Introduction to Mathematical Economics

Wilson Mixon

June 14, 2017

This book is dedicated to the *Maxima* development team.

Preface

The title page of this book is a bit misleading. I (Wilson Mixon) am not the author of most of the book's material. I am grateful to have received the permission of Professors Anthony L. Ostrosky, Jr. and James V. Koch to use their textbook [16] as the basis for this project. I have edited their material slightly and have added some material. Mainly, I have incorporated material from the *Maxima* open-source computer algebra system. Also, I have extended the original discussion of sets and lists.

I hope that this project reflects well on the work of Professors Ostrosky and Koch. I accept responsibility for any shortcomings that I have introduced in this rendering of their work.

It is fitting that the bulk of this preface should be in the words of the original authors:

> . . . Economists are called on with increasing frequency to apply their logic and tools to a variety of societal problems: pollution; depletion of natural resources; crime; urban sprawl; taxation; food production; the depreciating dollar–the list swells daily. The ability of the economist to speak to these problems reflects a well-developed body of theory, modes of analysis that emphasize logic, and sophisticated quantitative tools. . . .
>
> Mathematics has played a central role in enabling economists to rigorously state their theorems, with emphasis on logical inference, and in enabling them to . . . [test] the empirical validity of their theories. The primary aim of this book is to show fledgling students how, where, when, and why they can appropriately utilize mathematics in economics and business. The student who masters the mathematical tools presented in this book will not

> only be able to read and apply [much] of the "language" that modern economic theory uses, but also he or she will learn (and perhaps for the first time understand) a great deal of economic theory. If the readers of this book are similar to the students at Illinois State university [where the authors taught],then one can expect that this combination of mathematics and economics will turn on lights, open doors, and yield deeper understandings. Many are the "math econ" students who have suddenly exclaimed, "Now I see what my Econ 100 instructor really meant!"
>
> We do not claim to present all the many applications of mathematics to economics and business in this book. This book is a well-defined one-semester introductory approach to the use of mathematics in economics and business. ...
>
> [Even so, an] outstanding feature of this book is the plentiful use of examples and applications. Each chapter contains a large section that is entirely devoted to applications of the mathematical tools; one entire chapter is devoted to specific applications of matrix algebra. Several examples, such as Stigler's famous "diet problem," are used on a number of occasions in order to demonstrate the power of applied mathematics. ...
>
> The organization of the book is based on the view that a thorough review of basic precalculus mathematics and algebra is the correct place to start. The differential calculus, with its many applications, is then introduced. Maximization and minimization techniques are plentifully' used and illustrated. The integral calculus is the next. major topic; two chapters are devoted to its exposition and application. Finally we cover matrix algebra, and devote an entire chapter to linear programming and input-output analysis in a matrix-algebra context. The overall organization of the book stresses a building-block approach, whereby each newly introduced topic depends on the topics previously covered.

Much of the material in this text is produced by *Maxima*, the open-source computer algebra system. The text is, however, self-contained. Some of the references to *Maxima* might affect the flow of the text, but the content can be accessed without the use of *Maxima*.

This text serves as an introduction to some of the mathematics that pertains to economics. It is not a full introduction to *Maxima* or to the *wxMax-*

ima user interface. To incorporate *Maxima* into your study of this material, look at these two sites: http:// maxima.sourceforge.net/ and http:// andrejv.github.io/wxmaxima/. Both sites have links to documentation.[1]

Another site that readers of this text should visit is http://statmath.wu.ac.at/leydold/maxima/ [9]. This site contains the text *Introduction to Maxima for Economics*, which has a quite complete introduction to both *Maxima* and *wxMaxima*. It also provides an briefer development of much of the material in this text. Finally, it contains some more advanced material, especially the treatment of ordinary differential equations.

I am coauthor (with Michael Hammock) of a textbook that develops microeconomic theory more fully than the confines of the present text allow. That text, *Microeconomic Theory and Computation*[7], also provides more detail on the use of *Maxima* than is provided here.

[1] You probably will not want to download *wxMaxima*. Windows and MacOS users should download an executable file that will install *wxMaxima*. Linux users can access *Maxima* and *wxMaxima* from their repositories.

Acknowledgements

This effort is dedicated to the *Maxima* team, which maintains and continually improves this remarkable piece of software. Among the members of that team, I thank Robert Dodier and Andrej Vodopivec for encouraging my early efforts to produce material to illustrate *Maxima*'s power and usefulness. A third team member, Gunter Königsmann, continues to provide help and insight in the use of the *wxMaxima* graphical user interface which greatly facilitates the use of *Maxima*, especially by newcomers.

I thank Michael Hammock for working with me on the previously-mentioned text and for maintaining a website that contains a large amount of material on the use of *Maxima* in economic analysis (www.wxmaximaecon.com).

Finally, and above all, I thank my wife Barbara. A suggestion that she made a few years ago pointed me toward *Maxima*. Since then, I've repaid her by bending her ears about the most recent tidbit that I've discovered by working on various projects and by providing material for her to proofread. Remarkably, she continues to encourage me and even offers to proofread new material. Who would believe such a report?

Contents

Chapter 1

The Role and Power of Mathematics

Mathematics is a rigorous and well-defined study of the structures, configurations, and interrelationships that characterize the world in which human beings live. Mathematics provides an exacting language that articulates the essential characteristics of a wide range of situations so that the key aspects of those situations can be dispassionately examined.

Modern mathematics is "economical" in the best sense of the word. It clearly states the barebones assumptions that underpin a relationship. In addition, it highlights the logical processes that characterize the relationship. Finally, it states any conclusions that are implied by the relationship in a clear and concise form.

Economics is filled with topics that are amenable to mathematical analysis. Relationships can be specified to relate production cost to output, output to inputs, wages to worker productivity, and so forth. Further, analysts often wish to establish the general nature of conditions required to minimize the cost of achieving a certain objective, or to maximize the output of a particular productive process. Sometimes, we seek specific values as well as knowledge of the requisite conditions. In still other circumstances, the analyst may seek evidence regarding how much change will occur in sales when the firm alters the amount of advertising it is undertaking, knowing that the statistics-based evidence will be somewhat imprecise.

Mathematical analysis can apply to abstract concepts as well as to concrete

ones. An important member of this category is utility, which is central to economic analysis. Utility is a highly abstract concept, but salient aspects can be represented mathematically. The relevant point is that mathematics is capable of dealing with a wide range of relationships that confront analysts and decision-makers.

The versatility of mathematics is apparent. Equally important is its power. With the help of mathematics, analysts can address problems that can be given only a cursory glance with a strictly verbal analysis. Indeed, we can make certain statements mathematically that with verbal language either cannot be made at all or that must be made only in an awkward fashion. The next section addresses a revealing historical example.

1.1 Stigler's Diet Problem

In 1945, George Stigler [19] addressed the so-called "diet problem." He sought the least expensive combination of foods available to consumers that would enable them to satisfy the recommended daily dietary allowances established by the Food and Nutrition Board of the National Academy of Sciences. That is, he set out to determine the cheapest way to obtain the nutrients that individuals need to sustain life.

We sketch Stigler's approach here, in order to illustrate salient aspects of mathematical analysis. Chapter XX returns to this problem when we consider linear and nonlinear programming techniques. The equation below shows the function that is to be minimized. The total daily cost of a subsistence diet, C, is the sum of the amounts spent on each of eighty goods, where the amount spent on Good j is $Pj \cdot Xj$. The price is Pj and the quantity is Xj.[1]

$$C = P1 \cdot X1 + P2 \cdot X2 + \cdots + Pn \cdot Xn$$

[1]For reasons that will become clear later, we do not use subscripts. Much of our work will involve commands that are written in text, so that subscripts are hard to enter. Multiplication is indicated with centered dots ($\cdot$) so that $P2$ is a variable name. In contrast, $P \cdot 2$ is a product, with the variable name being P. We will not be entirely consistent in our usage. In some settings, subscripts provide for easier interpretation. The context will typically make the usage clear.

In Stigler's specification, (up to) eighty types of food were to be combined to define the least-cost diet. Suppose that item 1 is peanut butter and the price of peanut butter is \$0.04 per gram. Then the cost of peanut butter in this diet is \$.04 X1, where X1 is the number of grams of peanut butter in a daily diet. Again, the sum off all such terms is C, the cost of the diet.

As noted, however, the diet must satisfy a set of constraints that dietitians had established. Nine such constraints were identified and incorporated into Stigler's analysis. The following expression shows what one of the nine equations in this model might look like: $a11 \cdot X1 + a12 \cdot X2 + \cdots + a1n \cdot Xn \geq 3000$, where 3000 is the number of calories required.

This equation is the first constraint–one that identifies minimum caloric requirement. The first of the two numbers attached to each coefficient a indicates the constraint number (1 here). The second identifies the food number. The other eight constraints look much like this one. For the first term in (1.2), therefore, X1 is peanut butter, which provides about 7 calories per gram, so $a11 \approx 7$.

Stigler's problem, addressed in the early 1940s, involved repeatedly evaluating combinations of eighty foods to ensure that they satisfied the nine constraints and then determining the cost. Not surprisingly, Stigler concluded (p. 310) that his approach to solving the problem was " ... experimental because there does not appear to be a direct method of finding the minimum of a linear function subject to linear conditions." That is, Stigler was forced to find a solution by hit-or-miss methods.

Before discussing the tentative nature of this conclusion, we briefly summarize the findings. Stigler determined that the established nutritional requires could be satisfied for about \$60 per year (in 1944 dollars, about \$800 per year in 2014 dollars). The major food items in the diet were these: wheat flour, cabbage, spinach, pancake flour, and pork liver.

Stigler's research is instructive at a number of levels. The direct implications of the results, taking them at face value, are important. This is a classic economic minimization problem with real world consequences. A second observation is that advances in applied mathematical analysis greatly facilitated the analysis that Stigler pioneered. Finally, changes in our understanding of nutrition require reexamination of the analysis.

- The first and obvious implication of the analysis is that the diet is

not very palatable. Not many individuals would find a diet limited to Stigler's set of ingredients to be very tasty.

- Second, the meaning of the seemingly direct phrase "the cost of food" is not as clear-cut as one might hope. Spending on food in the United States is around $3500 *per capita*. We buy much more than subsistence, and a casual reference to "the cost of food" refers to something quite different from the cost of a subsistence diet. We eat more, sometimes too much more, than subsistence requires and, more importantly, we select foods that have attractions other than mere subsistence.

- Third, linear programming was invented in 1939 in the Soviet Union and later in the United States. Its existence was unknown to Stigler, however, because of secrecy surrounding World War II. By 1947, however, the technique was published and had become widely used. Thus, within the decade in which Stigler's work was published, more efficient solutions became available. Even low-cost personal computers today provide us the ability to solve systems like Stigler's analytically, rather than by approximation. The same is true for more complex systems that consist of nonlinear relationships. Indeed quite sophisticated software is routinely included in programs like *Microsoft Excel* and the open-source *LibreOffice Calc*. Also, it is part of computer algebra systems, which the next section discusses.

  The end result is that advances in computing power coupled with new mathematical methods, have made solutions to problems like the one that Stigler pioneered routine. These advances now guide decisions in manufacturing, in communications, and in transportation.

- Fourth, we have learned more about the characteristics of foods since Stigler attacked his diet problem. For example, nutritionists now are aware that liver, though a good source of thiamine and manganese, protein, vitamin A, and numerous other nutrients, is very high in cholesterol. Changing knowledge of nutrition is one reason that analysts have updated Stigler's findings.

- Fifth, the prices of foods relative to those of other goods and services has declined and relative prices have changed among types of foods.

- Finally, we now have more foods to choose from for our new cost-minimizing diet.

In reexamination, Bassi [1] estimated a *per-capita* cost of about $160 in 1975 dollars (about $670 dollars at the 2014 general price level). Bassi's mix looks a lot like Stigler's, except that red beans play a larger role and beef kidneys have replaced pork liver as the predominant meat source (again pointing out the distinction between a subsistence diet and a palatable diet).

1.2 Analysis Using a Computer Algebra System

One product of the confluence of mathematical advances and increased computing power is the Computer Algebra System (CAS). A CAS is a mathematical toolkit that can be used as a simple (or advanced) calculator. It can also be used to find solutions to symbolic mathematical expressions and, if these expressions cannot be solved, to produce simulations that provide insights into the expressions' implications.

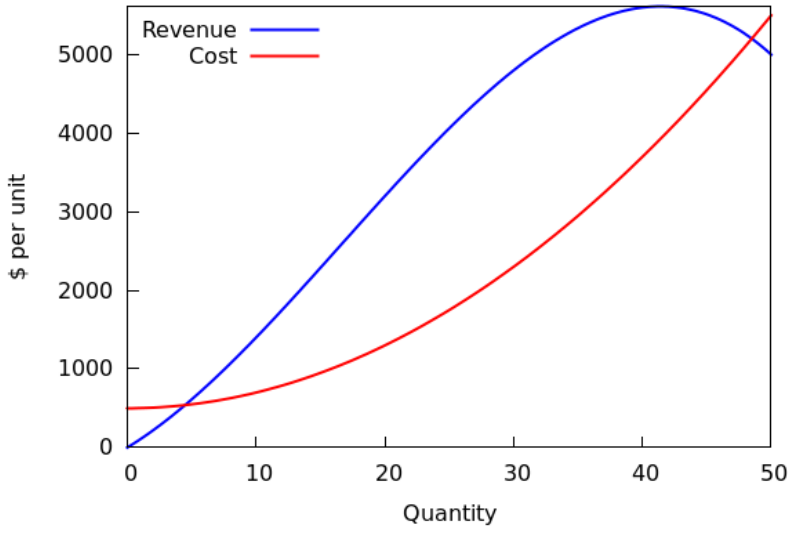
Both proprietary CAS programs (ones that must be purchased) and open source programs (one need not pay to use them) exist. The two most widely-used proprietary CAS programs are *Mathematica* and *Maple*. The most widely-used open source program, and the one used in the remainder of this book, is *Maxima*. An analyst who has gained skill in using any one of these three programs can quickly transfer that skill to the use of either of the other two. See citeMeglicki.

1.2.1 The *wxMaxima* User Interface

The *Maxima* CAS offers a selection of user interfaces, the most popular one being *wxMaxima*. The remainder of this section consists of a quick overview of *wxMaxima*, based on a screen shot of a small notebook. We will not actually begin to learn how to use *wxMaxima* here, but instead simply preview a few of its possibilities. The appendix to this text provides more detail. Also, the website that accompanies this book (http://www.wxmaximaecon.com/)

```
(%i9) Revenue: 100*q + 5*q^2 - 0.1*q^3; Cost: 500 + 2*q^2$
(%o8) -0.1 q^3+5 q^2+100 q

(%i10) wxdraw2d( user_preamble="set key left",
        xlabel="Quantity", ylabel="$ per unit",
        key="Revenue",explicit(Revenue,q,0,50),
        color=red, key="Cost",explicit(Cost,q,0,50))$
```



(%t10)

```
(%i11) [find_root(Revenue - Cost, q, 0, 20),
       find_root(Revenue - Cost, q, 20,50)];
(%o11) [4.486456043643644,48.49471574640573]
```

Figure 1.1: wxMaxima Interface

provides links for downloading *Maxima*, which includes *wxMaxima* and for getting started in *wxMaxima*. Also, see [7], Chapters 1 and 2.

Figure 1.1 shows three input/output cells. The input is entered as text. Commands are ended by semicolons (if resulting output is to be printed) or dollar signs (if resulting output is to be suppressed). Once a set of commands is to be executed, ctrl-enter generates the output.

The first cell shows a simple way to assign a name to an expression (the two commands in the first cell). The second cell graphs the expression(s), and the third cell shows how to find a root of an expression, in this case "Revenue = Cost." At this point, just observe the general nature of the workbook

without concern regarding particulars of the commands.

Briefly consider the third cell, which involves the determination of one of the two quantities at which revenue and cost are equal. This is a "break-even" quantity, the minimum quantity that the firm must sell in order to cover costs. The solution is "numerical," meaning that we just looked for a numerical value, not an algebraic solution. Later, will we return to examples like this one, but we also use *Maxima* to examine the nature of solutions, including determining the quantity that yields maximum profit. This figure is the first of many that follow. Most of the cells show the commands that generate a set of results and those results.

The last part of Figure 1.1 is a horizontal line. This is a "cursor"–*wxMaxima* interprets any keyboard entry on such a line as the beginning of a new set of one or more commands.

1.3 The Diet Problem in *Maxima*

We close this brief introduction to *Maxima* by looking back to the Stigler diet problem. For the sake of clarity and simplification, Stigler reported a subset of his larger model in order to focus on the logic of his approach. Zhou provides an extensive development of Stigler's analysis of that subset, and we use Zhou's notation.[2] The cell below shows that we wish to minimize z, which is the cost function, subject to a set of the eight constraints that apply (actually 13 constraints because we add five nonnegativity constraints) in this simplified representation of Stigler's analysis. Each constraint in a linear expression. The result is a set of five values for the five foods ($x1$, $x2$, $x3$, $x4$ and $x5$) that enter into this analysis and the value z = 0.109. Chapter xx returns to this issue and interprets the values more fully..[3]

This introduction to *wxMaxima* is necessarily cursory. Even so, we have achieved the following: assigning names to expressions, graphing those ex-

[2]Wenxiao Zhou, "A Discussion on 'The Stigler Diet Problem' by Applying the Simplex Method & GAMS,"22 April 2013, http: //www. unc.edu/ marzuola/ Math547_S13/ Math547_S13_Projects/W_Zhou_Section003_SimplexMethod.pdf.

[3]This presentation of the input/output cell, where it appears as a single figure, will not be used beyond this point. Rather, the input will appear after the identifier (%i), and output will appear after the identifier (%o) except when the output is a graph. Graphs will follow the input and will be numbered sequentially through each chapter.

```
(%i15) load(simplex)$
       minimize_lp(z,[c1,c2,c3,c4,c5,c6,c7,c8,
         x1>=0,x2>=0,x3>=0,x4>=0,x5>=0]);
(%o16) [0.109,[x5=0.0486,x4=0.00511,x3=0.0112
,x2=0.00859,x1=0.0355]]
```

pressions, using a numerical method to determine a critical value (analytical methods follow in subsequent material), and solving a linear programming problem.

1.4 Summary

Mathematics underlies the analysis of many of the issues that business decision makers, public policymakers, and economists address. This chapter discusses the application of mathematical analysis as it applies to an important issue, the cost of a subsistence diet. This example shows how mathematics can be applied, and it shows that the mathematical tools are becoming increasingly sophisticated. Coupled with the power of modern computers, this increased sophistication has broadened the purview of applied mathematical analysis.

One important advance of the past few decades has been the development of computer algebra systems. These systems extend analysts abilities by solving complex problems, by allowing for the examination of a range of scenarios, and by producing simulations of systems that defy formal solution. The remainder of this book displays some of these features by applying the *Maxima* open-source computer algebra system.

The development and use of mathematical tools to solve business and economic problems has expanded very rapidly in recent years. A course covering the materials presented in this book is now required of business administration, accounting, marketing, and economics majors in most colleges and universities. It behooves the student who wishes to be well prepared and efficient to master the mathematics that appear in the following chapters, not only because it probably will be necessary in order to graduate, but also because mathematics will prove to be very useful in later employment. Mathematical economics can open new doors to those who take the time to master its essentials.

1.5 Questions and Problems

1. Mathematics can be used in very useful ways in analyzing many business and economics problems. This fact has led some enthusiasts to contend that "If you can't measure it, it isn't worth knowing about," or "If you can't measure it, it doesn't exist." Write a critique of these statements.

2. Many historians claim that they do not use models in writing history and in arriving at conclusions about historical phenomena. Is it possible to analyze something without having an underlying model? Will hard work produce insights and generalizations if you do not have a model? Explain.

3. Those who use mathematical tools in the analysis of business and economics problems frequently contend that it is possible to say things with mathematics that could not be said verbally. Is this true? Can the reverse be true?

4. A frequent criticism of business and economics models is that they do not fit the real world with precision. The real world nearly always seems to be somewhat different from the world outlined in the model. Is this a valid criticism? Can we construct models that precisely relate to a particular situation, or to all situations?

5. Business and economic models that employ mathematics are occasionally criticized on the grounds that they employ unrealistic assumptions. For example, economists assume that individuals maximize utility. Some criticize this assumption as being unrealistic. Are realistic assumptions necessary when one is using mathematical models?

6. Related to Question 5: For a number of reasons, we should suspect that Stigler's conclusions are wrong. Does that mean that the analysis should be ignored? Explain.

Chapter 2

Variables, Sets, Lists, and Relations

Chapter 1 used Stigler's diet problem to demonstrate that the appropriate use of mathematics can provide valuable insights into an important issue. With mathematics we can isolate and examine the crucial forces operating in an increasingly complex world.

The world's complexity is what drives analysts to use mathematics. At the same time, this complexity bedevils the analyst. The variety of possible interactions in any complex systems means that no decision-maker can consider all of the factors that might influence that decision. Even so, leaving out pertinent information can lead to serious error. It is humanly impossible for any individual to provide a complete description of the richness, complexity, and variety that characterize the world.

Realistic, successful analysis of a problem that faces a decision-maker requires that the analyst isolate the key aspects of reality in that problem. That is, the analyst must abstract and simplify, always taking care to retain those factors that are deemed crucial to the situation at hand. For example, the availability of iron ore is a crucial factor in the production of steel, whereas the religious affiliations of steel workers is probably an irrelevant factor.[1]

[1]We say "probably" because in some contexts this might not be true. If production is a team activity is a team effort and if religious composition affects team performance either by promoting cohesion or by providing insights from various viewpoints, then it should be taken into account. Witness communes like the Amana community.

Skillful analysis of a problem results in a *theory*. A theory is an abstract set of relationships from which we can derive meaningful propositions.

A good theory delineates the crucial forces that are at work in a situation, the circumstances under which those forces are related, the nature of those relationships, and the probable result of the interaction of those forces. Here is a theory of your success: Students who are more intelligent, and who study more, will generally earn higher grades than others not as well equipped or prepared. Such a theory forthrightly states two of the most important factors that determine student grades, and also indicates the relationship between these two factors and grades. This theory is probabilistic, in that it identifies tendencies but does not insist that, for any two individual, the one that is both a bit more intelligent and a bit more diligent will inevitably earn better grades. Furthermore, as stated here this theory says nothing about the interrelationships and tradeoffs between intelligence and hours of study in terms of expected grades.

The language and the component parts of theories require further examination, for it is our ability to construct and use abstract theoretical relationships that determines our ability to make intelligent choices and decisions. This statement is itself a theory. Some might argue that intuition is all that matters: "Good decision-makers are born, not trained." The theory stated here does ignore intuition. It does not, however, deny a role for intuition. Suppose that the analysis defines the general nature of some relationships but does not specify the precise value of some of its parameters. Intuition can be combined with formal analysis at this point.

Analyzing economic and business problems often involves combining several theories into a *model*. We can express an economic model as a series of mathematical equations, but we need not do so. Many models are developed verbally, although such models often suffer from lack of precision. A model identifies the factors and influences that are important in a situation, and delineates the relationships among those factors and influences. A *model* is best thought of as a systematic presentation of interrelated theories.

2.1 Variables

The remainder of this text focuses on mathematical models that relate to economic activities. The analysis is phrased in terms of expression of how

variables relate to each other. Most of these variables, like income or a price, can be quantified. Others, like utility, can be ordered but not quantified in a meaningful way. A *variable* is a quantity that can assume different values at different points of observation.

The magnitude of a variable can assume various values. For example, the gross national product (GDP) of the United States could be 100, 1000, 1500, or, indeed, any positive magnitude. Because a variable's magnitude can assume various different values, a variable must be represented by a general symbol. Hence the price of a pizza might be represented by the symbol p, while the tax rate might be represented by the symbol r. Chapter 1 represented the magnitudes of the 80 different foods as $X1$, $X2$,..., $X80$. The letters at the end of the alphabet, such as u, w, x, y, and z, commonly symbolize the magnitudes of variables. This, however, is a matter of convention rather than of necessity. In *Maxima*, we often use a string of letters to name a variable. Thus *cons* might be the name assigned to consumption.

Variables can be either *cardinal* or *ordinal*. The values assigned to cardinal numbers have meaning: the difference between \$2 and \$5 is \$3. Ordinal variables define order alone: we may judge one person to be friendlier or happier than another but we cannot assign a specific value to the difference in the levels of friendliness or happiness.

We may classify cardinal variables as being either *continuous* or *discrete* in terms of the magnitudes that are permissible for those variables. A continuous variable is one that can assume any value within a given interval of values. Annual income is a continuous variable. A discrete variable is one that can assume at most a limited number of values within a given interval of values. The number of siblings in your family is a discrete variable.

Consider some examples of continuous and discrete variables. The examples reveal that the distinction, while quite real, can sometimes be ignored with impunity. When this distinction can be safely ignored depends, of course, on theoretical considerations.

- Gross Domestic Product (GDP) is a continuous variable in that it can assume any positive magnitude. Hence 1,543.324 billion dollars is one possible magnitude. In practice, however, the actual reports and computations of GDP are typically restricted to values that are truncated to a specific number of decimal places. The value 18036.649... become

$18036.65, a discrete value when GDP is reported in billions of dollars and round to the second decimal place.

- Flipping a coin many times and recording the outcome (1 = heads, 0 = tails) generates values of a discrete value; the value x=22/3, for example, cannot be observed. Despite this fact, applying the normal distribution, which describes a continuous variable, can still provide meaningful analysis of the probabilities associated with the number of heads per toss when the number of tosses becomes large, as you learned in statistics.

- Sometimes it is not clear whether the relevant variable is discrete or continuous. Consider the number of cars in a household. The number is an integer, but how often cars are traded is continuous. At the market level, the relevant variable can be treated as continuous, for two reasons. First, if a small fraction of households change the number of cars per household, the result will be a very small change in the number of new cars purchased. The same is true of a small change in the age at which households trade for a new car. Technically, the number of new cars sold per year is an integer (though the average number sold per month or per day is not). Even so, when the total number sold is in the millions, the relevant variable can be treated as continuous without compromising the analysis.

Variables may also be classified as to whether or not they are endogenous or exogenous in nature. An endogenous variable is one whose value is to be determined inside the model being used. An exogenous variable is one whose value is taken as given; its value is determined by forces that are outside the model.

The magnitude of endogenous variables is explicitly examined and determined by the model. A supply-and-demand model determines the price of goods and services. Price is endogenous variable in such a case. If, however, government mandates a price for the good, then price becomes an exogenous variable in the model.

Recall the simple model of income determination from your macroeconomics principles class. The $Y = C + I + G$ equation (where Y is a measure of national income, C is private consumption expenditures, I is business investment expenditures, and G is governmental purchase of goods and services)

determines national income. In the simplest models, C is considered to be endogenous, while I and G are considered to be exogenous. That is, the value of C is determined inside the model, whereas the values of I and G are taken as given. (The *wxMaxima* workbook that accompanies this section illustrates this example of a model.)

2.2 Equations, Roots, and Constants

Like the simple income determination model above, mathematical models are usually expressed in the form of equations. An equation is a statement that asserts the equality or equivalence of two (or more) mathematical expressions. Each equation must contain at least one variable. For example, the expression $2 \cdot X = 10$, is one statement about the variable X. Of all possible values, 5 is the only one for which this statement to be true.

The previous example, 5 is a *critical root*, or a *solution value*. Critical root(s) or solution value(s) is (are) the value(s) of the variable(s) of an equation that cause(s) the equation to hold true.

Many examples of critical roots (solution values) occur in the field of economics. The equilibrium price that clears the market, the magnitudes of inputs and outputs that maximize profits, and the dollar value of the consumption of private individuals that leads to an equilibrium level of GDP are all examples of critical roots.

Some equations are characterized by mathematical terms that never change in value. You are undoubtedly aware that the value of π (the Greek letter pi) is a constant that is equal to 3.14159.... The value of π never changes. Another example of a constant is e, the base of the system of natural logarithms, which is equal to 2.71828 .... A numerical constant is a magnitude that is fixed and does not change in value. When a constant is joined to a variable, that constant is often referred to as the coefficient of that variable.

Many equations include parameters that act as numerical constants in a limited fashion. A parametric constant or parameter acts as a constant only within the context of a particular equation or problem, but may assume a different constant value in other equations or problems.

An example sharpens the difference between numerical constants and parametric constants. Assume that the number attending a St. Louis Cardinals baseball game is given by the equation $Q = 50,000 - b \cdot P$. The number 50,000 is the stadium's seating capacity and is a constant, at least until major construction occurs. The number attending will generally be values within a given interval of values that cannot exceed 50,000. The number in attendance is less than this 50,000 and depends on the price, P, and the change in Q per one-unit change in price, $-b$. The value of the parameter b depends on many things that are exogenous to this simple theory: the day of the week, whether the game is critical to a pennant race, the identity of the pitcher, and the identity of the visiting team among others.

It is customary to use letters at the beginning of the Latin alphabet (for example, a, b, c, and d) or of the Greek alphabet (α, β, γ, and δ) to symbolize parameters in a particular equation. As an example, or first use of *Maxima*, consider a simplified Keynesian national-income model to illustrate the various terms and definitions that we have developed in this subsection. The cell below defines the four equations and shows the condition that must be met for equilibrium to occur. The first three are simple theories of behavior, so they are behavioral equations. The last of these four equations, entered into the `solve( )` command, is an equilibrium condition. The result is a variable named Yeq, the equilibrium value of Y.

The four commands below state the components of the model and determine the condition for equilibrium as follows. Notice some aspects of this entry. First, we use a `fixed font` to indicate *Maxima* commands, which must consist entirely of text characters. Second, the first three commands end with dollar signs, so that they produce no printed output (though the variables name on the left of the colon in each is assigned to the expression on the right of the colon and these assignments remain in *Maxima*'s memory). The fourth command ends with a semicolon, so that the result of executing this command is printed.

```
(%i) C:a+b*Y$ I:I0$ G:G0$ solnY: solve(Y=C+I+G, Y);
```

$$\textbf{(\%o)}\quad [\mathbf{Y} = -\tfrac{\mathbf{I0+G0+a}}{\mathbf{b-1}}]$$

The result of executing this command is assigned the name `solnY`. [2] The

[2] Any allowable name would do; some names like `value`, are reserved and cannot be applied, though `Value` could be—*Maxima* is case-sensitive.

output is shown as the single item in a list. The textttsolve command always produces a list. *Maxima* follows mathematical conventions, so the output is not always as you might expect to see it. In the example above, we can restate the expression as follows: $Y = (a + I0 + G0)/(1 - b)$, so that the multiplier $1/(1 - b)$ becomes apparent.

Suppose that we have values for the parameters a and b and for the exogenous values $I0$ and $G0$. The following input/output combination shows a set of values and the implied values of Y and C. The first command below assigns the name `Yeq` to the equilibrium income. The second command substitutes a set of parameter values into `Yeq` and assigns the name `Yeq0` to the result. The third command substitutes the parameters and the equilibrium output level into the consumption function, yielding the equilibrium consumption level. The final command provides a check, to confirm that total spending sums to the equilibrium output level.[3]

```
(%i) Yeq:  rhs(solnY[1]);
     Yeq0 :  subst([a = 150, b=0.75, I0=25, G0=20], Yeq);
     C0:subst([a=150,b=0.75,Y=Yeq0],C);      C0+25 +20;
```

(%o) $-\frac{\mathbf{I0+G0+a}}{\mathbf{b-1}}$ (%o) **780.0** (%o) **735.0** (%o) **780.0**

The above four-equation model has two endogenous variables (Y and C) and two exogenous variables (I and G), the values of which are assumed to be determined outside this model. The consumption function illustrates the use of two parameters ($a = 150$ and $b = 0.75$). The values of the exogenous variables are also numerical constants. The final command states and solves the equilibrium condition.

2.3 The Real Number System

As we have seen, variables, constants, and parameters usually take on numeric values. We can classify numeric values in terms of their position on the real number line.

[3]When we refer to variables in general, we use standard mathematical notation, so that the variables appear like this: $C = a + b \cdot Y$). When the variables are names assigned in *Maxima* input, we use the fixed font, like this: `C = a + b*Y`.

Consider the positive integers (1, 2, 3,...), the negative integers (- l, -2, -3, ...), and zero. All these values may be found on the real number line portrayed below. A real number line has the following characteristics: (1) The origin (location of zero) on the real number line is arbitrarily chosen. (2) The units of measurement on the real number line are arbitrarily chosen. (3) A positive or negative direction along the real number line is indicated by the sign of the number; this sign reflects the location of a particular point relative to the origin. (4) The ordering relation among the numbers on the real number line is that, if $x < y$, then the point x lies to the left of point y on the real number line. The number line in Figure 2.1, generated by *Maxima*, shows three integers, the values $-\pi$ and π, the constant e, the fraction 2/3, and the square root of 5. Confirm that these values are in the correct sequence.

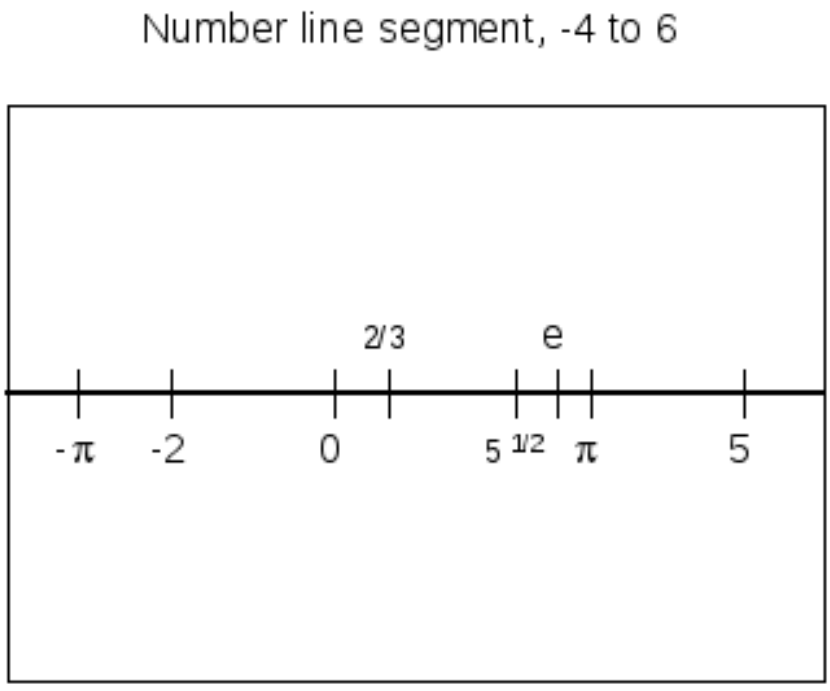


Figure 2.1: Real Number Line, Segment

The gap between any two whole, integer values found on the real number

line may be partially filled with rational numbers. A rational number like 2/3 results from the division of one integer by another, provided that the denominator is not equal to zero. The number 2/3, expressed as the quotient of the integers 2 and 3, is more commonly known as a fraction. Any integer may be expressed as the quotient of some two integers. Therefore every integer is a rational number. For example, $5 = 10/2 = 5/1$.

The remaining gaps on the real number line are filled by irrational numbers. Irrational numbers cannot be expressed as the quotient of two integers. An example is the value of π, which is $3.14159\ldots$. The square root of 5 ($\sqrt{5}$) is another example of an irrational number, $\sqrt{5} = 2.236067\ldots$.

To summarize: A rational number is the quotient of two integers, the denominator not being equal to zero; an irrational number cannot be expressed as the quotient of two integers; and a fraction is a rational number that is not an integer.

The rational and irrational numbers together form the real number system. The one-to-one relationship between the real number system and the real number line means that we may use the terms "real number" and "point" interchangeably. A real number is a point on the real number line. We omit complex numbers, which involve $i = \sqrt{-1}$. Complex numbers can occur in some analysis of dynamic systems, but take us beyond the purview of this text

2.4 Sets and Set Theory

We could paraphrase the preceding paragraph as follows: The set of rational numbers can be combined with the set of irrational numbers to form a larger set, the set of real numbers. In this paraphrase, "set" is used casually. The concept of the set, however, can be defined much more precisely. The theory based on that concept, developed in the latter part of the nineteenth century, forms the foundation of much of modern mathematics. A thorough treatment of set theory would require an enormous amount of work. However, you can acquire a working knowledge of the basic concepts of set theory with considerably less effort.

We begin with a formal definition. A set is a collection of distinct, well-defined objects. Here "well-defined" means that any object either is an in-

stance of the set or it is not, and a condition exists for determining which of these is true. A set may be defined by either the "roster method" or the "set-builder method." Consider a simple set: A = {1, 2, 3, 4, 5}. This is an application of the **roster** (or **enumeration**) method: the elements of the set are listed. Such a set necessarily has a finite number of elements. Notice the notation: A set's name is typically a capital letter, and the rule for identifying its elements is enclosed in curly brackets.

In the **set-builder** (or **definition**) approach, these brackets contain a rule for identifying the elements Consider B = $\{x|0 < x < 100\}$. Read this as "x such that x's value is between 0 and 100 and does not include 0 or 100." Infinitely many real numbers qualify, so a set that is constructed with the set-builder method can (but need not) be infinite.

Maxima uses the roster method, so the sets that it manipulates are finite, though they can be very large. The three commands below build sets A, B, and C. The resulting output appears below the commands. Compare the third command with the third output line. *Maxima* has removed the elements that repeat. It does this because a set consists entirely of distinct elements; repetitions are not allowed. Also observe that *Maxima* writes the elements of set B in alphabetic order, not in the order in which they were entered. An important aspect of a set is the sequence does not matter.

```
(%i) A:1, 2, 3, 4, 5, 6, 7, 8, 9;
     B:red, orange, yellow, green, blue, indigo, violet;
     C:1,2,3,4,5,6,7,8,9,8,7,6,5,4,3,2,1;
(%o) {1, 2, 3, 4, 5, 6, 7, 8, 9}
     {blue, green, indigo, orange, red, violet, yellow}
     {1, 2, 3, 4, 5, 6, 7, 8, 9}
```

Consider one more case, one in which we might be tempted to think that the set contains a single element. We enter these named expressions: `X: a/c + b/c`, `Y: a/c + b/c` and `Z: (a + b)/c` as *Maxima* commands. The first two expressions, assigned the names X and Y are equivalent to each other and have the same form. The third expression, Y, has the same value but *not the same form.* Therefore, it is a distinct element of the set X,Y,Z. Either of these two equivalent commands generates the relevant set: `{X,Y,Z}` or `set(X,Y,Z)`.[4] The result is $\{\frac{b+a}{c}, \frac{b}{c} + \frac{a}{c}\}$.

[4]We have distributed the commands through this paragraph, rather than placing them

An element that is in a set is said to be a *member* of that set. Membership is typically signified with the symbol $\in$. Read A $\in$ G as "set A is an element of set G," or as "set A is a member of set G," or as "set A belongs to set G." The notation $\notin$ indicates that the element does not belong to the set.

The command ***elementp(A,x)*** instructs *Maxima* to determine whether x belongs to A. In the example below, *Maxima* informs us that 5 is an element of A and that 11 is not an element of this set, where `A:{1,2,3,4,5,6,7,8,9}` and the two test commands are these: `elementp(5,A)` yields **true**, and `elementp(11,A)` yields **false**.

An important special case is the **null set**, sometimes called the *empty set*. This set, which contains no elements, is denoted by the symbol $\emptyset$. The names of NBA basketball players whose height is under 5 feet is the null set, as is the set of years in which the real GDP of the United States grew at an annual rate greater than 50 percent.

2.4.1 Set Algebra

Sets can relate to each other in various ways. Consider the following relationships: equality, subsets, union, intersection, universality, and complementarity.

Equality

Two sets $S1$, and $S2$ are said to be equal or identical if and only if $S1$, and $S2$ have exactly the same elements. The next exhibit considers four sets, S, A, B, and C. Sets A, B, and C are subsets of S, but only C contains all of S's elements. Therefore, $C = S$, while $A \neq S$ and $B \neq S$. Note that we have entered some repetitions that do not appear in the first output line, which identifies set S. Each element of a set must be unique.

```
(%i) [S: {1,2,3,4,5,6,7,8,9,9,8,7},
        A:odds(S), B:evens(S), C:S];
(%o) [{1, 2, 3, 4, 5, 6, 7, 8, 9}, {1, 3, 5, 7, 9},
        {2, 4, 6, 8}, {1, 2, 3, 4, 5, 6, 7, 8, 9}]
```

together and following them with output. When this option seems to provide a better flow, we will use it. See the accompanying workbook for the input/output cell.

The foregoing material requires defining a function named `evens()` and another named `odds()`. The definition of these functions, which provides a way for *Maxima* to emulate the set-builder approach to set building, is sketched in the workbook that accompanies this chapter. For a more complete treatement, see [13].

Next we confirm that *Maxima* can discover the relationships between sets that we have asserted. The following commands are entered: `[is(S=S), is(S=C), is(A=S), is(B=S), is(B=C)]`. *Maxima* treats each of these `is( ... )` statements as a condition to be evaluated. The resulting output is the answers that we expect: [**true, true,false, false, false**].

Subsets

Set A is said to be a subset of set S if and only if every element of A also belongs to S. The notation is this: $A \subset S$ is read "Set A is a subset of set S." In the example above, $A \subset S$, $B \subset S$, and $C \subset S$. We confirm these assertions with the following commands (both the commands and the output are entered as lists): `[subsetp(A,S), subsetp(B,S), subsetp(C,S)]`. As expected, the result is [**true, true, true**]: all of the named sets are subsets of S.

Union

A new set may be formed by the union of two sets. Let $S1$ and $S2$ be any two arbitrary sets. The union of $S1$ and $S2$ consists of the elements that are in $S1$, in $S2$, or in both $S1$ and $S2$. The notation is $S1 \cup S2$, which is read "$S1$ union $S2$." In the example above, $A \cup B = C$. In the next example, *Maxima* determines the union of these two sets: all integers from 1 through 10 and the even integers from 2 through 20.

The commands below create a set that consists of the even integers between 1 and 20, set `S1`. The second set, `S2`, consists of the squares of the integers between 1 and 10. The third set consists of the union of `S1` and `S2`.

```
(%i) S1:setify(makelist(i,i,1,10));
     S2:setify(makelist(i^2, i, 1,10));    union(S1,S2);
```

(%o) $\{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}, \mathbf{7}, \mathbf{8}, \mathbf{9}, \mathbf{10}\}$
(%o) $\{\mathbf{1}, \mathbf{4}, \mathbf{9}, \mathbf{16}, \mathbf{25}, \mathbf{36}, \mathbf{49}, \mathbf{64}, \mathbf{81}, \mathbf{100}\}$
(%o) $\{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}, \mathbf{7}, \mathbf{8}, \mathbf{9}, \mathbf{10}, \mathbf{16}, \mathbf{25}, \mathbf{36}, \mathbf{49}, \mathbf{64}, \mathbf{81}, \mathbf{100}\}$

Intersection

Another way to form a new set is via the intersection of two or more sets. The intersection of two sets (the definition easily extends to more than two) $S1$ and $S2$ consists of the elements that are in both $S1$ and $S2$. The notation is $S1 \cap S2$ is read "S1 intersection S2." Formally, $S1 \cap S2$ is equivalent to $\{x|x \in S1 \textit{ and } x \in S2\}$. The intersection of S1 and S2 above is the set {2, 4, 6, 8, 10 }, which executing the command `intersection(S1,S2)` confirms.

The intersection of sets that share no common elements is the *null set*; such sets are said to be *disjoint*. The command `intersection({red, yellow, green}, {up, over, out})` generates this output: `{}`, which is *Maxima*'s notation for the null set.

Set Difference

The union operator defines elements that are members of any of two or more sets. The intersection operator defines elements that two or more sets share in common. A third, related operator is the *set difference operator*, which defines elements that are in the first set but not in the second set. Order matters. A formal definition is this: Given any two arbitrary sets $S1$ and $S2$, the set difference of $S1$ and $S2$ consists of the set of all elements that belong to $S1$ but not to $S2$. Formally, $S1 - S2 = \{x|x \in S1 \textit{ and } x \notin S2\}$.

Consider these three sets: $X : \{1, 2, 3\}, Y : \{3, 4, 5\}$, and $Z : \{1, 2, 5, 6, 7\}$. The command `setdifference(X,Y)` produces the set that consists of the elements that are in X but not in Y: {1,2}. In contrast, `setdifference(Y, X)` produces the set that consists of elements of Y that are not in X: {4, 5}. Other examples based on these sets appear in the workbook that accompanies this chapter.

Multiple Sets

Unlike union and intersection, set difference is a binary operation: it compares two sets. Both union and intersection can be applied to more than two sets. Refer to the three sets above. The union of the three is $X \cup Y \cup Z = 1, 2, 3, 4, 5, 6, 7$. The intersection of these sets is $X \cap Y \cap Z = \{\}$, the empty set.

Universal Sets

A *universal set* includes all elements that are allowed by definition. If the set is potential results of flipping a coin, then the universal set is {heads, tails} (assuming the coin never lands on its edge). Thus, a universal sets is a complete listing of all elements or outcomes that can be associated with a particular action or situation.

If the contents of a set A is known and if the universal set is given, then we can deduce the contents of a second set that *complements* the first set. The complement to set A is $A' = U - A$, where the prime indicates. Alternatively, the complementary set A' is $\{x | x \in U \text{ and } x \notin A\}$.

2.4.2 Set Geometry: Venn Diagrams

The algebraic relationships between sets that can be illustrated visually by means of Venn diagrams, as shown in Figure 2.2 below. This exhibit shows the six relationships that involve any two sets and the universe of which they are a part. The construct of Venn diagrams can extend to any number of sets. The relationships are these:
a) Both A and B are subsets of the universe U, and B is a subset of A. $B \subset A$; $A \subset U$; $B \subset U$.
(b) The union of A and B contains all elements of A and all elements of B: $A \cup B$.
(c) The intersection of A and B contains all element of both A and B. In this case A and B have no common elements; they are disjoint: $A \cap B = \emptyset$.
(d) The intersection of A and B, when A and B contain some common elements: $A \cap B$.
(e) The set difference $A - B$ shows all elements that are in A but not in B.
(f) The complement to A, A', shows all elements of U that are not in A.

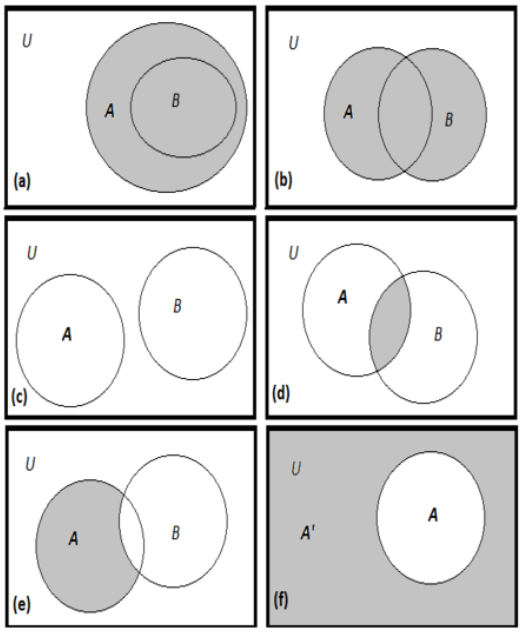


Figure 2.2: Venn Diagrams

Exercise 2 - 1

1. Using set notation, specify each of the following.
   (a) The set of all integers greater than -5 but less than 5.
   (b) The set of all prime numbers from 0 to 25.
   (c) The set of all real numbers greater than 0.
   (d) The set of all even numbers that are also ?prime? numbers, in the sense that they cannot be divided by any integer to obtain another integer.

2. Let $S = \{1, 2, 3\}$, $T = \{3, 4, 5\}$, $V = \{3, 2, 1\}$, and the universal set $U = \{, 2, 3, 4, 5\}$. Which of the following statements are correct? If a statement is incorrect, correct it.

(a) $S = T$	(b) $S = V$	(c) $3 \in S$
(d) $4 \in V$	(e) $S \subset V$	(f)$T \subset S$
(g) $V \not\subset T$	(h) $S \cup T \neq U$	(i) $S \cap T = U$
(j) $V \cap T = \emptyset$	(k) $S \cup V = S$	(l) $U - S = T$
(m) $V' = T$	(n) $U - S = U - V$	(o) $S \cup V \cup T = U$

3. Let $A = \{1, 2, 3\}$, $B = \{2, 3, 4, 5\}$, and $C = \{1, 3, 5\}$. Verify that the following assertions are correct for these sets:
   (a) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, and
   (b) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

4. Use Venn diagrams to verify 3(a) and 3(b).

5. If $A \subset B$ and $C \subset D$, does this mean that $(A \cup B) \subset (C \cup D)$? Explain.

6. If $A \subset B$ and $C \subset D$, does this mean that $(A \cap B) \subset (C \cap D)$? Explain.

7. Use Venn diagrams to show when the following expressions are correct.
   (a) $A \cup B \cup C = C$
   (b)$A \cap B \cap C = C$ and
   (c) $A \cap B \cap C = \emptyset$

8. In Figure 2.3, the universe is U. The sets $S1$, $S2$, and $S3$ are as indicated. Area VI is the difference between U and areas I, II, III, IV, and IV. Identify each of the following.
   (a) The two sets that are disjoint.
   (b) The area(s) corresponding to $S1 \cup S2$.
   (b) The area(s) corresponding to $S1 \cap S2$.
   (c) The area(s) corresponding to $S1'$.
   (d) The area(s)–if any–corresponding to $S1 \cap S2 \cap S3$.
   (f) The complement to the area defined in (d).
   (f) The area(s)–if any–corresponding to $S1 \cup S2 \cup S3$.
   (g) The complement to the area defined in (f).
   (h) $S1 - S3$.
   (i) $S3 - S1$.

2.4.3 Set Theory: The Formal Algebra

We can formally translate the Venn-diagram analysis of the previous section into a series of laws that define the algebra of sets. These laws lack the intu-

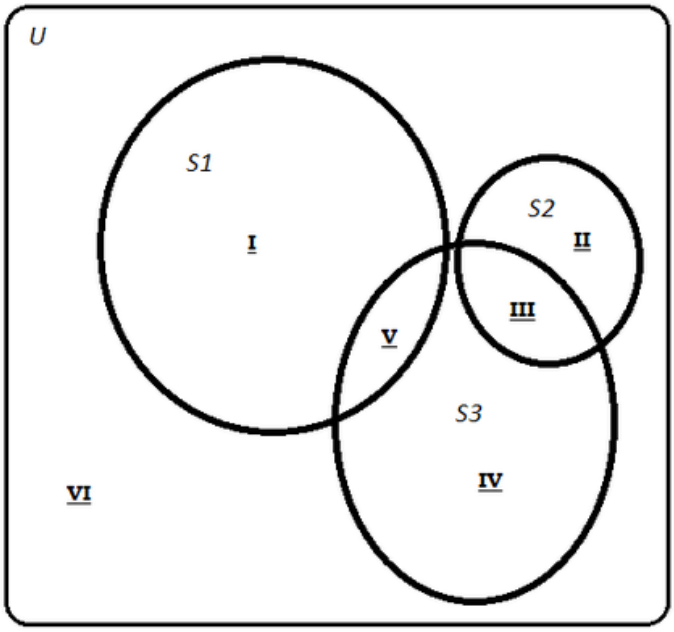


Figure 2.3: Three Sets

itive appeal of the Venn diagrams. However, the fact that they are algebraic rather than graphical in character is advantageous in extended applications of set theory. Throughout, we assume the existence of three sets–A, B, and C–that are subsets of the universal set U. You might recognize the similarity of many of these laws to those that provide the foundation for standard algebra.

- Commutative Laws
  (a) $A \cup B = B \cup A$
  (b) $A \cap B = B \cap A$

- Associative Laws
  (a) $A \cup (B \cup C) = (A \cup B) \cup C$
  (b) $A \cap (B \cap C) = (A \cap B) \cap C$

- Distributative Laws
  (a) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
  (b) $A \cup (B \cap C) = (A \cup B)) \cap (A \cup C)$

- Idempotent Laws
  (a) $A \cup A = A$
  (b) $A \cap A = A$

- Identity Laws
  (a) $A \cup \emptyset = A$
  (b) $A \cup U = U$
  (c) $A \cap U = A$
  (d) $A \cap \emptyset = \emptyset$

- Complement Laws (a) $A \cup A' = U$
  (b) $A \cap A' = \emptyset$
  (c) $(A')' = A$
  (d) $U' = \emptyset$
  (e) $\emptyset' = U$

- DeMorgan's Laws
  (a) $(A \cup B)' = A' \cap B'$
  (b) $(A \cap B)' = A' \cup B'$

2.4.4 Ordered and Unordered Pairs

In set theory, the two-element set $\{x, y\}$ is equal to the two-element set $\{y, x\}$. That is, $\{x, y\} = \{y, x\}$. The pair (x, y) is therefore said to be an *unordered pair*, in that the ordering is irrelevant. Contrast this to an ordered pair, in which $(x, y) \neq (y, x)$ and the ordering of elements x and y is crucial. More formally, given two elements x and y, a pair (x, y) is said to be an *ordered pair* if $(x, y) \neq (y, x)$ unless $x = y$.

As an example in which ordering matters greatly, consider the ordered pairs consisting of the wins followed by the losses of an athletic team. For example, the ordered pair $(2, 12)$ would represent the 2-win, 12-loss record, which is quite different than a very successful $(12 - 2)$ record. Ordered pairs are enclosed in parentheses–$(2, 12)$ is not the same as $(12, 2)$–and unordered pairs appear within curly brackets–{2,12} is the same as {12,2}.

We can extend the concept of ordered elements in order to distinguish between ordered and unordered triples, quadruples, and so forth. Consider a list of the last four U. S. Presidents of the twentieth century.They are Carter,

Reagan, Bush, and Clinton. As members of a set, the following would be true {Reagan, Bush, Carter, and Clinton} = {Clinton, Bush, Reagan, Carter}.

2.5 Lists

A historian would not consider the two orderings in the set of presidents to be the same. They would insist on a list of the presidents: [Reagan, Bush, and Clinton] would be a chronological list; some other ranking might result in a different list. A *list* is an ordered n-tuple of elements, and any of these elements may consist of text strings, numbers, mathematical expressions, sets, or other lists. Lists are the basic building blocks for computer algebra systems like *Maxima*. The command `pList: [Carter, Reagan, BushI, Clinton,BushII,Obama]` produces a list of presidents and assigns it the name `plist`. With this information in Maxima's memory, the command `plist[3]` produces the output `BushI`.[5]

Lists can be used to assign names to expressions. These expressions can involve computation or they can consist of strings. Also, a list can contain another list or a set. The following set of input and resulting output shows a four-item list. Each item in the list is bound to a member of a set of four names. The command that creates the lists ends with a $, so that printing is suppressed. The individual items are then recalled and printed. The four commands in the second input line result in the four lines of output, one for each of the named items in the list.

```
(%i) [a,b,c,d]:["Some text", {x[1],x[2]}, [p,q], log(10.0)]$
     a; b; c; d;
```

(%o) Some_text (%o) $\{\mathbf{x_1}, \mathbf{x_2}\}$ (%o) $[\mathbf{p}, \mathbf{q}]$ (%o) **2.30258509299**

2.5.1 Creating Lists with Commands

Often rather than entering the items in a list by hand, it is safer and easier to create the list using the `makelist()` command. This command, which we used earlier, requires an expression in terms of a counter variable (any name

[5] We repeat that, unlike sets, lists allow multiple copies and that order matters.

will do), then the counter variable itself, and finally a start and end value (additional arguments can be inserted; see the *Maxima Manual*). Each of the three commands below creates a list, to which a name has been assigned. The names can be used to recall the list or items in the list.

```
(%i) xList:  makelist(i,i,0,10);
     sqrtList:  makelist(sqrt(x),x,0,10);
     halfList:  makelist(x/2, x, 0, 10);
```

(%o) $[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$
(%o) $[\ 0, 1, \sqrt{2}, \sqrt{3}, 2, \sqrt{5}, \sqrt{6}, \sqrt{7}, 2^{\frac{3}{2}}, 3, \sqrt{10}\ ]$
(%o) $[0, \frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, 3, \frac{7}{2}, 4, \frac{9}{2}, 5]$

Maxima contains many built-in operations like `sqrt()`, `sin()`, and `log()` that can be applied directly to the items in a list, as below. Also, some operations like dividing all items by the same constant or raising them to the same power can be applied directly to a list. The next three commands show a quick way to create the items in `sqrtList` and `halfList`, along with a way to square each item in `xList`.

```
(%i) sqrt(xList); xList/2; xList^2;
```

(%o) $[0, 1, \sqrt{2}, \sqrt{3}, 2, \sqrt{5}, \sqrt{6}, \sqrt{7}, 2^{\frac{3}{2}}, 3, \sqrt{10}]$
(%o) $[0, \frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, 3, \frac{7}{2}, 4, \frac{9}{2}, 5]$
(%o) $[0, 1, 4, 9, 16, 25, 36, 49, 64, 81, 100]$

Some operations cannot be applied to a list this way. *Maxima*'s `map()` command can be applied in such cases. This command can also be used instead of some of those that we have already seen. The next input/output group shows how to apply some basic operations to two lists of equal length (be aware that you must avoid illegal operations like dividing by zero). Note the use of quotation marks to indicate the binary operation that is being conducted.

```
(%i) A: [2,4,6,8]$ B:[1,5,7,9] map("+",A,B); $
     map("-",A,B);map("*",A,B);map("/",A,B); map("^",A,B);
```

(%o) $[3, 9, 13, 17]$ (%o) $[1, -1, -1, -1]$
(%o) $[2, 20, 42, 72]$ (%o) $[2, \frac{4}{5}, \frac{6}{7}, \frac{8}{9}]$
(%o) $[2, 1024, 279936, 134217728]$

The first command below takes the factorial of each term in `xList`. The `map()` command requires the following: a specification of the operation to be applied, in quotation marks, and the list or lists to which the operation applies.

Operations or functions can be mapped onto a single function. The next input/output group shows two ways to return the factorial of the values in `xList`. The first approach directly applies the factorial command, !. The second approach is to create a named function `fact(x)` and then to map that function onto the values. Be aware of the slight differences in the syntax.

```
(%i) map("!",xList);
     fact(x):=factorial(x)$ map(fact, xList);
(%o) [1, 1, 2, 6, 24, 120, 720, 5040, 40320, 362880, 3628800]
(%o) [1, 1, 2, 6, 24, 120, 720, 5040, 40320, 362880, 3628800]
```

To illustrate the second approach in more detail, we create a function with the command below and than map the function onto `xList`.

```
(%i) f(x):=a*x^b$ exprList:  map(f, xList);
(%o) [0, a, a 2^b, a 3^b, a 4^b, a 5^b, a 6^b, a 7^b, a 8^b, a 9^b, a 10^b]
```

The subst command can be used to determine the values of the expressions above, given specified values of the parameters a and b.

```
(%i) valueList:  subst([a=5,b=2], exprList);
(%o) [0, 5, 20, 45, 80, 125, 180, 245, 320, 405, 500]
```

We assigned the list a name, `valueList`, so individual items can be extracted. To extract the fourth item in `exprList` and its counterpart in `valueList`, we use the command `[exprList[4], valueList[4]]`. Note the use of brackets to indicate the item number. Placing a set of commands inside a list causes *Maxima* to place the output into a list: $[a\,3^b, 45]$.

2.5.2 Describing a List

Specific information about a list is often useful. For example, if the list is long, we might wish to know the number of items that it contains. We might wish to know the sum of the items in the list, or their mean value, or their standard deviation. The commands below result in output that shows that `xList` contains 11 items, the sum of which is 55, and the mean of which is 5. The (sample) standard deviation is $\sqrt{11}$.

```
(%i) xLength:  length(xList);
     xSum:sum(xList[i],i,1, xLength); xMean:(xSum/xLength);
     sqrt(sum((xList[i]-xMean)^2,i,1,xLength)/(xLength-1));
```

(%o) **11** (%o) **55** (%o) **5** (%o) $\sqrt{\mathbf{11}}$

To determine the minimum or maximum value, we must use the `apply` command, as below. The workbook provides details regarding this operation.

```
(%i) [apply(min,xList), apply(max,xList)];
```

(%o) $[\mathbf{0}, \mathbf{10}]$

Maxima offers a module, `descriptive`, that computes these values. The purpose of doing this by applying the formulas directly is to illustrate the manipulation of lists. Note that the names for the minimum and maximum values are `smin` and `smax`. If you work much with large data lists, then descriptive will be of value.

```
(%i) load(descriptive)$
     mean(xList); std1(xList); smin(xList); smax(xList);
```

(%o) **5** (%o) $\sqrt{\mathbf{11}}$ (%o) **0** (%o)**10**

2.5.3 A Matrix as a List of Lists

In general, a matrix is a list of lists, all of which must have the same length. Thus a matrix is rectangular. Chapter 9 details the uses of matrices to

conduct mathematical analysis. Here, we limit their use to creating tables. Consider the simple example above, in which we listed the minimum and maximum values of x. With a matrix, we could add a list of names to these values, as shown below.

```
(%i) matrix(["Minimum x", "Maximum x"],
     [apply(min,xList), apply(max,xList)]);
```

$$(\%o)\ \begin{bmatrix} \textbf{Minimum x} & \textbf{Maximum x} \\ 0 & 10 \end{bmatrix}$$

2.6 Relations

Any ordered pair (triple, quadruple, and so forth) of values constitutes a relation. Given the ordered n-tuple $(x, y, z, \ldots)$ a relation among the variables exists whenever every set of values for any of $n-1$ of the variables implies one or more values for the remaining variable. For example, suppose that $x + 2 \cdot y - 1.5 \cdot z^2 = 0$. Then specifying values for any two of the variables implies one or more values for the third.

Using the command `expr:x + 2*y - 1.5*z^2 = 0` we enter this expression. We then assign values to two of the variables and solve for the third variable. Assigning values to x and y can result in more than one z value. Assigning values to x and z implies a single y value. Likewise, assigning values to y and z implies a single x value.

To support these assertions, we use three commands that solve the expression for one of the two variables in terms of the other two. The `subst` commands below determine the implications of given values of y and z for x, of x and z for y, and of x and y for z. For x and y, singe values result; but for z, a list of two values is reported.

```
(%i) expr:  x + 2*y - 1.5*z^2 = 0;
     subst( [y=4, z = 3], solve(expr, x) );
     subst( [x=8, z=4], solve(expr, y) );
     subst( [x=1, y=3], solve(expr, z) );
```

$$(\%o)\ -\mathbf{1.5\,z^2} + \mathbf{2\,y} + \mathbf{x} = \mathbf{0} \quad (\%o)\ [\mathbf{x} = \tfrac{\mathbf{11}}{\mathbf{2}}]]$$

$$(\%o)\ [\mathbf{y} = \mathbf{8}] \quad (\%o)\ [\mathbf{z} = -\tfrac{\sqrt{\mathbf{2}}\sqrt{\mathbf{7}}}{\sqrt{\mathbf{3}}}, \mathbf{z} = \tfrac{\sqrt{\mathbf{2}}\sqrt{\mathbf{7}}}{\sqrt{\mathbf{3}}}]$$

2.7 Questions and Problems

1. Some historians claim that they do not use models in writing history and in arriving at conclusions about historical phenomena. Is it possible to analyze something without having an underlying model? Will hard work produce insights and generalizations if you do not have a model? Explain.

2. Indicate whether the sets of numbers below can be properly described as being any or all of the following: integers, fractions, rational numbers, real numbers, irrational numbers.
   (a) {-5, -1, 2, 4}
   (b) {4/3,1/2,-3/8,11/12}
   (c) $\{\sqrt{(2)}, \sqrt{(3)}, \pi, \sqrt{(11)}\}$.

3. Refer to the three sets in (2). Suppose that we replace {} with [], indicating that the three quadruplets are list, not sets. How would this change affect the interpretation of the values?

4. Apply the *Maxima* command `sort` to these lists: (a) [-5, -1, 2, 4], (b) [4/3,1/2,-3/8,11/12], and (c) $[\sqrt{(2)},\sqrt{(}3),\pi,\sqrt{(}11])$. What do you perceive is *Maxima*'s default direction of sorting? If you're curious about reversing this order, see the `sort` command in the manual. (In *wxMaxima*, click on the word `sort` in any command and hit the F1 key.)

Chapter 3

Rectangular Coordinates and Functions

This chapter builds on material from Chapter 2. Chapter 2 shows that variables may be related via mathematical expressions. This chapter focuses on a subset of those expressions, functions. It examines three types of functional relationships. The first, explicit functions, exist when the value of some variable is determined by an explicit relationship between that variable and the value(s) of one or more other variables. The second, implicit functions, exist when a set of two are more variables must jointly satisfy the condition that a mathematical expression imposes. We saw an example of an implicit function at the end of Chapter 2. Finally, two or more variables' values can be bound by the fact that each of these variables is functionally related to another set of one or more variables, which are called *parameters*. Our variables are said to be related via parametric equations (sometimes called freedom equations).

Much of our illustrative analysis involves just two variables. In some cases, the important relationship actually involves just two variables. In other cases, the relationship can involve more than two variables, but the method of approach can be outlined with the two-variable case and then extended. Because of the importance of the two-variable case, this section begins with an extension of the number line that Chapter 1 developed, showing two variables' values simultaneously with the use of rectangular coordinates. Occasionally, we extend the analysis to include a third dimension. Furthermore, the reasoning involved in creating rectangular coordinates in a plane can be

extended to any dimension, a topic that we address in Chapter 7 and again in Chapter 9, which treats linear algebra.

3.1 Rectangular Coordinates

The concepts of the real number line and ordered pairs enable us to the *rectangular* (or Cartesian) *coordinate system*. Suppose that we have two real number lines that are perpendicular to each other. The two dotted lines in the next figure are suitable examples. The intersection of these two real number lines is designated the origin of our coordinate system. The horizontal line called the x axis) and the vertical line (called the y axis) together form the coordinate axes.

Before treating the nature of the rectangular coordinate system more completely, we address a few details regarding the commands and the resulting output. The `draw2d` command generates Figure 3.1, a two-dimensional figure with rectangular coordinates. The "`wx`" prefix instructs `wxMaxima` to place the resulting graph in the workbook. Setting the `xaxis` and `yxaxis` options to true produces the two dotted lines that define the quadrants of the space. We name the axes (using `xlabel` and `ylabel`) and turn off the listing of values (using `xtics` and `ytics`). Then we set the ranges for X and Y. Finally, we create four labels that describe points in the quadrants. Each `label` command contains some text and the coordinates to which the labels attach.

```
(%i) wxdraw2d( title="Four Quadrants",
     xaxis=true, yaxis=true, xlabel="X", ylabel="Y",
     xtics=false, ytics=false,xrange=[-6,6],yrange=[-6,6],
     label(["*I (+,+)",5,5]),label(["*II, (-,+)",-5,5]),
     label(["*III (-,-)",-5,-5 ]),
     label(["*IV (+,-)",5,-5]))$
```

Both of the coordinate axes have the basic properties of any real number line. Points to the right of the origin on the, x axis, and upward on the y axis indicate positive values; points to the left of the origin on the x axis and downward on the y axis represent negative values.

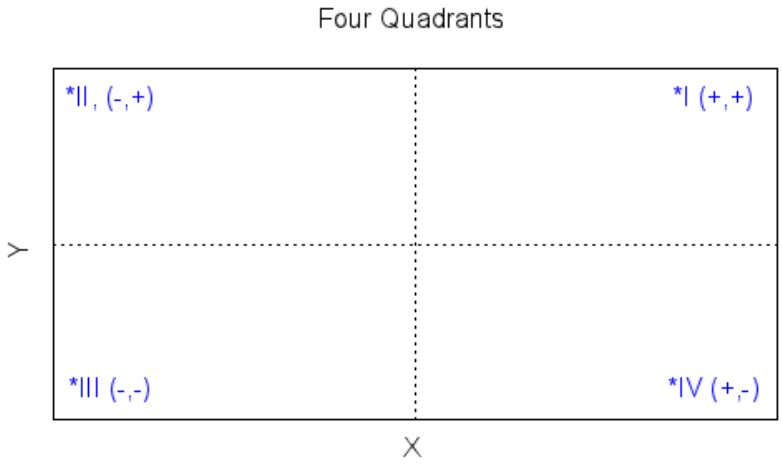


Figure 3.1: Rectangular Coordinates: The Four Quadrants

Each ordered pair of real numbers is represented by a unique point on the plane that is formed by the x and y axes. Examples appear in Figure 3.1 as asterisks. Given an ordered pair (a, b), the x coordinate, or *abscissa* of the variable x on the x axis, is always the first element in the ordered pair. The second element of the ordered pair is the y coordinate, or *ordinate* of the variable y on the y axis. This means that the point given by the ordered pair (a, b) is not the same as the point given by the ordered pair (b, a) unless $a = b$.

The x and y coordinates that comprise an ordered pair indicate the location of the point given by that ordered pair. As Figure 3.1 demonstrates, when the elements of ordered pair (a, b) are both positive, then the point that corresponds to this ordered pair lies in the first quadrant, that is, the area of the coordinate system that lies to the right of the y axis and above the x axis.

When the signs of the ordered pair (a, b) are (-, +), then the point in question lies in the second quadrant, to the left of the y axis and above the x axis. When the signs of the ordered pair (a, b) are both negative, then the point lies in the third quadrant, to the left of the y axis and below the x axis. Finally, when the signs of the ordered pair (a, b) are (+, -), the point lies in the fourth quadrant, to the right of the y axis and below the x axis. By convention, quadrant numbers are indicated by Roman numerals.

The two real number lines that form the coordinate system need not have the same units of measurement. That is, the units in which they are measured need not be the same. Indeed, often they cannot be the same. One axis might measure profit and the other axis number of units sold, so the units for one is dollars per time period and the other is some measure of physical units per time period (the two time periods need not be the same). Or one axis might measure price in terms of dollars per unit and the other axis quantity in terms of physical units. This latter example describes the coordinate axes that we use to conduct supply-and-demand analysis.

3.2 Functions

Consider these two equations: $y = x^2$ and $y^2 = x$. Both are equations, but for only one of the two is y a *function* of x. When $y = x^2$, each value of x implies a single value of y. Such is not the case when $y^2 = x$, for in that case a particular x value can be related to either of two y values. For example, when $x = 4$, y can be either -2 or +2, for squaring either yields a value of 4 and satisfies the equation. We summarize this introduction with a formal definition: A *function* is a relation (a set of ordered pairs) such that no two ordered pairs have the same first element.

A function is denoted by $y = f(x)$, which is read "y is a function of x" (not "y equals f times x"). Other functional notations that are frequently used include $g(x)$, $h(x)$, and $F(x)$. To create a functional expression in *Maxima*, use notation like this: `f(x, a) := (a*x^2)`. The next cell shows this expression when a's value is not specified, again when each of two values of a is specified, and finally when both a and x are specified. All five commands are placed into a list, so the output also appears in a list. Be aware that the values entered must be entered in the order specified in parentheses, x first and then a.

```
(%i) [f(x,a) := a*x^2, f(x,3), f(5,3), f(x,-3), f(5,-3)];
```

$$\textbf{(\%o)}\quad [\mathbf{f}(\mathbf{x},\mathbf{a}) := \mathbf{a}\,\mathbf{x}^{\mathbf{2}},\quad \mathbf{3}\,\mathbf{x}^{\mathbf{2}},\quad \mathbf{75},\quad -\mathbf{3}\,\mathbf{x}^{\mathbf{2}},\quad -\mathbf{75}]$$

A function is just a special case of a relation. A function exists when no two ordered pairs have the same value for x, but different values for y. The

definition of a function implies that it must be a relation. The reverse, however, is not true. Not every relation is a function, since in a relation there may be several ordered pairs that exhibit the same value of x, but different values of y. Functions are therefore a subset of relations.

A different letter must be used to denote each function used in a particular problem. For example, if quantity demanded q_d and quantity supplied q_s are different functions of price p, then the functional notation used must reflect the fact that two different functions exist. Hence one might write $q_d = f(p)$ and $q_s = g(p)$.This indicates that the two functions are not equivalent, that is, that $f(p) \neq g(p)$ in general. Of course for one p value, the one that corresponds to market equilibrium, $f(p) = g(p)$.

Given the functional relationship $y = f(x)$, the value of variable y depends on the value of variable x. Once x takes on a particular value, y's value is determined. Variable y therefore depends on x, and is referred to as the *dependent variable*, whereas variable x is the *independent variable*. A function like $y = f(x)$ is often referred to as an *explicit function*: y's value is explicitly related to x's value in a way that $f(x)$ defines. Later in this section, we consider two other ways that relationships between x and y might be specified.

3.2.1 Domain and Range

Often only certain values are permissible for both the independent and dependent variables. Most relationships in business and economics involve real numbers, not imaginary values. Some, like units of output, cannot assume negative values; others, like profits, can assume negative values. Let output be defined by $f(x)$ and profits by $g(x)$. A formal statement of the limits that we observed is that the domain of $f(x)$ consists of 0 and the positive real numbers. The profit function $g(x)$ is limited to the real numbers, its domain. In general, given the $y = f(x)$, the set of all values that x may assume is the *domain* of the function.

Likewise, sometimes only certain values are permissible for the dependent variable. Let $y = x^2$.Then, regardless of what real number value the independent variable x assumes, the value of y cannot be negative. The range of variable y is prescribed and limited. In general, given $y = f(x)$, the set

of all values that variable y may assume is referred to as the *range* of the function.[1]

3.2.2 Implicit Functions and Parametric Equations

As noted, for a function of the form $y = x^2$, y is an explicit function of x. The dependent variable y has a value that is uniquely defined by a value of the independent variable x. We also noted that $y^2 = x$ is not such a function. We can, however, express the latter as y^2–$x = 0$ and assign that expression the name $F(x, y)$. This is an *implicit* function. It expresses a relationship between x and y but does not assign roles of dependence to either.

Yet another possibility is that x and y are determined by equations that involve a third variable, which we label t. The value of t can be treated as a *parameter* in the two equations and by letting t vary over some range, we can deduce the behavior of x and y and, therefore, how they relate to each other. We need not, and perhaps cannot, deduce an equation that relates x and y. An importanat economic example is this: a large set of production functions allow for both output and cost to be stated in terms of the employment level of a single input but does not generate an expression of output in terms of cost or *vice versa*.

The graph below shows how `draw` handles an explicit function and an implicit function.[2] For the first function, $y = x^2$, the domain is the entire real number line and the range is the nonnegative portion of the real number line. The second graph shows the implicit function y^2–$x = 0$. For this implicit function, x can take on only nonnegative real numbers. This function implies that $y = \sqrt{x}$, which involves imaginary numbers for $x < 0$. The values of y can range over the entire real number line.

```
(%i) first:gr2d(title="y = x^2",explicit(x^2, x, -5,5))$
     second:gr2d(title = "y^2 = x",
```

[1] In the `draw` command in *Maxima*, the options `xrange` and `yrange` define the ranges over which values are to be graphed. Thus, xrange must be within the function's domain and yrange must be within the function's range, as those terms are defined in the text.

[2] The two `gr2d` commands create scenarios that are assigned names. The result of each is that an object is stored in *Maxima*'s memory. Executing `wxdraw` with the object names inside the parentheses causes *Maxima* to graph the two named scenarios. As an exercise, replace `gr2d` with `wxdraw2d` and graph each of these expressions separately.

```
     implicit(y^2=x,x,0,5,y,-sqrt(5),sqrt(5)))$
wxdraw(first,second, columns=2)$
```

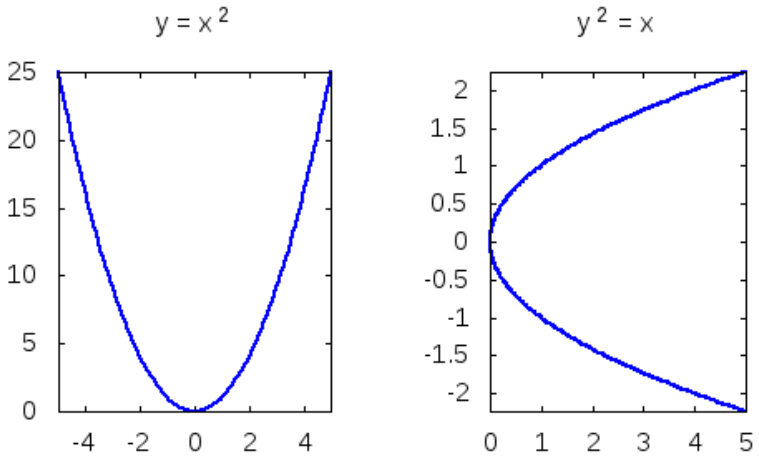


Figure 3.2: Explicit and Implicit Functions

Exercise Set 2.1

1. For each of the following, determine the range of the dependent variable y and the domain of the independent variable x.
   (a) $y = 8 + x$ (b) $y = \sqrt{x}$ (c) $y = \sqrt{(4 - x^2)}$
   (d) $y = \frac{1}{x^2-1}$ (e) $y = \frac{1}{8-x}$

2. Write out a few (a, b) that are consistent with the following expressions. Indicate which expressions can be cast as explicit functions with y as the dependent variable.
   (a) $y = x^2$ (b) $y = x^4$ (c) $y^2 = x$ (d) $y^3 = x$ (e) $y^4 = x$
   (f) $y = \sqrt{(x)}$ (g) $y = 1/x$ (h) $y = \pi \cdot x^2$ (i) $x^2 + y^2 = 4$
   (j) $y + x^2 = 1$ (k) $y^2 + x = 9$ (l) $x = 3$ (m) $y = 1/2$

3. Use the `wxdraw2d` command to graph the expressions in (h) and (k) from the list above.

4. The total revenue, which is defined as $TR = P \cdot Q$ of a firm per day, is a function of its daily sales Q. Assume that the firm's output capacity is 10 to units per day. What are the domain and range of TR if the price is defined as $P = 1200/\sqrt{(Q)}$? (This is the *inverse demand curve*.)

5. A supply curve is a functional relationship between quantity supplied and price. Graph the following supply schedule. Be careful when you label your axes.

```
(%i)  Qlist:["Q, units per week :"
          ,1000,2000,4000,7000,11000];
      Plist:  ["P, $ per week :  ", 5,6,7,8,9];
      matrix(Qlist,Plist);
```

$$(\%o)\ \begin{bmatrix} \textbf{Q, units per week}: & 1000 & 2000 & 4000 & 7000 & 11000 \\ \textbf{P, \$ per week}: & 5 & 6 & 7 & 8 & 9 \end{bmatrix}$$

3.2.3 Composite Functions

Given two functions $f(x)$ and $g(z)$, it may be possible to create a new function by means of an operation known as *composition*. Given $y = f(x)$ and $x = g(z)$, the *composite function* or composition of $f(x)$ and $g(z)$ is the function

$h(z)$ for which $y = f(g(z)) = h(z)$. The functional relationship $y = f(g(x))$ is read as, "the function f of a function g, of z," and indicates that variable y is a function of variable x, which itself is a function of variable z. To evaluate the function $h(z)$,one must first compute $g(z)$and then evaluate $f(x)$ at the point $g(z)$.The function $h(z)$ is defined only at those points z for which $g(z)$ is in the domain of $f(x)$.

EXAMPLE: Given that $y = f(x) = \sqrt{x}$ and $x = g(z) = z + 1$, the composite function is given by $y = f(g(z)) = \sqrt{z+1}$.

Note that the function $g(z)$ is defined for all real values of z, whereas the function $f(x)$ is defined only for those values such that $x > 0$. This means that the composite function $f(g(z))$ can be satisfied only when $(z + 1) \geq 0$. (The symbol $\geq$ means, "greater than or equal to," while the symbol $\leq$ means "less than or equal to.")

WARNING: Be careful in reading the composite function notation: $f(g(z))$ is read "f of g of z." It is not the product $f(x) \cdot g(z)$. Therefore, the composite function $f(g(x))$ is $\sqrt{z+1}$; it is not $\sqrt{x} \cdot (z + 1)$.

The input/output cell below illustrates the fact that *Maxima* understands composite functions.

```
(%i) [f(x):=sqrt(x), g(z):=z+1, f(g(z))];
```

$$(\%o)\quad [\mathbf{f}(\mathbf{x}) := \sqrt{\mathbf{x}},\quad \mathbf{g}(\mathbf{z}) := \mathbf{z}+\mathbf{1},\quad \sqrt{\mathbf{z}+\mathbf{1}}]$$

Using the function notation shown above, `f(x) :=` an expression and greatly facilitate the evaluation of a function for a number of values. Suppose that we wish to evaluate for values of x from 51 through 58, or for a single value of x. The first command below creates a list of x values and binds that list to a name. The second command creates the functional expression. The third command applies the function to the list of x values and assigns the result to the name `yList`. The the `matrix` command creates a table of values. The final command applies the function to a single value, $x = 81$. The results appear as a table and as a single value, $5 \cdot 9 = 81$.

```
(%i) xList:makelist(50 + 1.0*i,i,1,)$ f(x):=5*sqrt(x)$
yList:  f(xList)$ matrix(xList,yList); f(81);
```

$$(\%o)\quad \begin{bmatrix} 51.0 & 52.0 & 53.0 & 54.0 & 55.0 & 56.0 & 57.0 & 58.0 \\ 35.707 & 36.055 & 36.4 & 36.742 & 37.08 & 37.416 & 37.749 & 38.078 \end{bmatrix}$$

$$(\%o)\quad \mathbf{45}$$

Exercise Set 2.2 Evaluate these expressions by hand and again with *Maxima*.

1. Given that $f(x) = 100 + 7x$ find (a) $f(0)$, (b) $f(5)$, and (c) $f(-10)$.
2. Given that $f(x) = 10 - 4x$ find (a) $f(1)$, (b) $f(10)$ (c) $f(a + h)$.
3. Given that $f(x) = x^2 + 4x - 6$ find (a) $f(0)$, (b) $f(10)$, (c) $f(-2)$.
4. Given that $f(x) = 1/x^2$ find (a) $f(2)$, (b) $f(-4)$, (c) $f(x + h)$.
5. Given that $f(x) = 2^x$ find (a) $f(0)$, (b) $f(3)$, (c) $f(-3)$.
6. Given that $f(x) = x^2 - 2x + 2$ show that $f(-2) \neq -f(2)$.
7. Given that $f(x) = x^2$ show that $f(x + h) - f(x) = h(2x + h)$.
8. Given that $f(x) = (1 + x)/(1 - x)$ show that $f(1/x) = -f(x)$ and $f(-1/x) = 1/f(x)$.
9. Given that $f(x) = x?(x - 1)$ show that $f(x + 1) = f(-x)$.
10. Given that $f(x) = x^2 + 8x - 3$ and $g(z) = 2$ find $f(g(z))$.
11. Given that $f(x) = 4x - x^2$ and $g(z) = 1/x$ find $f(g(z))$ and $g(f(x))$.

3.2.4 Functional Forms

This section discusses specific functional forms that we use subsequently. For now, we limit the discussion to functions of two variables, typically of the form $y = f(x)$.

Polynomial Functions

A polynomial is a very general functional form that can represent many relationships. A polynomial function $y = f(x)$ is defined for all real values of x by an equation of the form $y = a_0 \cdot x^0 + a_1 \cdot x^1 + a_2 \cdot x^2 + \cdots + a_n \cdot x^n$. Because $x^0 = 1$ and $x^1 = x$, we can rephrase this expression in this more

familiar form: $y = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \cdots + a_n \cdot x^n$. Often, this more compact representation is useful:

$$y = f(x) = \sum_{i=0}^{n} a_i \cdot x^i.$$

The symbols, a_0, a_1, ..., a_n are coefficients of the polynomial.

It can be useful to refer to polynomials in terms of their degree. The **degree of a polynomial function** is the highest nonnegative integer power of any independent variable in the polynomial. Consider these examples:

- $y = 5 \cdot x + 3$ is a first-degree polynomial.
- $y = 5 \cdot x^2$ is a second-degree polynomial.
- $y = 5 \cdot x + 5 \cdot x^2$ is also a second-degree polynomial.
- $y = 5 \cdot x + 5 \cdot x^2 + 5 \cdot x^5$ is a fifth-degree polynomial.

Exponents and Bases

In considering polynomial functions, we introduced the concept of the degree, or highest power, of a polynomial. The power to which a variable (or a number) is raised is referred to as the *exponent* of that variable or number. The expression $x \cdot x \cdot x \cdot x \cdots x^n$ represents the product of n x's, all of which have the same value. The variable x is referred to as the *base* of the term x^n, and the letter n is called the exponent of variable x.

The cell below states some of the laws that govern the use of exponents and gives an example of each of these laws in operation. Using the operations that are reported in the output, use *Maxima* to determine the results of the following: $(x^m)^n$, x^m/x^n, $(x/y)^n$, $(x \cdot y)^{1/n}$, $x^{1/n}/y^{1/n}$, $x^{m/n}$.[3]

[3]We repeat a note made earlier: We use the `matrix` command throughout the book to create tables. Chapter xx address the applications of `matrix` to mathematical analysis. Notice that a matrix is made up of a number of lists, all of which have equal length. The lists can consist of values, strings of text, or (as above) combinations of the two.

Note the present of `radcan` in the third example. Copy this cell, remove the `radcan` —don't forget to remove both parentheses—and see what difference occurs. The term `radcan` stands for radical canonical.

```
(%i) matrix(["Operation", "General Result", "Example"],
     ["x^0",x^0, concat("8^0 = ",8^0)],
     ["x^m*x^n",x^m*x^n,x^5*x^3] ,
     ["(x*y)^n",radcan((x*y)^n), (x*y)^3] ,
     ["x^(-n)",x^(-n),x^-2],["x^(1/n)",x^(1/n),x^(1/2)]);
```

Operation	General Result	Example
x^0	1	8^0 = 1
x^m*x^n	x^{n+m}	x^8
(x*y)^n	$x^n y^n$	$x^3 y^3$
x^(-n)	$\frac{1}{x^n}$	$\frac{1}{x^2}$
x^(1/n)	$x^{1/n}$	$\sqrt{x}$

Specific Polynomial Functions

Figure 3.3 shows an example of four specific types of polynomial functions, all of which will be used in examples in this book and all of which are frequently used to analyze and illustrate economic and business issues. The general expressions for the four are these: constant, $y = x^0$; linear, $y = a + b \cdot x$; quadratic, $y = a + b \cdot x + c \cdot x^2$; and cubic, $y = a + b \cdot x + c \cdot x^2 + d \cdot x^3$.

```
(%i) wxdraw2d(xaxis=true,yaxis=true, yrange=[-15,15],
     key="A constant function", explicit(3, x, -5, 5),
     color = black, key="A linear function",
     explicit(2 + 0.75*x, x, -5, 5), color=red,
     key="A quadratic function",
     explicit(-2 - 1.5*x + 0.25*x^2, x, -5, 5),
     color = gray40, key="A cubic function",
     explicit(-5 + 0.5*x - 0.1*x^2 + 0.05*x^3, x, -5, 5))$
```

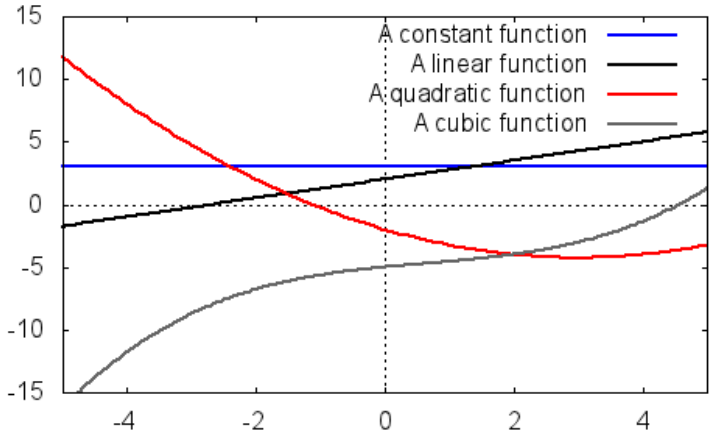


Figure 3.3: Four Polynomial Functions

We have already seen examples of constant and linear functions, both in the simple Keynesian model. We will encounter all of these functional forms as we encounter various models, such as those of cost curves, that require the additional flexibility that terms with larger exponents provide.

