all that demand. But what if firms had limited capacity? That is, what if a given firm couldn't meet the demand of an entire market because it simply couldn't produce that much in the requisite amount of time? Although a complete answer requires techniques beyond the scope of this text, an intuitive understanding of why limited capacity aids firms in avoiding the Bertrand trap can be seen from our earlier analysis.

Similar to what we did in our discussion of lock-in, we'll focus on why the Bertrand equilibrium—all firms price at unit cost—cannot be an equilibrium when there are binding capacity constraints. As before, consider two firms, A and B. Suppose market demand at their common unit cost of c is $D(c)$. Assume, however, neither firm can produce that much (although, perhaps, combined they can). We need to show that pricing at unit cost is *not* a best response to your rival's pricing at unit cost. To see this, suppose B is pricing at unit cost and its capacity is C_B, where, as just assumed, $C_B < D(c)$. If A charges unit cost, it makes no profit. Suppose A charged above unit cost. Although all customers would prefer to buy from B, B can handle only C_B of the demand. The remaining $D(c) - C_B$ of demand is unmet and those customers will have no choice but to turn to A. Consequently, A will make sales and at a price above unit cost; hence, A will make a profit. A profit beats no profit, so we see that pricing at unit cost is *not* a best response for A to B's pricing at unit cost. Be-

Limited Capacity: *By limiting capacity, firms can avoid the Bertrand trap.*

cause we don't have a situation of mutual best responses, both firms pricing at unit cost is *not* an equilibrium. Because pricing below cost is a dominated strategy, we are left to conclude that the equilibrium, whatever it is, involves the firms earning positive profits. In other words, by having limited capacity, the firms avoid the Bertrand trap.

This result helps to explain why fierce price competition often breaks out in dying industries (industries for which demand is going away). Before their industry was dying, firms in the industry had a healthy amount of capacity; but not so much that they were in the Bertrand trap. As the industry dies (*i.e.*, demand shrinks), one or more firms in the industry become capable of handling all the demand that remains. Price competition will get fiercer as a consequence because the firms have fallen into the Bertrand trap.

Having a Long Horizon

The sixth condition that caused firms to fall into the Bertrand trap was that they played myopically; that is, they did not take into consideration future play among them. Not surprisingly, if firms ensure that condition doesn't hold, they will be able to avoid the Bertrand trap. Because the consequences of having a long horizon on the play of games is a very important concept, with applications beyond the Bertrand model, we will devote the next section to it.

5.4 Repeated Interactions

Many strategic situations are repeated over time. For instance, setting prices or advertising levels is something that Coke and Pepsi do repeatedly. The gas station at one corner of a busy intersection competes day in and day out with its rival across the street. Or, more mundanely, I interact with the same people over and over again, as do you, albeit with a different set of people. As we will see, strategic interactions are quite different when they are repeated over time and players care about the future than when they are played only once or players ignore the future.

To illustrate the power of repetition, let's begin with a concrete example, namely infinite repetition of the advertising game of Figure 5.1. (Why "infinite"? you might ask; the answer and what we really mean by infinite will become clear later in this section.) We refer to the game shown in Figure 5.1 as the *stage game*. Each period, the players play the stage game as shown and receive the payoffs that correspond to the outcome of the stage game they play. So, for instance, if, in the 12th period of play, Row advertises and Column does not, then Row gets \$125,000 in that period and Column gets \$50,000 in that period.

A Review of Discounting and Present Value

Because players are getting paid over time, we need to *discount* payoffs that will be received in the future. *Starting* in any given period, the players value the payoffs of that period at face value, but they discount future periods' payoffs. Recall from basic finance if r is the interest rate across periods, then a payment of x dollars to be received one period in the future has a *present value* of

$$\frac{x}{1+r}$$

dollars. That is, x dollars to be paid one period in the future is worth the same as

$$\frac{x}{1+r}$$

dollars paid today. So for example, if $x = 11$ and $r = .1$ (*i.e.*, the interest rate is 10%), then we see that the promise of \$11 in one period's time is worth \$10 today (= \$11/1.1).

If a payoff, x, is to be received *two* periods hence, then it's present value is

$$\frac{x}{(1+r)^2}.$$

This is can be seen as follows. Suppose we were living *one* period hence, then the promised future payoff would,then, be only one period hence and, thus, worth

$$\frac{x}{1+r}$$

one period forward from today. What's the value today of receiving $x/(1+r)$ in the next period? It's

$$\frac{\frac{x}{1+r}}{1+r} = \frac{x}{(1+r)^2}$$

using the logic of the previous paragraph. We can repeat this logic as many times as necessary; hence, we can conclude that the present value of x to be received t periods in the future is

$$\frac{x}{(1+r)^t}.$$

Suppose you will receive a sequence of payoffs, x_0, x_1, x_2, ..., forever, where x_t is the payoff received t periods in the future (hence, x_0 is the payoff received today). What is the present value of this sequence? It is the sum of the present values of the individual payoffs. Since the present value of x_t is

$$\frac{x_t}{(1+r)^t},$$

that sum is

$$x_0 + \frac{x_1}{1+r} + \frac{x_2}{(1+r)^2} + \cdots + \frac{x_t}{(1+r)^t} + \cdots = x_0 + \sum_{t=1}^{\infty} \frac{x_t}{(1+r)^t},$$

where, recall, the notation $\sum_{t=1}^{\infty}$ means the sum of the terms from the 1st to the "infinity-th."

Now suppose that all the x's in the sequence of payoffs are the same, just x. Then this last formula becomes

$$x + \sum_{t=1}^{\infty} \frac{x}{(1+r)^t} = x + x \sum_{t=1}^{\infty} \frac{1}{(1+r)^t}$$

(recall the distributive rule lets us pull a common term out of a sum). As you recall from basic finance, the sum in the last expression is equal to $1/r$; that is,[7]

$$\sum_{t=1}^{\infty} \frac{1}{(1+r)^t} = \frac{1}{r}.$$

Consequently, the present value of receiving x each period forever, starting today, is given by

$$x + \frac{x}{r}. \tag{5.4}$$

Note that x/r is the present value of receiving x each period forever, *starting tomorrow*.

Obtaining Cooperation in a Prisoners' Dilemma Game

Now let's return to game theory and the game in Figure 5.1. What the players in this game would like to do is achieve the *cooperative outcome*, which is neither player advertises and both get a payoff of \$100,000. We saw above that, if they play just once (or are myopic), then they fail to achieve the cooperative outcome; instead, both players play their dominant strategy of advertise, leading them to an undesirable outcome in which each gets a payoff of \$75,000.

But what if the game were repeated infinitely and the players weren't myopic? In this case, observe it is feasible for one player to "punish" another for advertising through that player's *future* play. In particular, if one of the players

[7] **OPT** To derive this, it will prove convenient to let $\delta = 1/(1+r)$. So the value of the sum, call it S, is given by

$$S = \sum_{t=1}^{\infty} \delta^t. \tag{♠}$$

Multiply both sides by δ, which yields

$$\delta S = \sum_{t=1}^{\infty} \delta^{t+1} = \sum_{t=2}^{\infty} \delta^t. \tag{♣}$$

Subtracting (♣) from (♠) yields

$$S - \delta S = \delta.$$

Hence,

$$S = \frac{\delta}{1-\delta} = \frac{\frac{1}{1+r}}{1-\frac{1}{1+r}} = \frac{\frac{1}{1+r}}{\frac{1+r-1}{1+r}} = \frac{1}{r}.$$

deviates from the cooperative outcome (*i.e.*, advertises), then the other player punishes the first by refusing to cooperate in the future.

How, formally, does this work? Consider the following, two-part strategy:

1. Provided no player has ever advertised in the past or it is the first period, then do *not* advertise.

2. If a player has ever advertised in the past, then advertise.

Observe that if both players adhere to it, then they are seeking to sustain cooperation (not advertising) by the threat of *reverting* to non-cooperative play (advertising). Recalling the concept of Nash equilibrium, both players will adhere to this strategy if playing this strategy is a best response to the other player's playing it; that is, if they represent mutual best responses. Do they? To answer this, we need to know what a player gets by adhering to the strategy and what the player would get by deviating. If a player adheres, then, because this player is anticipating the other will adhere, this player anticipates getting \$100,000 every period. From expression (5.4), this has a present value of

$$\$100,000 + \frac{\$100,000}{r} = \text{value of adhering}.$$

If a player deviates (*i.e.*, advertises), then, because this player is anticipating the other will adhere, this player anticipates getting \$125,000 this period (this player has advertised, while the other has not); but, in future periods, this player will only get \$75,000 per period—the payoff if they revert to advertising in every future period. From our review of present value, we know this stream of payoffs has a present value of

$$\$125,000 + \frac{\$75,000}{r} = \text{value of deviating}.$$

Adhering is a best response to adherence if and only if the value of adhering exceeds the value of deviating; that is, if and only if

$$\$100,000 + \frac{\$100,000}{r} \geq \$125,000 + \frac{\$75,000}{r},$$

which is equivalent to

$$\frac{\$25,000}{r} \geq \$25,000, \tag{5.5}$$

which must hold provided the interest rate doesn't exceed 100% (*i.e.*, $r \leq 1$).

Value of a Long Horizon: *Infinitely repeated play can allow the players to achieve better outcomes than they could if the stage game were played just once.*

Myopic play—not caring about the future—can be viewed as equivalent to there being an infinite interest rate. If the interest rate were infinite, then expression (5.5) wouldn't hold, so cooperation wouldn't be possible; precisely the same conclusion we obtained analyzing the stage game back in Section 5.1.

One question you might have is whether a player would actually follow through on part 2 of this strategy; that is, would a player revert to advertising

every period in order to punish a rival player who advertised? The answer is yes. To see why, observe that if you think your rival is going to advertise—and do so regardless of what you do—then your best response (indeed, dominant strategy) is to advertise too.

Avoiding the Bertrand Trap Through Repetition

Let's now employ our knowledge of repeated games to see how having a long horizon can help firms avoid the Bertrand trap. Assume all of the conditions—except myopic play—of the Bertrand model apply. We'll consider a situation with F firms, $F \geq 2$. Assume that if the F firms are all charging the same price, then each firm gets $1/F$th of demand at that price.

Our goal, similar to that of the previous subsection, is to see whether the firms can sustain a cooperative outcome. What would be the cooperative outcome in this situation? It would be for all firms to charge the price a monopoly firm would. Call that price p^*. Observe that, in the cooperative outcome, each firm would get

$$\frac{1}{F}D(p^*) \times (p^* - c) = \frac{\pi^*}{F},$$

where, as before, $D(p^*)$ denotes the *market* demand at price p^* and c is the constant unit cost. Observe, too, the implicit definition of π^*,

$$\pi^* = D(p^*) \times (p^* - c);$$

the quantity π^* is the profit a monopoly firm would make in this market.

As we did in the previous subsection, consider a two-part strategy:

1. Provided no firm has ever set its price below the monopoly price, p^*, in the past or it is the first period, then set price at p^*.