Determining a Polynomial's Coefficients from Points

If we can observe any two points on a straight line, then we can compute the line's slope and intercept using techniques that you learned in high-school algebra. Likewise, if we can observe any three points, then we can compute the coefficients of a quadratic equation. Finally, any four points determine the coefficients of a cubic equation.[4] The output below shows an example of using *Maxima* to determine the coefficients of a quadratic equation and plotting that equation.

```
(%i) [x0,y0]:[10,5]$ [x1,y1]:[20,50]$ [x2,y2]:[30,45]$
     solve([y0=a+b*x0 + c*x0^2, y1=a+b*x1+ c*x1^2,
     y2=a+b*x2+c*x2^2],[a,b,c]);
     expression:  subst(%[1], y = a+b*x+c*x^2 ) ;
```

[4]Generally, if the number of points equals the degree of the polynomial *plus* 1, then the polynomial's coefficients can be computed.

```
wxdraw2d( xaxis=true,title="A Quadratic Equation",
  explicit(rhs(expression),x,0,40), point_type=
  filled_diamant,point_size = 2, color=red,
  points([ [x0,y0],[x1,y1],[x2,y2] ]));
```

(%o) $[[\mathbf{a} = -\mathbf{90}, \mathbf{b} = \mathbf{12}, \mathbf{c} = -\frac{\mathbf{1}}{\mathbf{4}}]]$ (%o) $\mathbf{y} = -\frac{\mathbf{x^2}}{\mathbf{4}} + \mathbf{12\,x} - \mathbf{90}$

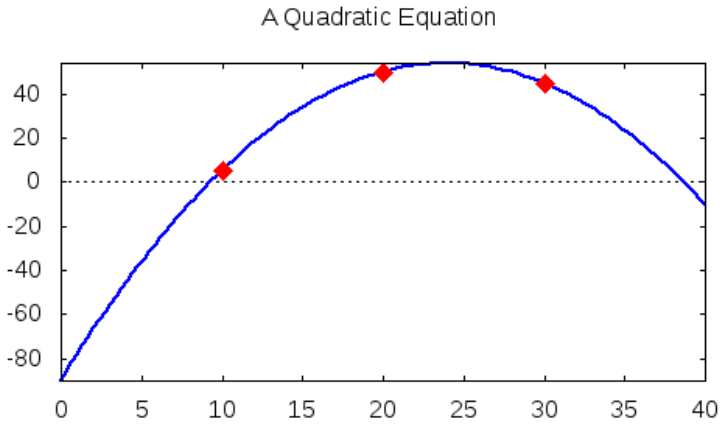


Figure 3.4: Fitting a Cubic Polynomial to Three Points

The values of the (x, y) pairs were selected arbitrarily. For each x value, this must be true: $y = a + b \cdot x + c \cdot x^2$, so the first list in the `solve` command is a list of three points at which that statement must be true if a, b, and c are indeed the proper coefficients. The second list is the list of unknowns for which we seek values. The `solve` output consists of a list with another list embedded. The expression that we name `expression` is obtained by extracting the inside list, by using the [1] after the %. The % symbol refers to the output that results from the most recent command. Thus we are inserting this list: $[a = -90, b = 12, c = -1/4]$ into the expression for the quadratic equation. As exercises, repeat this process in order to define the coefficients of a linear polynomial and a cubic polynomial and to graph the implied polynomials.

Determining the Coefficients from Points

If we can observe any two points on a straight line, then we can compute the line's slope and intercept using techniques that you learned in algebra. Like-

wise, if we can observe any three points, then we can compute the coefficients of a quadratic equation. Finally, any four points determine the coefficients of a cubic equation. The output below shows an example of using *Maxima* to determine the coefficients of a quadratic equation and to plot that equation.

```
(%i)  [x0,y0]:[10,5]$ [x1,y1]:[20,50]$ [x2,y2]:[30, 45]$
      solve([y0=a+b*x0 + c*x0^2, y1=a+b*x1+ c*x1^2,
      y2=a+b*x2+c*x2^2],[a,b,c]); expression:  subst(%[1],
      y = a+b*x+c*x^2 );
      wxdraw2d( xaxis=true,title="A Quadratic Equation",
        explicit(rhs(expression),x,0,40),
        point_type=filled_diamant,point_size = 2, color=red,
        points([ [x0,y0],[x1,y1],[x2,y2] ])); $
```

(%o) $[[\mathbf{a} = -\mathbf{90}, \mathbf{b} = \mathbf{12}, \mathbf{c} = -\frac{\mathbf{1}}{\mathbf{4}}]]$ (%o) $\mathbf{y} = -\frac{\mathbf{x^2}}{\mathbf{4}} + \mathbf{12\,x} - \mathbf{90}$

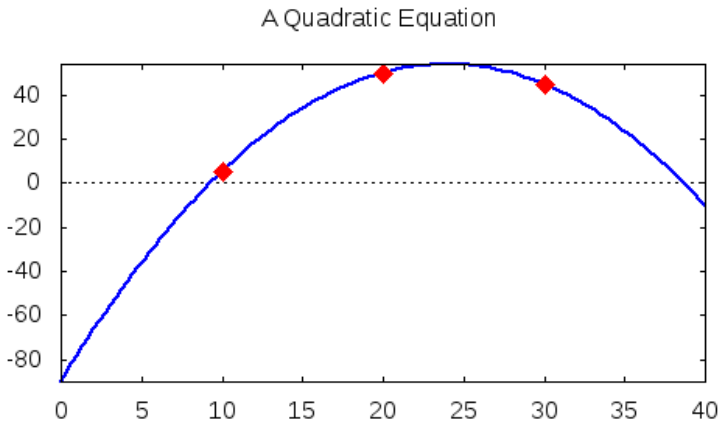


Figure 3.5: Fitting a Polynomial Through Three Points

The values of the (x, y) pairs were selected arbitrarily. For each x value, the following must be true: $y = a + b \cdot x + c \cdot x^2$. Therefore, the first list in the `solve` command is a list of three points at which that statement must be true if a, b, and c are indeed the proper coefficients. The second list contains unknowns for which we seek values, a, b, and c. The `solve` output consists of a list with another list embedded. The expression that we name `expression` is obtained by extracting the inside list, by using the [1] after

the %. The % symbol refers to the output that results from the most recent command. Thus we are inserting this list: `[a = -90, b = 12, c = -1/4]` into the expression for the quadratic equation. As exercises, repeat this process in order to define the coefficients of a linear polynomial and a cubic polynomial and to graph the polynomials.

Sometimes, we know the value at a point and we know (or have a good estimate of) the slope at that point and we wish to use that information to determine the linear function that passes through that point. The cell below shows an example, where a line with a slope of -2.25 passes through the point $(20, 100)$. The first command solves the relevant equation and assigns the solution the name `soln`. The second command substitutes $a = 145$ into the expression $y = a - 2.25 \cdot x$ to yield the expression for the linear equation. The `wxdraw2d` command draws the line over a range of x values, yielding Figure 3.6. The drop line shows the designated point through which the line passes.

```
(%i) [soln:  solve(100 = a -2.25*20,a),
     expr:subst(soln, a - 2.25*x)];
     wxdraw2d( xaxis=true, xlabel="x",ylabel="y",
      key=string(expr),explicit(expr,x,0,70), color=black,
      line_width=1, line_type = dots, points_joined=true,
      points([[0,100],[20,100],[20,0]]) )$
(%o) [[a = 145], 145 - 2.25 x]
```

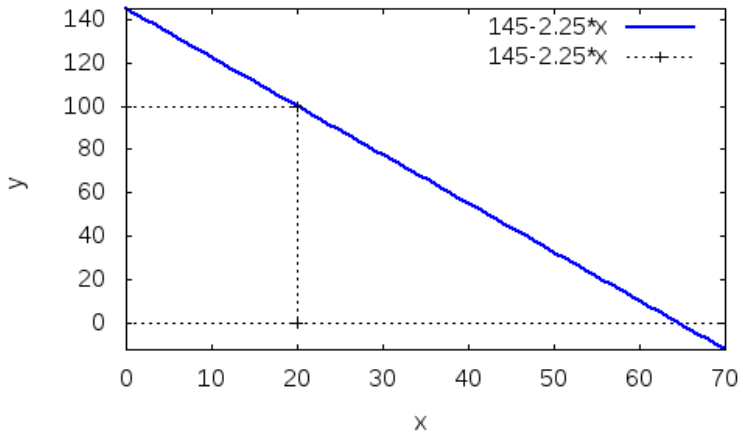


Figure 3.6: A Fitted Line

More on Quadratic Functions

The quadratic form is general enough to apply to many economics and business issues, so examining it in more detail is warranted. We begin with the *quadratic formula*, which is shown below.

```
(%i) solve(a + b*x + c*x^2, x);
```
$$(\%o)\quad [x = -\frac{\sqrt{b^2-4ac}+b}{2c},\quad x = \frac{\sqrt{b^2-4ac}-b}{2c}]$$

This formula provides a rule to find the roots of a quadratic equation. That is, is shows the values of x for which $y = 0$.[5] We see that two such values seem to occur. We say "seem to" because our attention is limited to real numbers. If $b^2 - 4 \cdot a \cdot c < 0$, then no real solutions occur. Also, if only positive x values make economic sense, then the number of solutions can be 0, 1, or 2. The examples below show three quadratic equations. The first, named `p1`, has no real solutions; the second, `p2`, has two real solutions but only one for $x > 0$; and the third, `p3`, has two real solutions for positive values of x. numbers. The term $b^2 - 4 \cdot a \cdot c$ is called the *discriminant*: it discriminates between equations with no real solutions and those with at least one real solution.

```
(%i) [p1, p2, p3]:  [50 - 5*x + 0.5*x^2, 10 + x/3 - x^2/3,
       10 - 10*x + x^2];
     wxdraw2d( xaxis=true, xlabel="x", ylabel = "y",
      key = "No real solution", explicit(p1, x, -5, 10),
      color=red, key = "One positive solution",
      explicit(p2, x, -6, 10), color = black,
      key = "Two positive solutions",
      explicit(p3,x,-5,10))$
```
$$(\%o)\quad [0.5\,x^2 - 5\,x + 50,\quad -\frac{x^2}{3} + \frac{x}{3} + 10,\quad x^2 - 10\,x + 10]$$

An Economic Application: Total and Marginal Cost Curves

A firm's cost curve is often represented by a cubic equation like the one graphed below. If $TC = 50 + 20 \cdot x - 2 \cdot x^2 + 0.25 \cdot x^3$, then its marginal cost

[5] Factoring a quadratic or completing the square are two other ways to find its root. Every quadratic equation can, however, be solved by using the quadratic formula.

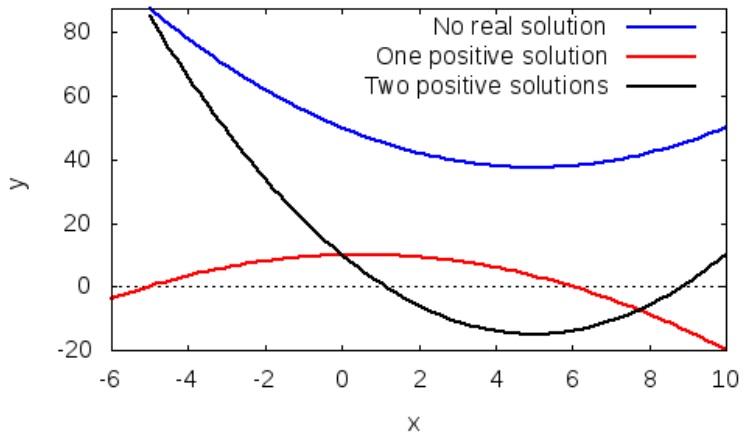


Figure 3.7: Three Quadratic Equations

curve is $MC = 20 - 4 - x + 0.75 \cdot x^2$. Figure 3.8 shows this illustration of how the curves relate to the quantity produced, x. Average cost is simply TC/x.

```
(%i) total:gr2d(xlabel="x",ylabel="$", yrange=[0,350],
     ytics=100,key="Total Cost",
      explicit(50+20*x-2*x^2+0.25*x^3,x,0,10 ) )$
     perunit:  gr2d( xlabel="x" ,ylabel="$ per unit",
      yrange=[0,60], ytics=20, key="Marginal Cost",
      explicit(20 - 4*x + 0.75*x^2,x,0,10 ),
      color=red, key="Average Cost",
      explicit((50+20*x-2*x^2+0.25*x^3)/x,x,0,10))$
     wxdraw(total, perunit, dimensions = [480,480])$
```

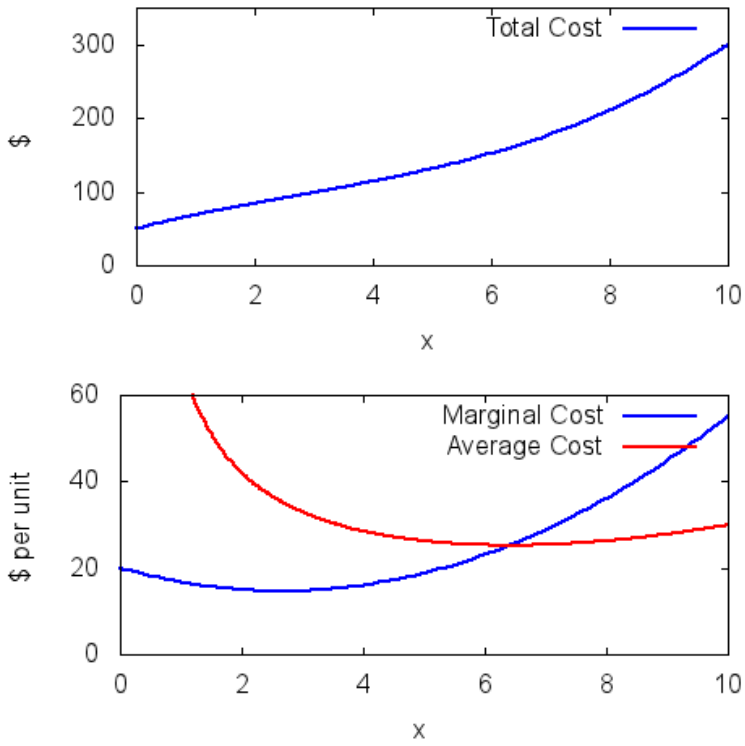


Figure 3.8: Cost Curves

Exercise Set 3.5

Graph the following functions and determine their critical values. Solving first can guide the drawing of the graph in that it indicates the range of x values for which $f(x)$ is to be graphed.

1. $x^2 - x - 6 = 0$
2. $x^2 - 25 = 0$
3. $x^2 + 6 \cdot x + 8 = 0$
4. $3 \cdot x^2 + 7 \cdot x - 3 = 0$
5. $x^2 - 4 \cdot x + 4 = 0$
6. $x^2 + x - 12 = 0$

7. $x^2 - 5 \cdot x + 3 = 0$

8. $x^3 - 5 \cdot x^2 - x + 5 = 0$

Exponential and Logarithmic Functions

Polynomials are useful in conducting analysis, but often other functional forms serve better. Two related classes of functions, exponential functions and logarithmic functions, often serve to represent the relationships that represent economic phenomena.

First, consider exponential functions. Earlier in this chapter, we introduced the concept of a power function, which we defined as a variable raised to a constant power, for example, $y = x^2$ or $y = x^(1/2) = \sqrt{x}$. We extend our use of exponents to the case in which the exponent itself may be a variable, for example, $y = 3^x$ or $y = a \cdot 3^{1/x}$. In such cases, the base (which is 3) is fixed and the exponent contains the variable. In general, an exponential function has the form, $y = a \cdot b^x$ where b is the fixed base such that $b > 0$, $b \neq 1$, and x is an independent variable that is any real number.

The first restriction on b's value, is to preclude roots that are imaginary numbers. The second restriction, reflects that fact that 1, raised to any power, is still 1. Figure 3.9 shows the values of y for a range of x values given three values of b: 0.5, 0.75, 1.25, and 1.5. You should experiment with other values. Why do all of the curves pass through (0,1)? Observe that all of these functions' values are in the first and second quadrants: for all x values.

```
(%i) wxdraw2d(user_preamble="set key left", xlabel="x",
       ylabel="y = b^x", key = "b = 0.5",
       explicit(0.5^x, x, -2, 5), color=red,key =
       "b = 0.75", explicit(0.75^x, x, -2, 5), color=black,
       key = "b = 1.25", explicit(1.25^x, x, -2, 5), color=
       orange, key= "b = 1.5", explicit(1.5^x,x,-2,4),
       color=black,line_width=1,line_type=dots,
       key="(0, 1)", points_joined=true,
       points([[-2,1],[0,1],[0,0]]), dimensions=[480,360])$
```

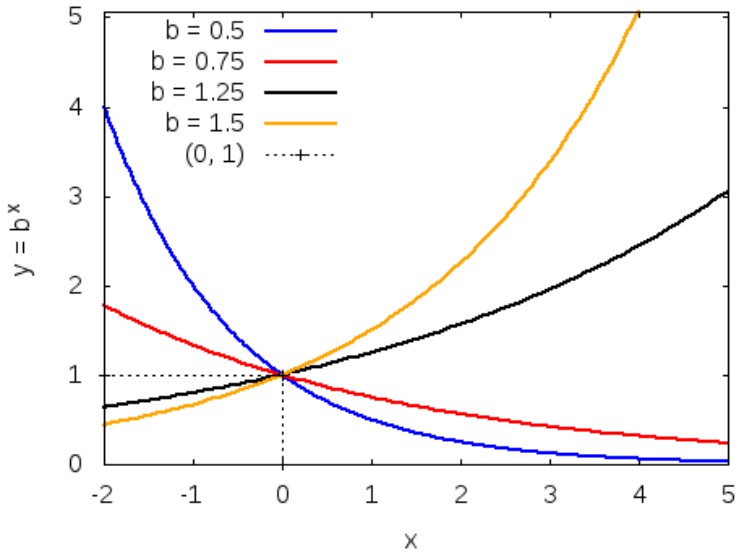


Figure 3.9: Four Exponential Functions

An especially important value of b is the "Euler number," 2.718 . . . , the base of the natural logarithms system. Consider the expression $y = (1 + 1/x)^x$. The increments to 1, which are $1/x$, become increasingly small as x increases. At the same time, however, the size of the exponent increases in inverse proportion to the change in the increment size.1 The table shows that $f(x)$ converges on the value of e as the increment size becomes small and the exponent becomes large). This process involves the concept of a limit, which Chapter 4 develops more formally.

```
(%i) (x):=(1 + 1/x)^x$ xList:  [1,2,3,10,20, 100]$
     iList:1/xList$ yList:float(map(f,xList))$
     matrix(cons("Value of x:  ",xList),
       cons("Increment size:  ", iList),
       cons("Value of y:  ", yList));
```

(%o)

Value of x :	1	2	3	10	20	100
Increment size :	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{10}$	$\frac{1}{20}$	$\frac{1}{100}$
Value of y :	2.0	2.25	2.3703	2.5937	2.6532	2.7048

The *logistic function* is an important case of an exponential function. This function has been used to model growth of forests and wildlife, and it has

been applied to learning curves and to the fraction of a population that has adopted a new product or technology. The function is $f(t) = K/(1+e^{1+a+b\cdot x})$. The value K defines the upper limit. The parameters a and b ($b < 0$) determine the function's curvature and height properties. Euler's number is e, and t is time. The cell below defines a function for which $K = 0.8$, $a = 0$, and $b = -0.2$, shows some values of that function, and graphs the function. The `map` command is used to map the function onto the list of t values.

```
(%i) f(t):= 0.8/(1 + exp(-0.2*t))$
     tList:  makelist(t-10, t, 0, 40,5)$
     yList:  map(f, tList)$
     matrix( cons("t",tList), cons("y",yList) );
     wxdraw2d( xlabel="t",ylabel="y", yrange=[0,1],
      yaxis=true, key="A Logistic Curve",
      explicit(f(t),t,-10,30 ) )$
```

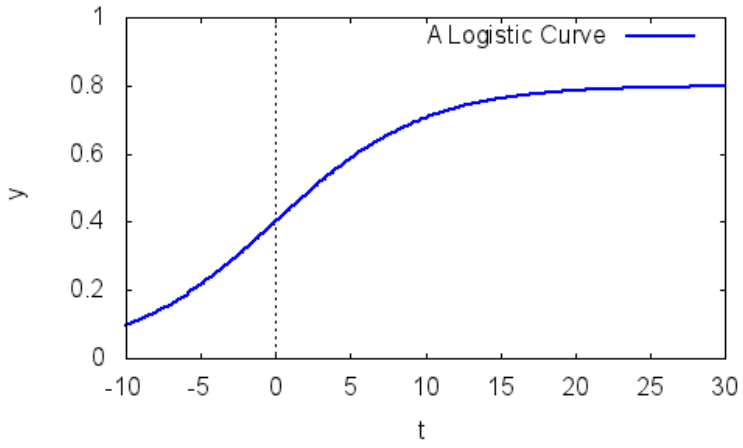


Figure 3.10: A Logisitic Curve

For this set of values, the population has been growing and has reached one-half its maximum size when observations begin ($t = 0$). When y is relatively small, the per-period growth rate is high and, at first, increasing. Then the growth rate decelerates, asymptotically approach zero. In this illustration y is only very slightly less than K, its maximum value, by year 30.

We can generalize the simple exponential function by attaching a coefficient a to the base b and another coefficient c to the variable x. Now our expression is $y = a \cdot b^{c \cdot x}$. The graphs below show that a affects the function's height; specifically a is the function y-intercept. The coefficient c affects the curvature.

```
(%i) f(x,a,b,c) := a*b^(c*x)$
     avals :gr2d(title="a*b^c*x; b=3, c=1", yaxis=true,
      xlabel="x",ylabel="y", yrange=[0,6], xtics=2,
      key="a = 1",explicit( f(x,1,3,1),x,-2,2), color=red,
      key="a=2",explicit(f(x,2,3,1),x,-2,2) )$
     cvals :  gr2d(title="a*b^c*x; a=1, b=3", yaxis=true,
      xlabel="x",ylabel="y", yrange=[0,6], xtics=2,
      key="c = 1",explicit(f(x,1,3,1),x,-2,2), color=red,
      key = "c = -1/2",explicit(f(x,1,3,-1/2),x,-2,2) )$
     wxdraw(avals, cvals, columns=2)$
```

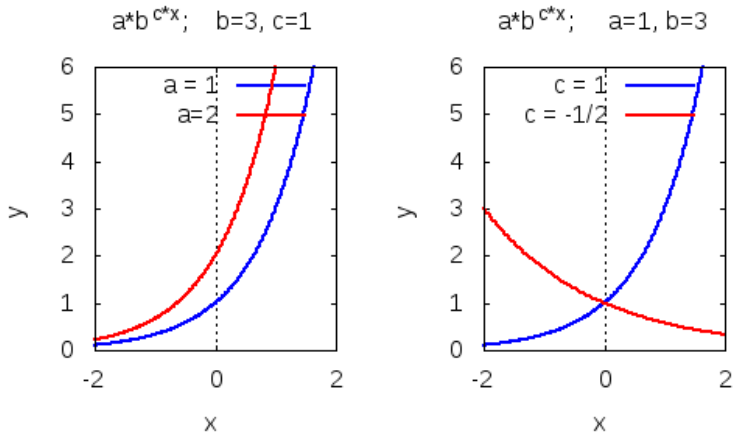


Figure 3.11: Four Exponential Functions

Exercise Set 2.6

1. Sketch on the same set of axes the graphs of the exponential functions $y = b^x$ for these values of b: 4, 8, and 12.

2. Sketch on the same set of axes the graphs of the exponential functions $y = a \cdot b^x$, with $a = 5$ and $c = 1$, for these values of b: 0.3, 1, and 7.

3. Draw the graph of this logistic curve $y = \frac{3}{1+e^{1-2 \cdot x}}$. Graph the function for $-2 \leq x \leq 5$.

Logarithmic functions are closely related to exponential functions. An exponential function of the form $y = f(x) = b^x$ is a "one-to-one" function in the sense that for each value of variable x there is one and only one value of variable y, and *vice versa*. Any function that exhibits a one-to-one relationship also has associated with it an inverse function. We develop inverse functions systematically later in this chapter. Logarithmic functions can be defined in terms of exponential functions, as follows: Given $x = b^y$ where $b > 0$ and $b \neq l$, we refer to $y = log_b x$ as the logarithmic function of x to the base b.

This definition implies that the *logarithm* of a number is the *exponent* to which a *base* must be raised in order to yield the original number. In general, it is true that $x = b^b \Leftrightarrow log_b x = y$. To repeat, logarithm, base, and exponent relate as follows: the *logarithm* of some variable x to the *base* b is the power to which we must raise the base b in order to yield the value x.

EXAMPLES: The following common logarithms are used frequently.

1. $10^0 = 1 \Leftrightarrow log_{10} 1 = 0$
2. $10^1 = 10 \Leftrightarrow log_{10} 10 = 1$
3. $10^2 = 100 \Leftrightarrow log_{10} 100 = 2$
4. $10^3 = 1000 \Leftrightarrow log_{10} 1000 = 3$
5. $10^{-3} = 0.001 \Leftrightarrow log_{10} 0.001 = -3$
6. $10^{-2} = 0.01 \Leftrightarrow log_{10} 0.01 = -2$
7. $10^{-1} = 0.1 \Leftrightarrow log_{10} 0.1 = -1$

The allowable bases for logarithms are positive, so raising a positive number to any power yields a positive number. This fact prohibits logarithms of negative numbers. We can extract the logarithm of quite small positive

values like 0.0001, for which the base-ten logarithm is -3, as Example 5 above shows.[6]

The logarithmic function permits any positive real number as the base b. Even so, almost all analysis is conducted with $b = 10$ or $b = e \approx 2.718$. Ten is the base for a system called "common logarithms," and e is the base for a system called "natural logarithms" or "Napierian logarithms" (Napier is the mathematician who developed this system).

A widely-used convention in science texts is to use the notation $logx$ to refer to common logarithms and lnx to refer to natural logarithms. In contrast mathematics textbooks use $logxto$ refer to natural logarithms and $log_{10}x$ to refer to refer to common logarithms. We use the mathematics convention with one modification, the use of parentheses, $log(x)$ rather than $logx$. This modification accommodates the way the *Maxima* works: it can take logarithms of expressions as well as numbers, so `log` is a command for which the argument must be entered into parentheses. The result of the commands log(100), log(100.0) and log(100.0)/log(10.0) appear below.

```
(%i) matrix(["log(100)","log(100.0)","log(10.0)",
     "log(100.0)/log(10.0)"],[ log(100),log(100.0),
     log(10.0),log(100.0)/log(10.0)]);
```

$$\text{(\%o)} \begin{bmatrix} \mathbf{log}(\mathbf{100}) & \mathbf{log}(\mathbf{100.0}) & \mathbf{log}(\mathbf{10.0}) & \mathbf{log}(\mathbf{100.0})/\mathbf{log}(\mathbf{10.0}) \\ \mathbf{log}\,(100) & 4.6051 & 2.3025 & 2.0 \end{bmatrix}$$

The command `log(100)` results in what appears to be the command itself. Actually, *Maxima* has evaluated this value and returned its exact representation, which is kept in *Maxima*'s memory. Entering 100.0, rather than 100 causes `Maxima` to report a floating-point representation of the natural logarithm of 100.0. Finally, we see that $log(100.0)/log(10.0)$ is the *common logarithm* of 100.0.

If you require common logarithms, it is easy to define a function that provides these values, as the next cell shows. We name the function that returns some values of $log_{10}(x)$. The `float` command ensures that the results are reported as floating-point values. To show how this function works, we apply it to the values in Example 1 above, all of which appear in xList.

[6] Actually, *Maxima* does evaluate the logarithms of negative values, but the results are imaginary numbers.

(%i)
```
log10(x) := float(log(x)/log(10))$ matrix(
xList:[.001, .01, .1,1,10,100,1000], log10(xList));
```

(%o)
$$\begin{bmatrix} 0.001 & 0.01 & 0.1 & 1 & 10 & 100 & 1000 \\ -2.9999 & -1.9999 & -0.999 & 0.0 & 1.0 & 1.9999 & 2.9999 \end{bmatrix}$$

The cell below shows some important laws that govern the behavior of logarithms. The command `logexpand:super` forces *Maxima* to expand some of the expressions rather than just evaluating them and reporting them as they were originally expressed. The last column applies for natural logarithms, reflecting the general relationship that whenever $x = b^y$, then $log_b x = y$.

(%i)
```
logexpand:super$ exprList:[a*b,a/b,1/b,a^b,1/a,exp(a)]$
logList:log(exprList)$ matrix(exprList,logList);
```

(%o)
$$\begin{bmatrix} a\,b & \frac{a}{b} & \frac{1}{b} & a^b & \frac{1}{a} & e^a \\ \log(b)+\log(a) & \log(a)-\log(b) & -\log(b) & \log(a)\,b & -\log(a) & a \end{bmatrix}$$

The relationship between the graph of an exponential function such as $y = b^x$ and its inverse function $y = log_b x$ can be illustrated graphically. The graph below shows this symmetry, using the exponential function $y = e^x$ and its inverse function $y = log(x)$.

(%i)
```
wxdraw2d(xaxis=true, yaxis=true, yrange=[-5,5],
user_preamble="set key bottom",xlabel="x",ylabel="y",
 key="y=x",explicit(x,x, -5, 5) , color = red,
 key="y=e^x", explicit(exp(x), x, -5, 5),
 line_width=3, color=orange, key="y = log(x)",
 explicit(log(x),x, - 5, 5))$
```

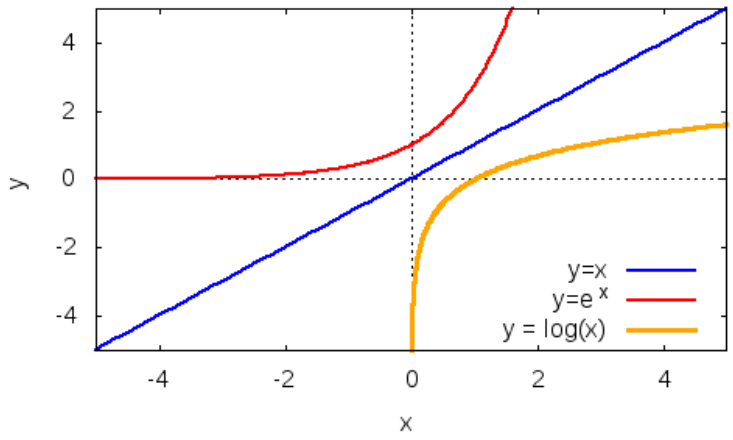


Figure 3.12: Exponential and Logarithmic Functions

Exercise Set 3.7

1. Express the following logarithmic functions in terms of exponents, and the exponential functions in terms of logarithms. (a) $6^2 = 36$ (b) $10^4 = 10000$ (c) $log_{10}(0.0001) = -4$ (d) $log_2(8) = 3$ (e) $4^3 = 64$ (f) $log_3(27) = 3$ (g) $log_2(x) = 9$ (h) $8^{1/3} = 2$ (i) $5^{-2} = 1/25$ (j) $log_2(64) = x$

2. Determine the range of values of x for which the following functions are defined. (a) $y = log(x + 8)$ (b) $y = 1og(8 - x)$ (c) $y = log(x^2 - 4)$ (d) $y = log(25 - x^2)$ (e) $y = log(x - 9)$ (f) $y = log(x^3 + 8)$

3. Using the laws concerning the use of logarithms express each of the following as a single logarithm: $log(x) + log(y) + log(z)$.

4. Use the relationship $log(e^x) = x = e^{log(x)}$, where $log(x)$ is the natural logarithm, to simplify the following. (a) $e^{log(x^5)}$ (b) $e^{log(5 \cdot x^5)}$ (c) $log(e^{x^2})$ (d)$log(\frac{x^2}{e^2})$ (e) $e^{2 \cdot log(5 \cdot x)}$ (f) $e^{-log(x^2)}$ (g) $log(4 \cdot x^3) \cdot e^{x^5}$ (h) $e^{3 \cdot log(x^2) + x^5}$

5. Pareto's law of the distribution of incomes says that the fraction of individuals N from a given population whose incomes exceed x dollars

is given by $N = a/x^b$. Pareto suggested that the value of b was approximately 1.50. If $a = 40,000$, what is the number of individuals whose incomes exceed $x = \$60,000$?

6. Given the general exponential form $y = f(x) = a \cdot b^{c \cdot x}$, graph (and compare) the shapes of the curves as follows:
(a) Select values for a and c such that $a > 0$, $c > 0$. Now graph the function for four different b values: 0, 0.5,1, 1.5. Place all inside the same `draw`-generated graph. Label each curve and select a different color for each. Set the line width to 2. Note that $b = 0$ is not allowed as the base for an exponential function. The graph will reveal why.
(b) Select values such that $0 < b < 1$, and $c > 0$. Draw two curves: for $a < 0$ and for $a > 0$. Place both inside the same `draw`-generated graph. Label each curve and select a different color for each. Set the line width to 2.
(c) Select values such that $a > 0$ and $b > 0$. Draw three curves: for $c < 0$, for $c = 0$, and for $c > 0$. Follow the instructions in (a).
[Create this expression in wxMaxima: `f(x,a,b,c) := a*b^(c*x)`. For the graphs, let x range from -2 to 2 in all cases. Insert `yrange=[-0.5,2]` and `line_width=2`.The last two insertions produce the output in a relatively bold relief. Label the x and y axes.]

Inverse Functions

Previous sections contain references a function that is the *inverse* of another function. We now refine our notion of the character of an *inverse* function. Consider a function $y = f(x)$ that is strictly increasing or decreasing for all values of x from some point a to another point b. A function increases (decreases) throughout this range of values is said to a *monotonically increasing* (decreasing) function. A function $f(x)$ is said to be monotonically increasing (decreasing) if $f(x_j) > f(x_i)$ ($f(x_j) < f(x_i)$) whenever $x_j > x_j$. The graph below shows two linear functions. Both are monotonic over the range depicted. Indeed these function are monotonic over the entire real number line.

```
(%i) [expr1, expr2]:[ 4*x - 5, 25 - 4*x]$
     increase:  gr2d(xaxis=true,yrange=[-10,30],xlabel="x",
```

```
 ylabel="y", key=string(expr1),explicit(expr1,x,0,8))$
decrease:  gr2d( xaxis=true,yrange=[-10,30],xlabel="x",
 ylabel="y",key=string(expr2),explicit(expr2,x,0,8))$
wxdraw(increase,decrease,columns=2)$
```

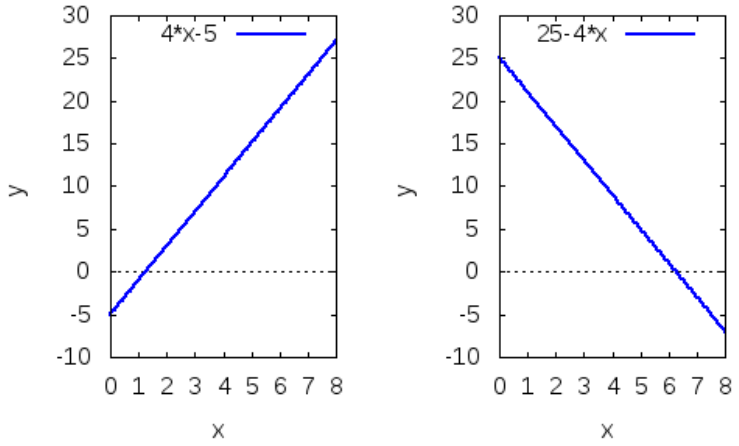


Figure 3.13: Two Monotonic Functions

Each of the functions illustrated above is a "one-to-one mapping." That is, for every value of x in the interval (a, b), there is one and only one value of y in the interval (c, d) such that $y = f(x)$. Hence a change in the value of x always yields a unique and different value of y, and *vice versa*. The mathematical notation for this relationship is the following: $x = f^{-1}(y) = g(y)$. This is read, "x is an inverse function of y." This new function, $x = f^{-1}(y) = g(y)$ is called an *inverse function*. Its domain is the interval (c, d), which is the range of $f(x)$. The range of $g(y)$ is the interval (a, b), the domain of x for $f(x)$.

The cell below shows the graphical relationship between a function f and its inverse function. We can obtain the graph of one from the graph of the other. First, however, we have *Maxima* derive the inverse function and recognize that this function's domain is limited to the positive real numbers. We define the expression that constitutes this function and assign it the name `expr`. Then we solve $y = f(x)$ for x. The solution involves $log(y)$ and, therefore, the inverse function's range (which must be the original function's domain) is limited to the positive real numbers.

```
(%i) [expr:  y=125*(2^x), soln:  solve( expr,x)[1]];
```
(%o) $[\mathbf{y = 125\,2^x}, \mathbf{x = \frac{\log\left(\frac{y}{125}\right)}{\log(2)}}]$

Before we graph the function and its inverse, we determine the range of the inverse function. Doing so helps guide the graphing of the two functions.

```
(%i) [y0, y1]:  [subst(x=0,expr), subst(x=5, expr)];
```
(%o) $[\mathbf{y = 125}, \mathbf{y = 4000}]$

Figure 3.14 shows that the two functions, except for scale, are mirror images of each other. We have added reference lines that take into account the difference in scale for the two variables. In the first, the reference line is $y = 4000 \cdot x$, and in second, the line is the equivalent, $x = y/4000$. We have extended the axes slightly from the values shown above in order to emphasize the similarity of these two functions.

```
(%i) initial:  gr2d(user_preamble="set key left",
      xlabel="x", ylabel="y", ytics=1000, key = "f(x)",
      explicit(rhs(expr),x,0,6) , color=black,line_width=1,
      key="",explicit(1000*x,x,0,6) )$
     inverse:  gr2d(user_preamble="set key left",
      xlabel="y", xtics=4000, yrange=[0,6], ylabel="x",
      key="g(y)",implicit(soln, y,rhs(y0),8000,x,0,6),
      color=black,line_width=1,key="",
      explicit(y/1000,y,0,rhs(y1)*1.6) )$
     wxdraw(initial,inverse, columns=2)$
```

For emphasis, we repeat that f^{-1} is not the same as $(f)^{-1}$, which is the inverse of the expression on the right-hand side of the initial function. The difference between the two is shown below, as created with *Maxima*'s `print` command.

$$\textbf{The inverse of } \mathbf{f(x)} \;=\; \frac{1}{y} = \frac{1}{125\,2^x}. \textbf{ The inverse function is } \frac{\log\left(\frac{y}{125}\right)}{\log(2)}.$$

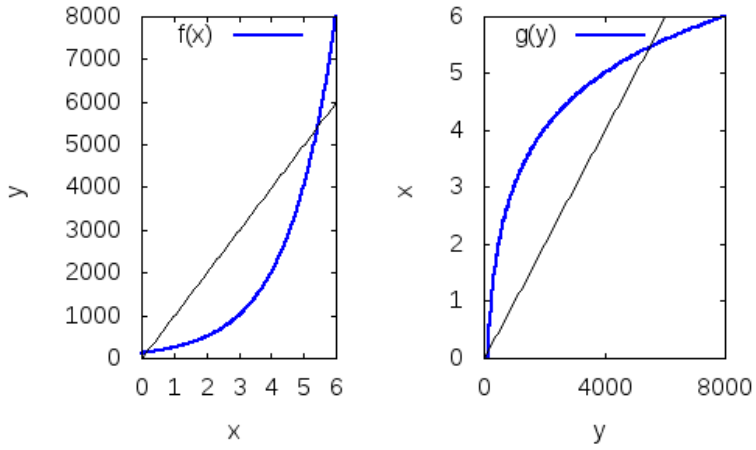


Figure 3.14: A Function and Its Inverse

3.3 Summation and Multiplication

In many business and economic problems, one is required to sum or multiply a large number of terms (numbers, variables, parameters, coefficients, and so forth). A shorthand notation that makes use of the uppercase Greek letter sigma, Σ, enables us to write lengthy sums in a more compact form. The uppercase Greek letter pi, Π, serves the same role for multiplication. The statements below were created in *Maxima*, using the `print`, `sum`, and `product` commands. Note that *Maxima*, like all software, has its conventions about output. In this case, it reports sums and products in reverse order.

```
(%i) print("Interpret", 'sum(a[i],i,1,10), " as " ,
        sum(a[i],i,1,10))$
     print("Interpret ",'product(a[i],i,1,10), " as ",
        product(a[i],i,1,10))$
```

Interpret $\sum_{i=1}^{10} a_i$ **as** $a_{10}+a_9+a_8+a_7+a_6+a_5+a_4+a_3+a_2+a_1$

Interpret $\prod_{i=1}^{10} a_i$ **as** $a_1\,a_2\,a_3\,a_4\,a_5\,a_6\,a_7\,a_8\,a_9\,a_{10}$

The first table below shows the results of the following instructions (see the workbook for details):

- Use makelist to create a list of the numbers 1 through 10.
- Make a second list of the squared values of these numbers.
- Sum the second through sixth values in the list of squared values.
- Sum all ten squared values.
- Sum k times each of the ten squared values. The illustrates the *homogeneity property* of summation.
- Create a list of first differences, beginning with the second squared value.
- Sum the list of nine first differences, confirming that this sum equals the last value *less* the first value in the original list of squared values. This result is an example of the *telescoping property* of addition. For any list $[a_1, a_2, \cdots, a_n]$, it must be true that $(a_2 - a_1) + (a_3 - a_2) + \cdots - (a_n - a_{n-1)} = a_n - a_1$.

List of terms	[1,2,3,4,5,6,7,8,9,10]
List of squared terms	[1,4,9,16,25,36,49,64,81,100]
Sum, x^2[2] thru x^2[6]	90
Sum of all items	385
*Sum of k*all items*	385 *k*
Differences x^2[i]/x^2[i−1] 2 thru 10]	[3,5,7,9,11,13,15,17,19]
Sum, differences x^2[i] − x^2[i−1] 2 thru 10	99

The next table shows the effects of the same sets of commands, except that `product` replaces `sum`, and that ratios of adjacent values replace differences between adjacent values.

Observe that the homogeneity property of multiplication implies that multiplying each of a set of values by k and then multiplying the result yields a result that is k^n as large as the products of the initial values.

Observe also that a variant of the telescoping property also applies. The product of the ratios created as in the table below equals the terminal value

in the list of original values. Why this must be so is easily determined by reference to the second-to-last line: striking out values from left to right leaves only the numerator of the last item, which is the final value of the original list.

List of terms	[1,2,3,4,5,6,7,8,9,10]
List of squared terms	[1,4,9,16,25,36,49,64,81,100]
Product, x^2[2] thru x^2[6]	518400
Product, all items	13168189440000
*Product, k*all items*	$13168189440000\,k^{10}$
Ratios x^2[i]/x^2[i−1] 2 thru 10]	$[4,\frac{9}{4},\frac{16}{9},\frac{25}{16},\frac{36}{25},\frac{49}{36},\frac{64}{49},\frac{81}{64},\frac{100}{81}]$
Product, ratios x^2[i]/x^2[i−1] 2 thru 10	100

Exercise Set 2.8

1. For each of the following functions, sketch the graphs of f and f^{-1} and determine the range and domain for both the function and its inverse.
   (a) $y = 5 + 4 \cdot x$ (b) $y = 3 - 2 \cdot x$ (c) $y = \sqrt{2 \cdot x + 8}$
   (d) $y = \frac{x}{x+1}$ (e) $y = x$ (f) $y = 2 \cdot x - 3$

2. Evaluate each of the following sums, both by hand and using *Maxima*.
   (a) $\sum_{j=40}^{44} j$ (b) $\sum_{k=1}^{4} 2^{k-1}$ (c) $\sum_{i=0}^{5} (-1)^i$
   (d) $\sum_{k=1}^{4} \frac{k-1}{k+1}$ (e) $\sum_{i=1}^{8} (3 \cdot i + 5)$

3. Rewrite each of the following in summation notation.
   (a) $1 + x + x^2 + x^3 + x^4$ (b) $1 + 2 + 4 + 8 + 16 + 32$
   (c) $1 + \frac{a}{x} + \frac{a^2}{x^2} + \frac{a^3}{x^3} + \cdots + \frac{a^n}{x^n}$

4. Evaluate each of the following products, both by hand and using *Maxima*.
   (a) $\prod_{j=40}^{42} j$ (b) $\prod_{j=2}^{5} x^j$ (c) $\prod_{i=5}^{5} i^2$
   (d) $\prod_{i=1}^{4} (3 \cdot i + a)$

3.4 Questions and Problems

1. Solve the equation $x^2 + 5 \cdot x + 6 = 0$ for x using the quadratic formula.

2. Refer to the equation $y = 4 \cdot x - 5$.
   (a) Find the values taken on by the dependent variable y when independent variable x takes on integer values between 1 and 3. Do this twice, first by hand and then with *Maxima*. Create a named list of x values and then create a functional expression for y and map that function onto the list.
   (b) Graph the function $y = 4 \cdot x$–5, either by hand or using *Maxima*.

3. Graph the equation $y = x^2 - 2 \cdot x + 1$. What kind of functional form is this? What is the range of this function? What is its domain?

4. Given the following price and quantity demanded data for pizza.

Quantity per day	50	60	70	80	90
Price, $ per pizza	15	14	13	12	11

   (a) Find the equation of the demand function for pizza such that $Q = f(P)$.
   (b) What are the slope and the ordinate (intercept) of the demand function? Beware: Q is the "y" variable given the way the function is expressed.

5. The gross domestic product (GDP) of Nuevo Laredo, a relatively less developed country, was valued at 100 billion pesetas in the year 2015. The leaders of the country feel that a realistic growth target is for GDP to grow to 220 within seven years (the year 2022). Consider GDP as a function of time t. Which of the following functional relationships most accurately portrays a growth path that is consistent with the leaders' belief? (a) $GDP = 100 + 2 \cdot t$; (b) $GNP = 100 + 20 \cdot t$; or (c) $GNP = 100 \cdot e^{0.1125 \cdot t}$. Use `wxdraw` to graph the three expressions. Let $t = 0$ for the year 2015. Specification (c) is consistent with a compounded annual growth rate of 11.25%. Given your knowledge of economic growth rates (or on the results of a quick Google search), do you think any of these growth paths is likely to occur?

6. The demand function for carpenters in Portland is given by $Q = 270,000 \cdot W^{-2}$, where Q = quantity of carpenters employed, measured in worker hours per time period, and W = wage rate of carpenters, measured in dollars per hour. How many carpenter hours will result if W = \$30 per hour? What kind of functional form is this demand function? (FYI, -2 is the elasticity of demand for labor with respect to W.)

7. Profit maximization requires that a firm equate marginal cost (MC) with marginal revenue (MR). The marginal cost curve for the production of mathematical economics textbooks is given by $MC = 8+0.01 \cdot Q$, where Q is the number of textbooks produced. The marginal revenue realized from the sale of textbooks is given by $MR = 104 - 0.05 \cdot Q$. (a) Find the profit-maximizing output and sales of the firm. (b) If the demand curve is $Q = 2080 - 10 \cdot P$, where P is the product price, what price must this firm charge to sell the quantity that you determined? (c) Determine the inverse demand curve $P = g(Q)$ and use Maxima to graph this function along with MC and MR.

8. The Numerical Control Company (NCC) produces control units that attach to machine tools. These tools automatically control a tool's operation more precisely than is possible with manual control. NCC has developed a new model, the DX3. NCC knows the identity of firms to which it sells and it is confident that eventually 10,000 units will be purchased. Cost estimates indicate that if 7,000 are adopted within five years, the DX3 will be profitable. Adoption follows the path defined by the logistic function $p(t) = k/(1 + e^{-0.2 \cdot t})$. Here $p(t)$ is the percent of maximum adoptions that will have occurred by time t, $k = 100$ is the maximum percent, and t is the number of years.
   (a) If NCC is right about the maximum value, will this project be profitable?
   (b) In approximately how many years will $p(t) = 99$ percent?

9. Given the production function $Q = f(L, K)$, where Q = output of the firm in units, and L and K are the number of units of labor and capital, respectively, that the firm chooses to hire. The specific form of the production function of the firm is $Q = 3 \cdot \sqrt{L} + 5 \cdot \sqrt{K}$. The marginal production of labor and the marginal product of capital are respectively equal to $MPL = 1.5/sqrtL$ and $MPK = 2.5/\sqrt{K}$. (Chapter 4 shows

why.) The wage that must be paid to a unit of labor is symbolized by w and is equal to \$1 per unit. The price of capital per unit, r, is \$2. The input prices and the product price are beyond the firm's control, and they are not affected by the firm's decisions regarding employment or output levels. Thus, the firm treats these values as constants.

The price of the output is symbolized by P and is \$4 per unit. Profit maximization requires that the firm choose each of the input levels so that the input price equals the product price $\times$ the marginal product of that input. The price of this price-taking firm's product *times* the marginal product of an input is that input's *value marginal product* (VMP). The firm must satisfy $w = VMPL$ and $r = VMPK$, where $VMPL = P \cdot MPL$ and $VMPK = P \cdot MPK$. Determine:
(a) the profit-maximizing magnitudes of L, K, and Q, and
(b) How much profit will the firm will earn if it selects the values determined in (a).

10. Suppose that the average annual earnings for members of a profession is represented by this function $log(y) = f(x) = 12.5 + 0.07 \cdot x - .001 \cdot x^2$, where y is average earnings and x is years of experience.
(a) Graph this function, with y (not $log(y)$ on the vertical axis.
(b) Assume that the average age at which members of this profession begin to practice is 25. What factor(s) might account for the downturn in earnings at $x = 35$?

Chapter 4

Limits, Continuity, and Differentiability

We now begin to explore calculus-based analysis. The calculus consists of two parts, *differential calculus* and *integral calculus*. Differential calculus, as the name suggests relates to examining differences. Its concern is with how y changes in response to a change in x. Integral calculus, as its name might suggest, involves integrating—putting parts back together. If we have a function that defines y's changes in terms of x's changes, then integration can help use to move from the changes (differences) into the value of x (level) at a given value of x. The operations of differentiation and integration are the inverse of each other, just as addition and subtraction, and multiplication and division, are the inverse of each other.

We use the methods of differential calculus when we know the form of the function that states the level of a variable in terms of the level of one or more other variables. Often, the important question is how much a change (changes) in the independent variable(s) will affect the dependent variable's value. This analysis involves taking derivatives. For equations with one independent variable, the derivative is actually the slope of the tangent to the graph of that function, as we see later in this chapter.[1]

[1] In at least one inportant respect, this representation of the use of differential calculus is incomplete. We often illustrate economic models that involve specific expressions. The comments here apply to those cases. In addition, however, differential calculus allows us to develop general principles, so that insights can be applied even when the functional form is not known.

We use the methods of integral calculus when we know the form of the function that describes changes and we wish to know the form of the function that describes the level of the variable in question. An important use of integral calculus is related to the fact that we can use an integral to find the area under a particular curve or function.

This chapter and the next three address the differential calculus and its applications. Chapter 8 introduces integral calculus. Before we learn how to take derivatives and to use them, we develop the concepts of limits and continuity. The concept of limits is critical to understanding how to interpret derivatives. Whether or not a function is continuous determines whether differential calculus methods can be applied.

4.1 Limits

We have already seen one example of a limit. When we examined the logistic function, we saw that a population grows toward a limiting value. Thinking of that example gives us a basis for a formal definition of the term. Begin with a function $y = f(x)$. If, as x approaches some value, x_0, and as a result $f(x)$ approaches some number A, then A is said to be the limit of $f(x)$ as x approaches x_0. The standard notation for the preceding statement is this:

$$\lim_{x \to x_0} f(x) = A.$$

The remainder of this section considers a series of examples. These examples illustrate the concept of the limit. They also introduces some important aspects of limits, aspects that we will encounter in subsequent analysis.

Example: Consider $y = x^2$, with $x_0 = 3$. The table below shows values of x that approach 3 from below and then values that approach 3 from above. These lists were constructed by subtracting values in `diffList` from 3 and by adding those value to 3, respectively. Then corresponding lists of y values were constructed by inserting *smallList* and `largeList` into $f(x)$. For ease of reading, all values are placed into matrices.

```
(%i) f(x):=x^2$
     diffList:  [3, 2, 1, 0.1, 0.01, 0.001, 0.001, 0.0001]$
     smallList:  3 - diffList$ largeList:  3 + diffList$
```

```
belowList:  f(3-diffList)$ aboveList:  f(3 + diffList)$
matrix( cons("x",smallList), cons("y", belowList));
matrix( cons("x",largeList), cons("y", aboveList));
```

(%o) $\begin{bmatrix} x & 0 & 1 & 2 & 2.9 & 2.99 & 2.999 & 2.999 & 2.9999 \\ y & 0 & 1 & 4 & 8.41 & 8.9401 & 8.994 & 8.994 & 8.9994 \end{bmatrix}$

(%o) $\begin{bmatrix} x & 6 & 5 & 4 & 3.1 & 3.01 & 3.001 & 3.001 & 3.0001 \\ y & 36 & 25 & 16 & 9.61 & 9.06 & 9.006 & 9.006 & 9.0006 \end{bmatrix}$

From either direction, as x approaches 3, $f(x)$ approaches 9. Maxima's `limit` command can be used to confirm this result:

(%i) `limit(f(x),x,3);` (%o) **9**

Example. As a second example, consider a series for which the important limit is positive infinity. Suppose that a value grows according to the function $y = y_0 \cdot (1 + r/n)^n$. In this function y_0 is the initial value of variable y, r is a per-period growth rate, n is the number of times that the growth process compounds per period, and t is the number of periods. For our illustration, we set $r = 1$, so that with no compounding the value doubles in one year. The question we address is the limiting value of the expression as n becomes very large (and the term r/n becomes very small). The cell below shows compounding annually ($n = 1$), semiannually ($n = 2$), and so forth up to hourly ($n = 8765.81$. The results show that by the time that compounding occurs daily ($n = 365$ for most years), the value is close to $e = 2.718\ldots$. The final entry, 8765.8 indicates daily compounding, with the number of days allowing for leap years.

```
(%i) g(n) := (1 + 1/n)^n$
     nList:  float([1,2, 4,12,52,365,8765.81 ]);
     yList:  map(g,nList);
     [limit(g(n),n,inf), float(limit(g(n),n,inf)) ];
```

(%o) [**1.0**, **2.0**, **4.0**, **12.0**, **52.0**, **365.0**, **8765.8**]
(%o) [**2.0**, **2.25**, **2.4414**, **2.613**, **2.6925**, **2.7145**, **2.7181**]
(%o) [e, **2.7183**]

The last line above shows the result of applying the `float` command with n approaching infinity. The first result shows the exact value; the second shows its floating-point representation.

Example. Consider a slight variation of the example above. Suppose that a sum of money accumulates for t years at an annual rate of r. The sum plus interest is compounded n times per year. The next cell shows the formula for the value after t years and it shows the limit of that formula as the number of compounding periods per year approaches infinity. This result shows that *Maxima*'s `limit` command is not restricted to finding numerical values. It can also find limits of expressions, where the limit is another expression.

```
(%i)  [ V: S*(1+r/n)^(n*t),     limit(V,n,inf)];
(%o)  [(r/n + 1)^(n t) S,       e^(r t) S]
```

Example. Next consider the expression function $y = (x^2 + 1)/(x–1)$. This function's domain is limited, in that y is undefined for $x = 1$. It might be useful, however, to determine y's value as x approaches 1 from above and from below.[2] We show the results of `limit` when the limit is approached for above (`plus` in the command below) and from below (`minus` in the second command line below. The two commands are in a list. The graph shows the behavior of this function as x approaches 1 in each direction. Note the limited `yrange` values. As an exercise, execute the commands in this cell after having removed the `yrange` option.

```
(%i) [limit((x^2+1)/(x-1),x,1,plus),
      limit((x^2+1)/(x-1),x,1,minus)];
     wxdraw2d(yrange=[-100,100],xlabel="x", ylabel="y",
      key="(x^2 + 1)/(x - 1)",
      explicit( (x^2+1)/(x-1), x, 0, 3 ) )$
(%o) [inf, -inf]
```

Example. Another case in which a limit's value depends on the direction of approach is a *step function*. Consider $y = |x|/x$. The limit of z as x approaches 0 is either 1 or -1, depending on the direction, as the output shows and as Figure 4.2 illustrates. This example offers in important lesson:

[2] We leave as an exercise the determination of y values for x values that are ever closer to 1. Follow the development of the first example.

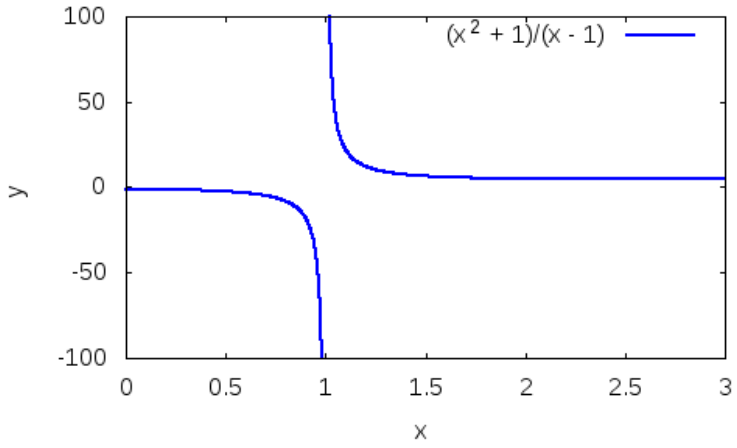


Figure 4.1: No Unique Limit

The limit is not necessarily the same as the value at x_0. For $x = x_0$, the function value y is not defined.

```
(%i) [z :abs(x)/x, limit(z,x,0,plus), limit(z,x,0,minus)];
     (yrange=[-2,2], xlabel="x",ylabel = "|x|/x",
       explicit(z,x, -2,2) )$
```

(%o) $[\frac{|x|}{x}, \quad 1, \quad -1]$

The definition of a limit would be quite difficult to apply in many situations. Fortunately, a number of theorems that involve combinations of relatively simple functions remove the requirement that we derive the limit from definition in each application. Using those theorems can greatly simplify the evaluation of limits for seemingly difficult functions. Furthermore, *Maxima* can directly apply these theorems and supply the limits to a large array of functions. Suppose that a relationship between y and x can be decomposed into two functions, $f(x)$ and $g(x)$. Furthermore, suppose that the following are the limits for these functions as $x \to x_0$: $f(x) \to A$ and $g(x) \to B$. Then the following are true:

- If the expression is $f(x) + g(x)$, then its limit is $A + B$.
- If the expression is $f(x) - g(x)$, then its limit is $A - B$.

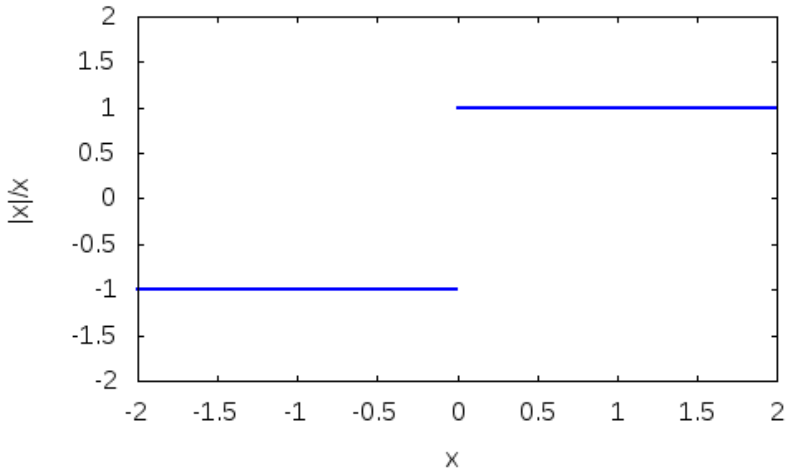


Figure 4.2: A Step Function

- If the expression is $f(x) \cdot g(x)$, then its limit is $A \cdot B$.
- If the expression is $f(x)/g(x)$, then its limit is A/B.
- If the expression is $K \cdot f(x)$ where K is a constant, then its limit is $A + B$.

The next display shows two functions of x, It evaluates each function as x approaches 100.0. It then evaluates the limits of sum, difference, product, and ratio of the two functions as x approaches 100.0. Finally, it evaluates the product of the two functions and the constant K. The function $f(x)$ returns the common (base 10) logarithm for 100.0.[3] The function $g(x)$ is, of course, a quadratic function of x.

```
(%i) [f(x):=log(x)/log(10.0), g(x):=0.05*x^2-2*x+20];
     nameList:  ["A", "B", "sum", "difference", "product",
       "ratio","times K"]$
     limitsList:  [limit(f(x),x,100.0),limit(g(x),x,100.0)
```

[3]The common logarithm of 100.0 is 2.0; the value 1.9999 results from rounding error in the `float` command. Maxima retains the exact value in its memory and applies that value to any calculations that involve $f(100.0)$, such as the five limits that are reported in the table.

```
    limit(f(x)+g(x),x,100.0),
    limit(f(x)-g(x),x,100.0),limit(f(x)*g(x),x,100.0),
    limit(f(x)/g(x),x,100.0),
    limit(K*f(x)*g(x),x,100.0)]$
matrix(nameList, limitsList);
```

(%o)
$$\begin{bmatrix} A & B & \text{sum} & \text{difference} & \text{product} & \text{ratio} & \times K \\ 1.9999 & 320.0 & 322.0 & -318.0 & 640.0 & 0.00625 & 640.0\,K \end{bmatrix}$$

EXERCISE 3.1

In each of Exercises 1 through 15, evaluate the limits. Try to determine the limits yourself before having *Maxima* provide the value. To keep the exercises compact we use $\lim_{x\to A} f(x)$ rather than the equivalent

$$\lim_{x\to A} f(x).$$

1. $\lim_{x\to 2} 2\cdot(x-1)$
2. $\lim_{x\to 4}(x^2+4\cdot x)$
3. $\lim_{x\to -1}(3\cdot x^3-2\cdot x^2+10)$
4. $\lim_{x\to 0}(\frac{x^2-8}{x^2+2\cdot x+1})$
5. $\lim_{x\to 2}(\frac{x^2-4}{x^2+4})$
6. $\lim_{x\to 0}((x^2-4)\cdot(x^3+4\cdot x-3)\cdot(4\cdot x+7)$
7. $\lim_{x\to 3}(x^2-5)^2$
8. $\lim_{x\to -1}\frac{x^2-3\cdot x^3}{2\cdot x+1}$
9. $\lim_{x\to 0}(\frac{x+4}{x-4})$
10. $\lim_{x\to 1} 10^{-x}$
11. $\lim_{x\to 0} 8^x$
12. $\lim_{x\to 1}(2-\frac{x^2}{4})$
13. [check] $\lim_{x\to 3}\frac{(x^3-3\cdot x^2+c\cdot x-24)}{(x-6)}$
14. [check] $\lim_{x\to A} f(x)$
15. $lim_{x\to 0}\frac{x^2+2\cdot e^{x^2}}{e^{x^3}}$

4.2 Extensions of the Limit Concept

The preceding section shows how a limit can be evaluated. The examples show three especially important cases: one is which the limit of the function is undefined (approaches either infinity or negative infinity) at some value of x; one at which the limit at a specified value of x differs, depending on

whether the x value is approached from above or from below, and one in which the limit is found when x itself approaches either positive or negative infinity. The implication of the first case, where the limit approaches positive or negative infinity, is direct: the function of x is not defined at that value. This section addresses the second.

4.2.1 Right-hand and Left-hand Limits

The definition of a limit states that as variable x approaches some finite number, the value of function f(x) approaches some finite number A. That is, $\lim_{x\to 0} f(x) = A$. The limit of a function from above may be the same as the limit taken from below. As have already seen, however, that need not be the case. Thus the left-hand limit may be one value, A, and the right-hand limit may be another value, B.

Only when both the left-hand limit and the right-hand limit exist, and they are equal to each other are we able to state that a (single) limit of a function exists. Therefore the function $f(x) = |x|/x$ does not have a uniquely defined limit at $x = 0$. Therefore, we must ensure that both limits are defined and that they are equal. Fortunately, *Maxima* checks for this. The next exhibit shows the results of a command to find the limit of the function $f(x) = |x|/x$ that does not specify direction and two commands that do specify direction. Without this specification, *Maxima* returns "und" which indicates that the value of the limit is undefined.

```
(%i) [limit(abs(x)/x,x,0), limit( abs(x)/x,x,0,plus),
     limit( abs(x)/x, x, 0, minus) ];
(%o) [und, 1, -1]
```

4.2.2 Infinite Limits

Our definition of a limit stipulates that both x and $f(x)$ approach finite numerical values (x_0 and A, respectively) in the limit. It is possible, nonetheless, that either we wish to consider cases in which x_0 becomes either arbitrarily small or arbitrarily large. In such cases, $g(x)$ can either approach a finite constant, or it can approach a value A that has no limits. First, consider

the simple function $g(x) = A/x$. Its infinite limits are defined in the next exhibit.[4]

```
(%i) g(x,A):= A/x;
     [limit(g(x,A),x,inf),limit(fgx,A),x,minf),
        limit(g(x,A),x,0,plus), limit(g(x,A),x,0,minus)];
     [limit(g(x,2),x,inf),limit(g(x,2),x,minf),
        limit(g(x,2),x,0,plus), limit(g(x,2),x,0,minus)];
     [limit(g(x,-2),x,inf),limit(g(x,-2),x,minf),
        limit(g(x,-2),x,0,plus), limit(g(x,-2),x,0,minus)];
```

(%o) $\mathrm{f}\left(\mathbf{x}, \mathbf{B}\right) := \frac{\mathbf{A}}{\mathbf{x}}$ (%o) $[\mathbf{0}, \mathbf{0}, \mathbf{infinity}, \mathbf{infinity}]$
(%o) $[\mathbf{0}, \mathbf{0}, \infty, -\infty]$ (%o) $[\mathbf{0}, \mathbf{0}, -\infty, \infty]$

This exhibit shows how the function is defined. The commands specify three sets of limits. In the first set A's sign is not specified; in the second, $A > 0$, and in the third $A < 0$. In each case we seek limits as x becomes a very large positive number (`inf`) and a very large negative number (`minf`). Also, we seek limits as x approaches zero from above and from below. In all cases, $1/x$ becomes quite small (approaches 0) as x becomes either very large or very small. Thus the limit of $1/x$ is zero as x approaches either infinity or negative infinity.

The behavior as x approaches zero depends on the sign of A. If that sign is not specified, *Maxima* provides the ambiguous response `infinity` which does not specify a sign. The second and third output lists show the source of this ambiguity: whether the limiting value of A/x is positive or negative for a specified limiting value of x depends on A's sign. Figure 4.3 adds insight into the behavior of this function.

```
(%i) positiveA: gr2d(xaxis=true, yrange=[-10,10],
        xlabel="x",ylabel="1/x", key="A = 2",
        explicit(f(x,2),x,-20,20))$
     negativeA: gr2d( xaxis=true, yrange=[-10,10],
        xlabel="x", ylabel="1/x", key="A = -2",
        explicit(f(x,-2),x,-20,20))$
     wxdraw(positiveA, negativeA);
```

[4] In *Maxima*, we use the notation `g(x,A):=A/x`, with both A and x treated as variables. *Maxima* does not distinguish between variables and parameters.

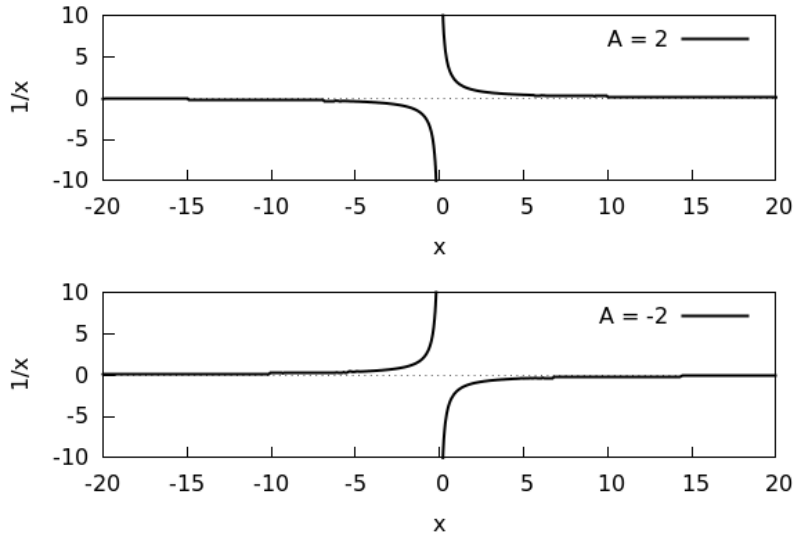


Figure 4.3: Graphs of A/x

In all monotonic functions, as x becomes either quite large or quite small, $f(x)$ can also grow without limit, either toward infinity or toward negative infinity. Consider the behavior of $f(x) = \pm\sqrt{x}$ and $g(x) = \pm log(x)$. For both, negative values of x are not in the domain, but x can grow without limit. As x grows without limit, so do the values of these two functions, as the next exhibit shows.

```
(%i) fList:  [sqrt(x), -sqrt(x), log(x), -log(x)];
     limit(fList,x,inf);
```

(%o) $[\sqrt{x}, -\sqrt{x}, \log(x), -\log(x)]$ (%o)$[\infty, -\infty, \infty, -\infty]$

A problem can arise if a function consists of a ratio of two expressions, both of which grow without limit. In many cases, the problem appears more imposing than it is in reality. Consider $f(x) = (4 \cdot x^2 + 2 \cdot x - 3)/(x^3 - 6 \cdot x + 2)$. Both the numerator and the denominator grow without limit as $x \to \infty$, so the behavior of $f(x)$ is not obvious. One can see, however, that the highest-powered term is in the denominator. Dividing both denominator and numerator by x^3 yields an expression that has a limiting value of $0/1 = 0$. A legitimate question is whether *Maxima* recognizes this relationship. As the next exhibit shows, it does.

(%i) `limit( (4*x^2+2*x-3)/(x^3-6*x +2),x,inf);` (%o) **0**

Maxima can handle much more complex relationships. Let $f(x) = \sqrt{x}/\log(x)$. Both the numerator and the denominator grow without limit, and no simple factoring can separate the growth of the two. The next exhibit shows some values of this function's components and of the function. These suggest that $f(x)$ grows without limit. *Maxima* is able to evaluate this limit and confirms what the listed values suggest. (The first item in each list identifies the values that are displayed.)

(%i)
```
xList:  ["x", 2.0, 10.0, 100.0, 1000.0];
sqrtxList:  sqrt(xList); logxList:  log(xList);
ratioList:sqrtxList/logxList;
limit(sqrt(x)/log(x),x, inf);
```
(%o) $[\mathbf{x}, \mathbf{2.0}, \mathbf{10.0}, \mathbf{100.0}, \mathbf{1000.0}]$
(%o) $[\sqrt{\mathbf{x}}, \mathbf{1.4142}, \mathbf{3.1622}, \mathbf{10.0}, \mathbf{31.622}]$
(%o) $[\log(\mathbf{x}), \mathbf{0.693}, \mathbf{2.3025}, \mathbf{4.6051}, \mathbf{6.9077}]$
(%o) $[\frac{\sqrt{\mathbf{x}}}{\log(\mathbf{x})}, \mathbf{2.0402}, \mathbf{1.3733}, \mathbf{2.1714}, \mathbf{4.5778}]$
(%o) ∞

4.2.3 An Economic Application: The Cobb-Douglas Function and the Constant-Elasticity of Substitution Function

.

We saw one economic implication of a limit above when we established that compound growth can, in the limit, be represented as an exponential function of time. The limit can also be used to show the relationship between two functions that are frequently used to illustrate economic principles and as the basis for empirical work.[5] The two are the Cobb-Douglas function $x^{\alpha} * y^{1-\alpha}$, and the Constant-Elasticity of Substitution (CES) function $(\frac{a}{x^b} + \frac{1-a}{y^b})^{-1/b}$. For the CES function the elasticity of substitution between the two inputs x and y is $s = 1/(b+1)$. For the Cobb-Douglas function $s = 1$. If $s = 1$ in the CES function, however, $b = 0$, and the CES function is undefined. Even so,

[5]This relatively advanced material can be omitted without loss of continuity.

we can evaluate the CES function as b approaches 0. Doing so reveals the relationship between these two functions.[6]

```
(%i) [CES:(a/x^b + (1-a)/y^b)^(-1/b), tlimit(CES,b,0)];
```

(%o) $\left[\frac{1}{\left(\frac{1-a}{y^b}+\frac{a}{x^b}\right)^{\frac{1}{b}}}, \quad x^a\, y^{1-a}\right]$

The first output entry shows the CES function. The second entry shows the limit of this function as $b \to 0$. The result is the Cobb-Douglas function.

4.3 Continunity

The first step in determining and interpreting the derivative of a function is to have a clear understanding of the concept of a limit. The second step involves using concept of continuity of a function. Very roughly speaking, a continuous function on a particular interval is one whose graph can be drawn without lifting one's pencil or pen from the paper in that interval. Figure 4.4 below shows two continuous functions and two discontinuous functions.[7]

```
(%i) [f1:50 + 10*x - 0.1*x^2, f2:150+2*abs(50-x),
      f3:  if x <=50 then 20+3*x else 50+ x,
      f4:  x/(x-40)]$
     continuous:gr2d(title="Continuous",explicit(f1,x,0,100),
       color=black,explicit(f2, x, 0, 100) )$
     discontinuous:  gr2d(title="Discontinuous",
       yrange=[-100,200], explicit(f3, x, 0, 50),
       explicit(f3,x,50.001,100),
       color=black,explicit(f4,x,0,100) )$
     wxdraw(continuous, discontinuous, columns=2)$
```

[6]The `tlimit` command is used instead of `limit`. This command creates a Taylor series representation of the function and then evaluates the limit of that series. We use this because on some installations of *Maxima*, applying the `limit` command to this exprssion resulted in an overflow error.

[7]The second of the two `explicit` commands used to generate the first discontinuous graph contains a "fudge" term in that the second segment begins at 50.001 and not 50. The reason for this is that `draw` would connect the points if 50 were both the endpoint of the first segment and the beginning point of the second segment.

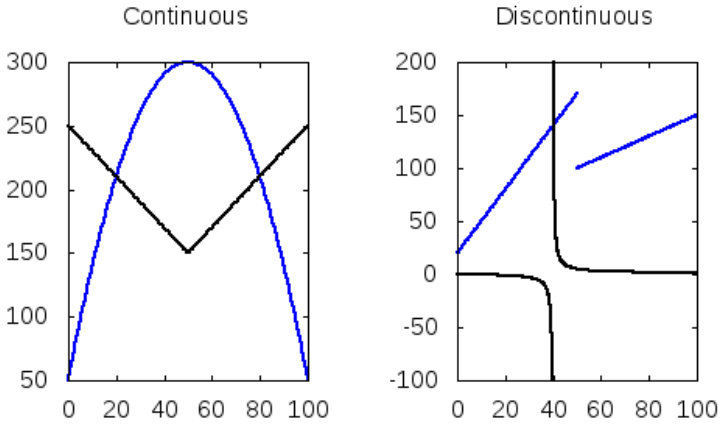


Figure 4.4: Continuous and Discontinuous Functions

Despite the kink at $x = 50$, the v-shaped function in the first panel is continuous. Both functions in the second panel exhibit discontinuities. Following a more formal definition of continuity, we catalog the types of discontinuities, one at $x = 40$ and the other at $x = 50$. Look at the expressions for these two functions and determine why the discontinuities exist.

The following definition applies: A function $y = f(x)$ is said to be *continuous* at $x = x_0$ if the following three requirements are fulfilled: (1) $f(x_0)$ exists such that point x_0 is in the domain of the function, (2) $\lim_{x \to x_0} f(x)$ exists, and (3) $\lim_{x \to x_0} f(x) = f(x_0)$. Failure to meet these conditions can fall into one of four categories: removable, jump, infinite, and essential.[8]

4.3.1 Removable Discontinuity

A removable discontinuity typically occurs when the domain of the independent variable does not include a point (or points). Such discontinuities can, in principle, be removed by reference to the limit of the expression. Consider

[8]This category list is used in the MIT Open Course "Single Variable Calculus": http://ocw.mit.edu/courses/mathematics/18-01sc-single-variable-calculus-fall-2010/index.htm. Actually the name applied to what we call essential is "other (ugly)."

$$y = \left(\frac{1}{1 + 1/x^2}\right)^x.$$

The domain of x does not include $x = 0$. Accordingly, when we attempt to evaluate $y(0)$, the result is an error message.

```
(%i) f(x):= (1/ (1 + 1/x^2))^x$          f(0);
(%o) expt: undefined: 0 to a negative exponent.
       #0: f(x=0) – an error.
       To debug this try: debugmode(true);
```

The next cell shows that $y = f(x)$ does have a defined limit. Whether we approach $x = 0$ from above or below, $y = 1$. We can add this point to the graph to complete the represent of $f(x)$ over the selected range of x values. In this case, the limit is "on the line," but this need not be the case. The endpoints on the two sections of the curve in Figure 4.5 are chosen so as to represent the function but to leave space for the point.

```
(%i) y0:  limit(f(x),x,0);
     wxdraw2d(explicit(f(x),x,-2,-0.005),
       explicit(f(x),x,0.005,2), point_type=circle,
       points( [ [0,y0] ]) )$
(%o) 1
```

The next exhibit shows the new, augmented function for which the domain spans the real number line. Selected values of this function, named *faug(x)*, appear as the output. The discontinuity has been removed.

```
(%i) faug(x) := if x = 0 then 1 else f(x)$
     [faug(-1/2), faug(0),faug(1), faug(2)];
(%o) [√5,  1,  1/2,  16/25]
```

An example much like the one above is the one that applies to periodic compounding: We have already evaluated the limit of this function when $n \to \infty$, finding that the result is $y_0 \cdot e^{r \cdot t}$. Use *Maxima* to confirm that the limit as $n \to 0$ is y0. Also confirm that the function is not defined at $n = 0$ (of course, $n < 0$ makes no economic sense). The fact that n = 0 is not part of the domain actually makes economic sense: if n = 0, compounding never occurs.

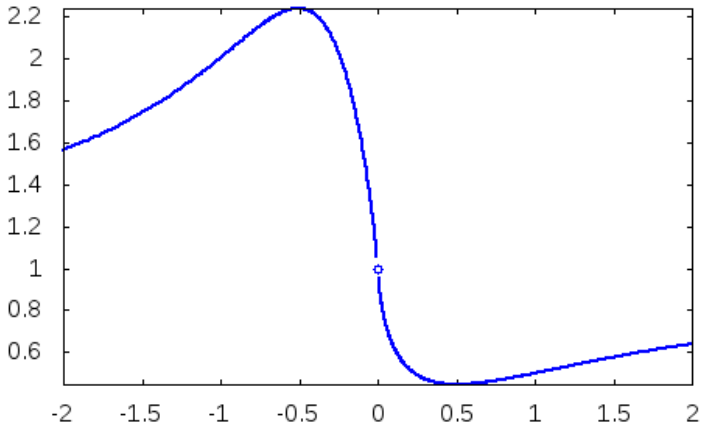


Figure 4.5: A Removable Discontinuity

4.3.2 Jump Discontinuity

A *jump discontinuity* occurs when the right-hand and left-hand limits exist but the two are not equal. One such discontinuity appears in Figure 4.4. The function $f(x) = |x|/x$, which we considered when examining limits, is another case.

4.3.3 Infinite Discontinuity

Each of the two functions in the next exhibit exhibits an *infinite discontinuity*. As x approaches a critical value (0 here) the functions' values become either very large or very small. The limits themselves may differ or they may be the same (that is ∞ or $-\infty$). Do not conclude that the fact that the limit of $1/x^2$ is the same from either direction implies continuity. Of the three conditions that must be met for a function to be continuous $1/x^2$ fails the first two: 0 is not in the domain, and the limit is no finite value.

```
(%i) kill(f,g)$ f(x):=1/x$ g(x):=1/x^2$
     ["f(x) limits:",limit(f(x),x,0),limit(f(x),x,0,plus),
     limit(f(x),x,0,minus)];["g(x) limit:",limit(g(x),x,0)];
     different:gr2d(yrange=[-100,100],xaxis=true,
       yaxis=true,title="f(x)=1/x",explicit(f(x),x,-1,1))$
```

```
same:gr2d(yrange=[0,100],xaxis=true,yaxis=true,
  title="g(x)=1/x^2", explicit(g(x),x, -1,1))$
wxdraw(different,same, columns=2)$
```

(%o) [**f(x)limits** :, **infinity**, ∞, $-\infty$] **(%o)** [**g(x)limit** :, ∞]

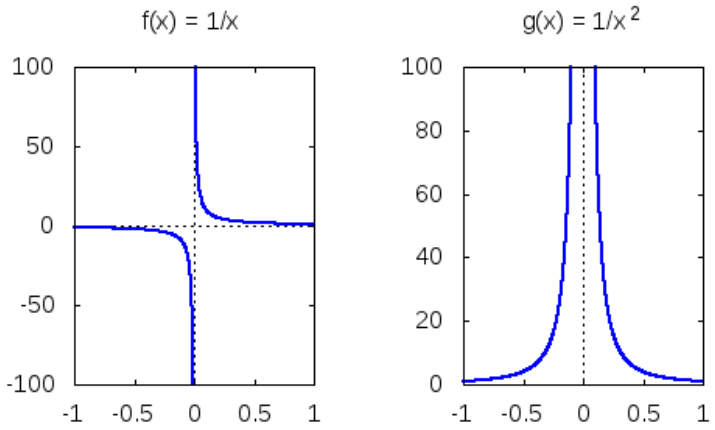


Figure 4.6: Infinite Discontinuities

The first output line above reminds us that *Maxima* cannot determine the limit of $1/x$ as $x \to 0$ unless the direction is provided. The second output line shows the same is not true of $1/x^2$: $g(x) \to \infty$ regardless of the direction from which zero is approached.

4.3.4 Other (Ugly) Discontinuity

Some functions have no limits as they approach a critical value. A commonly-used representative for these is $\lim_{x \to x_0} sin(1/x)$. As Figure 4.7 shows, this function fluctuates increasing wildly between -1 and 1 as x $\to$ 0.

A large number of functions exhibit this behavior. Fortunately, many of the ones that do are functions that depict oscillations of physical systems as time goes to infinity, and are not likely to apply to phenomena typically analyzed with economic models. Even so, the possibility of such behavior should be kept in mind.

```
wxdraw2d(yrange=[-1,1.5], xtics=-.1,0,.1,
       key="Approaching 0 from below",
       explicit(sin(1/x),x,-.1,0),color=black,
       key="Approaching 0 from above",
       explicit(sin(1/x),x,0,.1) )$
```

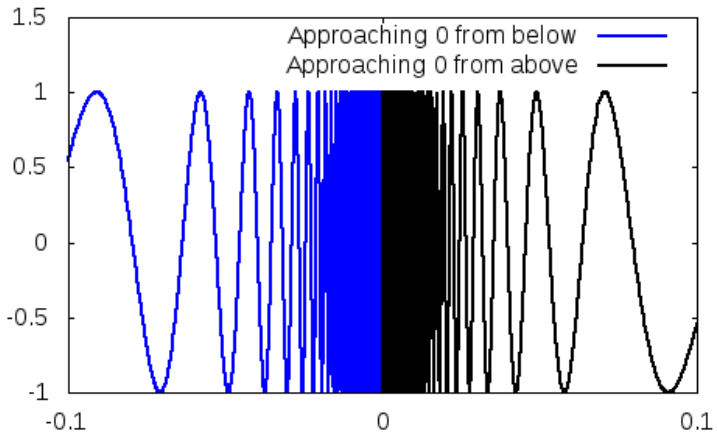


Figure 4.7: An "ugly" discontinuity

Exercise 3-3

Determine the values of x for which the following functions are continuous. If a discontinuity exists, determine the type of the discontinuity, and correct any removable discontinuities.

1. $f(x) = 3 \cdot x^3 + 2 \cdot x^2 + x + 1$
2. $f(x) = x/(x+1)$
3. $f(x = (x^3 - 27)/(x^2 - 9)$
4. $f(x) = (x^2 - 3 \cdot x + 2)/(x - 2)$
5. $f(x) = (x^2 + x - 2)/(x - 1)^2$
6. $f(x) = (x^2 - 4 \cdot x - 21)/(x - 7)$
7. $f(x) = \sqrt{4 - x^2}$
8. $f(x) =$
9. $f(x) = \frac{x^2+5 \cdot x+6}{x^2+4 \cdot x+4}$
10. $f(x) = 8/(x - 4)$
11. $f(x) = 5/(1 - 2^x)$
12. $f(x) = e^x$
13. $f(x) = \frac{1}{x \cdot (x-4)}$
14. $f(x) = (x - 6)/6$
15. $f(x) = \frac{x+4}{x^2+2 \cdot x-8}$
16. $f(x) = 1/(x^2 + 1)$
17. $f(x) = (x = 4)/(3 \cdot x^2 - 27)$
18. $f(x) = (x^2 - 16)/(x + 4)$
19. $f(x) = \sqrt{x}$
20. $f(x) = 1/(2^{e^x} - 2)$

4.4 The Derivative of a Function

The concepts of a limit and of a continuous function constitute the foundation for our study of the derivative of a function. Consider the function $y = f(x)$. We now focus on the effect that an incremental change in x–which we denote as Δx– has on y. The magnitude of any change in y, denoted Δy, that does occur depends not only on the magnitude of Δx, but also on the specific form of $f(x)$.

Consider first the simple linear form. In Chapter 2, a linear function was graphed as a line with a constant slope. We found that the slope of a straight line that passed through any two points (x_1, y_l) and (x_2, y_2), or in functional notation $[x_1, f(x_1)]$ and $[x_2, f(x_2)]$, was given by the quotient

$$m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

This definition of the slope m is hereafter referred to as the *difference quotient*. The exhibit below, for $y = 10 + 2 \cdot x$, reports an initial value of x, $x_1 = 5$, the function, and the initial value of y, $y_1 =$. Then it produces a table, the first row of which is a list of increasingly small Δx values. The

second row shows the implied x value; the third row shows the implied y values. The last two rows show the implied changes in y, Δy, and the quotients, $\Delta y/\Delta x$. For this linear function all values on the last row equal the coefficient of x.

```
x1:5, f(x):=10 + 2*x, y1:f(x1)]$
    DeltaxList:  [2.00, 1.00, 0.50, 0.25, 0.10, 0.01]$
    xList:  x1 + DeltaxList$ yList:  f(xList)$
    DeltayList:yList-y0$ dydxList:DeltayList/DeltaxList$
    matrix(cons("x changes",DeltaxList),
      cons("x values",xList), cons("y values ", yList),
      cons("ychanges", DeltayList),
      cons("difference quotients",dydxList) );
```

x changes	2.0	1.0	0.5	0.25	0.1	0.01
x values	7.0	6.0	5.5	5.25	5.1	5.01
y values	24.0	22.0	21.0	20.5	20.2	20.02
y changes	4.0	2.0	1.0	0.5	0.199	0.0199
difference quotients	2.0	2.0	2.0	2.0	1.9999	1.9999

We cannot always use linear functions in decision and choice problems. When the underlying function is nonlinear, we must determine the slope of the function at a particular point of interest, since the slope of the function differs at different points on that function. Consider the nonlinear function $y = f(x) = x^2$. We commence by selecting a particular point on the graph of that function, namely $(3, 9) = (x_1, y_1)$. Then, as in the preceding example, we generate a list of values of a variable x^2 by specifying changes for which the absolute values decrease as we move toward the middle column of the table. The table shows that the difference quotient moves toward 6 as we move closer to the point (3, 9). That quotient is not defined at 9 because that value implies $\Delta x = 0$.