2. If a firm has ever set its price below p^* in the past, then set price at unit cost, c.

Would all firms following this strategy constitute a Nash equilibrium? That is, is this strategy a best response to itself being played by $F - 1$ other firms? To answer this, we need to know what a firm gets by adhering to the strategy and what it would get by deviating. If a firm adheres, then, because this firm anticipates its $F - 1$ rivals will adhere, this firm anticipates getting π^*/F every period. From expression (5.4), this has a present value of

$$\frac{\pi^*}{F} + \frac{\pi^*/F}{r} = \text{value of adhering}.$$

If a firm chooses to deviate, what would be its best deviation? Answer, pricing just below p^* and stealing the entire market. Because its price is just below p^*, its profit is arbitrarily close to the monopoly profit, π^*. In fact, it is sufficiently close that we can treat its payoff if it deviates as π^*. However, in all subsequent

periods, its profit will be zero because firms are pricing at unit cost. The present value of this stream of payoffs is, therefore,

$$\pi^* + \frac{0}{r} = \pi^* = \text{value of deviating}\,.$$

Adhering is a best response to adherence if and only if the value of adhering exceeds the value of deviating; that is, if and only if

$$\frac{\pi^*}{F} + \frac{\pi^*/F}{r} \geq \pi^*\,,$$

which, multiplying both sides by F/π^*, will be true if and only if

$$1 + \frac{1}{r} \geq F\,,$$

which, doing a bit of algebra, is equivalent to

$$\frac{1}{r} \geq F - 1\,. \tag{5.6}$$

If $F = 2$ (*i.e.*, it's a situation of duopoly), then expression (5.6) will hold true provided the interest rate does not exceed 100%. If $F = 3$, then that expression holds true if the interest rate does not exceed 50%. More generally, that expression holds true if

$$\frac{1}{F-1} \geq r\,. \tag{5.7}$$

From expression (5.7), we see that *the more firms there are in the industry, the smaller the range of possible interest rates for which the firms can sustain the cooperative outcome*. In other words, the more firms there are in the industry, the less likely it is that they will be able to sustain the cooperative outcome. If they can't sustain the cooperative outcome, then they are stuck in the Bertrand trap. Observe that this result helps to justify why competition is generally less fierce in more concentrated industries (*i.e.*, fewer firms makes it more likely the firms will avoid the Bertrand trap).

Effect of Number of Firms: *The more firms in an industry, the more likely it is that the firms will find themselves stuck in the Bertrand trap (i.e., making zero profit).*

Why Infinite Repetition?

One question that you might have is why the focus on *infinitely* repeated games? What about games that are repeated a fixed number of periods?

The answer is that if any of the games we've considered is repeated a fixed number of times, then the only equilibrium is repetition of the equilibrium of the stage game. Why is this? The answer can be seen by recalling that what made players cooperate *today* was the threat of reverting to a bad outcome *tomorrow* if players failed to cooperate today. But if there is no tomorrow, then there is no threat, and, hence, no cooperation.

To be slightly more rigorous, consider the last period of play. For a game such as the advertising game of Figure 5.1, the players will clearly play their dominant strategies—that is, both will advertise—because it's impossible for the to suffer any future repercussions from having done so. Now consider the players' thinking in the period before the last, the penultimate period. Each player knows that both players will play their dominant strategies in the last period *no matter what happens today*. Hence, there are no repercussions for not cooperating in the penultimate period either, which means the players will play their dominant strategies (*i.e.,* both will advertise). This "unraveling" argument works all the way back to the first period; knowing that, in all subsequent periods, the players will play their dominant strategies, there can be no motive not to play your dominate strategy today.

Known Game End: *If players know precisely when the game will end, then it is not possible to sustain cooperative play.*

The same reasoning applies to Bertrand competition that is repeated only a fixed number of times. In the last period, with no future periods to induce cooperative behavior, the firms will be in the Bertrand trap and, thus, all price at unit cost. In the penultimate period, recognizing that no matter what is done today, they will all price at unit cost tomorrow, there is nothing to deter firms from undercutting their rivals on price; but if there is nothing deterring them from doing so, then the logic of the Bertrand trap dictates that they will end up pricing at unit cost. Again, this unraveling argument works all the way back to the first period.

What Do We Really Mean by Infinite?

At this point, you might object by noting that nothing lasts forever. People die. Firms go out of business. At some point, life on Earth will become impossible.[8] Even if humans somehow escape the Earth's fate, most cosmological theories hold that the Universe itself has only a finite life. Given all this, what do we mean by an infinitely repeated game?

Obviously, we can't mean literally infinitely repeated. But if so, then don't the arguments from the previous subsection rule out any cooperation? The answer is no, the reason being that the unraveling argument given above relied on the players knowing when the game would end; that is, what would be the last period. In many situations, the players don't know when the game will end. How does this help? Well as long as the players think there is a chance the game will continue to the next period, to the period after that, and so forth, there is a future to deter the players from acting uncooperatively today.

To be more rigorous, suppose that, conditional on reaching any given period, the probability of playing the stage game again in the next period is $\beta > 0$. Because β is a probability, $\beta \leq 1$. If $\beta = 1$, then the game literally lasts forever, so we're interested in $\beta < 1$. What's the value today of receiving x next period if the probability of there being a next period is β? Well, if we were *certain* to

[8] The time remaining for life on this planet is unknown. But because the Sun is becoming increasingly hotter, calculations suggest that surface water will be gone from our planet in one billion years. Five billion years from now, the Sun will become a red giant and scorch the Earth.

get x, it would be

$$\frac{x}{1+r}.$$

But if we're uncertain—as we are because $\beta < 1$—then we have to multiply that amount by the probability we get it; that is, not only do we need to discount for the time until payment, but also for the uncertainty of payment. So, x in one period's time when there is only a probability of β of there being a next period is worth, today,

$$\frac{x\beta}{1+r}.$$

What about if x is to be paid us in two periods? We know the present value is

$$\frac{x}{(1+r)^2},$$

but again we need to discount for the uncertainty of payment. What's the probability of there being a period two periods hence? Well, its probability β one period hence and probability β one period hence after that, so the probability of a period two periods hence is $\beta \times \beta = \beta^2$. So the value discounted for *both* time and uncertainty—the *expected present value*—is

$$\frac{x\beta^2}{(1+r)^2}.$$

Generalizing, the expected present value of x t periods hence is

$$\frac{x\beta^t}{(1+r)^t} = x\left(\frac{\beta}{1+r}\right)^t. \tag{5.8}$$

In essence, rather than the discount factor being $1/(1+r)$ it is $\beta/(1+r)$. Let's define the *risk-adjusted interest rate*, ρ, by the expression

$$\frac{1}{1+\rho} = \frac{\beta}{1+r},$$

which, following some algebraic manipulation, tells us

$$\rho = \frac{1+r}{\beta} - 1. \tag{5.9}$$

Suppose that we are to receive x every period starting today and every period until the game ends. What is the expected present value of that stream? Using expression (5.8) and the definition of ρ, it is

$$x + \sum_{t=1}^{\infty} x\left(\frac{\beta}{1+r}\right)^t = x + x\sum_{t=1}^{\infty}\frac{1}{(1+\rho)^t} = x + \frac{x}{\rho}. \tag{5.10}$$

Observe if payments started tomorrow (one period hence), then the expected present value would be x/ρ. *The punchline to all this is that, by calculating the*

risk-adjusted interest rate, we can simply repeat all the analysis we did in the previous subsections using ρ rather than r.

A couple of further points. One, observe that the probability of the game lasting forever is zero. To see this, recall that the probability of it lasting t periods into the future is β^t. Because $\beta < 1$, β^t is shrinking in t. Indeed, as t goes towards infinity, β^t is going to zero; that is, there is zero probability of the game lasting forever. For example, suppose $\beta = .9$. Then the probability of the game lasting 10 periods is .35; the probability of it lasting 100 periods is .000027; and the probability of it lasting 1000 periods is 1.74×10^{-46}. That last probability is on the order of the probability that if we picked one atom randomly from all the atoms on Earth that atom would just happen to be one of the atoms making up your nose; in other words, essentially zero.

The second point is to recall that if the interest rate is too high, then it is not feasible to sustain cooperative play. From expression (5.9), observe that as β gets smaller—that is, the probability of the game surviving goes down—then ρ gets bigger. In other words, the smaller the probability of the game surviving, the harder it is to sustain cooperative play. This helps explain why price competition can be particularly fierce in dying industries: Because the probability of the game surviving far into the future is so low, the risk-adjusted interest rate is extremely high (the future is heavily discounted), and, thus, cooperation is hard to sustain.

Repeated Play in Everyday Life

There are many reasons people cooperate or behave nicely, but one of the more important reasons is that they know they are playing a repeated game with an unknown end date. The importance of this should not be underestimated. Some everyday examples:

Paying bills. Firms are, of course, legally required to pay their bills. But they have considerable discretion in how promptly they pay. Hence, in real life the following often occurs. One firm—call it A—owes another firm—call it B, but A knows B may soon go out of business. Hence, the importance A places on maintaining a good relation with B is reduced (*i.e.*, in the terms of the previous analysis, ρ is high). Hence, A will tend to delay paying its bills or make B press it to pay; that is, A will cease cooperating with B.

The lame-duck problem. Your coworkers' incentives to cooperate with you are greater when they think they will be interacting with you in the future than when they know they won't. Studies find, therefore, that people have a harder time on the job once they've announced they're leaving soon. In particular, managers are often advised not to give their subordinates too much advanced notice of their departure; subordinates work hard for a manager, in part, because they anticipate future interactions.

A tough reputation. Sometimes repeated play allows you to develop a reputation for being tough in the following sense. The possibility of future interactions makes it credible that you'll carry out actions in the stage game that wouldn't actually be in your interest were the stage game played just once, but which, if others thought you would carry them out, would make others play in a manner more to your liking. For example, certain countries try to develop a reputation for never negotiating with hostage takers (*e.g.*, terrorists). If there were only, ever, going to be one hostage-taking situation, then it would be better to negotiate than not. But knowing that, terrorists will be tempted to seize hostages to achieve their objectives through subsequent negotiations. On the other hand, if potential hostage-taking situations are likely in the future, then by not negotiating today, you discourage *future* would-be hostage takers from seizing hostages. In business, firms may seek to develop a reputation for being tough competitors to deter potential entrants from coming into their markets; it might not make sense to beat up on an entrant if the stage game were played once, but if it deters future entrants, then it could be a sensible strategy.

Summary | 5.5

This chapter introduced game theory as a tool to assess strategic situations. Although we have only scratched the surface of what game theory can offer in terms of understanding strategic interactions, we nevertheless saw that it could help us predict how others will play and how we should play in response.

We analyzed two games in depth, the Prisoners' Dilemma (the advertising game of Figure 5.1) and the Bertrand Model. The first of the two introduced the concept of dominant strategy. We also saw in that game that strategic interactions could yield outcomes unfavorable to both players.

To properly analyze the Bertrand Model, a pricing game, we needed to introduce the concept of Nash equilibrium; that is, players playing mutual best responses. Like the Prisoners' Dilemma, the outcome of the Bertrand Model was unfavorable to the players. For this reason, we dubbed it the Bertrand Trap.

We also noted that there were ways firms could avoid the Bertrand Trap: (i) differentiate their products; (ii) make it difficult for consumers to learn prices; (iii) raise switching costs; (iv) limit capacity; (v) obtain a cost advantage; or (vi) exploit their repeated interactions.

The last of these, repeated interactions, we showed was a powerful concept and could promote cooperative play in otherwise non-cooperative situations.

Key Concepts

- Elements of a game (players, actions, & payoffs)
- Dominant strategy

- Prisoners' Dilemma
- Best response
- Nash equilibrium
- Dominated strategy
- Bertrand model
- Avoiding the Bertrand trap
- Repeated games
- Stage game
- Infinite versus finite repetition (unknown versus known end of game)

Mathematical Appendices

Algebra Review | A1

This appendix reviews those aspects of algebra useful for successful completion of a course in managerial economics.

Functions | A1.1

We begin with a quick review of functions.

Definition. *A function takes each element of one set of numbers and maps it to a unique number in a second set of numbers.*

For example, the function $f(x) = 5x + 3$ takes real numbers and maps them into real numbers. For instance, $f(1) = 8$; that is, the function maps 1 into 8. Likewise $f(3) = 18$. Note that for any x, $f(x)$ is unique (*e.g.*, there is only one value for $f(4)$, namely 23).

The set of numbers that can serve as an argument for a function—that is, the set of numbers for which the function is defined—is known as the *domain* of the function. The set of numbers to which this set is mapped is known as the *range* of the function. For example, because square root is defined for non-negative numbers only, the function $g(x) = \sqrt{x}$ has a domain equal to $\{x|0 \leq x < \infty\}$ (*i.e.*, the set of numbers between 0 and infinity). Because, as discussed below, $\sqrt{x} \geq 0$ for all x in the domain of square root, we see that its range is also the set of non-negative numbers. As another example, the function $h(x) = |x|$ has a domain equal to all real numbers and range consisting of all non-negative numbers.[1]

As a convention, $f(\cdot)$ denotes the function and $f(x)$ denotes the particular value of the function evaluated at x.

Although $f(x)$ is a unique value, there is no guarantee that there is a unique x that solves the equation $f(x) = y$ because more than one element of the domain can map to a single element of the range. For instance, $|5| = |-5| = 5$. So, if we define $h(x) = |x|$, then there is no single x that solves the equation $h(x) = 5$. In some cases, however, there is always a unique x that solves the equation $f(x) = y$. For instance, if $f(x) = 5x + 3$, then there is only one x for each y, namely $x = (y3)/5$. When there is always a unique x that solves $f(x) = y$ for all y, we say that $f(\cdot)$ is an *invertible function*. Sometimes we just

[1] Recall that $|x|$ denotes the absolute value of x. That is, $|x| = x$ if $x \geq 0$ and $|x| = -x$ if $x < 0$ (*e.g.*, $|5| = 5$ and $|-4| = 4$).

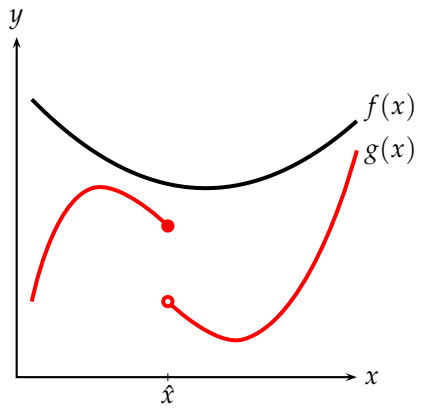

Figure A1.1: The function $f(\cdot)$ (in black) is continuous. The function $g(\cdot)$ (in red) is not—it has a jump (point of discontinuity) at $\hat{x}$.

say its "invertible." The standard notation for the inverse mapping is $f^{-1}(\cdot)$. So, for example, if $f(x) = 5x + 3$, then $f^{-1}(y) = (y3)/5$. Observe that the following rule holds for functions and their inverses:

Rule 1. *For an invertible function* $f(\cdot)$, $f\left(f^{-1}(y)\right) = y$ *and* $f^{-1}\left(f(x)\right) = x$.

A function, $f(\cdot)$, is *increasing* if, for all x_0 and x_1 in its domain, $x_1 > x_0$ implies $f(x_1) > f(x_0)$. A function is *decreasing* if, for all x_0 and x_1 in its domain, $x_1 > x_0$ implies $f(x_1) < f(x_0)$. Following the graph of a function from left to right, it is increasing if its graph goes up. It is decreasing if its graph goes down. The term *monotonic* will be used to refer to a function that is either an increasing function or a decreasing function. Observe that a monotonic function must be invertible.

Rule 2. *Monotonic functions are invertible.*

A function is *continuous* if you can draw its graph without picking up your pen or pencil. See Figure A1.1. Technically, for any $\hat{x}$ in the domain of the function $f(\cdot)$ if we fix a $\varepsilon > 0$ (but as small as we like), then there exists a $\delta > 0$ such that

$$|f(x) - f(\hat{x})| < \varepsilon \text{ if } |x - \hat{x}| < \delta\,.$$

Basically, this says that the graph can't jump at $\hat{x}$

Exponents | A1.2

Let n be a whole number (*i.e.*, $n = 0$ or 1 or 2 ...).[2] Let x be a real number. The expression x^n is shorthand for multiplying x n-times by itself; that is,

$$x^n = \underbrace{x \times \cdots \times x}_{n \text{ times}} .$$

Note the convention that $x^0 = 1$ for all x. The n in x^n is the *exponent*. Sometimes x^n is described in words as "taking x to the nth power."

Addition and multiplication of exponents

One rule of exponents is

Rule 3. $x^n \times x^m = x^{n+m}$.

For example, $2^2 \times 2^3 = 2^{2+3} = 2^5$. (As is readily verified: $2^2 = 4$, $2^3 = 8$, $4 \times 8 = 32$,and $2^5 = 32$.) Rule 3 is readily verified:

$$\begin{aligned} x^n \times x^m &= \underbrace{x \times \cdots \times x}_{n \text{ times}} \times \underbrace{x \times \cdots \times x}_{m \text{ times}} \\ &= \underbrace{x \times \cdots \times x}_{n+m \text{ times}} \\ &= x^{n+m} . \end{aligned}$$

A second rule of exponents is

Rule 4. $(x^n)^m = x^{n \times m} = x^{nm}$.

Observe the convention that the multiplication of two variables (*e.g.*, n and m) can be expressed as nm. As an example, $(2^2)^3 = 2^6$. (As is readily seen:

[2] Actually, one does not need to limit n to the whole numbers; it is certainly valid to let n be any real number. That is, for instance, $x^{1/2}$ is a valid expression (it is the same as $\sqrt{x}$). Even x^π is a valid expression. When, however, n is not a whole number, the interpretation of x^n is somewhat more involved. It is worth noting, however, that all the rules set forth in this section apply even if n is not a whole number.

$2^2 = 4$, $4^3 = 64$, and $2^6 = 64$.) Rule 4 is readily verified:

$$\begin{aligned}(x^n)^m &= \underbrace{x^n \times \cdots \times x^n}_{m \text{ times}} \\ &= \underbrace{\underbrace{x \times \cdots \times x}_{n \text{ times}} \times \cdots \times \underbrace{x \times \cdots \times x}_{n \text{ times}}}_{m \text{ times}} \\ &= \underbrace{x \times \cdots \times x}_{n \times m \text{ times}} \\ &= x^{nm}.\end{aligned}$$

The rules for adding and multiplying exponents obey the usual rules of arithmetic; hence, for instance:

Rule 5. $x^{n+m} = x^n x^m = x^m x^n = x^{m+n}$

Rule 6. $x^{nm} = x^{mn}$, *so* $(x^n)^m = (x^m)^n$.

Rule 7. $(xy)^n = x^n y^n$.

Rule 8. $(x^n y^m)^p = x^{np} y^{mp}$.

Rule 9. $x^{p(n+m)} = (x^{n+m})^p = (x^n x^m)^p = x^{np} x^{mp} = x^{np+mp}$.

Rule 10 (Rule of the common exponent I)**.** $z^p x^n y^n = z^p (xy)^n$.

Rule 11 (Rule of the common exponent II)**.** $z^{p+n} x^n = z^p (zx)^n$.

Rule 12 (Rule of the common exponent III)**.** $z^{pn} x^n = (z^p x)^n$.

Errors to be avoided

The following are common *mistakes* to be avoided:

No no 1. $(x+y)^n \neq x^n + y^n$*; that is, exponentiation is* not *distributive over addition. For example,* $(2+3)^2 = 5^2 = 25$*, whereas* $2^2 + 3^2 = 4 + 9 = 13$.

No no 2. $x^{pn} y^n \neq x^p (xy)^n$. *For example,* $2^6 \times 3^2 = 2^{3\times 2} 3^2 = 64 \times 9 = 576$*, whereas* $2^3 (2 \times 3)^2 = 8 \times 36 = 288$. *(Also recall Rule 12.)*

Negative exponents

The notation x^{-n} means $(1/x)^n$ (obviously, $x \neq 0$). Observe that, because

$$x^n x^{-n} = x^n \left(\frac{1}{x}\right)^n = \left(x \times \frac{1}{x}\right)^n = 1^n = 1 = x^0$$

(where the third inequality follows from Rule 8), we have that x^{-n} is the multiplicative inverse of x^n. It is also true that

$$x^n \times \frac{1}{x^n} = 1\,.$$

Because multiplicative inverses are unique, this means

Rule 13. $(1/x)^n = 1/x^n$.

All of the rules considered so far for exponents remain true with negative exponents (as long as the numbers being raised to the negative exponents are not zero).

It is readily seen that

Rule 14. $(x/y)^n = (y/x)^{-n}$.

Observe

$$(x+y)^{-n} = \frac{1}{(x+y)^n}$$

not

$$\frac{1}{x^n} + \frac{1}{y^n}\,;$$

that is,

No no 3. *Observe*

$$(x+y)^{-n} \neq \frac{1}{x^n} + \frac{1}{y^n}\,.$$

For example, $(2+3)^{-2} = 5^{-2} = 1/25$, *whereas* $1/2^2 + 1/3^2 = 1/4 + 1/9 = 13/36$.

Square Roots | A1.3

The positive square root of a number $x \geq 0$ is denoted $\sqrt{x}$; that is, $\sqrt{x}$ is the positive number such that $\sqrt{x} \times \sqrt{x} = x$. Observe the convention that $\sqrt{x} \geq 0$ (and equal to zero only if $x = 0$).

Because the product of two negative numbers is positive (*e.g.*, $-2 \times (-2) = 4$), each number $x \geq 0$ also has a negative square root. It is denoted $-\sqrt{x}$. Hence, as an example, the positive square root of 3 is denoted $\sqrt{3}$ and the negative square root of 3 is denotes $-\sqrt{3}$.

Rules for square roots

By definition

Rule 15. $\left(\sqrt{x}\right)^2 = x$ *and* $\sqrt{x^2} = |x|$. *That is, squaring and taking square roots are inverse operations.*

Also by definition, the square root of xy is $\sqrt{xy}$. Observe, too,

$$\left(\sqrt{x}\sqrt{y}\right)^2 = \left(\sqrt{x}\sqrt{y}\right)\left(\sqrt{x}\sqrt{y}\right) = \sqrt{x}\sqrt{x}\sqrt{y}\sqrt{y} = xy\,,$$

where the second equality follows because multiplication is associative (*i.e.*, we are free to rearrange the order of the terms). But this says, then, that $\sqrt{x}\sqrt{y}$ is also the square root of xy; we may conclude:

Rule 16. $\sqrt{xy} = \sqrt{x}\sqrt{y}$.

This rule is useful for simplifying expressions. For instance, $\sqrt{8} = \sqrt{4\times 2} = \sqrt{4}\sqrt{2} = 2\sqrt{2}$.

The following is a common mistake to avoid:

No no 4. $\sqrt{x+y} \neq \sqrt{x} + \sqrt{y}$*; that is, taking square roots is* not *distributive over addition. For example,* $\sqrt{9+16} = \sqrt{25} = 5$*, whereas* $\sqrt{9} + \sqrt{16} = 3 + 4 = 7$.