The difference quotient method helps us find the slope of a line between two points. It measures an average rate of change rather than the slope of the function at a specific point. Given two points (x_1, y_1) and (x_2, y_2), $\Delta y/\Delta x$ measures the average rate of change in y that occurs over the interval per unit change in x. Commands are omitted because they are virtually the same as those for the preceding exhibit.

x changes	**x values**	**y values**	**y changes**	**difference quotients**
-2	1	1	-8	4
-1	2	4	-5	5
-0.5	2.5	6.25	-2.75	5.5
-0.2	2.8	7.8399	-1.16	5.8
-0.1	2.9	8.41	-0.589	5.8999
-0.01	2.99	8.9401	-0.0598	5.9899
0	3	9	0	$---$
0.01	3.01	9.06	0.06	6.0099
0.1	3.1	9.61	0.61	6.1
0.2	3.2	10.24	1.24	6.2
0.5	3.5	12.25	3.25	6.5
1	4	16	7	7
2	5	25	16	8

Both the linear example and the quadratic example illustrate that as the change in variable x, Δx, becomes increasingly small, approaching 0 in the limit, the difference quotient approaches some finite value as a limit. In the case of $y = x^2$, as $x \to 3, \Delta y/\Delta x \to 6$. This is true whether we approach $x = 3$ from above or from below. When this is the case, this limit is called the *derivative* of $y = f(x)$ with respect to x. This derivative is denoted dy/dx and is defined as follows: Given the function $y = f(x)$, the derivative of y with respect to x, is

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

provided that the limit exists.

This definition of a derivative still measures a rate of change; however, the rate of change is an infinitesimally small change in variable x. For that reason, a derivative may be thought of intuitively as being taken at a particular point on a curve. We apply the definition of a derivative to the functions $y = 10 + 2 \cdot x$ and $y = x^2$.

For the linear function

$$\frac{dt}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{(10 + 2 \cdot (x + \Delta x) - (10 + 2 \cdot x)}{\Delta x} =$$

$$\lim_{\Delta x \to 0} \frac{(10 + 2 \cdot x + 2 \cdot \Delta x) - (10 + 2 \cdot x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{(2 \cdot \Delta x)}{\Delta x} = 2.$$

This confirms that the slope of a linear function is the same for all values of the independent variable.

For the second function, $y = x^2$

$$\frac{dt}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{(x + \Delta x)^2 - x^2}{\Delta x} =$$
$$\lim_{\Delta x \to 0} \frac{x^2 + 2 \cdot x \cdot \Delta x + (\Delta x)^2 - x^2}{\Delta x} =$$
$$\lim_{\Delta x \to 0} \frac{2 \cdot x + \Delta x}{\Delta x} = 2 \cdot x.$$

The function's rate of change is directly related to x's value. For $x = 3$, for example, $dy/dx = 2 \cdot 3 = 6$.

As these examples indicate, the derivative of a function is, in general, a new function derived from the original function. If the original function was a function of variable x and only of x, then the derivative is also a function of x (but of no other variable). In the first, linear case the derivative is a *degenerate function* of x: $dy/dx = 2 \cdot x^0$.

We briefly introduce *Maxima*'s command to determine a derivative. We will put it to considerable use as we proceed. The command for the type of expression that we have encounter so far is simple. It has just two arguments. This are the expression itself and the identity of the independent variable: `diff(expression, x)`. The next exhibit applies `diff` to the two expressions that have analyzed so far.

```
(%i) [diff(10 + 2*x, x), diff(x^2, x)];
(%o) [2,     2 x]
```

Exercise 3-4

For the following functions, find the derivative, $dy/dx = df(x)/dx$, by evaluating the limit of the difference quotient. Then use *Maxima* to confirm your results.

1. $f(x) = x^3$
2. $f(x) = x^2 + 3 \cdot x + 4$
3. $f(x) = 4 \cdot x - 1$
4. $f(x) = 4 - 3 \cdot x$
5. $f(x) = a \cdot b \cdot x$
6. $f(x) = b \cdot x^2$
7. $f(x) = 144 - 32 \cdot x$

4.4.1 Geometric Interpretation of the Derivative

The formal definition of a derivative can be illustrated geometrically. The geometric interpretation of a derivative also leads us directly to the idea of a derivative as the slope of a tangent line at a point on a curve. We begin with the geometric interpretation of a difference quotient. Consider the function $y = f(x) = x^2$, whose graph is illustrated below. A list of four Δx values, two negative and two positive, is used to create a list of four x values. Those values, in turn, are used to create a list of four y values. The four chords below connect (3, 9) and the four $(3 + \Delta x, f(3 + \Delta x))$ points. The slopes of these four chords are the difference quotients. Visual examination shows that as Δx approaches 0, the slope of the resulting chord becomes closer to the slope of $f(x)$ at $x = 3$.

```
(%i) limit(((x+h)^2-x^2)/h,h,0); f(x):= x^2$ x0:3$
     deltaList:[-4, -3, 3, 4]$ xList:x0+deltaList$
     yList:  f(xList)$
     wxdraw2d( explicit(f(x),x,-2,8), line_width=1,
       color=black,points_joined=true,
       points([[3,f(3)] ,[xList[1],yList[1]]]),
       points([ [3,f(3)] ,[xList[2], yList[2] ] ]),
       points([ [3,f(3)] ,[xList[3], yList[3]]]),
       points([ [3,f(3)] ,[xList[4], yList[4]]]),
       color=orange, line_width=2,
       key="Tangent line",
       explicit(-9 + 6*x,x,0,7),dimensions=[480,480])$
(%o)  2x
```

The slope of the line that is tangent to $f(x)$ at $x = 3$ is the limiting value of chords like these two, when $\Delta x \to 0$.

The `limit` command confirms that the limiting ratio of Δy to Δx is $2 \cdot x$. This command, using h to denote Δx, shows that the limiting slope of the chords is indeed $2x$. The direction from which $\Delta x \to 0$ does not affect the outcome and, therefore, does not need to be specified in the `limit` command.

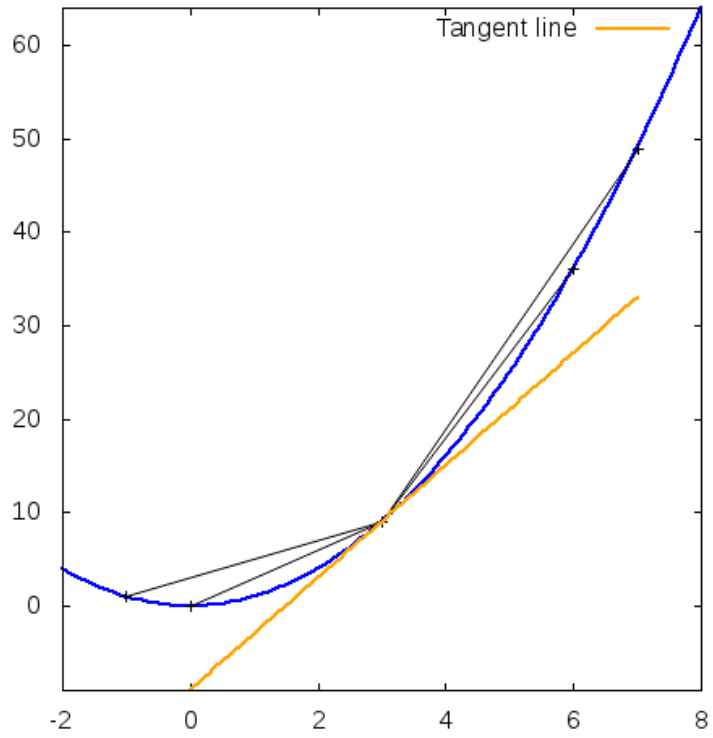


Figure 4.8: Derivative as Tangent

4.4.2 Continuity and Differentiability

Not all functions can be differentiated. For a function $y = f(x)$ to differentiable at a point, it must be continuous at that same point ($f(x)$ must exist). Thus continuity at a point is a necessary, but not sufficient, condition for a derivative to exist at that point. Recall our recoverable discontinuity for

$$y = \left(\frac{1}{1 + \left(\frac{1}{x}\right)^2} \right)^x$$

Refer to Figure 4.5 to recall this function. The value $x = 0$ is not part of this function's domain. Maxima can return a derivative for this function, the messy term below. As it turns out, however, this derivative cannot be evaluated for $x = 0$. Trying to evaluate dy/dx at $x = 0$ yields the same error

message that resulted in trying to evaluate f(0). As an exercise, confirm that dy/dx does exist for other values of x.

```
(%i) f(x):= (1/(1 + (1/x)^2))^x;
     diff(f(x),x); subst(x = 0, %);
```

$$\text{(\%o)}\quad \mathbf{f}\,(\mathbf{x}) := \left(\frac{1}{1+\left(\frac{1}{x}\right)^2}\right)^{x}$$

$$\text{(\%o)}\quad \frac{\frac{2}{\left(\frac{1}{x^2}+1\right)x^2}-\log\left(\frac{1}{x^2}+1\right)}{\left(\frac{1}{x^2}+1\right)^{x}}$$

(%o) expt: undefined: 0 to a negative exponent.
-- an error. To debug this try: debugmode(true);

To repeat, continuity is necessary for the evaluation of a derivative at a point on a function. It is not, however, sufficient. Suppose that $f(x) = |x|$. This function is continuous at $x = 0$: $f(0)$ exists, $\lim_{\Delta x \to 0} f(x)$ exists, and $\lim_{\Delta x \to 0} f(x) = f(0) = 0$. Even so, no unique value. The derivative of $f(x)$ is $dy/dx = |x|/x$. As from the left the term $|x|/x$ approaches -1; as from the right, $|x|/x$ approaches 1. At $x = 0$, it is not defined. In the graph below, any line that passes through the point (0,0) with a slope between -1 and 1 would be "tangent" to this function. Thus the derivative of $f(x)$ has no unique value at $x = 0$.

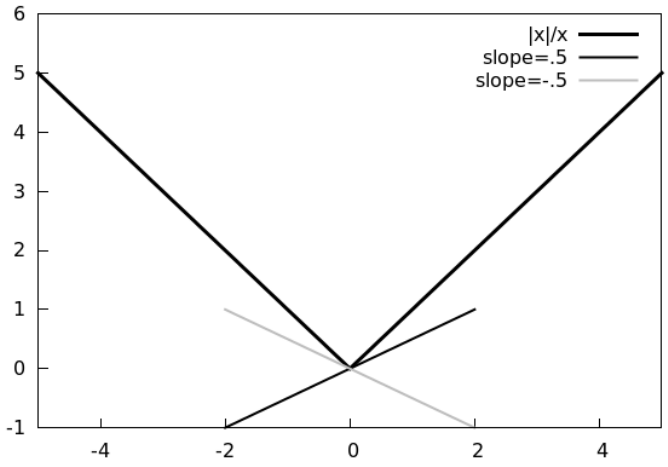


Figure 4.9: The absolute value function

Chapter 5

Differentiation: Univariate Functions

Marginal analysis is the backbone of much of modern economic theory and its application. Marginal analysis examines the effects on other variables of small changes in a particular exogenous variable or a set of such values. For example, it can be used to analyze the effect of a price change on the quantity demanded for a good, other things equal. It can also be used to address the effect of a simultaneous change in price and income, again other things equal. Furthermore, once the analytical framework is defined, the implications of changes in the "other things" can be addressed.

We can express the rate of change in one variable in response to small changes in another variable as the first derivative of the function involved. Chapter 4 introduced the concept of a derivative and its geometric interpretation. This chapter derives and utilizes rules that will assist us in differentiating functions of one variable. Mastery of the technique of differentiation does not just enable us to speak to specific problems. This mastery also enhances our understanding of a wide range of concepts such as marginal cost, marginal revenue, marginal utility, elasticity, and population growth.

5.1 Rules for Differentiation

Finding the derivative of a function would be tedious if it required us to compute the derivative as the limit of a difference equation, following its

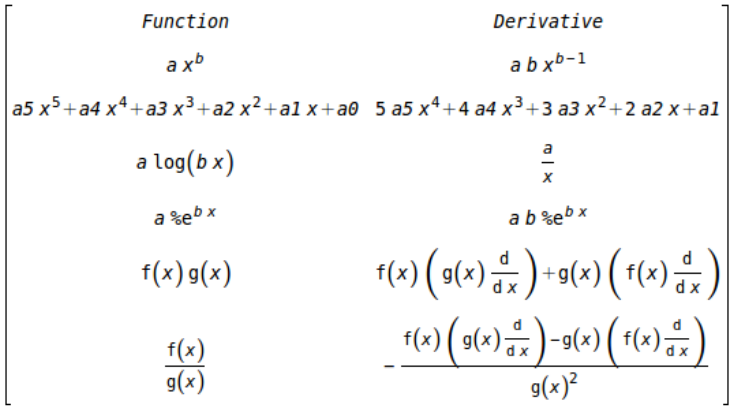


Figure 5.1: Some functions and their derivatives

definition in Chapter 4, in every case. Fortunately a set of time-saving rules enables us to find the derivative without referring to the difference quotient. Furthermore, computer algebra systems know these rules. Figure 5.1 shows some of the most important rules. We discuss them briefly here and more fully below.

The first and second functions above are closely related in that both are subsets of a larger class of functions, polynomials. For a monomial, $a \cdot x^b$ (a polynomial with a single term), the derivative is $dy/dx = b \cdot a \cdot x^{b-1}$, as indicated in the first row in Exhibit 1. A polynomial is the sum of any number of such terms, and its derivative is the sum of terms like term's derivative. The second line illustrates this with a fifth-degree polynomial. Note that the constant term $a0$ vanishes. This result illustrates an important general aspect of sums and differences: The derivatives of sums (differences) of $f(x)$ and $g(x)$are sums (differences) of $\frac{df(x)}{dx}$ and $\frac{dg(x)}{dx}$.

The third line shows the derivative for a simple logarithmic function. The fourth line does the same for a simple exponential function. The fifth and sixth lines require attention. These relate to the product and quotient of two functions. For the product $f(x) \cdot g(x)$, taking the derivative requires taking the derivative of $g(x)$and multiplying this derivative by $f(x)$, and then taking the derivative of $f(x)$ and multiplying that derivative by $g(x)$. The quotient rule is quite similar except for the squared term in the denominator.

Exercise 5.1. Find dy/dx for the following. Solve these expressions by applying the rules in Exhibit 1 and then with *Maxima*.

1. $y = 1/2$
2. $y = 1000$
3. $y = e^x$
4. $y = \pi$
5. $y = x^{n+1}$
6. $y = x^2$
7. $y = x^{3/2}$
8. $y = -x^2$
9. $y = -x^{-0.5}$
10. $y = x^2 \cdot log(x)$
11. $y = e^x / \sqrt{(x)}$

5.1.1 Polynomials

A quadratic equation illustrates the process of taking a derivative of a polynomial and of interpreting that derivative. Recall that the derivative is the slope of the function. The expression below says that the derivative is negative for $x < 10$ and positive for $x > 10$. Thus, reaches a minimum value at $x = 10$, as the graph confirms.

The Maxima command to determine the expression for the derivative and to print that expression is this:

```
print("For ", f(x) := 0.5*x^2 - 10*x + 100,
" the derivative is df(x)/dx =", diff(f(x), x) )$
```

The resulting output provides the expression for dy/dx:

For $\mathrm{f}(x) := 0.5x^2 - 10x + 100$ the derivative is $df(x)/dx = 1.0x - 10$.

The commands to plot $f(x)$ and its derivative are these:

```
original:  gr2d( xlabel="x",ylabel="f(x)",
    explicit(f(x),x,0,20))$
derivative:  gr2d( xlabel="x",ylabel="dy/dx",xaxis=true,
    explicit(diff(f(x), x), x, 0, 20))$
wxdraw(original,derivative), wxplot_size=[480,480];
```

The graphs in Figure 5.2 show the values of the quadratic function and its derivative. The two are drawn separately because the y-axis units are different.

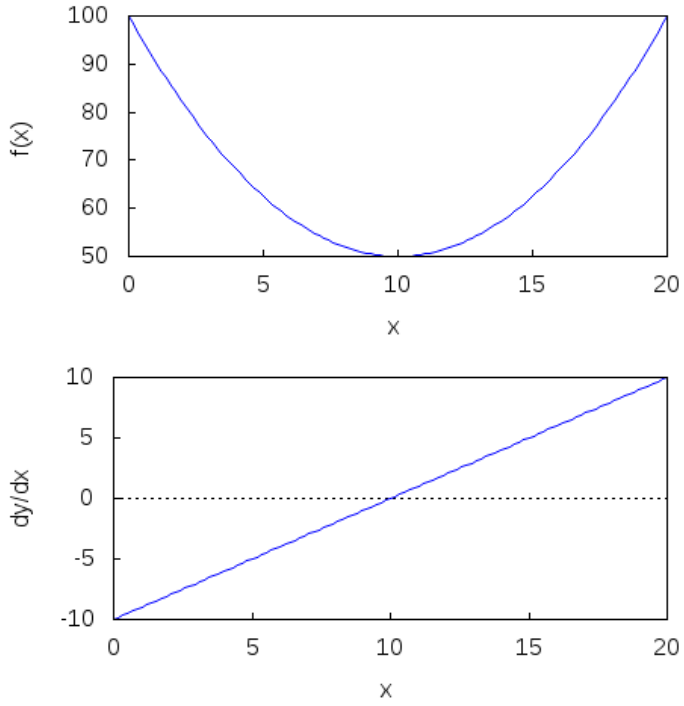


Figure 5.2: Graphing a quadratic function and its derivative

5.1.2 Logarithmic Functions

Recall that logarithms are defined only for positive values of a variable. Also recall, that the logarithm of a variable is a monotonic function of that variable. Specifically, it is an ever-increasing function. Therefore, we should expect the derivative to be positive for all values of the variable. The command `diff(100*log(50*x), x)` yields the result $\frac{100}{x}$. Remember that the *Maxima* command `log(x)` refers to the natural logarithm, not the common logarirthm.

The result is consistent with our expectation. The result tells us two more things about the derivative: unlike the original function, the derivative is monotonic but ever-decreasing, and its value is independent of the coefficient of x. The second result reflects the fact that and the derivative of the constant log(50) is zero. As before, the graphs in Figure 5.3 confirm that both the initial function and its derivative are monotonic.[1]

5.1.3 Exponential Functions

Exponential functions can exhibit either growth or decay. In either case, both the function and its derivative are monotonic. For a function showing growth (positive exponent), the derivative is positive, monotonic, and accelerating, as Figure 5.4 shows. For a function that shows decay, the derivative is negative but increasing toward zero. The rate of approach declines as x increases. (Again, commands are not reported.)

5.1.4 Product and Quotient Rules

Consider the product and then the ratio of these two terms: $\sqrt{(x)}$ and $e^{0.05 \cdot x}$. The next two exhibits show the functions, the derivatives of the component parts, and the derivative of the product or ratio. In each case, the third line contains two equivalent statements. The first is the direct result of applying the rule, and the second is the result of stating that expression in canonical form. Deriving the expression on the third line of each output from the material on the first two lines is left as an exercise.

[1] The commands to generate this figure are much the same as the ones used previous and are, therefore, not shown.

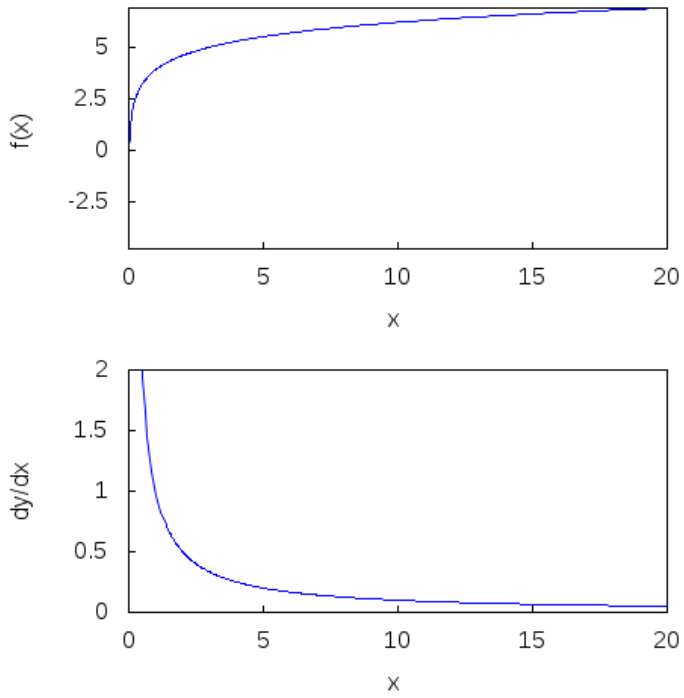


Figure 5.3: A logarithmic function and its derivative

These commands

```
y:sqrt(x)*exp(.05*x);[diff(exp(.05*x),x),diff(sqrt(x),x)];
[dydx:  diff(y,x), radcan(dydx)];
```

produces this output

$$\sqrt{x}\,\%e^{0.05x}$$

$$[0.05\%e^{0.05x}, \frac{1}{2\sqrt{x}}]$$

$$[0.05\sqrt{x}\,\%e^{0.05x} + \frac{\%e^{0.05x}}{2\sqrt{x}}, \frac{(x+10)\ \%e^{\frac{x}{20}}}{20\sqrt{x}}].$$

To analyze the ratio, use these commands

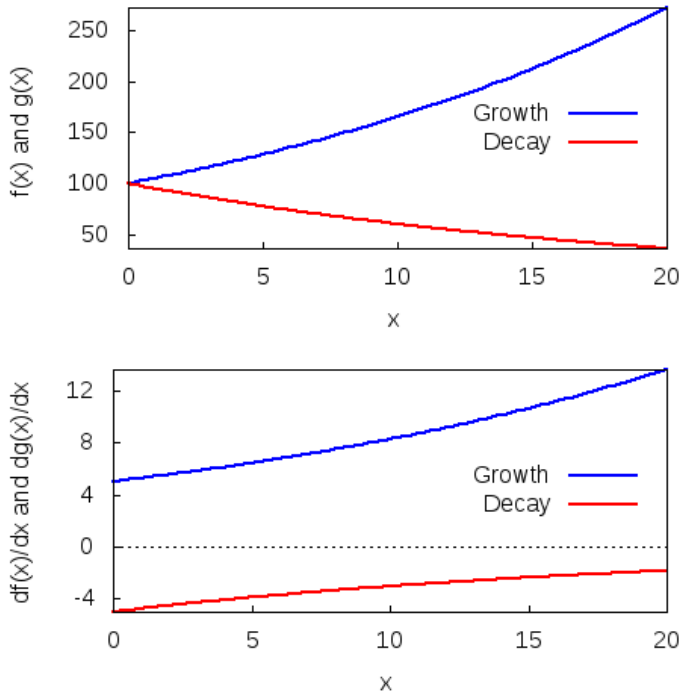


Figure 5.4: Two exponential functions and their derivatives

```
z:sqrt(x)/exp(.05*x);[diff(exp(.05*x),x),diff(sqrt(x),x)];
[dzdx:  diff(z,x), radcan(dzdx)];
```

to generate these results

$$\sqrt{x}\,\%e^{-0.05x}$$

$$[0.05\%e^{0.05x}, \frac{1}{2\sqrt{x}}]$$

$$[\frac{\%e^{-0.05x}}{2\sqrt{x}} - 0.05\sqrt{x}\,\%e^{-0.05x}, -\frac{(x-10)\ \%e^{-\frac{x}{20}}}{20\sqrt{x}}].$$

Figure 5.5 graphically depicts the behavior of these two expressions and of their derivatives over the indicated ranges. For the product, the function is monotonic and the derivative is always positive. For small values, the function y grows at a decreasing rate, but for larger values the exponential term's rapid growth dominates the damping effect of the relatively slow growth of

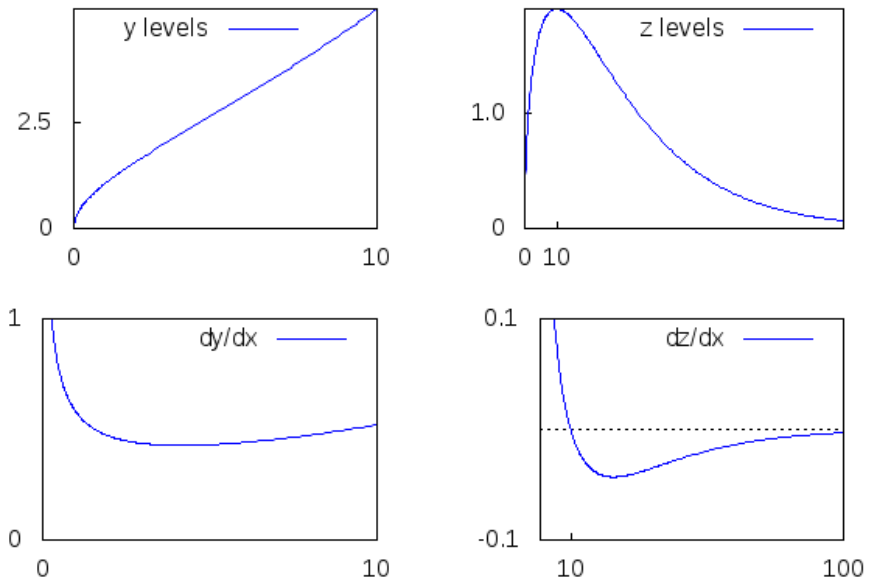


Figure 5.5: Graphs of y and z and of their derivatives

$\sqrt{(x)}$. The ratio of these two functions of x is not monotonic: at first, the growth of the square root term dominates. Eventually, however, the exponential term, now in the denominator, overwhelms the growth in the square root term. The behavior of z is reflected in its derivative, which begins with positive values but falls to zero when , and remains negative thereafter. For large values of x, z is asymptotically approaching 0 as is its derivative (from below).

5.1.5 The Chain Rule

Some of the illustrations above are cases that can be stated like this: $y = f(u)$ and $u = g(x)$. *The chain rule* for deriving $\frac{dy}{dx}$ is $\frac{d(f(x))}{dx} = \frac{d(f(u))}{du} \cdot \frac{d(g(u))}{dx}$. The example below shows a case of this in which $f(u) = \sqrt{u}$ and $g(x) = a + b \cdot x^n$. The expression for which the derivative is to be determined is $\sqrt{b\,x^n + a}$. The commands below determine $\frac{dy}{dx}$ twice, one by using the product defined above and again by instructing *Maxima* to evaluate the expression directly. The first list of commands below defines the two functions and the derivatives.

The second list of commands calculates $\frac{dy}{dx}$ in the two ways indicated above.[2]

```
[y:sqrt(u),dydu:diff(sqrt(u),u),u:a+b*x^n,dudx:diff(u,x)];
[dydu*dudx, diff(''y,x)];
```

The output consists of two lists. The first list shows $y = f(u)$ and its derivative and then $u = g(x)$ and its derivative. The second list contains $\frac{dy}{dx}$ first as a product of two derivatives and then as produced by *Maxima*. Confirm the equivalence of the two expressions.

$$[\sqrt{u}, \frac{1}{2\sqrt{u}}, b\,x^n + a, bn\,x^{n-1}]$$

$$[\frac{bn\,x^{n-1}}{2\sqrt{u}}, \frac{bn\,x^{n-1}}{2\sqrt{b\,x^n + a}}]$$

The chain rule can be extended to any number of functions.

5.1.6 Trigonometric Functions

Business and economic problems involving trigonometric functions are not as frequent as problems that involve the functions that we have developed thus far. We present a few of the rules for differentiating trigonometric functions primarily to point out that *Maxima* can produce derivatives for functions like those in the table below.

[2]The "quote-quote" operator, '', instructs *Maxima* to evaluate y rather than treating it as a noun. Confirm that removing this operator causes the derivative to be evaluated as zero.

Function Name	Function	Derivative
sine of x	$\sin(x)$	$\cos(x)$
sine of u	$\sin(u)$	$\cos(u)\,\text{derivative}(u, x, 1)$
cosine of x	$\cos(x)$	$-\sin(x)$
tangent of x	$\tan(x)$	$\sec(x)^2$
sine of x squared	$\sin(x)^2$	$2\cos(x)\sin(x)$
sine of the square of x	$\sin(x^2)$	$2x\cos(x^2)$
sine of x times cosine of x	$\cos(x)\sin(x)$	$\cos(x)^2 - \sin(x)^2$

5.1.7 Inverse Functions

Assume that the function $y = f(x)$ is either strictly increasing or strictly decreasing and continuous on an interval (a, b). If the function $y = f(x)$ is such that permissible values of x always uniquely determine specific values of y, then the function $y = f(x)$ has an inverse function of the form $x = f^1(y) = g(y)$. That is, the function $y = f(x)$ serves to define a new function g whose value at each point y is the number x such that $y = f$. This means that

not only does a given value of x yield a unique value of y because $y = f(x)$, but also that a given value of y yields a unique value of x because $x = g(y)$. There is a one-to-one correspondence between y and x.

Consider the example that appears in the display below. For positive values of x, f is monotonic. The commands below do the following. The commands in the first line specify f and determine the expression for $\frac{df}{dx}$. The commands in the second line define the inverse function g. The third line of commands determines the expression for $\frac{dg}{dy}$ and substitutes f where appropriate to confirm that $dg/dy = \frac{1}{df/dx}$.

```
(%i36)  assume(x>0,y>0) $        f : sqrt(5*x - 300)$        dydx: diff(f,x)$
        solve(y = sqrt(5*x-300), x)$             g: rhs(%[1])$
        dxdy : diff(g, y)$                                   subst(y=f,dxdy)$
        print("The expression for f(x),  x > 0 is ", f, " which is montonic")$
        print("Use the chain rule to determine the derivative, df/dx = " , dydx)$
        print("Solve y = sqrt(5*x - 300) for x. This yields g, which is ", g)$
        print("The expression for dg/dy, is ", dxdy, "which, after substitution is ",
           subst(y=f,dxdy),",")$
        print ("     which equals 1/(df/dx)")$
```

The expression for f(x), x > 0 is $\sqrt{5\,x-300}$ *which is montonic*

Use the chain rule to determine the derivative, df/dx = $\frac{5}{2\sqrt{5\,x-300}}$

*Solve y = sqrt(5*x - 300) for x. This yields g, which is* $\frac{y^2+300}{5}$

The expression for dg/dy, is $\frac{2\,y}{5}$ *which, after substitution is* $\frac{2\sqrt{5\,x-300}}{5}$,

which equals 1/(df/dx)

Exercise 4.3

Find dy/dx for the following. Remember that $log(x)$ refers to the natural logarithm, unless another base is specified. Find the derivatives by hand and check your work with *Maxima*.

1. $y = u^2 + 3 \cdot u + 7$ when $u = x^2 - 7$
2. $y = u^2$ when $u = 1/x^2$
3. $y = u^3 + 4$ when $u = x^2 + 2 \cdot x$
4. $y = u^{1/2}$ when $u = \frac{x^2-3}{x^2+4}$
5. $y = u^3 + 4$ when$u = v^2 + 2 \cdot v$ and$v = x^2$
6. $y = (x^2 + 4)^2$
7. $y = \sqrt{3 \cdot x}$
8. $y = (2 \cdot x^3 + 1)^3$
9. $y = (4 \cdot x^2 + x^4 - 1)^5$
10. $y = (\frac{x^2-1}{2 \cdot x^2+1})^3$
11. $y = (x^2 + 3)^4 \cdot (2 \cdot x^3 - 5)^3$
12. $y = \frac{x}{x+1})^5$
13. $y = log_a(x^2 + 1)$
14. $y = log_1(x^2 + 1)^2$
15. $y = log_b(2 \cdot x^3 + 3 \cdot x)$
16. $y = logx$
17. $y = log(x^2)$
18. $y = log(x^2 + 1)$
19. $y = log(3 \cdot x^2)$
20. $y = log(x^2 + x - 1)^3$
21. $y = log(\frac{1+x^2}{x^2-1})^2$
22. $y = \frac{logx}{x}$
23. $y = [log(x^2 + 2)] \cdot (x^3 + 3)$
24. $y = a^x \cdot x^a$
25. $y = a^{1+x^2}$
26. $y = 3^{x-2}$
27. $y = log(1/x)$
28. $y = x^2 \cdot log(x)$
29. $y = a^{1+x^2}$
30. $y = 3^{x-2}$
31. $y = e^x$
32. $y = a^x \cdot e^x$

5.2 Higher-Order Derivatives

The function that we obtain when taking the derivative of some function, $f(x)$, is also a function of x. This function may also have a derivative. This process can be continued indefinitely. Some functions have a non-zero derivatives up to a finite order; others have a limitless number of such derivatives. The next exhibit shows one of each. A cubic function has three non-zero derivatives; an exponential growth function has a limitless number. We look at the first three derivatives for each.

The exhibit below shows the commands required to determine the relevant derivatives and to create a matrix that displays these derivatives, along with the resulting output matrix. The first line of commands defines the two functions, y as a function of x and z as a function of t (time). The next two lines create the first-order through fourth-order derivatives for y. Note the syntax of the `diff( )` command: insert the expression, then the independent variable, and finally the order of the derivative. Specifying the order is optional for first-order derivatives. The following two lines do the same for z. Finally, the last three lines create the output matrix output.

```
y : a*x^3 + b*x^2+c*x$    z : A*exp(r*t)$
diffy : diff(y,x)$  diffy2:diff(y,x,2)$
    diffy3: diff(y,x,3)$ diffy4: diff(y,x,4)$
diffz : diff(z,t)$  diffz2:diff(z,t,2)$
    diffz3: diff(z,t,3)$ diffz4: diff(z,t,4)$
matrix(["Order","y's derivatives", "z's derivaties"],
    ["0",y, z], ["1", diffy,diffz], ["2", diffy2,diffz2],
    ["3",diffy3,diffz3], ["4",diffy4,diffz4]    );
```

Order	y's derivatives	z's derivaties
0	$a\,x^3+b\,x^2+c\,x$	$\%e^{r\,t}\,A$
1	$3\,a\,x^2+2\,b\,x+c$	$r\,\%e^{r\,t}\,A$
2	$6\,a\,x+2\,b$	$r^2\,\%e^{r\,t}\,A$
3	$6\,a$	$r^3\,\%e^{r\,t}\,A$
4	0	$r^4\,\%e^{r\,t}\,A$

The "0^{th} order" derivative is the original function. The fourth-order derivative of y is zero, as is any higher-order derivative. All derivatives of z are positive if $r > 0$. Use the rules developed above to confirm that the entries in each row are the derivatives of the expressions in the preceding row.

We use the following notation to indicate order: $\frac{d^n y}{dx^n}$ is the nth-order derivative of y with respect to x. Because *Maxima*'s commands are entered as text, we adjust by using names like the ones in the preceding exhibit: `diffy2` is $\frac{d^2 y}{dx^2}$, and so forth.[3] In later references, we simplify "first-order derivative" to "first derivative" and use similar references to higher-order derivatives. We seldom have occasion to refer to fourth derivatives or higher.

The exhibit above shows a cubic function and three derivatives (four, actually, but the fourth derivative is 0). Figure 5.6 shows an example. It graphs the function and each of the three derivatives. The expression for the function and each derivative appears above the graph.

Suppose that this cubic function defines the amount of output a firm produces each week as a function of worker hours (maybe in thousands). Increasing employment up to just over seven units increases output. Beyond that employment level, total output falls. Now look just below this graph at the graph of its first derivative. That function defines the rate of output change as employment increases. At first the derivative is both positive and

[3]The names are arbitrary. Any legal name can be entered.

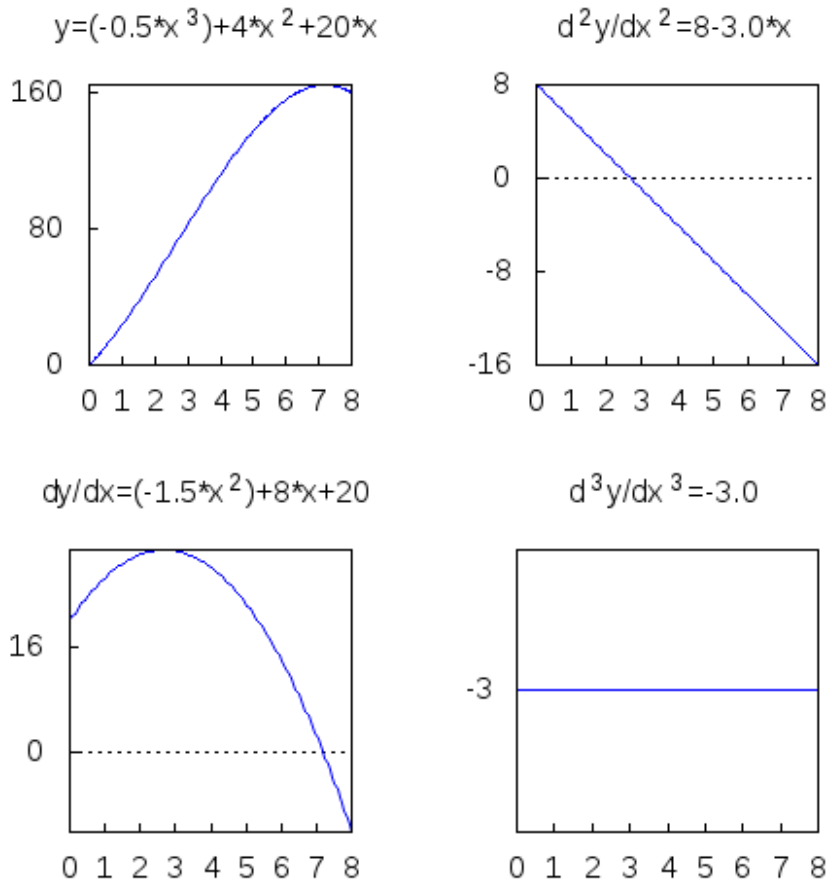


Figure 5.6: A cubic function and three derivatives

increasing. For x greater than 2.5 or so, the derivative remains positive but decreases. Look again at the original function and observed that for small values of x, $f(x)$ is accelerating, rising at an increasing rate. For larger values of x, $f(x)$ decelerates and finally decreases in value. An exercise: determine the value at which the first derivative equals zero.

The second derivative's graph (upper right panel) shows that the first derivative changes sign at $x = 8/3$. The second derivative's derivative (the third derivative of f, is a constant. The derivative of a constant is zero, so the fourth and all higher-order derivatives for this function are zero.

These graphs are in separate panels rather than in a single graph for an

important reason. The units of the functions are quite different. Suppose that the original function does represent production. Then the original function is expressed in physical units per time period. The first derivative is in physical units per time period per worker hour. The second derivative is in physical units per time period per worker hour per worker hour, and so forth. The units for higher-order derivatives can become quite unwieldy. Fortunately, as previously noted, we seldom go beyond the third derivative.

Exercise 4.4
Determine the following for each of these functions: $\frac{dy}{dx}$, $\frac{d^2y}{dx^2}$, and $\frac{d^3y}{dx^3}$.

1. $y = 8 \cdot x^3$ 2. $y = 8 \cdot x^{1/2}$ 3. $y = x^4 + x^2 + 1$
4. $y = e^x$ 5. $y = log(x)$

5.3 Economic Applications of Derivatives

As noted, the concept of the derivative is central to economic analysis. This section introduces the application of derivatives to a set of microeconomic topics, to a macroeconomic topic, and to issues involving growth. Many of these topics will be revisited in later chapters, where we extend the analysis that appears here.

5.3.1 Production, Cost, Revenue, Profits, and Elasticity

This section looks at production in the short run and relates production to cost. It then looks at total cost curves and per-unit cost curves. Next, demand for the firm's product and revenue are introduced. Then we look at profit maximization. Finally, the concept of elasticity is introduced and demand elasticity is related to profit maximization.

The Production Function

Many important general insights relating to production can be garnered by referring to a specific function that represents production in a relatively abstract manner. Suppose for now that output q is determined by a function $q = f(L)$, where L (labor) is a variable input. Economics often refer to

this function as the total product of labor (TPL) function. Two important per-unit functions can be derived from this function. The output per unit of labor, or average product of labor is $APL = q/L = f(L)/L$. The change in output per one (small) unit change in labor is the marginal product of labor, $MPL = \frac{dq}{dL} = \frac{df(L)}{dL}$.

The law of diminishing returns states that, beyond some employment level, MPL begins to decrease. That is $\frac{dq}{dL}$ decreases. Chapter 6 returns to this analysis, extending it to cases in which more than one input is treated as variable.

One of the simplest of such production functions is the Cobb Douglas function: $q = A \cdot L^a \cdot K^{1-a}$, where q is the amount that the firm produces each period, L is the number of worker units employed each period, and K is the number of units of capital employed each month. The coefficient A reflects the level of technology and is constant at any point in time. For our purposes, units do not matter and are ignored.[4] In input/output exhibit below shows the total product, average product, and marginal product functions for a Cobb-Douglas production function. A little manipulation reveals that the APL and MPL functions can be stated as follows: $APL = A \cdot \left(\frac{K}{L}\right)^{1-a}$ and $MPL = a \cdot A \cdot \left(\frac{K}{L}\right)^{1-a}$, so that $MPL = a \cdot APL$. This simple relationship between APL and MPL is a characteristic of this function and does not generalize to all production functions.

```
f : A*L^a*K^(1-a)$ APL: f/L$   MPL : diff(f,L)$
table_form([
["Total Product","Average Product","Marginal Product"],
     [f, APL, MPL]])$
```

Total Product	Average Product	Marginal Product
$A K^{1-a} L^{a}$	$A K^{1-a} L^{a-1}$	$a A K^{1-a} L^{a-1}$

The graphs in Figure 5.7 show the total and per-unit functions. One aspect of this function that might cause concern is that MPL decreases from the beginning and is always below APL. In principles classes you probably saw a different configuration, with MPL rising at first, until diminishing marginal product sets in. The failure of the Cobb-Douglas to generate this result is a

[4] "Our purposes" are purely illustrative. For applications to specific issues, of course, knowing the units is crucial.

failure but probably not a serious one. Most production occurs in the range where diminishing marginal product is present; in fact, in the range for which $APL > MPL > 0$. For the Cobb-Douglas case, $APL > MPL > 0$ for all values of L.[5]

```
(%i62)  [f0,APL0,MPL0]:subst([A=1,K= 1000, a = 0.7], [f,APL,MPL]);
        total: gr2d(xlabel="L", ylabel="Units produced, q",
            line_width=2, explicit(f0,L,0.1, 500)   )$
        perunit: gr2d(line_width=2, xlabel="L",
            ylabel="q per unit of labor",yrange=[0,4],key="APL",
            explicit(APL0,L,0.1, 500), color=orange, key="MPL",
            explicit(MPL0, L,  0.1, 500))$
        wxdraw( total,perunit),wxplot_size=[480,480]$
```

(%o59) $[7.9432\,L^{0.7}, \frac{7.9432}{L^{0.3}}, \frac{5.5602}{L^{0.3}}]$

(%t62)

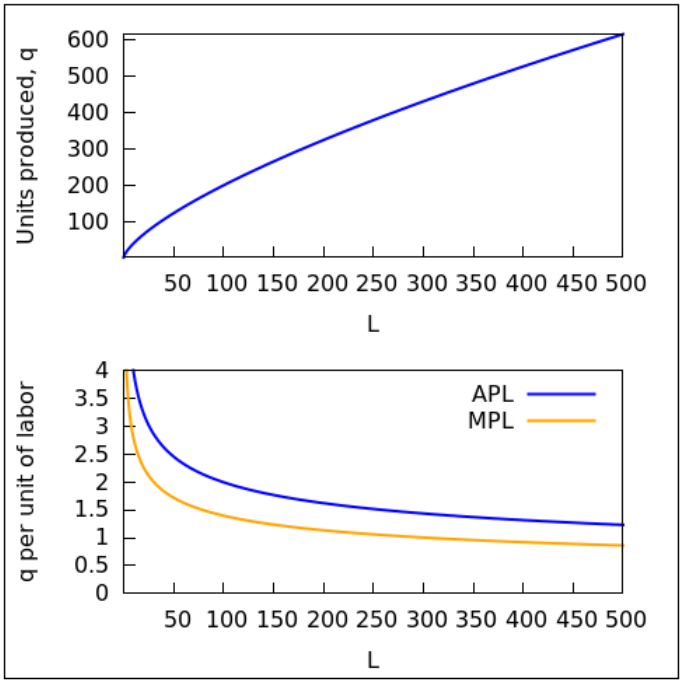


Figure 5.7: TPL, APL, and MPL for Cobb-Douglas production function

[5] In the *Maxima* commands, the L range does not begin at 0 for the per-unit functions. The values of APL and MPL are not defined for $L = 0$.

Production and Cost

For relatively simple production functions, we can derive the firm's total cost function from the production function. We demonstrate the relationship using the Cobb-Douglas function. The resulting cost function is somewhat inflexible, owing to the fact that the Cobb-Douglas function is homogeneous. See [7] for a more flexible specification.

For some purposes, we posit a functional form for a cost function without tying it to a production function. This *ad hoc* approach provides flexibility but gives up some analytical rigor. We illustrate this approach with a cubic cost function.

Costs for Cobb-Douglas Production. For the Cobb-Douglas function we can derive cost curves that express total cost (and both average and marginal cost) explicitly in terms of the quantity produced.[6]

To derive a total cost curve requires three steps. First, the command `solve(f = q,L)` relates quantity to output, with this result:

$$[L = \frac{q^{\frac{1}{a}} K^{1-\frac{1}{a}}}{A^{\frac{1}{a}}}].$$

Now, define the total cost function with the command `TC: subst%, w*L + r*K)`, which uses the output from the previous command to substitute for L in the expression $TC = w \cdot L + r \cdot K$, where w and r are unit costs of labor (L) and capital (K). The resulting expression is this rather daunting expression:

$$\frac{q^{\frac{1}{a}} w\, K^{1-\frac{1}{a}}}{A^{\frac{1}{a}}} + rK.$$

Finally, substitute values for K, w, r, and the production function parameters, using the command `subst([K=1000,r=1,w=1,A=1,a=.7], TC)`.[7] The resulting expression $0.051794q^{1.4285} + 1000$, is quite easy to interpret. The fixed cost is 1000 per period, and variable cost increases by 1.4285 percent per one-percent change in q (reflecting the uniformly decreasing marginal returns that we noted earlier).

[6]For some other production functions, such explicit cost functions might not exist. Often, however, cost curves can be drawn using a parametric approach, with one of the inputs serving as the parameter.

[7]Any values would do for K, r, w, and A. The value of a must be between 0 and 1.

Figure 5.8 shows the per-unit cost curves for this production function (see the workbook for more detail). Both the average variable cost and the marginal cost increase monotonically.

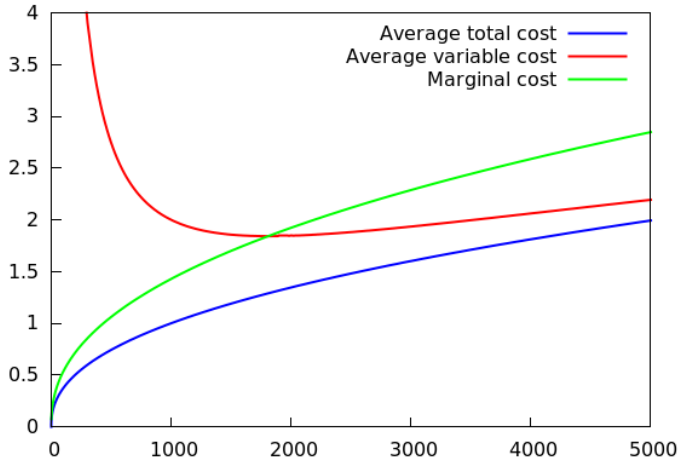


Figure 5.8: Per-unit cost curves for a Cobb-Douglas production function

A Cubic Cost Function. In actual production, increasing marginal returns may occur over some range before diminishing marginal returns set in. Therefore, both average variable cost and marginal cost curves may be U-shaped. A cubic total cost function is the simplest one that can illustrate such cost curves. We use such a curve for the following analysis.[8] The cost functions are as follows:

Total Cost	Average Cost
$0.05q^3 - 1.7q^2 + 25q + 150$	$0.05q^2 - 1.7q + \frac{150}{q} + 25$

Average Variable Cost	Marginal Cost
$0.05q^2 - 1.7q + 25$	$0.15q^2 - 3.4q + 25$

Figure 5.9 shows the curves that these expressions generate. Both TC and TVC increase throughout the range of production. For small output rates, the rate of cost increase diminishes, but after $q \approx 10$ the rate increases. The

[8] We could begin with a cubic production function and establish the relationship between output and cost, but the process is rather involved. See Hammock and Mixon [7].

MC curve reflects this aspect of the TC curve, decreasing for small q and increasing for larger q. Both AC and AVC exhibit the U-shape that economic theory suggests one should observe. The two converge as q increases, with the vertical distance between them equally the value of AFC. The AFC curve contains no useful analytical information and will be omitted from subsequent analysis.

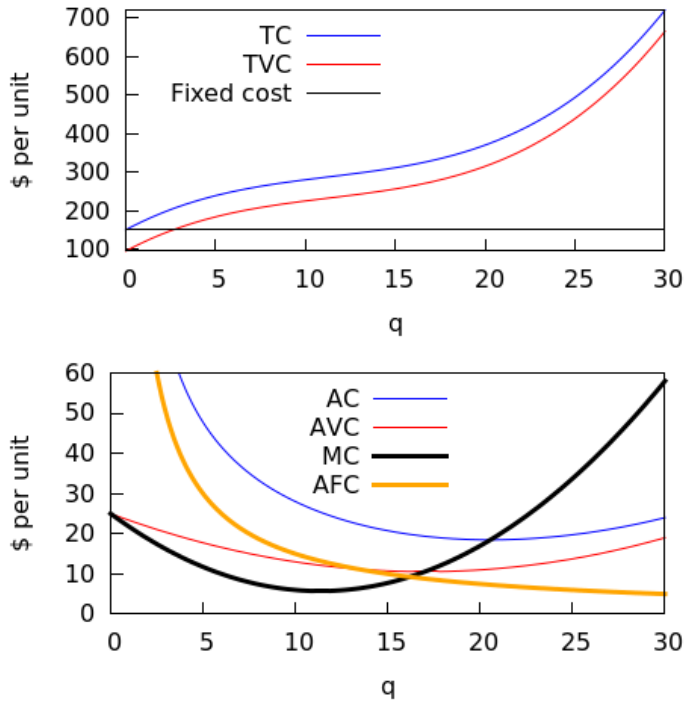


Figure 5.9: Total and Per-unit cost curves, cubic cost function

Figure 5.9 reveals three quantities that can be of interest: the quantity at which MC's sign changes, the quantity at which $MC = AVC$ (and AVC achieves is minimum value), and the quantity at which $MC = AC$ (and AC achieves its minimum value. We see below that these values are approximately 11.333, 17, and 20.551 respectively. In the following analysis, require that *Maxima* return decimal representations (via the command `numer:true`).

The first input line defines the slope of MC twice, once by taking the first derivative of MC and then by taking the second derivative of TC. As must be the case, the two results are identical. The commands are these: `[MCslope:`

`diff(MC,q), diff(TC,q,2)]`. The result is this list of two identical items: $[0.3q - 3.4, 0.3q - 3.4]$.

The second input line contains three applications of `solve( )`, to determine the three values indicated above. The result of `solve(MC = AVC...)` is of note, because *Maxima* produces a list of values, only one of which is real. The commands list is as follows: `soln1:solve(MCslope,q); soln2: solve(MC = AVC,q); soln3:solve(MC=AC,q);`. The first command in the list yields $[q = 11.33]$, the quantity at which the marginal cost curve reaches its minimum value. The second command yields $[q = 0, q = 17]$. The first solution reflects a peculiarity of the cubic functional form and has no economic import. The second shows the quantity at which AVC reaches its minimum value.

The final input item in the list of commands, `solve(MC=AC,q`, determines the quantity at which the marginal cost curve intersects the average cost curve: $q = 20.55]$.[9]

Demand and Revenue

The general expression for the demand curve is $q = f(p)$, where f shows the quantity demanded at each price. Often, it is easier to work with the inverse function, $p = g(q)$. An important reason for using the inverse demand curve is that total revenue is $TR = p \cdot q$. The illustration here treats a case in which a homogeneous product is sold in a single market and the same price applies to each unit. More complex demand curves can be specified with suitable extensions of this simple representation.

The illustration in this section uses two types of demand curves, linear and constant-elasticity (the concept of elasticity is the one you encountered in your principles course and which the next section revisits). The demand curves are of this form: $q = a + b \cdot p$ and $q = A \cdot p^e$. In the first specification, a is the horizontal intercept (the quantity demanded when $p = 0$ and b is the demand curve's slope ($b < 0$). In the second specification, A is a constant (the quantity when $p = 1$) and e is the price elasticity of demand ($e < 0$). For this specification, no horizontal intercept exists.

[9] As the accompanying workbook shows, this command produces three solutions, only one of which is real and, therefore, of consequence for this analysis.

The three lines of input below show expressions for two demand curves, named `fL` and `fCE`. The second input line provides *Maxima* with information it requires in determining the inverse functions. The third input line produces expressions for the inverse demand curves.

```
 [fL: a + b*p, fCE: A*p^e];
declare(e, noninteger)$ assume(q>0, A>0)$
[solve(fL=q,p), solve(fCE=q, p)];
```

The resulting output consists of two lists. The first, $[bp + a, p^e A]$, is the original expressions for the quantity demanded. The second, $[[p = \frac{q-a}{b}], [p = \frac{q^{\frac{1}{e}}}{A^{\frac{1}{e}}}]]$, is the pair of inverse demand curves.

The inverse of the linear demand curve can be written as $p = (a/b) + (1/b) \cdot q$, so that (a/b) is the vertical intercept (graphs appear below) and $(1/b)$ is the inverse demand curve's slope. For the constant-elasticity demand curve, the inverse is $p = (1/A^(1/e)) \cdot q^(1/e)$. The following input shows the inverse demand curves and the associated total revenue curves and marginal revenue curves. We use functional notation, `pL(q,a,b)` for example, to facilitate graphing. The first two input lines specify the inverse demand curves. The second pair of lines specifies total revenue functions. The "quote-quote" operator (`''`) is used to force *Maxima* to evaluate the products in the second pair of commands and the derivatives in the third pair.

```
pL(q, a, b) := a/b + q/b$
pCE(q,A,e) := q^(1/e)/A^(1/e)$
TRL(q, a,b):=''(expand(q*pL(q,a,b)))$
MRL(q,a,b) := ''(diff(TRL(q,a,b),q))$
TRCE(q, A,e):=''(q*pCE(q,A,e))$
MRCE(q,A,e) :=''(diff(TRCE(q,A,e),q))$
```

The resulting expressions are these:

$$\begin{bmatrix} & Linear,\, L & Constant\, Elasticity,\, CE \\ Price & \frac{q}{b} - \frac{a}{b} & \frac{q^{\frac{1}{e}}}{A^{\frac{1}{e}}} \\ Total\, Revenue & \frac{q^2}{b} - \frac{aq}{b} & \frac{q^{\frac{1}{E}+1}}{A^{\frac{1}{E}}} \\ Marginal\, Revenue & \frac{2q}{b} - \frac{a}{b} & \frac{\left(\frac{1}{e}+1\right) q^{\frac{1}{e}}}{A^{\frac{1}{e}}} \end{bmatrix}.$$

We specify a set of parameters: $a = 60$ and $b = -2$ for the linear demand curve, and $A = 1200 and e = -1.5$ for the constant-elasticity demand curve. Figure 5.10 shows the graphical representation of these two inverse demand curves and the associated marginal revenue curves.

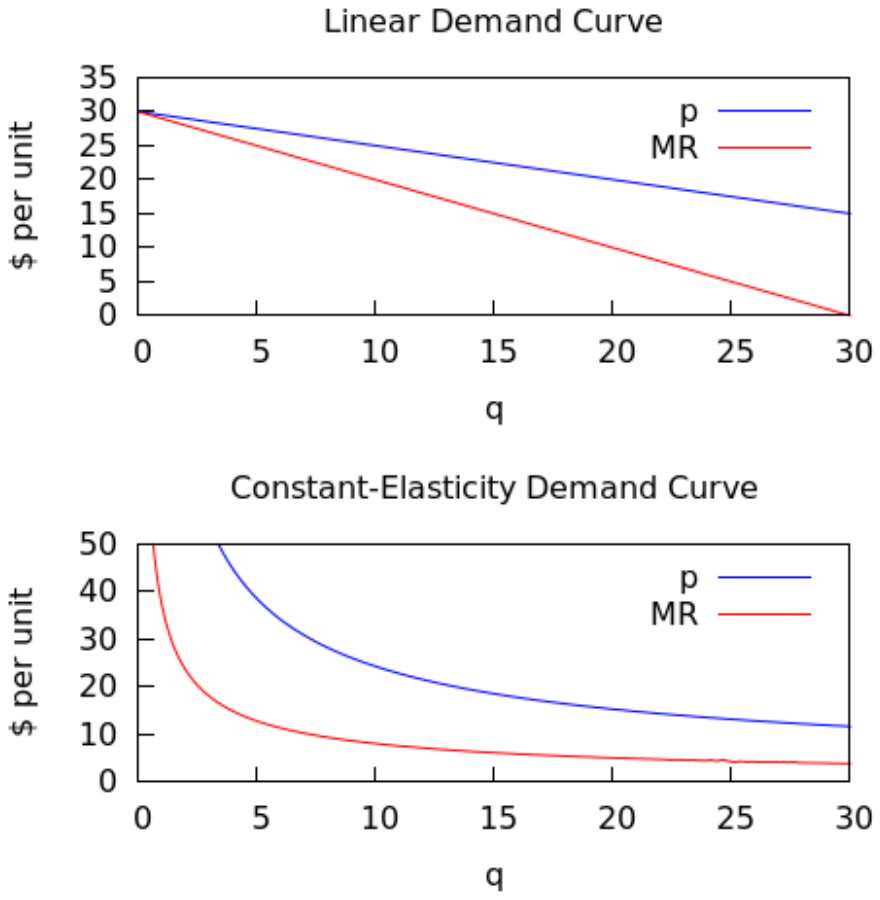


Figure 5.10: Inverse demand curves and marginal revenue curves

Profits

Recall from principles of microeconomics that producing the quantity at which $MR = MC$ yields maximum profit for a firm. We combine the cost information above with the CE demand curve to determine how this firm's

profit level relates to quantity and, in doing so, we confirm that the equality of MR and MC does correspond to maximum profit (or minimum loss).

Figure 5.11 shows the total and per-unit curves that relate to the cost function above, combined with the constant-elasticity demand curve.[10] The top panel, which shows total values, reveals that the firm earns losses at all output levels because total revenue lies below total cost. The firm does, however, cover its variable costs so that producing at a rate between 0 and just under 25 units results in a lower loss than producing nothing and losing the per-period fixed cost, \$150. Minimum loss appear to occur at approximately $q = 15$. We determine the actual value below.

The per-unit curves tell the same story. The inverse demand curve (p) is below average cost but above average variable cost. Maximum profit (minimum loss here) occurs where the marginal revenue and marginal cost curves intersect, apparently around $q = 14$. The combination of a polynomial cost function and constant-elasticity demand curve does not yield an analytical solution, so we cannot directly solve $MR = MC$ to determine the profit-maximzing quantity. Rather, we use a numerical root-finding option in this command: `xOptimal: find_root(MC-MRCE(q,1200,-1.5),q,5,15)`, which yields a value of $q = 13.701$. Using the command `pCE(xOptimal,1200,-1.5)` to plug the optimal quantity into the demand curve shows that the loss-minimizing price is \$19.72. Finally, the command `PCE(xOptimal,1200,-1.5)` yields the result that the profit level is -\$31.78 per time period.

[10] See the accompanying workbook for the commands that generate these graphs.

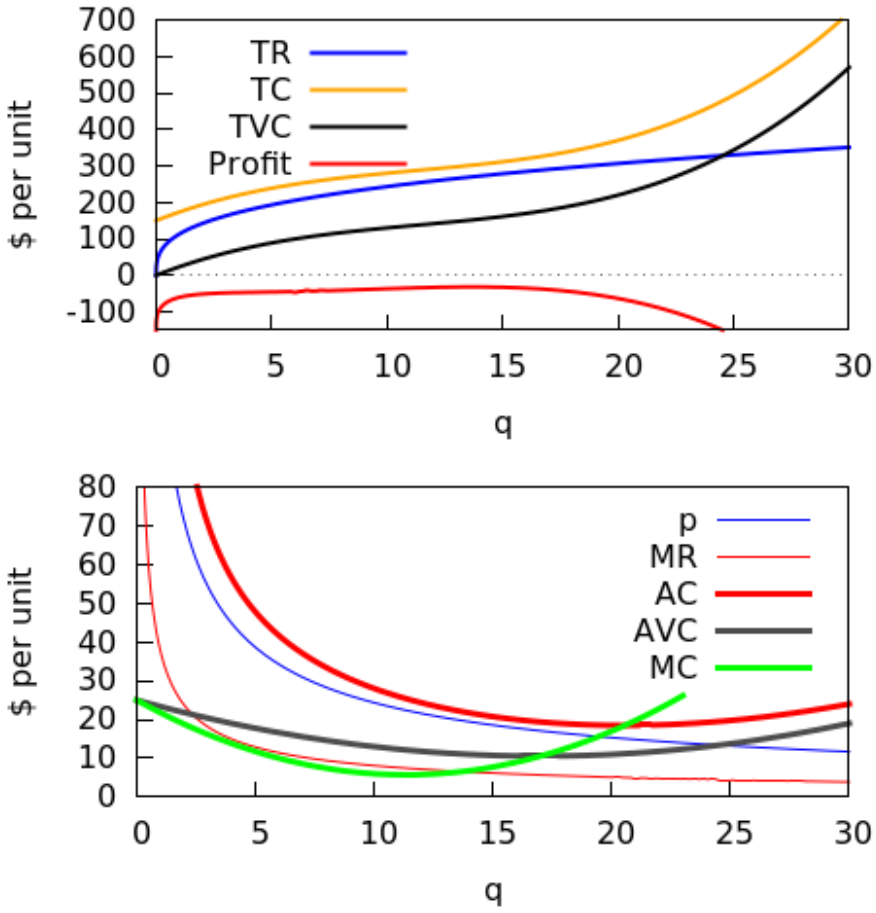


Figure 5.11: Revenue, costs, and profit

Chapter 6

Differentiation: Multivariate Functions

The material up to this point is developed mostly in terms of functions of a single variable. This chapter extends the analysis to functions that involve several variables. It focuses on differentiation, but in doing so it addressed pertinent aspects of functions of more than one variable. Fortunately, most of what we have learned up to this point applied with relatively little modification.

6.1 Partial Differentiation

This chapter addresses functions of two or more independent variables. Consider $z = f(x, y)$, where z is the dependent variable and x and y are independent variables. Given any permissible values of x and y, we can determine a value for z. For example, if $z = x \cdot y$, and $x = 2$ and $y = 4$, then $z = 8$.

This chapter deals with functions of the form $y = f(x1, x2, ..., xn)$, where the second part of each independent variable's name distinguishes it from the other independent variables. That is $x1$ is a variable, not a value of a variable named x.[1]

[1]Recall that we use this approach to naming to accommodate the fact that computer algebra systems typically require that data entry involves text characters. Often in later material, we use x_1, rather than $x1$, where the subscript adds clarity.

The execution of partial differentiation, also called taking a partial derivative, is quite similar to its single-variable counterpart, and we proceed in the same way. Suppose that we wish to examine the effect of a change in the value of $x2$ with all other independent variables keeping their initial variables. Then differentiation with respect to $x2$ alone involves taking the limit of a difference ratio as follows:

$$\frac{\partial y}{\partial x2} \equiv \lim_{\Delta x2 \to \infty} \frac{\Delta y}{\Delta x2} = \lim_{\Delta x2 \to \infty} \frac{f(x1, x2 + \Delta x2, \cdots, xn) - f(x1, x2, \cdots, xn)}{\Delta x2}.$$

The derivative taken in this expression measures the rate of change of y with respect to $x2$ and is referred to as the *partial derivative* of y with respect to $x2$. The partial derivative indicates the effect of a change in a single independent variable on the dependent variable, with all other independent variables in the function held constant while this particular derivative is taken.

The process of partial differentiation is denoted by the variant form of the lower-case Greek letter delta (δ), namely ∂. Suppose that we have a function of the form $y = f(x1, x2)$. We represent the partial derivative of y with respect to $x1$ by $\frac{\partial y}{\partial x1}$. Notation regarding partial derivatives is not completely standard. Any of the following can be used to mean the same as $\frac{\partial y}{\partial x1}$: $\frac{\partial y}{\partial x1}$, f_1, f_{x1}, y_1, or y_{x1}. The context usually makes the meaning clear.

6.2 Rules of Differentiation

Fortunately the rules of differentiation that we derived previously can also be used to find partial derivatives when we deal with functions of several independent variables. The major alteration in our differentiation procedure is that all independent variables not explicitly involved in the differentiation are treated as constants and are differentiated accordingly. *Maxima*'s command reflects this similarity: the same `diff()` command applies to both simple derivatives and partial derivatives.

Look at two examples. The first is $u = x^2 + 4 \cdot x \cdot y + y^2$. The commands `u_x : diff(u(x,y), x)` and `u_y: diff(u(x,y), y)` generate this result:

$$[4y + 2x, 2y + 4x].$$

We use `u_x` and `u_y` as names for these partial derivatives, $\partial u/\partial x = 4y + 2x$ and $\partial u/\partial y = 2y + 4x$. It is apparent in this case that the value of both of the partial derivatives depend on the values of both x and y.[2]

For the next example, $v = 2 \cdot y/x + 4 \cdot x/y$, the partial derivatives are also functions of both variables, as the table below shows. The table below shows the original function and the two partial derivatives. Both $\partial v/\partial x$ and $\partial v/\partial y$ are rather involved functions of both variables, x and y. When we look at higher-order partial derivatives, we will discover a way to determine more about how values of the partial derivatives change as x and y change.

$$\begin{bmatrix} \texttt{v(x,y)} & \texttt{v\_x} & \texttt{v\_y} \\ \frac{2y}{x} + \frac{4x}{y} & \frac{4}{y} - \frac{2y}{x^2} & \frac{2}{x} - \frac{4x}{y^2} \end{bmatrix}$$

We can generalize the results of the previous examples. Suppose we have a function $y = f(x1, x2 \ldots, xn)$, where the x's are independent variables that are independent of one another. It is in principle possible to compute a partial derivative of y with respect to each of the x's. A total of n such partial derivatives exist.[3]

Exercise Set 6.1.
In each case compute all first-order partial derivatives. Then confirm your results using *Maxima*.

1. $f(x, y) = x^4 + y^4 - 4 \cdot x^2 \cdot y^2$
2. $f(x, y) = x \cdot y + x/y$, for $y \neq 0$
3. $f(x, y) = \sqrt{x^2 + y^2}$
4. $f(x, y) = (x + y)/(x - y)$ for $x \neq y$
5. $f(x, y) = e^{x^2+x \cdot y}$
6. $f(x, y, z) = x \cdot y \cdot z$
7. $f(x, y) = x/y^2 + y/x^2$
8. $f(x, y) = log_e(x \cdot y + y^2)$
9. $f(x, y) = x \cdot cos(y) + y \cdot cos(x)$
10. $f(x, y) = \sqrt{x \cdot y}$
11. $f(x, y) = \frac{x^2/2+3 \cdot y^2}{x \cdot y + y \cdot z}$

6.2.1 Geometric Interpretation of the Partial Derivative

We learned in Chapter 4 that when $y = f(x)$, the derivative of y with respect to x can be given a geometric interpretation: dy/dx is the slope of

[2]This need not be so: consider, for example $z = a \cdot x + b \cdot y^2$. Then $\partial z/\partial x = a$, a constant, and $\partial z/\partial y = 2 \cdot y$, a function of y alone.

[3]The workbook that accompanies this chapter illustrates this point with a function of four variables.

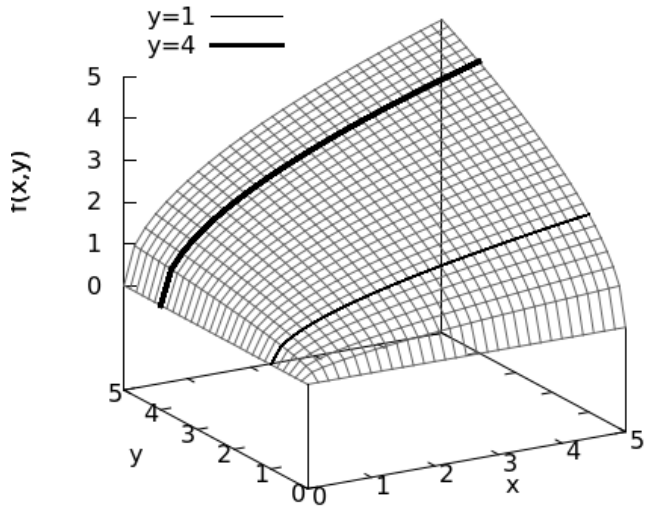


Figure 6.1: Graph of $\sqrt{x \cdot y}$

the curve of that function at a point in a plane. Similarly, a partial derivative has a geometric interpretation. Consider the function of two variables $f(x, y) = \sqrt{x \cdot y}$. The partial derivatives are $\partial f(x, y)/\partial x = (1/2) \cdot \sqrt{y}/\sqrt{x}$ and $\partial f(x, y)/\partial y = (1/2) \cdot \sqrt{x}/\sqrt{y}$. Thus, both partial derivatives are (1) positive and (2) directly related to the value of the other variable. Figure 6.1 shows this surface for a range of x and y values.

The surface in Figure 6.1 confirms our expectation. Holding y at given value defines a line on the surface that shows z values for the range of x values. Likewise, holding x at a given value defines a line that shows z values for the range of y values. As y increases, the level of z for each x value increases; also the slope of the line that shows z relative to x becomes steeper. Likewise, as x increases, z increases at each y value and the slope increases at each value of y. Compare the slopes of the lines labeled "y = 1" and "y = 4" that indicate levels at which y is held constant.

6.2.2 Differentials

One purpose of analysis is to explain, at least in a qualitative sense, how changes in one or more independent variables affect the value of a dependent variable. To approach this question in an intuitive fashion, we begin with a tautology: $\Delta y = (\Delta y/\Delta x) \cdot \Delta x$. As it stands this says that the change in y equals the change in y per unit change in x, multiplied by the change in x. This, as it stands, tells us nothing. Suppose, however, that we replace with its limiting value, dy/dx. Now, we can say that $\Delta y \approxeq (dy/dx) \cdot \Delta x$. This expression says that, at a given value of x, the line tangent to the function approximates the effect of a change in x on y.

One step remains in defining the differential. As Δx becomes a small number, the product $(dy/dx) \cdot \Delta x$ becomes closer to Δy. Thus, for sufficiently small change dx, the expression becomes very nearly accurate. Consider an example. Let $y = x^3$. Change x from 2 to 2.01 and determine both Δy and dy. At $x = 2$, $y = 8$, and $dy/dx = 3 \cdot x^2$. At $x = 2.01$, $y = 2.01^3 = 8.1206$, so $\Delta y = 8.1206 - 8 = 0.1206$. The predicted change using the differential is $dy = 3 \cdot x^2 \cdot 0.01 = 0.12$.

We extend the concept of a *differential* to include functions of two or more independent variables. We begin by noting that a partial derivative measures the rate of change of the dependent variable with respect to an infinitesimal change in one of the independent variables, all other independent variables held constant. A *total differential* of a function, however, is a linear approximation of the rate of change of the dependent variable when all of the independent variables change by an infinitesimal amount.

The total differential is the sum of the changes in the dependent variable caused by simultaneous infinitesimal changes in all the independent variables. Suppose that $y = f(x1, x2, \cdots, xn)$. Then the total differential of y is this: $dy = f_1 \cdot dx1 + f_2 \cdot dx2 + \cdots + f_n \cdot dxn$, where $dx1$, $dx2$, ..., dxn indicate small changes in the n independent variables. These changes occur independently of each other.

Extend the example above by making $z = x^3 + y^2$. The total differential for z is $dz = 3 \cdot x^2 \cdot dx + 2 \cdot y \cdot dy$. Begin with $x = 2$ and $y = 3$, so that $z = 8 + 9 = 17$ and $dz = 12 \cdot dx + 6 \cdot dy$. Let $\Delta x = 0.01$ and $\Delta y = 0.01$. Then the approximation for dy is $12 \cdot 0.01 - 6 \cdot 0.01 = -0.18$, so that the approximate value of y is now 17.18. The actual value is 17.181.

6.2.3 Maxima Notation

To see how *Maxima* treats a differential, we revisit the preceding paragraph. The command `z: x^3 + y^2` creates the expression to be evaluated. To find that initially $z = 17$, use this command: `subst([x=2,y=3],z)`. Likewise, to see that z's value changes to approximately 17.181, use the command `subst([x=2.01,y=3.01],z)`.

Now use the command `diff(z)` to define the total differential of this expression:

$$2y\ \text{del}(y) + 3x^2\ \text{del}(x).$$

We can substitute the values of x and y and evaluate the differential with the command `subst([del(x)=.01, del(y)=.01,x=2,y=3],%)`, getting 0.18 as the result. Note that *Maxima* treats del(x) and del(y) as variables.

Exercise Set 6.2
Find the total differentials for these functions by hand and with *Maxima*.

1. $z = x^2/2 + x^3/3$
2. $z = x^2 + x + y + y^2$
3. $z = log_e(x + y)$
4. $z = x \cdot x \cdot y$
5. $w = e^{x^2} + y^2 + z^2$
6. $w = z^2 \cdot (2 \cdot x + 3 \cdot y)$

6.2.4 Total Derivatives of Composite Functions

The previous section developed the concept of a *total differential*. This section develops the concept of a *total derivative*. The slight difference in terminology represents a substantial difference in the type of function with which we are dealing. Our discussion of partial derivatives and total differentials relied on the assumption that the independent variables were independent of one another. Often, however, such an assumption cannot be fulfilled. This section considers three general cases in which the assumption is violated. It also considers the case in which no dependent variable can be identified but in which two are more variables are bound by an *composite function*.

Case 1: Interrelated Independent Variables
Given the function $z = f(x, y)$, where it also is true that $y = g(x)$. Hence $z = f(x, g(x))$, which is a *composite function* in which x is the only independent variable. An important underlying assumption of a partial derivative is no longer fulfilled when we encounter a function such as $z = f(x, g(x))$. Variable

$y = g(x)$ cannot remain constant when x varies. That is, x and y are not independent of each other. In a case like this, $dz = \partial f/\partial x \cdot dx + \partial f/\partial y \cdot dy$ is the total differential of z. We can divide both sides of this expression by dx to obtain $dz/dx = \partial f/\partial x + \partial f/\partial y \cdot dy/dx$. The term dz/dx is the *total derivative* of z with respect to x.

This total derivative has two parts. The first part, $\partial f/\partial x$, measures the change in z brought about by changes in x, all other variables held constant. The second part, $\partial f/\partial y \cdot dy/dx$ measures the change in z brought about by change in variable x that works through intermediate variable y. The first part of this expression is often called the *direct effect*, while the second part is called the *indirect effect*. The indirect effect takes account of the fact that changes in variable x affect variable y, which in turn affects variable z. In general if $z = f(x, y1, y, \ldots, yn$, then

$$\frac{dz}{dx} = \frac{\partial z}{\partial x} + \frac{\partial z}{\partial y1} \cdot \frac{dy1}{dx} + \frac{\partial z}{\partial y2} \cdot \frac{\partial dy2}{dx} + \cdots + \frac{\partial z}{\partial yn} \cdot \frac{dyn}{dx}.$$

Consider three examples, the first of which we also develop with *Maxima*. Let $z = f(x, y) = x^2 + 2 \cdot x \cdot y + y^2$. The total derivative is $dz/dx = \partial z/\partial x + (\partial z/\partial y) \cdot (dy/dx)$. In our example, $y = g(x) = e^{3 \cdot x}$, so that $dy/dx = 3 \cdot e^{3x}$. We confirm that the rule for evaluating dz/dx yields the same result as inserting $g(x)$ into $f(x, y)$ and evaluating the result. We proceed with three sets of commands, which generate the table below.

$y^2 + 2xy + x^2$	$2y + 2x$	$2y + 2x$
$\%e^{6x} + 2x\,\%e^{3x} + x^2$	$6\%e^{6x} + 6x\,\%e^{3x} + 2\%e^{3x} + 2x$	
$3\%e^{3x}\ (2y + 2x) + 2y + 2x$	$6\%e^{4x} + 6x\,\%e^{3x} + 2\%e^{x} + 2x$	

The table above shows the result of using *Maxima* to determine this total derivative. The first input line defines z without specifying how x and y are related, and it takes the two partial derivatives. *Maxima* responds as if x and y are independent of each other. The commands are these:
`[z : x^2 + 2*x*y + y^2, dzdx01: diff(z,x), dzdy:diff(z,y)]`. The name `dzdx01` is assigned to the derivative with respect to x because, we will determine this derivative two more times in the next two lines of input.

The second input line, `[z:subst(y=exp(3*x),z),dzdx02:diff(z,x)]`, adds the information that y is a function of x and takes the total differential of z, which is the second item in the second line of output.

The third input line, `[dzdx03: dzdx01 + dzdy*diff(%e^(3*x),x), expand(subst(y=%e^x,dzdx03))]`, confirms that $dy = \frac{dy}{dx} \cdot dx$. The value of $\frac{dy}{dx}$ is determined by the command `diff(%e^(3*x),x)` a value that is used in the final step. The `subst` command informs *Maxima* that $y = e^{3 \cdot x}$.[4]

For a relatively simple example like this one, using *Maxima* provides little advantage. In a more complex example, however, *Maxima* can both simplify the process of extracting the total derivative and reduce the probability of errors.

The second example is the function $z = e^{x^2-y^2}$, with $x = 2 \cdot y^3$. For this function, $\frac{dz}{dy} = \frac{\partial z}{\partial y} + \frac{\partial z}{\partial x} \cdot \frac{dx}{dy} = e^{x^2-y^2} \cdot (-2 \cdot y) + e^{x^2-y^2} \cdot (2 \cdot x) \cdot (3 \cdot 2 \cdot x^2) = 12 \cdot x \cdot y^2 \cdot e^{x^2-y^2}$.

Finally, let $w = f(x, y, z) = 2 \cdot x^2 + 3 \cdot y^3 + 4 \cdot z^4$, where $y = e^x$ and $z = log_e x$. In this case, $\frac{dy}{dx} = e$ and $\frac{dz}{dx} = 1/x$, so $\frac{dw}{dx} = f_x + f_y \cdot \frac{dy}{dx} + f_z \cdot \frac{dz}{dx}$, or $\frac{dw}{dx} = 4 \cdot x + 9 \cdot y^2 \cdot e^x + 16 \cdot z^3 \cdot \frac{1}{x}$).

Case 2. "Independent" Variable(s) Affected by One or More External Variables

Chapter 5 introduced explicit expressions in which x and y are both functions of a third variable the "parametric" expressions. The case considered here is similar. Suppose that z is a function of two (or more) such variables. Let $z = f(x, y)$, where $x = g(t)$ and $y = h(t)$. This is another variant of a composite function, where $z = f(g(t), h(t))$. The value of z can change when the value of either x or y (or both) change independently of t (due to a parameter shift, for example) or when a changed value of t causes the change(s) in x and y. We consider two examples.

First, suppose that $f(x, y) = x^2 \cdot y^3$, where $z = \frac{t^2}{2}$ and $y = 3 \cdot t^3$. In this case $\frac{dz}{dt} = f_x \cdot \frac{dx}{dt} + f_y \cdot \frac{dy}{dt} = (2 \cdot a \cdot x \cdot y^3) \cdot t + (3 \cdot x^2 \cdot y^2) \cdot 9 \cdot t^2$. This expression simplifies to $\frac{dz}{dt} = 2 \cdot a \cdot x \cdot y^3 \cdot t + 27 \cdot (x \cdot y \cdot t)^2$.

For the second example, we refer to *Maxima*. The expression to be evaluated is $w = g(x) = e^{x \cdot y \cdot z}$. The command `w: exp(x*y*z)` assigns the name `w` to this expression. With this information alone, the command `diff(w)`

[4]This command is embedded in the `expand` command to make the items in the second and third lines of output easier to compare. It is not required.

produces this output:

$$xy\,\%e^{xyz}\,\mathrm{del}(z) + xz\,\%e^{xyz}\,\mathrm{del}(y) + yz\,\%e^{xyz}\,\mathrm{del}(x).$$

Suppose, however, that $x = s^2 + t^2$, $y = s \cdot t$, and $z = \sqrt{t}$. In this case, none of the three variables can change while the others remain constant. The command `depends([x,y],[s,t],z,t)` tells *Maxima* that x and y depend on both s and t, and that z depends on t alone. It does not specify the nature of the dependencies.

Now the command `diff(w)` produces this output:

$$\%e^{xyz}\left(xy\left(\frac{d}{dt}z\right) + x\left(\frac{d}{dt}y\right)z + \left(\frac{d}{dt}x\right)yz\right)\mathrm{del}(t) +$$

$$\left(x\left(\frac{d}{ds}y\right)z + \left(\frac{d}{ds}x\right)yz\right)\%e^{xyz}\,\mathrm{del}(s),$$

which shows that five derivatives must be evaluated in order to produce dw in terms of s and t. The rather long `subst` command
`subst([diff(x,s)=diff(s^2+t^2,s), diff(x,t)= diff(s^2 + t^2,t), diff(y,s)= diff(s*t,s), diff(y,t)= diff(s*t,t), diff(z,t) =diff(sqrt(t),t) ], diff(w))` provides the necessary information for evaluation of dw. Now the command `diff(w)` generates this result:

$$\left(2tyz + sxz + \frac{xy}{2\sqrt{t}}\right)\%e^{xyz}\,\mathrm{del}(t) + (2syz + txz)\,\%e^{xyz}\,\mathrm{del}(s).$$

We are not quite finished, because x, y, and z are in the expression. One more set of substitutions, `subst([x=s^2+t^2,y=s*t,z=sqrt(t)],%)` gives us the result that we seek:

$$\left(2s\,t^{\frac{5}{2}} + \frac{3s\sqrt{t}\,(t^2+s^2)}{2}\right)\%e^{s\,t^{\frac{3}{2}}\left(t^2+s^2\right)}\,\mathrm{del}(t) +$$

$$\left(t^{\frac{3}{2}}\left(t^2+s^2\right) + 2s^2\,t^{\frac{3}{2}}\right)\%e^{s\,t^{\frac{3}{2}}\left(t^2+s^2\right)}\,\mathrm{del}(s).$$

The coefficient of $del(t)$ is $\partial w/\partial t$, and the coefficient of $del(s)$–on the second line– is $\partial w/\partial s$.[5]

[5]The workbook shows that placing the definitions of s and t directly into w and executing `diff(w)` yields the same result.

Exercise 6.3
Evaluate the derivatives below, first by following the approach used above and then with *Maxima*.

1. For $z = x^2 \cdot y + x \cdot y^2$, when $y = x^3$, find dz/dx.
2. For $z = e^{x^2 \cdot y} \cdot x^2$, when $y = \sqrt{(x)}$, find dz/dx.
3. For $z = log_e x$, when $y = 8 \cdot x + 8$, find dz/dx.
4. For $z = x/y + x \cdot y$, when $x = 2 \cdot t$ and $y = t^2$, find dz/dt.
5. For $z = \sqrt{x + 8 \cdot y}$, when $x = 2 \cdot t + 4$ and $y = t^3 + t$, find dz/dt.
6. For $w = x^2 \cdot y^2 + x \cdot y \cdot z + y^2 \cdot z^2$, when $x = 2 \cdot t,\ y = 2 \cdot t + 2,$ and $z = t^2$, find dw/dt.
7. For $z = \sqrt{x^2 + y^2}$, when $x = x - tF$ and $y = x + t$, find $\partial z/\partial s$ and $\partial z/\partial t$.
8. For $z = (x^2 + y^2)^3$, when $x = s + t$ *and* $y = 25 - t$, find $\partial z/\partial s$ and $\partial z/\partial t$.
9. For $z = log_e(x \cdot x \cdot z)$, when $x = s^2 \cdot t,\ y = s \cdot t^2,$ and $z = s \cdot t$, find $\partial z/\partial s$ and $\partial z/\partial t$.

Case 3. Implicit Functions
Many of the functions used in the study of business and economics are *implicit functions* of the form $F(x, y) = 0$. Examples include isoquants, indifference curves, budget constraints, and iso-cost curves. Also, demand and supply curves can be expressed as implicit functions. Sometimes, as in the case of a demand curve, the implicit function can be rephrased as an explicit function, but this need not be so. For example, an indifference curve that slopes upward over some range (at the margin, on good has become a "bad") cannot be rephrased this way. We encounter some examples below. Standard notation is to express an implicit function of and as follows: $F(x, y) = 0$, with upper-case F replacing the lower-case f.

Reasoning that we have developed implies that $dF = \frac{\partial F}{\partial x} \cdot dx + \frac{\partial F}{\partial y} \cdot dy$. $F(x, y) = 0$, a constant, so dF must equal zero. Therefore, dy/dx must equal $-F_x/F_y$, given that $F_y \neq 0$. This important result is the *Implicit Function Theorem*. As an example, suppose that $F(x, y) = x^2 + y^2 - 16$, which can also

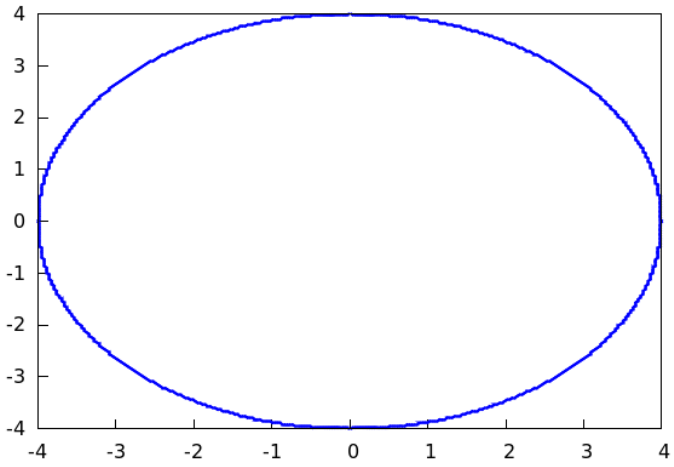


Figure 6.2: An implicit function

be written as $x^2+y^2 = 16$. The table below shows, $F_x = 2 \cdot x$ and $F_y = 2 \cdot y$, so $dy/dx = -x/y$. Thus the function slopes downward when x and y have the same sign and upward when they have opposite signs (notation: in Maxima we use `Fx` and `Fy` to indicate F_x and F_y). The derivative dy/dx is not defined when $y = 0$. Figure 6.2 confirms the observations made above.[6]

```
 table_form([["F", "Fx", "Fy","-Fx/Fy"],
Fx (x,y) := ''(diff(F(x,y),x)),
Fy (x,y) := ''(diff(F(x,y),y)),  -Fx(x,y)/Fy(x,y)] ])$
```

F	Fx	Fy	−Fx/Fy
$F(x,y) := x^2 + y^2 - 16$	$Fx(x,y) := 2x$	$Fy(x,y) := 2y$	$-\frac{x}{y}$

As a second example, suppose that $F(x, y) = e^x + e^y - 20$. Following the same process as before, we can establish that $dy/dx = -e^x/e^y = -e^{x-y}$. Both e^x and e^y are positive, so $dy/dx < 0$ for all values of both variables. Neither variable's value can exceed $log_e 20$ (why?). As x approaches that limit, y must become very small, and vice versa. Figure 6.3 illustrates these conclusions.