Recall that the sum of the lengths of any two sides of a triangle must exceed the length of the remaining side. For this reason, this last "no no" is sometimes called the triangle inequality (can you see why?) and stated as

Theorem 1 (Triangle Inequality)**.** $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ *and equal only if x or y or both is (are) zero.*

Equations of the Form $ax^2 + bx + c = 0$ | A1.4

If a polynomial equation of the form $ax^2 + bx + c = 0$, a, b, and c constants and x variable, has a solution, then the solution or solutions are given by the *quadratic formula*:

Theorem 2 (Quadratic formula)**.** *Let a, b, and c be constants and x an unknown. If*

$$ax^2 + bx + c = 0$$

has a solution in x, then its solution or solutions are given by the formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}\,. \tag{A1.1}$$

Some quick observations:

1. Because one cannot take the square root of a negative number, we see that $ax^2 + bx + c = 0$ can only have solutions if $b^2 \geq 4ac$.
2. The equation $ax^2 + bx + c = 0$ will have one solution only if $b^2 = 4ac$. If $b^2 > 4ac$, then it will have two solutions.

Examples:

- $x^2 + 5x + 4 = 0$. $b^2 = 25 > 16 = 4ac$, so two solutions exist:

$$\frac{-5 \pm \sqrt{25 - 16}}{2} = \frac{-5 \pm 3}{2} = -1 \text{ or } -4\,.$$

- $x^2 + 2x + 1 = 0$. Observe $b^2 = 4 = 4 = 4ac$, so there is only one solution: $x = -b/(2a) = -2/2 = -1$.

Lines | A1.5

A *line* has the formula $y = mx + b$, where m and b are constants and x and y are variables. This way of writing a line is called *slope-intercept* because m is the *slope* of the line and b is the *intercept* (the point on the y axis at which the line intersects the y axis). If $m > 0$, the line slopes up (*i.e.*, rises as its graph is viewed from left to right). If $m < 0$, the line slopes down (*i.e.*, falls as its graph is viewed from left to right).

Consider an expression of the form

$$Ax + By = C\,,$$

where A, $B \neq 0$, and C are constants. This, too, is an expression for a line, which we can see by rearranging:

$$By = -Ax + C\,; \text{ hence, } y = -\frac{A}{B}x + \frac{C}{B}\,.$$

To summarize:

Rule 17. *An expression of the form $Ax + By = C$ is equivalent to the line $y = mx + b$, where $m = -A/B$ and $b = C/B$.*

The line through two points

The line through two distinct points (x_1, y_1) and (x_2, y_2) is the line satisfying the equations:

$$y_1 = mx_1 + b \text{ and } y_2 = mx_2 + b\,.$$

Notice that if we subtract the first equation from the second, we get:

$$y_2 - y_1 = mx_2 - mx_1 = m(x_2 - x_1)\,,$$

from which it follows that

$$m = \frac{y_2 - y_1}{x_2 - x_1}\,.$$

Substituting this back into the second equation, we get

$$y_2 = \frac{y_2 - y_1}{x_2 - x_1}x_2 + b\,.$$

Hence,

$$b = y_2 - \frac{y_2 - y_1}{x_2 - x_1} x_2 = \frac{y_2(x_2 - x_1)}{x_2 - x_1} - \frac{x_2(y_2 - y_1)}{x_2 - x_1} = \boxed{\frac{x_2 y_1 - x_1 y_2}{x_2 - x_1}}.$$

To conclude:

Rule 18. *The line through the points (x_1, y_1) and (x_2, y_2) has*

$$\textit{slope} = \frac{y_2 - y_1}{x_2 - x_1} \textit{ and intercept} = \frac{x_2 y_1 - x_1 y_2}{x_2 - x_1}.$$

One might worry, in the last rule, about what happens if $x_1 = x_2$. In this case the line is completely vertical (assuming the usual orientation of having y on the vertical axis).

Parallel lines[3]

If $y = mx + b$ is a line and (x_1, y_1) is a point *not* on that line, then we can find a line parallel to $y = mx + b$ that passes through (x_1, y_1). Call this parallel line $y = Mx + B$. A parallel line has the same slope, so $M = m$. This permits us to find B by subtraction:

$$B = y_1 - mx_1.$$

Logarithms | A1.6

Often we need to solve equations of the form $A^x = B$, where A and B are constants and x is the unknown to be solved for. This section explores, *inter alia*, how such equations are solved.

The natural logarithm

For reasons that we won't pursue here, the constant e, which I will define in a moment, arises a lot in mathematics. This constant is defined to be[4]

$$e = \sum_{t=0}^{\infty} \frac{1}{t!}, \tag{A1.2}$$

where $t!$—read t *factorial*—is defined as

$$t! = \begin{cases} 1, \text{ if } t = 0 \\ 1 \times \cdots \times t, \text{ if } t = 1, 2, \ldots \end{cases}.$$

[3] The math concept, not the album by Blondie.

[4] An alternative—but equivalent—definition is

$$e = \lim n \to \infty \left(1 + \frac{1}{n}\right)^n.$$

This is why the formula for continuous compound interest is expressed in terms of e.

Calculations show that $e \approx 2.718281828$. Because e is an irrational number, no exact decimal representation exists for it.

The number e is called the *base of the natural logarithm*. Define the function $\exp(x) = e^x$.

Observe that $\lim_{x\to-\infty} e^x = 0$ and $\lim_{x\to\infty} e^x = \infty$. Moreover, it can be shown that e^x is increasing and continuous. Hence, for every $y > 0$, there exists a unique x such that $e^x = y$.

The function $\ln(\cdot)$

The function $\ln(\cdot)$ is defined as the inverse of $\exp(\cdot)$. That is,

$$\ln\big(\exp(x)\big) = \ln\left(e^x\right) = x\,.$$

The function $\ln(\cdot)$ is called the *logarithm* or *log*.[5] Because $\exp(x)$ is a positive number for all x, it follows that the function $\ln(\cdot)$ is defined for positive numbers only.

Recall that, for every $y > 0$, there exists a unique x such that $e^x = y$. Taking logs of both sides, we see that $x = \ln(y)$; that is, in this case, x is the log of y.

Proposition 27. $\ln(xy) = \ln(x) + \ln(y)$.

Proof: Recall there exists a unique number a and a unique number b such that $x = e^a$ and $y = e^b$. We thus have

$$\begin{aligned}
\ln(xy) &= \ln(e^a e^b) \\
&= \ln(e^{a+b}) && \text{(Rule 3)} \\
&= a + b && (\ln(\cdot) \text{ is the inverse of } \exp(\cdot)) \\
&= \ln(x) + \ln(y)\,.
\end{aligned}$$

■

Proposition 28. $\ln(x/y) = \ln(x) - \ln(y)$.

Proof: Define a and b as in the previous proof. Observe

$$\begin{aligned}
\ln(x/y) &= \ln\left(\frac{e^a}{e^b}\right) \\
&= \ln(e^a e^{-b}) && \text{(Def'n of negative exponent)} \\
&= \ln(e^{a-b}) && \text{(Rule 3)} \\
&= a - b && (\ln(\cdot) \text{ is the inverse of } \exp(\cdot)) \\
&= \ln(x) - \ln(y)\,.
\end{aligned}$$

■

[5] To be technical, I should say the log with respect to e (the base).

Proposition 29. $\ln(x^\gamma) = \gamma \ln(x)$.

Proof: Define a as in the previous proof. Observe

$$\begin{aligned} \ln(x^\gamma) &= \ln\left((e^a)^\gamma\right) \\ &= \ln(e^{a\gamma}) && \text{(Rule 4)} \\ &= a\gamma && (\ln(\cdot) \text{ is the inverse of } \exp(\cdot)) \\ &= \gamma \ln(x)\,. \end{aligned}$$

■

Solving equations with exponents

Consider the motivating example, namely that we wish to solve the equation $A^x = B$ for x. Using Proposition 29, this is equivalent to solving

$$x \ln(A) = \ln(B)\,.$$

Dividing both sides by $\ln(A)$, we find

$$x = \frac{\ln(B)}{\ln(A)}\,. \tag{A1.3}$$

System of Equations | A2

Often in economics we need to know the solution to a system of equations. For instance, we could have that demand is given by

$$D = \alpha - \beta p$$

and supply is given by

$$S = \eta + \gamma p\,,$$

where D is the quantity demanded as function of price, p, S is the quantity supplied as a function of price, and the Greek letters are non-negative constants. If the market is in equilibrium, then demand and supply must be equal; that is, the price must be such that $D = S$. To solve for that price, we need to solve the system of equations above.

A Linear Equation in One Unknown | A2.1

Recall that if x is an unknown variable and A, B, and C constants ($A \neq 0$), then the equation

$$Ax + B = C$$

is solved by the following procedure:

1. Subtract B from both sides of the equation (equivalently, add $-B$ to both sides). This yields
$$Ax = C - B\,.$$
2. Divide both sides of this last equation by A (equivalently, multiply both sides by $1/A$). This yields the solution,
$$x = \frac{C - B}{A}\,.$$

Observe this method be used even if, rather than x, we have some invertible function of x, $f(x)$. That is, consider the equation

$$Af(x) + B = C\,.$$

Employing the two steps just given yields

$$f(x) = \frac{C - B}{A}\,.$$

Letting $f^{-1}(\cdot)$ denote the inverse function, we can invert both sides to obtain

$$x = f^{-1}\left(\frac{C-B}{A}\right).$$

A2.2 Two Linear Equations in Two Unknowns

Consider the system

$$Ax + By = C \text{ and} \tag{A2.1}$$
$$Dx + Ey = F, \tag{A2.2}$$

where A–F are constants and x and y are the two unknowns that we wish to solve for. Recall, from Section A1.5, that expressions like (A2.1) and (A2.2) are formulæ for lines. Hence, solving this system is equivalent to finding the point, (x, y), at which those two lines cross.

Because of this equivalence, it follows that, if the lines dont cross—are parallel—then no solution exists. Recall two lines are parallel if they have the same slope. From Section A1.5, the slopes of these two lines are A/B and D/E, respectively (recall Rule 17). Hence, if $A/B = D/E$, there is no solution.[1]

Assume $A/B \neq D/E$. A solution can, then, be found by the *method of substitution*, which has the following steps:

1. Write the first equation, expression (A2.1) in slope-intercept form:

$$y = -\frac{A}{B}x + \frac{C}{B}. \tag{A2.3}$$

2. Substitute the righthand side of this last equation, expression (A2.3), for y in the second equation of the system (*i.e.*, for y in (A2.2)):

$$Dx + E\left(-\frac{A}{B}x + \frac{C}{B}\right) = F.$$

3. Combine terms in x:

$$\left(D - \frac{EA}{B}\right)x = F - \frac{EC}{B};$$

or, simplifying,

$$\frac{BD - EA}{B}x = \frac{BF - EC}{B}.$$

[1] To be technical, if it is also true that $C/B = F/E$, so the two lines have the same intercept, then they are the same line. In which case, one could say that there are an infinite number of solutions, because all points on this common line satisfy the two equations.

4. Solve this last expression for x:

$$x = \frac{BF - EC}{BD - EA}. \tag{A2.4}$$

5. Substitute for the x in expression (A2.3) using the righthand side of expression (A2.4):

$$y = -\frac{A}{B} \times \frac{BF - EC}{BD - EA} + \frac{C}{B};$$

simplifying,

$$\begin{aligned} y &= \frac{-A(BF - EC) + C(BD - EA)}{B(BD - EA)} \\ &= \frac{CD - AF}{BD - EA}. \end{aligned}$$

For example, recall the demand and supply equations with which we began this appendix:

$$\begin{aligned} Q &= \alpha - \beta p \text{ and} \\ Q &= \eta + \gamma p, \end{aligned}$$

where Q denotes the equilibrium quantity traded (recall, in equilibrium, demand and supply are equal). The first equation is already in slope-intercept form. Substituting into the second yields:

$$\alpha - \beta p = \eta + \gamma p.$$

Solving for p:

$$p = \frac{\alpha - \eta}{\beta + \gamma}.$$

(Note, because prices must be non-negative, this tells us that there can be an equilibrium only if $\alpha \geq \eta$.) Substituting the equilibrium price back into demand, we find:

$$Q = \frac{\alpha\gamma + \beta\eta}{\beta + \gamma}.$$

Calculus | A3

It is beyond a simple appendix to teach calculus fully. Here we seek to review the main points. The intention is that it be understandable by anyone with a good mastery of high school mathematics.

Calculus is divided into two parts. One, known as *differential calculus*, is concerned with the rates at which things change. For instance, marginal cost, *MC*, is the rate at which total cost changes as we increase production.

The second part, known as *integral calculus*, is concerned with areas under curves. For instance, the area under the demand curve from 0 to x units is the total benefit consumers derive from the x units.[1]

The Derivative | A3.1

Think about speed (you may wish to read, first, the discussion on pages 37–38). The speed at which you are traveling at a moment in time, t, is best estimated by

$$\text{speed} = \frac{D(t+h) - D(t)}{h},$$

where $D(\cdot)$ is your distance traveled as a function of the time spent traveling and h is some very small increment of time (*e.g.*, a small fraction of an hour, such as a second). Indeed, the smaller we can make h, the better our estimate will be. In fact, the ideal estimate is

$$\text{speed} = \lim_{h \to 0} \frac{D(t+h) - D(t)}{h},$$

where "lim" means the limit of that ratio as h approaches zero. For instance, if $D(t) = S \times t$, then

$$\frac{D(t+h) - D(t)}{h} = \frac{S \times (t+h) - S \times t}{h} = \frac{S \times h}{h} = S.$$

Clearly, the last expression doesnt depend on h, so the limit of it as h goes to zero is S.

[1] As a technical note, there are some qualifications to this statement that I am skipping over.

Definition. *The* derivative *of a function $f(\cdot)$ at x is*

$$\lim_{h\to 0} \frac{f(x+h)-f(x)}{h}\,, \tag{A3.1}$$

assuming that limit exists.

For our purposes in this text, we generally will assume that the limit exists.

Definition. *A function is said to be* differentiable *if a derivative exists at every point in its domain.*[2]

Again, we will typically limit our attention to differentiable functions in this text.

We denote derivatives in a number of ways. Specifically, the derivative of $f(x)$ can be denoted as $f'(x)$ (read "f prime of x") or as $df(x)/dx$. If we have previously stated that $y = f(x)$, then we can also denote the derivative of $f(\cdot)$ as dy/dx. For some functions, such as cost functions, $C(\cdot)$, and revenue functions, $R(\cdot)$, we use a prefix "M," for *marginal*, to denote the derivative. That is, for example, the derivative of $C(\cdot)$ is $MC(\cdot)$.

Properties of derivatives

Proposition 30. *If $f(x) = K$, K a constant, for all x, then $f'(x) = 0$ for all x.*

Proof: OPT The numerator of the fraction in expression (A3.1) is always zero. ■

Proposition 31. *If*

$$b(x) = \alpha f(x) + \beta g(x)\,,$$

where α and β are constants and $f(\cdot)$ and $g(\cdot)$ are differentiable, then

$$b'(x) = \alpha f'(x) + \beta g'(x)\,.$$

Proof: OPT Observe expression (A3.1) can be written as

$$\alpha\frac{f(x+h)-f(x)}{h} + \beta\frac{g(x+h)-g(x)}{h}\,.$$

Now take limits. ■

Proposition 32 (Product rule)**.** *If $b(x) = f(x)g(x)$, $f(\cdot)$ and $g(\cdot)$ differentiable, then*

$$b'(x) = f'(x)g(x) + f(x)g'(x)\,.$$

[2]The domain of a function, recall, is the set of values for which the function is defined.

Proof: OPT Observe

$$\begin{aligned}\frac{f(x+h)g(x+h)-f(x)g(x)}{h} &= \frac{f(x+h)g(x+h)-f(x)g(x)}{h} \\ &\quad + \underbrace{\frac{f(x+h)g(x)-f(x+h)g(x)}{h}}_{+0} \\ &= g(x)\frac{f(x+h)-f(x)}{h} + f(x+h)\frac{g(x+h)-g(x)}{h}\,.\end{aligned}$$

Now take limits (note $\lim_{h\to 0} f(x+h) = f(x)$). ■

Proposition 33 (Chain rule)**.** *If $b(x) = f\big(g(x)\big)$, $f(\cdot)$ and $g(\cdot)$ differentiable, then $b'(x) = f'\big(g(x)\big)g'(x)$.*

Proof: OPT Observe

$$\begin{aligned}\frac{f\big(g(x+h)\big)-f\big(g(x)\big)}{h} &= \frac{f\big(g(x+h)\big)-f\big(g(x)\big)}{g(x+h)-g(x)} \times \frac{g(x+h)-g(x)}{h} \\ &= \frac{f\big(g(x)+\eta\big)-f\big(g(x)\big)}{\eta} \times \frac{g(x+h)-g(x)}{h}\,,\end{aligned}$$

where $\eta = g(x+h)-g(x)$, so $g(x+h) = g(x)+\eta$. Observe that $\lim_{h\to 0}\eta = 0$; hence, the limit of the righthand-side of the last expression is $f'\big(g(x)\big)g'(x)$, as claimed.[3] ■

Note the chain rule can be applied repeatedly.

The following two results are stated without proof.[4]

Theorem 3.

$$\lim_{h\to 0}\frac{e^h-1}{h} = 1\,.$$

Theorem 4.

$$\lim_{\varepsilon\to 0}\frac{\ln(1+\varepsilon)}{\varepsilon} = 1\,.$$

Proposition 34. *The derivative of e^x with respect to x is e^x (i.e., $de^x/dx = e^x$).*

Proof: OPT Observe

$$\frac{e^{x+h}-e^x}{h} = e^x\frac{e^h-1}{h}\,,$$

[3] In case you were wondering, yes, for continuous functions, it is true that $\lim_{x\to z} f(x)g(x) = f(z)g(z)$.

[4] The curious can consult any decent calculus text. A specific citation is §2.6 of D.W. Jordan and P. Smith's *Mathematical Techniques*, 2nd ed., Oxford: Oxford University Press, 1997

where I have used Rule 3 to factor out e^x. The result then follows from Theorem 3. ■

Proposition 35. *The derivative of* $\ln(x)$ *with respect to* x *is* $1/x$ *(i.e.,* $d\ln(x)/dx = 1/x$*).*

Proof: OPT Observe

$$\begin{aligned}
\frac{\ln(x+h)-\ln(x)}{h} &= \frac{1}{h}\ln\left(\frac{x+h}{h}\right) && \text{(Proposition 28)}\\
&= \frac{1}{h}\ln\left(1+\frac{h}{x}\right)\\
&= \frac{1}{x\varepsilon}\ln(1+\varepsilon) && \text{(making the substitution } h = x\varepsilon\text{)}\\
&= \frac{1}{x}\times\frac{\ln(1+\varepsilon)}{\varepsilon}\,.
\end{aligned}$$

Because $\varepsilon = h/x$, it goes to zero as h goes to zero. Hence, the result follows from Theorem 4. ■

Proposition 36. *The derivative of* x^z *with respect to* x *is* zx^{z-1} *(i.e.,* $dx^z/dx = zx^{z-1}$*).*

Proof: OPT Let $f(y) = e^y$ and $g(w) = z\ln(w)$. Using Proposition 29, we have

$$\ln(x^z) = z\ln(x) = g(x)\,.$$

Because $\exp(\cdot)$ and $\ln(\cdot)$ are inverse functions, we have

$$x^z = \exp\big(\ln(x^z)\big) = e^{g(x)} = f\big(g(x)\big)\,.$$

The chain rule tells us that $dx^z/dx = f'\big(g(x)\big)g'(x)$; hence,

$$\frac{d}{dx}x^z = e^{g(x)}\frac{z}{x} = x^z\frac{z}{x} = zx^{z-1}\,.$$

Note the first equality follows from Propositions 34 and 35, the second because log and exp are inverse functions, and the last from Rule 3. ■

Probability Appendices

Fundamentals of Probability | B1

This appendix reviews a number of fundamental concepts related to probability.

A *trial* (alternatively, experiment or observation) is a situation in which one outcome will occur out of a set of possible outcomes. For example, a flip of a coin is a trial and its possible outcomes are heads and tails. Other examples are:

- A toss of a die is a trial and its possible outcomes are the numbers one through six.
- A toss of a pair of dice is a trial and its possible outcomes are the 36 possible combinations of the two die faces.
- The price of a barrel of crude oil a year from now is a trial and its outcomes are all non-negative dollar amounts.
- Whether a given manufactured good is defective is a trial and its possible outcomes are "defective" and "not defective."
- The number of defective products in a production run of 1000 is a trial and its possible outcomes are the integers between 0 and 1000 (inclusive).

We are often interested in the *probability* that a given outcome of a trial will occur; that is, how likely that outcome is. For example, the probability of the outcome heads when a coin is flipped is 1/2; that is, we believe that it is equally likely the coin will land heads as it will tails. A belief, moreover, that is supported by our past experiences with flipping coins. Similarly, if ...

- ... the trial is the toss of a die, then the probability of getting a "four" is 1/6.
- ... the trial is the toss of a pair of dice, then the probability of getting a "three" on the first die and a "two" on the second is 1/36.
- ... the trial is the toss of a die-like cube, four sides of which are black and two of which are white, then the probability of getting a black side is 2/3 (= 4/6)

The greater the probability, the more likely that outcome is. Hence, for instance, in the last example the probability of getting a black side (2/3) is

greater than the probability of getting a white side (1/3), reflecting that a black side is twice as likely to appear as a white slide.

There are certain rules that probability over the outcomes of a trial must satisfy. To illustrate these rules, let N denote the number of possible outcomes in the trial, let $n = 1, 2, \ldots, N$ index the possible outcomes, let ω_n denote the nth outcome, and let $P\{\omega_n\}$ denote the probability of the nth outcome. Then the rules are

1. For each outcome n, $P\{\omega_n\} \geq 0$; and
2. $P\{\omega_1\} + P\{\omega_2\} + \cdots + P\{\omega_N\} = 1$. This can equivalently be expressed as
$$\sum_{n=1}^{N} P\{\omega_n\} = 1\,, \tag{B1.1}$$
where $\sum$ means sum and the expression is read as "the sum from $n = 1$ to N of $P\{\omega_n\}$."

The first rule is the requirement that probabilities be non-negative. The second rule is the requirement that if we add up the probabilities of all the possible outcomes, then this sum must equal one. If the probability of an outcome is zero, then its occurrence is impossible. If the probability of an outcome is one, then its occurrence is certain.

How outcomes get assigned probabilities is a tricky question, and one which has engendered much philosophical debate. For practically minded people, however, there are essentially three methods of assigning probabilities to outcomes. They are *logically*, *experimentally*, and *subjectively*.

Logic applies, for instance, in deciding that the probability of *heads* is 1/2: It seems clear that *heads* and *tails* are equally likely, so—since probabilities sum to one—$P\{\textit{heads}\} = 1/2$.

Experimentally refers to probabilities that are drawn from data. For instance, experimentally applies to how seismologists determine that there is a .63 probability of a "major" earthquake in the San Francisco Bay Area within the next thirty years:[1] The geological record reveals that major quakes tend to follow cycles. Currently, the Bay Area is in a stage of a cycle, which roughly two out of three times in the past has ended with a major quake within the next thirty years.