[6] If the two variables have the same units, then this is the formula for a circle with radius = 4 and with the center at zero. *Maxima* does not assume that units are the same. If you want the output to look circular, add this option: `proportional_axes=xy`.

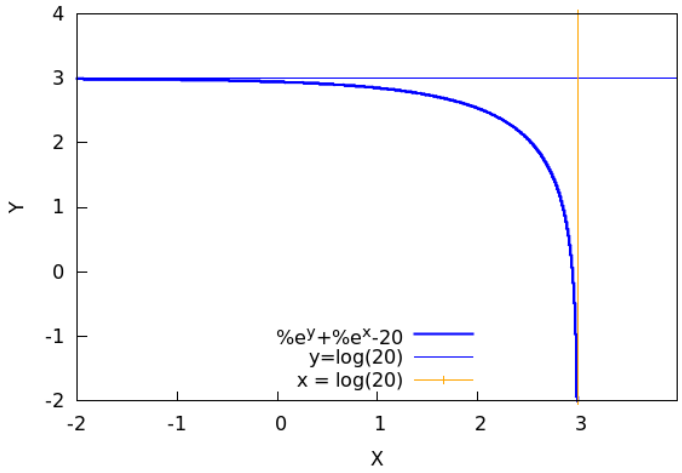


Figure 6.3: Graph of $e^x + e^y - 20$

Finally, consider $F(x, y)y^3 + x \cdot y - 12$. Following the steps in the preceding examples reveals that $dy/dx = -y/((3 \cdot y^2 + x))$, as reported in the following table..

$F(x,y)$	Fx	Fy	dy/dx
$y^3 + xy - 12$	y	$3y^2 + x$	$-\frac{y}{3y^2+x}$

This expression is much more difficult to evaluate than then preceding two. Graphical analysis helps us to understand it better. Figure 6.4 shows that the relationship between x and y is not monotonic. For the values of x and y that yield the upper curve, the derivative is monotonic ($dy/dx > 0$), but for the values that generate the lower curve the derivative is no longer monotonic. If, however, our attention is limited to positive x values, then the implications are simpler: Both the initial function and its derivative are monotonic: dy/dx is negative but it approaches zero as x increases.

Partial derivatives can be defined for implicit functions of more than two variables. If $F(x, y, z) = 0$, then $\partial z/\partial x = -F_x/F_z$. The other two partial derivatives can be defined in like fashion. The next display shows an implicit function of three variables. It also shows F_x and F_y, along with $\partial y/\partial x$. As the entry in the fourth column shows, for this function $\partial y/\partial x$ depends on the values of all three variables.

$F(x,y,z)$	Fx	Fy	$\partial y/\partial x$
$y^2\,z^2 + \sqrt{xy}$	$\frac{y}{2\sqrt{xy}}$	$2y\,z^2 + \frac{x}{2\sqrt{xy}}$	$-\frac{y}{2\sqrt{xy}\left(2y\,z^2+\frac{x}{2\sqrt{xy}}\right)}$

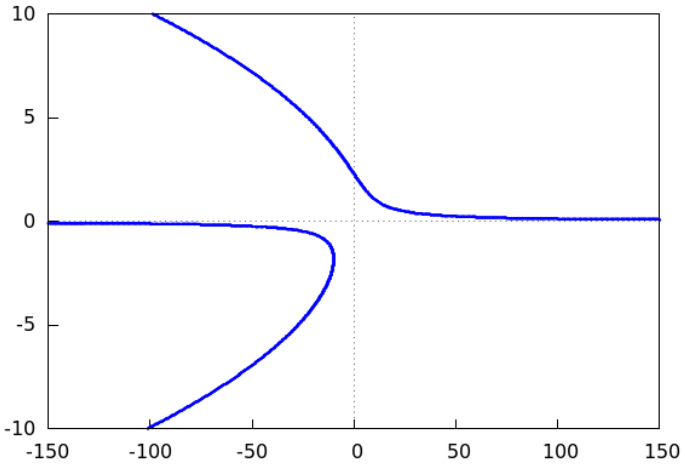


Figure 6.4: Graph of $x^3 + x \cdot y - 12$

6.3 Higher-order Partial Derivatives

Given the differentiable function z = f (x,y), we have seen that the process of partial differentiation produces two new functions, namely $\partial z/\partial x = f_x = \partial f(x,y)/\partial x$ and $\partial z/\partial y = f_y = \partial f(x,y)/\partial y$. Each of these two functions is itself a function of variables x and y, so we can differentiate them once again in order to obtain the rate of change of the partial derivative with respect to either x or y.

The partial derivative of a partial derivative is referred to as a *higher-order partial derivative*. Standard notation for the case in which we take a "second-order partial derivative" (the partial derivative of a partial derivative) is as follows:

$$z_{xx} = \frac{\partial}{\partial x}\left(\frac{\partial z}{\partial x}\right) = \frac{\partial^2 z}{\partial x^2} = \frac{\partial^2 f}{\partial x^2} = f_{xx}$$

$$z_{yy} = \frac{\partial}{\partial y}\left(\frac{\partial z}{\partial y}\right) = \frac{\partial^2 z}{\partial y^2} = \frac{\partial y^2 f}{\partial y^2} = f_{yy}$$

$$z_{xy} = \frac{\partial}{\partial y}\left(\frac{\partial z}{\partial x}\right) = \frac{\partial^2 z}{\partial y \partial x} = \frac{\partial^2 f}{\partial y \partial x} = f_{xy}$$

$$z_{yx} = \frac{\partial}{\partial x}\left(\frac{\partial z}{\partial y}\right) = \frac{\partial^2 z}{\partial x \partial y} = \frac{\partial^2 f}{\partial x \partial y} = f_{yx}$$

We refer to z_{xy} and z_{yx} as *cross (or mixed) partial derivatives*. They result when one differentiates function z first with respect to one variable and then with respect to the other variable. For example, let $z = 5 \cdot x^2 \cdot y$. Then we find z_{xy} in two steps. First we differentiate z with respect to x, and obtain $10 \cdot x \cdot y$. Then we differentiate $10 \cdot x \cdot y$ with respect to y, and obtain $10 \cdot x$, which is z_{xy}.

In many cases, $z_{xy} = z_{yx}$. When this is the case, it makes no difference whether we first differentiate the function with respect to x, then with respect to y, or *vice versa*. We obtain the same result. In general, $z_x y = z_y x$ when *Young's Theorem* applies.

> Young's Theorem: If a function $f(x, y)$, its two first-order partial derivatives, and both cross partial derivatives are continuous, then $f_{xy} = f_{yx}$.

This theorem may be generalized to functions of n variables. It enables us to disregard the ordering of our differentiation when we find cross partial derivatives, provided that the continuity property holds as outlined.

The table below illustrates these derivatives with the function $f(x, y) = \sqrt{(x \cdot y + x \cdot y^2 + x^2 \cdot y)}$. The first line shows the original function. The next two lines show the first partial derivatives. The fourth and fifth lines show the second partial derivatives. The last two lines show the cross partial derivatives, which equal each other, in accordance with Young's Theorem.

Name	Expression
z	$\sqrt{xy} + x\,y^2 + x^2 y$
z_x	$\frac{y}{2\sqrt{xy}} + y^2 + 2xy$
z_y	$\frac{x}{2\sqrt{xy}} + 2xy + x^2$
z_{xx}	$2y - \frac{y^2}{4(xy)^{\frac{3}{2}}}$
z_{yy}	$2x - \frac{x^2}{4(xy)^{\frac{3}{2}}}$
z_{xy}	$\frac{1}{2\sqrt{xy}} - \frac{xy}{4(xy)^{\frac{3}{2}}} + 2y + 2x$
z_{yx}	$\frac{1}{2\sqrt{xy}} - \frac{xy}{4(xy)^{\frac{3}{2}}} + 2y + 2x$

Exercise 6.4
Use implicit differentiation to show $\partial z/\partial x$ and $\partial z/\partial y$ for these expressions.

1. $2 \cdot x^2 + 3 \cdot y^2 + 4 \cdot z^2 = 24$
2. $log_e x \cdot y \cdot z = 10$
3. $a \cdot x + b \cdot y + c \cdot z = e$
4. $e^x + e^y + e^z = 1000$
5. $3 \cdot x^2 - 4 \cdot y - z^2 + x^2 \cdot y \cdot z^2 = 20$
6. $-x^2 - 4 \cdot y^2 + 2 \cdot z^3 = 60$
7. $log_e x + log_e y + log_e z = e^x$
8. $x + y - log_e z = 0$
9. $(x^2 + 8 \cdot y \cdot z) \cdot (x^3 + 5) = 8$
10. $x^3 + y^3 + z + x \cdot y + 2 \cdot x \cdot y^2 + 3 \cdot y \cdot z^3 = 0$

Given the following expressions, determine the indicated higher-order partial derivatives.

11. For $z = x^2 + 2 \cdot x \cdot y + y^2$, find z_{xx} and z_{yx}
12. For $z = 4 \cdot x^2 \cdot y^2$, find z_{xx} and z_{yy}
13. For $z = e^(x^2 + y^2) + 4 \cdot x^3 \cdot y^2$, find z_{xx} and z_{yy}
14. For $z = x/y - y/x$, find z_{xx} and z_{yy}
15. For $z = 2 \cdot x^2 + y^2 - 4 \cdot x - 8 \cdot x \cdot y^2$, show that $z_{xy} = z_{yz}$
16. For $z = (x - y)/(x + y)$, show that $z_{xy} = z_{yx}$
17. For $z = log_e(x^2 + y^2)$, show that $z_{xx} + z_{yy} = 0$
18. For $z = e^(x^2 + y^2) + 4 \cdot x^3 \cdot y$, show that $z_{xyy} = z_{yxy} = z_{yyx}$

6.4 Applications of Partial Derivatives

The remainder of this chapter consists of two applications of partial derivatives to economic analysis. The first application is to partial elasticities of demand. The second relates to an optimal combination of advertising media.

6.4.1 Partial Elasticities

As we have seen, the elasticity is the percentage change of one variable per one-percent change in some other variable. As such, this measure of response is valuable precisely because it is unit-free. We need not know the units of either variable in order to make a meaningful statement about how responsive one variable's value is to the other variable's value. Furthermore, when the elasticity in question is the price elasticity of demand, we have established an important relationship between elasticity of demand and marginal revenue,

that $MR = P \cdot (1 + 1/Ep)$, where Ep is the price elasticity of demand, a negative number.

This section extends the analysis of elasticity to include cross-price elasticities and income elasticity. Regarding the latter, we establish an important feature: depending on whether this elasticity is greater than 1 or less than 1, consumers will spend a larger or smaller share of income on the good if their incomes increase.

Consider the general expression for a demand curve, $qa = f(pa, pb, m)$, where qa is the quantity of Good A, pa is the price of Good A, pb is the price of Good B, and m is money income. We define the elasticities as follows: $Eaa = \partial qa/\partial pa \cdot pa/qa$, $Eab = \partial qa/\partial pb \cdot pb/qa$, and $Eam = \partial qa/\partial m \cdot m/qa$. The first of this set is the "own-price" elasticity of demand, the second is the cross-price elasticity, and the third is the income elasticity. If $Eaa < -1$, the demand for this good is price elastic; if $0 < Eaa < -1$, the demand is price inelastic.

If $Eab > 0$, then Good B is a substitute for Good A; otherwise the two goods or complements (or unrelated, if $Eab = 0$). Finally, if $Eam > 1$, then the budget share of Good A moves in the same direction as income (such goods are sometimes called superior goods). If $0 < Eam < 1$, the quantity purchased of Good A is directly related to money income (such goods are *normal goods*; superior goods are also normal goods). If $Eam < 0$, Good A is an *inferior good*.

We illustrate these relationships with two demand curves. One is linear (constant slope) and the other exhibits constant elasticity. First, consider a linear case in which the quantity of Good x (qx) is a function of the prices of Goods x and y (px and py), money income (m), and the number of potential buyers: $qx = b_0 + b_1 \cdot px + b_2 \cdot p + b_3 \cdot m + b_4 \cdot n$.

A detailed interpretation of this expression would require a great deal of information: physical units of x and y, the time period length, the monetary unit in which prices and income are stated, and the relevant measure of the number of potential buyers. Suppose, for example, qx is the number of riders on a regional transportation system, measured in thousands of riders per week, px is the fare per ride, py is the per-mile cost of operating a private automobile, m is average per-family income in the region, and n is the number of people living within a specified distance of a station.

The exhibit below shows a stylized linear demand function for a transit system's services. The list named `paramsEst` is a hypothetical set of estimated values for the parameters. The first output line shows the general expression for this demand function. The second output line shows the expression given the estimated parameters. Also, it shows the number of rides per period (3950.0) that is estimated, given the values of the independent variables.[7]

```
qx : b0 + b1*px + b2*py + b3*m + b4*n;
paramsEst : [b0=5000, b1 = -400,  b2=300, b3=-0.01, b4 = 0.002]$
[qxEst: subst(paramsEst,qx),
qx0: subst([px=4, py=0.5, m=40000, n=400000], qxEst)];
b2 py+b1 px+b4 n+b3 m+b0
  [300 py-400 px+0.002 n-0.01 m+5000,3950.0]
[Exx : -400*4/qx0, Exy: 300*0.5/qx0,
 Exm: -0.01*40000/qx0, Exn: 0.002*400000/qx0];
[-0.405,0.038,-0.101,0.203]
```

The bottom three lines show the implied elasticities, along with the calculations required to generate them.

Supposing that qx is stated in 1000s of riders per period, interpret the values in the fourth input line as follows: px, the per-ride fare is \$4; py, \$0.50, is the per-mile cost of operating a private automobile; m, \$40,000, is annual per-household income in the relevant area; and n, 400,000, is the number of people living within a defined distance from the transit line. From the estimated linear relationship we can determine that the demand curve slopes downward, that transit rides and automobile rides are substitutes, that transit rides are an inferior good, and that adding 1 person to the area generates 2 more rides per period.

Expressing these values in terms of elasticities offers a number of advantages. One is that doing so makes comparison with similar analyses done by others is easier. Such comparisons can indicate whether the study has been conducted in a proper fashion (best available set of measures for variables and the proper set of variables in the model, for example). Also, elasticities provide some relevant information more directly than the original coefficients. In particular, $-1 < Exx < 0$ implies that the transit authority could increase

[7] The values used here are stylized, but they are based on research by one of the authors of the original edition of this text. See Ostrosky and Kuhn, p. 160.

revenues by raising the fare. That $Exy = 0.038$ indicates a weak relationship between automobile operation cost and ridership.

Next consider a *constant-elasticity demand function*, $qx = A \cdot px^{Exx} \cdot py^{Exy} \cdot m^{Exm}$ (other variables could be included).[8] At all values of the independent variables, the elasticities are the same. This constancy of the elasticities implies that the slopes change as variables' values change, as Figure 6.5 demonstrates.

To determine the slopes, take the first partial derivatives of the qx function. To use Maxima, apply this list of commands:

```
[diff(qx(px,py,m,A,Exx,Exy,Exm),px),
diff(qx(px,py,m,A,Exx,Exy,Exm),py),
diff(qx(px,py,m,A,Exx,Exy,Exm),m)].
```

The resulting expressions are in the following table.

$$\begin{bmatrix} dqx/dpx = & Exx\, m^{Exm}\, px^{Exx-1}\, py^{Exy} A \\ dqx/dpy = & Exy\, m^{Exm}\, px^{Exx}\, py^{Exy-1} A \\ dqx/dm = & Exm\, m^{Exm-1}\, px^{Exx}\, py^{Exy} A \end{bmatrix}$$

A little manipulation of the results above reveals that the slopes are these: $dqx/dpx = Exx \cdot qx/px$, $dqx/dpy = Exy \cdot qx/py$, and $dqx/dm = Exm \cdot qx/m$.

The next table shows qx values for selected combinations of prices and income, given these parameters: $A = 0.012$, $Exx = -1.5$, $Exy = 0.75$, and $Exm = 1.2$. Comparing the quantity in the second row with that in the first row allows computation of the arc own-price elasticity. Likewise, the values in the third and fourth rows provide the values for calculating the arc cross-price elasticity and the arc income elasticity. The accompanying workbook shows the computations using the commonly-used midpoint formula. The arc elasticity values are, respectively, -1.475 (compared to the point elasticity of -1.5), 0.751 (compared to the point elasticity of 0.75), and 1.19 (compared to the point elasticity of 1.2).

prices and income	quantity
$py = 4, m = 100, px = 2$	3.014263717811496
$py = 4, m = 100, px = 3$	1.640757346405055
$py = 5, m = 100, px = 2$	3.563393273053673
$py = 4, m = 150, px = 2$	4.903325869367984

[8] A is a scaling factor. Mathematically, it is the value of qx when all independent variables equal 1. Given that physical, temporal, and monetary units can be selected at will, A's value can be selected for convenience.

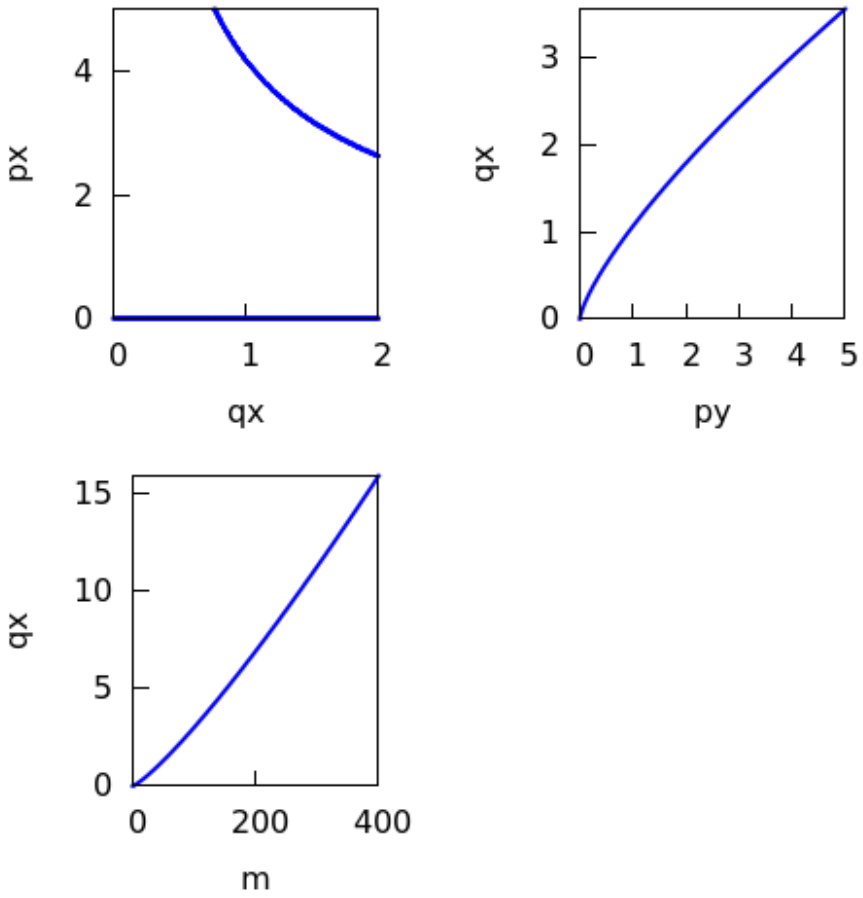


Figure 6.5: Views of a constant-elasticity demand function

6.4.2 Selection of Advertising Media

For a second example, consider a hypothetical firm Eddie's Electronics Emporium (EEE with a fixed budget of $B0$ dollars that it is willing to spend on advertising in either of two media, television (TV) or newspapers (N). The price of each unit of television advertising and the price of each unit of newspaper advertising are pt and pn. The firm can make any select any combination of television and newspaper spots that do not violate the budget constraint, given by $B0 \geq pt \cdot TV + pn \cdot N$.

Assume that the firm chooses to spend all the $B0$ dollars at its disposal. Then $B0 = pt \cdot TV + pn \cdot N$. This constraint can be phrased in terms of the number of newspaper ads, given the number of TV ads: $N = B0/pn – pt/pn \cdot TV$, so the slope is $dN/dTV = -pt/pn$.

The effect of advertising is governed by a function $S = F(TV, N)$, where S is per-period sales. We expect that adding to either type of advertising increases sales: $f_T V = \partial S/\partial TV > 0$ and $f_N = \partial S/\partial N > 0$. Furthermore, we expect diminishing returns to either of the advertising media: $\partial^2 S/\partial TV^2 < 0$ and $\partial^2 S/\partial N^2 < 0$.

The total differential of the advertising sales function is $dS = \partial S/\partial TV \cdot dTV + \partial S/\partial N \cdot dN$. Define an " isosales curve" as the locus of all combinations of television and newspaper advertising that yield a specified constant level of sales. That is, an isosales curve is given by the equation $S0 = f(TV, N)$. Along any isosales curve, the change in sales is 0, and we can therefore rewrite the differential as $0 = \partial S/\partial TV \cdot dTV + \partial S/\partial N \cdot dN$, so the slope of the isosales curve is $–dN/dTV = f_T V/f_n$. We establish below that the firm's optimal allocation of its advertising budget occurs when the combination of TV and N satisfied this condition: $pt/pn = f_T V/f_n$.

EEE budgets $10,000 per period to advertise its sales and repair services. Each TV spot costs $1000, and a spot in the local newspaper costs $100. The function that relates sales to advertising is $S = 10000 \cdot TV^0.2 \cdot N^0.6$. The display below shows expressions for the sales function and the advertising budget (first output list). It also shows the results of setting $pt/pn = f_T V/f_n$. The first result is that, for this pair of expressions, TV and N are used in a fixed ratio. The second result is that $TV = 5/2$, (implying that $N = 30 \cdot 5/2 = 75$.

```
subst([TV=5/2, N=75], S);
wxdraw2d(xrange=[0,9], yrange=[0,150],xlabel="TV", ylabel="N",
 key = "B=$10000", implicit(B,TV,0,9,N,0,100),  color= black,
 key="S = $160190", implicit(S-160190,TV,0,9,N,0,150),color=red,
 key="S = $120000", implicit(S-120000,TV,0,9,N,0,150),color=gray,
 key="S = $200000", implicit(S-200000,TV,0,9,N,0,150) )$
```

$1.6018\,10^5$

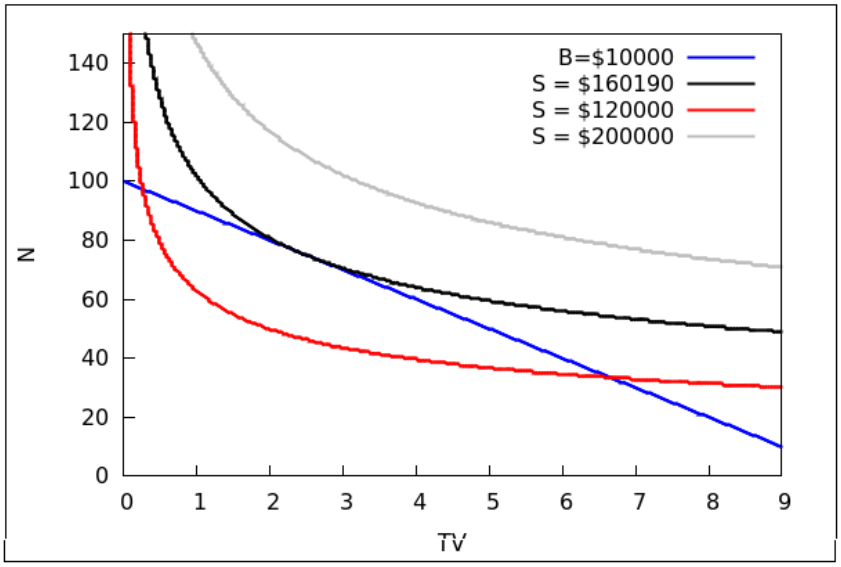


Figure 6.6: Budget line, isosales curves, and optimal mix

```
[S : 10000*TV^0.2*N^0.6, B:1000*TV+100*N-10000];
[Nsoln: solve(diff(S,N )/diff(S,TV)=1/10, N),
    TVsoln: solve(subst(Nsoln, B), TV)];
```

$$[\,10000\,N^{0.6}\,TV^{0.2}\,,1000\,TV+100\,N-10000\,]$$

$$[\,[\,N=30\,TV\,]\,,[\,TV=\frac{5}{2}\,]\,]$$

Figure 6.6 shows the highest sales level that EEE can achieve with $B = 10000$. It also shows the budget line, along with three isosales lines. One of the isosales lines shows various combinations of TV and N that would yield $S = \$160,190$. Only one of these combinations can be achieved with $B = \$10000$. The second, lower isosales line is for $S = \$12000$. Either of the intersections of this line with the budget line would consistent with $B = \$10000$, but would be inefficient. The third isosales line, for $S = \$20000$, cannot be attained with this budget.

At the northwest point of intersection of the isosales line $S = \$200000$, $f_T V/f_N > pt/pn$, so $f_T V/pt > f_N/pn$. This implies that the increase in sales from increasing TV expenditures by a small amount exceeds the decrease in sales from reducing newspaper spending by a small amount. A rightward movement along the budget line increases sales, until the tangency point is reached. At that combination, the marginal sales per dollar is the same for both media.

Consider one aspect of the sales function, that it is *homogeneous*. In general, if $g(x, y)$ is homogeneous of degree k, then this is true: $g(k \cdot x, k \cdot y) = k^n \cdot g(x, y)$, where k is a positive value and n measures the degree of homogeneity. Apply this to the sales function: $(k \cdot N)^0.2 \cdot (k \cdot TV)^0.6 = k^0.8 \cdot N^0.2 \cdot TV^0.6 = k^0.8 \cdot S$. Therefore, our sales function is homogeneous, degree 0.8. We will encounter homogeneity in later analysis of production. Look at Figure 6.6: The three isosales lines all have the same slope along any ray (constant N/TV). This is an important aspect of homogeneous functions.

Suppose that we double TV and N. As a result, as the next table shows, sales increase to \$278,900, the second entry. The third entry shows that $2^{0.8} \cdot \$160190 = \$278900$. Here, $k = 2$. If we start from column 2 and let $k = 1/2$, then the fourth entry equals the second entry times $(1/2)^{0.8}$.

$$\begin{bmatrix} S(5/2,\ 75) & S(5,\ 150) & S(5/2, 75) \cdot 2^{0.8} & S(5,\ 150) \cdot (1/2)^{0.8} \\ 1.6019 \cdot 10^5 & 2.789 \cdot 10^5 & 2.789 \cdot 10^5 & 1.6019 \cdot 10^5 \end{bmatrix}$$

Before leaving this illustrative example, we consider some managerial aspects of this analysis. To say that sales are related to a promotional bundle in a specific way is not to say that the firm knows this relationship. It does not. Therefore, we are not predicting that EEE will employ precisely this combination. Rather, this analysis offer a *normative* (prescriptive) framework. EEE will gain the maximum possible benefit from spending on promotion if it happens to employ this combination. In order to gain from is promotional spending, EEE should be considering the marginal gain from each advertising medium. If $\partial S/\partial TV > \partial S/\partial N$, then EEE will gain by moving some funds from newspaper ads to TV ads; the reverse is true if the inequality sign is reversed.

EEE's ignorance is one reason that the precise values of TV and N are unlikely to be employed. A second reason is lumpiness. While $TV = 5/2$ might not be impossible (5 ads every two weeks, for example), divisibility is

limited. Even so, the prescriptive statement in the preceding paragraph is still correct, even though the equality of marginal returns per dollar spend probably cannot be exactly equalized.

This analysis could be extended to determine the conditions that must hold for EEE to have budgeted the profit-maximizing amount of money to promotion.

6.4.3 Taxation in Competitive Markets

One of the first applications of the model of competitive markets is a demonstration that the incidence of either an excise (per-unit) tax or an *ad valorem* (per-dollar) tax depends on demand an supply conditions, not on the nominal incidence of the tax. The accompanying workbook illustrates this result with linear and constant-elasticity demand and supply curves. Here, we establish the conditions that determine how much of an excise tax falls on buyers and how much on sellers.[9]

Using the inverse demand and supply curves facilitates this analysis. Let $pd = pd(x)$ and $ps = ps(x)$ where pd and ps indicate the heights of the demand and supply curve at each output rate x. We assume that both curves are monotonic and that $d(pd)/dx < 0$ and $d(ps)/dx \geq 0$. We define an implicit function $F(x,t) = pd(x) - ps(x) - t = 0$. That is, the per-unit tax is a "wedge" between the price that consumers pay and the price that sellers receive. The Implicit Function Rule implies that[10]

$$\frac{dx}{dt} = -\frac{F_t}{F_x} = \frac{1}{pd_x - ps_x}.$$

Our interest is in $d(pd)/dt$ and $d(ps)/dt$. Multiplying dx/dt by $d(pd)/dx = pd_x$ yields an expression for dp/dt,

$$\frac{d(pd)}{dt} = \frac{pd_x}{pd_x - ps_x}.$$

In terms of economic impacts, this expression is the change in the price that buyers pay, given the quantity that is determined above. We could equally

[9] We follow Bishop[2].

[10] The notation pd_x and ps_x refers to $d(pd)/dx$ and $d(ps)/dx$.

well have multiplied by $d(ps)/dx)$ with the resulting expression being

$$\frac{d(ps)}{dt} = \frac{ps_x}{pd_x - ps_x}.$$

This is the change in the price that sellers receive.

The expressions above indicate the signs of the prices in the presence of a tax. Commonly, these relationships are expressed in terms of elasticities. For an price elasticity, E, the value is $E = (dx/dp) \cdot (p/x)$ so that $dp/dx = p(/E \cdot x)$. Making these substitutions for the supply curve and the demand curve leads to these relationships (copied from *wxMaxima*):

$$\frac{Eps}{Eps - Epd}, \frac{Epd}{Eps - Epd}.$$

The first term is the effect of the tax on the price that buyers pay, and the second shows the effect on the price that sellers receive. The ratio of these two, Eps/Epd, the ratio of the fraction of the tax that is passed to buyers to the fraction that is passed to sellers.

The difference between the price paid and the price received is $1 \cdot t$. Confirm that subtracting the effect of the tax on the price sellers receive from the effect on the price that buyers pay yields 1. Also, divide the first term by the second to confirm that the ratio of the two effects is Eps/Epd.[11]

6.4.4 Production Theory

The illustration above that addresses the effects of advertising on sales is mathematically much like *production theory*. This section develops more carefully some salient aspects of production theory. It begins with concept of a *production function*. A production function defines the maximum output that a firm can obtain from any given set of inputs that it uses. Assume a production function of the form $Q = F(L, K)$, where Q = output, L = units of labor, and K = units of capital. The marginal product of labor (MPL) measures the change in output that results when a very small change is

[11] The accompanying workbook treats taxation in more detail. It considers two specific functional forms for the demand and supply curves and illustrates the points that this section sketches.

made in the amount of labor being used, the amount of capital being held constant: $MPL = \partial Q/\partial L = f_L$. Likewise, the marginal product of capital is $MPK = \partial Q/\partial K = f_K$.

Optimization

The preceding illustration depicts a firm as having a fixed budget and seeking to generate maximum sales based on an optimal use of two advertising media. More generally, we can represent firms as setting out to gain the largest output for a given cost by selecting the ideal mix of inputs, given a set cost. Not surprisingly, if the firm uses two inputs, K and L, for which the unit costs are r and w, then the firm will attain the result if it uses the mix that spends the allocated cost $C = w \cdot L + r \cdot K$ on a combination such that $f_K/f_L = r/w$.

Suppose, however, that the firm wishes to select a quantity and then find the lowest-cost input mix consistent with that quantity. This problem is called the *dual* to the one above. Figure 6.7 shows an isoquant (for $Q = 200$ units in this illustration; see the accompanying workbook) and three isocost lines. Line $TC1$ represents a cost that is inconsistent with producing the specified output level. Line $TC2$ represents a cost at which either of two points (where the isoquant intersects $TC2$) is consistent with the required output level but at an unnecessarily high cost.

Finally, the isocost line $TC0$ allows the output level to be produced, given that labor and capital are combined as indicated by the point of tangency of $TC0$ and the isoquant (at approximately $L = 25$ and $K = 15$. As in the case of maximizing output subject to a cost constraint, minimizing cost subject to an output constraint requires that the ratio of marginal products equal the ratio of input prices. The negative of the slope of the isoquant is called the *marginal rate of technical substitution (mrts).*

Homogeneous and Homothetic Functions

As we have seen a function is homogeneous if changing all inputs in the same proportion, k , so that their ratio remains the same, causes output to increase by k^n, where n is the degree of homogeneity. Functions with the special condition that $n = 1$ are said to exhibit *linear homogeneity.* (Beware: The

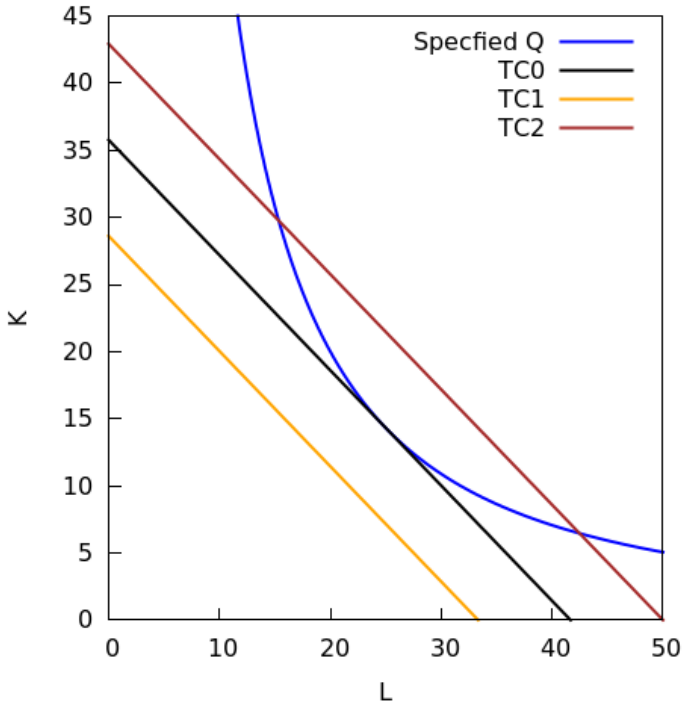


Figure 6.7: An isoquant and three isocost lines

functions themselves need not be linear.) The following function illustrates important aspects of homogeneity: $Q = A \cdot L^a \cdot K^b$. This function is often called the *Cobb-Douglas production function*, and we denote it as such.[12] For this function, $A \cdot (k \cdot L)^a \cdot (k \cdot K)^b = A \cdot k^{a+b} \cdot Q$. Thus, changing L and K by a factor k causes Q to change by k^{a+b}.

Figure 6.8 shows three sets of isoquants. Each set depicts the production of the following quantities: 5, 10, and 15. The titles indicate the degree of returns to scale. The first function is $Q = L^{.3} \cdot K^{.2}$, which implies $n < 1$. In production, this situation is called *decreasing returns to scale*. The second production function is $Q = L^{.6} \cdot K^{.4}$, so that $n = 1$. This production function is said to exhibit *constant returns to scale*. Finally, the third function is $Q = L^{.75} \cdot K^{.5}$, so that $n = 1.25$ and the function exhibits *increasing returns to scale*. In all cases $a/b = 3/2$.

In all cases, a ray along which the ratio of K and L is the same on each of three isoquants is the same is added. Each ray's slope is 2. That is $K = 2 * L$. The ray is just for comparison, and any common slope will do. In all cases, the isoquants are "parallel displacements" of each other: that is, along any ray like the one shown, the *mrts* is the same on all isoquants. When $n = 0.6$, the isoquants become increasingly farther apart, given constant increments to output.

Compare the length of line segments along the ray between the first two isoquants with the length between the second isoquants. With decreasing returns to scale, the second segment is longer. With constant returns to scale, the two segments are of equal length. With increasing returns to scale, the second segment is shorter. A less obvious point is that the isoquant slopes are the same at all nine points of intersection.

Homogeneous functions are a subset of another important class of functions, those that are *homothetic*. A function is homothetic if the *mrts* for any two inputs is the same for all values of the inputs as long as they are used in the same proportion. Such a function need not be homogeneous. The exhibit below shows a function that is not homogeneous but is homothetic. This exhibit show that this function is not homogeneous, because no term of the form $k^n \cdot Q$ can be extracted.[13]

[12] Strictly the Cobb-Douglas function requires that $a+b = 1$. We extend the terminology to include the slightly more flexible expression.

[13] The accompanying workbook generalizes this function and goes into some detail about

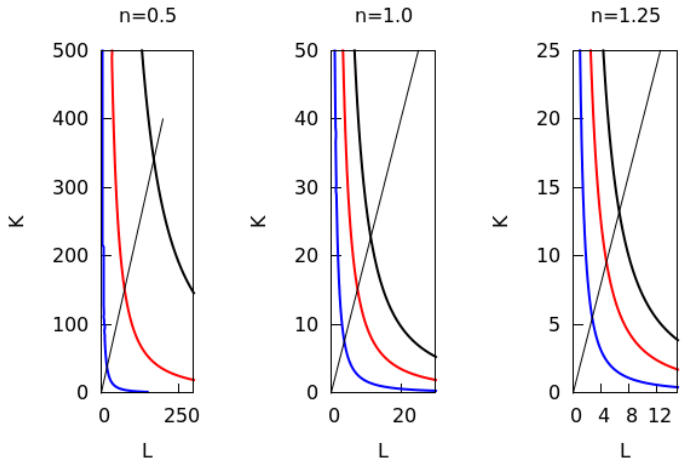


Figure 6.8: Varying returns to scale

```
Q(L, K) := 16*L^.6*K^.4- L^1.2*K^0.8;
Q(k*L, k*K);
```

$$Q(L,K) := 16\,L^{0.6}\,K^{0.4} - L^{1.2}\,K^{0.8}$$

$$16\,(k\,K)^{0.4}\,(k\,L)^{0.6} - (k\,K)^{0.8}\,(k\,L)^{1.2}$$

For small values of L and K, Q increases as L and K employment increases in a fixed proportion. Eventually, however, the second term dominates and output decreases. Thus, the function is not homogeneous. The second output line confirms this feature of the function. As noted in the preceding paragraph, the expression cannot be factored so that the output that results from replacing the original and values with a multiple of those values leads to an expression that is proportional to the original value.

This function is, however, *homothetic*. As the fourth column below shows, the slope of the isoquant equals $(3/2) \cdot (K/L)$. That is, the isoquant's slope depends only on the ratio of inputs and is not affected by scale.

$$\begin{bmatrix} MPL & MPK & MPL/MPK = & mrts \\ -\frac{6K^{\frac{4}{5}}L^{\frac{3}{5}} - 48K^{\frac{2}{5}}}{5L^{\frac{2}{5}}} & \frac{32L^{\frac{3}{5}} - 4K^{\frac{2}{5}}L^{\frac{6}{5}}}{5K^{\frac{3}{5}}} & -\frac{K^{\frac{3}{5}}\left(6K^{\frac{4}{5}}L^{\frac{3}{5}} - 48K^{\frac{2}{5}}\right)}{L^{\frac{2}{5}}\left(32L^{\frac{3}{5}} - 4K^{\frac{2}{5}}L^{\frac{6}{5}}\right)} & \frac{3K}{2L} \end{bmatrix}$$

the implication of a non-homogeneous function like this one.

A production function need not be homothetic. Consider the function

$$\mathrm{Q}(L,K) := L^b \log(K)$$

where $0 < b < 1$. We change both inputs by a factor k, using this command: `Q(k*L, k*K)`. The resulting expression is $k^b (\log(K) + \log(k)) L^b$. We cannot derive an expression of the form $k^n \cdot Q$)

The table shows the marginal products of labor and capital and the marginal rate of technical substitution. We cannot derive an expression in the form $mrts = m \cdot (K/L)$, so this function is not homothetic.

$$\begin{pmatrix} MPL & MPK & mrts \\ b \log(K) L^{b-1} & \frac{L^b}{K} & \frac{bK \log(K)}{L} \end{pmatrix}$$

6.4.5 Homogeneity and Euler's Theorem

Euler's theorem states that for a homogeneous function $f(x1, x2, \ldots, xn)$, the relationship between the partial derivatives and the function's value is as follows: $x1 \cdot f_{x1} + x2 \cdot f_{x2} + \cdots + f_{xn} = n \cdot f(x1, x2, \ldots, xn)$. This theorem provides important insights regarding economic issues. For discussion, use $Q = f(L, K)$ as the relevant production function. First, consider the question of how to account for output. In general, $f(L, K)$ is such that we cannot uniquely assign part of the output to each input.

Suppose, however, that a linear homogeneous function provides a good approximation to reality. Then the following is true: $L \cdot MPL + K \cdot MPK = Q$. This provides an accounting framework for thinking about how to attribute output to the economy's resources. It does not say anything about incomes. Suppose, however, that all firms are price-takers in both input and output markets. Then each firm maximizes profits by hiring labor and capital in amounts such that each resource's *value marginal product* (VMP) equals the unit cost of employing the resource. Recall that $VMPL = p \cdot MPL$ and $VMPK = p \cdot MPK$, where p is the price for which the firm sells output. At the margin (given the optimal input combination), $w = VMPL$, so $MPL = w/p$, the "real" wage. Likewise, $MPK = r/p$. These equalities imply that $L \cdot w/p + K \cdot r/p = Q$ so that $L \cdot w + K \cdot r = p \cdot Q$. Unlike the expression in the preceding paragraph, this is a positive statement. It says that given the conditions above the value of the total output, $p \cdot Q$, will

be distributed by market forces in a specific way, a way that is called the *marginal productivity theory of income distribution.*

This theory of distribution might provide useful insights into the general working of an economy that is largely based on free markets. For various reasons, outcomes at firm or industry level can deviate from this result. Suppose that production functions of the firms in an industry exhibit linear homogeneity, but that each firm faces a downward-sloping demand curve for its product.

In this case, a firm's marginal revenue is $mr = p \cdot (1 + 1/E)$, where p is the price that the firm charges and E, a negative number, is the elasticity of demand at price. This firm will employ L and K in quantities such that the marginal revenue product, not the value of the marginal product, equals the unit cost of the input. That is, employment will occur at levels such that $p \cdot (1+1/E) \cdot MPL = w$ and $p \cdot (1+1/E) \cdot MPK = r$. This result implies that $L \cdot w + K \cdot r = p \cdot (1 + 1/E) \cdot Q$. Suppose that $E = -5$.[14] Then the amount paid to L and K is four-fifths of the total product cost. The remainder is profit.

Second, suppose that a firm is large enough relative to one or more of its input markets that its employment decision affects the market input price. The firm has what is loosely called "monopsony power." In this case the firm's marginal input cost is$mic = w \cdot (1+1/Esupply)$, where is the price elasticity of supply to the firm and w is the input price (for a type of labor). Such a firm will employ this type of labor in an amount such that $VMPL = mic$. The variable mic is the marginal input cost. It is, roughly, the amount by which total cost of the relevant input changes per one-unit change in the amount of the input the firm employs. In this case, as in the preceding one, the payments to L and K do not sum to the value of the product, so the firm receives a profit.

Third, production functions need not homogeneous of degree 1. Hammock and Mixon show that in a competitive industry all firms except the firm (or firms) at the margin operate in the area of *decreasing returns to scale*–the degree of homogeneity is less than 1. Suppose that $n = 0.9$ is the degree of homogeneity. In that case $w \cdot L + r \cdot K = 0.9 \cdot p \cdot Q$. The remaining ten percent

[14] E is the price elasticity of demand for the firm's product, not for the industry product. Unless the industry is a perfect monopoly, E will be a larger negative number than the elasticity of demand for the industry product.

of the revenue is the profit that accrues to the owner of the inframarginal firm or to the owner of some specialized factor (like location) that makes the firm's cost lower that the cost incurred by the marginal firm(s).

Finally, of course, production functions need not be homogeneous at all. Euler's theorem applies only to homogeneous functions. Having said this, however, be aware that being descriptively inaccurate is not the same as being useless. If production at an aggregate level is well approximated by a linear homogenous production function, then the marginal productivity theory of distribution might provide the best model with which to start thinking about this important issue.

6.4.6 Two More Elasticities

We have encountered four elasticity measures: price elasticity of demand, cross-price elasticity of demand, income elasticity of demand, and price elasticity of supply (of an input). This section introduces two elasticity measures that are often used in developing and applying production theory. The first is the elasticity of substitution, which relates to the ease with which firms can change their input mix in response to changes in relative input prices. Second, a set of and output elasticity values indicate how responsive output is to changes in the individual inputs, holding constant the employment of the other inputs.

The Elasticity of Substitution

The elasticity of substitution, which we denote $Esub$, can be thought of as a measure of an isoquant's slope–more accurately, as a measure of its curvature. Figure 6.9 shows two "unit isoquants," isoquants for which the quantity produced is one unit. The family of production functions that generated these is homogeneous, so all other isoquants are parallel to these. These are from two different production functions. In one case, $Esub = 2$ and the isoquant exhibits moderate curvature. For the other, $Esub = 0.2$ and the curvature is quite pronounced.

To be more precise, we define the elasticity of substitution as the ratio of the percentage change in the K/L to the percentage change in the $mrts$, the slope of the isoquant. For the first case in Figure 6.9, the percentage

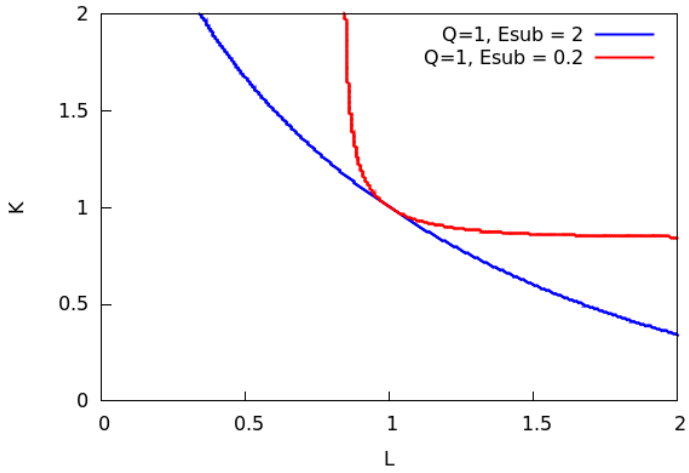


Figure 6.9: Two unit isoquants

change in K/L is twice the percentage change $mrts$. For the second case, the percentage change in K/L is only 0.2 that of the change in $mrts$.

The elasticity of substitution takes on special importance when we recall that optimizing firms select the input combinations that equate their $mrts$ with the ratio of the relative input prices (more generally, their relative marginal input costs–see the preceding section). Thus, we can redefine the elasticity of substitution as

$$Esub = \frac{d\left(\frac{K}{L}\right)}{d\left(\frac{w}{r}\right)} \cdot \frac{\frac{w}{r}}{\frac{K}{L}}.$$

This is the percentage change that the firm makes in its input ratio per one-percent change in the ratio of the input prices. Given that $w \cdot L$ and $r \cdot K$ are the costs of the inputs, consider the implications of $Esub$'s value on how total cost is distributed. If $Esub = 1$, then a one-percent change in the ratio of input costs is exactly offset by a one-percent change in the opposite direction, so that the share of cost that goes to each input type remains the same.

If, however, $Esub \neq 1$, then a change affects income distribution between the inputs. Specifically, if $Esub > 1$, then a change in w/r causes labor's share to move in the opposite direction. Likewise, if $Esub < 1$, then a change w/r causes labor's share to move in the same direction that w/r changes. Among other implications, the value of a firm's $Esub$ could affect a union's ability

to negotiate for higher wages.

The Cobb-Douglas production function, introduced above, exhibits $Esub = 1$ for all values of L and K. Indeed, one reason that the Cobb-Douglas production function developed is that Douglas, the economist, sought a function for which labor's share remained constant.[15] His collaborator, Cobb (a mathematician), pointed Douglas to the function that bears their names. For this function, $Q = A \cdot L^a \cdot K^{1-a}$. Therefore, $MPL = a \cdot A \cdot L^a \cdot K^{1-a} = a \cdot (Q/L)$. Likewise, $MPK = (1-a) \cdot (Q/L)$. Together, these imply that

$$\frac{w}{r} = \frac{MPL}{MPK} = \frac{a}{1-a} \cdot \frac{K}{L}.$$

We leave proof that, therefore, $Esub = 1$ as an exercise.

In part, because of the implication of for income shares, economists have developed expressions for production function that allow to take values other than 1. The most widely-used function is the Constant Elasticity of Substitution (CES) function:

$$A \cdot \left(\frac{a}{L^b} + \frac{1-a}{K^b} \right)^{-1/b}.$$

The table below shows this function, the two marginal product functions, and the ratio of these marginal product functions. This ratio, which is the $mrts$, can be restated as

$$mrts = \frac{1}{1-a} \cdot \left(\frac{K}{L} \right)^{1+b}.$$

$$\begin{bmatrix} CES\,function,\, Q = & \frac{A}{\left(\frac{a}{L^b} + \frac{1-a}{K^b}\right)^{\frac{1}{b}}} \\ MPL = & aA\,L^{-b-1} \left(\frac{a}{L^b} + \frac{1-a}{K^b}\right)^{-\frac{1}{b}-1} \\ MPK = & (1-a)\,A\,K^{-b-1} \left(\frac{a}{L^b} + \frac{1-a}{K^b}\right)^{-\frac{1}{b}-1} \\ mrts = & \frac{a\,K^{b+1}\,L^{-b-1}}{1-a} \end{bmatrix}$$

To determine that the elasticity of substitution is is quite direct. In the exhibit below, the expression `mrtsCES` restates $mrts$ in terms of a variable

[15] Cobb's reading of labor history indicated to him that this constancy had persisted, at least approximately, for a long period of time.

named α, where $\alpha = K/L$. It shows that

$$mrts = \frac{\alpha}{1-\alpha} \cdot k^{1+b},$$

so that $mrts$ is a monotonic function of α. The second output line shows that

$$\frac{d\ mrts}{d\alpha} \cdot \frac{\alpha}{mrts} = 1 + b.$$

The inverse function rule implies , therefore, that

$$Esub = \frac{d(K/L)}{d\ mrts} \cdot \frac{mrts}{K/L} = \frac{1}{1+b}.$$

Refer to Figure 6.9 to recall the implications of the elasticity for the shape of isoquants for the CES production function. Also, note that as $b \to 0$, the elasticity of substitution approaches 1, the Cobb-Douglas value.

```
mrtsCES : radcan(subst(K=alpha*L,    MPL/MPK));
diff(mrtsCES,alpha)*(alpha/mrtsCES); Esub = 1/%;
```

$$-\frac{a\,\alpha^{b+1}}{a-1}$$

$$b+1$$

$$Esub = \frac{1}{b+1}$$

Output Elasticity

Output elasticities measure the proportionate response in total output that is elicited by proportionate changes in the quantity of one input, all other inputs being held constant. Denote a general production function as $Q = f(X1, X2, \ldots, Xn)$. The output elasticity with respect to the i^{th} input is defined as $(\partial Q/\partial Xi) \cdot (Xi/Q)$. It is useful to rephrase this expression as $(\partial Q/\partial Xi)/(Q/Xi)$, which is the ratio of the marginal product of this input to its average product.

The *Maxima* output below this paragraph shows the result of this computation to the CES production function. The results are not in a form that is

easily interpreted. Note, however, that $\left(\frac{a}{L^b} + \frac{1-a}{K^b}\right) = \left(\frac{A}{Q}\right)^b$ can be deduced from the production function.

$$\frac{a}{L^b\left(\frac{a}{L^b} + \frac{1-a}{K^b}\right)}, \frac{1-a}{K^b\left(\frac{a}{L^b} + \frac{1-a}{K^b}\right)}$$

Therefore, we can phrase the two entries as

$$\frac{a}{(L \cdot A/Q)^b} = \frac{a}{A^b}\left(\frac{Q}{L}\right)^b$$

and

$$\frac{1-a}{(K \cdot A/Q)^b} = \frac{1-a}{A^b}\left(\frac{Q}{K}\right)^b.$$

As b approaches zero ($Esub \to 1$), these two output elasticity values approach a and $1-a$ as the CES production function approaches the Cobb-Douglas function, a conclusion that we approach below.

L'Hopital's Rule, CES, and Cobb-Douglas

Entering the command `limit(Q(L,K,A,a,b, b, 0)` into Maxima should return a result that can be phrased as the Cobb-Douglas function, but this aspect of *Maxima* is not fully reliable (in the author's experience). We can, however, apply *L'Hopital's Rule* and let *Maxima* do some of the work for us.

L'Hopital's Rule: Let one of the following be true for two functions, $f(x)$ and $g(x)$ both functions have either 0, ∞, or $-\infty$ as a limit as both approach a command value. a. That is,

$$\lim_{x \to a} f(x)/g(x) = 0/0,$$

$$\lim_{x \to a} f(x)/g(x) = \infty/\infty,$$

or

$$\lim_{x \to a} f(x)/g(x) = -\infty/-\infty.$$

The value of a can be any real number, ∞ or $-\infty$.

In any of these cases,

$$\lim_{x \to a} f(x)/g(x) = \lim_{x \to a} f'(x)/g'(x).$$

The terms $f'(x)$ and $g'(x)$ are the first derivatives of the two functions.

Rather than working with the CES function, we consider its logarithm, which *Maxima* provides:

$$\log(A) - \frac{\log\left(\frac{a}{L^b} + \frac{1-a}{K^b}\right)}{b}.$$

The second term is a ratio of two terms, both of which approach ∞ as $b \to \infty$. We can see that if $g(b) = b$, then $g'(b) = 1$ for all values of b, so for all b, the limit of $g'(b) = 1$. Therefore we can focus on $f'(b)$. The derivative of

$$-\log\left(\frac{a}{L^b} + \frac{1-a}{K^b}\right)$$

is

$$-\frac{a \cdot K^b \cdot \log(L) + (1-a) \cdot \log(K) \cdot L^b}{(a-1) \cdot L^b - a \cdot K^b}.$$

As $b \to 0$, this expression approaches

$$a \cdot \log(L) - a \cdot \log(K) + \log(K).$$

We know that the limiting value of $log(A)$, a constant, is just $log(A)$). Therefore the limiting value of

$$\log(A) - \frac{\log\left(\frac{a}{L^b} + \frac{1-a}{K^b}\right)}{b}$$

is

$$log(A) + a \cdot \log(L) - a \cdot \log(K) + \log(K).$$

Taking the anti-logarithm of this expression (reversing the effect of having taken the logarithm above) yields (after a couple of manipulations) this result: $A \cdot K^{1-a} \cdot L^a$, the Cobb-Douglas production function. Therefore, this function is the limiting case of the CES function when $b = 0$, which corresponds to $Esub = 1$.

6.5 Questions and Problems

1. (i) For each function below, determine $\partial z/\partial x$, $\partial z/\partial y$, $(\partial^2 z)/(\partial x^2)$, $(\partial^2 z)/(\partial y^2)$, and $(\partial^2 z)/(\partial x \partial y)$, if they exist. Compare your answers with those produced by *Maxima*.

   a. $z = 6 \cdot x^2 \cdot y$ b. $z = 0.3 \cdot x^4/y^2$ c. $z = 0.3 \cdot x^x/y^x$
   d. $z = y^{x+y}$ e. $z = e^{x+1} \cdot y^2$ f. $z = log_e 5 \cdot x^4 \cdot y$
   g. $z = log_e x^y$ h. $z = A \cdot x^a \cdot y^{1-a}$

   (ii) Assuming that x and y might be inputs and z might be output, determine which of these functions, if any, are plausible candidates to represent production.

2. One of the earliest applications of regression analysis to economic behavior, by Richard Stone, estimated that the demand for beer in Great Britain pre-World War II was approximately $Q = 180 \cdot Y^{-0.02} \cdot P^{-1.04} \cdot R^{0.94}$, where Q is quantity, Y is aggregate real income, P is the mean retail price of beer, and R is the mean retail price of other goods and services.

   a. Determine the estimated income, own-price, and cross-price elasticities of demand.

   b. If the price of beer rises in a given year, does this estimated demand curve predict that total spending on beer (which equals total revenue of beer sellers) will rise or fall? If the average price of a pint was £0.1 during the sample period, what is the estimated *marginal revenue* from beer sales?

   c. Given this demand specification, one can deteremine marginal revenue but not total revenue based on the information given above. Explain.

3. Apex Electronics produces a generic hand calculator that is sold under various store brands. The price at which it can sell is \$2.50 per unit. AE's total cost is $TC = \$(500 + Q/4 + Q^2/5000)$. Determine the quantity that AE should sell and its profit level.

4. With continuous compounding the present value PV of \$1 to be received t years from now is given by $PV = \$1 \cdot e^{-r \cdot t}$. Determine $\partial PV/\partial t$ and $\partial PV/\partial r$. Evaluate both expressions given that the initial values are $t = 15$ and $r = 0.04$.

5. Suppose that the "production function" that relates performance on an accounting examination (G) to hours spend studying (H) and intelligence (I, however it might be measured) is

$$G = \frac{125}{\%e^{-\frac{I}{100}} + \%e^{-\sqrt{H}} + 1}.$$

   a. Use *Maxima* to draw this function over the ranges $0 \leq H \leq 20$ and $95 \leq I \leq 125$.

   b. Determine the expressions for $\partial G/\partial H$ and $\partial G/\partial I$. Confirm that both "inputs" exhibit diminishing marginal returns.

   c. Determine the values of G and the two "marginal products" when $H = 15$ and $I = 120$.

   d. Confirm that $G_{HI} = G_{IH}$ and that this result is a quite small positive value. Explain what a positive value means in this setting. Does this seem plausible to you? Explain.

Chapter 7

Optimization: Maximization, Minimization, and Constraints

Economic analysis assumes that actors have objective functions. These functions include the utility function that represents a consumer's preferences and the profit function that represents the outcome of a firm's actions. Furthermore, this analysis often begins with the assumption that the actors attempt to maximize (or, for some functions, minimize) the value of these functions.

The actions of the decision-maker are nearly always constrained by limitations such as the amount of money or time (or both), or some minimum acceptable level of performance or output. When such constraints apply, the actor generally cannot achieve the maximum value of the objective function. Rather, the actor is assumed to *optimize*. That is, the actor finds the set of options that yield the maximum (or minimum) attainable value of the function, subject to the constraint.

We have inserted an important and somewhat controversial behavioral assumption here, that the actor is an optimizing agent. This is the standard approach for neoclassical economics. This approach can be thought of as either positive or normative. That is, we can interpret the results of the model as behavior in which we should expect actors to engage (positive). Alternatively, we may be interested in investigating the conditions that are required for optimization and using those to prescribe behavior for someone who is seeking to optimize in a particular setting (normative).

This chapter initially demonstrates how one may find the maximum or min-

imum value(s) of a differentiable function. This finding is then generalized to the case in which one wishes to find the maximum or minimum value(s) of a differentiable function that subject to some constraint(s). This generalization provides an apparatus for dealing with a wide range of problems in practical and theoretical situations.

7.1 Extreme Value(s): Functions of One Variable

We begin by examining functions of a single variable. We specify conditions under which such a function's value is increasing, decreasing, or remaining constant. Also, for a function that is changing in a given direction, we examine changes in the rate of increase or decrease.

We have examined a number of functions that either increase or decrease monotonically. Figure 7.1 shows two pairs of curves, representing the following functions over a range of x values: $y = 4 + 2 \cdot x$ and $y = 4 - x$, for $-5 \leq x \leq 5$ and $y = e^{0.5 \cdot x}$ and $y = e^{-0.5 \cdot x}$ for $-3 \leq x \leq 3$. In all four cases, dy/dx has the same sign for all x values. For each linear function,dy/dx is constant, and the values of y can increase or decrease without limit. For the exponential functions, dy/dx depends on x, and these functions can extend indefinitely in one direction only, with $y = 0$ as their lower limit.

Figure 7.2 shows two functions and reports their expressions. Both of these parabolic functions are non-monotonic. The first function achieves a *maximum* value at $x = 5$. For $x < 5$, $dy/dx > 0$; for $x > 5$, $dy/dx < 0$, and for $x = 5$, $dy/dx = 0$. The second function achieves its *minimum* value at $x = 5$. Now the first two statements regarding dy/dx are reversed, but at $x = 5$, $dy/dx = 0$, as before. When a function has a single (local) extreme value, establishing the value of x at which $dy/dx = 0$ reveals the value of x that generates an extreme value of y. Whether that extreme is a maximum value or a minimum value is indicated by the behavior of dy/dx for smaller and larger values of x. We return to this point below.

Of course, many functions have more than one extreme value. The first of the two graphs in Figure 7.3 is that of a cubic polynomial. This function has a single *local* maximum value, when x is about 1. It has a single *local* minimum value, when x is about 7. For this function, y increases without

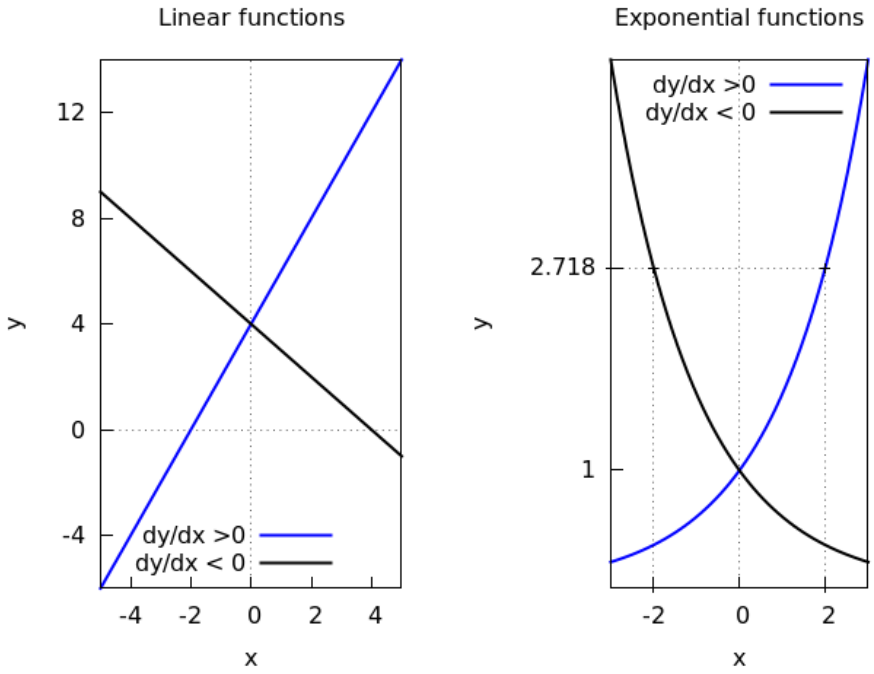


Figure 7.1: Two pairs of monotonic functions

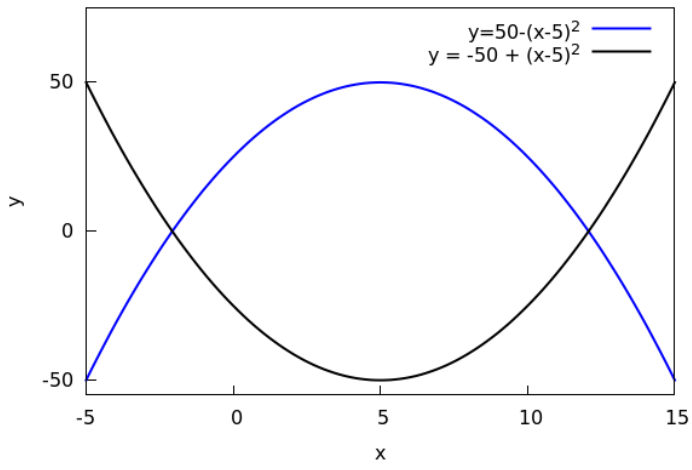


Figure 7.2: Two parabolas

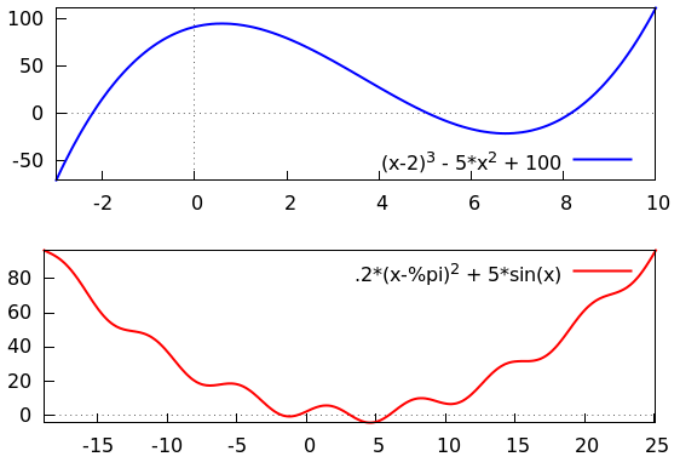


Figure 7.3: Multiple extreme values

bound as $x \to \infty$, and it decreases without bound as $x \to -\infty$. Therefore, it has no *global* maximum or minimum value.

The second graph shows a function that combines a sine function with a simple polynomial. This function has numerous local extreme values. Like the cubic function, it has no global maximum: it increases without limit as x becomes either very large or very small. It does, however, have a single global minimum, at $x \approxeq 5$. In addition, it has a number of local extreme values, both maxima and minima. The *wxMaxima* workbook for this chapter provides more detail.

7.2 Inflection Points and Concavity

The first derivative, dy/dx, measures the slope (rate of change) of a curve at a given point. It reflects whether the original function is increasing or decreasing at that point. The second derivative, d^2y/dx^2 , measures the rate of change of the slope of the function $f(x)$. The second derivative reflects whether the function $f(x)$ is increasing at an increasing (decreasing) rate or decreasing at an increasing (decreasing) rate.[1]

[1]A commonly used physical interpretation of the first and second derivatives relates to a moving automobile. The first derivative of distance with respect to time measures

We define concavity as follows: Take as given that dy/dx and d^2y/dx^2 exist for all x in some interval. Then, if $d^2y/dx^2 > 0$ the curve (x) is said to be concave upward at x. If $d^2y/dx^2 < 0$, then the curve $f(x)$ is said to be concave downward at x.

Figure 7.4 shows the two graphs from Figure 7.2 and adds three tangent line segments to each of the two graphs. The first graph illustrates the case in which the curve of the function is concave downward. At $x = 1$, the first point of tangency on the graph, the first derivative is positive and the second derivative is negative. This means that the function is increasing, but at a decreasing rate. At $x = 9$, $dy/dx < 0$ and $d^2y/x^2 < 0$, indicating that the function is decreasing at an increasing rate, that is, that the slope of the function is becoming increasingly negative. Between these two values, at $x = 5$, $dy/dx = 0$, and $f(x)$ achieves its maximum value.[2]

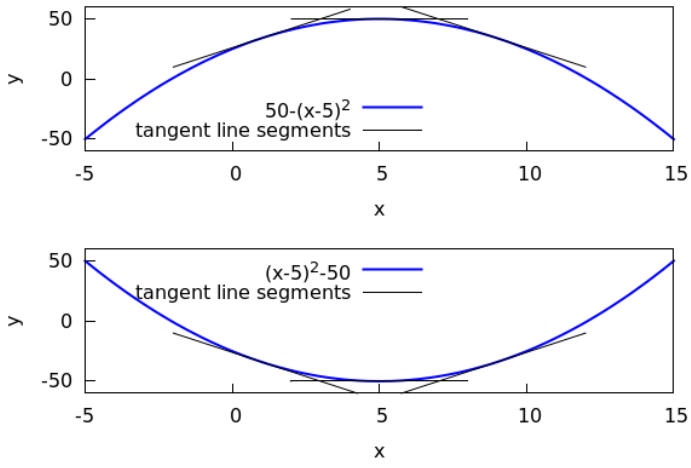


Figure 7.4: Points of tangency

The second graph depicts the case where the curve of the function is concave upward. At $x = 1$, $dy/dx < 0$, but $d^2y/dx^2 > 0$. This implies that the

velocity ("speed"), while the second derivative of distance with respect to time measures acceleration. An economic example relates to the general price level (say GDP Deflator value): The inflation rate is the first derivative, and the annual rate at which inflation is changing is the second derivative

[2]For the first function, $d^2y/dx^2 = -2$ for all x values. For the second function, the corresponding value is 2.

function is decreasing at a decreasing rate; that is, that the slope of the function is becoming less negative. When $x = 9$, $dy/dx > 0$ and $d^2y/dx^2 > 0$. The function is now increasing at an increasing rate.

When the concavity of the function changes from downward to upward or from upward to downward at a value of x, this point on the function is called an *inflection point* or *point of inflection*. Figure 7.5 illustrates the two different types of points of inflection. In the upper portion of the first column, the point of inflection occurs where the concavity of the function changes from downward to upward, at $x = 20$. The point of inflection in the upper portion of the second column occurs where the concavity of the function changes from upward to downward, again at $x = 20$.[3]

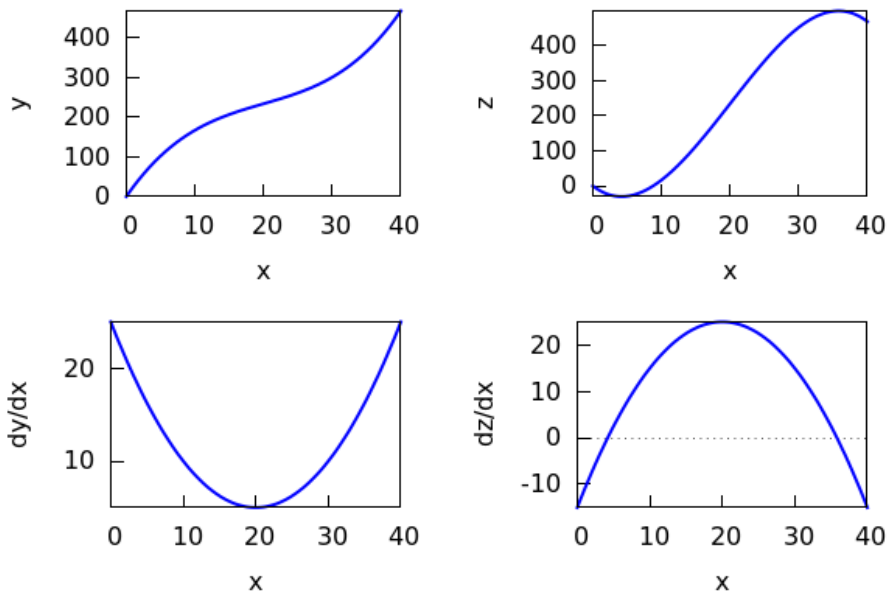


Figure 7.5: Inflection points

Points of inflection have definite implications for the first and second derivatives. We can see in the lower portion of the first column that the point of inflection is the minimum value of dy/dx when concavity is changing from downward to upward. Analogously, the point of inflection in the lower portion of the second column is the maximum value of dz/dx in the case in

[3]The two functions are $\frac{x^3}{60} - x^2 + 25 \cdot x$ and $-\frac{x^3}{30} + 2x^2 - 15 \cdot x$.

which the concavity of the function is changing from upward to downward.

Whether the concavity of the function is changing from downward to upward or upward to downward, the second derivative of the function equals zero. Hence, when $x = 20$, both d^2y/dx^2 and d^2z/dx^2 in Figure 7.5 equal zero. This reflects the facts that function y is changing from increasing at a decreasing rate to increasing at an increasing rate, and function z is changing from increasing at an increasing rate to increasing at a decreasing rate.

We can make a more general statement about points where the concavity of a function is changing. Given a differentiable function $y = f(x)$, for which dy/dx and d^2y/dx^2 are also continuous, the following must hold. If the value of the function d^2y/dx^2 is changing from a negative value to a positive value, or from a positive value to a negative value, then there must be one point at which the value of the function $d^2y/dx^2 = 0$.

A warning is in order. Given the function $y = f(x)$ that has a point of inflection at $x = x0$, then it must be true that $d^2y/dx^2 = 0$ when $x = x0$. *The converse is not true.* The fact that $d^2y/dx^2 = 0$ when $x = x0$ is does not guarantee the existence of a point of inflection when $x = x0$. Therefore $d^2y/dx^2 = 0$ is a *necessary condition* rather than a *sufficient condition* for identifying a point of inflection. Consider a simple example: $y = 4 \cdot x$. Then, for all values of x, $dy/dx = 4$ and $d^2y/dx^2 = 0$. The graph of $y = 4 \cdot x$ is a straight line, however, so it contains no points of inflection.

Consider $y = x^4$. For this expression, $dy/dx = 4 \cdot x^3$ and $d^2y/dx^2 = 12 \cdot x^2$. Solving $12 \cdot x^2 = 0$ yields $x = 0$. This does not imply, however, that a point of inflection occurs at $x = 0$. A point of inflection exists only if the concavity of the function changes from downward to upward or from upward to downward. We can determine whether this is the case by evaluating d^2y/dx^2 for $x < 0$ and for $x > 0$. When $x < 0$, $12 \cdot x^2 > 0$. Likewise,when $x > 0$, $12 \cdot x^2 > 0$. Hence $d^2y/dx^2 > 0$ both for values of x that are less than zero and for values of x that are greater than zero. Therefore, the concavity of the function is not changing. For a point of inflection to exist in this case, the sign of d^2y/dx^2 must change when we go from values of x less than 0 to values greater than 0. Hence there is no point of inflection at $x = 0$.

7.3 Finding Maxima and Minima

The value(s) of x for which $f(x)$ attains a maximum or a minimum are referred to as *extreme values*. It is necessary, as suggested earlier, to distinguish between absolute (or global) and relative (or local) extreme values. We begin with four definitions. First, we define an absolute (global) maximum: Let $y = f(x)$ be a real-valued function defined on a set S of real numbers. Then the function $f(x)$ has an absolute maximum at $x = x0$ if $f(x0) \geq f(x)$ for all x in S. The definition of an absolute minimum is similar: Let $y = f(x)$ be a real-valued function defined on a set S of real numbers. Then the function $f(x)$ has an absolute minimum at $x = x0$ if $f(x0) \leq f(x)$ for all x in S.

The definitions of a relative (local) maximum and a local minimum are also similar. Given $y = f(x)$ as above, the function $f(x)$ has a relative (local) maximum at $x = x0$ if $f(x0) \geq f(x)$ for all values of x in a *neighborhood*. Finally, the function $f(x)$ has a relative (local) minimum at $x = x0$ if $f(x0) \leq f(x)$ for all values of x in a neighborhood. A neighborhood of $x0$ is an interval containing $x0$. Formally, a δ neighborhood of $x0$ is the interval $(x0-\delta, x0+\delta)$.

Figure 7.6 shows three graphs that represent functions over a range of values (not specified). The first of the three has a relative maximum value near the center of the range of x values, which is also the absolute minimum value within this range of values (but not necessarily a global maximum–the graph does not indicate y's behavior outside this range of values), and the same absolute maximum value at either end of the range of x values. The second function exhibits a relative momimum value near the center of the range of x values, which is also the absolute maximum value in this range, and the same absolute minimum value at either end of the range of x values. Again, we cannot conclude how y behaves outside of this range from the graph.

The third graph exhibits the following behavior: as x increases in value: the absolute minimum, a relative (local) maximum, a relative minimum, and the absolute maximum for this range of x values. As before, the graph does not allow us to infer that we have found a global maximum for all x values.

Two tests enable us to identify extreme points. One, the *first derivative test*, is only a necessary condition for an extreme point. Even when satisfied, the first derivative test does not guarantee that an extreme point exists. The other, the *second derivative test*, when it accompanies a first derivative test, is a sufficient condition for an extreme point. When this test is satisfied, an

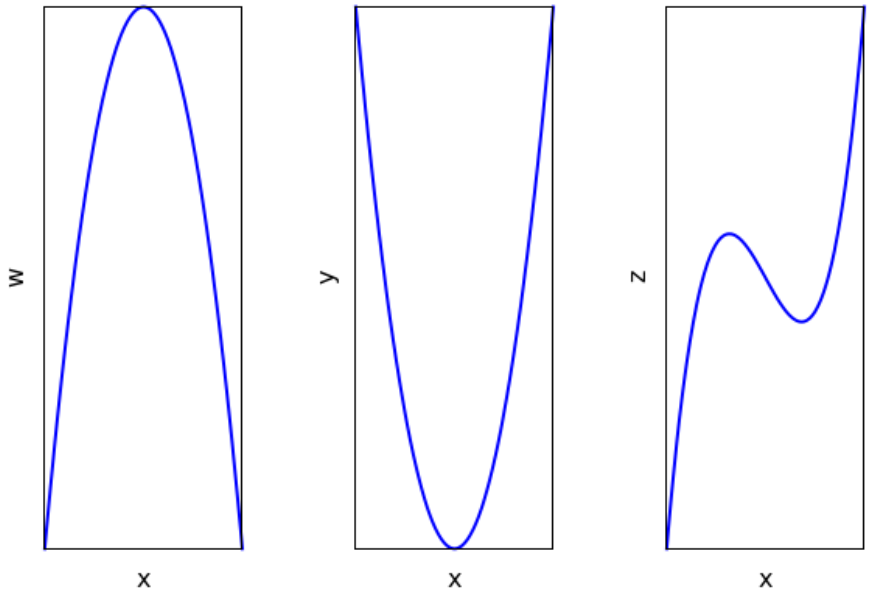


Figure 7.6: Local extreme values

extreme point exists. These tests do not say whether the extreme value is absolute or relative. They can say that, at least, a relative extreme value has been determined.

7.3.1 The First Derivative Test

The first derivative test consists of a set of three steps. First, given that the first derivative of a function $f(x)$ is $df(x)/dx = 0$ when $x = x0$, solve the equation $df(x0)/dx = 0$ for its critical root(s). Next, examine each critical root separately. If, within a given interval, $df(x)/dx$ changes its sign at a value $x = x0$, then an extreme point on $f(x)$ has been identified.

The third step is to establish which of the following occurs. If $df(x)/dx > 0$ for $x < x0$ and $df(x)/dx < 0$ for $x > x0$, then a relative (local) maximum exists at $x = x0$. Alternatively, if $df(x)/dx < 0$ for $x < x0$ and $df(x)/dx > 0 for x > x0$, then a relative (local) minimum exists at $x = x0$. Finally, if $df(x)/dx > 0$ for $x \neq x0$ or if $df(x)/dx < 0$ for $x \neq x0$, then no relative extreme point exists at $x = x0$.

Consider this simple example: $y = x^2$, for which $dy/dx = 2 \cdot x$. This expression has a single root, $x0 = 0$. For $x > x0$, $dy/dx > 0$, and for $x < x0$, $dy/dx < 0$. We can, therefore, conclude that this function reaches a local minimum value at $x = x0 = 0$.

A second example involves the two cubic polynomials defined in the table below. The first polynomial's derivative has two roots. To determine that the first results in a local minimum, we determine that the derivative's value at x = 8.7197 is f(8.7197) = -0.898 and that f(8.9197) = 0.890. Thus f(x) is decreasing for $x < 8.8197$ and increasing for $x > 8.8197$. We leave confirming that the second root, x = 31.181 results in maximum value for f(x) as an exercise.

Function	Derivative	Solution(s)
$-\frac{2 \cdot x^3}{15} + 8 \cdot x^2 - 110 \cdot x$	$-\frac{2 \cdot x^2}{5} + 16 \cdot x - 110$	$[x = 8.8197, x = 31.18]$
$\frac{x^3}{30} - 2 \cdot x^2 + 40 \cdot x$	$\frac{x^2}{10} - 4 \cdot x + 40$	$[x = 20.0]$

The second polynomial's derivative has a root at $x = 20$, but $dg(x)/dx > 0$ for $x < 20$ and also for $x > 0$. For example, at $x = 19.8$, $dg(x)/dx = 0.001$; also, at $x = 20.1$, $dg(x)/dx = 0.001$. Thus, $x = 20$ corresponds to an inflection point for $g(x) = x^3/30–2 \cdot x^2 + 40 \cdot x$. Figure 7.7 confirms the results that we have derived using information in the table above.

Exercise 7.1

For each expression below, find any extreme points that exist and determine whether each such point is a relative maximum, a relative minimum, or a point of inflection. Confirm your results with Maxima graphs.

1. $y = x^2–4 \cdot x + 16$
2. $y = x^3–6 \cdot x^2 + 9 \cdot x$
3. $y = x \cdot e^x$
4. $y = x \cdot (x - 1)^2$
5. $y = (x - 1)^3 + 8$
6. $y = x + 1/x$
7. $y = x(2 \cdot x) - 2 \cdot x$
8. $y = x^3/3–x^2 + x + 1$
9. $y = x^3$

For each of the following, find the absolution maximum and/or minimum in the designated intervals. Graph each function.

10. $y = x^2$, where $-8 \leq x \leq 16$
12. $y = (25 - 3 \cdot x)^{0.5}$ where $0 \leq x \leq 3$
13. $y = (x - 8)^2$, where $-2 \leq x \leq 4$
14. $y = 150 - 0.8 \cdot x$, where $0 \leq x \leq 10$

14. (a) When an automobile travels s miles per hour, the cost per mile (in dollars) of operating the automobile is $O = \frac{s^2}{5000} - \frac{s}{50} + \frac{189}{200}$. At what speed

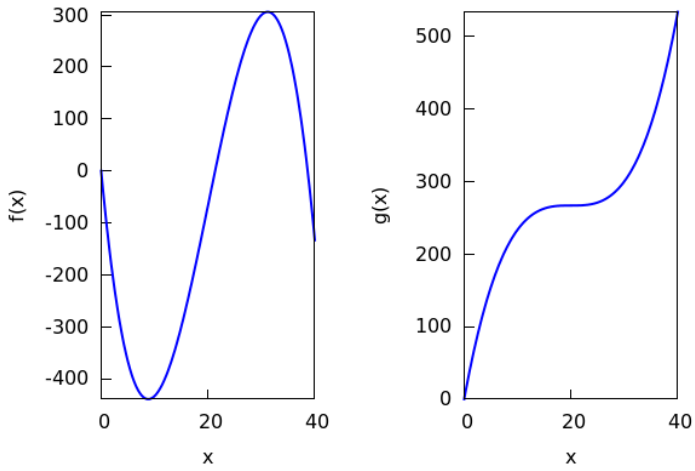


Figure 7.7: Two extreme points and an inflection point

is the cost per hour minimized? (b) Suppose that the cost of the driver's time is added to this equation, and that this cost is $30/s$. Determine the new cost-minimizing speed.[4]

15. The demand curve for a firm's product is $q = 8–p$, where q is the number of units sold and p is the price per unit. (a) What price should the firm charge if it chooses to maximize its total sales revenue ($p \cdot q$)? (b) This demand curve implies that the total revenue function is $TR = p \cdot q = 8 \cdot q - q^2$. Confirm that the marginal revenue function is $mr = d(TR)/dq = 8 - 2 \cdot q$. (c) Suppose that marginal cost is constant at $mc = 1$. What quantity maximizes the firm's profit? What price must the firm charge in order to sell this quantity.

7.3.2 The Second Derivative Test

Earlier in this chapter, we used the second derivative to indicate whether a given function was concave upward or concave downward. We now use the second derivative to determine whether the critical roots found by the first derivative test actually relative generate maxima or minima. Application of this test is a two-step process.

[4]At the time of writing (June 2016), self-driving trucks are being considered. One advantage of this type of truck is that it could save fuel costs by traveling at lower speeds than are economical when drivers' wages must be taken into account.

First, given that the first derivative of a function $y = f(x)$ exists, solve the equation $dy/dx = 0$ for its critical roots. (This step is identical to the first step of the first derivative test.) Next, if the second derivative d^2y/dx^2 also exists, then one of the following three conditions must hold:
(a) If $d^2y/dx^2 < 0$, then the function $f(x)$ has a relative maximum at $x = x0$.
(b) If $d^2y/dx^2 > 0$, then the function $f(x)$ has a relative minimum at $x = x0$.
(c) If $d^2y/dx^2 = 0$, then the second derivative test fails. We must return to the first derivative test to ascertain whether a relative maximum or minimum exists.

Figure 7.7 illustrates the second derivative test graphically. In the first panel, when $x = 8.8197$, $d^2y/dx^2 = 8.9443$ (condition b), so this point corresponds to a local maximum value. When $x = 31.18, d^2y/dx^2 = -8.9443$ (condition a), so this point corresponds to a local minimum value. In the right-hand graph, at $x = 20$, $d^2y/dx^2 = 0$, (condition c), so the second derivative test cannot detect whether or not a relative maximum or minimum exists at this point. Hence we cannot be certain what we have, based solely on these two tests.

Exercise 7.2
Find the extreme values of the following functions, and determine by use of the second derivative test whether they are maxima or minima. Confirm your results with *Maxima* graphs.

1. $y = x^2$–$8 \cdot x + 10$	2. $y = x \cdot (6 - x)^2$	3. $y = x^2 + 8$
4. $y = x^4 - 2 \cdot x^2 + 6$	5. $y = x \cdot e^(-x)$	6. $y = x + 1/x$
7. $y = x^3/3 + x^2/2 + 12 \cdot x$	8. $y = x/(x + 1)$	9. $= 1/(x + 4)$

7.4 Maxima and Minima: Functions of Two Independent Variables

The previous section dealt with the finding of extreme points for functions of a single independent variable. We now extend this discussion to include functions of two independent variables. We defer a discussion of how to identify extreme points in functions of more than two independent variables until our work with matrix algebra.

As above, we use the term *absolute extreme point* synonymously with *global extreme point*. This underlines the fact that an absolute extreme point is

global in nature with respect to the function in question. We also occasionally refer to a relative extreme point as a *local extreme point*. This emphasizes the fact that a given function may have several extreme points, one for each locality or neighborhood of the function. Only one of these local extreme points can be a global extreme point, however (unless some of the points happen to have the same extreme value).

As with the case of a single independent variable, we define our terms. First, a global maximum: Let $z = f(x, y)$ be a real-valued function defined on a set S of ordered pairs of real numbers. When $x = x0$ and $y = y0$, the function $f(x, y)$ has an absolute (global) maximum if $(x0, y0) \geq f(x, y)$for all (x, y) in S.

Likewise, for a global minimum: Let $z = f(x, y)$ be a real-valued function defined on a set S of ordered pairs of real numbers. When $x = x0$ and $y = y0$, the function $f(x, y)$ has an absolute (global) minimum if $f(x0, y0) \leq f(x, y)$ for all (x, y) in S.

Relative maxima and minima are defined in like fashion. Let $z = f(xy)$ be a real-valued function defined on a set S of ordered pairs of real numbers. When $x = x0$ and $y = y0$, the function $f(x, y)$ has a relative (local) maximum if $f(x0, y0) \geq f(x, y)$ for all $(x.y)$ in the immediate vicinity or neighborhood of $x0, yo)$ in S.[5]

Figure 7.8 illustrates two functions. Each has two independent variables. The function on the left reaches a relative maximum at a point $(x0, y0)$ within the range shown. Likewise, the function on the right reaches a relative minimum at a point $(x0, y0)$ in the range shown. The gray planes are at the maximum and minimum values of z. The blue curves show the effect of setting $x = x0$ to determine z values for the y values in the indicated range. The black curve reverses the roles of x and y. The intersection of these two curves is also the point of tangency to the plane.

The slope of the black curve is $\partial z/\partial x$, and the slope of the blue curve is $\partial z/\partial y$. At extreme values of these functions, the two partial derivatives equal zero. In general for a function $f(x, y)$, which has continuous first partial derivatives, the following is true: The function reaches a relative extreme value when $\partial z/\partial x = \partial z/\partial y = 0$.

[5]A neighborhood of $(x0, y0)$ is a circular area around $(x0, y0$. Formally. δ neighborhood of $(x0, y0)$ is the disk $(x - x0)^2(y - y0)^2 \leq \delta^2$.

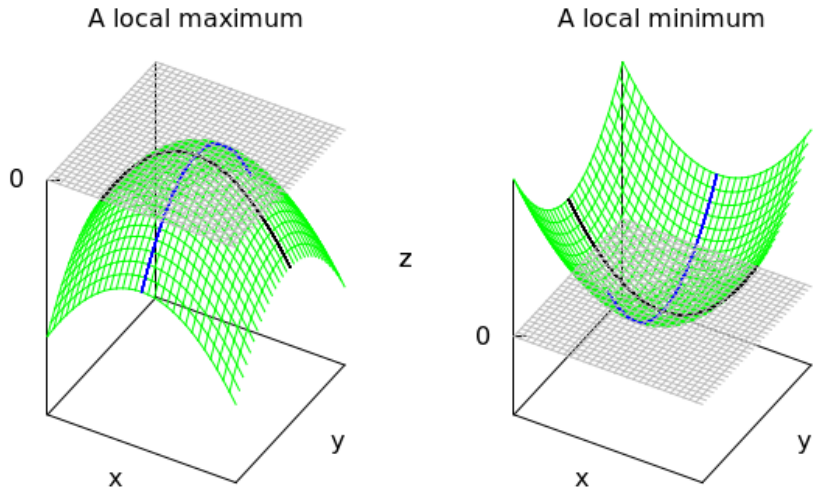


Figure 7.8: Local maximum and minimum

Figure 7.8 shows that the condition $\partial z/\partial x = \partial z/\partial y = 0$ can indicate that either a local maximum or a local minimum has been achieved. We emphasize *may have*, because two other possibilities exist. The function could have reached an inflection point, just as we saw could happen with a single independent variable. Furthermore, it could reach a *saddle point*.

To envision a saddle point, imagine yourself at a point on a three dimensional shape. When you look in one direction you appear to be at the top of a hill (where the slope is zero). Now, turn 90 degrees. You are still at a point where the slope is zero, but the shape running from the back to the front of the saddle is now a valley and you are at the minimum. Figure 7.9 illustrates the case of a saddle point. Standing at the point indicated by +, when you face in the y direction, you are atop the surface (at least locally). When you face in the x direction you are (at least locally) at the minimum point on the surface.

Figure 7.10 shows a more complicated picture. The function here is $z = x^3 \cdot y^2$, for which $\partial z/\partial x = 3 \cdot x^2 \cdot y^2$ and $\partial z/\partial y = 2 \cdot x^3 \cdot y$. When either $x = 0$ or $y = 0$, both partial derivatives equal zero, but no local extreme values are apparent. In particular, observe that moving from negative to positive values

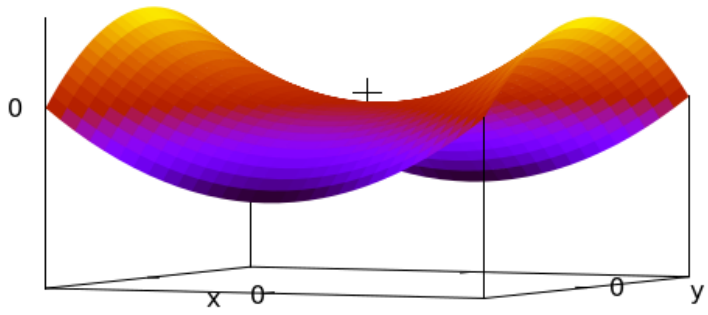


Figure 7.9: A saddle point

of x reveals a line of inflection points. At each of these points, $\partial z/\partial x = 0$ but on either side of the line $\partial z/\partial x > 0$.That is $\partial z/\partial x$ does not change signs.

7.4.1 Second Order (Sufficient) Condition

In the case of functions of one independent variable of the form $y = f(x)$, $f'(x) = 0$ was a necessary condition for an extreme point. Also $f"(x) < 0$ and $f"(x) > 0$ were sufficient conditions for the existence of a maximum and minimum point, respectively.

The second-order (sufficient) condition for the existence of an extreme point in the case of a function of two independent variables is analogous, though more extensive. A second partial derivative such as f_{xx} considers the shape of a surface only in reference to the plane XZ. Similarly f_{yy} considers the shape of a surface only in reference to the plane YZ.

Neither second partial derivative considers the shape of any cross section of the surface. For example, f_{xx} ignores the YZ plane as well as the XY plane. This means that we cannot rely on the sign of the second partial derivative

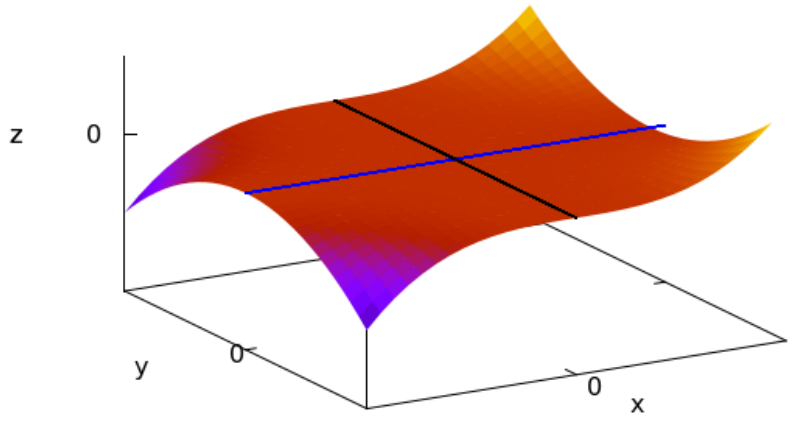


Figure 7.10: Figure with inflection points

to identify extreme points, as we did in the case of functions of only one independent variable. For example, $f_y = 0$ and $f_{yy} < 0$ is not a guarantee of a maximum point because we have not also considered f_{xx} and f_{xy}.

Applying the second-order test consists of the following:
1. Given that the first partial derivatives of $z = f(x, y)$ exist, set them equal to zero. That is, find the equations for $\partial z/\partial x = 0$ and $\partial z/\partial y = 0$. Solve these two equations for their critical roots. We label these values $x = x0$ and $y = y0$.

2. Determine whether f_{xx} and f_{yy} exist at $(x0, y0)$. If both exist, then one of the following conditions must hold at this point:
(a) When $f_{xx} \cdot f_{fyy} - (f_{xy})^2 > 0$, $f_{xx} < 0$, and $f_{yy} < 0$, we have a *relative maximum.*
(b) When $f_{xx} \cdot f_{fyy} - (f_{xy})^2 > 0$, $f_{xx} > 0$, and $f_{yy} > 0$, we have a *relative minimum.*
(c) When $f_{xx} \cdot f_{fyy} - (f_{xy})^2 < 0$, we have a *saddle point.*
(d) When $f_{xx} \cdot f_{fyy} - (f_{xy})^2 = 0$, the second-order test *fails.* A relative extreme point may exist. The second-order test, however, cannot indicate whether or not that is the case. One must examine the original function $z = f(x, y)$

in the neighborhood of $x = 0, y = y0$ in order to determine whether a local extreme point appears to exist.

Summary of conditions for unconstrained extreme value of $z = f(x, y)$

First-order condition

$$f_x = 0, f_y = 0$$

Second-order condition

(a) *Maximum*	$f_{xx} \cdot f_{yy} - (f_{xy})^2 > 0$	and $f_{xx}, f_{yy} < 0$
(b) *Minimum*	$f_{xx} \cdot f_{yy} - (f_{xy})^2 > 0$	and $f_{xx}, f_{yy} > 0$
(c) *Saddle point*	$f_{xx} \cdot f_{yy} - (f_{xy})^2 < 0$	
(d) *Test fails*	$f_{xx} \cdot f_{yy} - (f_{xy})^2 = 0$	

Three examples follow.

Example 1. Find the extreme value of $z = f(x, y) = 8 - x^2 - y^2$, if it exists.
$f_x = -2 \cdot x$ and $f_y = -2 \cdot y$ have a single solution, $x = 0$ and $y = 0$.
Therefore, the function may have an extreme value at (0,0).
Furthermore, $f_{xx} = -2, f_{yy} = -2$ and $f_{xy} = 0$.
Thus, the second order condition is $f_{xx} \cdot f_{yy} - (f_{xy})^2 = 4 - 0 > 0$, $f_{xx} < 0$, and $f_{yy} < 0$, indicating a maximum value at (0,0).

Example 2. Find the extreme value of $z = f(x, y) = x^3 + y^3 - 3 \cdot x \cdot y$, if it exists.
$f_x = 3 \cdot x^2 - 3 \cdot y = 0$ and $f_y = 3 \cdot y^2 - 3 \cdot x$ have two solutions, one when $x = 0$ and $y = 0$ and another when $x = 1$, and $y = 1$.
Therefore, the function may have extreme values at (0,0) and (1,1).
$f_{xx} = 6 \cdot x, f_{yy} = 6 \cdot y$ and $f_{xy} = -3$.
First, consider the (0,0) case:
The second order condition is $f_{xx} \cdot f_{yy} - (f_{xy})^2 = (6 \cdot x) \cdot (6 \cdot y) - (-3)^2 = 0 - 9 < 0$, indicating a saddle point at (0,0).
Now, consider the (1,1) case:
The second order condition is $f_{xx} \cdot f_{yy} - (f_{xy})^2 = (6 \cdot x) \cdot (6 \cdot y) - (-3)^2 = 36 - 9 > 0$, indicating a local extreme value at (1,1). Because $f_{xx} = 3 \cdot x$ and $f_{yy} = 3 \cdot y$, z reaches a minimum point at (1,1).

Figure 7.11 shows this function over the relevant range.
The blue lines show z values for x when $y = 0$ and for y when $x = 0$. The blue lines intersect at the saddle point.
The black lines show z values for x when $y = 0$ and for y when $x = 0$. The black lines intersect at the local minimum value.

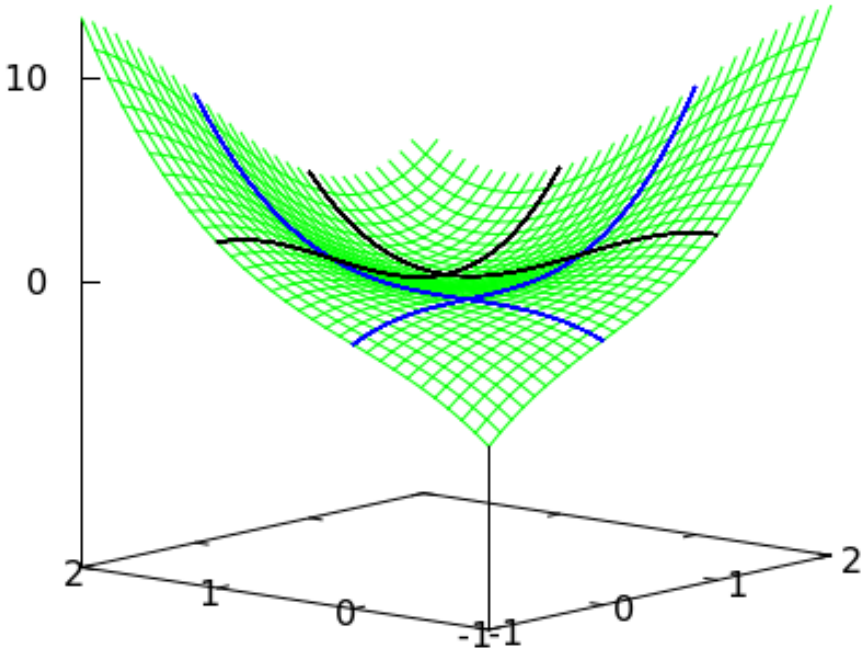


Figure 7.11: Example 2 figure

Example 3. Find the extreme value of $z = f(x, y) = x^2 + y^2$, if it exists.
$f_x = 2 \cdot x$ and $f_y = 2 \cdot y$ have a single solution, $x = 0$ and $y = 0$.
Therefore, the function may have an extreme value at (0,0).
Furthermore, $f_{xx} = 2, f_{yy} = 2$ and $f_{xy} = 0$.
Thus, the second order condition is $f_{xx} \cdot f_{yy} - (f_{xy})^2 = 4 - 0 > 0$, $f_{xx} > 0$, and $f_{yy} > 0$, indicating a minimum value at (0,0).

Exercises 7.3
Find the extreme values of the following functions. If possible, determine by use of the second derivative test whether each is a relative maximum or a relative minimum.

1. $z = z^2 + (y - 4)^2$
2. $z = x^2 - x \cdot y + y^2 - 2 \cdot x + y$
3. $z = x^2 - 2 \cdot x \cdot y + y^2$
4. $z = x^2 + y^2 - 2 \cdot x - 2 \cdot y - x \cdot y + 4$
5. $z = x^3 - 3 \cdot x + y^3 - 12 \cdot y + 6$
6. $z = x^2 + y^2 + x \cdot y + 5 \cdot x + 4 \cdot y$
7. $z = x^2 + 2 \cdot y^2 - 4 \cdot x + 8 \cdot y$

7.5 Maxima and Minima Subject to Constraints

Rare is the decision-maker who makes decisions without reference to constraints. Business people and consumers alike have limited budgets, resources. and time. As a consequence, many of the most realistic maximization and minimization problems in business and economics involve finding an extreme point subject to one or more constraints.

For example, the task of a salesperson may be to maximize the sales in a territory subject to a budget that limits the salesperson's ability to travel and service that territory. An academic administrator may wish to construct a schedule of courses that maximizes the usage of classrooms for certain key time periods during the day. However, the administrator must do so without violating constraints on how many classes can be offered, how many classes can be offered in a single time slot, and so forth. The number of decision-making problems that involve constrained maximization or minimization is as large and diverse as the world itself.

A constraint acts as a prohibiting, limiting agent in an optimization problem. That is, the constraint reduces the feasible or workable area of the objective function Figure 7.12 show two views of an objective function (blue surface)

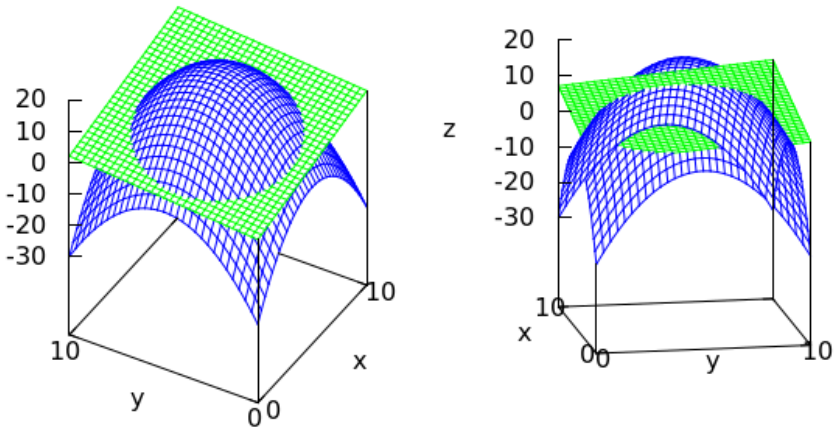


Figure 7.12: Two views of a constraint

and a constraint (green plane). The objective is to move as high up the hill as the constraint allows. The graph does not provide enough information to determine the highest feasible value of z.[6]

In general, a constraint must result in an extreme point whose value is less than or equal to the extreme value obtained when the same objective function is maximized in the absence of the constraint. Similarly, imposing a constraint on a minimization problem must result in an extreme point whose value is greater than or equal to the value obtained when the same objective function is minimized in the absence of the constraint.

We generally try to solve a constrained optimization problem by one of two methods. The first involves substituting the constraint into the objective function, then proceeding as if one were maximizing or minimizing an unconstrained function. This method seems straightforward. Unfortunately, it often it becomes complicated and quite troublesome when the objective function and constraint(s) are something other than very simple functions.

[6]We determine below that $(x = 7/5, y = 16/5)$ is the combination that yields the largest value of $z, z = 19/5$.

Hence the most popular method of maximizing or minimizing in the face of a constraint is by the use of *Lagrangian multipliers*.

7.5.1 Lagrange Multipliers

Assume an objective function in the form $z = f(x, y)$that is to be maximized or minimized subject to the constraint given by $g(x, y) = 0$. We now term a new objective function that contains both the original objective function and the constraint:

$L =$	$L(x, y, \lambda) =$	$f(x, y)+$	$\lambda \cdot g(x, y)$
The	The	The	The
Lagrangian	New	Original	Constraint
Expression	Objective	Objective	
	Function	Function	

The Greek letter λ (lambda) is a newly created unknown variable that has the property at being able to apply to the constrained objective function precisely the same first-order condition applied when an extremum is found in the absence of a constraint. The new variable λ has an important interpretation: it equals the change in the objective function per unit change in the constraint.

First-order Conditions

We determine the optimal values of x and y in three steps: First we differentiate the new objective function in with respect to x, y, and λ. Then we set these partial derivatives equal to 0. Finally, we solve this system of three equations for the three unknown values.

Thus, we create a system of these three equations:

$$L_x = f_x - \lambda \cdot g_x = 0,$$

$$L_y = f_y - \lambda \cdot g_y = 0,$$

and

$$L_\lambda = f_\lambda - \lambda \cdot g_\lambda = 0.$$

We can solve these equation for the critical roots of the function $L(x, y, \lambda)$. Note that the last of the three first-order conditions is actually nothing more than the constraint that must be satisfied when the extreme point is found.

Example 1. We now apply this approach to the function and constraint that generate Figure 7.12. The two expressions, stated as *Maxima* output, are $\mathrm{f}\,(x, y) := -\left((x-5)^2 + (y-5)^2\right) + 20$ and $g(x, y) := x + y/2 - 3$.[7]

The four commands below create the Lagrangian expression and determine the first partial derivatives.

```
L : f(x,y)- %lambda*g(x,y);
Lx:  diff(L,x);
Ly:  diff(L,y);
Llambda:  diff(L,%lambda);
```

The resulting output is the following four expressions.

$$-\lambda\left(\frac{y}{2} + x - 3\right) - (y-5)^2 - (x-5)^2 + 20$$

$$-2\,(x-5) - \lambda$$

$$-2\,(y-5) - \frac{\lambda}{2}$$

$$-\frac{y}{2} - x + 3$$

The command `soln:  solve([Lx,Ly,Llambda],[x,y,%lambda])[1];` instructs *Maxima* to solve the three partial derivatives and to return a list of values. That list is assigned the name `soln`. The result is this list of values:

$$[x = \frac{7}{5}, y = \frac{16}{5}, \lambda = \frac{36}{5}].$$

Inserting these values into $f(x, y)$ shows the maximum value of z given this constraint. Use the command `subst(soln,f(x,y));` to generate the result 19/5, the maximum attainable value of z.[8]

[7] When an expression is entered into *Maxima*, it is treated as being equal to zero unless another value is expressly entered.

[8] The accompanying workbook relaxes the constraint by a small amount and shows that the value of λ closely approximates the resulting change in z's value.

Example 2. Maximize the utility function $u = x \cdot y$ subject to the budget constraint given by $m = px \cdot x + py \cdot y$. As above, we create the Lagrangian expression. We use these commands: `[u : x*y, L : u + %lambda*(m - px*x - py*y)]`. The commands are in brackets to create this output list:

$$[xy, \quad \lambda\,(-pyy - pxx + m) + xy].$$

Commands to create a list of first-order conditions result in this output:

$$[y - \lambda px, \quad x - \lambda py, \quad -pyy - pxx + m].$$

You should derive these conditions as an exercise. Next the command `soln: solve([Lx,Ly,Llambda],[x,y,%lambda])[1]` creates the following list of solutions and assigns it the name `soln`:

$$[x = \frac{m}{2px}, y = \frac{m}{2py}, \lambda = \frac{m}{2px\ \ py}].$$

The first two entries are the (uncompensated) demand curves for x and y. Note that this consumer's income would be equally divided between the two goods: $px \cdot x = m/2 = py \cdot y$. The Lagrangian multiplier λ can be interpreted as the marginal utility of income. Given this function, this value is constant for a given set of prices.[9]

Now, suppose that $m = 100$, $px = 2$, and $py = 5$. The command `subst([m = 100, px=2, py = 5], soln)` performs the (in this case simple) calculations to yield $[x = 25, y = 10, \lambda = 5]$. We see that the consumer does spend $m/2 = 50$ on each good. Also, λ is a positive constant so the consumer gains 5 "units" of utility per one-unit increase in m, no matter what the consumer's income level might be.

Often a minus sign (-) is used in front of the constraint in a Lagrangian expression. Making this change does not affect the critical roots of the independent variables in the original objective function. There is an intuitive explanation or why the sign of the constraint term is of no consequence. The value of the constraint term, when the objective function is being maximized or minimized, as appropriate, is equal to zero. Whether we add or subtract zero is of no consequence.

[9]Beware of two possible errors in interpreting this specific result. First, this constancy is a characteristic of this class of utility functions and should not be treated as a general result. Second, numerical values have no meaning when utility: the measures are subjective.

Second-order (sufficient) test

The method of Lagrange identifies only those values of the independent variables that satisfy first-order or necessary conditions for an extreme point. These values may or may not actually represent an extreme point. A second-order test is necessary to provide further information on this matter. The second-order test is as follows.

1. Given: $L_x = L_y = 0$ at $x = x0, y = y0$. Given also: L_{xx}, L_{yy}, and L_{xy} exist at $x = x0, y = y0$.

2. Then, one of the following conditions must hold:

(a) If $L_{xx}L_{yy} - (L_{xy})^2 > 0$, and both L_{xx} and L_{yy} are negative, then we have a relative maximum at $x = x0, y = y0$.
(b) If $L_{xx}L_{yy} - (L_{xy})^2 > 0$, and both L_{xx} and L_{yy} are positive, then we have a relative minimum at $x = x0, y = y0$.
(c) If $L_{xx}L_{yy} - (L_{xy})^2 < 0$, then the second-order test fails and is incapable of indicating whether or not a relative extreme point exists. A relative extreme point may exist. One must analyze the function $z = f(x, y)$ in the neighborhood of $x = x0, y = y0$ in order to ascertain whether a local extreme point exists at $x = x0, y = y0$.

The analysis of such complicated cases is one of the areas in which a computer algebra system becomes especially useful. Many points in the neighborhood can typically be evaluated quickly, providing insights into the function's behavior in what might be a critical region.

The second-order test outlined above is quite similar to the second-order test described for the case when an unconstrained extreme point is being sought. There is, however, an important difference, Assume that $L_{xx}L_{yy} - (L_{xy})^2 \leq 0$. In the unconstrained case, a saddle point exists when $f_{xx}f_{yy} - (f_{xy})^2 < 0$, and an extreme point may exist when $f_{xx}f_{yy} - (f_{xy})^2 <= 0$. In the constrained case, however, we can say nothing about the existence of a saddle point when $L_{xx}L_{yy} - (L_{xy})^2 < 0$. An extreme point may exist when $L_{xx}L_{yy} - (L_{xy})^2 = 0$ as well as when $L_{xx}L_{yy} - (L_{xy})^2 < 0$ in the constrained case.

Summary of Conditions for Constrained Extremum: $z = f(x, y)$ subject to $g(x, y) = 0$.

First-order condition		
	$L_x = 0, L_y = L_\lambda = 0$	
Second-order condition		
(a) *Maximum*	$L_{xx} \cdot L_{yy} - (L_{xy})^2 > 0$	and $L_{xx}, L_{yy} < 0$
(b) *Minimum*	$L_{xx} \cdot L_{yy} - (L_{xy})^2 > 0$	and $L_{xx}, L_{yy} > 0$
(c) *Test fails*	$L_{xx} \cdot L_{yy} - (L_{xy})^2 \leq 0$	

Example 1. Find the extremum of $z = x^2 + y^2 - 4 \cdot x - 4 \cdot y + 7$ subject to $x + y = 4$.

Form the Lagrangian function $L = x^2 + y^2 + 2 \cdot x + 2 \cdot y + 4 + \lambda \cdot (x + y - 4)$.

Derive the first-order conditions:
$L_x = 2 \cdot x - 4 - \lambda = 0$,
$L_y = 2 \cdot y - 4 - \lambda = 0$, and
$L_\lambda = x + y - 4 = 0$.

Both L_x and L_y equal λ, so $2 \cdot x - 4 = 2 \cdot y - 4$, implying that $x = y$ when the constraint is satisfied. This implies that $x = y = 2$.

The second-order expressions are $L_{xx} = 2, L_{yy} = 2$, and $L_{xy} = 0$. Therefore $L_{xx} \cdot L_{yy} - (L_{xy})^2 = 4 - 0 > 0$ which ensures that an extreme value exists. Furthermore, $L_{xx} > 0$ and $L_{yy} > 0$ indicate that a minimum value has been found.

The value of λ is 0, implying that relaxing the constraint would move us no closer to the unconstrained maximum. As it happens, this constraint is irrelevant: The constrained optimum is the local minimum. Repeat this example, setting $x + y = 4$ and confirm that $\lambda \neq 0$; also, interpret the new value.[10]

Example 2. Find the extremum of $z = x \cdot log(y)$ subject to $x + y = 4$. The original expression and the Lagrangian expression are these:

$$[x \ \sqrt{y}, \qquad x \ \sqrt{y} - \lambda \, (y + x - 4)].$$

The associated first-order conditions involve setting the following derivatives equal to zero:

$$[\sqrt{y} - \lambda, \qquad \frac{x}{2\sqrt{y}} - \lambda, \qquad - y - x + 9].$$

[10]The workbook for this chapter shows z and the plane $z = -1$ confirming that z is tangent to the plane at this input combination.

The solution to this system of equations for x, y, and λ are

$$[x = 6, y = 3, \lambda = \sqrt{3}].$$

The terms that relate to the second-order test–f_{xx}, f_{yy}, and f_{xy} are

$$[0, \quad -\frac{x}{4y^{\frac{3}{2}}}, \quad \frac{1}{2\sqrt{y}}].$$

The second-order test relates to the sign of

$$0 \cdot -\frac{1}{2\sqrt{3}} - \left(\frac{1}{2\sqrt{3}}\right)^2,$$

which is $-1/12$. Therefore, we cannot be certain that we have found an extreme point.

We can use *Maxima*'s ability to carry out simulations to provide evidence regarding the nature of this solution. The commands `xList:makelist(8/3 + i/6,i, -4,4);` `yList: 4 - xList;` `zList: float( xList* sqrt(yList));` can be used to generate this table of values.

x	2	$\frac{13}{6}$	$\frac{7}{3}$	$\frac{5}{2}$	$\frac{8}{3}$	$\frac{17}{6}$	3	$\frac{19}{6}$	$\frac{10}{3}$
y	2	$\frac{11}{6}$	$\frac{5}{3}$	$\frac{3}{2}$	$\frac{4}{3}$	$\frac{7}{6}$	1	$\frac{5}{6}$	$\frac{2}{3}$
z	2.828	2.933	3.012	3.061	3.079	3.06	3.0	2.89	2.721

The center entry (8/3, 4/3) is the solution that we found above. It is also the largest value of z in this neighborhood of (x, y) combinations.

Exercise 7.4
Solve the following constrained optimization problems by the method of Lagrange multipliers. Use *Maxima* to confirm your computations.

1. $z = 2 \cdot x^2 + y^2$ subject to $x + y = 1$
2. $z = x^2 - 2 \cdot x \cdot y + y^2$ subject to $x + y = 2$
3. $z = x^2 + 4 \cdot y^2 + 24$ subject to $x - 4 \cdot y = 10$
4. $z = 4 \cdot x^2 + x \cdot y + 3 \cdot y^2$ subject to $x + 2 \cdot y = 21$
5. $z = 6 \cdot x^2 - x \cdot y + 5 \cdot y^2$ subject to $2 \cdot x + y = 24$
6. $z = 3 \cdot x^2 + y^2 - 2 \cdot x \cdot y - 8$ subject to $x + y = 1$

7.6 Economic Applications

We now apply the tools that we have developed in this chapter to specific types of problems in business and economics.

7.6.1 Profit Maximization: Price-Taking Firms

A fundamental problem in business and economics is that of determining the conditions that must hold for a firm to maximize its profits. We focus on a stylized firm that produces a specific, well-defined product (either a good or a service). In this section, we look at a firm that purchases its inputs and sells its output in perfectly competitive markets. That is, the firm's purchases of inputs, and its sales of the output it produces, are sufficiently small and that its output level does not appreciably affect the prices of either inputs or outputs. We refer to such firms as *price takers*.

Let $p = f(x)$ be the firm's inverse demand function. Assume that p, the price of the firm's product, is a constant, unaffected by x, the firm's output rate. The firm's total sales revenue TR of the firm is $TR = p \cdot x$, or $TR = x \cdot f(x)$. Therefore, we can write $TR = g(x)$.

The firm's total cost TC function is $TC = h(x, k)$, where k is per-period fixed cost, that part of cost that is not affected by the output rate. Thus, total cost is the sum of fixed cost variable variable costs, those that do depend on x's value.

The profit of the firm, π, is defined as total revenue minus total cost, and equals $TR - TC$. We find the conditions for a profit-maximizing level of output by satisfying the first-order and second-order conditions for an unconstrained maximum, namely, $d\pi/dx = 0$ and $d^2\pi/dx^2 < 0$. Differentiating $TR - TC$ with respect to x and setting this derivative equal to zero yields $d\pi/dx = dTR/dx - dTC/dx = 0$ which implies that the firm achieves maximum profit if it produces the quantity at which $dTR/dx = dTC/dx$. In terms that you have seen before, the firm achieves maximum profit by producing the quantity at which *marginal revenue* (dTR/dx) equals *marginal cost* (dTC/dx).

For the price-taking firm $p = dTR/dx =$ marginal revenue. Therefore, such a firm maximizes its profits by selecting the output rate at which $MC = dTC/dq = p$.

Figure 7.13 depicts the situation of a firm in a perfectly competitive market (*i. e.*, a price-taking firm) The total cost and total revenue functions shown in the top panel intersect twice, initially at output level $x1$ and subsequently at output level $x3$. At these two levels of output, total profit is equal to zero. The intervals $(0, x1)$ and $(x3, \infty)$ represent negative profit, whereas the interval $(x1, x3)$ represents positive profit. The total profit curve in the same panel reflects these considerations.[11]

The first-order condition for profit maximization, $d\pi/dx = 0$, implies that the slope of the total revenue curve must be equal to the slope of a tangent to the total cost curve. In the top panel of Figure 7.13, this equality occurs at $x = x1$.[12] The lower panel confirms that the profit maximization condition can also be phrased in terms of the equality of marginal revenue and marginal cost.[13]

The lower panel shows the quantity at which the firm's average cost is minimized. For the price-taking firm, this is the quantity at which *profit per unit* is maximized. One might be tempted to think that this quantity is one to be produced. It is not. As the graph shows, for each unit between $xminac$ (for x minimum ac)and $x2$, the price exceeds the marginal cost. That means that each of these units adds to the profits that would be earned at $x = xminac$. The lesson is general: Focus on the marginal values.

7.6.2 Profit Maximization: Price-Searching Firms

Many firms produce output that is a less than perfect substiute for that of other firms. These firms must simultaneously determine the price to charge and the quantity to sell. That is, they must search for the point on their

[11]The units on the axes are stated in some detail. They emphasize that output has two aspects, a physical unit and a time period. Also, note that the monetary units on the y axes differ for the two graph. The first shows revenue, cost and profit (or loss) per time period. The second shows per-unit revenue cost and revenue per unit produced.

[12]Depending on the curvature of the TC and the price level, it is possible that a second tangency point can occur for $x < x1$. If so, this tangency point is consistent with minimizing profits or, equivalently, with maximizing losses.

[13]From a managerial point of view, the marginal revenue, marginal cost comparison might be more useful. Firms are likely to have some data on how changes in output affect their cost levels. Knowing the values of total cost of a product is more problematic, especially for firms that produce more than one product.

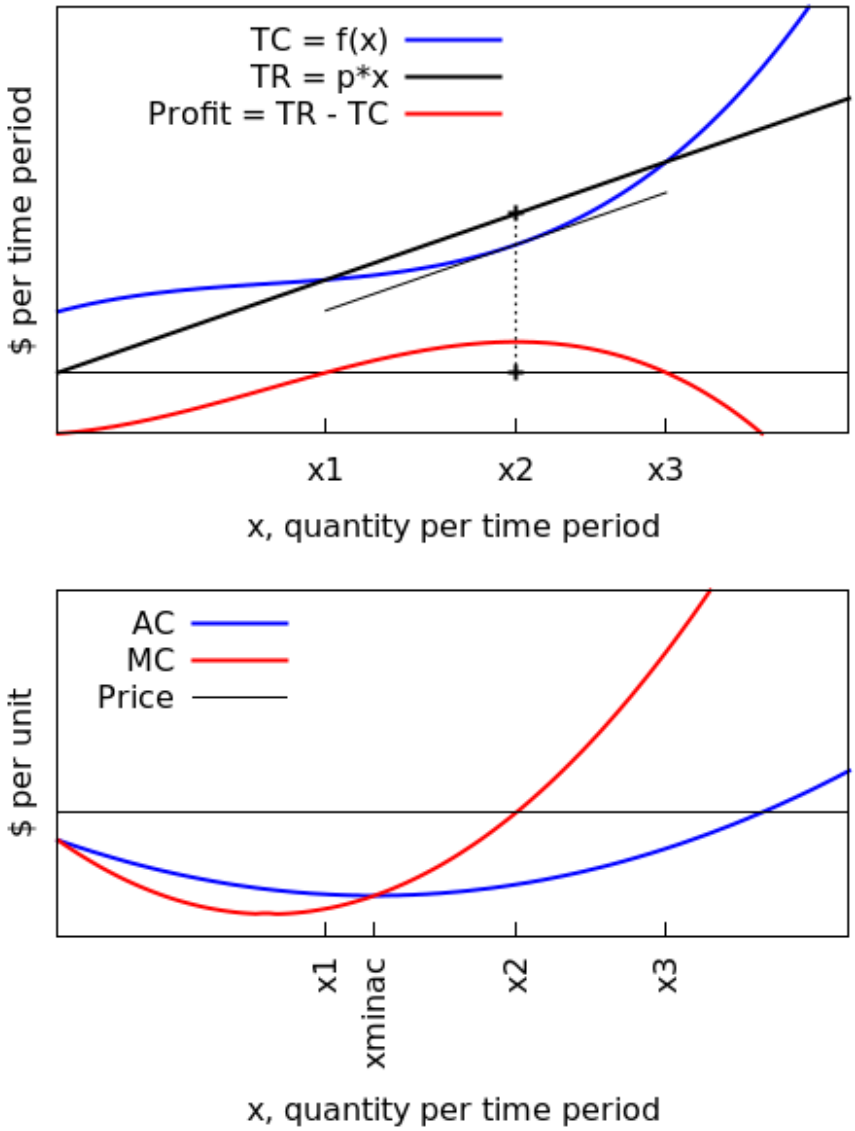


Figure 7.13: Cost, revenue, and profits

demand curve that yields maximum profit.[14]

Let the imperfectly competitive firm's demand function for its product x (in inverse form) be $p = f(x)$, where $f_x < 0$. The firm's total cost function TC is the same as above. Hence the firm's profit function is given by $\pi = f(x) \cdot x - TC$. The first-order condition for profit, as before, is that the firm produce a quantity such that $d\pi/dx = dTR/dx - dTC/dx = 0$. That is, the firm, like its price-taking counterpart, must select the quantity at which marginal revenue equals marginal cost.

Also, for this quantity to correspond to maximum, not minimum profit, the condition $d^2\pi/dx^2 < 0$. This is equivalent to requiring that $d^2TR/dx^2 - d^2TC/dx^2 < 0$ or $d^2TR/dx^2 < d^2TC/dx^2$ must be satisfied. That is, the slope of the marginal revenue curve be less than the slope of the total cost curve. The marginal cost curve must cut the total cost curve from below.

Figure 7.14 illustrates profit maximization for the imperfectly competitive (price searching) firm. The firm's total revenue function is no longer a ray through the origin, because price decreases as output increases. The lower panel shows this fact more expressly. The black curve shows the price that this firm can charge for each possible output rate per period of time. It also shows the implied marginal revenue, which is less than the price. Aside from the divergence of marginal revenue from price (average revenue), the analysis proceeds as before.

Example. That the Co-op Bookstore considers that to maximize profit on its sales of Adam Smith's *The Wealth of Nations* is quite appropriate. The (inverse) demand for the book is given by $p = 0.1665 \cdot x^2 - 0.175 \cdot x + 50$ (dollars). The implied marginal revenue is $MR = -0.4995 \cdot x^2 - 0.35 \cdot x + 50$ The per-unit cost is a constant, \$25 per book. So $TC = 25 \cdot x$ and marginal cost is 25. The accompanying workbook shows the demand, marginal revenue, and marginal cost curves.

The first derivative of the profit function, $\pi = -0.1665 \cdot x^3 - 0.175 \cdot x^2 + 25 \cdot x$ is $d\pi/dx = -0.4995 \cdot x^2 - 0.35 \cdot x + 25$. Setting this expression equal to zero yields two values, one of which is negative. The positive value is $x = 6.7329$ units per week, which we name $x1$. If the firm is thinking of per-unit sales

[14] If the firm could know its demand curve, then its task would be simply to set the profit-maximizing price (or, equivalently, to sell the profit-maximizing quantity). In fact, firms do not know the demand curve for their products, so "price searching" is more descriptive that "price setting."

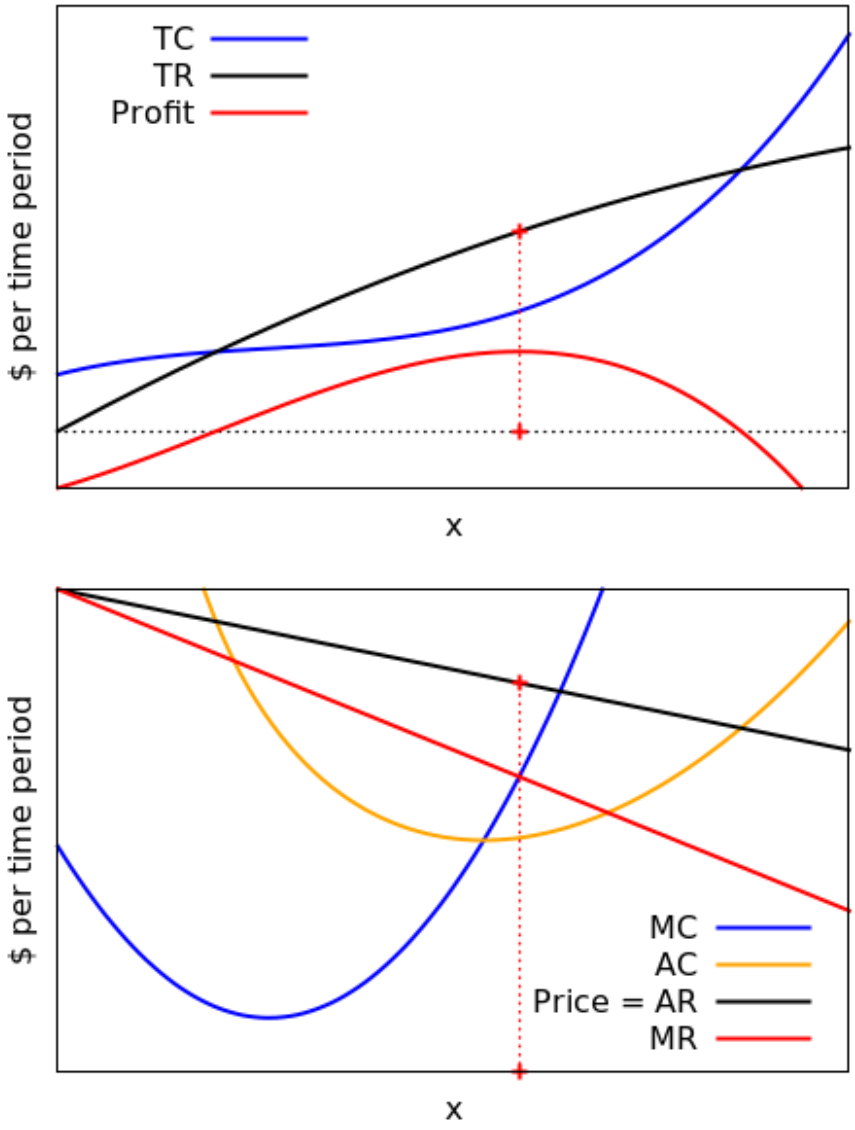


Figure 7.14: Demand, cost, and profits Revenue, cost, and profits

over a long time, then fraction units make sense. If it is thinking of this week, only a discrete number of books can be sold.

Fortunately, *Maxima* lets us have it both ways. The following three commands determine price, output, and profit if the bookstore sells 6, 6.7329, or 7 books per week:

```
subst(x=x1,[x,p,TR-TC]);
subst(x=floor(x1),[x,p,TR-TC]);
subst(x=ceiling(x1),[c,p,TR-TC]);.
```

The results are these.

Quantity	Price	Profit
6.7329	41.273	109.57
6	42.956	107.73
7	40.616	109.31

The second derivative of the profit function is $-0.999 \cdot x - 0.35$. For any of the three x values above, this expression is negative. We have, therefore, determined a profit-maximizing quantity.

7.6.3 Production: Marginal and Average Products

Consider a production function of the form $Q = f(L, K)$, where Q =output, and L and K are the labor and capital inputs, respectively. Assume a short-run situation such that the amount of capital is fixed at K = K0.[15] Thus $Q = f(L)$, given $K = K0$. The functional form f reflects current technology.

The *average product of labor* APL is given by

$$APL = \frac{Q}{L} = \frac{f(K0, L)}{L} = \frac{f(L)}{L}.$$

The marginal product of labor MPL is given by

$$MPL = \frac{d(f(L))}{dL} = f_L(L).$$

[15] Hammock and Mixon, Chapter 6, treats production in more general terms.

As Figure 7.15 illustrates, the MPL curve cuts the APL curve at the APL curve's highest point. That is, $MPL = APL$ at the value of L for which APL is at a maximum. We can demonstrate this mathematically by showing the conditions under which APL is at a maximum. We differentiate $APL = Q/L$ with respect to L using these to *Maxima* commands: `depends(Q,L)` and `diff(Q/L, L)`.[16] The result is $\frac{\frac{d}{dL}Q}{L} - \frac{Q}{L^2}$. Setting this derivative equal to zero and dividing through by L, which has a positive value, implies that $MPL = APL$ is a necessary condition for APL to achieve an extreme value.

To confirm that this value is a maximum requires evaluating the second derivative of $APL(L)$. The derivative supplied by *Maxima* is this:

$$\frac{\frac{d^2}{dL^2}Q}{L} - \frac{2\left(\frac{d}{dL}Q\right)}{L^2} + \frac{2Q}{L^3}.$$

The second term is $2 \cdot MPL/L^2$ and the third term is $2 \cdot \frac{Q}{L} \cdot \frac{1}{L^2} = 2 \cdot APL/L^2$. Because the first-order condition requires that $APL = MPL$, these sum to zero. Therefore, the sign of the first term determines the sign of second derivative of $APL(L)$. The numerator of that term, the first derivative of $MPL(L)$, is negative for all L. The test confirms what the graph shows: at $L = L1$, the average product of labor reaches its maximum value.[17]

7.6.4 Production and Cost

This section derives the firm's cost curve from the production function that the previous section develops. This derivation illustrates the nature of the dual relationship between production and cost. That is, a firm's cost curve is derived from its production function, so that employment decisions and costs are uniquely related.

Mathematically, however, matters are not quite that simple. The production function that governs our hypothetical firm is $Q = 30 \cdot L + 5 \cdot L^2 - 0.2] \cdot L^3$, a cubic function. For such a function Q is not monotonically related to L. If,

[16] Of course, for this simple application of the quotient rule, *Maxima*'s use is not required.

[17] The accompanying workbook shows a slightly more flexible, though probably less plausible, production function. For it, a quite small value of L results in minimum APL, and a larger value results in a maximum. We shall see that, even if the APL-minimizing value of L exists, it is likely not relevant to economic analysis.)

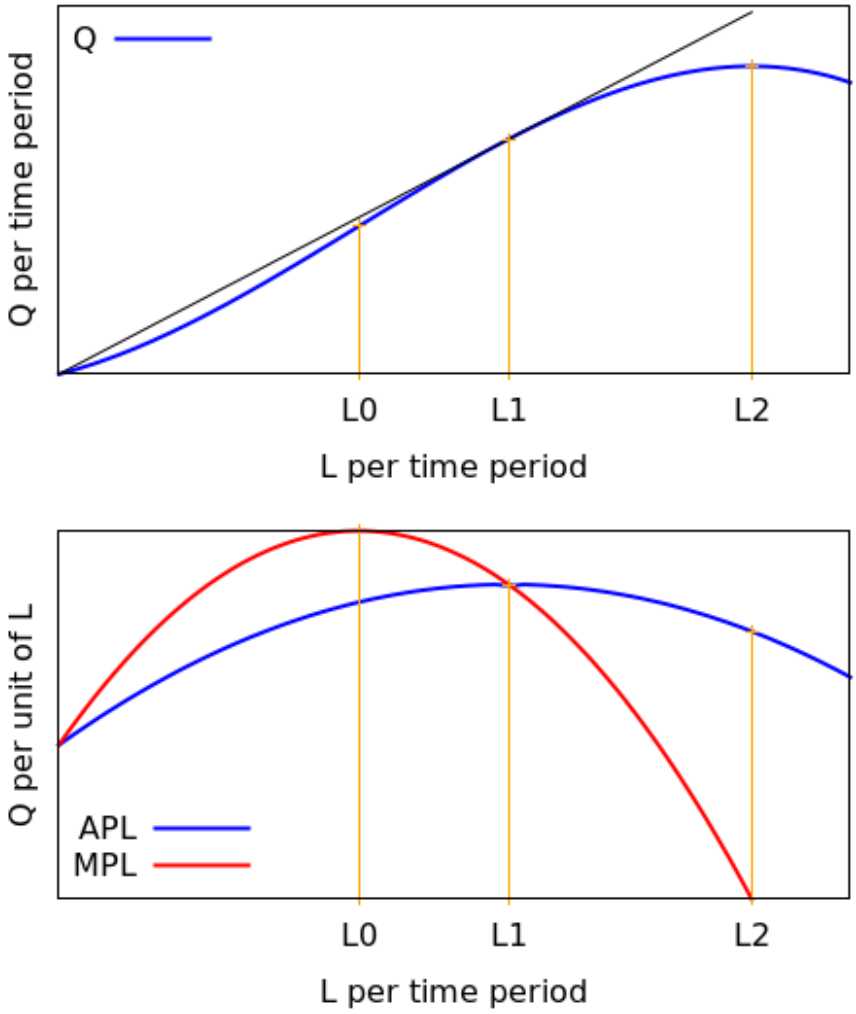


Figure 7.15: Total and per-unit output

however, maximizing behavior is added, the difficulty vanishes, because no maximizing employer would hire so much L that the MPL becomes negative.

As a result of this observation, we can map employment of L to both cost and output. Therefore, Q and TC are bound by a *parametric* relationship.[18] For the cubic production used here, the maximum value of output is $Q = 1003.6$ units, with $L = 19.262$ units of labor. We restrict all graphs to this range.

To begin, we define the parametric relationships between L and cost, $\text{TC}(L) := 400 + 500L$ for total cost, $\text{MC}(L) := \frac{500}{-0.6L^2+10L+30}$ for marginal cost, and $\text{AC}(L) := \frac{500L+400}{-0.2L^3+5L^2+30L}$ for average cost. We also define a parametric relationship for average variable. Be aware that these are emphatically *not* the usual total cost, marginal cost, and average cost functions; these relate costs to output and not to employment.

Figure 7.16 shows total and per-unit cost curves. These curves show cost as a function of quantity. They do so via the application of commands like this one: `parametric(Q(L),AC(L),L,.1,Lmax)`, which appear inside the `draw2d` command. This command places Q values from the production function on the x axis and TC values from the cost (as a function of L) function on the y axis. Other commands do the same for the per-unit cost curves. Observe that the resulting cost curves behave much as the stylized curves that we graphed earlier, using an *ad hoc* cubic cost function.

Figure 7.16 also contains a price line, which is appropriate for a price-taking firm. We know that profit-maximization is attained by producing Q such that $MC = p$. The graph shows that this equality holds for two values. Earlier analysis points out the the first of this pair is consistent with loss maximization. Thus, a quantity $Q \approx 950$ seems to be the profit-maximizing quantity.

We do not have an explicit expression to relate price and Q. Some economic theory, however, solves our problem. Remember that a price-taking firm maximizes its profit by employing its variable input at a level such that $price \cdot \partial Q/\partial L$ equals the wage rate. In our example, we can phrase this requirement as `MPL(L)*12- 500`. We apply the `find_root` command to determine this value which is $Q = 917.53$, with an employment level $L = 15.404$. Confirm that this firm earns a profit of approximately \$2908.

[18] We limit the analysis to the short run, where the employment of a single input varies. More advanced treatments allow multiple variable inputs. See Hammock and Mixon, Chapter 7, and "Key to Textbooks."

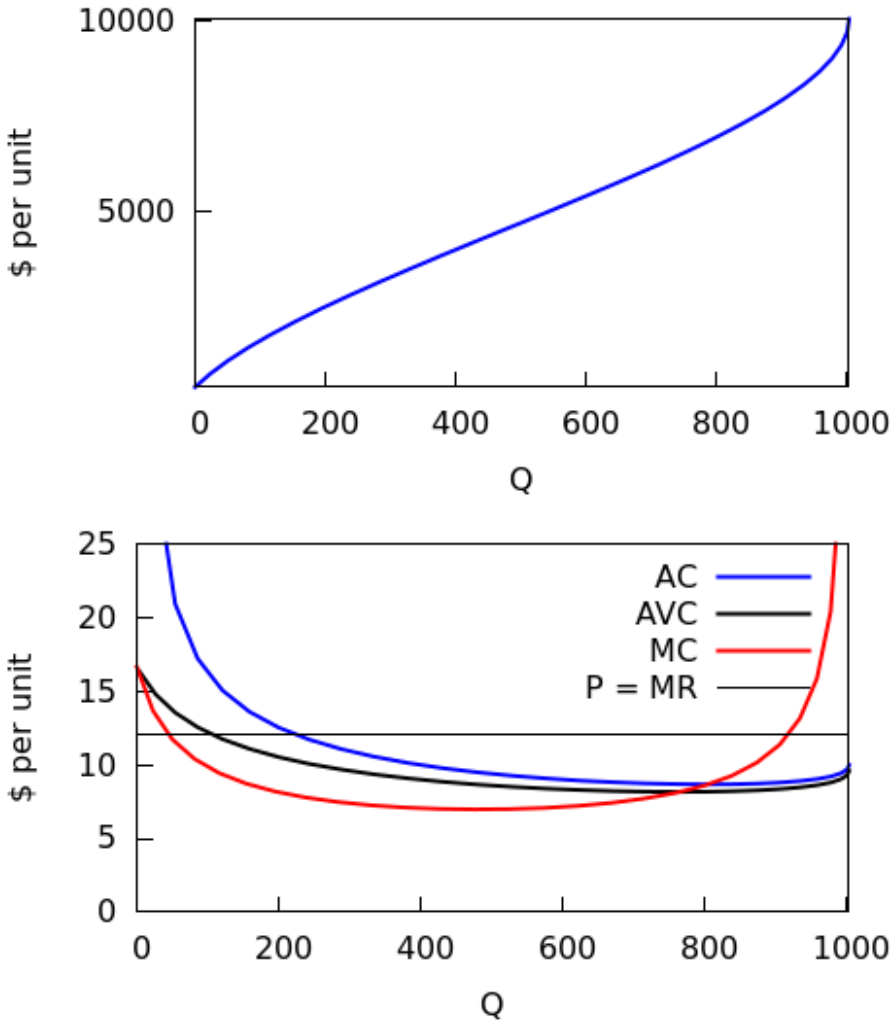


Figure 7.16: Total and per-unit cost curves

We can now see why production by a price-taking firm occurs within the range of decreasing marginal product. In this example, production occurs where $MPL = 41.666$ and $APL = 59.562$. More generally, the following relationships hold: $MC = w/MPL$ and $AVC = w/APL$. Profit-maximization requires that $MC = p$, so $MPL = w/p > 0$, where w/p is the "real wage" that the firm pays. The price level must exceed the minimum value of AVC, so $MC > AVC$, implying that $MPL < APL$.[19]

7.6.5 Taxation

Chapter 6 examines the effects of imposing a tax on either the sale or the purchase of a single good or service when both buyers and sellers are price takers. It shows that whether the tax is nominally imposed on buyers or on sellers is beside the point. Either way, the effects on the price paid by the buyers and the price received by the sellers is determined by the relative values of the price elasticity of demand and the price elasticity of supply.

This section examines the case in which the seller in question is a price-searcher. More specifically, we examine the polar case of the pure monopolist. In the case of pure competition (price takes on both sides of the market), the analysis proceeds by solving a system of two equations (the demand and supply curves) and then imposing and tracing the effects of a change. (This approach is another example of *comparative statics* analysis.) When examining the case of price takers' markets, we did not look at the behavior of any single firm or buyer; the effects of their behavior are summarized in the supply curve and the demand curve respectively. Therefore, we did not have to apply any optimization conditions.

With a single price-searching seller, however, the firm's optimizing behavior is central to the analysis. As before, we follow the analytical approach that Bishop [2] provides. Begin by defining the objective function: the function $\pi = TR(x) - TC(x) - t \cdot x$ is to be maximized.

The first-order condition for maximizing π is $TR_x - TC_x - t = 0$. The second-

[19] If the firm is a price-searcher, then employment can occur at a lower L, where MPL might exceed APL. The accompanying workbook shows how this can happen. Likewise, if the firm faces an upward-sloping supply curve of labor, it might employ an amount of labor such that $MPL > APL$.

order condition is that $\pi_{xx} = TR_{xx} - TC_{xx} < 0$.[20] That is, the marginal revenue's slope must be less than that of the marginal cost. In most cases, we expect MR curves to slope downward ($TR_{xx} < 0$) and marginal cost curves to be either horizontal or upward sloping ($TC_{xx} \geq 0$).

In order to determine the nature of dp/dt, we first apply the implicit function theorem to the first-order condition and determine dx/dt. First, $d\pi_x/dx = TR_{xx} - TC_{xx}$. And, of course, $d\pi_t = 1$. Therefore,

$$\frac{dx}{dt} = \frac{1}{TR_{xx} - TC_{xx}}.$$

Multiplying this term by dp/dx yields the result that we seek:

$$\frac{dp}{dt} = \frac{p_x}{TR_{xx} - TC_{xx}},$$

where p_x is the slope of the inverse demand curve. Both the numerator and the denominator are negative, so $dp/dt > 0$.

This formulation allows direct analysis of the case in which the demand curve and the marginal cost curve are linear. If $p = \alpha + a \cdot x$, then $MR = \alpha + 2 \cdot x$. Likewise, if $AVC = b + c \cdot x$ then $MC = b + 2 \cdot x$. The slopes of MR and MC are $2 \cdot a$ and $2 \cdot c$. The table below summarizes these aspects and implications of the linear price and average variable cost functions. The results imply that $dp/dt = a/(2 \cdot (a - c))$. Recalling that $a > 0$ and $c \geq 0$, these values imply that $dp/dt < 1$, a result that is similar to the case of price-taker markets.

We can say a bit more. Suppose that $c = 0$.Then $dp/dt = 1/2$. For $c > 0$, $dp/dt < 1/2$. As with price-taker markets, the relative values of b and c determine the effect of the tax on the prices paid and received.

$$\begin{bmatrix} p & p\,\text{Slope} & TR & MR & MR\,\text{Slope} \\ ax + \alpha & a & a\,x^2 + \alpha x & 2ax + \alpha & 2a \\ === & === & === & === & === \\ AVC & AVC\,\text{slope} & TVC & MC & MC\,\text{Slope} \\ cx + b & c & c\,x^2 + bx & 2cx + b & 2c \end{bmatrix}$$

[20] Also, the local maximum profit must be larger than at any other output rate. In particular, it must exceed $\pi(0)$, the profit (positive or negative) that the firm earns when $x = 0$.

Unlike the case of price-taker markets, the result above is not general. We can easily find a case in which $dp/dt > 0$. Suppose that the demand curve has constant elasticity at each price and that marginal cost is constant (its slope is 0). Recalling that $MR = p \cdot (1 + 1/Epd)$, setting $MR = MC$ implies that $p = MC/(1 + 1/Epd)$. Here, Epd is the price elasticity of demand.

For a price-searching firm to produce an appreciable amount of at good with this type of demand curve, marginal revenue must be positive. That requires that the demand curve must be elastic. That is, $Eps < -1$. This, in turn, implies that $1 + 1/Epd < 1$. Because Epd is a constant, $\frac{1}{1+1/Epd}$ is a constant markup that maximizes profits. If a firm's sale is taxed, then the tax-inclusive marginal cost is another constant, $c+t$, and the price will rise by $t \cdot \frac{1}{1+1/Epd} > t$.

Figure 7.17 shows the impact of a \$1 excise tax on the sale of products in market with linear demand and cost curves and in a market with constant elasticity and constant marginal cost. The details are in the workbook that accompanies this chapter.

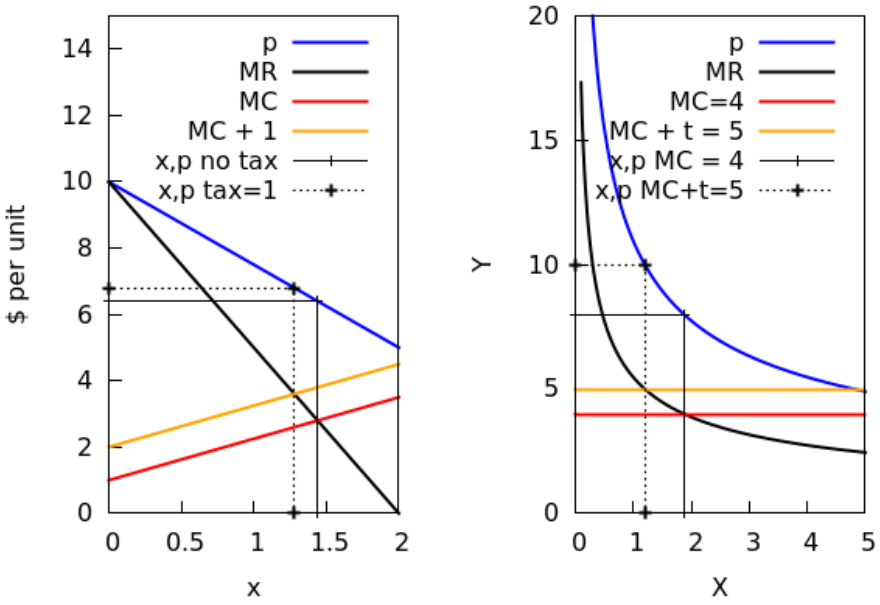


Figure 7.17: Tax effects

In the linear case, the price increase is clearly less than that vertical displacement of the marginal cost curve (which in this illustration is \$1). In

some sense, therefore, we can say that the buyers and the single seller share the cost of the tax. In the second example, however, the price rises by more than the tax (twice as much in this case because we use $Epd = -2$), so the question of "sharing" cannot be stated in simple terms.

One might be tempted to think that the monopolist gains from having a tax imposed, because the price rises by more than the tax. This inference is incorrect, because it ignores two effects of the tax. First, the monopolist must pay the tax and, second, the monopolist's output decreases. In this illustrative example the monpolist's profit (less fixed cost, which we ignore) falls from \$7.5 to \$6.0.[21]

This illustrative case that results in $dp/dt > 1$ is not the only case in which this result can occur. Bishop [2] provides the general conditions that lead to this result. Also Bishop addresses the impact of *ad valorem* taxes.

7.6.6 Inventories and Reordering

Most business firms live in a world where their production is not perfectly synchronized with their sales. Any particular firm therefore typically maintains some sort of inventory of unsold units of its output. There is a direct relationship between the number of units of inventory and the cost of keeping that inventory. Hence the firm wants to maintain as small an inventory as possible and still be able to meet anticipated customer orders. At the same time, however, there are costs associated with starting up production and reordering when the firm's inventory is depleted.

As a result, the firm must balance these two types of costs when it decides about how large an inventory to keep and how often to reorder. A large inventory increases inventory storage costs but reduces reorder costs. A small inventory decreases inventory storage costs but increases reorder costs. The optimal inventory (that inventory that minimizes the sum of storage and reorder costs) must take both types of costs into account.

Let Q be the expected sales of the firm in units in a particular time period, which we will designate a year. Suppose that $Q = 60,000$; this implies that

[21] Initially the firm sells 1.875 units for \$8 per unit and incurs a variable cost ($AVC = MC$) of \$4. Its profit is $1.875 \cdot 8 - 4 \cdot 1.875 = 7.5$. With a \$1 tax, the profit falls to $1.2 \cdot 10 - 1.2 \cdot 5 = 6$.

the firm expects to sell 60,000 units over the space of the next year. Let us further assume that these sales will be spaced evenly throughout the year, so that 60,000/ 12 = 5000 units will be sold each month.

Storage Cost. Let U represent the number of units that the firm receives when it reorders. This means that the average number of units the firm has in its inventory (assuming that the sales of the units are spaced evenly throughout time) is $U/2$. There are costs associated with maintaining a unit of inventory in terms of protection, storage, and so forth. Let c represent the cost of maintaining a unit of inventory for one year. Hence $\cdot U/2$ is the total cost of maintaining an average inventory of $U/2$ units.

Reordering cost. Assume that two separate types of costs are associated with reordering to replenish the inventory. The first type of cost is fixed in nature and does not vary with the size of the order. The cost of recording an order (which presumably does not depend on the size of the order) is an example of this type of cost. We represent this fixed cost by the letter f. The second type of cost varies directly and proportionately with the size of the order and covers the incremental cost of shipping and packaging each unit in the order.[22] Let b refer to the incremental cost associated with reordering each of U units.

The total cost of reordering in a specific instance is equal to the sum $f+b\cdot U$. Since a total of Q units is eventually needed for sale, and U units are reordered each time, a total of Q/U reorders are made during the year. This means that the total cost of reordering during the entire year is given by $(f+b\cdot U)\cdot(Q/U)$.

Total cost (storage and reordering). The total inventory cost (TIC) associated with storing and reordering is $TIC = c\cdot U/2+U+(f+b\cdot U)\cdot Q/U$, which can be rewritten as $TIC = c \cdot U/2 + f \cdot Q/U + b \cdot Q$. Note that b is not a coefficient of U.

To determine the cost-minimizing number of orders requires that we find the optimal value of U (the order size), because the number of orders placed is Q/U. Thus, we require the solution to $dTIC/dU$, which is $dTIC/dq = c/2 - 2 \cdot f \cdot Q/U^2$. Setting this expression equal to zero and solving yields the expression $U^2 = 2 \cdot f \cdot Q/c$, so that $U = \sqrt{(2 \cdot f \cdot Q/c)}$. This result is commonly called *the square root law of inventory management.* Because b is

[22]Proportionality is not required. If the relationship between order size and order cost were more complicated, the analysis would proceed in the same way. The exact nature of the solution would, of course, differ somewhat.

not a coefficient of Q, its value has no effect on the optimal inventory level (or, equivalently, the optimal number of orders per year).

To confirm that we have found a cost-minimizing order size, evaluate the second derivative, which is $d^2TIC/dU^2 = \frac{2Qf}{U^3}$. All terms in this expression are positive, so the value of U that satisfies the first-order condition corresponds to a minimum value of TIC.

We determine the optimal value of U, given the following: $f = 500, Q = 60000$. That value is $U \approx 1732.05$, which implies that the optimal number of orders, N, is 34.64. Of course, the actual number must be an integer, and might be constrained by provider-imposed restrictions. To see what other values of N imply, substitute $N = Q/U$ into TIC, so that

$$TIC = \frac{2N^2 f + Qc + 2NQb}{2N}.$$

Using the coefficient values above and $b = 1$ generates the relationship that Figure 7.18 depicts.

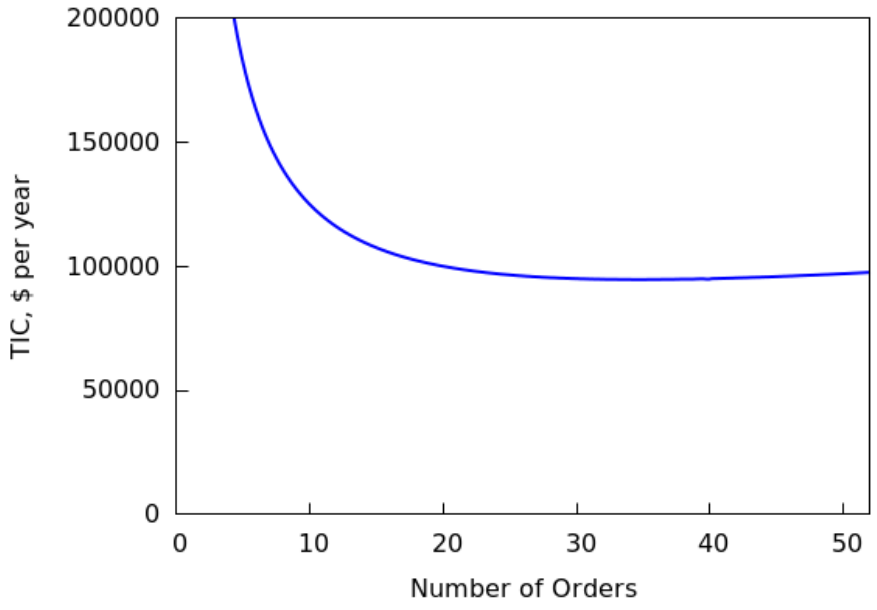


Figure 7.18: TIC vs. N

For this particular set of parameter values, TIC is not very sensitive to N

for $N > 25$ or so. Of course, this result is due to the parameters this example and is not general.

7.6.7 The Optimizing Firm

The applications section of Chapter 6 developed an example of the output-maximizing, cost-minimizing firm in detail. This section demonstrates how we can use the more sophisticated tools of this chapter to extend the problems we introduced earlier. We revisit the firm that produces an output by combining two inputs, capital (K) and labor (L). The per-unit cost of the two inputs are r and w respectively.[23]

We look at the firm's behavior from two equivalent and complementary viewpoints. First, we suppose that the firm has a fixed budget (cost that it can incur) and determine the conditions that must pertain if that budget is to be used to produce that maximum output. Then we suppose that the firm's output level is pre-determined and we determine the conditions that must pertain if that output level is to produced at the lowest possible cost. It will come as no surprise that the conditions in the two cases are the same. This result reflects the *duality* of these two approaches.

Maximizing Output Subject to a Cost Constraint

The production function for a representative firm that takes the form $Q = f(L, K)$, where Q =output, L = labor, and K = capital. The representative firm has $C0$ dollars to spend on inputs and faces a cost constraint given by $C0 = w \cdot L + r \cdot K$. The task of the firm is to maximize $f(K, L)$ subject to the constraint. We construct a Lagrangian function that reflects these facts: $W(L, K, \lambda) = f(L, K) - \lambda(C0 - w \cdot L - r \cdot K)$

Setting the first partial derivatives of W with respect to L, K, and λ equal to zero, we obtain three first-order conditions:

- $W_L = f_L - \lambda \cdot w = 0$
- $W_K = f_K - \lambda \cdot r = 0$
- $W_\lambda = C0 - w \cdot L - r \cdot K = 0$

[23]The two-input limitation is for demonstration only. The methods easily extend to any number of inputs.

The first two conditions above imply that $f_L.w = \lambda$ and $f_K/w = \lambda$, so $f_L/w = f_K/r$. This condition implies that at the marginal the return per \$ must be the same for all inputs. Alternatively, we can state the results as implying that $f_L/f_K = MPL/MPK = w/r$, that the marginal rate of technical substitution between the two inputs must equal the ratio of the marginal costs of employing the inputs.

Finally, the condition $\lambda = f_L/w = f_K/r$ provides an interpretation of the Lagrangian multiplier in this setting. Here, λ is the marginal product per dollar spent. That is, it is the change in output per (small) change in cost.

Cost Minimization Given a Level of Output

Assume that the representative firm wishes to minimize the cost of producing a certain level of output Q0. That is, the firm wishes to minimize its cost function $C = w \cdot L + r \cdot K$ subject to the constraint that it must produce $Q0$ units of output, where $Q = f(L, K)$. The Lagrangian function to be minimized is $V(L, K, \mu) = w \cdot L + r \cdot K + \mu \cdot (Q0 - f(L, K))$. Setting the first partial derivatives of V with respect to L, K, and μ equal to zero, we obtain this set of first-order conditions.

- $V_L = w - \mu \cdot f_L = 0$
- $V_K = r - \mu \cdot f_K = 0$
- $V_\mu = Q0 - w \cdot L - r \cdot K = 0$

The first two first-order conditions yield implications that are identical to the implications of the first-order conditions that we derived above. Also, we can determine that $\mu = 1/\lambda = w/MPL = r/MPK$. For this firm w is the marginal input cost of labor, the cost of acquiring one more unit of labor, and MPL, as always, the the change in output from a small change in L. Therefore, the ratio of the two is the ratio of the cost change to the output change, which is the marginal cost. Therefore, $\mu = MC$.

Profit Maximization

Now, we can reexamine the implications of profit maximization. We have established that profit maximization requires producing a quantity such that $MR = MC$. We now know that $MC = w/MPL = r/MPK$ if the firm's output is to be produced in a least-cost fashion. These conditions combine to imply that the firm must employ labor and capital at rates such that $MR =$

$w/MPL = r/MPK$. This implies that the profit-maximizing firm employs the two inputs at rates such that $w = MR \cdot MPL$ and $r = MR \cdot MPK$. These are the *marginal revenue products* of the two inputs, $MRPL$ and $MRPK$. If the firms are price-takers in their output markets $p = MR$ and the employment levels are such that $w = p \cdot MPL$ and $r = p \cdot MPK$. These are the *values of marginal products* of the two inputs, $VMPL$ and $VMPK$.[24]

7.7 Questions and Problems

1. Find the extrema, if such exist, and determine whether each extremum is a maximum or a minimum.

   a. $z = f(x, y) = 2 \cdot x^3 + y^2$

   b. $z = f(x, y) = x^2 + x^2 \cdot y + y^2$

   c. $z = f(x, y) = x^3 - x \cdot y^2$

   d. $z = f(r, s) = r + 2 \cdot r^2 + s - s^3$

   e. $z = f(P, Y) = 12 \cdot P \cdot Y - P \cdot Y^2$

   f. $z = f(L, K) = 10 \cdot L^{0.75} \cdot K^{0.25}$

2. Heinz Westphal, Vintner, imports Rhein and Mosel wines. The value of the wine V increases as time passes according to the following formula: $V = 6 \cdot 2.5^{\sqrt{t}}$, where t = ageing time in years. The present value of the wine PV, given a discount rate r and continuous appreciation, is $PV = V \cdot e^{-r \cdot t}$.

   (a) How long should Westphal hold the wine before selling it in order to maximize the present value of the wine? That is, what t maximizes PV? State your solution as a general expression in terms of r.

   (b) If $r = 0.08$, what is the corresponding t that maximizes PV?

3. The Des Moines Packing Company has a total cost (TC) function of the form $TC = f(M, L)$, where M = unbutchered meat in pounds and

[24] If the firm is a price-searching in one or more input markets, then the conditions for profit maximization change slightly.

L = hours of labor. Specifically, $TC = 3 \cdot M + 7 \cdot L$. The production function for finished, butchered meat (Q) in pounds is $Q = 2M^{0.5} \cdot L^{0.5}$. The Des Moines Packing Company wishes to produce 10,000 pounds of finished, butchered meat in this time period.

(a) Find the quantities of M and L that minimize the cost of doing so.

(b) With respect to the production function, do diminishing marginal returns exist with respect to M, L, or both?

(c) Does the production function exhibit increasing, decreasing, or constant returns to scale? (d) Does Euler's theorem apply?[25]

4. Determine (if possible) whether the following functions are monotonically increasing or decreasing. If neither, determine whether points of inflection exist and, if so, where; or whether extreme points exist and, if so, where; and whether the extreme points, if any, are maxima or minima.

a. $y = f(x) = 2 \cdot x^2$

b. $y = 6 + 0.15 \cdot x$

c. $y = 6 \cdot x^2 + 2 \cdot x + 1$

d. $y = a + b \cdot x + c \cdot x^2 + d \cdot x^3$

e. $y = 10 + 5 \cdot x + 2 \cdot x^2 - x^3$

f. $y = sin(x)$

g. $y = sin(x^2)$

h. $y = 15 - x + 2 \cdot x^2 + x^3$

i. $y = 100/x^2$

[25] At first glance, this question should bother you. Isn't the ratio of meat to finished meat close to 1? Econometricians estimating production functions like this one run into a "dominant variable" problem. One of the variables, like unbutchered meat here, varies so closely with output that the impacts of the other variables cannot be detected very accurately. This problem is one of detection and estimation, however, and not one of existence. Just as butchered meat cannot be produced without unbutchered meat, it cannot be produced without other inputs like labor and capital. It is possible that the elasticity of subsitution is much lower than 1, which is the elasticity of substitution in this example. Replacing this function with, say, a constant-elasticity-of-substitution function would not change the nature of the analysis, just its difficulty.

5. An electric power company has two generator plants, which we label A and B. The total cost functions in each plant are given by $TCA = 8 + 5QA - QA^2 + O.5 \cdot QA^3$, and $TCB = 2 + 2 \cdot QB + QB^2$, so $TC = TCA + TCB$, where TC = total cost. $Q = QA + QB$ is kilowatt hours generated (in thousands). The company wishes to minimize the cost of generating any given amount of electricity. How should it allocate production between the two generating plants if it must produce 10,000 kilowatt hours ($Q = 10$)?

6. Harold Hedonist has a utility function of the form $U = QA \cdot QB$, where U = utility, and QA and QB are quantities of two different goods or services. Mr. Hedonist, who has \$100 to spend in this time period, faces parametric prices such that $PA = \$1$ and $PB = \$2$. Use the Lagrange multiplier technique to determine the utility-maximizing quantities of QA and QB.

7. Assume that $f(x)$ is a monotonic transformation of x given that $f(x_1) > f(x_0)$ whenever $x_1 > x_0$ was previously greater than x_0. Monotonic transformations are order-preserving. With respect to ordinal utility maximization, the maximization of a monotonic transformation of a utility function yields exactly the same results as the maximization of the original utility function. Demonstrate that maximizing $U = QA^2 \cdot QB^2$ subject to $100 = QA + 2 \cdot QB$ yields the same results as those found in the preceding problem.

8. The West Mifflin Ford dealership expects to sell 1000 new Mustangs during the next year. These sales will be evenly spaced throughout the year. The cost of storing an unsold Mustang for 1 year is \$1500. The cost of placing a new order for new Mustangs is \$700 plus \$250 per new automobile ordered.

   (a) (a) What is the optimal size of order that West Mifflin should place when it orders new Mustangs?

   (b) How many such orders should West Mifflin place during this year?

   (c) Determine total inventory cost for the two integer values nearest to the computed value.

   (d) How much would total inventory cost change if Ford allows no more than 24 orders per year?

9. The state of Taxonia wishes to maximize the total tax revenue T that it receives from a per-unit tax of amount t per unit that it is going to place on the output of Monopoly, Inc., to which Taxonia grants monopoly status. The total revenue TR function of Monopoly, Inc., is given by $TR = 6 \cdot Q - Q^2$, while its total cost (TC) function in the absence of a sales tax is given by $TC = 2 \cdot Q$, where Q is units of output.

   (a) What tax per unit will maximize total tax receipts for the state of Taxonia?

   (b) How much tax revenue (T) will this tax raise?

   (c) What are Monopoly Inc.'s profit-maximizing price and output in this situation?

Chapter 8

Integral Calculus

The previous four chapters addressed differential calculus. For a function such as $y = f(x)$, we learned how to find a new function, dy/dx, which we termed the *derivative* of the function y with respect to the variable$x$ We found that the first derivative of a function is the slope of the function at a particular point.

This chapter introduces the second main branch of the calculus: the integral calculus. The integral calculus is distinctive in two specific ways. First, it enables us to define and measure the concept of area; for example, the area under a curve. As with the differential calculus, we apply the limit notion when we work with an integral. A few of the very many applications of the integral calculus to business and economics include: measuring consumer surplus, determining the total amount of depreciation that a firm will realize in a specific time span, measuring the deadweight loss due to monopoly, and finding total product when one knows only marginal product or total variable cost when one knows marginal cost.

The second way in which the integral calculus is notable is that the technique of integration is operationally the *inverse* of differentiation. Whereas in the differential calculus a function is given and one must find the corresponding derivative, in the integral calculus the derivative of the function is given and we must work backward to find the original function. This relationship is useful, for example, when we have some knowledge of the expression for marginal sales revenue, but seek an expression for total sales revenue. The backward direction of this process of integration is the reason that an integral

of a function is often called the antiderivative of that function.

8.1 The Definite Integral

We begin our analysis by considering the definite integral, which is defined as follows: If $F(x)$ is a function such that $dF(x)/dx = F'(x) = f(x)$ is its derivative for a given interval on the x axis, then $F(x)$ can be defined to be the antiderivative or integral of $f(x)$.

The process of integration is symbolized as follows: $\int f(x)dx) = F(x) + C$. The left-hand side of this equation is read, "the integral f of x with respect to x." The elongated $\int$ symbol is an *integral sign* (which, as we shall soon see, implies the summation of continuous values). We call $f(x)$ the *integrand*. That is, $f(x)$ is the function that is being integrated. The symbol dx indicates that we are integrating with respect to a variable x. On the right-hand side of this equation, $F(x) + C$ is the *indefinite integral*, while C itself is any arbitrary *constant of integration*.

The inclusion of the dx symbol on the left-hand side of this equation may seem superfluous. However, we cannot omit it, since it indicates the variable with respect to which we are integrating. If dx were absent, then the equation would be just as incomplete as a derivative that was written $df(x)/d$. One might presume that the differentiation is taking place with respect to x. However, that should not be assumed and may not be true.

In the process of finding the antiderivative or integral of a given function $f(x)$, we produce another function of x, namely $F(x) + C$. This, too, is analogous to the process of differentiation, where we differentiated a function of x and thereby generated a second function that was, in general, also a function of x.

When we find the antiderivative or integral of a function, it will in general not be unique. That is, many alternative functions could have the same derivative. For example, when $y = F(x) = 2 \cdot x$, $dy/dx = 2$; however, when $y = F(x) = 2 \cdot x + 1000$, $dy/dx = 2$ also. Thus, $2 \cdot x$ is an antiderivative of 2, and $2 \cdot x + 1000$ is also an antiderivative of 2. More generally, so also is $2 \cdot x + C$ where C is any constant, whether or not C's value is specified.

Hence, an infinite number of antiderivatives (integrals) can be associated with a particular derivative. By way of contrast, in the past four chapters,

when we found the derivative of a function, that derivative was unique. For example, if $y = F(x) = 2 \cdot x^2$, then $dy/dx = 4 \cdot x$, which is unique. There is no other value or function that is the first derivative of the function $y = 2 \cdot x^2$.

We stress the non-uniqueness of an antiderivative (integral), because it is important to the understanding of the process of integration. In general, the indefinite integral of $f(x) = 2 \cdot x$ would be

$$F(x) = x^2 + C, \text{ for } \frac{dF(x)}{dx} = \frac{d(x^2 + C)}{dx} = 2 \cdot x = f(x).$$

Geometrically, $y = x^2 + C$ represents a family of curves that are parallel to one another, but have a vertical displacement from one another.

Figure 8.1 illustrates such a family of curves for the function $F(x) = x^2 + C$. Unless we know the value of the arbitrary constant C, we cannot determine the unique antiderivative of a given function. When additional information is supplied concerning the value of the constant C, we state that the initial conditions or *boundary conditions* have been specified. In the example depicted in Figure 8.1, if we are given the initial condition that $x = 0$, and the value $F(x) = F(0) = 3$, then the value of the constant C is determined. $F(0)$ now equals $F(x) + C = F(0) + C = 3$, so $0^2 + C = 3$, and $C = 3$. Thus $F(x) = x^2 + C$ becomes $x^2 + 3$. Figure 8.1 shows $F(x) + C$ for three values of C: 0, 20, and -20.

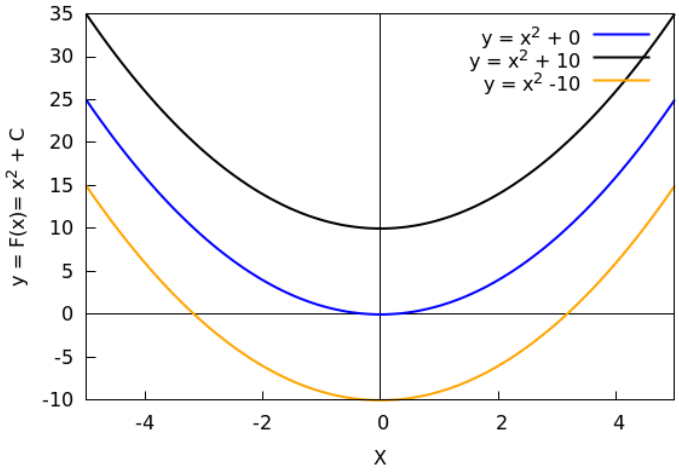


Figure 8.1: Graphical representation of the constant of integration

The $F(x)$ term of the indefinite integral $F(x) + C$ is entirely a function of x and has no definite numerical value. That is why $F(x) + C$ is referred to as the *indefinite integral*. In the absence of additional information, the value of $F(x) + C$ is unknown and indefinite.

You must pay careful attention to the notation you use when you are finding an antiderivative or integral. Whereas $f'(x)$ denoted a derivative of a function in the differential calculus, $F(x)$ is the antiderivative or original function in the integral calculus, and $f(x)$ now refers to the derivative.

In describing the process of integration in the equation $\int f(x)dx) = F(x)+C$, we used a function of the form $y = f(x)$. Functions involving the letter x are customarily used to illustrate the process of integration, just as we used functions involving the letter x to illustrate the process of differentiation. There is nothing magical about the symbol x. We could use any other letter, such as, s, t, u, v, or w, to illustrate integration with equal validity.

8.2 Rules and Properties Relating to the Integral

Chapter 5 presented a series of rules that greatly simplified the task of finding the derivative of a function. This section states a series of rules that will help you integrate a wide variety of functions.

Differentiation and integration have an inverse relationship to each other, so many of the rules relating to integration are closely related to the rules for differentiation. Specifically, it is often the case that one need only reverse a specific rule for differentiation to get the needed rule for integration. Nevertheless, this is not always the case. Some integrals are not easy to evaluate. It is customary to use published tables to assist one in evaluating integrals because of the multitude of mathematical forms that are involved.[1] In the next few pages, therefore, we consider only a few of the rules that you can use to find integrals.

[1]Enter "integral table" into your browser's search engine to get a list of tables. Computer algebra systems like *Maxima* contain extensive tables.

8.2.1 The Power Rule

Suppose that $f(x) = x^n$. Then $\int f(x)dx = \int x^n dx = \frac{1}{n+1} \cdot x^{n+1} + C = F(x)$. We prove this directly by taking the derivative of $F(x)$. This proof simply applies the definition of the integral as the antiderivative. Remembering that the differential of $y = f(x)$ is $dy = f_x dx$, the differential of the right-hand side of the power-rule equation above is

$$\frac{d\left(\frac{1}{n+1} \cdot x^{n+1} + C\right)}{dx} = \frac{n+1}{n+1} \cdot x^n = x^n, \text{ so } f_x dx = x^n dx.$$

The proof above correctly suggests that the derivative of an integral must always be equal to the integrand. That is, if the correct integration has been performed, $d(F(x) + C)/dx$ must be equal to $f(x)$.

Examples

1. $\int x^5 dx = \frac{1}{6} \cdot x^6 + C$ Check: $\frac{d(x^6/6+C)}{dx} = x^5$
2. $\int x dx = \frac{1}{2} \cdot x^2 + C$ Check: $\frac{d(x^2/2+C)}{dx} = x$
3. $\int dx = \int 1 dx = \int x^0 dx = x + C$ Check: $\frac{d(x^1/1+C)}{dx} = x^0/1 = 1$
4. $\int \sqrt{x} dx = \int x^{1/2} dx = \frac{2}{3} \cdot x^{3/2}$ Check: $\frac{d(\frac{2}{3} \cdot x^{3/2}+C)}{dx} = x^{1/2} = \sqrt{x}$
5. $\int \left(\frac{1}{x^2}\right) \cdot dx = \int x^{-2} dx = -\frac{1}{x} + C$
   Check: $\frac{d(-(1/x)+C)}{dx} = \frac{d(-x^{-1}+C)}{dx} = x^{-2} = 1/x^2$

The power rule of integration explicitly requires that $n \neq -1$. The following example demonstrates why this restriction is necessary. Let us try to find the integral of $f(x) = l/x$. $\int 1/x)dx = \int x^{-1} dx = (1/0) \cdot x^0$. Hence, when $n = -1$, the power rule no longer applies because the integral is undefined due to division by 0. The following rule deals with this type of situation.

8.2.2 The General Logarithmic Rule

The general logarithmic rule states that $\int \frac{1}{x} dx = \int x^{-1} dx = log(|x|) + C$, for $x \neq 0$. Here, *log* refers to the natural logarithm.

To prove this result, note that the differential of the right-hand side of the equation describing the general logarithmic rule is $d(log(|x|) + C)$, which is equal to $(l/x)dx$. We started our integration with $(l/x)dx$ on the left-hand side, so the general logarithmic rule is proved.

The antiderivative in the general logarithmic rule constains an absolute-value sign. This is used because logarithms do not exist for negative values of any variable x. When working with a problem in which we are certain that the domain of a variable consists only of positive values, we may omit the absolute-value sign.

8.2.3 The General Exponential Rule

Suppose that a base a is raised to the x power. That is $y = f(x) = a^x$. In this case $\int ydx = \frac{a^x}{log(a)} + C$, where log denotes the natural logarithm. To prove this assertion, recall that $d(a^x)/dx = a^x \cdot log(a)$.

An important special case of the general exponential rule is the rule for which $a = e$ the base of natural logarithms. Because $log(e) = 1$, $\int e^x dx = e^x + C$.

Before we develop additional rules that facilitate dealing with exponential functions that are not of the precise form $y = a^x$ or $y = e^x$, we confront a problem that often confuses people who are learning about integration. We have already mentioned that in order to get rid of the arbitrary constant of integration, initial conditions (boundary conditions) must be specified. In many cases the initial condition is the value of the arbitrary constant itself. For example, $\int 2 \cdot xdx = x^2 + C$. If the initial condition is $x = 0$ and $F(0) = 3$, then the constant of integration would also be 3: $F(0) = 0^2 + C = 3$, so $C = 3$.