Subjectively is a kind way to refer to probabilities that are guesses. For example, the probability that a brand new technology will prove popular in the market place is not a number that is readily derived logically or from past experience. The executives of the company launching this new technology, however, must make guesses about the probability in order to make good decisions about the product's launch.

[1]The United States Geological Survey (USGS) website (http://earthquake.usgs.gov/regional/nca/ucerf/) as of January 22, 2011.

Events | B1.1

Often one is interested in combinations of outcomes known as *events*. An event is a collection—or *set*—of outcomes that satisfy some criteria. An event is said to occur if one of the outcomes that satisfies these criteria occurs. Suppose, for instance, the trial is the roll of a die, then an example of an event would be getting a number less than or equal to four. This event is the set of outcomes 1, 2, 3, and 4 and it occurs if a 1, 2, 3, or 4 is rolled. If the trial is the sexes of a couples first two children, then an event is "the first child is a girl," which is the set of outcomes (girl, girl) and (girl, boy). Another event for this trial is "the sexes of the two children are the same," which is the set of outcomes (girl, girl) and (boy, boy).

As a rule, I will denote events by capital italic letters (*e.g.*, A, B, etc.).

The probability of an event is the sum of the probabilities of the outcomes in that event. For example, the probability of rolling a number less than or equal to 4 is

$$\frac{2}{3} = P\{1\} + P\{2\} + P\{3\} + P\{4\}\,.$$

The probability of the first of a couple's two children being a girl is

$$\frac{1}{2} = P\{girl, girl\} + P\{girl, boy\}\,.$$

Formally, the probability of an event A can be expressed as

$$P(A) = \sum_{\omega \in A} P\{\omega\}\,, \tag{B1.2}$$

where "$\omega \in A$" means ω is one of the outcomes that make up A and the expression is read as "the sum over outcomes in A of the probabilities of those outcomes."

Often one is interested in comparing the probabilities of events. *Equality* and *contained in* are two relations between events commonly of interest. Two events, A and B, are equal—denoted $A = B$—if they contain the same outcomes. From the definition of the probability of an event, it follows that

Proposition 37. *If A and B are events and $A = B$, then $P(A) = P(B)$.*

Note the converse is *not* true; that is, the fact that two events have the same probability does not imply they are the same event.

One event, A, is *contained in* another event, B—denoted $A \subseteq B$—if every outcome in A is also in B.[2] For example, the event "both of a couple's two children are girls" is contained in the event "their first child is a girl." Or, for instance, the event "an even number is rolled" is contained in the event "a number greater than or equal to two is rolled." If $A \subseteq B$, every outcome in A is also in B, but B may contain outcomes that are not in A; hence, it follows that

[2] You may have previously seen the term "subset of" used for "contained in."

Proposition 38. *If A and B are events and $A \subseteq B$, then $P(A) \leq P(B)$.*

Again, the converse is *not* true; that is, $P(A) < P(B)$ does not imply that A is contained in B. Proposition 38 does, however, imply that if $P(A) > P(B)$, then A can*not* be contained in B.

Although Proposition 38 may seem obvious, the truth is that studies have shown that many people make mistakes in this area. For example, suppose I tell you that Linda grew up in a politically liberal household in San Francisco and that Linda herself was involved with a number of progressive causes while she was an undergraduate at Berkeley. Which of the following two statements is more likely to be true about Linda?

1. Linda is a bank teller; or
2. Linda is a feminist and a bank teller.

Most people choose statement 2, but that is incorrect:[3] the event Linda is a feminist and a bank teller is contained in the event Linda is a bank teller (feminist or not); or, to put it differently, all feminist bank tellers are bank tellers, but not all bank tellers are feminist bank tellers. So, from Proposition 38, we know that statement 1 is more likely to be true than statement 2.

Often we want to know what the probability is that an event will *not* happen. This is equivalent to asking what is the probability that one of the outcomes that do *not* make up the event will occur. If A is an event, let A^c be the event made up of the outcomes that do not make up A (A^c is called the *complement* of A). From (B1.1), we know that

$$\begin{aligned} 1 &= \sum_{n=1}^{N} P\{\omega_n\} \\ &= \underbrace{\sum_{\omega \in A} P\{\omega\}}_{\text{sum over events in } A} + \underbrace{\sum_{\omega \in A^c} P\{\omega\}}_{\text{sum over events } \textit{not} \text{ in } A} \\ &= P(A) + P(A^c)\,. \end{aligned}$$

This, in turn, establishes

Proposition 39. *The probability that an event A does* not *occur (equivalently, that event A^c* does *occur) is $1 - P(A)$*

For example, the probability that "at least one child is a boy in a pair of children" is most easily calculated by recognizing that this event is the complement of the event "both children are girls." The probability that "both children are girls" is $1/4$, so the probability that "at least one child is a boy" is $3/4$.

Sometimes we want to know the probability that two events will *both* occur. The event that both events occur is, itself, an event. If A and B are two events,

[3] This is known as the *conjunctive fallacy*.

then the event that they both occur is denoted $A \cap B$ (read "A *intersection* B"). An outcome is contained in $A \cap B$ if it is in both A and B. In other words, every outcome in $A \cap B$ is in A and it is in B. It follows that

$$A \cap B \subseteq A \text{ and } A \cap B \subseteq B\,.$$

From Proposition 38, we can therefore conclude that

Proposition 40. *If A and B are two events, then the probability of their* both *occurring, $P(A \cap B)$, does not exceed the probability that A occurs nor does it exceed the probability that B occurs (i.e., $P(A \cap B) \leq P(A)$ and $P(A \cap B) \leq P(B)$).*

That is, the probability of both events occurring cannot exceed the probability of one event occurring (and the other event either occurring or not occurring). This is another way to view the "Linda" problem above. One event is "Linda is a bank teller" and another is "Linda is a feminist." Statement 2—"Linda is a feminist bank teller"—is the intersection of these two events; hence, from Proposition 40, it follows that the probability of statement 2 cannot exceed the probability of statement 1.

Sometimes two events cannot possible happen together (*e.g.*, the event "the next person I meet is a man" and the event "the next person I meet is pregnant").[4] We denote this by writing $A \cap B = \emptyset$, where $\emptyset$–called the *null set*—denotes the "event" that no outcome occurs. Since some outcome must occur, the null set represents an impossible event.[5] Reflecting the idea that $\emptyset$ is an impossible event, we define $P(\emptyset) = 0$.

The probability of $A \cap B$ is calculated by summing up the probabilities of all outcomes common to A and B (*i.e.*, by employing formula (B1.2)—except substituting $A \cap B$ for A in that expression).

In some situations, we want to know the probability that either event A, event B, or both will occur. That is, we want to know the probability of an outcome occurring that is contained in either A, B, or both. This defines a new event, which we denote as $A \cup B$ (read "A *union* B"). For example, if the trial is the roll of a die, then the union of the events "an even number is rolled" and "a prime number is rolled"[6] is the event "a number larger than one is rolled," since the outcome one is the only outcome not in at least one of the two events (the even numbers are 2, 4, and 6, while the prime numbers are 2, 3, and 5). Note that if an outcome is common to both events it is only counted once in their union. For example, the union of "an even number is rolled" and "a prime number is rolled" is $\{2, 3, 4, 5, 6\}$ and not $\{2, 2, 3, 4, 5, 6\}$. Because every outcome in A must, by definition, be in $A \cup B$, it follows from Proposition 38 that

[4] Bad Arnold Schwarzenneger movies notwithstanding.

[5] The null set is sometimes referred to as the *empty set*.

[6] Recall a prime number is a whole number greater than one that is divisible only by one and itself (*e.g.*, 3 is prime—it is divisible by 1 and 3 only—whereas 4 is not prime—it is divisible by 2 in addition to being divisible by 1 and 4).

Proposition 41. *If A and B are two events, then the probability of one, the other, or both occurring, $P(A \cup B)$, is at least as great as the probability of either one alone occurring. That is, $P(A) \leq P(A \cup B)$ and $P(B) \leq P(A \cup B)$.*

That is, for example, the probability of rolling an even number cannot exceed the probability of rolling an even number, or a prime number, or both.

The event $A \cup B$ may contain outcomes that are common to both A and B. Since we don't double count such outcomes it follows that

Proposition 42. *If A and B are events, then $P(A \cup B) \leq P(A) + P(B)$.*

To see why Proposition 42 is true, observe

$$\begin{aligned} P(A \cup B) &= \sum_{\omega \in A \text{ only}} P\{\omega\} + \sum_{\omega \in A \cap B} P\{\omega\} + \sum_{\omega \in B \text{ only}} P\{\omega\} \\ &\leq \underbrace{\sum_{\omega \in A \text{ only}} P\{\omega\} + \sum_{\omega \in A \cap B} P\{\omega\}}_{P(A)} + \underbrace{\sum_{\omega \in A \cap B} P\{\omega\} + \sum_{\omega \in B \text{ only}} P\{\omega\}}_{P(B)} \\ &= P(A) + P(B)\,. \end{aligned}$$

From this last expression, we see that the difference between $P(A \cup B)$ and $P(A) + P(B)$ is that the latter counts $\sum_{\omega \in A \cap B} P\{\omega\} = P(A \cap B)$ twice. It follows, therefore, that we can get the correct value for $P(A \cup B)$ by subtracting $P(A \cap B)$ from $P(A) + P(B)$. We can, thus, conclude

Proposition 43. *If A and B are two events, then the probability of one, the other, or both occurring, $P(A \cup B)$, is given by the formula:*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)\,.$$

Two events are *mutually exclusive* if they cannot both occur; that is, if $A \cap B = \emptyset$. For example, the events "roll an even number" and "roll an odd number" are mutually exclusive (no number is both even and odd). Given that $P(\emptyset) = 0$, it follows from Proposition 43 that

Proposition 44. *If A and B are two* mutually exclusive *events, then $P(A \cup B) = P(A) + P(B)$.*

Conditional Probability | B2

Often we want to know the probability that one event will occur given that we know another event has occurred or will occur. For example, what is the probability that the next unit off the assembly line will be defective given that two of the last one hundred were defective? Or what is the probability that it will rain this afternoon given that it is cloudy this morning? These probabilities are called *conditional probabilities* and they are denoted $P(A|B)$—read "probability of A conditional on B"—where B is the event we know has occurred or will occur and A is the event in whose probability we are interested. For instance, suppose someone asked what is the probability that it will rain this afternoon given that (conditional on) its being cloudy this morning? We would denote this as

$$P(\text{rain in afternoon}|\text{cloudy in morning})\,.$$

To better understand conditional probability, let the trial be the roll of a die and let A be the event "a six is rolled"[1] and let B be the event "an even number is rolled." Suppose I secretly rolled the die and told you that I had rolled an even number (*i.e.,* B has occurred). What would you imagine the probability of my having rolled a six to be (*i.e.,* what is $P(A|B)$)? You would probably reason as follows: There are three even numbers—2, 4, and 6—each of which is equally likely; hence, the probability of a six having been rolled given that the number rolled is even is $1/3$. This reasoning is correct. Now suppose I asked you what the probability that I rolled a five is given that I have rolled an even number (*i.e.,* A is now "a five is rolled"). You would respond zero, because five is not even and, so, is an impossible event given that the number rolled is even. Finally, suppose I reversed the situation and told you that I had rolled a six and asked you what the probability is that I have rolled an even number (*i.e.,* A is now "an even number is rolled" and B is "a six is rolled"). You would respond one, because six is an even number.

You should be able to verify that each of the probabilities you calculated above satisfy the following rule:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}\,. \tag{B2.1}$$

[1] This is also an outcome. Note an event can contain a single outcome. Some authors refer to events made up of single outcomes as *simple events*.

This is, in fact, the formula for conditional probability (provided $P(B) > 0$).[2]

Although the examples of conditional probability given above were trivial, conditional probability is often harder than it seems. For example, suppose I asked you what is the probability that a couple with two children has at least one boy given that you know they have at least one girl. Most people's intuition is that the answer is $1/2$. This, however, is wrong. To see why, let A denote the event "at least one child is a boy" and let B denote "at least one child is a girl." The goal is to calculate $P(A|B)$. Writing out B in terms of the outcomes it contains, we see that it is (girl, girl), (boy, girl), and (girl, boy). Hence $P(B) = 3/4$. Similarly writing out A, we see that it is the outcomes (boy, boy), (girl, boy), and (boy, girl). Hence $A \cap B$ is the outcomes (girl, boy) and (boy, girl); hence $P(A \cap B) = 1/2$. Using (B2.1), we obtain $P(A|B) = 1/2 \div 3/4 = 2/3$; that is, the probability of at least one boy given that you know the couple has at least one girl is $2/3$. The reason most peoples intuition fails is that they incorrectly reason that, since one child is a girl, they wish to determine the probability that the other child is a boy. What they are forgetting is that there are three ways for a couple to have at least one girl, *two out of three of which* involve the couple also having one boy.

In Section B1.1, we saw that the *unconditional probability* (*i.e.*, given no additional information) of at least one of a couples two children being a boy was $3/4$. Given the information that at least one child is a girl, we saw that the *conditional probability* that they have at least one boy is $2/3$, which is less than $3/4$. What has happened is that, upon receiving the information that at least one child was a girl, we *revised* or *updated* our beliefs about the probability that the couple has at least one boy. In general, we say that a person is revising or updating his or her beliefs about an event when he or she uses new information to calculate a conditional probability. The probability of the event prior to receiving the information is called the *prior probability* and the probability of the event after receiving the information is called the *posterior probability*. Note that the posterior probability need not be lower than the prior probability: If, for instance, the event was the probability that a couple with two children has two boys, then you would update your beliefs upon learning that at least one of the children was a boy from $1/4$ (the prior probability) to $1/3$ (the posterior probability). As you will see, calculating revised or updated beliefs is one of the primary uses of conditional probability.

The rules given in Appendix B1 also hold true for conditional probabilities:

Proposition 45. *Let ω denote an arbitrary event and let A, B, and C denote events. Assume $P(B) > 0$. Then the following are all true.*

(i) $P\{\omega|B\} \geq 0$ *and* $P(A|B) \geq 0$.

(ii) $\sum_{\omega \in B} P\{\omega|B\} = 1$.

(iii) If $A = C$, *then* $P(A|B) = P(C|B)$.

[2]The case in which $P(B) = 0$ is of no interest, because, then, $P(A|B)$ is the probability of A given that something impossible has or will occur; that is, it is a nonsensical quantity.

(iv) *If* $A \subseteq C$, *then* $P(A|B) \leq P(C|B)$.

(v) *The probability that event* A *does* not *occur conditional on* B *is* $1 - P(A|B)$; *that is,* $P(A^c|B) = 1 - P(A|B)$.

(vi) $P(A \cap C|B) \leq P(A|B)$ *and* $P(A \cap C|B) \leq P(A|B)$.

(vii) $P(A \cup C|B) \geq P(A|B)$ *and* $P(A \cup C|B) \geq P(A|B)$.

(viii) $P(A \cup C|B) = P(A|B) + P(C|B) - P(A \cap C|B)$, *from which it follows that*

  (a) $P(A \cup C|B) \leq P(A|B) + P(C|B)$; *and*

  (b) $P(A \cup C|B) = P(A|B) + P(C|B)$ *if* $A \cap B$ *and* $C \cap B$ *are* mutually exclusive *events.*

For future reference, observe that we can rewrite expression (B2.1) in the following ways:

$$P(A \cap B) = P(A|B) \times P(B)\,;$$

or, reversing the roles of A and B,

$$P(A \cap B) = P(B|A) \times P(A)\,. \tag{B2.2}$$

Independence | B2.1

In the previous section we noted that the difference between $P(A)$ and $P(A|B)$ reflected what learning that event B had or would occur told us about the probability that event A would occur. If $P(A) < P(A|B)$, then learning B caused us to increase the probability with which we believed A would occur. If $P(A) > P(A|B)$, then learning B caused us to decrease the probability with which we believed A would occur. What if $P(A) = P(A|B)$? Then B tells us *nothing* about the probability that A will occur. In this last case, we say that A and B are *independent events*.

If A and B are independent, then

$$P(A) = P(A|B) = \frac{P(A \cap B)}{P(B)}\,,$$

where the second equality comes from expression (B2.1). Ignoring the middle term, if we multiply both sides of this last expression by $P(B)$, we obtain

$$P(A) \times P(B) = P(A \cap B)\,.$$

From this, we can conclude

Proposition 46. *If* A *and* B *are* independent *events, then* $P(A \cap B) = P(A) \times P(B)$.

Be careful! Proposition 46 applies only to *independent* events. If A and B are *not* independent events, then $P(A) \times P(B) \neq P(A \cap B)$. A common mistake is to presume two events are independent, when they are not, and to then calculate the probability of their both occurring as the product of each occurring. For example, suppose there are two safety systems, A and B. Suppose that the probability of A failing is $1/1000$ and the probability of B failing is $1/1000$. Is the probability of disaster (both failing) 1 in 1 million? Probably not. If, for instance, A can fail due to a power surge and so can B, then a single power surge might cause both to fail simultaneously. That is, the probability of both systems failing is greater than 1 in a million because they are not truly independent.

Bayes Theorem | B2.2

In this section, we study one of the most important rules for updating probabilities, namely Bayes Theorem. To do so, though, we need to consider some additional definitions.

Let Ω (read "omega") be the set of all possible outcomes. That is, every outcome is contained in Ω (*i.e.*, $\omega \in \Omega$ for all outcomes ω). From this, it follows that every event is also contained in Ω (*i.e.*, $A \subseteq \Omega$ for all events A). Finally, because all outcomes are in Ω, it follows from expression (B1.1) on page 158 that $P(\Omega) = 1$ (which can be interpreted as "something is certain to happen").

We say that a list of events are *exhaustive* if collectively they contain all the possible outcomes. This is denoted by $A_1 \cup A_2 \cup \cdots \cup A_T = \Omega$, where A_t denotes one of T events. It can be shown that the following is true:

Proposition 47. *Let $A_1, \ldots, A_T$, and B be events. Suppose that $A_1, \ldots, A_T$ are* mutually exclusive *and* exhaustive, *then*

$$P(B) = P(A_1 \cap B) + \cdots + P(A_T \cap B)\,.$$

Recall expression (B2.2); that is, $P(A \cap B) = P(B|A) \times P(A)$. From that fact and Proposition 47, it follows that

Proposition 48 (Law of Total Probability)**.** *Let $A_1, \ldots, A_T$, and B be events. Suppose that $A_1, \ldots, A_T$ are* mutually exclusive *and* exhaustive, *then*

$$P(B) = P(B|A_1) \times P(A_1) + \cdots + P(B|A_T) \times P(A_T)\,. \tag{B2.3}$$

Let $A_1, \ldots, A_T$ be *mutually exclusive* and *exhaustive* events. From the definition of conditional probability (*i.e.*, expression (B2.1) above), we know that

$$P(A_t|B) = \frac{P(A_t \cap B)}{P(B)}\,. \tag{B2.4}$$

We also know that $P(A_t \cap B) = P(B|A_t) \times P(A_t)$ and that we can write $P(B)$ using the Law of Total Probability (*i.e.*, using the formula (B2.3) above). Substituting these two expressions into (B2.4) as appropriate, we obtain:

Theorem 5 (Bayes Theorem)**.** *Let $A_1, \ldots, A_T$, and B be events. Suppose that $A_1, \ldots, A_T$ are* mutually exclusive *and* exhaustive *and that $P(B) > 0$, then for any event A_t, we have*

$$P(A_t|B) = \frac{P(B|A_t) \times P(A_t)}{P(B|A_1) \times P(A_1) + \cdots + P(B|A_T) \times P(A_T)}. \tag{B2.5}$$

Example 25 [Paradox of the Three Prisoners]: Once upon a time, there were three prisoners: 1, 2, and 3. Each prisoner was kept in a separate cell and the prisoners could not communicate with each other. The prisoners knew that in the morning one of them would be executed, while the other two would be set free. Although the authorities knew which prisoner would be executed, the prisoners did not. Suppose each prisoner assumed she had a 1/3 probability of being executed. Suppose that Prisoner 3 could not wait the night to find out whether she was to be executed. She called a guard over and asked him "am I to be executed?"

The guard replied, "you know that tradition forbids me from telling you that."

"Well, then," said Prisoner 3, "could you at least tell me which of the other two prisoners will be set free?"

"Since there is no information in that, I dont see why not. Prisoner 1 will be set free," said the guard, who never lies.

Prisoner 3 sat back in her cell and reflected, "well, then, it's between Prisoner 2 and me. But, wait, this means I have a 50% chance of being executed! Woe is me!" As Prisoner 3 sank into depression, it suddenly occurred to her that had the guard told her Prisoner 2 would be set free, she still would have calculated her probability of being executed as being 1/2. "Wait a minute," she thought, "that means that regardless of what the guard told me, my probability of being executed was 1/2; which means that my initial probability of being executed should have been 1/2 instead of 1/3 . I must have made a mistake somewhere!"

Indeed, Prisoner 3 has made a mistake. To see what her mistake was, as well as what the true probability of her execution is, let's employ Bayes Theorem. To do so, we need one more assumption: Suppose that if both Prisoners 1 and 2 are to be set free, then the probability is 1/2 that the guard tells Prisoner 3 that Prisoner 1 will be released and the probability is 1/2 that the guard tells Prisoner 3 that Prisoner 2 will be released. In terms of the formula above, we can think of A_t as the event "Prisoner t will be executed" and we can think of B as the event "Prisoner 3 is told Prisoner 1 will be set free." The question we seek to answer is what is the probability that Prisoner 3 is to be executed given that Prisoner 1 is to be set free; that is, what is $P(A_3|B)$? The data we have been given to answer this question can be expressed as

1. $P(A_1) = P(A_2) = P(A_3) = 1/3$ (the *unconditional* probability of Prisoner t's being executed is 1/3).

2. $P(B|A_1) = 0$ (the guard never lies, so he won't say Prisoner 1 is to be set free if he is really doomed).

3. $P(B|A_2) = 1$ (if Prisoner 2 is to be executed, the Prisoner 3 will *certainly* be told that Prisoner 1 will be set free).

4. $P(B|A_3) = 1/2$ (if Prisoner 3 is to be executed, then, by assumption, the probability that the guard tells Prisoner 3 that Prisoner 1 is to be set free is $1/2$).

Inserting these data into Bayes Theorem, we have that $P(A_3|B)$—the probability that Prisoner 3 will be executed given she was told Prisoner 1 was to be set free—equals

$$P(A_3|B) = \frac{\frac{1}{3} \times \frac{1}{2}}{0 \times \frac{1}{3} + 1 \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3}} = \frac{1}{3}.$$

That is, having learned that Prisoner 1 will be set free, the probability that Prisoner 3 will be executed is $1/3$. As the guard said, the knowledge that Prisoner 1 will be freed is uninformative with respect to whether Prisoner 3 will be executed (note this also means that it is independent of whether Prisoner 3 will be executed). Prisoner 3's mistake was to forget that the guard would always tell her that one of her fellow prisoners would be released regardless of who was to be executed. In other words, asking which of her fellow prisoners will be released is equivalent, in terms of estimating her own probability of execution, to asking will one of her fellow prisoners be released. Since only one prisoner will be executed, this second question cannot yield an informative response; hence, neither can her original question.