It is possible, however, for the initial condition to have a value other than that of the constant of integration. Exponential functions sometimes furnish examples of this phenomenon. Consider the integral $\int e^x dx$ with the initial condition $F(0) = 3$. Thus $F(x) = \int e^x dx = e^0 + C$. Hence $F(0) = e^0 + C = 3$, or $1 + C = 3$, and so $C = 2$. That is, $F(0) = 3 \neq C = 2$. Therefore we should not always assume that the constant of integration and the initial condition of the function are identical.

8.2.4 Important Properties of Integration

In order to progress in our evaluation of integrals, we state some important theorems that enable us to develop further useful techniques and rules for integration.

The **additive property** is this: If both $f(x)$ and $g(x)$ are integrable, then the integral of $u(x) = f(x) + g(x)$ is the sum of the integrals of $f(x)$ and $g(x)$. More generally, the integral of a sum of a finite number of functions is equal to the sum of the integrals.

$$\int (f(x) + g(x))dx = F(x) + C1 + G(x) + C2 = F(x) + G(x) + C,$$

where $C = C1 + C2$, and more generally,

$$\int \sum_{i=1}^{n} f_i(x)dx = \sum_{i=1}^{n} F_i(x) + C,$$

where $F_i(x)$ is the antiderivative for $f_i(x)$ for all $i = 1, 2, \ldots, n$ and $C = \sum_{i=1}^{n} C_i$, the sum of the arbitrary constants of integration.

The **multiplicative property** is this: $\int K \cdot f(x)dx = K \cdot \int f(x)dx = K \cdot F(x) + C$, where K is any constant.

The **linearity property** combines the additive and multiplicative properties to this: a linear combination of n functions has an integral that equals the linear combinations of the integrals of the individual functions.

$$\int \sum_{i=1}^{n} K_i \cdot f_i(x)dx = \sum_{i=1}^{n} K_i \cdot F_i(x) + C,$$

where $F_i(x)$ is the antiderivative for $f_i(x)$ for all $i = 1, 2, \ldots, n$; $C = \sum_{i=1}^{n} C_i$, the sum of the arbitrary constants of integration; and K_i is the coefficient for $f_i(x)$.

Examples Confirm each by taking the derivative of the integral.

1. $\int -4 \cdot x^2 dx = -4 \cdot \int x^2 dx = -\frac{4}{3}x^3 + C$

2. $\int(3 \cdot x^2 - 5 \cdot x + 1)dx = 3 \cdot \int x^2 dx - 5 \cdot \int x dx + \int 1 dx =$
   $x^3 + C1 - \frac{5}{2} \cdot x^2 + C2 + x + C3 = x^3 - \frac{5}{2} \cdot x^2 + x + C$
3. $\int(8/x)dx = 8 \cdot \int(1/x)dx = 8 \cdot log(|x|) + C$
4. $\int(2 \cdot e^x + x^{-2})dx = 2 \cdot \int e^x dx + \int x^{-2} dx = 2 \cdot e^x - \frac{1}{x} + C$

8.2.5 Integration by Substitution

Our rules of integration, and the theorems stated above, deal with relatively uncomplicated integrands. As we have seen, however, more complicated integrands do exist and cannot be handled by the rules and theorems. We need a process by which we can transform a complicated integrand into the simple integrands utilized in the rules and theorems. One such process is known as the *substitution method of integration*, which "substitutes" a new variable of integration for the original variable. The object of the substitution is to transform the complicated integrand into one of the simple integrands with which our rules and theorems can deal.

The technique of integration by substitution is applicable whenever we can transform the original integral $\int f[g(x)]dx$ as follows:

$$\int f(x)dx = \int f[g(x)] \cdot g'(x)dx = \int f(u) \cdot \frac{du}{dx}dx = \int f(u)du.$$

The substitution involved requires the replacement of $g(x)$ by u and the replacement of $g'(x)dx$ by du. This substitution transforms the operation $\int dx$ into the operation $\int du$. Now, integrating with respect to the variable u, yields an indefinite integral that is a function of u, such as $F(u) + C$. We can then transform this indefinite integral back into a function of the original variable, x, by making the opposite substitution, that is, replacing u with $g(x)$ and replacing du with $g'(x)dx$. This accomplishes the needed integration by means of substitution.

We can now rewrite our four integration rules for the cases in which substitution is carried out:

1. The power rule: $\int u^n du = \frac{u^{n+1}}{n+1} + C,\ for\ n \neq -1$

2. The general logarithmic rule: $\int \frac{1}{u} du = log_e|u| + C$
3. The general exponential rule: $\int a^u du = \frac{a^u}{log_e a} + C$
4. The exponential rule, base e: $\int e^u du = e^u + C$

Integration by substitution is connected to the use of the chain rule in differentiation. Integration is, as we have pointed out, the reverse of differentiation. This means that when we introduce a new function $u = g(x)$ in the process of integration, the usual checking process (by which we ascertain whether our integral is correct) must utilize the chain rule of differentiation. That is, since integration by substitution involves the introduction of a new function u, which is a function of x, the checking process must use the function-of-a-function rule (the chain rule) in order to return us to the original function.

Examples

1. Evaluate the integral $\int 2 \cdot (e^{2 \cdot x} + 1)^2 \cdot e^{2 \cdot x} dx$.
   Let $u = e^{2 \cdot x} + 1$. Then, $du/dx = 2 \cdot e^{2 \cdot x}$ or $dx = \frac{1}{2 \cdot e^{2 \cdot x}} du$.
   Our integral, stated in terms of u is $\int 2 \cdot u^2 e^{2 \cdot x} \cdot du/(2 \cdot e^{2 \cdot x})$.
   This simplifies to $\int u^2 du = \frac{1}{3} \cdot u^3 + C$.
   This, in terms of x is $\frac{1}{3} \cdot (e^{2 \cdot x} + 1)^3 + C$.

   A computer algebra system cannot determine what substitution works, but it can reduce the amount of work and, more importantly, reduce the chance of error. In this case, the following commands are entered:
   `y: 2*(e^(2*x) +1)^2*e^(2*x)$ 'integrate(y,x);`
   `changevar(%, u=(e^(2*x) +1), u, x);`
   Note the ' before `integrate`. This tells *Maxima* to state the result in a "noun" form–that is, not to evaluate it. The `changevar` command contains the noun form that the `'integrate` command creates, the definition of the u substitution, and the name of the original independent variable, x.
   The result consists of two items. The first is the unevaluated integral $2 \int \%e^{2x} \left(\%e^{2x} + 1\right)^2 dx$ (recall that $\%e$ is *Maxima*'s notation for the constant $e = 2.718\ldots$), and the second is the integral stated in terms of u, $\int u^2 du$. If you want *Maxima* to evaluate the result of the output above, enter this command: `integrate(u^2,u);`, producing the result $\frac{u^3}{3}$. Be aware that *Maxima* does not generate the constant of integration; you must keep in mind that it exists.

2. Evaluate $\int 3 \cdot x^2 \cdot (x^3 - 4)^2 dx$. Let $u = x^3 - 4$, so that $du/dx) = 3 \cdot x^2$, or $dx = du/(3 \cdot x^2)$.
   With the pertinent substitutions,
   $\int 3 \cdot x^2 \cdot (x^3 - 4)^2 dx = \int 3 \cdot x^2 \cdot u^2 \cdot (du/(3 \cdot x^2)) = u^2 du$.
   This final expression is evaluated as $u^3/3 + C = (1/3) \cdot (x^2 - 4)^2 + C$.

3. Evaluate $\frac{dx}{x-2}$. Let $u = x - 2$, so that $du/dx = 1$ or $du = dx$.
   Now, $\int \frac{dx}{x-2} = \int \frac{du}{u} = log(|u|) + C = log(|x - 2|) + C$.

The three examples above indicate that an appropriate substitution consists of two parts that are related to each other. One part of the substitution is the *derivative* of the other part of that substitution. It is also possible that integration by substitution may result in a constant multiple of $f(u)du$. As the examples below demonstrate, however, this does not present a problem. The homogeneous property allows us to factor this constant multiple and to place it in front of the integral sign.

Examples

1. $\int 2^{4 \cdot x} dx$. Let $u = 4 \cdot x$ so $du/dx = 4$ or $dx = du/4$.
   Now $\int 2^{4 \cdot x} dx = 2 \int e^u du/4 = (1/2) \int e^u du$
   $= (1/2) \cdot e^u + C = (1/2) \cdot e^{4 \cdot x} + C$.

2. $\int \frac{dx}{2 \cdot x - 5}$ Let $u = 2 \cdot x - 5$, so $dx = (1/2) \cdot du$.
   Now $\int \frac{dx}{2 \cdot x - 5} = \int \frac{1}{u} \cdot \frac{1}{2} \cdot du = \frac{1}{2} \cdot \int \frac{1}{u} \cdot du =$
   $\frac{1}{2} \cdot log(|u| + C = \frac{1}{2} \cdot log(|2 \cdot x - 5|) + C$.

3. $\int K^{4 \cdot x} dx$, where K is a constant. Let $u = 4 \cdot x$ so that $du/dx = 4$ or $dx = \frac{1}{4} \cdot du$.
   Thus, $\int K^{4 \cdot x} dx = \int K^u \cdot \frac{1}{4} \cdot du = \frac{1}{4} \int K^u du =$
   $\frac{1}{4} \cdot K^u \cdot \frac{1}{log(K)} + C = \frac{K^{4 \cdot x}}{4 \cdot log(K)} + C$.

In order for the technique of integration by substitution to work, we must always completely transform the original integrand from one that involves one variable, say x, to a completely different function involving another variable, say u. If substitution is impossible, or is carried out improperly so that a function or functions of two or more variables results, then we must try a new substitution, for there is no general way in which we can find the integral of this new quantity.

Example

Evaluate $\int \frac{2\cdot x-3}{x^2-3\cdot x}dx$. Try this: Let $u = 2\cdot x - 3$ so that $du/dx = 2$ or $dx = du/2$.
Now $\int \frac{2\cdot x-3}{x^2-3\cdot x}dx = \int \frac{u}{x^2-3\cdot x}\cdot\frac{1}{2}\cdot du$, which is not integrable, since the substitution created a new function with two variables, x and u. The proper substitution should have been $u = x^2 - 3\cdot x$.

Now $du/dx = 2\cdot x - 3)$ so $dx = \frac{du}{x-3}$. Complete the steps to confirm that the integral is $log(|u|) + C = log(|x^2 - 3\cdot x|) + C$.

It is often possible to decide on the appropriate substitution by simple observation of the original integrand. That ability, however, usually means that you have acquired the knowledge and foresight that seem to come only with experience, some trial and error, and hard work. Integration is generally considered to be a more difficult process to master than differentiation. The correct way to integrate a function is not always readily apparent. Also, if the substitution is carried out improperly so that a function or functions of two or more variables results, then you must try a new substitution. There is no completely general way to find the needed integral by means of substitution. All of these difficulties are reasons why tables of integrals are so useful, and one of the reasons that a computer algebra system can be a useful asset.[2]

Exercises 8.1
Integrate the following integrals. Determine how many of these integrals *Maxima* can evaluate directly, without the use of the `changevar` command.

1. $(4x^3 - 3x^2 + 2x - 6)dx$
2. $(\frac{x^3-3x+2}{x^2})dx$
3. $(4x-3)^2dx$
4. $4x\sqrt{2x^2+1}\ dx$
5. $\frac{x}{x^2-6}dx$
6. $\sqrt{2x+1}\ dx$
7. $\frac{x}{\sqrt{1-x^2}}\ dx$
8. $\frac{x+1}{x^2+2x+3}\ dx$
9. $(2x-5)^2$ dx
10. $(1-x)\ dx$
11. $\frac{x^5}{\sqrt{(1-x^6)}}\ dx$
12. $x^2\left(1-x^3\right)^2\ dx$
13. $e^{-x}\ dx$
14. $\frac{e^{1/x}}{x^2}\ dx$
15. $x\,e^{x^2+4}\ dx$
16. $x^2\,e^{x^3}\ dx$
17. $\frac{e^{3x}}{e^{3x}+3}\ dx$
18. $a^{8x}\ dx$

[2]Many online integral tables are available. For example: http://integral-table.com/.

8.2.6 Integration by Parts

Another method used to transform complex, seemingly unworkable integrands into more workable forms is *integration by parts*. Just as integration by substitution was seen to be the inverse of the chain rule for differentiation, the technique of integration by parts may be viewed as the inverse of the product rule for differentiation.

Integration by parts is applicable whenever the original integral $\int g(x)dx$ can be transformed as follows:$\int f(x) \cdot g'(x)dx = f(x) \cdot g(x) - f'(x) \cdot g(x)dx + C$. This expression can be converted into a more abbreviated form by making the following substitutions: $u = f(x)$, $v = g(x)$, $du = f'(x)\ dx$, and $dv = g'(x)\ dx$.

The expression now becomes $u\ dv = v \cdot u - v\ du + C$.

In review, the procedure that we must follow when attempting to evaluate an integral by parts is to transform any original integral of the form $\int g(x)dx$ into an integral in which the only term to evaluate is $f'(x)g(x)\ dx$. If we have made appropriate choices when we substituted for $f(x)$ and $g(x)$, then the transformed integral will be easier to evaluate than the original.

Examples

1. Evaluate $\int x \cdot e^x dx$. Let $u = x$ and $dv = e^x dx$. Then $du = dx$ and $v = e^x$.
   $\int x \cdot e^x dx = x \cdot e^x - \int e^x dx = x \cdot e^x - e^x + C$.
   The *Maxima* command `integrate(x*%e^x, x)` generates the result $(x - 1) \cdot e^x$, the same as above except for the constant of integration.

2. Evaluate $\int log(x)dx$. Let $u = log(x)$ and $dv = dx$. Then, $du = 1/x$ and $v = x$.
   $\int log(x)dx = x \cdot log(x) - \int x \cdot (1/x)dx = x \cdot log(x) - x + C$.

Unfortunately no general rule dictates the best way to transform a complex integral into a more pliable one. However, a few hints can reduce the difficulty of what is otherwise often a frustrating procedure.

First, when we make a substitution, we know that $f(x)$ and $g'(x)dx$ are the two terms that make up the left-hand side of the equation. $g'(x)dx$ is the differential of $g(x)$. This means that we should choose the differential of $g(x)$

for substitution purposes. However, that differential must be integrable so that we can find $g(x)$, which is part of the right-hand side of the equation.

Second, the object of transforming the integral is to produce an integral that is more amenable to ordinary rules of integration. Hence we should choose the most complicated substitution that is possible, yet integrable, for $g'(x)dx$. That is, a more complex substitution that is integrable is preferred to a simple substitution that is also integrable. The more complete the substitution, the easier $\int f'(x) \cdot g(x)dx$, the transformed function, will be to integrate. In everyday terms, you should accomplish as much as possible with the substitution.

Fortunately, tables of integrals can reduce the difficulty of this process, though practice to learn the general nature of functional forms is required. Perhaps more fortunately, computer algebra systems routinely contain large tables of integrals and can integrate many of the expressions that otherwise would take much time and subject one to a significant risk of error.

Exercise 7.2
Evaluate the following integrals. Confirm that *Maxima* can integrate all of these expressions.

1. $\int x \cdot log(x)dx$
2. $\int log(2 \cdot x)dx$
3. $\int x^2 \cdot e^{2 \cdot x+3}dx$
4. $\int x \cdot e^{2 \cdot x}dx$
5. $\int x^2/e^{3 \cdot x}dx$
6. $\int (x+4) \cdot log(x)dx$
7. $\int x \cdot e^{-x}dx$
8. $\int log(x)/\sqrt{(x)}dx$

8.3 Applications of the Indefinite Integral

We now look at several examples in which a marginal function, represented by a first derivative, is known. For example, we know marginal cost, which is $dTC/(dQ$. What we do not know is the total cost function itself, which a is $TC = f(Q)$ In brief, we shall now examine situations in which we know the derivative of a function, and can use that knowledge to construct the function itself.

8.3.1 Marginal Cost and Total Variable Cost

Marginal cost MC, the addition to total cost TC (and to total variable cost TVC) that occurs when an incremental unit of output is produced, is given by $MC = dTC/dq = dTVC/dQ$.

We can find the total cost function, where TC = C(Q), by integrating the marginal cost function with respect to output: $TC = \int MC dQ = TVC(Q) + K$, where K is the constant of integration. Given the economic setting of this example, K is the firm's per-period fixed cost.

As an example, let $MC = 25 - 10 \cdot Q + 3 \cdot Q^2$ and fixed cost = 100. Then total cost can be written as $\int MC dQ = \int (25 - 10 \cdot Q + 3 \cdot Q^2) dQ = TVC(Q) + K$ which implies that $TC = 25 \cdot Q - 5 \cdot Q^2 + Q^3 + 100$.

We can check this result by observing that $MC = dTC/dQ = 25 - 10Q + 3 \cdot Q^2$.

8.3.2 Marginal Revenue and Total Revenue

Marginal revenue MR is the change in sales revenue a firm obtains when the quantity that is sells changes by one (small) unit of output: $MR = dTR/dQ$, where TR = total sales revenue in dollars. If we know the form of the marginal revenue function, then we can find the total revenue function by integrating the marginal revenue function with respect to output: $TR = \int MR \, dQ = R(Q) + C$.

The arbitrary constant of integration C has a value of 0 in a total revenue function. This recognizes the fact that usually the firm will realize no revenue if it does not sell any of its output. We can therefore state the total revenue function as $TR = R(Q)$.

As an example, let the marginal revenue function $MR = 25 - 2 \cdot Q$. Then it follows that total revenue is given by $dTR = \int (25 - 2 \cdot Q) dQ = 25 \cdot Q - Q^2 + C$, which is $TR = 25 \cdot Q - Q^2$ because $C = 0$.

We can check this result by observing that $dTR/dQ = 25 - 2 \cdot Q$, which is our original MR function. Also, note that the average revenue, $AR = P = 25 - Q$ (the inverse demand curve), has one-half the slope of the MR curve, a characteristic of the linear demand (and marginal revenue) curve that we have noted before.[3]

[3]The inverse demand curve is the same as the average revenue curve when, and only

8.3.3 Demand Functions, Total Revenue, and Price Elasticity of Demand

Suppose that we have a good estimate of the point price elasticity of demand Ep and that we are confident that this elasticity will remain much the same over some relevant range of prices. Thus our analysis can proceed as if we know that $Ep = (dP/dQ) \cdot (P/Q) = (dP/P) \cdot (P/dP) = K$, where K is the constant value of Ep.

Some rearrangement yields $dQ/Q = K \cdot dP/P$. We can integrate both sides so that $\in dQ/Q = K \int dP/P$. The general logarithmic rule implies that $log(|Q|) + C1 = K \cdot log(|P|) + C2$. Both price and quantity are positive values, so the absolute values can be replaced with actual values. Rewrite this expression as $log(Q) = K \cdot log(P) + C3$, where $C3 = C2 - C1$.

Exponentiation of both sides of this expression yields $Q = e^{C3} \cdot P^K$, a general expression for the constant-elasticity demand curve. We can rewrite this expression as $Q = C \cdot P^K$, where $C = e^{C3}$ and $K = Ep$, the elasticity of demand.

Suppose that $K = -1.5$, $P = 2$, and $Q = 25$. Then $25 = C \cdot 2^{1.5}$ Solving $25 = C \cdot 2^{-1.5}$ for C yields $P \approx 70.7$. Therefore, the demand curve is approximately $Q = 70.7 \cdot P^{-1.5}$ over the relevant range of values.[4]

8.4 The Definite Integral

The integral calculus was introduced in order to measure the area under a curve. Archimedes (287 -212 B.C.) successfully utilized the "method of exhaustion" in order to find the approximate area contained in a region by placing inside that region a polygonal region that more or less approximated the original region of interest. Successive polygonal regions were then introduced, each with an additional side, in order to give a closer approximation

when, each unit of the product sells for the same price. That is, if any sort of price discrimination is practiced, this relationship does not hold.

[4]Confirm that $MR > 0$ over the range of values for which this expression is a reasonable representation of the demand curve. Suppose that you are advising a seller and that the seller tells you that $Ep = -0.75$, what advice would you offer. Note that $Ep = -0.75$ implies that $MR < 0$.

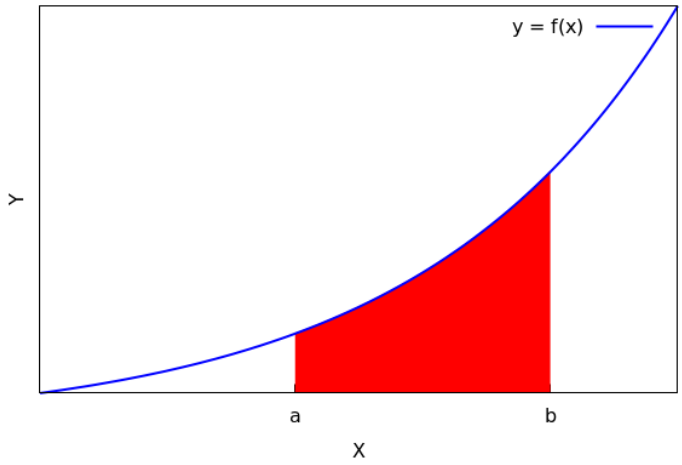


Figure 8.2: The closed interval [a,b]

of the original region of interest. Eventually, if the process were carried out long enough, the method of exhaustion would lead to a close approximation of the area of a particular region.

We use the method of exhaustion to develop an intuitive and visual idea of how the integral calculus is used to find the area under a curve. Instead of using a many-sided polygon, we use a rectangle (a four-sided polygon).[5]

Figure 8.2 illustrates the continuous function $y = f(x)$, where the domain of the function is the closed interval $[a, b]$. The problem confronting us is to calculate the shaded area, which is the area enclosed by the curve and the abscissa between points a and b. We refer to this area as A.

As an illustrative approximation to the area defined above, we divide the interval $[a, b]$ into n subintervals (where $n = 4$ in our example) as shown in Figure 8.3. Part (a) approximates the area under the curve by inscribing four rectangles below the curve between points a and b. Part (b) approximates the area under the curve between points a and b by inscribing four rectangles from above the curve. The left-hand boundary of each rectangle in part (a) has a minimum height of $y = f(x)$, whereas the right-hand boundary of each rectangle in part (b) has a height that represents the maximum value that $y = f(x)$ assumes in that subinterval.

[5]Later, we look at numerical methods for estimating areas that cannot be determined

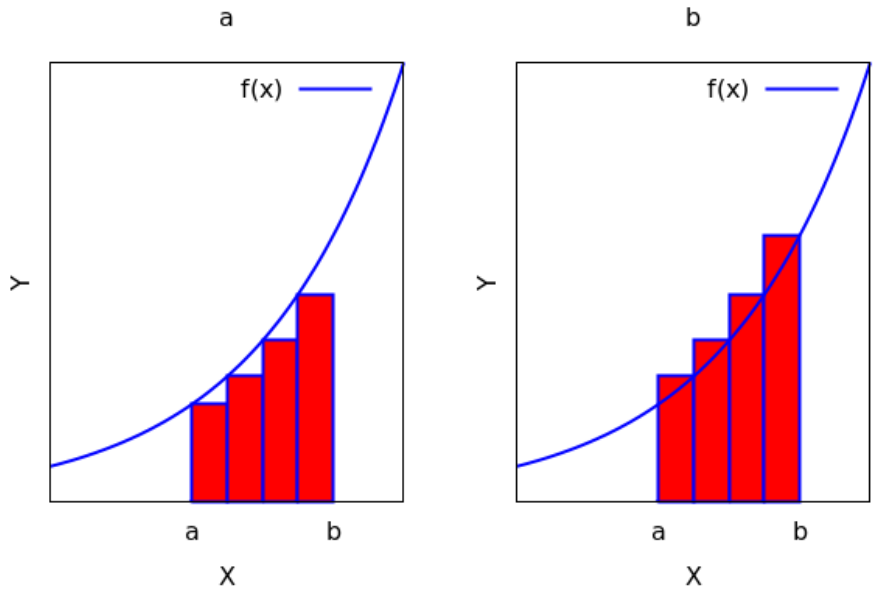


Figure 8.3: Approximating the area using rectangles: (a) approximation from below, and (b) approximation from above.

The area of a rectangle is given by the product of the height and the width of that rectangle. The first rectangle in Figure 8.3 has a height of $f(x_0)$ and a width of $\Delta x_0 = x_1 - x_0$. To generalize, the ith rectangle in part (a) has a height of $f(x_i)$ and a width of Δx_i. The area of the ith rectangle is given by $Area_i = f(x_i)\dot{\Delta}x_i$. The total area in the four rectangles between points a and b in part (a) is given by

$$A_n^- = \sum_{i=0}^{4} f(x_1) \cdot \Delta x_i.$$

We can see that this is an *underestimate* of the total area under the curve between points a and b.

In similar fashion, we can measure the area of each rectangle in Figure 8.3. This measure, which yields an *overestimate* of the area under the curve between points a and b, is equal to

$$A_n^+ = \sum_{i=1}^{5} f(x_1) \cdot \Delta x_i.$$

The two approximations to the area under the curve between points a and b are labeled A_i^- (underestimate) and A_i^+ (overestimate). It is apparent that $A_n^- < A < A_n^+$. The unshaded portions of the rectangles under the curve in part (a) and above the curve in part (b) are responsible for the differences between A_n^-, A, and A_n^+.

It is possible to achieve an even better approximation of the area under the curve in the closed interval $[a, b]$ by further subdividing that interval. Figure 8.4 illustrates the effects of increasing the number of subdivisions from 4 to 8. As n increases from 4 to 8, and Δx_i becomes smaller, areas A^- and A^+ differ less from each other, and also become closer approximations of the true area A.

In the limit, as $\Delta x \to 0$, both areas A^- and A^+ approach the true area A. That is, when the area of the inscribed rectangles in part (a) is equal to the area in the circumscribed rectangles in part (b), we have found the area under the curve A. This area is known as the *Riemann integral* or *definite integral*.

---

analytically. Those methods typically use trapezoids.

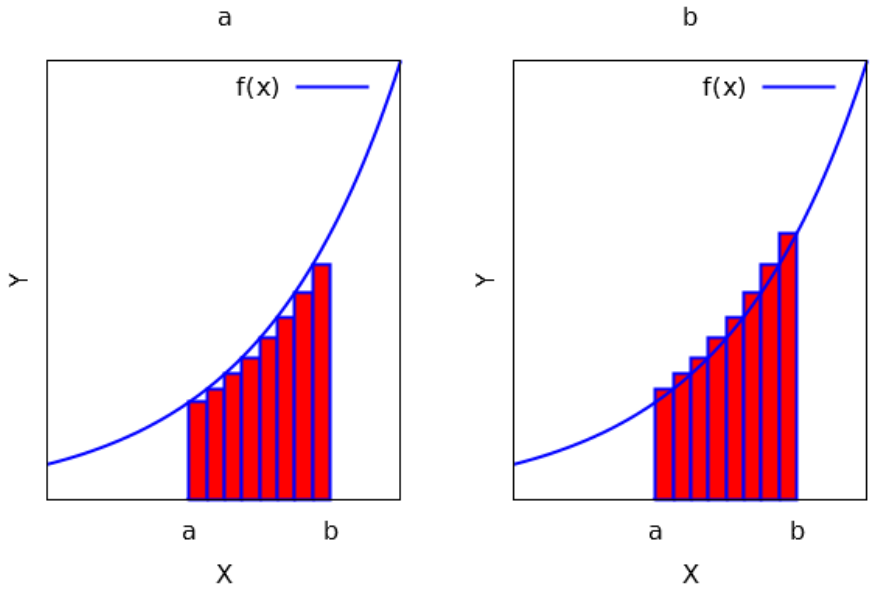


Figure 8.4: Approximating the area using rectangles: (a) approximation from below, and (b) approximation from above.

Definition: Let A_n^+ be the upper estimate and A_n^- the lower estimate of the area under the graph of $y = f(x)$ when the interval $[a, b]$ is divided into n subintervals. If

$$\lim_{n\to\infty} A_n^+ = \lim_{n\to\infty} A_n^-$$

the function $f(x)$ is said to be *Riemann integrable*, and A is said to be the Riemann or definite integral of $f(x)$ on $[a, b]$.

The expression $\int_a^b f(x)\ dx = A$ is read, "the integral of $f(x)$ from a to b is A." The letters a and b signify the limits of integration. That is, the lower limit or bound of the variable x is equal to a, and the upper limit or bound of variable x is equal to b. For example, the definite integral $\int_2^6 f(x)\ dx$ indicates that we shall integrate the function $y = f(x)$ between the values of 2 and 6 for variable x.

Four matters relating to our definition of the definite integral warrant additional discussion.

- First, the integral sign ($\int$) functions in place of the summation sign ($\sum$), which we have used so often previously in a wide range of different contexts. The integral sign indicates that the number of terms (or rectangles) to be summed is infinite–more precisely that $n \to \infty$. The integral is, therefore, a special case of a $\sum$-type summation.

- Second, the symbol representing change, Δx, has now been replaced by the integration notation dx and represents an infinitesimal change.

- Third, the indefinite integral with which we previously worked resulted in a function of variable x, whereas the definite integral results in a numeric answer that represents a specific area.

- Fourth, when we evaluate $\int_a^b f(x)\ dx$, the constant of integration that we encountered with the indefinite integral now disappears. This is the *second fundamental theorem of the calculus*, which we will shortly introduce, gives us the following result:

  $$\int_a^b f(x)dx = F(x)\Big|_a^b = F(b) + C - (F(a) + C) = F(b) - F(a).$$

  That is, when we integrate $f(x)$ over the interval from a to b, the constant of integration disappears.

The first fundamental theorem of the calculus relates $F(x)$ and $f(x)$. This is the theorem: Given an integrable function $f(t)$ on a closed interval $[a, b]$, that is, given $\int_a^b f(t)dt = F(t)$ if $a \leq x \leq b$, then the derivative of F_t exists at each value x and is equal to $f(t)$. That is, $F'(t) = f(t)$.

Examples

1. $\int_0^3 xdx = \frac{x^2}{2}\Big|_0^3 = 9/2 - 0 = 9/2$
2. $\int_{-1}^2 (x^3 - 3 \cdot x^2) = \frac{x^4}{4} - x^3\Big|_{-1}^2 = (\frac{16}{4} - 8) - (\frac{1}{4} + 1) = -\frac{21}{4}$
3. $\int_3^9 \frac{dx}{x} = log_e(9) - log_e 3 = log_3(9/3) = log_e(3)$
   The log_e rather than log is placed here as a reminder. As noted earlier, *Maxima* uses log to mean log_e and we follow this convention.

The next two examples illustrate the fact that when we use the change- of-variable technique in order to integrate a function, that is, when we integrate by substitution, we must always use new limits of integration.

Examples

1. $\int_3^{15} \frac{x}{2 \cdot x - 5}$.
   Let $u = 2 \cdot x - 5$. Then $du/dx = 2$, so $dx = du/2$. Note that when we integrate with respect to u, the new limits of integration are: When $x = 4$, $u = 3$ and when $x = 10$, $u = 15$. Thus

$$\int_3^{15} \frac{1}{2} \cdot \frac{du}{u} = \frac{1}{2} \cdot log(|u|)\Big|_3^{15} = (log(15) - log(3))/2 =$$

$$log(15/3)/2 = log(5)/2.$$

   Alternatively, before we evaluate the integral, we can convert the antiderivative back from u to x and then use the original limits of 4 and 10. That is,

$$\int_3^{15} \frac{1}{2} \cdot \frac{du}{u} = \frac{1}{2} \cdot log(|u|)\Big|_3^{15} = \frac{1}{2} \cdot log(|2 \cdot x - 5)\Big|_4^{10} =$$

$$(log(15) - log(3))/2 = log(5)/2,$$

   as before.

2. *Evaluate* $\int_0^2 3 \cdot x^2 \cdot (x^3 - 1)^2 \, dx$. Let $u = x^3 - 1$ so that $du/dx = 3 \cdot x^2$ or $dx = (1/3) \cdot x^2 \, du$. When we integrate with respect to u, the limits of integration become $u = -1$ when $x = 0$ and $u = 7$ when $x = 2$. Thus,

$$\int_{-1}^{7} u^2 \, du = \frac{1}{3} \cdot u^3 \Big|_{-1}^{7} = (343 + 1)/3 = 344/3.$$

As an exercise, convert u back to x and then use the original limits of 0 and 2 to get the same answer.

Dealing with Negative Areas

Consider the continuous function $y = f(x)$, which was depicted in Figure 8.2. We wish to find the area that lies under the curve, but above the x axis, between points a and b. That area is shaded in Figure 8.2.

It is possible, however, that a function may assume both positive and negative values. This means that the graph of such a function lies below the abscissa for some values of x, and above the x axis for other values of x. Figure 8.5 illustrates such a possibility. The area between the curve and the x axis for the interval $[c, d]$, which is indicated by "$-$": $f(x)dx$, is negative in sign. It is negative because the heights of the rectangles that are circumscribed or inscribed in that region are negative.

Definition. A negative area is the area measured by the definite integral that lies below the x axis and above the curve representing the function being integrated.

When we measure the area for the interval $[a, b]$, that is, when we find $\int_a^b f(x) \, dx$ in Figure 8.5, the positive and negative areas counteract each other. Specifically, the area for the interval $[c, d]$ is subtracted from the sum of the areas for the intervals $[a, c]$ and $[(d, b]$. However, if you are interested in the numeric or absolute value of these three areas, then you must sum the areas of the regions above the x axis, *minus* any areas of regions below the x axis ($-(A) = -A$). That is, the total or absolute area between the curve illustrated in Figure 8.5 and the x axis is, in the interval $[a, b]$, given by $\int_a^b |f(x)| \, dx$.

The absolute-value sign in $\int_a^b |f(x)| \, dx$ implies that the graph of $|f(x)|$, which is illustrated in Figure 8.5(b), coincides with the graph of $f(x)$ when

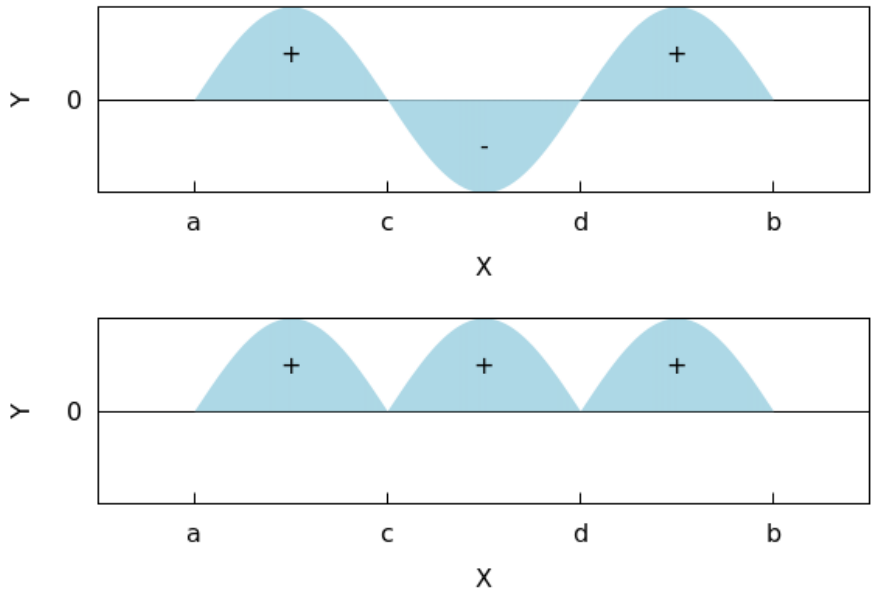


Figure 8.5: Integrating negative areas. (a) Sum of the areas' values. (b) Sum of the areas' absolute values.

$f(x) \geq 0$. When $f(x) < 0$ in some intervals, as is the case in part (b), we can obtain $f(x)$ by finding its mirror image with respect to the x axis. The area between the curve and the x axis in the interval $[c, d]$ is equivalent in absolute size in both parts (a) and (b). The area between the curve and the x axis in the interval $[c, d]$ in part (b) is the mirror image of the area between the curve and the x axis in the same interval in part (a).

As we demonstrate below, the function $|f(x)|$ is integrable on the interval $[a, b]$ whenever $f(x)$ is integrable on the same interval. That is, we can show that $\int_a^b |f(x)|\ dx$ is the sum of the positive areas minus the sum of the negative areas. Hence $\int_a^b f(x)\ dx = \int_a^c f(x)\ dx - \int_c^d f(x)\ dx + \int_d^b f(x)\ dx$.

Examples

1. Find the absolute value of the area bounded by the curve $y = x^3 - 6 \cdot x^2 + 8 \cdot x$ and the x axis, over the range $x = 0$ to $x = 4$. (See the shaded areas of Figure 8.6, left panel.)
   Area $= \int_0^2 (x^3 - 6 \cdot x^2 + 8 \cdot x)dx - \int_2^4 (x^3 - 6 \cdot x^2 + 8 \cdot x)dx =$

$\left(\frac{x^4}{4} - 2 \cdot x^3 + 4 \cdot x^2\right)\Big|_0^2 - \left(\frac{x^4}{4} - 2 \cdot x^3 + 4 \cdot x^2\right)\Big|_2^4 = 4 - (-4) = 8$
Be aware that the definite integral over the range 0 to 4 does *not* evaluate to 8. This integral is the sum of the two areas, which equals 0.

2. Find the area bounded by the curve $y = x^2 - 4 \cdot x$ and the x axis such that only positive values of x are permitted. See the shaded area in the right panel of Figure 8.6. This function has roots at $x = 0$ and $x = 4$ and is negative over the range defined by the two roots. Therefore, Area $= \int_0^4 dx = \left(\frac{x^3}{3} - 2 \cdot x^2\right)\Big|_0^4 = (64/3) - 32 - 0 = -32/3$, so the absolute value is $32/3$.

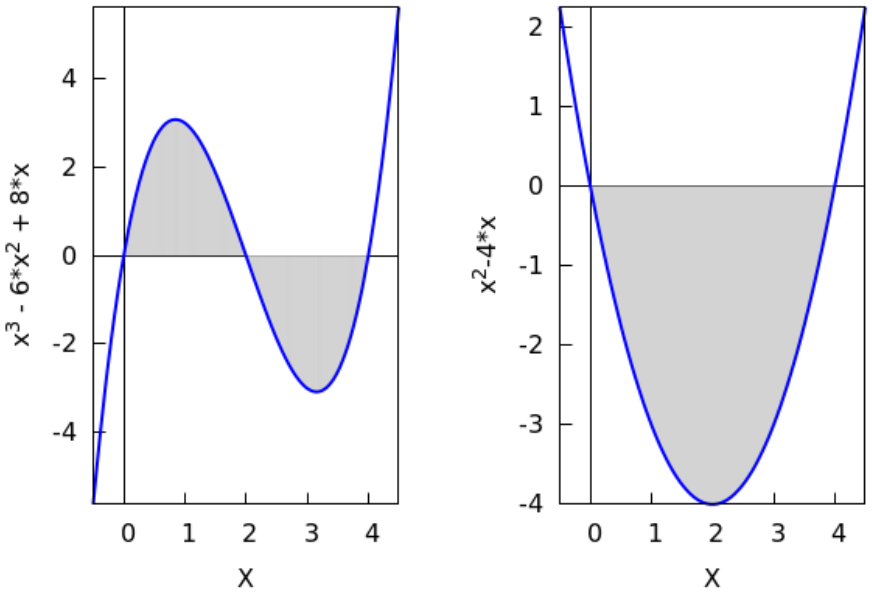


Figure 8.6: Graphs for Examples 1 and 2

8.4.1 Some Additional Properties of Integrals

We can now state some additional properties of integrals that are useful in practical situations.

1. If we interchange the limits of integration, the sign of the definite integral also changes. That is, $\int_a^b f(x)\ dx = -\int_a^b f(x)\ dx$.

   An example: $\int_0^3 x^2\ dx = x^3/3\big|_0^3 = 9 - 0 = 9$, and
   $$\int_3^0 x^2\ dx = x^3/3\big|_3^0 = 0 - 9 = -9$$

2. $\int_a^a f(x)\ dx = F(a) - F(a) = 0$

3. $\int_a^n f(x)\ dx = \int_a^b f(x)\ dx + \int_b^c f(x)\ dx + \cdots + \int_{n-1}^n f(x)\ dx$.
   This equation says that we can successfully divide a definite integral into a sum of finite subintegrals. This property warrants some discussion.

   First, consider a point that relates to the previous property. First, the limits of integration usually seem to suggest that one is counting some points or areas twice. For example, b appears first as the upper limit of the first subintegral. Then it appears as the lower limit on the second subintegral. (Likewise, for c, ..., $n-1$.) Is this a double counting of point b? The answer is no, because Property 2 demonstrates that the integral of a single point is zero. Therefore it is entirely appropriate to use a point such as b as both an upper and a lower limit in the process of integration.

   Property 3 also enables us to find the area under the graph of a function that is discontinuous. The definition of a definite integral explicitly states that the function being integrated must be continuous over the interval $[a, b]$, if it is $\int_a^b f(x)\ dx$ that we wish to find. This would seem to pose a problem if the function is discontinuous at one or more points within that interval $[a, b]$. However, Property 3 enables us to break the original integral into several subintegrals in order to avoid the problem of a discontinuous function.

   Consider Figure 8.7. Both panels depict functions that are discontinuous at point c. Hence it is impossible to integrate either of these functions over the entire interval $[a, b]$. The solution is to break the overall integral for the interval $[a, b]$ into two subintegrals, as follows: $\int_a^b f(x)\ dx = \int_a^c f(x)\ dx + \int_c^b f(x)\ dx$.[6]

[6]This discontinuity must be of either a jump or point variety in order to apply the method described here. It cannot be an infinite discontinuity.

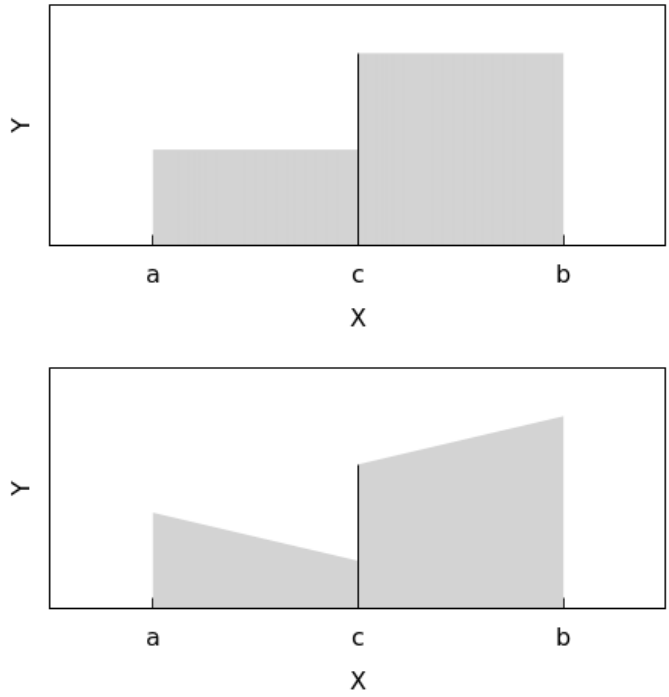


Figure 8.7: Functions wth a discontinuity

Integration over an interval requires continuity within that integral, so we can integrate over the interval $[a, c]$ and then separately integrate over the interval $[c, b]$, and sum the results. Both functions illustrated in Figure 8.7 are continuous within these subintervals, and therefore the integration can be carried out.

Evaluate the following integrals

1. $\int_0^3 x^2 \; dx$
2. $\int_{-1}^2 (2 \cdot x + 3 \cdot x^2) \; dx$
3. $\int_1^3 (x^2 - 3 \cdot x + 8) \; dx$
4. $\int_0^2 (8 \cdot x^3 + 6 \cdot x^2 - 2 \cdot x + 5) \; dx$
5. $\int_{-2}^4 (x - 1) \cdot (x - 2) \; dx$
6. $\int_{-1}^0 (x + 1)^2 \; dx$
7. $\int_{-2}^2 (x - 5)^4 \; dx$
8. $\int_2^4 \frac{4 \cdot x^3}{x^4 + 1} \; dx$
9. $\int_1^e log_e(x) \; dx$
10. $\int_4^9 \frac{1}{\sqrt{x}} \; dx$
11. $\int_{-1}^1 x^2 \cdot (x^3 + 1) \; dx$
12. $\int_0^3 e^{-2 \cdot x} \; dx$
13. $\int_0^1 x \cdot e^x \; dx$
14. $\int_{-1}^1 3 \cdot x^2 \cdot (x^3 - 4)^2 \; dx$
15. $\int_4^{10} \frac{2 \cdot x - 3}{x^2 - 3 \cdot x} \; dx$

8.4.2 Integrating the Area between Curves

An interesting and occasionally difficult problem arises when we wish to determine the area of a region that is formed and enclosed by curves. For example, consider Figure 8.8 How do we proceed when we wish to find either the area labeled A the area labeled B? We need only to reformulate slightly our basic approach to finding an integral in order to accommodate this situation.

Definition: Given: (1) two functions $f(x)$ and $g(x)$, both of which are integrable on the interval $[a, b]$. (2) $f(x) \geq g(x)$ in the interval $[a, b]$. Then $\int_a^b (f(x) - g(x)) \; dx$ is the area between the curves of these two functions.

In Figure 8.8, then, the area between functions $f2$ and $g2$, Area B, is $\int_a^b (f1 - g1) \; dx$.[7] Suppose, however, that we wish to compute the total area that $f2$

[7] Likewise for $f1$ and $g1$. This illustration is constructed so that Area A and Area B

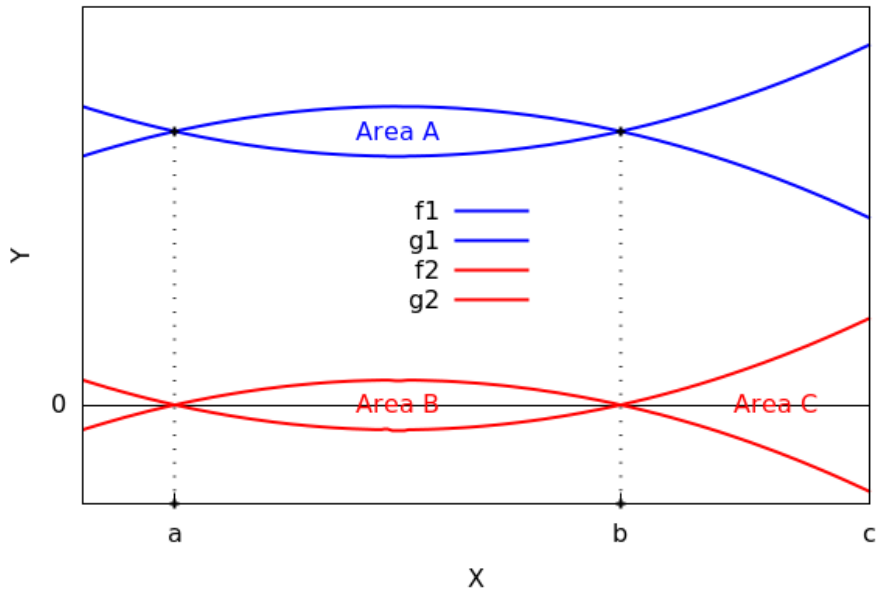


Figure 8.8: Areas between pairs of curves

and $g2$ bound between a and c. If our interest is in the absolute value of the sum of Areas B and C. In that case, integrating $(f2 - g2)$ will not work. We must integrate a function $|f2 - g2|$.

Definition. Given two functions $f(x)$ and $g(x)$, both of which are integrable on the interval $[a, b]$. The total absolute area between the curves of these two functions is given by $\int_a^c |f(x) - g(x)|\ dx$.

Example. The equations for the two expressions that define Areas B and C in Figure 8.8 are $f2 = -x^2/10 + 2 \cdot x - 5$ and $g2 = x^2/10 - 2 \cdot x + 5$. For these functions Area B = 94.281, Area B + Area C = -27.86, and absolute value of Area B plus Area C = 216.42. All values are computed in the accompanying workbook. *Maxima*'s `integrate` command does not evaluate $\int_a^c |f2 - g2|\ dx$. If b has been determined, then the absolute area can be computed as follows: $\int_a^c |f2 - g2|\ dx = \int_a^b |f2 - g2|\ dx - \int_b^c |f2 - g2|\ dx$.[8]

Whether you seek the value of the algebraic difference between two functions

are equal.

[8]Also, *Maxima*'s `romberg` command, which we consider below, can evaluate the expression in terms of absolute values.

or the absolute value depends on the question at hand. To consider an economic application, suppose that $f1$ is a marginal value function (whatever the units of value might be), and that $g1$ is a marginal opportunity cost (in the same units of value as $f1$). Suppose that we wish to determine the net value of x. Then, we seek the algebraic difference between the two functions, not the absolute value.

Thus, if b units could be produced, the net gain would be Area A less the small area to its left. Inspection reveals that this combined area is positive (in fact, its value is 80.474). Suppose, however, the x can be provided only in a quantity $x = c$. In that case, the total net benefit of x is Area A less the sum of the small area to the left of Area A and Area C, which appears to be negative (as the accompanying workbook shows, the value is -27.86.) Producing c units of x destroys value.

Occasions sometimes arise when it is easier to find the area between two curves if one conceptually interchanges the roles of the x and y axes. Sometimes, doing so is necessary. Consider the two curves illustrated in Figure 8.9. We wish to find the shaded area in the interval $[c, d]$ (*i.e.*, two values of y) between the two curves. We must write the relationship between variables x and y as $x = g(y)$ for the curve farthest to the left, because x values do not map to unique y values. Likewise the relationship between variables for the curve to the right is written as $x = f(y)$ The two functions $g(y)$ and $f(y)$ are integrable over the interval $[c, d]$, as follows: $\int_c^d$.

Often, when two curves bound an area, the choice of the variable with respect to which we integrate can be selected entirely as a matter of convenience. Either choice will yield the same value for the area. Two examples follow.

Example 1. Determine the size of the area bounded by the curves $y^2 = 2 \cdot x$ and $y = 2 \cdot x - 2$. (See Figure 8.10.) The two curves intersect at (2, 2) and (1/2, -1), which are found by solving the two equations simultaneously.

For $y^2 = 2 \cdot x$, y is not a function of x, but x can be stated as a function of y: $x = y^2/2$. Even so, we can integrate with respect to x by breaking "Area" into two parts. Between $x = 0$ and $x = 1/2$, the area is the difference between $\sqrt{2 \cdot x}$ and $-\sqrt{2 \cdot x}$ Between $x = 1/2$ and $x = 2$, the area is the difference between $\sqrt{2 \cdot x}$ and $2 \cdot x - 2$. Look at the first panel of Figure 8.10. Thus,

$$\text{Area} = \int_0^{1/2} \sqrt{2 \cdot x} - (-\sqrt{2 \cdot x}) + \int_{1/2}^2 \sqrt{(2 \cdot x)} - (2 \cdot x - 2)\ dx =$$

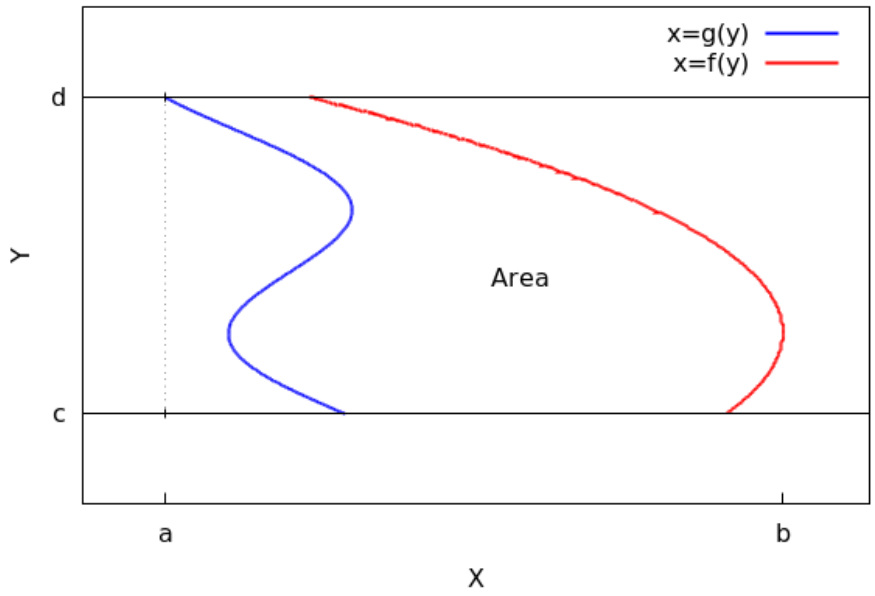


Figure 8.9: Integrating with respect to variable y

$$\tfrac{2}{3}\cdot(2\cdot x)^{3/2}\Big|_0^{1/2}+\left(\tfrac{1}{3}\cdot(2\cdot x)^{3/2}-x^2+2\cdot x\right)\Big|_{1/2}^{2}=\tfrac{27}{12}=\tfrac{9}{4}.$$

Alternatively, and much more easily, we can integrate with respect to y. The two curves are drawn with y as the independent variable in the second panel of Figure 8.10. The integral is this:

$$\text{Area}=\int_{-1}^{2}\left(\tfrac{y+2}{2}-\tfrac{y^2}{2}\right)\,dy=\tfrac{1}{2}\cdot\left(\tfrac{y^2}{2}+2\cdot y-\tfrac{y^3}{3}\right)\Big|_{-1}^{2}=\tfrac{27}{12}=\tfrac{9}{4}.$$

Example 2. Find the area bounded by the curves y= x^2 and $y=2\cdot x$. (See Figure 8.11.) Solving the two equations simultaneously, we find that the points of intersection are (0, 0) and (2, 4). Thus

$$\text{Area}=\int_0^2(2\cdot x-x^2)\,dx=(x^2-x^3/3)\Big|_0^2=4/3.$$

Alternatively, we can integrate with respect to y, adjusting for the proper limits, so that:

$$\text{Area}=\int_0^4\left(\sqrt{x}-\tfrac{y}{2}\right)\,dy=\left(\tfrac{2}{3}\cdot y^{3/2}-\tfrac{y^2}{4}\right)\Big|_0^4=\tfrac{4}{3}.$$

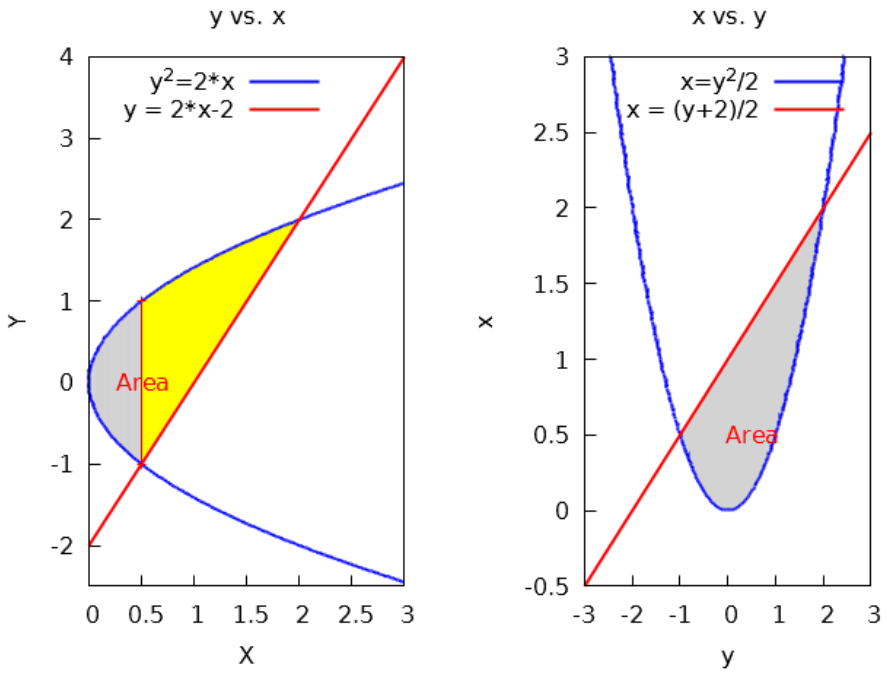


Figure 8.10: Area bounded by a pair of functions, two views

Exercise 8.4
Draw a sketch (either by hand or using *Maxima*) bounded by the following expressions. Compute the values by hand and check your solution with *Maxima*.

1. $y = x^3, y = 0, x = 0, x = 2$
2. $y = 9 - x^2, y = x + 3$
3. $y = 3 - x^2, y = -2 \cdot x$
4. $y = 6 - x, y = x + 2, y = 0$
5. $y = 6 - x, y = x + 2, y = 8$
6. $y = 6 \cdot x - x^2, y = x^2 - 2 \cdot x$
7. $y^2 = x, y = x/2 - 3/2$
8. $y = y = x^3, y = 2 \cdot x + 4, x = 0$
9. $y = x, y = 10{-}4{\cdot}x, y = 0, x = 0$

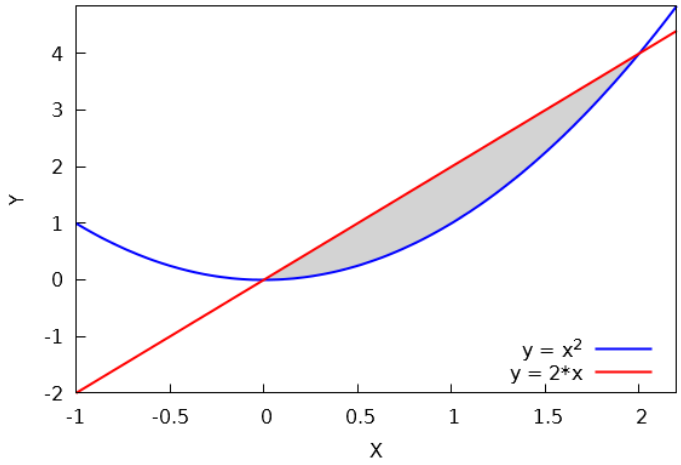


Figure 8.11: Graphs of $y = x^2$ and $y = 2 \cdot x$

8.4.3 Improper Integrals

Our examination of integration has thus far assumed a continuous function of the form $y = f(x)$, which is defined for the closed finite interval $[a, b]$. When an integral satisfies these restrictions, it is said to be a *proper integral*. This section discusses integrals that are said to be *improper*. Specifically, we examine situations in which the restrictions for a definite integral are relaxed.

We consider two general types of improper integrals. The first occurs when there are infinite limits of integration. The second occurs when there is an infinite integrand.

Case 1: Improper integral due to an infinite limit of integration
When the limits of integration are no longer finite, for example, when we wish to study the definite integral $\int_a^b f(x)\ dx$ as $a \to \infty$ and/or as $\to -\infty$ we have an *improper integral*. In such a case, it is not possible to find a finite value for the integral. This is because $F(\infty) - F(0)$ is meaningless, as are $F(b) - F(-\infty)$ and $F(\infty) - F(-\infty)$.

Definition. An improper integral with an infinite limit of integration is formally symbolized by

$$\int_a^{\infty} f(x)\ dx = \lim_{b\to\infty} \int_a^b f(x)\ dx \text{ or } \lim_{b\to\infty} F(x)\Big|_a^b.$$

Such an integral is said to be *convergent* when the limit exists and is finite, whereas it is said to be *divergent* when the limit does not exist.

We can use the definition in any particular case by initially finding $\int_a^b$, that is, by finding the indefinite integral $F(x)$. Second, we evaluate $F(x)$ for a and b, then find the limit as $b \to \infty$. If the limit is finite, then the integral exists and is convergent. If the limit is infinite, then the integral is diverging and has no finite value.

It is not uncommon to see an improper integral written without the limit notation in front of the integral. That is, instead of $\lim\limits_{b\to\infty} f(x)\ dx$ we often see the shorthand expression $\int_a^\infty f(x)\ dx = F(x)\big|_a^\infty$. This shorthand notation nevertheless must be evaluated with the limit concept held firmly in mind. This implicit step must be carried out, since the limit may be divergent, and if it is, the integral has no finite value.

The existence of an improper integral with an infinite limit for its upper bound does not change the fact that we are measuring the area under a curve. Figure 8.12 illustrates the graph of a function $y = f(x)$ where the upper limit of integration b is infinite. That is, $\int_a^b f(x)\ dx = \int_a^\infty f(x)\ dx$.

If the improper integral is convergent, that is, if the limit exists, then the shaded region under the curve is considered to be a finite area. However, if the improper integral is divergent, then a limit does not exist and the shaded area under the curve is infinite in size.

It is possible, of course, for the lower bound of integration to be infinite as well. In this case, the lower bound a tends to $-\infty$. We can define the improper integral $\int_{-\infty}^b f(x)\ dx$ as $\lim\limits_{a\to-\infty} \int_a^b f(x)\ dx$. We then use the usual procedure to determine whether the improper integral is convergent or divergent.

A more complicated case is the situation in which both limits of integration are, infinite; that is, we wish to find $\int_{-\infty}^\infty f(x)\ dx$.

Definition. An improper integral with both limits of integration infinite exists when, for any real number C,

$$\int_{-\infty}^{\infty} f(x)\ dx = \int_{-\infty}^{C} f(x)\ dx + \int_{C}^{\infty} f(x)\ dx = \lim_{a\to-\infty, b\to\infty} \int_a^b f(x)\ dx.$$

Both integrals, $\int_{-\infty}^C f(x)\ dx$ and $\int_C^\infty f(x)\ dx$, must be convergent in order for the improper integral $\int_{-\infty}^\infty f(x)dx$ to be convergent. If either of the two

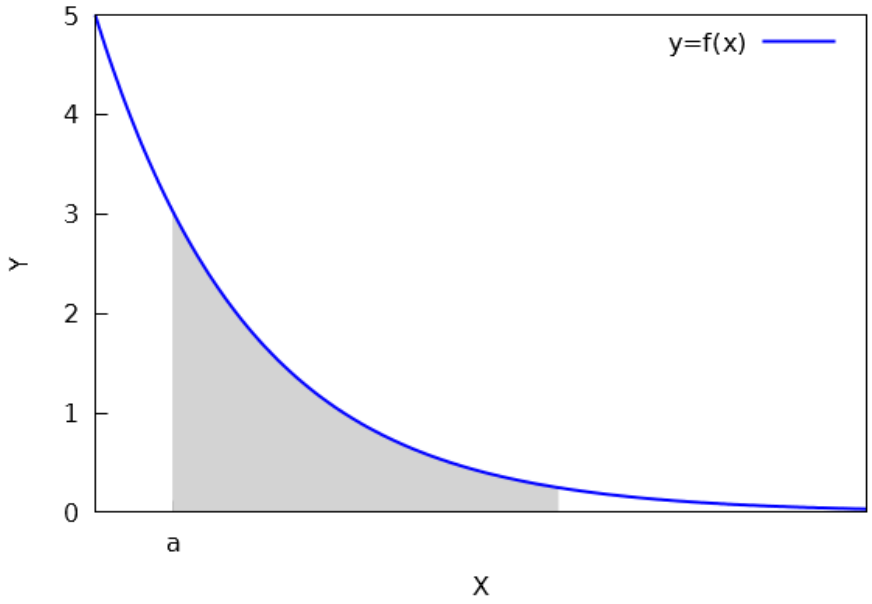


Figure 8.12: Improper integral: the case of an infinite limit

integrals is divergent, then the improper integral $\int_{-\infty}^{\infty} f(x)dx$ is divergent.

Examples

1. $\int_{-\infty}^{0} e^{3\cdot x}\ dx = \lim_{a\to-\infty} \int_{a}^{0} e^{3\cdot x}\ dx = \lim_{a\to-\infty} \frac{e^{3\cdot x}}{3}\Big|_{a}^{0} = 1/3 - 0 = 1/3$

2. $\int_{1}^{\infty} \frac{1}{\sqrt{x}}\ dx = \lim_{b\to\infty} \int_{1}^{b} dx/x = \lim_{\to\infty} 2\cdot\sqrt{x}\Big|_{1}^{b} = \lim_{b\to\infty} (2\cdot\sqrt{b} - 2)$
   This integral is divergent; its limit does not exist.

3. $\int_{-\infty}^{\infty} e^{x}\ dx = \lim_{a\to-\infty,b\to\infty} \int_{a}^{b} e^{x}\ dx = \lim_{a\to-\infty,b\to\infty} e^{x}\big|_{a}^{b} = \lim_{a\to-\infty,b\to\infty} (e^{b} - ea)$
   The term e^b grows without bound as b increases, so the last term does not have a limit. Therefore, the integral is divergent.

Instructing *Maxima* to attempt to evaluate the three intergrals yields these results. The command `integrate(%e^(3*x),x,minf,0);` yields the result $\frac{1}{3}$. Both the commands `integrate(1/x,x,1,inf);` and `integrate(%e^x,x,minf,inf);` yield the same warning, "defint: integral is

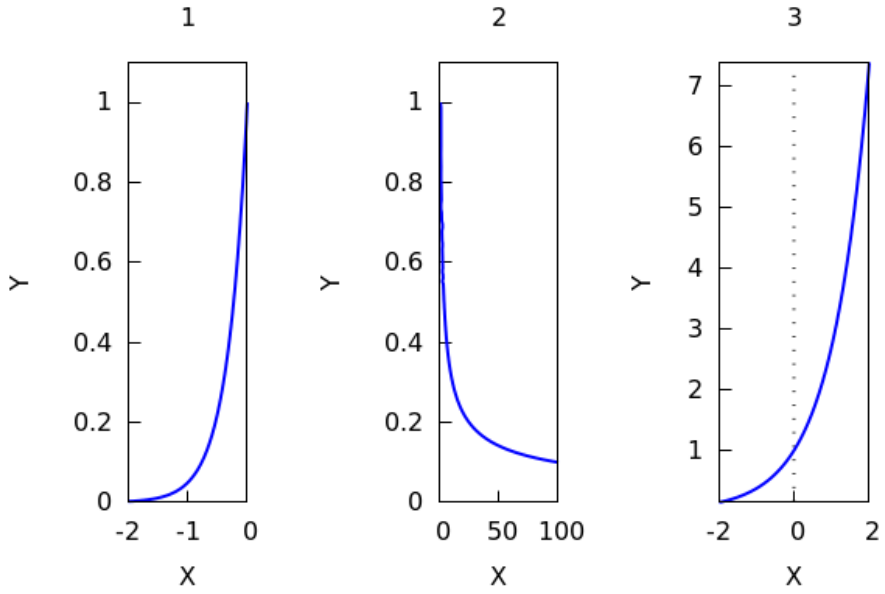


Figure 8.13: Three examples of improper integrals

divergent."

Case 2: Improper integral due to an infinite integrand
The second type of improper integral occurs whenever the integrand becomes infinite (due to an infinite discontinuity). It does not matter here that the limits of integration are finite. We need to examine three instances involving improper integrands.

The first instance involves an *infinite integrand at lower limit.*
Definition. Given: an integral $\int_x^b f(t)\ dt$ that exists for the interval $a < t \leq b$. Define a new function $F(x)$ such that: $F(x) = \int_x^b f(t)\ dt$ for $a < t \leq b$ The function $F(x)$ is said to be an improper integral at the point $x = a$ and is denoted by the symbol $\lim\limits_{x \to a} \int_x^b f(t)\ dt$.

The second instance to consider involves an *infinite integrand at the upper limit.* The definition above considers the circumstance in which the integrand becomes infinite at the lower limit of integration. It is also possible for the integrand to become infinite at the upper limit of integration:
$\lim\limits_{x \to b} \int_x^b f(t)\ dt$ for $a \leq t < b$.

Either of the improper integrals noted in the preceding paragraphs can be evaluated in the limit to determine whether the improper integral is convergent or divergent. This analysis duplicates that in which we were concerned with infinite limits to integration, but the integrand was finite.

Examples

1. $\int_0^3 \frac{dx}{x-3} = \lim_{b\to 3} \int_0^b \frac{dx}{x-3} = log_e(|x-3|)\Big|_0^b = \lim_{b\to 3} log_e(|b-3|) - log_e(|-3|)$
   The limit does not exist, and the integral is divergent.

2. $\int_0^1 \ dx = \lim_{a\to 0} \int_0^1 \frac{dx}{x} = \lim_{a\to 0} log_e(|x|)\Big|_a^1 = \lim_{x\to a}(log_e(|1|) - log_e(|a|)$
   The limit does not exist, and the integral is divergent.

3. $\int_0^1 \frac{dx}{\sqrt{(x)}} = \lim_{a\to 0} \int_a^1 \frac{dx}{\sqrt{(x)}} = \lim_{a\to 0} 2 \cdot \sqrt{x}\Big|_a^1 = \lim_{a\to 0}(2 - 2 \cdot \sqrt{a}) = 2$

Case 3: Improper integral due to an infinite discontinuity
Finally, the third instance involves an *infinite integral due to an infinite discontinuity*. The case sometimes arises in which a function $y = f(x)$ is discontinuous at some point c. Then the integral $\int_a^b f(x)\ dx$ is defined for the interval $[a, b]$, except at point c, when $a < c < b$. The additivity theorem relating to integrals tells us that we can write $\int_a^b f(x)\ dx = \int_a^c f(x)\ dx + \int_c^b f(x)\ dx$.

Only when both of the two integrals on the right-hand side of this equation converge can we be certain that the improper integral $\int_a^b f(x)\ dx$ also converges. It is not sufficient for only one of the integrals on the right-hand side of the equation to be convergent.

The analysis in the equation $\int_a^b f(x)\ dx = \int_a^c f(x)\ dx + \int_c^b f(x)\ dx$ relates to a situation in which one finite discontinuity exists. We can extend this equation to deal with the situation in which any *finite* number of such infinite discontinuities exist. For example, assume that $y = f(x)$ is infinitely discontinuous at points c and d, where $a < c$ and $d < b$. Then we have $\int_a^b f(x)\ dx = \int_a^c f(x)\ dx + \int_c^d f(x)\ dx + \int_d^b f(x)\ dx$. All three of the integrals on the right-hand side must be convergent in order for us to assert that the improper integral is convergent.

Examples

1. Evaluate $\int_1^5 dx/(x-2)^2$. The integrand is discontinuous at $x = 2$ so we restate it as $\int_1^5 dx/(x-2)^2 = \int_1^5 dx/(x-2)^2 + \int_1^5 dx/(x-2)^2$. Evaluating these integrals reveals, however, that both are divergent. If either is divergent, then so is $\int_1^5 dx/(x-2)^2$.

2. Evaluate $\int_1^5 dx/\sqrt{x^2-9}$. This integral poses two sources of difficulty. First, for $x < 3$ it's value is is a complex number. Second, it is discontinuous at $x = 3$. Evaluating the entire integral using $Maxima$ yields this result: $\log(3) - i\,\text{atan}\left(\frac{\sqrt{5}}{2}\right)$, which is a complex value. Integrating from 3 to 5 yields $log(3)$. Thus, the integral converges. Evaluating this interval by hand involves some trigonometric substitutions. See Mitchell [**?**].

Exercise 8.5. Evaluate the following integrals.

1. $\int_0^\infty \frac{x^2}{\sqrt{x^3+1}}\,dx$
2. $\int_3^3 dx/(3-x)$
3. $\int_1^\infty dx/x^2$
4. $\int_{-\infty}^0 (x \cdot e^x)\,dx$
5. $\int_{-\infty}^\infty x \cdot e^{-x^2}\,dx$
6. $\int_0^\infty e^{-x}\,dx$
7. $\int_0^1 log_e(x)\,dx$
8. $\int_{-1}^1 dx/x^4$
9. $\int_{-1}^8 \frac{dx}{1/x^3}$
10. $\int_{-1}^0 \frac{x}{x^2-1}\,dx$
11. $\int_1^\infty x \cdot log_e(x)\,dx$
12. $\int_0^2 \frac{x}{\sqrt{4-x^2}}\,dx$

8.5 Economic Applications

The concept of the definite integral pertains to numerous economics applications. The following examples are illustrative.

8.5.1 Consumer Surplus and Producer Surplus

The concept of *consumer surplus* is often used in applied welfare economics to evaluate the desirability of particular policy (or regime) options. For example, one can use the idea of consumer surplus to measure the loss that consumers realize as a result of the exercise of business and labor monopoly

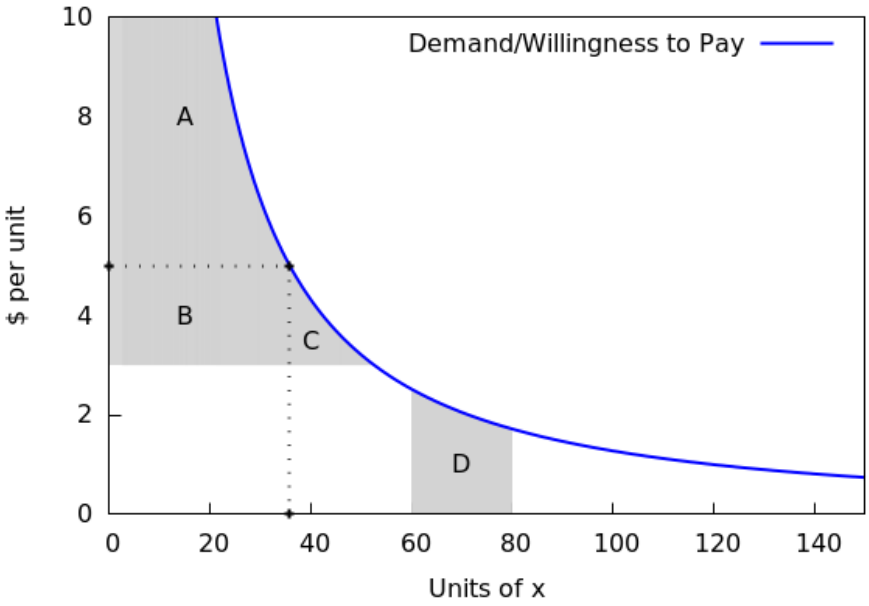


Figure 8.14: Consumer Surplus

power. One can also use consumer surplus to help make a decision about the desirability of building a new highway, a new lock and dam, or of expanding a wilderness area.[9] Figure 8.14 provides a framework for examining the basic aspects of consumer surplus.

Start with area A in Figure 8.14. The area of A (which extends off the graph) is the numerical value of consumer surplus when consumers can buy the quantity that they select at a price of \$5 per unit. Mechanically, this area is $\int_0^{x_1}(p-5)\,dx$, where x_1 is the quantity demanded when $p = 5$ ($x1 \approxeq 35.9$). The value of p at each quantity is the price that a consumer is willing and able to pay for that unit. Hence, the demand curve is also a willingness to pay curve.

For this demand curve, which is $x = p^{-3/4}$, area A is approximately \$17,218. We find this area by integrating the inverse demand curve, $p = (x/120)^{-4/3}$. The integral for the inverse of this inelastic demand curve is divergent, so we

[9]The concept of consumer surplus is more complex than the present illustration indicates. For details and references, see Hammock and Mixon [7]. Also, while the concept is fairly precise, applications in evaluating projects like those cited above must involve a significant degree of imprecision.

integrate from $x = 0.001$ to $x = x_1$). The exact value of area A is typically of little interest and often cannot be interpret in a meaningful way. Suppose that some small amount of x is a requisite for life. Then placing a value on the first few units per time period makes no economic sense. In any event, the question of the total value of consumer surplus is not of interest.

More interesting are areas B, C, and D. The area B + C can be determined in either of two ways. First, we could determine the value of x when $p = 3$ (the lower price here) and integrate the inverse demand curve from $x = 0.001$ to that value of x. This would yield area A + B + C. Subtraction provides our solution. Alternatively,we could integrate the demand curve as follows: $\int_3^5 x\ dp$, where x is the a function of p. For this example, the area is approximately \$86.05.

Consider two reasons that the price might be \$5 rather than \$3, or *vice versa*. First, a \$2 per unit tax might be imposed on a good for which the market price as \$3. Then area B is the tax revenue and area C is the deadweight loss (a loss to consumers but that does not accrue to anyone). Alternatively, the higher price might reflect monopoly privilege, in which case the seller gains area B, which the consumers lose area B + C. Again, area C depicts a deadweight loss.

Finally, look at area D. This area that would be of interest if a park were to be expanded, or some other service were to be provided without requiring a payment by users (added highway lanes, for example). Integrating the inverse demand curve over the range $x = 60$ to $x = 80$ would provide a measure of the value to this expansion to users and provides guidance as to whether the expansion is warranted. For this illustration, the value of area D is approximately \$41.47.

The concept of *producer surplus* is analogous to that of consumer surplus. For price-taking firms, we can define producer surplus as the difference between the price that producers receive for their product and the price that one or more of the firms must be paid to produce the marginal unit. In fact, this amount is the marginal cost of producing the additional output, either by expanded the output from incumbent firms or by the entry of more firms, or both.

In a market with price searchers, the producer surplus is the total revenue less the integral of the marginal cost curve. We limit out attention to the case of price takers. Before offering illustrative examples, we consider a

central feature of markets that consist of price searchers: the quantity that maximizes the sum of consumer and producer surplus is the equilibrium quantity.

Let $f(x)$ and $g(x)$ be the inverse demand and supply curves. The area between these is Surpluses $= \int_0^x (f(x) - g(x)\ dx = F(x) - G(x) + C$, where C is the unknown constant of integration. It is apparent upon inspection that the first-order condition for maximizing the "Surpluses" function is that $f(x) - g(x) = 0$. The remarkable conclusion is, therefore, that the equilibrium quantity generates maximum combined surpluses.[10]

We now consider two examples in which consumer surplus and producer surplus are computed. Figure 8.15 illustrates these two examples. Also, we consider an example in which areas under demand curves can be evaluated and interpreted but cannot be cleanly divided into consumer surplus and producer surplus.

Example 1. Find the consumer's surplus given that the demand and supply functions in a price-takers' market, for a particular commodity are these: Demand: $p = 30 - 2 \cdot x^2$, Supply: $p = 3 + x^2$.

The positive solution to this set of equations, $x = 3$ is the equilibrium quantity. Substituting that value into either the demand or supply curve yields $p = 12$ as the equilibrium price. The integrals that we seek to evaluate are these: CS $= \int_0^3 (30 - 2 \cdot x^2 - 12)\ dx = 36$ and PS $= \int_0^3 (12 - (3 + Q^2)\ dx) = 18$.

The meaning of producer surplus requires some elaboration. Which producers receive it, and how? Consider first a constant cost industry (horizontal long-run supply). In this case, the surplus does not exist. Next, suppose that an industry consists for a large number of identical firms and that as the number of firms increase, the prices of some inputs increase. Then the producer surplus accrues to the owners of those resources, with the firms earning zero economic profit. Finally, consider an industry of firms with different cost curves. Then the firms with costs lower than those of the marginal firms can earn profits. Also, owners of resources employed by the firms in the industry might receive part of the surplus. See [7], Chapter 9.

[10] This does not imply, however, that the quantity is optimal. The inverse demand and supply curves do not take into account external effects of producing or consuming the product. Also, willingness to pay is one criterion in evaluating the marginal value of a unit. Other criteria, such as paternalism, could enter one's analysis.

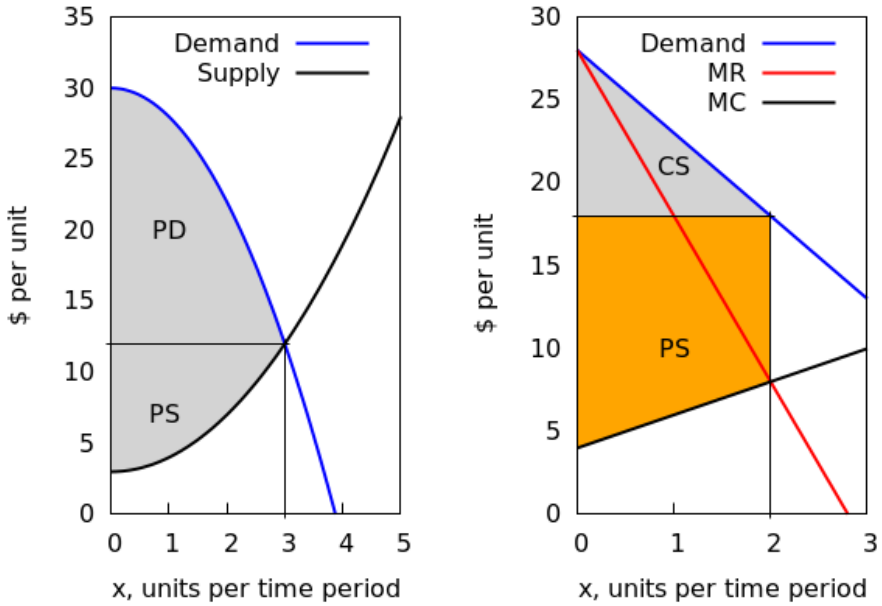


Figure 8.15: Consumer and producer surplus illustrations

Example 2. The inverse demand function for a particular commodity is $P = 28 - 5 \cdot x$, and the marginal cost of producing and marketing the commodity is $MC = 2 \cdot x + 4$. We determine the consumer surplus and the producer surplus.

We find maximum profit by setting marginal revenue equal to marginal cost and finding the profit-maximizing output, $x = 2$. Inserting this value into the demand curve yields $p = 18$. The relevant integrals are $\text{CS} = \int_0^3 (28 - 5 \cdot x - 18)\, dx = 10$ (the gray triangle in Figure 8.15) and $\text{PS} = \int_0^3 (18 - (2 \cdot x + 4))\, dx = 24$ (the orange trapezoid in Figure 8.15).

It is easy to confirm that the orange area is the same size as the triangle defined by the y axis, the marginal revenue curve, and the marginal cost curve. This common area is the amount that producing 2 units adds to the firm's profit. That profit is, therefore, PS - fixed cost.

Example 3. This is a stylized response to a rhetorical question that is often posed: What does it say about a society's values that elite athletes earn a multiple of the amount the school teachers earn? We will see that the correct (if not convincing) answer is, "Nothing." The wages paid to individuals

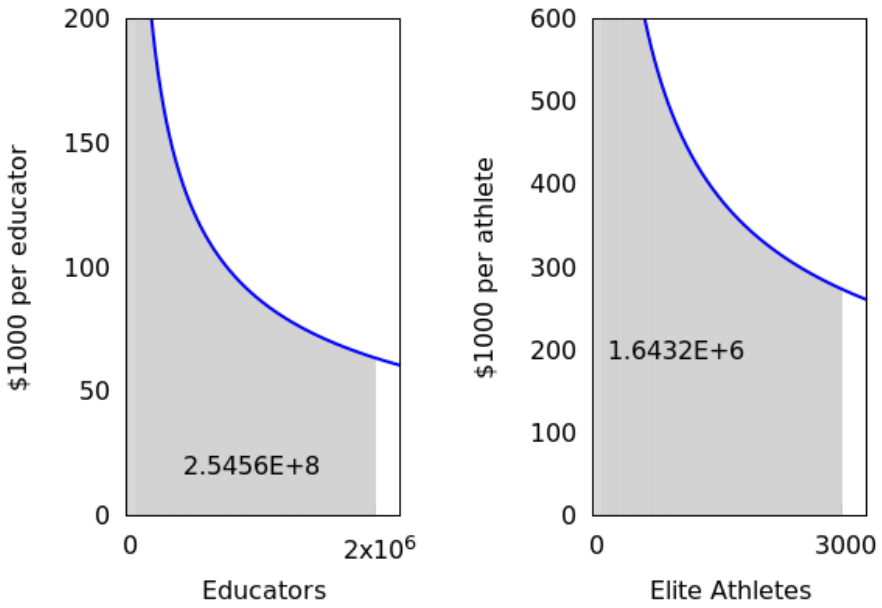


Figure 8.16: Illustrative marginal value functions

roughly (often very roughly) approximate the *marginal* value of their services. To determine the total value of these service (however that value is defined) requires integrating the marginal value function. An important proposition of elementary calculus is that one cannot determine total value by looking at the margin (derivative).

To illustrate this point, suppose that the marginal value function for educators is $MVedu = 50000/\sqrt{edu}$ and $MVath = 500/\sqrt{ath}$, where edu is the number of educators and ath is the number of elite athletes.[11]

For the values used here, the resulting marginal value of educators is approximately \$63,640 per year and that of elite athletes is approximately \$273,860 per year. The total areas under the marginal value curves are approximately \$254,560,000 and \$1,643,200, respectively. Thus the total value attributed to

[11]This example is purely illustrative but not purely fanciful. In a recent year, the number of K-12 teachers was about 2.4 million and the number of professional athletes was about 12,000. The median salary of teachers was around \$55,000, while that of profession athletes was around \$36,000. Accordingly, we presume that the very higly paid, elite athletes number well below 12,000. For our purposes, we set the number at 3,000 and the number of educators at 2,000,000.

the services of educators is about 154.92 times that attributed to the athletes (to emphasize: this is a *multiple* of 154.92, not 154.92 percent). All values are computed in *Maxima*. The accompanying workbook shows the details.

This example, unlike the previous two, does not offer a breakdown into consumer and producer surplus. This fact reflects the nature of the two labor markets in which these activities occur. Educators are employed by an amalgam of government and private agencies (mostly government). These agencies' employment and wage decisions are likely politically motivated. rather than aimed at maximizing a relatively simple objective function.

The market for athletes is likely even more complex. The value provided by athletes will be divided among the athletes, the owners of franchises, and spectators. Neither the competitive model of price takers nor the monopolistic model of price searchers applies well. Superstar models and tournament models both predict very high earnings for a few participants.[12]

8.5.2 The Normal Distributions

One of the cornerstones of modern statistics is the *normal distribution*. An astonishingly broad range of physical and human phenomena can be usefully represented by a normal distribution. A random variable x is said to have a normal distribution if its density function is given by the equation

$$N(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right),$$

where $\exp(\dots)$ is the same as $e^{(\dots)}$. Here μ is the mean of this population's values and σ is its standard deviation.

[12]The model as stated implicitly assumes that all of these athletes are identical. Both the superstar model and the tournament model indicate that, even with homogeneity, large differences will accrue. If the athletes are not quite identical, both models predict great rewards for relatively small differences in ability. Cyrenne [5] summarizes these models and applies them to salaries of ice hockey players. The model in the current example also ignores earnings differences due different individual bargaining abilities or due to endorsement earnings.

Garicano and Rossi-Hansberg [6] is a much more ambitious application and extension of the superstar model.

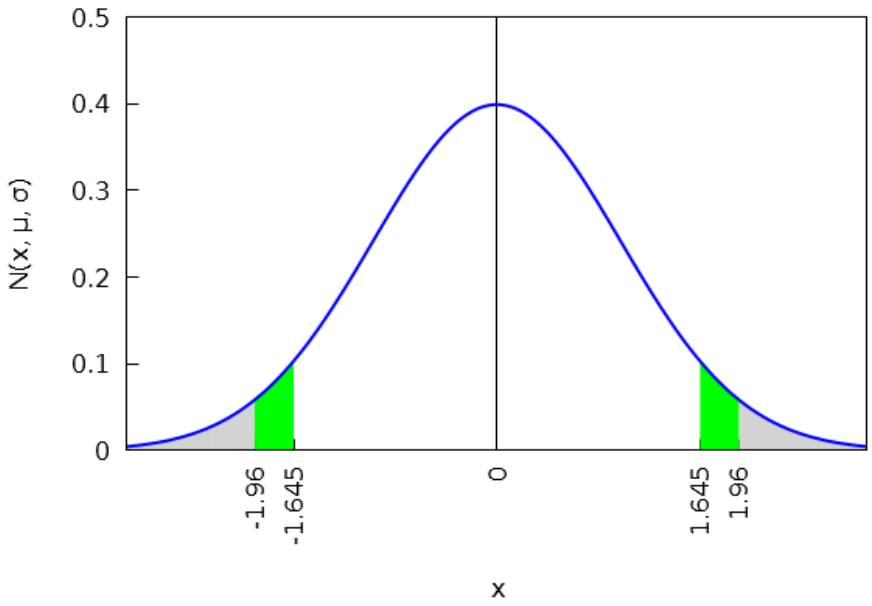


Figure 8.17: Standard normal distribution

The characteristic bell-shaped curve that represents the normal distribution is illustrated in Figure 8.17. The area under the normal curve is given by the improper integral $\int_{-\infty}^{\infty} N(x)\ dx$ and is equal to 1.

Figure 8.17 shows a special case of the normal distribution, one in which $\mu = 0$ and $\sigma = 1$. The area to the left of $x = -1.96$ is the fraction of the population with values less than this value; that area is approximately 0.025. The area above $x = 1.96$ is the same. Therefore, about 95 percent of the population has values between -1.96 and 1.96. Likewise, about 90 percent of the population has values between -1.645 and 1.645. These values can be obtained from a table of value.

Once we have entered the formula into *Maxima*, however, we can confirm that these values represent integrals. Consider the following commands and the resulting output:
`float(integrate(N(x,0,1), x,-1.645,1.645 ))` yields 0.90003, and
`float(integrate(N(x,0,1), x,-1.96,1.96 ))` yields 0.95.

We can confirm that the area under the curve equals 1. We do so with two commands, one evaluating the area to the left of 0 and the other evaluating

the area to the right: `float(integrate(N(x,0,1), x,minf,0 ))` yields 0.5, as does `float(integrate(N(x,0,1), x,0,inf ))`. We could have directly entered $-\infty$ (`minf`) and ∞ (`inf`) into the command, which would result in 1.0.

The commands above relate to the standard normal distribution. Once we have generated our function in *Maxima*, we can analyze a normal distribution without standardizing the units. Suppose, for example, that $\mu = 5$ and $\sigma = 0$, and that we wish to know the fraction of the population that has values less than 1. The command `float(integrate(N(x,5,3), x,minf,1 ))` provides the result, 0.091211, or 9.1211 percent.[13]

8.5.3 Capital Accumulation

Capital accumulation is the process of adding to a given stock of capital by the process known as *investment*. The capital stock in time t is designated by $K(t)$. The rate at which the stock of capital is being depleted or increased over time is given by the derivative, dK/dt. This variable is *net investment*, which is the amount of capital added to the stock *less* depreciation. Thus, $dK/dt = I(t) - D(t)$, so that the capital stock is $\int (I(t) - D(t))\ dt$, where conceptually, integration occurs from the beginning of time.

Over a finite period, $t = a$ to $t = b$ the relevant expression is

$$\int_a^b (I(t) - D(t))\ dt = K(t)\Big|_a^b = K(a) - K(b).$$

As an example, suppose that $K(0) = 100$, and $I(t) - D(t) = a \cdot e^{g \cdot t}$, where $t = 0$ is the initial period. The first panel shows the investment level for each time period, and the second panel shows the capital stock for each time period. Let $t1 = 20$, where $t1$ replaces b in the general expression for the capital stock.

The equation for the capital stock can be written as the command `K(K0, a,g,t1):=''(K0 + integrate(NetInv(t,a,g),t,0,t1))`, which yields this

[13]Creating a function as we have done is not necessary. *Maxima*'s `distrib` module provides these values and more. Furthermore, it does so for 25 continuous and discrete distributions, not just normal distributions.

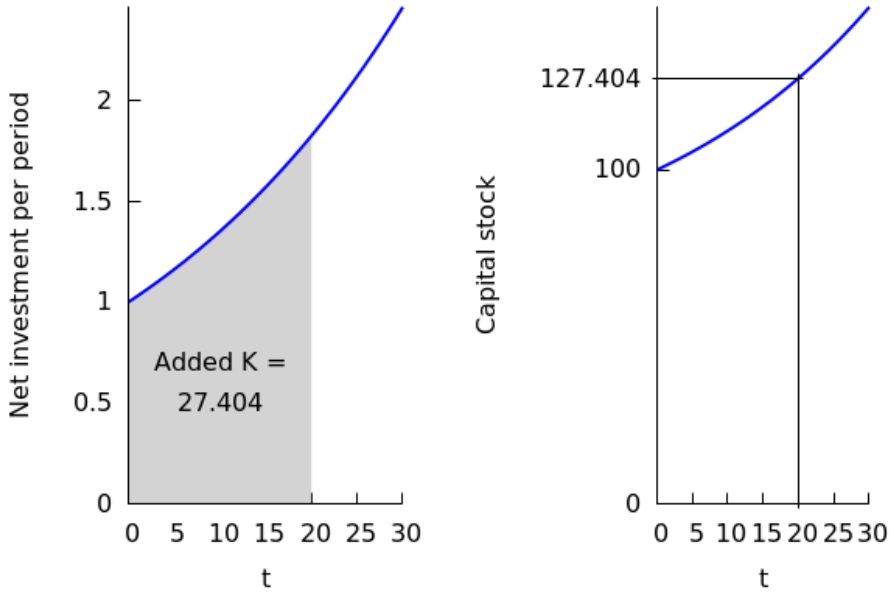


Figure 8.18: Net investment and capital stock growth

output:

$$\mathrm{K}\,(K0, a, g, t1) := a\left(\frac{\%e^{g\,t1}}{g} - \frac{1}{g}\right) + K0.$$

Recall that `%e` is *Maxima*'s notation for the constant $e(= 2.718\ldots)$. We use $K0 = 100$, $a = 1$, and $g = 0.03$. For $t1 = 20$, the value of the capital stock is approximately 127.404, so the stock has grown by 27.404 units during this twenty year period.

Observe how the value of ΔK appears in the two panels. In the flow panel on the left, it is an area: changes per year summed over the 20 years. In the stock panel on the right, it appears as a vertical distance.

8.5.4 The Solow Growth Model[14]

The preceding example shows that investment adds to a capital stock. This relationship holds for any firm (or even household), and it holds for the

[14]This material relates to the preceding application more than to the content of this chapter. It may be omitted without loss of continuity.

aggregate economy. This section sketches the Solow growth model, which illustrates how production, consumption, saving, and investment interact to determine an economy's capital stock and per-capita income. The development used here follows Mankiw [11].

The model begins with production. The model assumes that production is a function of two inputs, capital (K) and labor (L). Also, the production functions exhibits constant returns to scale, or in terms that Chapter 6 develops, it is homogeneous of degree 1. Formally, $Y = f(K, L)$ where Y is total output. Because $f(K, L)$ is homogeneous of degree 1, multiplying all inputs by the same value multiplies output by the same value. We multiply, both K and L by $1/L$, which causes Y to be multiplied by $1/L$. Hence, $Y/L = f(K/L, L/L)$ or $y = f(k, 1)$ where lower-case letters denote per-labor-unit values. (From now on, we refer to these as "per capita," which is exactly appropriate if L is proportional to population) The constant in $f(\ldots)$ is of no consequence, so we rewrite the per-capita production function as $y = f(k)$.

For illustration, we use the simple Cobb-Douglass production function $Y = A \cdot \sqrt{K \cdot L}$, which converts to $y = A \cdot \sqrt{k}$, where A is a technology index. Solow [18] does not use a specific functional specification.

The production function generates a marginal product of capital function: $mpk = dy/dk$. In the illustrative example $mpk = A/(2 \cdot \sqrt{k})$. We use this function below to determine income shares.

We limit our attention to a relatively simple model. This is a model of a closed economy without government. Thus total output is either consumed or saved: $Y = C + S$.[15] We restate these values in per-unit-of-labor terms, $y = c + s$, in order to relate them directly to the production relationships stated above.

Another assumption is that not only is the economy closed in terms of trade (no net exports), but it is also closed in terms of capital flows. This assumption implies that, in equilibrium, $y - c = s = i$ where y is per-capita output and c is per-capita consumption, so $s = y - c$ is per-capita saving. Saving is the only source of funds for investment, so per-capita saving equals per-capita investment: $s = i$.

[15] This is not as severe an assumption as it might appear. If part of the output is diverted to government, then government must spend on either consumption or investment.

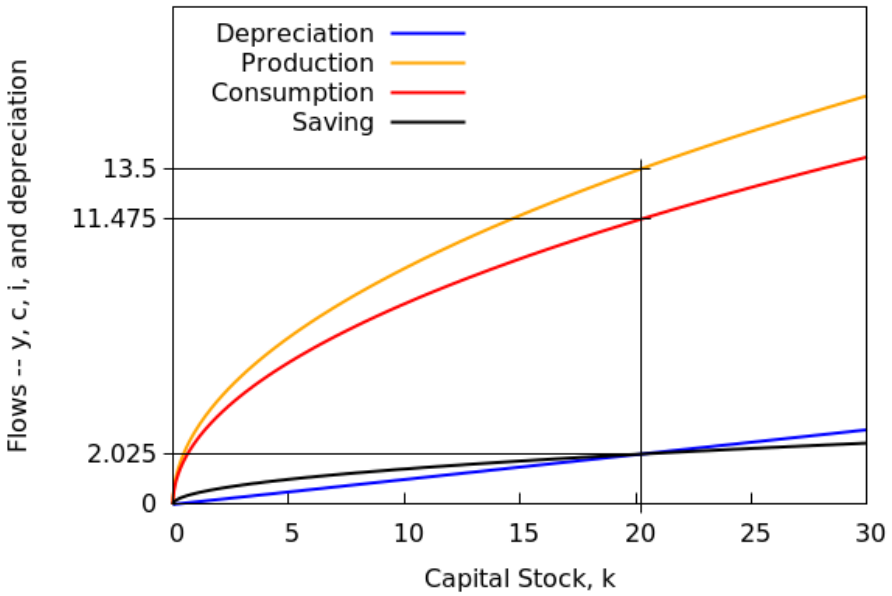


Figure 8.19: Solow growth model equilibrium

We can now define an investment function, which we name chg_k: $chg\_k = s \cdot f(k, A) - d \cdot k$, where d is the depreciation rate. (Investment is the change in the capital stock, k.) For the system to have attained equilibrium, the per-capita stock of capital must not be changing. This occurs only when $chg\_k = 0$ Figure 8.19 shows the nature of equilibrium for these parameter values: $A = 3$, $s = 0.15$, and $d = 0.1$. This economy saves 30 percent of total output, and therefore invests this amount. Each year ten percent of the capital stock must be replaced. The result is a capital stock of $k = 20.25$ For larger values, depreciation exceeds investment, and for smaller values investment exceeds depreciation. Per-capita output is 13.75 units per year, and per-capita consumption is 11.475 units per year.

We can determine that one-half of output accrues to labor and one-half to capital, due to the nature of the Cobb-Douglas function. Thus capitalists receive $13.5/2 = 6.75$ units of output per unit of labor. Of this, they must maintain the capital stock, so net-of-depreciation income is 4.725 times the number of units of labor. Each unit of labor receives 6.75 units of output per

year.[16]

8.5.5 Present Value

The *present value* (often called the *discounted value*) of a single payment A_t to be received t periods in the function is $A_t/(1+i/n)^{n \cdot t}$, where i is the discount rate (often an interest rate at which the entity that is to received the payment can borrow or lend), and n is the number of times that the discounting is compounded each period. In the limit as $n \to \infty$, this value, V, approaches $A_t \cdot e^{-i \cdot t}$.

This expression holds for any value of A at any number of periods t. Therefore, the present value of a discrete number of such would be the sum of the present values of the individual values. If, for example, A_1, A_2, and A_3 are to be received at the end one year, two years, and three years respectively, and compounding is instantaneous, then the present value of this flow is $V = A_1 \cdot e^{-i} + A_1 \cdot e^{-2 \cdot i} + A_1 \cdot e^{-3 \cdot i} = \sum_{t=1}^{3} A_t \cdot e^{-i \cdot t}$.

Change the nature of the flow by allowing the flow to be continuous at a rate A_t per year (or whatever period is specified). That is payment begins immediately and is continuously made throughout the year. The summation expression in the preceding paragraph now gives way to this integral: $V = \int_0^3 A_t \cdot e^{-i \cdot t} dt$. Keep in mind that each individual payment occurs in infinitesimally short period, with the rate being such that the sum of the payments in year t is A_t.

The example below compares two scenarios. In the first, lump sum of 1000, 3000, and 2500 are received at the ends of the first, second, and third years. In the second, the annual amounts are the same, but they are spread continuously throughout the year. The resulting present values are higher in the second case, because the payments are received earlier.

These *Maxima* commands apply in the first case. `[A1,A2,A3] :[1000, 3000, 2500], [V1, V2, V3 ]: [A1*exp(-0.1*1), A2* exp(-0.1*2), A3* exp(-0.1*3)], and V: V1+V2+V3;`. The resulting value is 5213.1, approximately.

[16] Alternatively, we can determine that the marginal product of capital, 1/3, is the payment to each unit of capital and that $(1/3) \cdot 20.25 = 6.75$ per unit of labor.

Now, spread the payments evenly over each of the three years. The resulting present value calculation is executed with these commands (`VC` for "value, continuous"): `[VC1: float( integrate( A1*exp(-0.1*t),t,0,1)), VC2: float( integrate( A2*exp(-0.1*t),t,1,2)), VC3: float( integrate( A3*exp(-0.1*t),t,1,2))] and VC: VC1 + VC2 + VC3`. The resulting value is 5687.5, approximately. As predicted, the fact that the payments are made throughout each year rather than at the ends increases their present value somewhat.

We can generalize this expression to allow for τ (the Greek letter tau) periods. Now $V = \int_0^\tau R(t) \cdot e^{-d \cdot t}$. Here $R(t)$ is a function of t, so the integral of this expression will depend on $R(t)$'s form. The discount rate is d. Consider a simple case in which R is the same each period. Now

$$V = \int_0^\tau D \cdot e^{-d \cdot t} = \frac{R}{r} \cdot (1 - e^{-d \cdot \tau}).$$

Suppose that $\tau = 2, R = 3000,$ and $d = 0.06$. Then the present value of this stream is $(3000/0.06) \cdot (1 - e^{0.12}) \approx 50000 \cdot (1 - 0.8869) \approx 5565$ dollars.

Figure 8.20 generalizes this calculation by letting τ range from 0 to 50 periods. Also, it shows the effect of increasing the discount rate from 0.06 to 0.08. The flattening of the two present value functions as the payment period is lengthened reflects the effect of compounding: the discounting process has an increasingly large impact as the successive payments move farther into the future. Likewise, the size of the discount factor becomes increasingly important as the length of the payment period increases.

Consider an application, due to Chiang [4]. A wine dealer holds a quantity of wine. This wine can be sold immediately for \$A, but holding it and selling it later will result in a higher price. Suppose that the wine's value increasing according to this function: $P = A \cdot e^{\sqrt{(t)}}$.[17]

The present value of the wine, sold in period t is $V(t) = P \cdot e^{-d \cdot t} = A \cdot e^{\sqrt{(t)}} \cdot e^{-d \cdot t} = A \cdot e^{\sqrt{t} - d \cdot t}$. For now, we assume that storage cost is zero, so the profit-maximizing dealer must simply choose the value of t that yields maximum present value. We can convert the expression to a linear-in-logarithms expression, $log(V(t) = log(A) + t^{1/2} - d \cdot t$.

[17]The precise functional relationship between price and age is not important. This one is used for convenience. The reasoning in this example extends to examples like fisheries and forests, in which the growth is a physical growth function rather that a price function. See McAfee [10], Chapter 4.

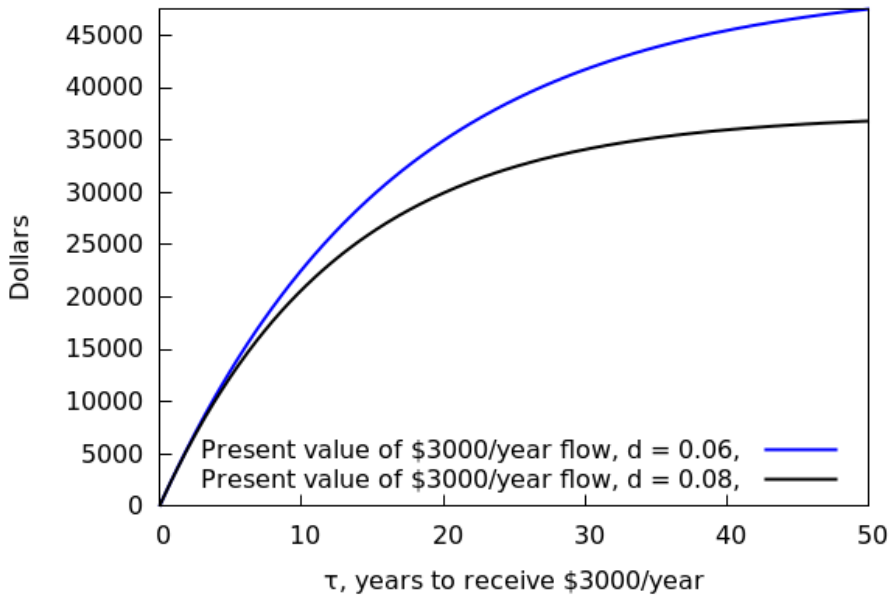


Figure 8.20: Present value, time, and the discount rate

The logarithmic rule for differentiation implies that

$$\frac{1}{V} \cdot \frac{dV}{dt} = \frac{1}{2} \cdot t^{-1/2} - r.$$

Because A is finite, this condition implies that $\sqrt{t} = \frac{1}{2 \cdot r}$. Squaring both sides of this equation shows that the optimal length of storage is $\frac{1}{4 \cdot r^2}$. A higher interest rate implies a shorter optimal storage period. This makes sense, for the interest rate is either what the dealer must pay to underwrite holding this wine or the rate that the dealer must forgo on alternative uses of funds, while the growing price is the return to this particular investment.

Add to this model the fact that storage costs are not zero. Suppose that each case of wine incurs a storage cost at a rate of s dollars per year. The present value of that stream of cost is $\int_0^t e^{-d \cdot t}\ dt = (s/r) \cdot (1 - e^{-d \cdot t})$. The present value of that stream of costs must be subtracted from the present value of the sale in time t (a single value, not a stream: a case can be sold only once).

The present value of the sale at time t, $N(t)$ is, therefore, $N(t) = A(t) \cdot e^{-d \cdot t} - (c/r) \cdot (1 - e^{-d \cdot t}) = (A(t) + (s/r) \cdot e^{-d \cdot t} - s/r$. We apply the product rule to

determine that

$$\frac{dN(t)}{dt} = \frac{dA(t)}{dt} \cdot e^{-d \cdot t} - d \cdot \left(A(t) + \frac{s}{d} \right) \cdot e^{-d \cdot t}.$$

This expression equals zero only if $\frac{d(A(t)}{dt} = d \cdot A(t) + s$. The term on the left-hand side is the gain in value from holding the asset one more period, *i. e.*, the marginal benefit from postponing sale. The term on the right-hand side is the marginal cost of postponing sale. This cost consists of increased discount due to the postponement plus the per-period storage cost. Thus, the familiar marginal cost=marginal benefit criterion for optimization must be satisfied.

8.6 Questions and Problems

1. An economy currently uses 500 million barrels of petroleum per year. With current technology, the use of petroleum is expected to grow at a rate of 8 percent per year, so that the growth pattern can be phrased as $P(t) = 500 \cdot e^{0.08 \cdot t}$. A promising computerized injection system promises to reduce the annual growth rate to 6 percent per year.

   a. State the new growth path mathematically.

   b. Plot the two growth paths using *Maxima*.

   c. Determine the amount of petroleum that the new technology promises to save over the next ten years.

2. The nonlinear inverse demand and supply curves for a product are $pd = \sqrt{225 - 5 \cdot x}$ and $ps = \sqrt{36 + 1.8 \cdot x}$. Use *Maxima*'s `find_root` command to confirm that the equilibrium quantity is approximately 27.794 units per time period and that the equilibrium price is approximately \$9.2752 per unit. Then determine the amount of surplus that accrues to consumers and to producers.

3. Megacorp is the only seller for a product, the inverse demand for which is $pd = 144 - x^2$. Megacorp's marginal cost is $mc = 48 + x^2/2$.

   a. Confirm that Megacorp will maximize its profits if it sells approximately 5.2372 units per time period for a price of approximately \$116.57.

b. Assuming that Megacorps does maximize its profits, compute the values for Consumer Surplus and for Producer Surplus

c. Determine that the efficient quantity, at which $p = mc$, is 8 units Also, integrate $pd - mc$ over the range 52372 to 8 to determine the deadweight loss due to Megacorp's producing its chosen quantity rather than $x = 8$.

4. The maker of a generic household appliance has this marginal cost function for the appliance: $mc = 0.00003 \cdot x^2 - 0.03 \cdot x + 20$, where x is the number of units produced each period. Furthermore, it incurs a fixed cost of \$15,000 per production period. It can sell each unit for \$30 (it cannot affect this price by changing x).

   a. Determine the total cost function per period.

   b. Determine the quantity that maximizes this firm's per-period profit.

   c. Given the answer to (b), find the area between the price and the marginal cost function. Use the available information to calculate the firm's per-period profit.

   d. Calculate total revenue. Use the total cost function from (a) and total revenue to confirm the solution in (c).

5. Currently 100,000 cars per hour use a stretch of highway at rush hours. Over the next few years, this value will grow, following this growth function: $g(t) = \frac{10000}{\sqrt{0.4t}}$. To what value will the number have grown in 3 years?

6. Consider this marginal revenue functions that apply over the relevant ranges for product y: MRy $= 10/(1 + y)^2$.

   a. Integrate MRy to determine the total revenue function. Use economic analysis to calculate the constant of integration.

   b. Determine the average revenue (inverse demand) function. Graph demand and marginal revenue over the range $y = 2$ to $y = 10$.

Chapter 9

Matrix Algebra

Chapter 1 used Stigler's diet problem to demonstrate the power of mathematics in formulating and solving an important problem. The crux of the diet problem is to find the least expensive combination of 80 foods available to a consumer that will satisfy nine recommended daily dietary allowances, as established by the Food and Nutrition Board of the National Academy of Sciences.

Formally, the problem is expressed as follows: Minimize $C = P_1 \cdot X_1 + C_2 \cdot X_2 + ... \cdots + P_{80} \cdot X_{80}$, where C is the cost of a diet that consists of 80 possible food items. This minimization problem is subject to a set of nine constraints, each of which corresponds to a minimum recommended dietary requirement i. The P's are prices of the goods, and the X's are the quantities.

The dietary restrictions are stated as a set of linear equations:

$$\begin{array}{llll} a_{11} \cdot X_1 + & a_{12} \cdot X_2 + & \cdots & a_{1,80} \cdot X_{80} = & R_1 \\ a_{21} \cdot X_1 + & a_{22} \cdot X_2 + & \cdots & a_{2,80} \cdot X_{80} = & R_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{91} \cdot X_1 + & a_{92} \cdot X_2 + & \cdots & a_{9,80} \cdot X_{80} = & R_9 \end{array}$$

Each term a_{ij} relates the number of units of requirement i that are provided by one unit of food j.

We have a system that consists of of 10 equations, the objective function that we are trying to minimize plus the nine constraint equations. The system

contains 80 unknown values, the quantities of the 80 different foods that can be consumed. This systems is an example of applied *linear algebra*.

Definition: *Linear algebra* is the study of systems of linear equations and the attempt to find a simultaneous solution for the unknowns of those equations, if such a solution exists.

It is important to note that linear algebra deals with linear equations. Linear equations are generally easier to deal with than are nonlinear equations. Nonlinear equations and nonlinear models often cannot be solved without the help of a computer.[1] It is also true, however, that we can usefully approximate many business and economics relationships with linear functional forms. Hence we are not severely disadvantaged by the fact that matrix algebra is restricted to the study, manipulation, and solution of linear equations.

9.1 Matrices and Vectors: Definitions

Begin with a very general case involving a system of m linear equations in n variables:

$$\begin{array}{cccc} a_{11} \cdot x_1 + & a_{12} \cdot x_2 + & \cdots & a_{1n} \cdot x_n = & c_1 \\ a_{21} \cdot x_1 + & a_{22} \cdot x_2 + & \cdots & a_{2n} \cdot x_n = & c_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} \cdot x_1 + & a_{m2} \cdot x_2 + & \cdots & a_{mn} \cdot x_n = & c_m \end{array}$$

The n variables, x_1 x_2, ..., x_n above are specifically aligned in a particular fashion. The variable denoted x_1 is the first variable and must appear in the first column. The variable x_2 is the second variable and must appear in the second column. Any variable x_j must appear in the j^{th} column.

Similarly, the subscript a_{ij} is definitive with respect to location in the system of equations. For example, a_{24} represents the coefficient of the variable that appears in the second row and the fourth column of the system of equations. In general, a_{ij} refers to the coefficient of the variable located in the i^{th} row and the j^{th} column.

[1] Computer algebra systems like *Maxima* provide tools for solving systems of nonlinear equations. Also, they provide tools for evaluating the behavior via simulations of systems that cannot be solved analytically.

Finally, the parameters c_1, c_2, ..., c_m are m in number and are unattached to any of the variables x_j. The constant c_3 belongs in the third row such that

$$\sum_{j=1}^{n} a_{3j} \cdot x_j = c_3.$$

More generally, for any particular row, i,

$$\sum_{j=1}^{n} a_{ij} \cdot x_j = c_i.$$

Definition. A *matrix* is a rectangular, ordered array of elements or entries consisting of numbers, parameters, or variables.

The existence of a matrix is usually signaled by the use of brackets [ ] or parentheses ( ). We use the bracket notation [ ] in this text.

Even though it is written in standard notation, the system of equations as expressed above can become cumbersome and unwieldy. Fortunately, this system of equations can be simply rewritten using a shorthand notation. Let $A \cdot X = C$ represent the system of equations that is displayed above, with

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and } C = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}$$

A, X, and C are matrices, and AX = C.[2] The elements in a particular matrix are not separated by commas, but by blank spaces. It is customary to symbolize a matrix by an upper-case (capital) letter, such as A, X, or C, whereas the elements in a particular matrix are customarily denoted by lower-case (small) letters, such as a, a_{ij}, and b. It is possible for the elements of the matrix to be numeric values, such as 5, 7, or 11.24. In this case, the numeric values are used in preference to lower-case letters.

[2]The same relationship could be written as $X \cdot A = C$. Doing so, however, would require rephrasing X and C as follows: $X = [x_1, x_2, \cdots, x_n]$ and $C = [c_1, c_2, \cdots, c_n]$.

The matrix labeled A above represents the coefficients of the variables in the system of equations. The A matrix has m rows and n columns. This can be contrasted with the variable matrix, labeled X, which consists of n rows and only one column. In general, there is no relationship between the number of rows and the number of columns in a matrix. The number of rows is not necessarily related to the number of columns, and *vice versa*. What is the case, however, is that the number of rows and the number of columns define the dimension (or order) of a matrix. For example, matrix A has m rows and n columns and is therefore said to be an $m \times n$ matrix (which is read, "m by n matrix "). The dimension of a matrix is always read rows first, columns second. A 5×7 matrix has five rows and seven columns, not *vice versa*. In an important special case in which $m = n$, for example, a 5 X 5 matrix, one is dealing with a *square matrix*.

We occasionally encounter the notation $A = [a_{ij}]$, which represents a matrix composed of the elements that take the form a_{ij}. The number of rows and columns is unspecified. Note well that $[a_{ij}]$, which represents a matrix, is not equivalent to a_{ij}, which represents a specific element in a matrix. That is, $[a_{ij}] \neq a_{ij}$ unless the dimensions of the matrix are 1×1.

A small comment on notation is in order. Either a_{ij} or $a_{i,j}$ may be used to indicate the value in the matrix element in the row i and column j. If we instruct Maxima to create a matrix of a's, using the command `genmatrix(a, 3, 3)`, the resulting matrix is

$$\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix}.$$

This differs from

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

only in terms of notation. The two are equivalent ways of saying the same thing. We use the latter, more compact, notation when doing so does not introduce a possibility of ambiguity.

It is instructive to rewrite Stigler's diet problem in matrix notation. We can denote the objective function that we seek to minimize by $P \cdot X = C$, and we can represent the constraint equations by $A \cdot X = R$, where

$$P = \begin{bmatrix} P_1 & P_2 & \cdots & P_{80} \end{bmatrix} \text{ is a } 1 \times 80 \text{ matrix,}$$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{80} \end{bmatrix} \text{ is an 80 x 1 matrix,}$$

$$\text{C} = [\text{C}] \text{ is a } 1 \times 1 \text{ matrix,}$$

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,80} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,80} \\ \vdots & \vdots & \vdots & \vdots \\ a_{9,1} & a_{9,2} & \cdots & a_{9,80} \end{bmatrix}$$

is a 9×80 matrix, and

$$R = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_9 \end{bmatrix}$$

is a 9×1 matrix.

Matrices X and C have the dimensions 80×1 and 9×1 respectively. Both matrices have only one column and are referred to as *column vectors*. Matrix P has the dimensions 1×80 and is referred to as a *row vector*.

We can use the concept of a vector to view a matrix as a series of related row and/or column vectors. Consider the matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

We can consider this matrix to be an ordered set of m row vectors,[3] namely,

$$A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}$$

It is absolutely essential that an $m \times n$ matrix be read as m rows by n columns. An $m \times n$ matrix is *not* equivalent to an $n \times m$ matrix except in the special circumstances in which $m = n$ and we have a square matrix. For example, we define matrices J and K as follows:

$$J = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \text{ and } K = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

Matrices J and K do not have identical dimensions. J is a 2×3 matrix, whereas K is a 3×2 matrix.

A matrix is an *ordered array of elements*, according to our definition. Each element of the matrix has an assigned location in the matrix. Any alteration of that assigned location will, in general, alter the matrix and the system of equations it represents. Consider the following system of equations.

$$\begin{aligned} 8 \cdot x_1 + 10 \cdot x_2 + 12 \cdot x_3 &= 1 \\ 3 \cdot x_1 + \qquad + 2 \cdot x_3 &= 0 \\ x_1 - 2 \cdot x_2 - 5 \cdot x_3 &= -5 \end{aligned}$$

The coefficient matrix of this system of equations is given by this matrix.

$$A = \begin{bmatrix} 8 & 10 & 12 \\ 3 & 0 & 2 \\ 1 & -2 & -5 \end{bmatrix}$$

The element in the second row and second column ($a_2 2$) of the coefficient matrix is 0 and must be included. Further, should we interchange the elements in the first and second columns of the first row, that is, should we

[3] We could also view the matrix as an ordered set of n column vectors.

interchange a_{11} and a_{12}, then the matrix would become the one below, $A*$.

$$A* = \begin{bmatrix} 10 & 8 & 12 \\ 3 & 0 & 2 \\ 1 & -2 & -5 \end{bmatrix}$$

Matrix $A*$ now represents the coefficients of this system of equations.

$$\begin{aligned} 10 \cdot x_1 + 8 \cdot x_2 + 12 \cdot x_3 &= 1 \\ 3 \cdot x_1 + \qquad + 2 \cdot x_3 &= 0 \\ x_1 - 2 \cdot x_2 - 5 \cdot x_3 &= -5 \end{aligned}$$

Matrices A and $A*$ are not the same; they represent different sets of coefficients.

We must finally observe that a matrix has no numeric value *per se*. One cannot state that a matrix has a value of 5, 7, 14, or any other number. A matrix is simply a shorthand, efficient method of writing an array of elements.

9.2 Matrix Operations

We have already seen that a matrix is a compact and logical way to write an array of elements. We represented a system of m linear equations in n variables as $A \cdot X = C$.

We have yet to indicate how one matrix is related to another. We have not touched on such questions as: Why did we choose to write X and C as column vectors rather than as row vectors? How do we multiply matrices? When are two matrices equal? Do the laws governing the addition and subtraction of real numbers also hold for matrices? This section addresses these and other matters.

More specifically, this section considers the following matrix algebra concepts: (1) equality, (2) addition, (3) subtraction, (4) the commutative and associative laws of addition and subtraction, (5) scalar multiplication, (6) matrix multiplication, and (7) the commutative and associative laws of multiplication.

9.2.1 Matrix Equality

Two matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ are said to be equal, such that $A = B$, if and only if A and B have the same dimensions *and* all corresponding elements in their arrays are identical. That is, $A = B$ if and only if $a_{ij} = b_{ij}$ for all i and j. The four examples that follow illustrate the nature of matrix equality.

1. $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ $A \neq B$ because $a_{21} \neq b_{21}$

2. $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$ $A = B$

3. $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$

   $A \neq B$ because A and B have different dimensions

4. $A = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ and $B = \begin{bmatrix} 1 \\ 0 \\ 8 \end{bmatrix}$

   $A = B$ if and only if $x_1 = 1$, $x_2 = 0$, and $x_3 = 8$

9.2.2 Addition and Subtraction of Matrices

We can add two matrices, $A = [a_{ij}]$ and $B = [b_{ij}]$, if and only if A and B have the same dimensions. $A + B = C$ such that $[a_{ij}] + [b_{ij}] = [c_{ij}]$, where $c_{ij} = a_{ij} + b_{ij}$ for all i and j. Matrix C has the same dimensions as A and B.

We can add two matrices of the same dimension, but we cannot add two matrices of different dimensions. For example, we can add a 2×3 matrix to another 2×3 matrix. We cannot, however, add a 2×3 matrix to a 3×2 matrix. The definition also tells us that addition involves adding corresponding elements in each matrix. That is, given that the two matrices are of the same dimension, we may add them by summing the corresponding elements of each matrix. This means that we add the element that is in the first row, first column of matrix A to the element that is in the first row, first column

of matrix B. The result is the element that appears in the first row, first column of the summed matrix C.

Likewise, we pair and add the elements in the first row, second column of each matrix, and so forth. Formally, we can write this process as follows:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} + \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,n} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,n} \\ \vdots & \vdots & & \vdots \\ b_{m,1} & b_{m,2} & b_{m,3} & b_{m,n} \end{bmatrix} =$$

$$\begin{bmatrix} b_{1,1}+a_{1,1} & b_{1,2}+a_{1,2} & \cdots & b_{1,n}+a_{1,n} \\ b_{2,1}+a_{2,1} & b_{2,2}+a_{2,2} & \cdots & b_{2,n}+a_{2,n} \\ \vdots & \vdots & & \vdots \\ b_{m,1}+a_{m,1} & b_{m,2}+a_{m,2} & \cdots & b_{m,n}+a_{m,n} \end{bmatrix}$$

To subtract matrix B from matrix A, replace the $+$ with $-$ in each of the cells.

The following six examples illustrate the processes of addition and subtraction of matrices.

1. $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, and $B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$, so $A + B = \begin{bmatrix} 6 & 8 \\ 10 & 12 \end{bmatrix}$

2. $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, and $B = \begin{bmatrix} 4 & 5 & 6 \\ 1 & 2 & 3 \end{bmatrix}$, so $A + B = \begin{bmatrix} 5 & 7 & 9 \\ 5 & 7 & 9 \end{bmatrix}$

3. $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, and $B = 6 \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$, so $A + B$ does not exist

4. $A = \begin{bmatrix} 4 & 8 \\ 10 & 12 \end{bmatrix}$, and $B = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, so $A - B = \begin{bmatrix} 3 & 6 \\ 7 & 8 \end{bmatrix}$

5. $A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$, and $B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, so $A - B = \begin{bmatrix} 0 & -2 & -2 \\ -4 & -4 & -6 \end{bmatrix}$

6. $A = \begin{bmatrix} 2 & 5 \\ 4 & 6 \end{bmatrix}$, and $B = \begin{bmatrix} 2 & 1 \\ 4 & 6 \\ 3 & 5 \end{bmatrix}$, so $A - B$ does not exist

Adding and subtracting matrices is governed by the *Commutative and Associative Laws for Matrix Addition and Subtraction*.

The Commutative Law. The order in which we add or subtract matrices is irrelevant. The commutative law demonstrates this fact. Observe that we treat subtraction as the addition of a negative number, so the proof for addition applies directly to subtraction.
Proof: $A + B = [a_{ij} + b_{ij}] = [a_{ij} + b_{ij}] = [b_{ij} + a_{ij}] = [b_{ij}] + [a_{ij}] = B + A$

The Associative Law. The associative law deals with situations in which three or more matrices are being added. We can apply the associative law to subtraction by considering subtraction to be the addition of a negative number. The law can be stated in terms of three matrices:
Given matrices A, B, and C, $A + (B + C) = (A + B) + C = A + B + C$.
Once the resulting matrix is defined, this process can be extended to more matrices, so the law applies to the addition or subtraction of any finite number of matrices.

Proof: $A + (B + C) = [a_{ij}] + [b_{ij} + c_{ij}] = [a_{ij} + b_{ij}] + [c_{ij}] \ = (A + B) + C = [a_{ij} + b_{ij} + c_{ij}] \ = A + B + C$

An example:
$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \quad C = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$
$A + (B + C) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = A + B + C$
Confirm that $(A + B) + C = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.

Scalar Multiplication. The process of multiplying a matrix by a number (called a *scalar* in matrix because it changes the scale of the matrix by changing the size of all elements proportionately) is referred to as scalar multiplication.

A formal definition of scalar multiplication is this: Given a matrix $A = [a_{ij}]$ and a scalar k, the scalar multiplication of k and A, written $k \cdot A$, is defined to be

$$k \cdot a = [k \cdot a_{ij}] = \begin{bmatrix} k \cdot a_{11} & k \cdot a_{12} & \cdots & k \cdot a_{1n} \\ k \cdot a_{21} & k \cdot a_{22} & \cdots & k \cdot a_{2n} \\ \vdots & \vdots & & \vdots \\ k \cdot a_{m1} & k \cdot a_{m1} & \cdots & k \cdot a_{mn} \end{bmatrix}.$$

Consider two examples:

1. Let $k = 5$ and $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$. Then $k \cdot A = 5 \cdot A = \begin{bmatrix} 5 & 10 \\ 15 & 20 \end{bmatrix}$.
2. Let $k = -3$ and $B = \begin{bmatrix} 1 & -2 & 3 \\ 4 & 5 & -6 \end{bmatrix}$. Then $k \cdot B = \begin{bmatrix} -3 & 6 & -9 \\ -12 & -15 & 18 \end{bmatrix}$.

Matrix Multiplication Like matrix addition and subtraction, matrix multiplication also has attendant dimensional requirements. However, these requirements differ from those imposed in matrix addition and subtraction. Two matrices A and B can be multiplied together to form the product $A \cdot B$ if and only if the *column* dimension of A is equal to the *row* dimension of B.[4]

Thus, we may multiply an $m \times n$ matrix by an $n \times p$ matrix. An $m \times n$ matrix A is said to be *postmultiplied* by an $n \times p$ matrix B in order to form a new $n \times p$ matrix C. Or we could equivalently state that $n \times p$ matrix B is premultiplied by $m \times n$ matrix A, once again yielding an $m \times p$ matrix C.

We can formally define the process of matrix multiplication as follows. Given: $A = [a_{ik}]$, an $m \times n$ matrix, and $B = [b_{kj}]$, an $n \times p$ matrix, where a_{ik}, is any element of A and b_{kj} is any element of B. Then: $A \cdot B = C$, an $m \times p$ matrix whose elements are

$$c_{ij} = \sum_{k=1}^{n} a_{ik} \cdot b_{kj}$$

for all $i = 1, 2 \ldots, m$ and $j = 1, 2 \ldots, p$.

Consider the matrices A and B, which the Maxima commands `A: genmatrix(a, 2, 2); B: genmatrix(b,2,1);` created:

$$\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \quad \begin{bmatrix} b_{1,1} \\ b_{2,1} \end{bmatrix}.$$

The product $A \cdot B$, generated by the command `A.B;`, is matrix C, a 2×1 matrix that consists of two sums:

$$\begin{bmatrix} a_{1,2}\, b_{2,1} + a_{1,1}\, b_{1,1} \\ b_{2,1}\, a_{2,2} + b_{1,1}\, a_{2,1} \end{bmatrix}.$$

[4]Note the use of the dot. Matrix multiplication of the sort discussed here is often called "dot multiplication" and the result is the "dot product."

The command `B.A;` instructs *Maxima* to premultiply a 2×2 matrix by a 2×1 matrix, which it cannot do. Accordingly, *Maxima*'s response to this command is

MULTIPLY MATRICES: attempt to multiply nonconformable matrices.

– an error. To debug this try: debugmode(true);.

You must clearly understand the following:

- The product of matrices A and B, $A \cdot B$ is read, "B is premultiplied by A" or "A is postmultiplied by B."
- In order to form the matrix product $A \cdot B$, the column dimension of A must be equal to the row dimension of B.
- If $A \cdot B$ is defined, then the result is a new matrix C that exhibits the row dimension of A and column dimension of B.
- That the product $A \cdot B$ is defined does not imply that the product $B \cdot A$ must also be defined.

Consider three examples.

1. If A is a 2×3 matrix while B is a 3×3 matrix, the $A \cdot B$ is a 2×3 matrix and $B \cdot A$ is not defined.
2. If A is a 1×3 matrix while B is a 3×1 matrix, the $A \cdot B$ is a 1×1 matrix and $B \cdot A$ is a 3×3 matrix. Even though both $A \cdot B$ and $B \cdot A$ are defined, they do not have the same dimensions.
3. Matrices A, B, and $A \cdot B$ are generated by these *Maxima* commands:
   `[A: matrix( [2,5], [7,1], [8,3] ),`
   `B:matrix([4,6,7], [9,10,11]), A.B]`. The matrices are these:

$$\begin{bmatrix} 2 & 5 \\ 7 & 1 \\ 8 & 3 \end{bmatrix} \cdot \begin{bmatrix} 4 & 6 & 7 \\ 9 & 10 & 11 \end{bmatrix} = \begin{bmatrix} 53 & 62 & 69 \\ 37 & 52 & 60 \\ 59 & 78 & 89 \end{bmatrix}.$$

   Name the third, matrix C. Then we can note that $c_{32} = \sum_{k=1}^{2} a_{3k} \cdot b_{k2} = 8 \cdot 4 + 3 \cdot 9 = 59$. Choose 2 or 3 other values in C and confirm that they are generated in like fashion.

The preceding examples indicate the dimensions of the matrix that results from matrix multiplication. Our approach to matrix multiplication has so far been mechanical. We shall now develop an intuitive understanding of matrix multiplication as well. In particular, we return to linear equations systems, with which this chapter began. We do so in order to address the logic of *matrix algebra*, or *linear algebra*.

Recall the general system

$$\begin{array}{cccc} a_{11} \cdot x_1 + & a_{12} \cdot x_2 + & \cdots & a_{1n} \cdot x_n = & c_1 \\ a_{21} \cdot x_1 + & a_{22} \cdot x_2 + & \cdots & a_{2n} \cdot x_n = & c_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} \cdot x_1 + & a_{m2} \cdot x_2 + & \cdots & a_{mn} \cdot x_n = & c_m \end{array}$$

We previously learned how to abbreviate this system of linear equations $A \cdot X = C$. Our definition of matrix equality enables us to state that matrix C equals the product $A \cdot X$ if and only if element c_i is given by $c_i = \sum_{k=1}^{n} a_{ik} \cdot x_k$ for $i = 1, 2, \ldots, m$.

The subscripts of the terms in this system lead intuitively to the definition of matrix multiplication. Specifically, we observe that the subscript k is used in both the a_{ik} and the x_k terms. This ensures that the number of columns in matrix A is the same as the number of rows in matrix X. In more detailed form, the matrix multiplication $A \cdot X = C$ involves the following:

$$\begin{array}{cccc} a_{11} \cdot x_1 + & a_{12} \cdot x_2 + & \ldots & a_{1n} \cdot x_n \\ a_{21} \cdot x_1 + & a_{22} \cdot x_2 + & \ldots & a_{2n} \cdot x_n \\ \vdots & \vdots & & \vdots \\ a_{m1} \cdot x_1 + & a_{m2} \cdot x_2 + & \ldots & a_{mn} \cdot x_n \end{array} = \begin{bmatrix} \sum_{k=1}^{n} a_{1k} \cdot x_k \\ \sum_{k=1}^{n} a_{2k} \cdot x_k \\ \vdots \\ \sum_{k=1}^{n} a_{mk} \cdot x_k \end{bmatrix} = \begin{array}{c} c_1 \\ c_2 \\ \vdots \\ c_m \end{array}$$

We now need to go from the somewhat familiar case above to the general case in which A is once again an $m \times n$ matrix and X is an $n \times p$ matrix. Any element c_{ij} of the new matrix $A \cdot X = C$ is given by $\sum_{k=1}^{n} a_{ik} \cdot x_{jk}$, where $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, p$.

In detail, the product $A \cdot X = C$ is given by the equalities below:

$$[a_{ik}] \cdot [x_{kj}] = \begin{bmatrix} \sum_{k=1}^{n} a_{1k} \cdot x_{k1} & \cdots & \sum_{k=1}^{n} a_{1k} \cdot x_{kp} \\ \sum_{k=1}^{n} a_{2k} \cdot x_{k1} & \cdots & \sum_{k=1}^{n} a_{2k} \cdot x_{kp} \\ \vdots & & \vdots \\ \sum_{k=1}^{n} a_{mk} \cdot x_{k1} & \cdots & \sum_{k=1}^{n} a_{mk} \cdot x_{kp} \end{bmatrix} = [c_{ij}],$$

as $i = 1, 2, \cdots, m$ and $j = 1, 2, \ldots p$.

In this case, the matrix of coefficients is applied to a matrix of veriables. Each column of the latter is the same length, but not (necessarily) the same variables. Suppose, for example, the coefficients relate to demographic data like age, years of schooling, and income. The age and location variables would likely be well-defined, but the analyst might have little reason to choose between two competing measures of income. The first column in the X matrix might have pre-tax income, and the second column might have post-tax income. Multiplying the (fixed-value) coefficient matrix A to the two-column X matrix would result in a two-column C matrix. The values in these columns would differ due to the different income measures.

Before addressing the nature of this issue more formally, we consider a simple hypothetical example. Consider three individuals who are the same age, have the same level of schooling, and have the same pre-tax and post-tax income, but behave somewhat differently. Specifically, suppose that their level of consumption of some product is defined by the following three equations:
`z1: 10 + 5*age - 2*years + 3*income,`
`z2: 12 + 4*age - 1.5*years+2.5*income, and`
`z3: 11 + 4.5*age - 1.8*years + 2.75*income`. Each of these is 30 years of age and has 12 years of schooling. Each also has a post-tax income of 35 and a pre-tax income of 40 (presumably in \$1000's per year). The three matrices below contain this data and the resulting values of z for these people.[5]

$$\begin{bmatrix} 10 & 5 & -2 & 3 \\ 12 & 4 & 0.5 & 2.5 \\ 1 & 4.5 & -1.8 & 2.75 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ 30 & 30 \\ 12 & 12 \\ 35 & 40 \end{bmatrix} = \begin{bmatrix} 241 & 256 \\ 225.5 & 238.0 \\ 210.65 & 224.4 \end{bmatrix}.$$

Note the two 1's in the X matrix. This is the value by which the constant term is multiplied. The A matrix is 3×4, and the X matrix is 4×2, so the C matrix is 3×2. We have two predicted levels of z for each person. The first column uses post-tax income to predict z,a nd the second column uses pre-tax income.

We now develop a more general treatment, one that provides a relatively simple way to remember what is supposed to be multiplied and what is

[5]The "people" are likely to representatives of some groups, perhaps identified by region or ethnicity. Coefficients would likely be estimates that have come from econometric studies.

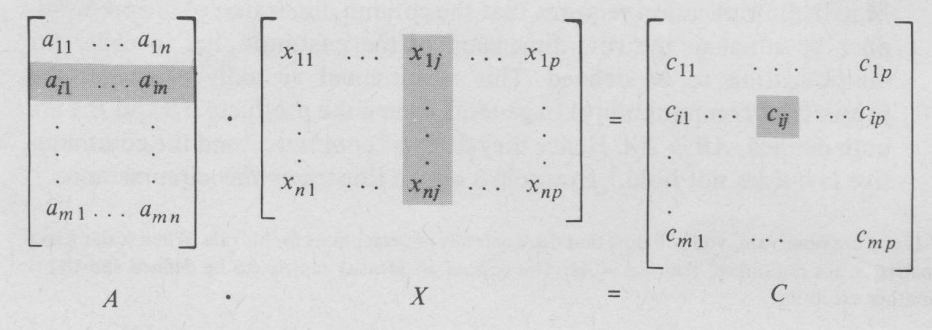


Figure 9.1: Matrix Multiplication

supposed to be added. Note that the number of columns in general matrix $A = [a_{ij}]$ is n, while the number of rows in matrix $X = [x_{kj}]$ is also n. Hence the multiplication is defined. The result of the multiplication $C = [c_{ij}]$ is a matrix of m rows and p columns. Matrix A must have i rows; matrix X must have j columns. The individual elements in matrix C are given by $c_{ij} = \sum_{i=1}^{n} a_{ik} \cdot x_{kj}$, with $i = 1, 2, \ldots m$ and $j = 1, 2, \ldots p$.

Thus, each element c_{ij} depends on the elements in row i of matrix A and the elements in column j of matrix X. Figure 9.1 demonstrates this relationship visually. The i^{th} row of matrix A times the j^{th} column of matrix X results in element c_{ij} (i^{th} row, j^{th} column) in matrix C. The i^{th} row in matrix A, the j^{th} column in matrix X, and element c_{ij} in matrix C are all enclosed in shaded boxes in Figure 9.1.

Consider the following three examples.

1. $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 0 & 2 \\ 5 & 6 & 3 \end{bmatrix}$

   $A \cdot B = \begin{bmatrix} 1 \cdot 1 + 2 \cdot 5 & 1 \cdot 0 + 2 \cdot 6 & 1 \cdot 2 + 2 \cdot 3 \\ 3 \cdot 1 + 4 \cdot 5 & 3 \cdot 0 + 4 \cdot 6 & 3 \cdot 2 + 4 \cdot 3 \end{bmatrix} = \begin{bmatrix} 11 & 12 & 8 \\ 23 & 24 & 18 \end{bmatrix}$

   $B.A$ is not defined.

2. $A = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \quad B = \begin{bmatrix} 3 & 1 \\ 2 & 0 \\ 1 & 2 \end{bmatrix} \quad A \cdot B = \begin{bmatrix} 10 & 7 \end{bmatrix}$

   $B \cdot A$ is not defined.

3. $A = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$ $B = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$ $A \cdot B = \begin{bmatrix} 10 \end{bmatrix}$, a scalar

$$B \cdot A = \begin{bmatrix} 3 & 6 & 9 \\ 2 & 4 & 6 \\ 1 & 2 & 3 \end{bmatrix}$$

9.2.3 Commutative, Associative, and Distributive Laws of Matrix Multiplication

Commutative Law. Matrix multiplication requires that the column dimension of the premultiplier be equal to the row dimension of the postmultiplier in order for multiplication to be defined. This requirement virtually eliminates the property of commutability. In general, even if the products $A \cdot B$ and $B \cdot A$ are both defined, $AB \neq BA$. Hence they do not "commute," and the commutative law does not hold.[6] Example 3 above illustrates this circumstance.

Associative Law. The associative law of matrix multiplication tells us that $A \cdot (B \cdot C) = (A \cdot B) \cdot C = A \cdot B \cdot C$, provided that the dimensional requirements for multiplication are satisfied. Figure 9.2 shows the dimensional requirements that must be satisfied.

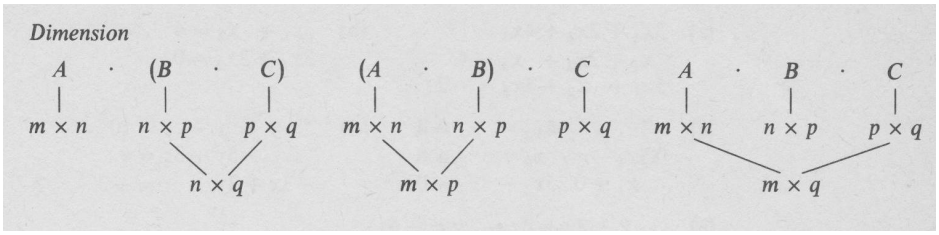


Figure 9.2: Conditions for associative property to hold

Exercise 9.1.

[6] You might have noted an exception to this rule. When scalar k and matrix A are multiplied, then $k \cdot A = A \cdot k$. The case of an identity matrix (to be defined shortly) is another exception. Actually, $k = k \cdot 1$, and 1 is a degenerate identity matrix, so this is actually just one exception. More later.

1. Find the coefficient matrix for each of the follow systems of linear equations.

   a. $3 \cdot x_1 + 2 \cdot x_2 + 4 \cdot x_3 = 17$, $x_1 + 2 \cdot x_2 + x_3 = 4$, and $5 \cdot x_1 + x_2 + 3 \cdot x_3 = -2$

   b. $x_1 + x_2 = 4$ and $3 \cdot x_1 + 2 \cdot x_2 = 0$

   c. $x + 2 \cdot y + 4 \cdot z - 2 = -6$, $-4 \cdot x + 2 \cdot w = 7$, $3 \cdot y + z - 4 \cdot w = 0$, and $-x - y + z = 6$

   d. $x + y - z = 10$, $-5 \cdot y + 3 \cdot z = 4$, and $-3 \cdot x + 2 \cdot y = -3$

2. State the dimensions of the following matrices.

   (a) $A = \begin{bmatrix} 1 & 2 & 1 \\ 3 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ (b)$B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$ (c) $C = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

   (d) $D = \begin{bmatrix} 0 & 19 & 9 \\ 0 & 26 & 12 \\ 0 & 33 & 15 \\ 2 & 7 & 4 \end{bmatrix}$ (e) $E = \begin{bmatrix} 9 & 12 & 15 \\ -9 & 5 & 1 \end{bmatrix}$ (f) $F = \begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix}$

   (g) $G = \begin{bmatrix} -9 & 5 & 1 \end{bmatrix}$ (h) $H = \begin{bmatrix} 1 & 2 & -1 & 2 & 1 \\ 3 & 0 & -6 & 0 & 1 \\ 0 & 0 & 8 & 0 & 1 \\ 2 & -11 & 10 & 15 & 5 \end{bmatrix}$

   For the next three sets of matrices, perform the indicated matrix operations whenever the matrices meet the required dimensional constraints.

3. Given that $A = \begin{bmatrix} -2 \\ 0 \\ 4 \end{bmatrix}$ $B = \begin{bmatrix} 1 & 1 & 2 \end{bmatrix}$ $C = \begin{bmatrix} 2 & 6 & 0 \end{bmatrix}$, find (a), $A \cdot B$ (b) $B \cdot A$, and (c) $A \cdot (B + C) = A \cdot B + A \cdot C$.

4. Given that $A = \begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix}$ $B = \begin{bmatrix} 3 & 5 \\ 2 & 0 \end{bmatrix}$ $C = \begin{bmatrix} 5 & 4 \\ 0 & -4 \end{bmatrix}$ find (a) $A + B$, (b) $A \cdot (B + C)$, (c) $A \cdot B \cdot C$, and (d) $C \cdot B \cdot A$.

5. Given that $A = \begin{bmatrix} 5 & 6 & 1 & 3 \\ 1 & 2 & 0 & -1 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & 1 & 4 \\ 1 & 3 & 1 \\ 1 & 0 & -1 \\ 0 & 2 & -1 \end{bmatrix}$ find $A \cdot B$.

9.3 Special Types of Matrices

Certain types of matrices occur with such frequency that they require additional study. Other matrices have elements or dimensions that sometimes present problems. This section considers the following specific situations: the identity matrix, the diagonal matrix, the scalar matrix, the null matrix, and the transpose of a matrix.

9.3.1 Identity Matrix

In the real number system, the number one (1) has the unique property that for any number a, we have $1 \cdot a = a \cdot 1 = a$, and in particular $1 \cdot 1 = 1$. In matrix algebra, the *identity* or *unit matrix* plays the same role as does the number 1 in ordinary algebra.

The formal defintion of the identiy matrix is this. The identity matrix of order n, denoted by the symbol I or I_n is a square matrix whose main, or principal, diagonal contains no other elements except the number 1. All other elements in the identity matrix are zeros. Thus, I_n is an identity matrix if

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

The subscript n on the identity matrix I_n, indicates the dimension of the matrix. We can see that the main diagonal of the identity matrix consists of the elements beginning with the element in the upper left-hand corner of the matrix and proceeding diagonally to the lower right-hand corner of the matrix. The main diagonal therefore contains the elements $a_{11}, a_{22}, a_{33}, \ldots, a_{nn}$. In the case at hand, every element of the main diagonal is the number 1.

The identity matrix is sometimes written in shorthand as

$$I = [\delta_{ij}], \text{ where } \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

and δ_{ij} is known as *Kronecker's delta*.

Two examples: $I_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

If A is a square matrix of order n, and I_n is the identity matrix, then $I \cdot A = A \cdot 1 = A$. The example below uses a 2×2 matrix A and a 2×2 identity matrix. The first line of command creates the two matrices. The second line contains a command to create a matrix that consists of four matrices and some text material. The resulting matrix is really a table and should not be used for analytical purposes.

```
I: matrix([1,0],[0,1])$ A:matrix([1,2],[3,4])$
matrix(["I","A","A.I","I.A"],[I,A,A.I,I.A]);
```

$$\begin{bmatrix} I & A & A.I & I.A \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} & \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} & \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \end{bmatrix}$$

When A is not a square matrix, say $m \times n$, $I \cdot A = A \cdot I = A$ still holds, but in a particular and limited way. The identity matrix in the term $I \cdot A$ is not the same identity matrix in the term $A \cdot I$. The dimensional requirements for an identity matrix differ, depending on whether we premultiply $(I \cdot A)$ or postmultiply $(A \cdot I)$. In the example below, A is a 3×2 matrix. It can be postmultiplied by a 2×2 identity matrix or premultiplied by a 3×3 matrix. In each case the result is the original matrix.

$$\begin{bmatrix} A & I2 & I3 & A.I2 & I3.A \\ \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 4 & 6 \end{bmatrix} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 4 & 6 \end{bmatrix} & \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 4 & 6 \end{bmatrix} \end{bmatrix}$$

Another circumstance of interest occurs when $A = I$. Then $I \cdot I = I$, which by iteration implies that $I^k = I$ for $k = 1, 2, 3, \ldots, n$. Thus an identity matrix raised to any power is in fact equal to itself. Such a matrix is an *idempotent* matrix. By definition, an idempotent matrix is one that, when raised to any power, does not change in value. A is an idempotent matrix if $A \cdot A = A$. Idempotent matrices are central to the development of regression analysis.

9.3.2 Diagonal Matrices

The concept of a *diagonal matrix* is directly related to the idea of the *principal diagonal* of a matrix. By definition, a diagonal matrix is (usually) a square matrix in which all elements both above and below the main diagonal are zero.[7] Matrix D is a general expression for a diagonal matrix:

$$D = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ 0 & \vdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}.$$

Consider four examples.

1. $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ is a diagonal matrix.

2. $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ is a diagonal matrix.
   Only one of nonzero value is required on the main diagonal.

3. $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{bmatrix}$ is not a diagonal matrix.
   Confirm that this is an idempotent matrix.

4. $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is a diagonal matrix.
   The identity matrix is an important special case of a diagonal matrix.

[7] The main diagonal or principal diagonal of a matrix consists of the elements that lie on the diagonal that runs from top left to bottom right. We limit our attention to square matrices.

9.3.3 Scalar Matrix

A *scalar matrix* is any square matrix S such that $S = \lambda \cdot I = \lambda[\cdot S_{ij}]$, where λ is any scalar. For example, the following matrix is a scalar matrix

$$S = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix} = 3 \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = 3 \cdot I$$

9.3.4 Null Matrix

In the real number system, zero has the unique property that for any number a, we have $0 \cdot a = a \cdot 0 = 0$. Also, $a + 0 = 0 + a = a$. In matrix algebra, the null or zero matrix plays a role similar to that of zero in the real number system. The definition of a null matrix is this: A null or zero matrix consists of elements that are all equal to zero. For example, all of the following three matrices are null.

$$O = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad O = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad O = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

As the second and third examples above indicate, the null matrix is not restricted to being a square matrix, as are the identity, diagonal, and scalar matrices. A square null matrix is idempotent.

The null matrix, like the number 0 in the real number system, has several unique qualities. For instance, the commutative law for addition of matrices holds when the null matrix and another matrix A are added if both the null matrix and the A matrix satisfy the usual dimensional requirements. That is,

$$\underset{m \times n}{A} + \underset{m \times n}{0} = \underset{m \times n}{0} + \underset{m \times n}{A} = \underset{m \times n}{A}$$

The commutative law with respect to the multiplication of matrices presents a more difficult problem when a null matrix is involved. The two products $A \cdot O = O$ and $O \cdot A = O$ both result in a null matrix. However, if matrix A is not square, then the products ($A \cdot O$ and $O \cdot A$) *will not commute*. That is,

$$\underset{m \times n \;\; n \times p}{A \cdot 0} = \underset{m \times p}{0} \neq \underset{q \times m \;\; m \times n}{0 \cdot A} = \underset{q \times n}{0}$$

On the other hand, if both matrix A and the null matrix are square, then the two products will commute. Thus,

$$\underset{m \times m \;\; m \times m}{A \cdot 0} = \underset{m \times m \;\; m \times m}{0 \cdot A} = \underset{m \times m}{0}$$

9.3.5 A Digression on Matrix Algebra

In three instances the intuition we have developed with reference to the algebra of numbers interferes with the proper application of matrix algebra.

Case 1. We have already discussed the situation in which the product $A \cdot B$ does commute in the algebra of numbers, but does not commute in matrix algebra. For example, $5 \cdot 6 = 6 \cdot 5 = 30$, while

$$\begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 3 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 6 & 10 \end{bmatrix} \neq \begin{bmatrix} 2 & 3 \\ 2 & 4 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 8 & 3 \\ 10 & 4 \end{bmatrix}.$$

Case 2. Given two real numbers a and b, we know from number algebra that if $a \cdot b = 0$, then $a = 0$ and/or $b = 0$. However, in matrix algebra, the product $A \cdot B = O$ does not imply that $A = O$ and/or that $B = O$. The following two examples illustrate this point.

$$A = \begin{bmatrix} 0 & 0 \\ -2 & 3 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 3 \\ 0 & 2 \end{bmatrix} \quad A \cdot B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = O$$

$$A = \begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix} \quad B = \begin{bmatrix} -3 & 6 \\ 1 & -2 \end{bmatrix} \quad A \cdot B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = O$$

Case 3. Given three real numbers a, b, and c, we know from number algebra that when $a \cdot b = a \cdot c$ (with $a \neq 0$), then $b = c$. Once again, however, this relationship does not hold in matrix algebra. For example, given matrices A, B, and C such that $A \cdot B = A \cdot C$, it does not follow that B and C are identical matrices such that $B = C$, as this example shows.

$$A = \begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix} \quad B = \begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix} \quad C = \begin{bmatrix} -4 & 2 \\ 4 & 4 \end{bmatrix} \quad A \cdot C = \begin{bmatrix} 8 & 14 \\ 16 & 28 \end{bmatrix} = A \cdot B,$$

but $B \neq C$.

9.3.6 The Transpose of a Matrix

We are now familiar with the meaning of the dimension of a matrix. The occasion sometimes arises, especially where matrix multiplication is concemed, when we need to form a new matrix whose rows and columns are interchanged in such a way that they reverse the dimension of the original matrix. Such a new matrix is referred to as the transpose of the original matrix.

More formally, given an $m \times n$ matrix labeled A, the transpose of A, denoted by A' or A^T, is a new matrix whose rows are the columns of A and whose columns are the rows of A. Thus, the new matrix has the dimension $n \times m$. That is, if

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = [a_{ij}] \quad \text{then}$$

$$A' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix} = [a_{ji}].$$

To repeat, the original matrix A has the dimension $m \times n$, while the transpose of A, A' or A^T , has the dimension $n \times m$, as the next four examples illustrate.

1. $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$, so $A' = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$.
   A is a 3×2 matrix; A' is a 2×3 matrix.

2. $B = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$ so $B' = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$.
   B is a 1×3 matrix, while B' is a 3×1 matrix. The transpose of an n-dimensional row vector is an n-dimensional column vector and *vice versa*.

3. $I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ so $I' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. That is, $I = I'$.

4. Likewise, for any square diagonal matrix: If $A = \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & \\ 0 & 0 & a_{33} \end{bmatrix}$, then
   $A' = \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & \\ 0 & 0 & a_{33} \end{bmatrix}$, or $A = A'$.

Transposed matrices exhibit four properties that are of immediate interest.

Property 1: $(A')' = A$. That is, the transposed matrix of an already transposed matrix is the original matrix. This proposition is easily demonstrated: $A' = [a_{ij}]' = [a_{ji}]$ and $(A')' = [a_{ji}]' = a_{ij}$.

Example:
$A = \begin{bmatrix} 1 & 5 \\ -3 & 7 \end{bmatrix}$ implies that $A' = \begin{bmatrix} 1 & -3 \\ 5 & 7 \end{bmatrix}$ and that $(A')' = \begin{bmatrix} 1 & 5 \\ -3 & 7 \end{bmatrix} = A$.

Property 2: $A \pm B)' = A' \pm B'$. More specifically, If A is an $m \times n$ matrix, and B is also an $m \times n$ matrix, then both $A \pm B)'$ and $A' \pm B'$ are also $n \times m$ matrices that have the same elements. In words, this property asserts that the transpose of a sum of two matrices is the sum of the transposes of those two matrices.

To prove this assertion, let $A = [a_{ij}]$ and $B = [b_{ij}]$, so that $C = A \pm B = [c_{ij}]$. Then $[c_{ij}]' = [c_{ji}] = [a_{ji}] \pm [b_{ji}] = [a_{ji}]' \pm [b_{ji}]'$.

We can extend this result to the addition or subtraction of any finite number of matrices such that $(A_1 \pm A_2 \pm \cdots \pm A_n)' = A_1' \pm A_2' \pm \cdots \pm A_n'$

The next example illustrates this property.
$A = \begin{bmatrix} 1 & 5 \\ 0 & 8 \\ 2 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 0 \\ 4 & 3 \\ -4 & 2 \end{bmatrix} \quad A + B = \begin{bmatrix} 2 & 5 \\ 3 & 10 \\ -2 & 2 \end{bmatrix}$
and
$(A + B)' = \begin{bmatrix} 2 & 3 & -2 \\ 5 & 10 & 2 \end{bmatrix} \quad A' = \begin{bmatrix} 1 & 0 & 2 \\ 5 & 8 & 1 \end{bmatrix}$
$B' = \begin{bmatrix} 1 & 3 & -4 \\ 0 & 2 & 1 \end{bmatrix} \quad A' + B' = \begin{bmatrix} 2 & 3 & -2 \\ 5 & 10 & 2 \end{bmatrix}$

Property 3: $(A \cdot B)' = B' \cdot A'$. Matrix A has the dimensions $m \times n$, and matrix B has the dimensions $n \times p$. Then $(A \cdot B)' = B' \cdot A'$. That is, the transpose of the product of these two matrices is the product of the individual transposes of these matrices *in reverse order*.

Proof: Let $A = [a_{ik}]$, $B = [b_{jk}]$, and $C = [c_{ij}] = A \cdot B$. Then

$$c_{ij} = \sum_k a_{ik} \cdot b_{kj} \quad \text{and} \quad c'_{ij} = c_{ji} = \sum_k a_{jk} \cdot b_{ki} = \sum_k a'_{kj} \cdot b'_{ik} = \sum_k b'_{ik} \cdot a'_{kj}.$$

We can extend result to include the product of any finite number of matrices: $A_1 \cdot A_2 \cdot \cdots \cdot A_n)' = A_n' \cdot \cdots \cdot A_2 \cdot A_1$.

The next three examples illustrate this property. The first example was created as the text was being printed. That is, the computations were done by hand. In the second and third, we used *Maxima* to create the matrices and to carry out the computations. In each of these two examples the *Maxima* output consists of three lists. The first list is the original matrices (A and B or A, B, and C). The second list contains the products of the original lists, the transpose of that product, and the transposes of the original matrices. The third list contains a single item, $B' \cdot A'$ or $C \cdot B' \cdot A'$. The accompanying workbook shows the commands.

1. $A = \begin{bmatrix} 3 & 2 \\ 0 & 5 \end{bmatrix} \quad B = \begin{bmatrix} 2 & 0 \\ 1 & 3 \end{bmatrix}$

   Then

   $A \cdot B = \begin{bmatrix} 8 & 6 \\ 5 & 15 \end{bmatrix} \quad (A \cdot B)' = \begin{bmatrix} 8 & 5 \\ 6 & 15 \end{bmatrix} \quad A' = \begin{bmatrix} 3 & 0 \\ 2 & 5 \end{bmatrix} \quad B' = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}$

   and $(B \cdot A)' = \begin{bmatrix} 8 & 5 \\ 6 & 15 \end{bmatrix}$. Thus $(A \cdot B)' = \begin{bmatrix} 8 & 5 \\ 6 & 15 \end{bmatrix} = B' \cdot A'$

2. A and B

   $$[\begin{bmatrix} 1 & 3 & -1 \\ 2 & 0 & 0 \\ 0 & -1 & 6 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ -1 & 2 \\ 1 & 3 \end{bmatrix}]$$

   $A \cdot B$, $(A \cdot B)'$, A', and B'

   $$[\begin{bmatrix} -3 & 3 \\ 2 & 0 \\ 7 & 16 \end{bmatrix}, \begin{bmatrix} -3 & 2 & 7 \\ 3 & 0 & 16 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 0 \\ 3 & 0 & -1 \\ -1 & 0 & 6 \end{bmatrix}, \begin{bmatrix} 1 & -1 & 1 \\ 0 & 2 & 3 \end{bmatrix}]$$

   $A \cdot B)'$

   $$\begin{bmatrix} -3 & 2 & 7 \\ 3 & 0 & 16 \end{bmatrix}$$

3. A, B, and C

   $$[\begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}, \begin{bmatrix} 0 & 3 & -1 \\ 1 & 0 & 2 \end{bmatrix}, \begin{bmatrix} 3 & 2 & 1 \\ 2 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}]$$

   $A \cdot B \cdot C$, $(A \cdot B \cdot C)'$, $(A \cdot B \cdot C)'$, A', B', and C'

   $$[\begin{bmatrix} 15 & -4 & 1 \\ 35 & -1 & -1 \end{bmatrix}, \begin{bmatrix} 15 & 35 \\ -4 & -1 \\ 1 & -1 \end{bmatrix}, \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 3 & 0 \\ -1 & 2 \end{bmatrix}, \begin{bmatrix} 3 & 2 & 1 \\ 2 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}]$$

$C' \cdot B' \cdot A'$

$$\begin{bmatrix} 15 & 35 \\ -4 & -1 \\ 1 & -1 \end{bmatrix}$$

Property 4: When the transpose of a square matrix results in the original matrix, the original matrix is said to be *symmetric about its main diagonal*. That is, if $A = A'$, then we have a symmetric matrix. The elements in matrix A that are above the main diagonal are a mirror image of the elements of matrix A that are below the main diagonal.

For example $A = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 3 & 5 \\ 4 & 5 & 6 \end{bmatrix}$ as is $I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. All identity matrices are symmetrical.

Exercise 9-1

Find the transpose of each of the following seven matrices.

1. $\begin{bmatrix} 1 & 3 & 4 \\ 5 & -1 & -1 \end{bmatrix}$ 2. $\begin{bmatrix} -12 & 5 \\ 0 & 8 \\ -5 & 4 \end{bmatrix}$ 3. $\begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}$ 4. $\begin{bmatrix} 4 \\ 0 \\ -3 \end{bmatrix}$

5. $\begin{bmatrix} 1 & 2 & 5 \end{bmatrix}$ 6. $\begin{bmatrix} 1 & -1 & 2 \\ 0 & 3 & 4 \end{bmatrix}$ 7. $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 4 & 4 & 4 \end{bmatrix}$

8. Given that $A = \begin{bmatrix} 2 & 4 \\ 1 & 2 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 3 \\ 0 & 5 \end{bmatrix}$, and $C = \begin{bmatrix} 1 & 0 \\ -1 & 3 \end{bmatrix}$ show that
(a) (A + B)' = A' + B', (b) $(A \cdot B)' = B' \cdot A'$,
(c) $(A \cdot B \cdot C)' = C' \cdot B' \cdot A'$ and (d) $(A')' = A$.

9.4 Determinants

Previous sections have demonstrated how it to write a linear equation system in shorthand by means of matrix algebra. For example, we developed the shorthand $A \cdot X = C$ to represent a typical system of linear equations. It's nice to be able to write a large system of equations in a concise, shorthand notation. However, the premium is on being able to solve that system of

equations for the values of the unknown variables represented by the vector X.

We can find solutions in a large number of situations. For example, when we have two linear equations in two unknowns, we can find the solution values of the unknown variables by setting one unknown variable equal to the other, substituting, and solving. It is apparent, nonetheless, that the process of substitution becomes exceedingly complex when many equations and unknowns are involved. Therefore we must further develop our matrix-algebra tools so that we can find the solution values for a large set of simultaneous linear equations.

The first step we must take is to master the concept of the determinant of a matrix. Once we have found the value of the determinant, we ordinarily know whether or not we can solve the system of equations in question, and we often can find the precise solution values. The *determinant* of a square matrix is a uniquely defined scalar (number) that is characteristic of that particular matrix. Determinants are denoted by vertical straight lines. Thus,

$$|A| = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

is a scalar (number). This scalar is said to the the determinant of of the n^{th} order.

The determinant is calculated by summing products of the matrix's terms in a specific fashion. Consider a 2×2 matrix

$$\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}.$$

Its determinant is $a_{1,1}\, a_{2,2} - a_{1,2}\, a_{2,1}$, the product of the terms in the diagonal *less* the product of the two off-diagonal terms.

This cross-multiplication process can be extended to a 3×3 matrix as the third example below illustrates. The third-order determinant consists of six terms, three of which are added and three of which are subtracted in the process of cross-multiplication. Using the same notation as in the 2×2 matrix, the determinant of a 3×3 determinant is

$a_{11} \cdot a_{22} \cdot a_{33} + a_{12} \cdot a_{23} \cdot a_{31} + a_{13} \cdot a_{21} \cdot a_{32} - (a_{31} \cdot a_{22} \cdot a_{13} + a_{32} \cdot a_{23} \cdot a_{11} + a_{33} \cdot a_{21} \cdot a_{12})$.

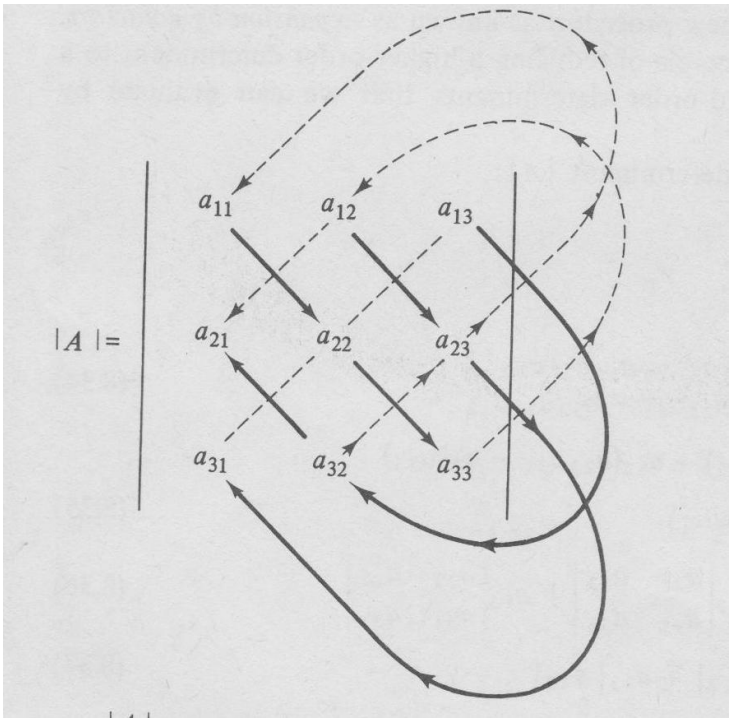


Figure 9.3: Calculating the value of a third-order determinant

Figure 9.3 illustrates how we can find the various products when a third-order determinant is involved. The solid lines in Figure 9.3 form a cross product of three elements, beginning in each case with an element in the top row and including two other elements that are each from a different row and column. The dashed lines also form a cross product of three elements, beginning in each case with an element from the bottom row and including two other elements, each of which is from a different row and column. The six products together determine the value of the determinant, with the solid-line products to be added and the dashed-line products to be subtracted.

Examples

1. $A = \begin{bmatrix} 4 & 2 \\ 1 & 5 \end{bmatrix}$, so $|A| = 4 \cdot 5 - 1 \cdot 2 = 20 - 2 = 18$.

2. $A = \begin{bmatrix} -3 & -4 \\ 1 & 5 \end{bmatrix}$, so $|A| = -3 \cdot 5 - 1 \cdot (-4) = -15 + 4 = -11$.

3. $A = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 2 & 4 \\ 4 & 1 & 3 \end{bmatrix}$, so $|A| = (1 \cdot 2 \cdot 3 + 0 \cdot 4 \cdot 4 + 0 \cdot 3 \cdot 1) -$
$(4 \cdot 2 \cdot 0 + 1 \cdot 4 \cdot 1 + 3 \cdot 3 \cdot 0) = 2.$

Evaluating Determinants of Orders Higher than Three. The cross-multiplication methods of evaluating determinants of orders two and three cannot be directly applied to determinants of orders higher than three. We can use another procedure to evaluate determinants of the fourth (and higher) orders. This new procedure is known as *expansion by cofactors*, and it operates on the principle of reducing a higher-order determinant to a series of second- or third-order determinants that we can evaluate by cross-multiplication.

We illustrate the general nature of this approach. We do not develop it fully, however, because computer algebra systems offer a much quicker and less error-prone approach to determining the value of the determinant of a large matrix. We illustrate this alternative immediately following our sketch of expansion by cofactors.

Consider once more a 3×3 matrix. We show that the determinant can be recast as a set of three smaller (2×2 matrices, each multiplied by one term in a given row or column. In our expansion, we use the terms of row 1. Any row or column could be used in this same way.

$$|A| = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix},$$

which we have determined to be

$$a_{1,1}\,a_{2,2}\,a_{3,3} - a_{1,2}\,a_{2,1}\,a_{3,3} - a_{1,1}\,a_{2,3}\,a_{3,2} + a_{1,3}\,a_{2,1}\,a_{3,2} + a_{1,2}\,a_{2,3}\,a_{3,1} - a_{1,3}\,a_{2,2}\,a_{3,1}.$$

We now extract the terms that we will use in our process ($a_{1,1}, a_{1,2}$, and $a_{1,3}$), yielding

$$a_{1,1}\,(a_{2,2}\,a_{3,3} - a_{2,3}\,a_{3,2}) - a_{1,2}\,(a_{2,1}\,a_{3,3} - a_{2,3}\,a_{3,1}) + a_{1,3}\,(a_{2,1}\,a_{3,2} - a_{2,2}\,a_{3,1}).$$

From the definition of the determinant, we can recast this expression in terms of three 2×2 matrices, each multiplied by the relevant cofactor:

$$a_{1,1} \cdot \begin{bmatrix} a_{2,2} & a_{2,3} \\ a_{3,2} & a_{3,3} \end{bmatrix} - a_{1,2} \cdot \begin{bmatrix} a_{2,1} & a_{2,3} \\ a_{3,1} & a_{3,3} \end{bmatrix} + a_{1,3} \cdot \begin{bmatrix} a_{2,1} & a_{2,2} \\ a_{3,1} & a_{3,2} \end{bmatrix}.$$

Label these smaller matrices $|A_{1,1}|, |A_{1,2}|$, and $|A_{1,3}|$. We can now restate $|A|$ as follows:

$$|A| = \sum_{j=1}^{3} (-1)^{1+j} \cdot a_{1,j} \cdot |A_{1,j}|.$$

The term $(-1)^{i+j} \cdot |A_{i,j}|$ is the cofactor of $a_{i,j}$. When $i+j$ is an even number, a value is added to the sum; when $i+j$ is an odd number, a value is subtracted from the sum. As noted above, any row value (i, not just $i = 1$) can be used. Also, the summation could be over the rows in a specified column.

The terms $|A_{i,j}|$ are called *minors*. The terms $(-1)^(i + j) + |A_{i,j}|$ are called signed minors or *cofactors*. We can define each cofactor as $|C_{i,j}| = (-1)^{i+j} \cdot |Ai, j|$. Then a general expression for the solution of a determinant by the method of cofactors can be written as

$$|A| = \sum_{j=1}^{n} a_{i,j} \cdot |C_{i,j}|$$

for any row i or

$$|A| = \sum_{i=1}^{n} a_{i,j} \cdot |C_{i,j}|$$

for any column j.

The use of cofactors can greatly simplify the computation of some determinants. Look at Example 3 above. All of the terms in row 1 except the first equal. Therefore, we can see, almost by inspection, that the determinant is $1 \cdot (2 \cdot 3 - 1 \cdot 4) = 2$ Judicious selection of row or column can make the computation much easier than the preceding discussion might suggest that it is. Even so, for $n > 3$, the use of a computer algebra system is recommended. Consider this 5×5 matrix

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{2}{3} & \frac{1}{2} & \frac{2}{5} & \frac{1}{3} & \frac{2}{7} \\ \frac{3}{4} & \frac{3}{5} & \frac{1}{2} & \frac{3}{7} & \frac{3}{8} \\ \frac{4}{5} & \frac{2}{3} & \frac{4}{7} & \frac{1}{2} & \frac{4}{9} \\ \frac{5}{6} & \frac{5}{7} & \frac{5}{8} & \frac{5}{9} & \frac{1}{2} \end{bmatrix}.$$

No row or column contains any zeros, and the computation, even with the application of the method cofactors will be arduous and subject to mistakes.

The solution, given almost instantaneously by *Maxima* is $\frac{1}{560105280000}$.[8]

The site http://www.purplemath.com/modules/minors.htm offers a slightly more expansive discussion of this topic. This site shows how to compute the value of a 4×4 determinant. It also offers some tips on how to manipulate a matrix so as to generate cells that have 0 as a value. These tips involve using the properties of determinants that we state below.

A digression on the differences between matrices and determinants. Matrices and determinants are not the same thing. A matrix, denoted by brackets or parentheses, has no numeric value. A matrix is a rectangular array of numbers, variables, and parameters. A determinant, on the other hand, does have a numeric value. A determinant is defined to be a scalar (number).

Matrices can be of any dimension and need not be square. Determinants must be square. A 2×3 determinant does not exist.

Properties of determinants. We can usefully apply the following properties when we work with determinants. These properties apply to determinants of any dimension.

Property 1. The determinant of a matrix A has the same value as the determinant of its transpose A'. Let

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}. \text{ Then } A' = \begin{bmatrix} a & c \\ b & d \end{bmatrix}.$$

For both of these matrices, the determinant is the same, $a \cdot d - b \cdot c$.

Property 2. Interchanging any two rows (or any two columns) of a determinant does not alter the absolute value of that determinant. It does, however, change the sign of the determinant.

For example $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = a \cdot d - b \cdot c$ but $\begin{vmatrix} b & a \\ d & c \end{vmatrix} = b \cdot c - a \cdot d$. Confirm that interchanging rows 1 and 2 of the initial matrix has the same effect as interchanging the columns.

Property 3. A determinant in which any two rows (or any two columns) are identical, or a determinant in which any two rows (or any two columns) are multiples of each other, has a value of zero.

[8]Three commands result in the creation of this matrix and the compuation of its determinant: h[i,j]:= i/(i+j), genmatrix(h,5,5), and determinant(%).

For example $\begin{vmatrix} a & b \\ k \cdot a & k \cdot b \end{vmatrix} = a \cdot b \cdot k - k \cdot a \cdot b = 0.$

Property 4. A determinant in which any row or any column has all zero elements has a value of zero.

For example $\begin{vmatrix} a & b \\ 0 & 0 \end{vmatrix} = 0 \cdot a - 0 \cdot b = 0.$

Property 5. Adding (or subtracting) a multiple of one row of a determinant to (from) another row of that determinant, or adding (or subtracting) a multiple of one column of a determinant to (from) another column of that determinant does not change the value of the determinant.

For example $\begin{vmatrix} a & b \\ c + k \cdot a & d + k \cdot b \end{vmatrix} = a \cdot d + a \cdot k \cdot b - (c \cdot b + k \cdot a \cdot b = a \cdot d - c \cdot b).$

Property 6. If every element in one row (or one column) is multiplied by a constant k, then the value of the determinant is also multiplied by k. For example,

For $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, |A| = a \cdot d - c \cdot b =$ and $\begin{vmatrix} a & b \\ k \cdot c & k \cdot d \end{vmatrix} = k \cdot a \cdot d - k \cdot c \cdot b = k \cdot |A|.$

By iteration, if we multiply the elements of two columns of a matrix M by k, then the determinant of the new matrix is $k^2 \cdot |M|$. Further, if we multiply all elements of an $n \times n$ matrix M by k the determinant of the resulting matrix is $k^n \cdot |M|$.

Exercise 9.3. Evaluate each of the following matrices. Confirm your solutions with *Maxima*.

1. $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 1 & 5 & 7 \end{bmatrix}$ 2. $\begin{bmatrix} 1 & 3 & 4 \\ 2 & 0 & 7 \\ 5 & 6 & 9 \end{bmatrix}$ 3. $\begin{bmatrix} 4 & 1 & 6 \\ 7 & 2 & 9 \\ 3 & 0 & 8 \end{bmatrix}$

4. $\begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}$ 5. $\begin{bmatrix} 2 & 1 & -3 & 4 \\ 5 & -4 & 7 & -2 \\ 4 & 0 & 6 & -3 \\ 3 & -2 & 5 & 2 \end{bmatrix}$ 6. $\begin{bmatrix} 1 & 1 \\ -3 & -3 \end{bmatrix}$

7. $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ 8. $\begin{bmatrix} 2 & 1 & 1 \\ 0 & 5 & -2 \\ 1 & -3 & 4 \end{bmatrix}$ 9. $\begin{bmatrix} 1 & 2 & -2 & 3 \\ 3 & -1 & 5 & 0 \\ 1 & 7 & 2 & -3 \\ 4 & 0 & 2 & 1 \end{bmatrix}$

9.5 The Inverse of a Matrix

Much of our work in this chapter has dealt with a system of n linear equations in n unknowns, such as

$$\begin{bmatrix} a_{11} \cdot x_1 + & a_{12} \cdot x_2 + & \cdots & +a_{1n} \cdot x_n = & c_1 \\ a_{21} \cdot x_1 + & a_{22} \cdot x_2 + & \cdots & +a_{2n} \cdot x_n = & c_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1} \cdot x_1 + & a_{n2} \cdot x_2 + & \cdots & +a_{nn} \cdot x_n = & c_n \end{bmatrix}$$

We have seen that we can write this system of equations as $A \cdot X = C$. We shall soon discuss the problem of solving such an equation system.

Solving a linear equation system of the form $A \cdot X = C$ seems easy at first glance. We are tempted to divide both sides of the equation system by A, giving us $X = C/A$. Unfortunately, we cannot do this. In matrix algebra, the division operator is not defined. That is, we cannot divide a matrix C by another matrix A. Instead, we must use a technique that involves finding the *inverse* of a matrix. Symbolically, in ordinary algebra, the follow equalities hold: $C/A = C \cdot A^{-1} = A^{-1} \cdot C$. That is, division can be restated as a multiplication process, where $1/C$ is the inverse of C. The process in matrix algebra is analagous to that in ordinary algebra, but it must be developed carefully.