The Paradox of the Three Prisoners illustrates one way Bayes Theorem can be useful; namely showing that what seems like information is not really information. Another way Bayes Theorem can be useful is in showing that what seems uninformative is actually informative; as the following example illustrates.

> **Example 26 [The Monty Hall Problem]:** This example is based on an old American television game show called *Let's Make a Deal*, which was hosted by a man named Monty Hall.[3] At one stage of the game, a contestant is asked to choose among three closed doors (numbers 1, 2, and 3). Behind one of the doors is a valuable prize (*e.g.*, a new car). Behind two of the doors are worthless joke prizes (*e.g.*, goats). The contestant does not know which door conceals the valuable prize, but he does know that the probability that the valuable prize is behind any given door is $1/3$. After choosing one of the three doors, Monty Hall has one of the unchosen doors opened to reveal one of the two joke prizes (Monty Hall knows which door conceals the valuable prize). Monty Hall then asks the contestant if he would like to switch his choice from the door he chose originally to the other closed door. If the contestant switches, then he gets the prize behind the door to which he switches. If he doesnt switch, then he gets the prize behind the door he originally chose. Assuming the contestant wants to maximize his

[3] I make no claim that the actual game show was played in the manner described here. This is simply a well-known puzzle.

> probability of winning the valuable prize, should he switch or should he stay with the door he originally chose?

Most people when confronted with this problem answer that it doesn't matter whether the contestant switches or not—the probability of winning the valuable prize will be the same; that is, seeing a door opened is uninformative. This, however, is incorrect: By switching the contestant *doubles* his probability of winning! Let's use Bayes Theorem to see why. Before doing so, we need one further assumption: if both the doors that the contestant did not select have joke prizes behind them, then Monty Hall is equally likely to open one as the other (remember Monty Hall only opens an unchosen door and only a door that has a joke prize behind it). For concreteness, suppose that the contestant chooses door #1 and Monty Hall opens door #2 to reveal a joke prize. Let A_t denote the event "the valuable prize is behind door number t" and let B denote the event "Monty Hall opens door #2." We want to know the probabilities that the valuable prize is behind door #1 or door #3 (since Monty Hall opens door #2 to reveal a joke prize, we know the valuable prize is not there); that is, we want to know $P(A_1|B)$ and $P(A_3|B)$. Indeed, since $P(A_1|B) = 1P(A_3|B)$, we need to determine only $P(A_3|B)$. The data we are given are

1. $P(A_1) = P(A_2) = P(A_3) = 1/3$.
2. $P(B|A_1) = 1/2$ (Monty is equally likely to open door #2 as he is to open #3 if both conceal joke prizes).
3. $P(B|A_2) = 0$ (Monty only opens a door that has a *joke* prize behind it).
4. $P(B|A_3) = 1$ (same reason).

Putting all this into Bayes Theorem yields

$$P(A_3|B) = \frac{1 \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3}} = \frac{2}{3};$$

hence, $P(A_1|B) = 1/3$. As claimed, the contestant doubles his probability of winning the valuable prize by switching.

Random Variables and Expectation | B3

Often we are interested in the payoffs associated with different outcomes. For instance, suppose you and a friend decided to wager a dollar on the outcome of a coin toss: heads you win a dollar; tails you lose a dollar. From your perspective, the outcome *heads* is associated with the payoff \$1 and the outcome *tails* is associated with the payoff −\$1. Traditionally, such payoffs are called *random variables* and are denoted as *functions* of outcomes. For example, if $x(\cdot)$ is a function that associates each outcome, ω, with a payoff, then, in the coin-toss example, one would write $x(\textit{heads}) = 1$ and $x(\textit{tails}) = -1$.

In general, one can dispense with keeping track of the outcomes; thus, instead of writing $x(\omega)$, one can just write x; and instead of writing "the probability of getting $x(\omega)$ is $P\{\omega\}$," one can just write "the probability of getting x is $P(x)$." In a sense, one can just think of payoffs as being outcomes.

The payoffs, since they are numbers, can be ordered from smallest to largest. Hence, if there are N possible payoffs, we would list them as $x_1, x_2, \ldots, x_N$, where it is to be understood that if $m < n$, then $x_m < x_n$.

Taking advantage of this ordering, we can write the probability $P(x_n)$ as just p_n. If there are N possible payoffs, then the list of probabilities $p_1, \ldots, p_N$ is called the *density* associated with the random variable (payoff) x.

Expectation | B3.1

When faced with an uncertain situation, say the gamble described in the previous section, we are often interested in knowing how much we can expect to win or how much we might win on average. This notion is captured in the concept of *expectation*. The *expectation of a random variable* is denoted $\mathbb{E}\{x\}$ and is defined as

$$\mathbb{E}\{x\} = p_1 \times x_1 + \cdots + p_N \times x_N \tag{B3.1}$$

for a random variable with N possible payoffs. In words, the expectation of a random variable is the sum, over the possible outcomes, of the product of each payoff and its probability.

One way to think about expectation is that the expectation of x is exceedingly close to your average[1] payoff if you were to repeat the uncertain situation

[1]Recall the (arithmetic) average of a set of numbers is their sum divided by the number of numbers in the set. For example, the average of 1, 3, and 8 is 4 $(= (1+3+8)/3)$.

a large number of times.[2] Another way to think about expectation is that the expectation of x is the "best guess" as to value that x will take; where "best guess" means the one with the least error.

As an example, the expected value of the gamble in which you receive \$1 if a coin lands heads but you pay \$1 (receive $-$\$1) if the coin lands tails is

$$\frac{1}{2} \times (-1) + \frac{1}{2} \times 1 = 0$$

dollars.

As a second example, suppose that your ice cream parlor makes \$2000 on a sunny day and \$1000 on a rainy day. Suppose that the probability of rain tomorrow is $1/5$, then you can expect to make

$$\frac{1}{5} \times 1000 + \frac{4}{5} \times 2000 = 1800$$

dollars tomorrow.

One can also take the expectation of a *function* of a random variable. For instance, suppose that you bet \$1000 on the toss of a coin: heads you win \$1000 and tails you lose \$1000. Suppose that you must pay 14% income tax on your winnings and you cannot deduct your loses from your income taxes. Then, from your perspective, you care about the following function of x:

$$f(x) = \begin{cases} x\,, \text{ if } x \leq 0 \\ .86x\,, \text{ if } x > 0 \end{cases}\,, \tag{B3.2}$$

because it gives your winnings in *after*-tax dollars. Your expected after-tax winnings are

$$\mathbb{E}\{f(x)\} = \frac{1}{2} \times (-1000) + \frac{1}{2} \times .86 \times 1000 = -70$$

dollars.

The general rule for the expectation of a function of a random variable is

$$\mathbb{E}\{f(x)\} = p_1 \times f(x_1) + p_2 \times f(x_2) + \cdots + p_N \times f(x_N)\,. \tag{B3.3}$$

Note that it is generally *not* true that $\mathbb{E}\{f(x)\} = f(\mathbb{E}\{x\})$; that is, it is generally *not* true that the expectation of the function equals the function of the expectation.

Some standard definitions concerning random variables and their expectations:

- The *mean* of a random variable is its expectation. That is, the mean of x is $\mathbb{E}\{x\}$.

[2] In fact, it can be proved that as the number of replications tends to infinity, the expectation will be arbitrarily close to the average you actually get. This is called the *Law of Large Numbers*.

- The *variance* of a random variable is the expectation of its deviation from its mean. That is, the variance of x—denoted $\mathrm{Var}(x)$—is

$$\mathrm{Var}(x) = p_1 \times \left(x_1 - \mathbb{E}\{x\}\right)^2 + \cdots + p_N \times \left(x_N - \mathbb{E}\{x\}\right)^2 .$$

- The (population) *standard deviation* of a random variable is the square root of its variance; that is, the standard deviation of x is $\sqrt{\mathrm{Var}(x)}$.

Distributions | B3.2

The *distribution* of a random variable x is a function that gives the probability that $x \leq y$ for each value y. If $F(\cdot)$ is the distribution function for the random variable x, then $F(y)$ is defined as

$$F(y) = \sum_{\{n|x_n \leq y\}} p_n ,$$

where $\{n|x_n \leq y\}$ is read as "the set of indices such that the value x_n is less than or equal to y." For example, if

$$x \in \{1, 2, 4, 7, 10\} \text{ and } p_n = \frac{n}{15} ,$$

then

$$F(3) = \frac{1}{15} + \frac{2}{15} = \frac{1}{5} ;$$

$$F(6.99) = \frac{1}{15} + \frac{2}{15} + \frac{3}{15} = \frac{2}{5} ; \text{ and}$$

$$F(7) = \frac{1}{15} + \frac{2}{15} + \frac{3}{15} + \frac{4}{15} = \frac{2}{3} .$$

Since, in this example, it is impossible to get an $x < 1$, $F(y) = 0$ for all $y < 1$. Since, in this example, we must get an $x \leq 10$, $F(y) = 1$ for $y \geq 10$.

$\int dx$ Continuous Distributions

Although so far we have defined probability in terms of discrete outcomes, we can define them in terms of continuous outcomes. Basically, this means considering distribution functions that are differentiable. Let $G(\cdot)$ be a differentiable distribution function. The derivative of a differentiable distribution function is called the *density function*. Let $g(\cdot)$ denote the density function; that is,

$$g(y) = \frac{d}{dy} G(y) .$$

Because integration is the "inverse" of differentiation, we know

$$G(y) = \int_{x_{\min}}^{y} g(x)dx\,,$$

where $x_{\min}$ is the smallest possible value of the random variable x.

For random variables with differentiable distribution functions, their mean and variance are defined, respectively, as:

$$\mathbb{E}\{x\} = \int_{x_{\min}}^{x_{\max}} xg(x)dx\,;\text{ and}$$

$$\mathrm{Var}(x) = \int_{x_{\min}}^{x_{\max}} \left(x - \mathbb{E}\{x\}\right)^2 g(x)dx\,,$$

where $x_{\max}$ is the largest possible value of x.

Two commonly used differentiable distribution functions are the normal and the uniform.

- The *normal distribution* has a density function given by

$$\sqrt{\frac{1}{2\sigma\pi}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\,,$$

  where σ (read "sigma") is the standard deviation of x, μ (read "mu") is the mean, π is the constant pi (*i.e.*, approximately 3.1416), and where $\exp(z)$ means the base of the natural logarithm (*i.e.*, $e \approx 2.7183$) raised to the zth power. A normally distributed random variable has a range from $-\infty$ to ∞. The probability of drawing such an x less than or equal to a given y is

$$\int_{-\infty}^{y} \sqrt{\frac{1}{2\sigma\pi}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)dx\,.$$

  This expression does not have a closed-form solution.

- The *uniform distribution* from 0 to Y has a density of $1/Y$. The mean of a random variable distributed uniformly from 0 to Y is $Y/2$. Its variance is $Y^2/12$. If x is distributed uniformly on between 0 and Y, then the probability of drawing an x less than or equal to a given y, $0 \le y \le Y$, is

$$\int_0^y \frac{1}{Y}dx = \frac{y}{Y}\,.$$

Index