To repeat, in matrix algebra, we must use the inverse of a matrix instead of dividing one matrix by another. Moreover, we must be careful in doing so, for (as we have already seen) matrix multiplication does not commute. Given

$$\underset{n \times n}{A} \cdot \underset{n \times 1}{X} = \underset{n \times 1}{C}$$

if an inverse does exist for A, then the solution for the X matrix is

$$\underset{n \times n}{X} = \underset{n \times n}{A^{-1}} \cdot \underset{n \times 1}{C} .$$

This is in general not equivalent to $X = C \cdot A^{-1}$.

We proceed by defining the inverse matrix and then considering how to go about determining its contents. Definition: If it exists, the inverse of a square

matrix A is another square matrix, denoted A^{-1}, that satisfies the relation $A^{-1} \cdot A = A \cdot A^{-1} = I$.

This definition of the inverse matrix is consistent with ordinary algebraic rules. For example, in ordinary algebra, $a \cdot a^{-1} = a^{-1} \cdot a = 1$. In the case of matrix algebra, it makes no difference whether A is premultiplied or postmultiplied by A^{-1}. The product that results is always the identity matrix I.

Our definition of the inverse of a matrix has two noteworthy implications. We state these without proof. See Perlis [17].

1. If an inverse of a matrix does exist, then this inverse is unique. That is, A^{-1} is the only matrix that, when multiplied by A, results in the identity matrix I.

2. An inverse matrix is that the commutative law of multiplication does hold when we multiply a square matrix by its inverse. That is, $A \cdot A^{-1} = A^{-1} \cdot A = I$.

   This is an important exception to the rule that, in general, matrix multiplication is not commutative.

9.5.1 Finding the Inverse of a Matrix if It Exists

The following theorem yields both the necessary and sufficient conditions for determining whether or not an inverse matrix exists, and how one can find it. Theorem: A square matrix A has an inverse matrix A^{-1} if and only if the determinant $|A| \neq 0$, in which case A is said to be *nonsingular*. The inverse is given by $A^{-1} = (1/|A| \cdot adjA$, where "$adjA$" refers to the adjoint of matrix A.

We emphasize three points concerning this theorem. First, a square matrix is a necessary, but not a sufficient condition, for an inverse matrix to exist. Second, if matrix A does have an inverse, A is said to be nonsingular. If matrix A does not have an inverse, then A is said to be singular. (We shall return to the concepts of singularity and nonsingularity when we solve systems of simultaneous equations.) Third, the condition $|A| \neq 0$ is a sufficient condition for an inverse to exist.

The theorem above uses the term *adjoint* of matrix A. We now define this new concept. Definition: The adjoint of matrix A, denoted by $adjA$, is the transpose of the cofactor matrix of A, which we encountered when computing determinants. More formally, let the cofactor matrix of matrix A be given by $C = [\,|A_{ij}|\,]$. Then the adjoint of A is given as

$$\begin{bmatrix} |C_{11}| & |C_{12}| & \cdots & |C_{1n}| \\ |C_{21}| & |C_{22}| & \cdots & |C_{2n}| \\ \vdots & \vdots & & \vdots \\ |C_{n1}| & |C_{n2}| & \cdots & |C_{nn}| \end{bmatrix}' = adjA = \begin{bmatrix} |C_{11}| & |C_{21}| & \cdots & |C_{n1}| \\ |C_{12}| & |C_{22}| & \cdots & |C_{n2}| \\ \vdots & \vdots & & \vdots \\ |C_{1n}| & |C_{2n}| & \cdots & |C_{nn}| \end{bmatrix}.$$

Remember that a cofactor of an element is a signed minor given by $|C_{ij}| = (-1)^{i+j} \cdot |A_{ij}|$. The cofactor is *not* the element of a particular row and column multiplied by the signed minor. That is, $|C_{ij}| =\neq a_{ij} \cdot (-1)^{i+j} \cdot |A_{ij}|$. Only when we expand by a particular row or column in order to find the determinant do we need to multiply the element and cofactor together.

Consider a 2×2 matrix, $A = \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}$. Then $|A| = 5$, $[\,|C_{ij}|\,] = \begin{bmatrix} 4 & -3 \\ -1 & 2 \end{bmatrix}$, and $adjA = \begin{bmatrix} 4 & -1 \\ -3 & 2 \end{bmatrix}$. Thus $A^{-1} = \frac{1}{5} \cdot \begin{bmatrix} 4 & -1 \\ -3 & 2 \end{bmatrix}$. To confirm that we have found the inverse, compute $A^{-1} \cdot A = \begin{bmatrix} 4/5 & -1/5 \\ -3/5 & 2/5 \end{bmatrix} \cdot \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix} = \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Now we consider a 3×3 matrix

$$M = \begin{bmatrix} 1 & 3 & 4 \\ 2 & 0 & 7 \\ 5 & 6 & 9 \end{bmatrix}.$$

To analyze this matrix, we use *Maxima*, which computes the determinant of matrix M. The value is 57. We use the command `adjM: adjoint(M)` to define the adjoint matrix and assign it a name.[9] The result is

$$adjM = \begin{bmatrix} -42 & -3 & 21 \\ 17 & -11 & 1 \\ 12 & 9 & -6 \end{bmatrix}.$$

[9]Maxima does not have a command to createthe matrix of cofactors, but this is just the transpose of the adjoint.

We can either divide the adjoint matrix by the determinant or use the command `invert(M)` to determine the inverse matrix, which is

$$M^{-1} = \begin{bmatrix} -\frac{14}{19} & -\frac{1}{19} & \frac{7}{19} \\ \frac{17}{57} & -\frac{11}{57} & \frac{1}{57} \\ \frac{4}{19} & \frac{3}{19} & -\frac{2}{19} \end{bmatrix}.$$

Finally, we can compute either $M^{-1} \cdot M$ or $M \cdot M^{-1}$ to confirm that the result is a 3×3 identity matrix.

Suppose that our matrix is

$$\begin{bmatrix} 1 & 0 & 4 \\ 2 & -3 & 1 \\ 6 & -9 & 3 \end{bmatrix}.$$

For this matrix (for which the values in the third row are 3 times their counterparts in the second row), the determinant is 0. Thus we cannot divide the elements of the adjoint matrix by the determinant in order to compute the elements of the inverse matrix. This matrix is singular and does not have an inverse matrix.

The three examples above illustrate two important points. First, the definition of an inverse matrix requires that $|A| \neq 0$. Not only does this mean that matrix A is nonsingular, but also it recognizes that division by zero is undefined. Therefore $|A| \neq 0$ is a sufficient condition for an inverse matrix to exist. Second, it is always possible to check whether the theorem concerning inverse matrices has been applied correctly. One need only multiply the alleged inverse and the original matrix. If the theorem has been applied correctly, the result must be the identity matrix.

Properties of Inverse Matices. Three properties of inverse matrices warrant mention.

1. If A^{-1} exists, the $(A^{-1})^{-1} = A$. That is, the inverse of an inverse matrix, if it exists, is the original matrix. For an example, consider M^{-1} in Example 2 above. Its inverse is

$$\begin{bmatrix} 1 & 3 & 4 \\ 2 & 0 & 7 \\ 5 & 6 & 9 \end{bmatrix},$$

   which is matrix M.

2. If A^{-1} and B^{-1} both exist, then $(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}$. That is, the inverse of the product of two matrices is equal to the product of their inverses in reverse order. This property generalizes to any number of matrices, so that $(A \cdot B \cdots Z)^{-1} = Z^{-1} \cdots B^{-1} \cdots A^{-1}$.

3. If A^{-1} exists, then $(A')^{-1} = (A^{-1})'$. That is, the inverse of the transpose is the transpose of the inverse.

Summary and Review: Finding the Inverse of a Matrix
By hand:

1. Determine whether or not the inverse matrix exists. That is, find $|A|$ If $|A| = 0$, then there is no inverse matrix.

2. Find the cofactor matrix. That is, find $C = [|C_{ij}|]$.

3. Find the adjoint of matrix A. That is, take the transpose of the cofactor matrix such that $C' = adjA$.

4. Divide $adjA$ by the determinant of A. That is,

$$A^{-1} = \frac{1}{|A|} \cdot adjA.$$

With *Maxima*: Use the command `invert(M)`, where `M` is the name that you have assigned to the matrix.

Exercise 9.4. For each of the following matrices, find the inverse if it exists.

1. $\begin{bmatrix} 2 & 1 \\ 0 & 5 \end{bmatrix}$ 2. $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ 3. $\begin{bmatrix} 2 & 1 & 1 \\ 0 & 5 & -2 \\ 1 & -3 & 4 \end{bmatrix}$ 4. $\begin{bmatrix} 2 & 1 & 3 \\ 3 & 0 & 1 \\ -1 & 1 & 4 \end{bmatrix}$

5. $\begin{bmatrix} -1 & 0 & 2 \\ 3 & 1 & -6 \\ -2 & -1 & 5 \end{bmatrix}$ 6. $\begin{bmatrix} 1 & 0 & -2 \\ -3 & -1 & 6 \\ 2 & 1 & -5 \end{bmatrix}$ 7. $\begin{bmatrix} 7 & 6 & 5 \\ 1 & 2 & 1 \\ 3 & -2 & 1 \end{bmatrix}$

8. $\begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix}$ 9. $\begin{bmatrix} 0 & 5 \\ 6 & 4 \end{bmatrix}$ 10. $\begin{bmatrix} 2 & 4 \\ -3 & -6 \end{bmatrix}$

9.6 Solving Simultaneous Linear Equations

By this juncture, we have become quite familiar with a simultaneous linear equation system of the form

$$\underset{n\times n}{A} \cdot \underset{n\times n}{X} = \underset{n\times 1}{C}.$$

If the inverse matrix A^{-1} does exist, then premultiplying both sides of $A\cdot X = C$ by A^{-1} yields $A^{-1}\cdot X = A^{-1}\cdot C$, or

$$\underset{n\times 1}{X} = \underset{n\times n}{A^{-1}} \cdot \underset{n\times 1}{C} = \underset{n\times 1}{D}.$$

Our definition of matrix equality tells us that the left side $n \times 1$ column vector of unknown variables represented by X must be equal to the right side $n \times 1$ column vector of solution values represented by D if the two sides are indeed equal.

We have also found that we can find an inverse matrix (such as A^{-1}), only if the matrix A is square. We stated this requirement by asserting that $A^{-1}\cdot A = A\cdot A^{-1} = I$, the identity matrix. This means that the number of equations is equal to the number of unknown variables.

Recall that when the value of the determinant of a matrix is zero, then you cannot find an inverse for that matrix. That is, $A^{-1} = \frac{1}{|A|}\cdot adj\,A$ and $A|| \neq 0$.

Thus, when $|A| \neq 0$, there is a unique solution for a linear equation system. Nonsingularity implies that an inverse can be found. When an inverse matrix can be found, then there is a unique solution. We may summarize the relationship between nonsingularity and the existence of a unique solution as follows: Nonsingularity implies the existence of an inverse and a unique solution and *vice versa*.

Consider two examples.

1. Given: $y-4\cdot x = 12$ and $y+3\cdot x = 5$. The commands `[A:matrix([1,-4 ], [1, 3 ]), C:transpose(matrix([12,5])), invA : invert(A),`

`D: invA.C];` create the commands that generate the following: A, the vector C, A^{-1}, and D. The resulting output follows:

$$\begin{bmatrix} 1 & -4 \\ 1 & 3 \end{bmatrix}, \begin{bmatrix} 12 \\ 5 \end{bmatrix}, \begin{bmatrix} \frac{3}{7} & \frac{4}{7} \\ -\frac{1}{7} & \frac{1}{7} \end{bmatrix}, \begin{bmatrix} 8 \\ -1 \end{bmatrix}$$

So, this system's solution is $y = 8$ and $x = -1$.

2. Given: $2 \cdot x_1 - x_2 - x_3 = 0, -x_1 + 4 \cdot x_2 - x_3 = 0, x_1 + x_2 = 8$. Then the X matrix, the C vector, X^{-1} and D are as below (created with the same set of commands as above).

$$\begin{bmatrix} 2 & -1 & -1 \\ -1 & 4 & -1 \\ 1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 8 \end{bmatrix}, \begin{bmatrix} \frac{1}{8} & -\frac{1}{8} & \frac{5}{8} \\ -\frac{1}{8} & \frac{1}{8} & \frac{3}{8} \\ -\frac{5}{8} & -\frac{3}{8} & \frac{7}{8} \end{bmatrix}, \begin{bmatrix} 5 \\ 3 \\ 7 \end{bmatrix}$$

The solution to this system is $x_1 = 5, x_2 = 3$, and $x_3 = 7$

Application: Ordinary Least Squares Estimation

In your statistics course you encountered the "normal equations." These equations solved for values of the parameters of a linear equation and were phrased in terms of the observed values of the independent variable(s) and the dependent variable, where the relationship is linear. That is, $y_t = b_0 + b_1 \cdot x_{1,t} + b_2 \cdot x_{2,t} + \ldots + b_k \cdot x_{k,t} + e_t$. A set of n values of the dependent and dependent variables are collected ($n > k + 1$). We can place these data into two matrices. The e_t are error terms that result from a random process; they are the "noise" while the first part of the equation is the "signal."

The first matrix, X consists of n rows and $n+1$ columns (the $+1$ is to accommodate the constant term, as our illustrative example below demonstrates). The second matrix, C consists of a single column of n values. In matrix notations, the relationship between X and Y is as follows: $Y = B \cdot X + E$ where E is a vector of n error terms. Statistics textbooks show that the ordinary least squares estimators for the values of the b terms (the coefficients) are derived as follows: $Bols = (X' \cdot X)^{-1} \cdot X' \cdot Y$. That is, the transpose of X, which is $k \times n$) is muliplied by X, which is $n \times k$, yielding a square $k \times k$ matrix, for which the inverse is computed. This inverse is multiplied by the product of the $X' \cdot X$ and Y matrices, which is a $k \times 1$ vector. The product of the inverse, a $k \times k$ matrix and this vector is another $k \times 1$ vector, which is the OLS estimators.

To derive a specific set of estimates (numerical values) from this set of estimators (rules), we require a data set. To illustrate the process, we look at a hypothetical example in which $y_t = b_0 + b_1 \cdot x_t + b_2 \cdot x_{2,t} + e_t$. The e_t terms are not observable. They can be estimated after the parameters have been estimated. The data are as follows:

$$\begin{bmatrix} 1 & 2 & 5 \\ 1 & 5 & 8 \\ 1 & 6 & 9 \\ 1 & 7 & 7 \\ 1 & 6.5 & 8 \end{bmatrix}, \begin{bmatrix} 1 \\ 2.1 \\ 3 \\ 4.5 \\ 7 \end{bmatrix}.$$

The first matrix contains the x values. This matrix includes a column of 1's which attach to the constant term, the estimated value of b_0. We are estimating three parameters, b_0, b_1, and b_2. We have five data points. Mechanically, this is enough to provide estimates. It is not nearly enough to provide reliable estimates, but it does illustrate the process.

Now we transpose X, and then mulitply this transpose by X. The result in the 3×3 matrix in the middle. Then we determine the inverse of that matrix.

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 5 & 6 & 7 & 6.5 \\ 5 & 8 & 9 & 7 & 8 \end{bmatrix}, \begin{bmatrix} 5 & 26.5 & 37 \\ 26.5 & 1.5610^2 & 2.0510^2 \\ 37 & 2.0510^2 & 283 \end{bmatrix}, \begin{bmatrix} 6.63 & 0.259 & -1.05 \\ 0.259 & 0.139 & -0.135 \\ -1.05 & -0.135 & 0.239 \end{bmatrix}$$

The first of the following pair of of matrices shows the result of multiplying X' by Y. Multiplying $(X' \cdot X)^{-1}$ by this matrix creates our set of OLS estimates for the coefficients: $(X' \cdot X)^{-1} \cdot (X' \cdot Y)$.

$$\begin{bmatrix} 17.6 \\ 1.0710^2 \\ 1.3610^2 \end{bmatrix}, \begin{bmatrix} 0.806 \\ 1.16 \\ -0.466 \end{bmatrix}.$$

Thus our estimated relationship is $y_t = 0.806 + 1.16 \cdot x_1 - 0.466 \cdot x_2$.[10]

Once we have computed the estimates for the coefficients, we can apply them to the data in order to see what values the models estimate. The final set of

[10]Be aware of the word "estimated." We do not know the true relationship (if any) between x and y and yielded these observed values.

output shows the estimates. It also repeats the Y vector for comparison, and it shows the size of the residuals–the differences between the OLS estimates and the actual values in this sample.

$$\begin{bmatrix} OLS\,estimates & Actual\,values & Residuals \\ \begin{bmatrix} 0.802 \\ 2.89 \\ 3.59 \\ 5.68 \\ 4.64 \end{bmatrix} & \begin{bmatrix} 1 \\ 2.1 \\ 3 \\ 4.5 \\ 7 \end{bmatrix} & \begin{bmatrix} 0.198 \\ -0.792 \\ -0.588 \\ -1.18 \\ 2.36 \end{bmatrix} \end{bmatrix}$$

Exercise 9.5. Write each of the following as matrices A, X and C, such that $A \cdot X = C$. For each, create the X^{-1} matrix and solve the system of equations. Confirm your solutions with *Maxima*.

1. $x_1 + 3 \cdot x_2 = 15$
   $x_1 - 2 \cdot x_2 = -3$
2. $2 \cdot x_1 + 3 \cdot x_2 = 10$
   $-4 \cdot x_1 + x_2 = -6$
3. $2 \cdot x_1 - 3 \cdot x_2 = 7$
   $3 \cdot x_1 + 5 \cdot x_2 = 1$
4. $10 \cdot x_1 - x_2 - x_3 = 0$
   $-x_1 + 12 \cdot x_2 - 2 \cdot x_3 = 0$
   $x_1 + 2 \cdot x_2 = 24$
5. $12 \cdot x_1 - 2 \cdot x_2 - x_3 = 0$
   $12 \cdot x_1 - 6 \cdot x_2 - x_5 = 0$
   $x_1 + x_2 = 16$
6. $x_1 + 2 \cdot x_2 - 3 \cdot x_3 = -1$
   $3 \cdot x_1 - x_2 + 2 \cdot x_3 = 7$
   $5 \cdot x_1 + 3 \cdot x_2 - 4 \cdot x_3 = 2$
7. $2 \cdot x_1 + x_2 - 2 \cdot x_3 = 10$
   $3 \cdot x_1 + 2 \cdot x_2 - 2 \cdot x_3 = 1$
   $5 \cdot x_1 + 4 \cdot x_2 + 3 \cdot x_3 = 4$
8. $x_1 + 2 \cdot x_2 - 3 \cdot x_3 = 6$
   $2 \cdot x_1 - x_2 + 4 \cdot x_3 = 2$
   $4 \cdot x_1 + 3 \cdot x_2 - 2 \cdot x_3 = 14$
9. $x_1 + 3 \cdot x_2 - 2 \cdot x_3 = 0$
   $2 \cdot x_1 - 3 \cdot x_2 + x_3 = 0$
   $3 \cdot x_1 - 2 \cdot x_2 + 2 \cdot x_3 = 0$

9.7 Maxima and Minima: Functions of n Independent Variables

Chapter 7 dealt initially with the means of finding extreme points for functions of one independent variable, and subsequently with the same consideration for functions of two independent variables. The case(s) of three or

more independent variables were said to flow directly from the one- and two-independent variable cases. This assertion was not formally demonstrated. With the help of matrix algebra, however, we can see how to identify extreme points when we deal with functions that have n independent variables.

9.7.1 First-Order (Necessary) Conditions

Consider a function of the form $z = f(x_1, x_2, \ldots, x_n)$, with the first partial derivatives of the function given by $f_1, f_2, \ldots, f_n$. In order for z to have extreme points, whether maxima or minima, it is necessary for it to be in a "stationary" position. That is, it is necessary that $f_1 = f_2 == f_n = 0$. This is the first-order condition for finding extreme points.

Chapter 7 demonstrated the first-order condition for extreme point(s) graphically for a function of one independent variable. This graphical representation had intuitive attractiveness, for it involved drawing a tangent to the curve of the function at all points at which the slope of that graph was equal to 0.

The analogous geometry for the case of two independent variables involves a three-dimensional diagram. A six-dimensional diagram is needed to illustrate extreme points when five independent variables are involved. In general, we need $n + 1$ dimensions to illustrate the case which involves n independent variables. It is difficult to draw intelligible three-dimensional diagrams; four or more dimensions strain both graphical talents and understanding. For that reason, we cannot illustrate the geometry of maxima and minima where n dimensions are involved.

9.7.2 Second-Order (Sufficient) Condition

We state without proof the second-order condition for an extremum of a function. Given that the first partial derivatives of $z = f(x_1, x_2, \ldots, x_n)$ exist and have been set equal to O for solution purposes, we must find the *Hessian determinant* (or simply *Hessian*) relating to the function. This Hessian determinant of a function $2 = f(x_1, x_2, \ldots, x_n)$ is denoted by $|H|$, and is composed of elements that are second-order partial derivatives of the function

such that

$$|H| = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & \vdots & & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{n} \end{bmatrix}.$$

Once we have found the Hessian determinant, one of the following conditions must hold:

a. When $|H_1|, |H_2|, \ldots, |H_n| > 0$, we have a minimum at the critical point.

b. When $|H_1 < 0$, $|H_2| > 0$, $|H_3| < 0$,..., (*i. e.*, alternating signs) we have a maximum at the critical point.

c. When neither (a) nor (b) is true, the test fails, and we must examine the function in the neighborhood of the critical point in order to determine whether an extreme point exists.

The terms $|H_1|, |H_2|, \ldots, |H_n|$ are *principal minors* of a Hessian determinant.

The Hessian determinant that is used in the second-order condition is a symmetric determinant. That is, the main diagonal of the Hessian consists of all second-order partial derivatives of the function with respect to the variables of the function; for example, $f_{11}, f_{22}, \ldots, f_{nn}$. The off-diagonal elements in the Hessian are composed of all mixed or cross-partial derivatives of the function, for example, f_{12} or f_{36}, where, according to Young's theorem, $f_{ij} = f_{ji}$.

The process for defining the Hessian minors is illustrated here for $|H_1|$, $|H_2|$, and $|H_3|$. The process continues up to and including H_n. (The $||$ indicates determinants, not absolute values.)

$$|H_1| = |f_{11}| = f_{11} \quad |H_2| = \begin{vmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{vmatrix} = \begin{vmatrix} f_{11} & f_{21} \\ f_{12} & f_{22} \end{vmatrix}$$

$$|H_3| = \begin{vmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{vmatrix} = \begin{vmatrix} f_{11} & f_{21} & f_{31} \\ f_{12} & f_{22} & f_{32} \\ f_{13} & f_{23} & f_{33} \end{vmatrix} \ldots$$

and so forth through $|H_n|$.

We can see that these results are consistent with the conditions that Chapter 7 developed for a single variable. In that case the Hessian contains a single element, f_{11}. When $|H_1| > 0$, we have found a local minimum, and when $|H_1| < 0$, we have found a local maximum. This is the same condition that Chapter 7 develped, that $f''(x) > 0$ indicates a minimum and $f''(x) < 0$ indicates a maximum.

Analogously, a function of two independent variables satisfied the second-order condition for a minimum in Chapter 7. If $f_{xx} \cdot f_{yy} - (fxy)^2 > 0$ and $f_{xx}, f_{yy} > 0$ at the critical point. A maximum existed if $f_{xx} \cdot f_{yy} - (f_{xy})^2 > 0$ and $f_{xx}, f_{yy} < 0$ at the critical point. This is precisely what the second-order condition for the Hessian determinants requires:

$$|H_1| = |f_{xx}| = f_{xx} > 0 \quad \text{and} \quad |H_2| = \begin{vmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{vmatrix} = f_{xx} \cdot f_{yy} - (f_{xy})^2 > 0$$

for a minimum. Similarly,

$$|H_1| = |f_{xx}| = f_{xx} < 0 \quad \text{and} \quad |H_2| = \begin{vmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{vmatrix} = f_{xx} \cdot f_{yy} - (f_{xy})^2 > 0$$

are required to establish a maximum.

A function with only one independent variable has only one principal minor. A function of two independent variables has only two principal minors. A function with n independent variables has n principal minors. We must examine each of those principal minors when we seek to determine whether an extreme point exists. If any one of those principal minors is found to have an incorrect sign, then we need go no further with the evaluation process.

It is wise to begin with H_1, proceed to H_2, and so forth. If, for example, we are testing for the existence of a maximum at a critical point, then the signs of the principal minors will alternate, beginning with a negative. If $H_1 < 0$ and $H_2 > 0$, but $H_3 > 0$, then a maximum point may not exist. In this case, the test fails, and we must examine the function in the neighborhood of the critical point in order to determine whether a maximum exists (a process in which a computer algebra system is quite useful). In any case, we need not go beyond $|H_3|$ to determine that the test has failed. Finally, if there are n independent variables, and n is an even number, then the sign of the n^{th} principal minor must be positive if a maximum exists. If n is odd, then the sign of the n^{th} principal minor must be negative for a maximum to exist.

Consider two examples.

1. Find the extreme point(s) for this function: $z = x^2 + x \cdot y + y^2 - 3 \cdot x + 2$.
   1^{st} order conditions: $z_x = 2 \cdot x + y - 3 = 0$ and $z_y = x + 2 \cdot y$.
   2nd order condition: $|H_1| = z_{xx} = 2 > 0$ and $|H_2| = \begin{vmatrix} z_{xx} & z_{xy} \\ z_{yx} & z + yy \end{vmatrix} = \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} = 3 > 0$, so z reaches a minimum point at (2,-1).

2. Find the extreme value(s) for $z = x_1^3 + 3 \cdot x_1 \cdot x_3 + 2 \cdot x_2 - x_2^2 - 3 \cdot x_3^2$.
   $z_1 = -3 \cdot x_1^2 + 3 \cdot x_3 = 0$, $z_2 = 2 - 2 \cdot x_2 = 0$, and $3 \cdot x_1 - 6 \cdot x_3 = 0$.
   1^{st} order conditions: $z_1 = -3 \cdot x_1^2 + 3 \cdot x_3 = 0$, $z_2 = 2 - 2 \cdot x_2 = 0$, and $3 \cdot x_1 - 3 \cdot x_3 = 0$.
   The critical points are (0,1,0) and (1/2,1,1/4).

   2nd order conditions for (1/2,1,1/4):
   $|H_1| = |z_1| = | - 6 \cdot x_1| = -6 \cdot x_1 = -3 < 0$
   $|H_2| = \begin{vmatrix} z_{11} & z_{12|} \\ z_{21} & z_{22} \end{vmatrix} = \begin{vmatrix} -6 \cdot x_1 & 0 \\ 0 & -2 \end{vmatrix} = 12 \cdot x_1 = 6 > 0$
   $|H_3| = \begin{vmatrix} -3 & 0 & 3 \\ 0 & -2 & 0 \\ 3 & 0 & -6 \end{vmatrix} = -18$. The critical value is a maximum.

   2nd order conditions for (0,1,0): $|H_1| = | - 6 \cdot x_2| = 0$.
   The test fails for this critical point. We must, therefore, go back to the function and examine around the neighborhood of the critical point. The critical point is neither a maximum nor a minimum.

Exercise 9.6. Determine the critical points, if any, that correspond to local maxima or minima for the following functions.

1. $z = 2 \cdot x^2 + y^2 - 2 \cdot x \cdot y + 5 \cdot x - 3 \cdot y + 1$

2. $z = 2 \cdot x_1 + x_1 \cdot x_2 + 4 \cdot x_2 + x_1 \cdot x_3 + x_3^2 + 8$

3. $z = 4 \cdot x_1 \cdot x_2 + 3 \cdot x_3 \cdot x_1^2 + x_2 \cdot x_3$

4. $z = x_1^2 + x_2^2 + 8 \cdot x_3^2 - x_1 \cdot x_2 + 10$

5. $z = x_1^2 + x_2^2 + x_3^2 + x_1 \cdot x_2 + x_2 \cdot x_3 - 3 \cdot x_1 - 8$

6. $z = x^3 + y^3 + z^3 - 3 \cdot x \cdot y \cdot z$

7. $z = 5 \cdot x^3 - 2 \cdot x \cdot y + 3 \cdot y^2$

9.8 Optimization: Maxima and Minima Subject to Constraints

We now extend the results of the previous section to deal with the situation in which we wish to identify extreme points in functions that have n independent variables, and in which the functions are subject to one or more constraints. The existence of one or more constraints implies two things: the relevant extreme value cannot be reached, and trade-offs exist. The latter is especially pertinent to economics, for it implies that an objective function z cannot typically be maximized by choosing any x_i level such that $\partial z/\partial x_i = 0$.[11]

9.8.1 The Case of One Constraint

Consider a function of n variables of the form $2 = f(x_1, x_2, \ldots, x_n$ subject to a constraint given by $g(x_1, x_2, \ldots, x_n) = 0$. We now form a new objective function of the form

$$L = L(x_1, x_2, \cdot, x_n, \lambda) = f(x_1, x_2, \ldots, x_n) - \lambda \cdot g(x_1, x_2, \ldots, x_n),$$

where λ is a *Lagrangian multiplier* whose value is to be determined by the maximization (minimization) process.

First-order (necessary) condition

We differentiate the new objective function given in the expression above with respect to the $n + 1$ variables $x_1, x_2, \ldots, x_n$, and λ, set these partial

[11] Dwight Lee, in personal conversation characterizes this result as follows: Anything worth doing is worth doing half-assed.

derivatives equal to 0, and solve for their critical roots. Only critical-root values can be extreme points. However, a critical-root value is not always an extreme point.

We need a second-order (sufficient) condition in order to make a firm judgment.

Second-order (sufficient) condition

Given that the first partial derivatives of L exist and have been set equal to 0 for solution purposes, we must find the **bordered Hessian** determinant relating to the function and its constraints. The bordered Hessian determinant of a function $z = f(x_1, x_2, \ldots, x_n)$, subject to $g(x_1, x_2, \ldots, x_n) = 0$, is denoted by $|H^B|$ and is composed of all second-order partial derivatives of the constraint such that

$$|H^B| = \begin{vmatrix} 0 & g_1 & g_2 & \cdots & g_n \\ g_1 & L_{11} & L_{12} & \cdots & L_{1n} \\ g_2 & L_{21} & L_{22} & \cdots & L_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ g_n & L_{n1} & L_{n2} & \cdots & L_{nn} \end{vmatrix}.$$

If all the first-order partial derivatives of the constraint and all the second-order partial derivatives of the function L exist at the critical point(s), then one of the following conditions must hold.
(a) When $|H_2^B|, |H_3^B|, \ldots, |H_n^B| < 0$, we have a minimum at the critical point.
(b) When $[H_2^B > 0|, |H_3^B l < 0, |F_4^B| > 0, \ldots$, we have a maximum at the critical point.
(c) When neither (a) nor (b) is met, the test fails, and we must examine the function in the neighborhood around the critical point in order to determine whether a constrained extreme value exists.

A bordered Hessian determinant is a symmetric determinant. It is simply a Hessian determinant that is bordered by the first partial derivatives of the constraint, and 0. The symmetry follows from the fact that a Hessian determinant, which is the major part of a bordered Hessian determinant, is also symmetric.

It is customary to denote the i^{th} bordered principal minor of a bordered Hessian determinant by the symbol $|H_i^B|$. Among the bordered principal

minors of the bordered Hessian are the following:

$$|H_2^B| = \begin{vmatrix} 0 & g_1 & g_2 \\ g_1 & L_{11} & L_{12} \\ g_2 & L_{21} & L_{22} \end{vmatrix} \quad |H_3^B| = \begin{vmatrix} 0 & g_1 & g_2 & g_3 \\ g_1 & L_{11} & L_{12} & L_{13} \\ g_2 & L_{21} & L_{22} & L_{23} \\ g_3 & L_{31} & L_{32} & L_{33} \end{vmatrix}.$$

The notation $|H_2^B|$ means that we must take the Hessian determinant $|H_2|$ and place around it the appropriate border. $|H_2^B|$ does not mean that we have a 2×2 determinant.

We should note carefully that the process of evaluating bordered Hessian determinants begins with $|H_2^B|$, not with $|H_1^B|$ We do not evaluate $|H_1^B|$ when we maximize or minimize subject to a single constraint. We shall shortly state a general rule that deals with this situation.

Consider two examples.

1. Find the maximum or minimum value(s) for the function $z = x_1^2 - 10 \cdot x_2^2$, subject to this constraint: $x_1 - x_2 = 18$.
   Our Lagrangian expression is $L = z = x_1^2 - 10 \cdot x_2^2 + \lambda \cdot (x_1 - x_2 - 18)$. The first-order conditions are $2 \cdot x_1 + \lambda = 0$, $-20 \cdot x_2 - \lambda = 0$, and $x_1 - x_2 - 18 = 0$, so the critical values are $x_1 = 20$, $x_2 = 2$, and $\lambda = -40$.

   Our second-order condition is

   $$|H_2^B| = \begin{vmatrix} 0 & 1 & -1 \\ 1 & 2 & 0 \\ -1 & 0 & -20 \end{vmatrix} = 18 > 0,$$

   so the constrained *maximum* value of z occurs when $x_1 = 20$ and $x_2 = 2$.

2. Find the values of x, y and z that yield a maximum or or minimum value of $w = 5 \cdot x^2 + 10 \cdot y^2 + z^2 - 4 \cdot x \cdot y - 2 \cdot x \cdot z - 36 \cdot y$.

   The first order conditions are
   $-2 \cdot z - 4 \cdot y + 10 \cdot x + \mu = 0$
   $20 \cdot y - 4 \cdot x + 2\mu - 36 = 0$
   $2 \cdot z - 2 \cdot x + 4\mu = 0$
   $4 \cdot z + 2 \cdot y + x - 12.$

The solution to the first order condition is

$$x = \frac{58}{49}, y = \frac{101}{49}, z = \frac{82}{49}, \mu = -\frac{12}{49}.$$

The second order conditions are

$$|H_2^B| = \begin{vmatrix} 0 & 1 & 2 \\ 1 & 10 & -4 \\ 2 & -4 & 20 \end{vmatrix} = -76$$

and

$$|H_3^B| = \begin{vmatrix} 0 & 1 & 2 & 4 \\ 1 & 10 & -4 & -2 \\ 2 & -4 & 20 & 0 \\ 4 & -2 & 0 & 2 \end{vmatrix} = -3528.$$

Thus, we determine that a constrained minimum point on w occurs at the critical values of x, y, and z.

9.8.2 The Case of Two or More Constraints

The problem of maximizing or minimizing a function of n independent variables subject to two or more constraints is analogous to the situation in the preceding section, in which there was only one constraint. In terms of a bordered Hessian determinant, we simply add an additional border for each effective constraint.

Consider a function of the form $z = f(x_1, x_2, \ldots, x_n)$, which is subject to m constraints ($m < n$) given by $g(x_1, x_2, \ldots, x_n) = 0$, $h(x_1, x_2, \ldots, x_n) = 0$, $\ldots$, $k(x_1, x_2, \ldots, x_n) = 0$. In order to find any extreme points that might exist, we form a new objective function of the form $L = L(x_1, x_2, \ldots, x_n, \lambda_1, \lambda_2, \ldots, \lambda_n)$ $= f(x_1, x_2, \ldots, x_n) + \lambda_1 \cdot g(x_1, x_2, \ldots, x_n) + \lambda_2 \cdot h(x_1, x_2, \ldots, x_n) + \ldots, \lambda_n \cdot k(x_1, x_2, \ldots, x_n)$, where the λ's are Lagrange multipliers.

First-Order (Necessary) Condition

This new objective function, L, is differentiated with respect to each of the $n + m$ variables $(x_1, x_2, \ldots, x_n, \lambda_1, \lambda_2, \ldots, \lambda_m)$. The resulting partial derivatives are set equal to 0 and solved for critical points. These critical points may or may not identify actual extreme points. We must apply a second-order (sufficient) condition to identify those critical points that are, indeed, extreme points.

Second-Order (Sufficient) Condition
Given that the first partial derivatives of L exist and have been set equal to zero for solution purposes, we must find the appropriate bordered Hessian determinant. In this case, the bordered Hessian is given by

$$|H^B| = \begin{vmatrix} 0 & \cdots & \cdots & 0 & k_1 & k_2 & \cdots & k_n \\ \vdots & & & \vdots & \vdots & \vdots & & \vdots \\ \cdot & \cdots & 0 & 0 & h_1 & h_2 & \cdots & h_n \\ 0 & \cdots & 0 & 0 & g_1 & g_2 & \cdots & g_n \\ k_1 & \cdots & h_1 & g_1 & L_{11} & L_{12} & \cdots & L_{1n} \\ k_2 & \cdots & h_2 & g_2 & L_{21} & L_{22} & \cdots & L_{2n} \\ \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\ k_n & \cdots & h_n & g_n & L_{n1} & L_{n2} & \cdots & L_{nn} \end{vmatrix}.$$

If all the first-order partial derivatives of the constraints and all the second-order partial derivatives of the function L exist at the critical point(s), then one of the following conditions must hold:

(a) When $|H^B_{m+1}|, |H^B_{m+2}|, \ldots, |H^B_{m+n}|$ all have the same sign, namely $(-1)^n$, we have a constrained minimum at the critical points.

(b) When $|H^B_{m+1}|, |H^B_{m+2}|, \ldots, |H^B_{m+n}|$ alternate in sign, where $|H^B_{m+1}|$ has the sign $(-1)^{m+1}$, we have a constrained maximum at the critical point.

(c) When the requirements of neither (a) nor (b) aremet, the test fails and we must examine the function in the neighborhood around the critical point in order to determine whether a constrained extremum exists.

The practice of beginning the evaluation of the bordered Hessians with something other than $|H_1|$ continues. The rule that guides this behavior requires that we begin with a bordered Hessian whose size is one bigger than the number of constraints. Hence, when m constraints exist, we begin our analysis of the bordered Hessians with $|H^B_{m+1}|$. We can now look back to our previous work and explain our previous choices in this regard. When $m = 0$ and no constraint exists, we begin with $|H_1|$ When $m = 1$, we begin with $|H^B_2|$ and so on.

The signs that are required for the successive bordered Hessian determinants follow a definite order. When we evaluate critical point(s) with respect to

a maximum, the bordered Hessians must alternate in sign. In the case in which two constraints exist, the sign of $|H^+m+1| = |H_3^B|$ must be negative, the sign of $|H_4^B|$ must be positive, and so forth. The rule is that the sign is given by $(-1)^{m+1}$. Hence, if $m = 2$, the sign of $|H_3^B|$ is $(-1)^3 = -1$, and $|H_3^B|$ is negative.

The sign determination for the case of a constrained minimum differs from that of a constrained maximum. When $m = 0$, all bordered Hessians must be positive. When $m = 1$, all bordered Hessians must be negative. When $m = 2$, all bordered Hessians must once again be positive. In general, the sign of all bordered Hessians must be the same as the sign of $(-1)^m$ if a minimum exists.

As an example, suppose that five constraints apply. Then all bordered Hessian determinants must have a negative sign, for $(-1)^5 = -1$, which is negative. We can see that when the number of constraints is odd, the signs must all be negative, whereas when the number of constraints is even, the signs must all be positive.

Exercise 8.7. Use Lagrangian multipliers to maximize or minimize the following functions subject to the indicated constraints.

1. $z = 2 \cdot x_1 + x_2 + 2 \cdot x_1 \cdot x_2$, subject to the constraint $2 \cdot x_1 + x_2 = 100$.
2. $z = 25 \cdot x_1 \cdot x_2 \cdot x_3$, subject to the constraint $x_1 + 2 \cdot x_2 + 4 \cdot x_3 = 180$.
3. $w = log_e(x) + log_e(y) + log_e(z)$, subject to the constraint $5 \cdot x + 2 \cdot y + z = 120$.
4. $z = x_1^2 + 4 \cdot x_2^2 + x_3^2 - 4 \cdot x_1 \cdot x_2 - 6 \cdot x_3$, subject to the constraint $x_1 + x_2 + x_3 = 21$

9.9 Questions and Problems

1. Write the following system of equations in matrix notation.
$$\begin{array}{lllll} a_{11}\cdot x_1+ & a_{12}\cdot x_2+ & a_{13}\cdot x_3+ & \cdots & +a_{1n}\cdot x_n = K_1 \\ a_{21}\cdot x_1+ & a_{22}\cdot x_2+ & a_{23}\cdot x_3+ & \cdots & +a_{2n}\cdot x_n = K_2 \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1}\cdot x_1+ & a_{n2}\cdot x_2 & a_{n3}\cdot x_3+ & \cdots & +a_{nn}\cdot x_n = K_n \end{array}$$

2. Given the following national income equations:

   1) (1) $Y = C + I + G$,

   (2) C=$200 + 0.7 \cdot Y$,

   (3) $I = 75 + 0.1 \cdot Y$,

   (4) $G = 100$

   (a) Find the equilibrium values of C, I, G, and Y.

   (b) What is the numeric value of the multiplier that we would use to determine the effects of autonomous changes in I on Y?

   (c) Suppose that autonomous investment increases from 75 to 100. What increase in Y results? (Show this result algebraically and also by means of matrix algebra.)

   (d) Demonstrate that your numeric solutions to parts (a) and (c) are general equilibrium solutions in the sense that they are internally consistent with each other in magnitude.

3. Find the equilibrium prices (P_i) and quantities (Q_i)for goods A, B, and C using inverse matrix algebra. Then use *Maxima* to check your answer.
   $Q_{D,A} = 8 - 2 \cdot P_A + 3 \cdot P_B - P_C \quad Q_{D,B} = 4 - 4 \cdot P_B + P_A + 3 \cdot P_C$
   $Q_{D,C} = 6 - P_C + 3 \cdot P_A + 3 \cdot P_B$
   $Q_{S.A} = 10 \quad Q_{S,B} = 2 \cdot P_A + 2 \quad Q_{S,C} = 8 + P_C$

4. Find the equilibrium position of a consumer, given the following utility function and budget information:
   $U = log(x) + log(y) + log(z)$
   $M =$ \$120 $\quad P_x =$ \$4 $\quad P_y =$ \$2 $\quad P_z =$4
   M is money income per period, (x, y, z) is the per-period consumption

levels of the three goods, and U is the utility level. The prices are P_x, P_y, and P_z.

Chapter 10

Linear Programming and Input-Output Analysis

Matrix algebra has several important uses beyond those demonstrated in Chapter 9. In particular, matrix algebra underlies two powerful quantitative techniques, linear programming and input-output analysis. Indeed, these are just extension of matrix algebra, but we devote a separate chapter to them only as a means of focusing attention on their importance to economists.

We begin by considering linear programming in a matrix algebra framework, and subsequently examine input-output analysis in the same fashion.

10.1 Linear Programming

Linear programming is a mathematical technique used to derive economically efficient solutions to problems that arise in a wide range of situations. Linear programming always involves the maximization or minimization of some function, subject to various constraints. For example, a school district may wish to minimize the cost of busing students from one location to another, with the students' initial locations considered as given. A firm might wish to maximize its output given that it faces certain input prices and has a finite limit on how much money it can spend on those inputs. Or a government agency may seek to minimize the cost of feeding, clothing, and equipping

an infantry division, subject to constraints concerning the location of the resources used and how much of each resource is required.

Linear programming is a mathematical technique whereby one maximizes or minimizes a linear function subject to various constraints, or side conditions, which are stated in the form of linear inequalities.

10.1.1 Stigler's Diet Problem Once Again

Chapter 1 considered a classic economic allocation exercise, Stigler's diet problem. We now recast the diet problem as a linear-programming problem that uses matrix algebra. The crux of the diet problem is to find the least expensive combination of foods available to consumers that will allow these same consumers to satisfy recommended daily dietary allowances established by the Food and Nutrition Board of the National Academy of Sciences. That is, Stigler's objective was to find the least expensive menu that would give the consumer required minimum levels of calories, vitamins, and so forth.[1]

Stigler allowed the consumer to choose a menu from among 80 possible foods. Let X_j refer to the quantity of food j that an individual consumes per time period. We can then represent the quantities of the 80 different foods consumed by the individual as $X_1, X_2, \ldots, X_{30}$. The X_j, with $j = 1, 2, \ldots, 80$, are decision variables, or unknowns, that are to be determined by solving the linear- programming problem.

Any attempt to determine the least expensive way to accomplish a given end must necessarily consider the prices of various alternatives. Let P_j represent the dollar price of food type j. Thus the prices of the 80 different foods that can be selected are represented by $P_1, P_2, \cdots, P_{80}$. For example, P_1 might be \$0.02 and represent the price of peanut butter per ounce. It follows that $P_1 \cdot X_1$, the price of peanut butter per ounce times the number of ounces purchased, is the consumer's total expenditure on peanut butter.

The equation below defines the total expenditure C that the consumer makes on the 80 different foods. This equation is referred to as the objective function; it is this function that we seek to minimize as a part of Stigler's diet

[1]Recall that, as a matter of historical accuracy, the formal technique of linear program was invented a few years after Stigler's research. Stigler used a series of approximations to derive a result that direct application of linear program would have yielded more easily.

problem. This objective function is *linear* in the X_j's, which we call the *decision variables.*Thus, our objective function is $C = \sum_1^{80} P_j \cdot X_j$.[2]

An obvious way to minimize C is to not spend any money on food, but this is not permissible, for such a menu plan would not satisfy the recommended daily dietary allowances, the constraints. More bluntly, it would kill the consumer. Let R symbolize a dietary requirement. Thus R_1 might represent the recommended intake of calories per individual per day. In Stigler's 1945 version of the diet problem, the recommended intake of calories per individual per day was 3000. Fewer calories would presumably be detrimental to the consumer's health. We represent the nine different dietary requirements of Stigler's problem by the variable names $R_1, R_2, \ldots, R_9$.

The diet problem is inherently challenging because two different foods seldom yield the same amount of nutrient per ounce of food. For example, in Stigler's 1945 exercise, 1 ounce of uncooked bacon yielded 186 calories, whereas 1 ounce of uncooked sirloin steak yielded only 88 calories. Let the symbol a_{ij} represent the number of units of nutrient i that are provided by 1 ounce of food j. Hence the a_{ij} for uncooked bacon (in terms of calories) is 186, while the analogous a_{ij} for uncooked sirloin steak is 88.

A consumer can satisfy a nutrient requirement by eating many different foods. The term $a_{ij} \cdot X_j$ represents the total number of units of nutrient i that are obtained when one consumes a given number of ounces of food j. For example, if one consumes 6 ounces of uncooked bacon, then $X_j = 6$, and since $a_{ij} = 186$, $a_{ij} \cdot X_j = 186 \cdot 6) = 1116$ calories.

We have previously noted that Stigler assumed that the individual must obtain at least 3000 calories per day from the foods consumed. The consumer can choose among the 80 foods in order to satisfy this requirement. We can write this constraint as follows, stating that the sum of all the calories the consumer derives from consuming various foods must be 3000 or greater: $a_{1,1} \cdot X_1 + a_{1,2} \cdot X_2 + \cdots + a_{1,80} \cdot X_{80} \geq 3000$.

[2]Nonlinear objective functions are not permissible in a linear programming problem. If the researcher attempts to represent the underlying phenomenon by a linear equation when it is actually nonlinear, then the results obtained will be inaccurate and unreliable. Fortunately, nonlinear programming techniques are now accessible, given advances in mathematics and in computing power. Even spreadsheet programs like *Excel* typically offer ways to address nonlinear systems. We illustrate nonlinear programming at the end of this section.

The nine daily dietary requirements that Stigler imposed on the consumer in his diet problem were reported in Chapter 1. Each of these nine requirements constitutes a constraint on the consumer's activities that takes the form of a linear inequality similar to the expression above that addresses caloric intake. The nine constraints are

$$\begin{array}{cccccc} a_{1,1} \cdot X_1 + & a_{1,2} \cdot X_2 + & \cdots & +a_{1,80} \cdot X_{80} & \geq R_1 \\ a_{2,1} \cdot X_2 + & a_{2,2} \cdot X_2 + & \cdots & +a_{2,80} \cdot X_{80} & \geq R_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{9,1} \cdot X_9 + & a_{9,1} \cdot X_2 + & \cdots & +a_{9,80} \cdot X_{80} & \geq R_9 \end{array}$$

The constraints or requirements n this set of equations are, like the objective function, linear in the decision variables X_j.

We must include one additional, seemingly obvious set of constraints, that $X_j \geq 0$ for all $j = 1, 2, \ldots, 80$. These restrictions are known as *nonnegativity constraints*. They explicitly restrict the solution values of the decision variables to be either zero or positive, thus eliminating the possibility of a nonsensical solution that might (for example) allow the individual to consume -5 ounces of bacon.

A formal expression of the problem at hand is that we must minimize our objective function C subject to the nine inequalities and the eighty nonnegativity conditions being satisfied.

We now show a variant of the Stigler solution, which also appears in Chapter 1. For purposes of exposition, Stigler focused on a subset of five foods and eight constraints. The solution to this subset yielded results that were close to that of the larger model. The problem consists of the following nine equations, with the first on being the objective function.

$$x5 + x4 + x3 + x2 + x1 \tag{z}$$

$$26.9x5 + 1.1x4 + 2.6x3 + 8.4x2 + 44.7x1 >= 3 \tag{c1}$$

$$1691x5 + 106x4 + 125x3 + 422x2 + 1411x1 >= 70 \tag{c2}$$

$$11.4x5 + 4x3 + 15.1x2 + 2x1 >= 0.8 \tag{c3}$$

$$792x5 + 138x4 + 36x3 + 9x2 + 365x1 >= 12 \tag{c4}$$

$$918.4x4 + 7.2x3 + 26x2 >= 5 \tag{c4}$$

$$38.4x5 + 5.7x4 + 9x3 + 3x2 + 55.4x1 >= 1.8 \quad \text{(c5)}$$

$$24.6x5 + 13.8x4 + 4.5x3 + 23.5x2 + 33.3x1 >= 2.7 \quad \text{(c6)}$$

$$217x5 + 33x4 + 26x3 + 11x2 + 441x1 >= 18 \quad \text{(c7)}$$

$$2755x4 + 5369x3 + 60x2 >= 75. \quad \text{(c8)}$$

The names of the expressions are in at the right. The expression z is the objective (cost) function, for which we seek a minimum value. The expressions $c_1 \cdots c_8$ are the constraints. The next set of commands loads the `simplex` program (more on this below) and enters the relevant command, which identifies the objective function, the eight constraints, and the nonnegativity conditions: `load(simplex)$ minimize_lp(z, [c1, c2, c3, c4, c5, c6, c7, c8, x1>=0, x2>=0, x3>=0, x4>=0, x5>=0])`.

The model yields the output in a list. The first item in the list is the value of z. The second item in the list is an embedded list of five x values:

$$[0.10904, [x5 = 0.048628, x4 = 0.0051128, x3 = 0.01125,$$

$$x2 = 0.0085915, x1 = 0.035456]].$$

The results show that the cost is \$0.10904 (which would correspond to about \$0.80 in 2016 dollars).

10.1.2 A Formal Definition of a Linear-Programming Model

The previous example, which stated Stigler's diet problem as a linear-programming problem, exhibits the three fundamental properties of any linear-programming problem:

1. The object of the problem is to find optimal values for the decision variables or unknowns in the problem.

2. The optimal values of the decision variables are such that they either minimize or maximize an explicit linear objective function.

3. The minimization or maximization solution of the objective function must be feasible. That is, the values of the decision variables in the optimal solution must satisfy both the linear inequality constraints and the nonnegativity constraints.

Our general linear-programming model has the following structure:]

Maximize or Minimize:
Z=$b_1 \cdot X_1 + b_2 \cdot X_2 + \ldots + b_n \cdot X_n$

Subject to:

$$\begin{array}{llll} a_{11} \cdot X_1 + & a_{12} \cdot X_2 + & \cdots & +a_{1n} \cdot X_n \quad (\leq, =\geq) \quad c_1 \\ a_{21} \cdot X_2 + & a_{22} \cdot X_2 + & \cdots & +a_{2n} \cdot X_n \quad (\leq, =\geq) \quad c_2 \\ & & & \vdots \qquad\qquad\quad \vdots \\ a_{m1} \cdot X_n + & a_{m2} \cdot X_2 + & \cdots & +a_{mn} \cdot X_n \quad (\leq, =\geq) \quad c_m \end{array}$$

and
$X_j \geq 0$ for all $j = 1, 2, \ldots, n$,
where n = the number of decision variables, and m = the number of constraints or side conditions.

Each of the decision variables (X_j's) must appear in the objective function of a linear-programming problem, Z. Only one sign ($\leq, =$, or $\geq$) can hold in any particular constraint. The b_j, a_{ij}, and c_i are the parameters of the model. The c_i's are the amounts of scarce resource i available for allocation ($i = 1, 2, \ldots, m$). The parameter a_{ij} represents the amount of resource i that is consumed by, or allocated to, each unit of decision variable j (for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$). The change in Z that results from a unit increase in X_j is represented by b_j. For example, if $b_j = -5$, then a one-unit increase in X_j decreases Z by five units.

We can write the general linear-programming model more compactly using matrix notation as follows:
Maximize or Minimize: $Z = B \cdot X$
Subject to:
$A \cdot X$ ($\leq, =\geq$), C and $X \geq 0$, where,

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, B = \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix}, C = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix}, A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

10.1.3 A Graphical Illustration

Although a graphical approach to solving linear programming problems is seldom very efficient in terms of either time or effort, the graphical approach

does offer a clear understanding of what linear programming actually involves. We use the graphical approach in this section to solve a relatively simple linear-programming problem. This should enable you to visualize what is happening when you encounter a more complex linear-programming solution.

We demonstrate the graphical approach to linear programming by a problem concerning a firm, Acme Manufacturing Company, that produces two products, X_1 and X_2. Acme is a price-taker in the market for both goods. The prices are \$20 per unit for X_1 and \$15 per unit for X_2.

Acme's production is limited by three resource constraints. Producing both X_1 and X_2 requires three inputs, which we label a, b, and c. Thus, Acme has two production functions: $X_1 = X_1(a, b, c)$ and $X_2 = X_2(a, b, c)$. For the relevant production period, Acme has only 60 units of input a, 24 units of input b, and 84 units of input c. Acme cannot augment the quantities of these inputs during the current production period.

For each unit of good X_1 that it produces, Acme uses 5 units of input a, 3 units of input b, and 12 units of input c. The production of a single unit of good X_2 requires 15 units of input a, 4 units of input b, and 7 units of input c. Acme cannot alter these input-output relationships during the current production period. It need not use all of any one of the inputs.

Acme's objective is to maximize the revenue R that it receives from the sale of goods X_1 and X_2, while at the same time not violating any resource constraint that it faces. We can state this problem in linear-programming terms as follows:

Maximize: $R = P_1 \cdot X_1 + P_2 \cdot X_2 = 20 \cdot X_1 1 + 15 \cdot X_2$,
subject to:
$5X1 + 15X2 \leq 60$ (input constraint a),
$3 \cdot X_1 + 4 \cdot X_2 \leq 24$ (input constraint b),
$12 \cdot X_1 + 7 \cdot X_2 \leq 84$ (input constraint c), and
$X_1, X_2 \geq 0$ (nonnegativity constraints).

Once we have formulated the linear programming problem, the next step in the graphical approach is to delineate the solution space. We do this by graphing the constraints (as shown in Figure 10.1). The constraints, when graphed, enclose a set of feasible (possible) solutions that constitute the *solution space*. Linear programming selects the point within that solution space,

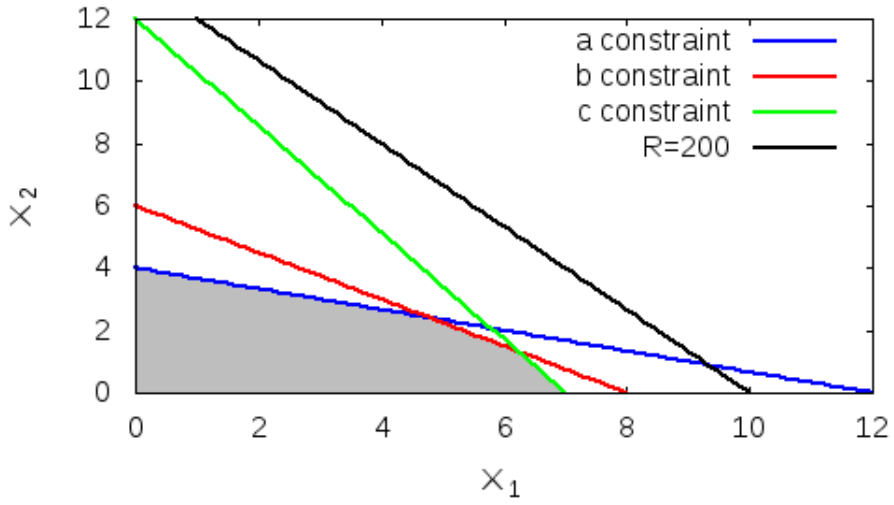


Figure 10.1: Graphic representation of the Acme production problem

or *feasible region*, that maximizes (minimizes) the value of the objective function.

If we ignore, for the time being, the less-than sign in the linear constraints of Acme's linear programming problem, and therefore treat these equations as if they had equals signs then we can begin to provide a visual representation of the linear-programming problem. From Chapter 2, we know that we can graph a straight line without great difficulty if we are given two points on that line. For example, if we know the horizontal intercept (abscissa) and vertical intercept (ordinate) of a line, then we can connect these two intercepts with a straight line and obtain the needed graph. We use this technique to define the solution space for Acme. Figure 10.1 illustrates this procedure.

Figure 10.1 also shows an isorevenue line (for $R = \$200$–any value will do). This line lies outside the feasible region, so this level of revenue cannot be attained.

The shaded area in Figure 10.1 has four corners (not counting the origin, where $R = 0$). Two of the points are *interior* and involve the production of positive amounts of both products. Two points are on the axes and involve the production of just one of the two goods. Figure 10.2 replicates Figure 10.1 and focuses on the feasible region.

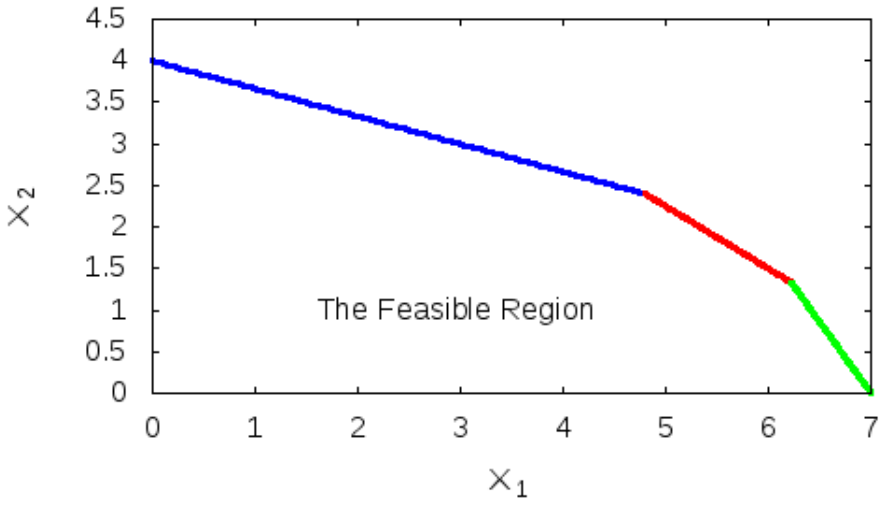


Figure 10.2: The feasible region of the Acme problem

The accompanying workbook shows that the interior solutions occur at (24/5, 12/5) and (56/9,4/3), or approximately (4.8, 2.4) and (6.22, 1.33).

An infinite number of points lie either within or on the boundary of feasible region. This means that an infinite number of possibilities confront us when we attempt to identify the solution that is optimal. To illustrate the search technique that linear programming performs, we consider a few specific points within the solution space. At the origin, as we have seen, $R = 0$. Acme can, however, produce positive quantities of goods X_1 and X_2 that generate some sales revenue, so the solution at the origin is not optimal. We can do better.

We assert that any movement from the origin that involves production of more of one good and no less of the other adds to revenue. Such moves are possible whenever the (X_1, X_2) combination is inside the bondary of the feasible region. Therefore, any point inside the solution space represents less production (and therefore less total sales revenue) than at least one point on the boundary of the solution space. Thus, the optimal solution lies on the boundary of the solution space, not inside it.

When we are considering points as candidates for the optimal solution, we can ignore any point inside the solution space. Knowledge of this fact substantially reduces the number of possible solutions with which we must con-

tend. We proceed therefore to consider only those solution points that are on the boundary. We begin with the values, $X_1 = 7, X_2 = 0$ which yields $R = \$140$. From this point, we move counter-clockwise to (6.22,1.33), at which $R = \$144.35$. The next move, to (4.8, 2.4) generates revenue of \$132.00. Finally, when no X_1 is produced and $X_2 = 4$, the revenue level falls to \$60. Therefore, among the corners, the largest revenue occurs where $X_1 = 6.22$ and $X_2 = 1.33$.

One of the corner points of the solution space is always an optimal solution to a linear-programming problem, so we are justified in ignoring the linear segments between these corners. This assertion may not be intuitively obvious, however, and requires more explanation.

Consider Figure 10.3, which graphs the constraints of the Acme linear programming problem. The solution space (feasible region) is indicated by the black border. Graphs that correspond to five values of the objective function, $R = \$20X_1 + \$15X_2$ also appear. Four of the five revenue levels are feasible; they are the values calculated above. For the highest of these five revenue values, the iso-revenue line is tangent to the feasible region boundary. Thus, a single value on that line is feasible. For $R > \$144.44$, no point is in the feasible region. Thus, the iso-revenue line for $R = \$160$, shown in yellow, cannot be attained.

In general, a corner of the solution space will always be an optimal solution to a linear programming problem. Only when the slope of the objective function is the same as the slope of a binding constraint will there be more than one optimal solution. In the unlikely event that the slope of the objective function in Figure 10.3 were the same as the slope of any one of the three line segments, then all the points on that line segment (including the corner points) would be optimal. Thus, the value of the objective function at the corner point is at least as high (or low, if the program involves minimization) as the points on the line segment, so only corner points need be considered.

An efficient linear-programming solution technique is ordinarily used to identify the corner points in a problem, evaluate them, and select the optimal solution from among those possibilities. The *simplex algorithm*, one such technique that is frequently used, is the one that *Maxima* used above to solve the subset of the Stigler diet problem. This algorithm is a search technique that identifies and evaluates the corner points in a problem. It repeatedly strives to find a better solution than the one at hand. When it reaches an

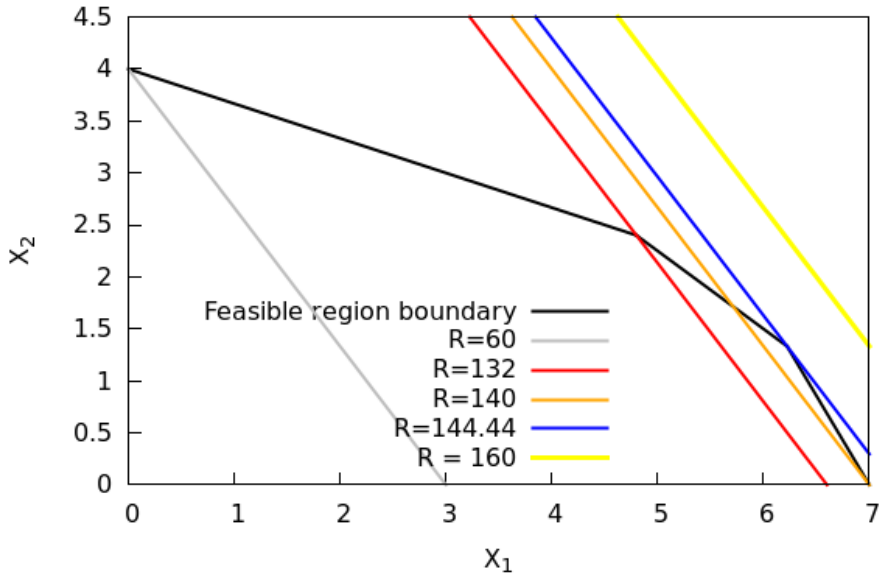


Figure 10.3: The feasible region and revenue levels for Acme

optimal point, such as the one that we identified in the Acme problem, it stops. A movement to any other corner would result in a worse solution.

The table below shows the objective function R and the three constraints.

$$\begin{bmatrix} R & 15X2 + 20X1 \\ C1 & 15X2 + 5X1 \leq 60 \\ C2 & 4X2 + 3X1 \leq 24 \\ C3 & 7X2 + 12X1 \leq 84 \end{bmatrix}$$

The following commands load the simplex module and execute the command to execute the linear programming tool: `load(simplex)$` `float(maximize_lp( R, [C1,C2,C3,X1>=0,X2>=0]));`. The `maximize_lp` command is embedded in a `float` command to generate easily-interpreted values rather than exact values. It is optional. The results are consistent with our computations above: [**144.44,[X2=1.3333,X1=6.2222**]]. The first item that the command reports is the value of the objective function. The second item is a list of the two production levels.

10.1.4 The Dual Problem

We can actually view every linear-programming problem as consisting of two separate problems. The original linear-programming problem is called the *primal* problem, while a second formulation of this original problem is known as the *dual* problem. The dual problem in linear programming often yields results that are quite useful to analyst.

Also, the dual problem is sometimes easier to solve than the original (primal) problem. Therefore, when the primal problem is intractable or just difficult to solve, we can sometimes solve the dual problem, then use the information from that dual to solve the primal problem.

Suppose that the original (primal) linear-programming problem is this:

Maximize: $Z = b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_n \cdot X_n$

Subject to m constraints:
$a_{11} \cdot X_1 + a_{12} \cdot X_2 + \cdots + a_{1n} \cdot X_n \leq C_1$
$a_{11} \cdot X_1 + a_{12} \cdot X_2 + \cdots + a_{1n} \cdot X_n \leq C_1$
$\vdots \quad \vdots \quad \vdots \quad \vdots$
$a_{11} \cdot X_1 + a_{12} \cdot X_2 + \cdots + a_{1n} \cdot X_n \leq C_1$
and
$X_j \geq 0$ for all $j = 1, 2, \ldots, n$.

The dual problem associated with this primal problem requires minimization. The problem is this:

Minimize: $W = d_1 \cdot Y_1 + d_2 \cdot Y_2 + \cdots + d_m \cdot Y_m$,
subject to

$$\begin{array}{cccc} a_{11} \cdot Y_1 + & a_{21} \cdot Y_2 + & \cdots & + a_{m1} \cdot Y_m \geq b_1 \\ a_{22} \cdot Y_2 + & a_{22} \cdot Y_2 + & \cdots & + a_{m2} \cdot Y_m \geq b_2 \\ \vdots & \vdots & & \vdots \quad \vdots \\ a_{1n} \cdot Y_1 + & a_{2n} \cdot Y_2 + & \cdots & + a_{mn} \cdot Y_m \geq b_n \end{array}$$

where $Y_1, Y_2, \ldots, Y_m$ are the dual variables.

A definite symmetry exists between a primal problem and its dual. When the primal problem involves the maximization of a function, then the dual problem involves the minimization of some function. When the constraints on the primal objective function require the firm to arrange its activities so that its input use and expenditures are less than or equal to certain constants, the constraints on the dual objective function require that certain of the

firm's activities be *equal to or exceed* certain constant levels. For example, if the primal problem is to maximize output given a cost constraint of $200, then the dual problem is to minimize the cost of producing a certain level of output, perhaps 10 units. We shall be more specific about the symmetry between the primal and dual linear-programming problems in a moment.

One of the most interesting results of many dual problems in linear programming is the idea of a shadow price. Decision-makers often wish to estimate the value of contributions that various inputs make to the optimal solution. On occasion, not all the available units of a particular input are used in the optimal solution. In such a circumstance, the decision-maker has little need for additional units of such inputs. Therefore the shadow price of such inputs is 0, because an additional unit of such an input would not alter the optimal solution. For example, the value of an extra seat in a classroom that already has many empty seats is 0. Similarly, the value of a second textbook to a student who already has one is 0 unless the student loses the first textbook.

Thus, a *shadow price* indicates the value at the margin that an input has for the objective function's activities.

We can, in many situations, interpret the shadow price of an input as the value of the marginal product of that input. This is an interesting and quite useful result, particularly in decision situations in which the input in question is not purchased in the market, or in which the connection between that input and the ultimate output seems quite distant at best. The shadow price in this case indicates the price that the decision-maker would be willing to pay for additional units of this input.

Economic planners in nonmarket economies such as the former Soviet Union have made considerable use of the shadow prices of linear programming. Since market prices often did not exist in the Soviet Union, it was difficult for planners to cost, price, and value things efficiently. Shadow prices provided some guidance in the absence of market signals.

Similarly, large organizations (the military services are a prime example) that do not vend their wares in a conventional fashion can use shadow-pricing techniques to increase the efficiency of many activities, such as purchasing, routing, and intra-branch transfers.

The Symmetry between the Primal and the Dual Problems

1. If the primal problem involves maximization, then the dual problem involves minimization, and *vice versa*.

2. If the primal problem involves $\leq$ constraints, then the dual problem involves $\geq$ constraints.

3. The coefficients of the variables in the primal objective function are the *right-hand constants* of the constraint equations in the dual problem.

4. The coefficients of the variables in the dual objective function are the *right-hand constants* of the constraint equations in the primal problem.

5. A new set of variables Y_i appears in the dual objective function and constraint equations. These Y_i values are the shadow prices of the inputs.

6. If the primal problem consists of n decision variables and m constraint equations, then the dual problem consists of m variables and n constraints.

7. The *coefficients* of the constraint equations in the primal problem are the same as the coefficients of the constraint equations in the dual problem except that the rows and columns are interchanged. That is, each a_{ij} now becomes afi. In matrix notation,
   $A = [a_{ij}]$, for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$
   in the primal problem, then $A' = [a_{ji}]$ is associated with the dual problem.

8. The nonnegativity constraints apply to all variables in both the primal and the dual problems.

9. The optimal solution is identical for both the primal and the dual problems.

Examples

1. Given that the primal problem is the Acme Manufacturing situation is *maximize*: $R = 20 \cdot X_1 + 15 \cdot X_2$

subject to:
$5 \cdot X_1 + 15 \cdot X_2 \leq 60$
$3 \cdot X_1 + 4 \cdot X_2 \leq 24$
$12 \cdot X_1 + 7 \cdot X_2 \leq 84$ and
$X1.X2 \geq 0$,
then the dual problem is
minimize: $C = 60v_a + 24 \cdot v_b + 84 \cdot v_c$
subject to:
$5 \cdot v_a + 3 \cdot v_b + 12 \cdot v_c \geq 20$
(The "value" to the firm of selling a unit of X_1 is attributed to the inputs, which could be used in producing X_2. Likewise for X_2 below.)
$15 \cdot v_a + 4 \cdot v_b + 7 \cdot v_c \geq 20$
$v_a, v_b \geq 0$
The optimal solution is: $X_1 = 6.22$, $X_2 = 1.33$ $R = 144.35$ $v_a = 0$, $v_b = 1.47$, and $v_c = 1.30$
The v's are shadow prices. The objective is to minimize cost, C, based on these shadow prices. The value of v_a is 0, because units of input a remain unused after the optimal solution has been implemented.

Here is the relevant information in the form of *Maxima* commands: `Cost: 60*va + 24*vb + 84*vc; Ca: 5*va + 3*vb +12*vc >= 20; Cb: 15*va + 4*vb + 7*vc >= 15;`. The command `minimize_lp( Cost, [Ca,Cb,va>=0, vb>=0, vc>=0]);` implements the simplex method and yields the following results (a `float` command was invoked also):

$$[144.44, [vc = 1.2963, vb = 1.4815, va = 0.0]].$$

The values differ slightly from those above, because *Maxima*'s output involves less rounding error.

A digression: We can return to the primal problem look at the shadow price in a slightly different way, one that relates to the Lagrangian multipliers that we encountered earlier. The commands below, sequentially add 1 unit of inputs a, b, and c, holding the other two at the initial levels: `C1alt: 15*X2+5*X1<=61$ C2alt:4*X2 + 3*X1 <= 25$ C3alt:7*X2 + 12*X1<=85$`. The results, after invoking a `float` command, are these:

$$\begin{bmatrix} [144.44, [X2 = 1.3333, X1 = 6.2222]] \\ [145.93, [X2 = 1.7778, X1 = 5.963]] \\ [145.74, [X2 = 1.2222, X1 = 6.3704]] \end{bmatrix}$$

Adding a unit of a changes nothing. Adding either b or c changes everything: X_1, X_2, and R all increase. Subtracting the initial R, 144.44, from each of the two higher values, yields the shadow prices of b and c.

2. Given that the primal problem is
minimize: $W = 2 \cdot X_1 + 5 \cdot X_2$
subject to:
$5 \cdot X_1 + 6 \cdot X_2 \geq 12$
$-3 \cdot X_1 + 4 \cdot X_2 \geq 10$
$X_1 + 5 \cdot X_2 \geq 8$
$2 \cdot X_1 + X_2 \geq 3$ and
$X_1, X_2 \geq 0$
then the dual problem is
maximize:
$U = 12 \cdot Y_1 + 10 \cdot Y_2 + 8 \cdot Y_3 + 3 \cdot Y_4$
subject to:
$5 \cdot Y_1 - 3 \cdot Y_2 + Y_3 + 2 \cdot Y_4 \leq 2$
$6 \cdot Y_1 + 4 \cdot Y_2 + 5 \cdot Y_3 + Y_4 \leq 5$
$Y_1, Y_2, Y_3, Y_4 \geq 0$

Exercise: Draw the four constraints in the primal problem for Example 2 and shade the feasible region. Use the graph to estimate the values of X_1 and X_2. Which constraints are binding? See the accompanying workbook to determine the exact values of X_1 and X_2 and to confirm that two of the Y values are zero.

10.1.5 Nonlinear Programming

We noted earlier that when one or more constraints, or the objective function, is not linear, then a linear programming solution may either be impossible or misleading. Fortunately, modern computers can examine nonlinear models, often quite rapidly. This is so, despite the fact that nonlinear programming typically involves many more computations. *Maxima* offers the nonlinear programming option `cobyla` (**c**onstrained **o**ptimization **by** **l**inear **a**lgebra). Bradley, *et al.* [3] provides a detailed overview of the nature of nonlinear programming and of the types of approaches to implementing such programming.

We provide a brief overview in two steps. First, we revisit the diet problem, comparing the simplex solution to that provided by `cobyla`. Then we set up and solve a simple nonlinear problem that appears in Bradley *et al.*

A nonlinear problem cannot be addressed with linear programming, but a linear problem can be addressed with nonlinear programming. We saw earlier that the simplex method could be implemented with the following commnds:

```
load(simplex)$
minimize_lp(z, [c1,c2,c3,c4,c5,c6,c7,c8,
x1>=0,x2>=0,x3>=0,x4>=0,x5>=0]);
```

where z is the objective function, and the brackets contain a list of constraints. The resulting output is $[0.10904, [x5 = 0.048628, x4 = 0.0051128, x3 = 0.01125, x2 = 0.0085915, x1 = 0.035456]$.
The interpretation appears early in this section

The cobyla counterparts are the following input:

```
load(fmin_cobyla);
fmin_cobyla(z, [x1,x2,x3,x4,x5], [0.031,0.01,0.01,0.005,0.005],
constraints = [c1,c2,c3,c4,c5,c6,c7,c8,
x1>=0,x2>=0,x3>=0,x4>=0,x5>=0],iprint=1);.
```

This input differs from the simplex input only in minor details, with one exception. The minor details are that the constraints are explicitly identified as such and that a code `iprint=1` is added in order to control the amount of detail that is reported. The more substantive difference is that this command requires a list of initial guesses of the values of $x_1, x_2, \ldots$. In general, the better guesses one can provide the better the method will work. In particular, close guesses reduce the number of iterations required to obtain the desired results.

As with the input, the output differs in detail from the simplex output:
$[[x1 = 0.035456, x2 = 0.0085915, x3 = 0.01125,$
$x4 = 0.0051128, x5 = 0.048628], 0.10904, 42, 0]$.
The output appears with the$x$ values first, followed the the value of z. The number 42 is the number iterations that were required. The 0 is a code that indicates that the process was completed without error.

We now consider a simple example that is not amenable to linear programming. We wish to maximize the function $z = 2 * x1 - x1^2 + x2$. Unfortunately cobyla is limited to minimization, but we can minimize $-z = -(2 * x1 - x1^2 + x2)$. The constraints are these: $x2 \leq 1.8$, $x_1^2 + x2^2 \leq 4$, $x1 \geq 0$, and $x2 \geq 0$. The command that we enter is `load(fmin_cobyla)$`

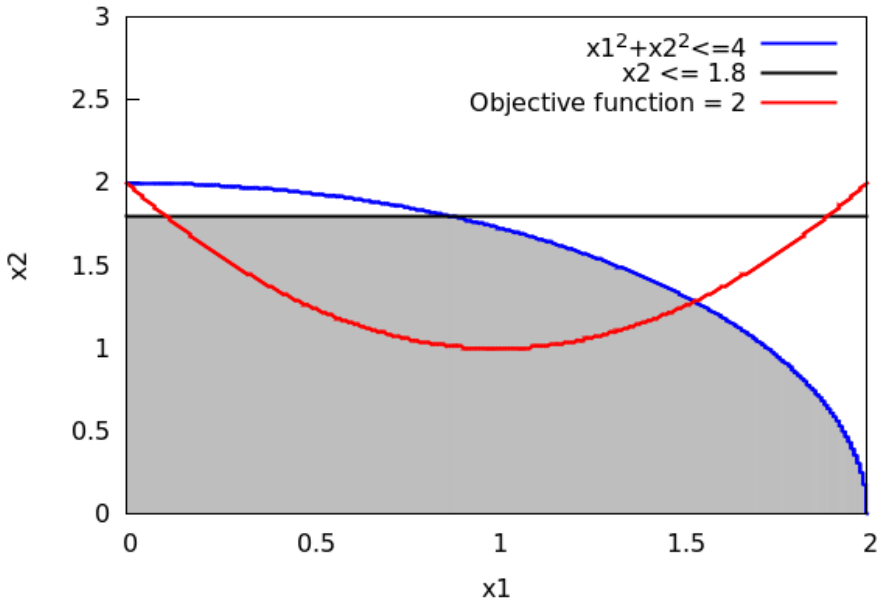


Figure 10.4: Feasible region and objective function, nonlinear

`fmin_cobyla(-(2*x1 - x1^|2 + x2), [x1,x2], [1,1.8], constraints = [x2<=1.8,x1^2 + x2^2<=4,x1>=0,x2>=0], iprint=1);`. We used Figure 10.4 to guess that $x1$ would be near to 1 and that $x2$ would equal 1.8.[3]

The resulting output,

$$[[x1 = 0.87178, x2 = 1.8], -2.7836, 20, 0]$$

confirms our expections. The value of $-z$ is -2.7836, so $z = 2.7836$ is the highest attainable value of our objective function.

10.2 Input-Output Analysis

The system of markets in the United States contains millions of separate and independent economic decision units–households, business firms, not-

[3]As an exercise, repeat this with guess of 10 and 10 or some other value. Confirm that the number of iterations increases.

for-profit organizations, and government agencies. Each decision unit is interested primarily achieving its own set of purposes, and seemingly pays little heed to the survival and behavior of most of the other decision units. Nonetheless, as Adam Smith persuasively demonstrated more than two centuries ago, the self-serving efforts of millions of independent economic decision-makers are somehow harnessed and drawn together by the functioning of an economic system.

Day in, day out, this economic system provides approximately the correct quantities of food, clothing, shelter, and other goods that consumers wish to purchase. (Correct, in the sense that persistent shortages and surpluses are seldom observed.) This is Smith's "invisible hand " at work, for no central planning agency wills this to take place. The individual decision-makers, perhaps intending only personal good, unwittingly (or otherwise) does public good as well.

How and why does the economic system hang together? What are the interrelationships between inputs and outputs that affect our everyday lives? The answers to questions like these are provided by general equilibrium analysis.

Adam Smith was an early practitioner of general equilibrium analysis, which explicitly includes and analyzes reactions and feedback effects among large numbers of variables. This analysis is based on the assumption that all markets and all decision-makers are affected by one another's actions to some degree. In general equilibrium analysis, the price of oil affects the price of gasoline, but it also affects the price of automobiles, plastic drinking cups, and the temperature at which you choose to heat your home.

This section introduces *input-output analysis*, a general equilibrium approach to production. Input-output analysis is entirely empirical in nature. It describes the relationship among various inputs and various outputs. Demand conditions play no part in input-output models, because input-output analysis examines only the relationship between inputs to the productive process and the outputs that result. The prices and quantities at which we might produce and sell such inputs and outputs and such factors as price elasticity of demand are not considered.[4]

[4]Attempts to generalize this analysis by allowing for demand considerations and for some elasticity of substitution in production constitute a set of techniques given the name "computable general equilibrium" models. See Mitra-Kahn [15] for an overview and history of these models.

The father of input-output analysis, Wassily Leontief, Leontief used input-output analysis to show how the production of one sector of the economy depends, to some degree, on the production of all other sectors of the economy. The input-output"tables" that Leontief produced described the numeric relationships between inputs used and the outputs produced in the American economy. This set of relationships was stated in physical terms (for example, tons of steel or gallons of gasoline) rather than monetary terms. The price of a ton of steel was not considered.

Most input-output models rest on three assumptions. The first is that no two commodities are produced jointly. Each firm or market is assumed to produce only one homogeneous product. Second, all inputs are employed in rigidly fixed proportions in production. The law of diminishing returns does not apply because input proportions never vary. Also, this assumption implies constant returns to scale in all production. Third, no external economies or diseconomies exist for any firm or market. Thus, the production by one firm cannot affect the technology that governs the production of any other product in either a positive of negative fashion.

Let the total production of any single industry during a particular period be represented by X_i. Industry i's production can be used as inputs in other productive processes, or it can be used to satisfy final consumption demand. If there are n different industries, then potentially n different industries can use output X_i,as an input. Therefore, we can write $X_i = X_{i1} + X_{i2} + \ldots + X_{in} + d_i$, where X_i = output of industry i, $X_{ij}j$ = output of industry i used as an input in industry j, and d_i = final demand for the finished goods and services of industry i.

We pointed out above that input-output analysis assumes that all production takes place under conditions of rigidly fixed proportions. Thus the amount of steel required to produce one car does not change, regardless of the number of cars produced. Let a_{ij} represent a *technical coefficient of production.* Specifically, a_{ij} is the (constant) number of units of input j that are required to produce one unit of output i. We can therefore express the production of industry i as follows:

$$X_i = \sum_{j=1}^{n} a_{ij} \cdot X_j + d_i \qquad i = 1, 2, \ldots, n,$$

where X_i = output of industry l, (1,7 = number of units of input j needed

to produce one unit of output i, X_j = output of industry j, and d_i = final demand for the finished goods and services of industry i.

In an economy that has n industries, there are $n \times n$ technical coefficients of production to consider, since each industry can potentially provide inputs to every other industry (including itself). Therefore, we can write a full system of linear equations that describes the input-output relationships for an economy composed of n industries:

$$\begin{array}{lllll} X_1 = & +a_{11} \cdot X_1 + & +a_{12} \cdot X_2 + & \ldots & +a_{1n} \cdot X_n + \quad d_1 \\ X_2 = & +a_{21} \cdot X_1 + & +a_{22} \cdot X_2 + & \cdots & +a_{2n} \cdot X_n + \quad d_2 \\ \vdots & \vdots & \vdots & & \vdots \quad \vdots \\ X_n = & +a_{n1} \cdot X_1 + & +a_{n2} \cdot X_2 + & \cdots & +a_{nn} \cdot X_n + \quad d_n \end{array}$$

These equations could equivalently be written in terms of the final demands for goods and services, the dis. This version of the input-output model indicates where the final goods and services in the economy are produced:

$$\begin{array}{lllll} d_1 = & (1 - a_{11} \cdot) X_1 - & a_{12} \cdot X_2 - & \cdots & -a_{1n} \cdot X_n \\ d_2 = & -a_{21} \cdot X_1 + & (1 - a_{22}) \cdot X_2 - & \cdots & -a_{2n} \cdot X_n \\ \vdots & \vdots & \vdots & & \vdots \\ d_i = & -a_{i1} \cdot X_1 - & a_{i2} \cdot X_2 - & \cdots & +(1 - a_{ii}) \cdot X_i \quad + \cdots - a_{in} \cdot X_n \\ \vdots & \vdots & \vdots & & \vdots \\ d_n = & -a_{n1} \cdot X_1 - & a_{n2} \cdot X_2 - & \cdots & +(1 - a_{nn}) \cdot X_n \end{array}$$

In matrix notation, this system of linear equations is expressed as

$$\begin{bmatrix} (1 - a_{11} & -a_{12} & \cdots & -a_{in} \\ -a_{21} & (1 - a_{22} & \cdots & -a_{2n} \\ \vdots & \vdots & & \vdots \\ -a_{n1} & -a_{n2} & \cdots & (1 - a_{nn}) \end{bmatrix} \cdot \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix},$$

or $(I - A) \cdot X = d$

In this equation I is an $n \times n$ identity matrix, A is the technical coefficient matrix, X is the n-industry variable matrix, and d is the final demand matrix. We frequently refer to the matrix $(I - A)$ as a *Leontief matrix*. Using matrix inversion, if $I - A$ is nonsingular, then we can find $(I - A)^{-1}$. This means that the unique solution for the X matrix is $X = (I - A)^{-1} \cdot d$.

An illustrative example. The table on the next page lists the sources of inputs, and the destinations of outputs, in a hypothetical economy. We could represent th this table's contents as a system of 11 simultaneous linear equations in 11 unknown values. Movements along any row show the output of an industry and where that output goes. For example, Row 6 addresses industry F. Two units its output go to industry A, six units to industry B, and so forth. Column 7 reveals that two units of industry D's output constitute the accumulation of inventories in industry F itself. Column 12 shows that he total production of industry F is 46 units.

A movement down any column in this table lists the inputs that each industry or sector receives from other industries or sectors. For example, column 5 indicates the inputs that industry E receives from other industries and sectors. Thus industry E uses five units of industry A's output, three units of industry B's output, five units of industry C's output, and so forth.

The "processing sector" of an input-output table (rows 1 through 6 and columns 1 through 6 contains all those industries that produce salable goods and services, such as cars, furniture, and toothpaste. The processing sector of most input-output tables is highly developed and may contain as many as 500 industries.

Columns 7 through 11 contain the "final demand" sector. For example, household purchases of goods and services, in column 11, total 14 units from industry A, 17 units from industry B, and so forth. Rows 7 through 11 contain the "payments sector" of the table. This sector shows the contribution of various owners of factor inputs (for example, households) to the production of each output. For example, households provide 19 units of their inputs, predominantly labor, to industry A, as recorded in row 11, column 1.

Compare this sector to the Keynesian aggregate expenditures equation $Y = C+I+G+X-M$, where Y is total spending, C is private-sector consumption, I is private-sector investment, G is government spending, X is exports, and M is imports.

The table also records total gross outlays (in row 12) and Total Gross Output (in column 12). The total gross outlay of inputs and the total gross output of goods and services are not equivalent to gross Domestic Product, which deliberately excludes intermediate outputs and inputs and concentrates only on the value of final goods and services. In constrast, total gross outlay and total gross input involve repeated double counting. This is not bad, however,

because the purpose of input-output analysis is to illustrate the connections of the economy, not to provide a measure of the value of total inputs used or outputs produced.

Input-output table

	Industry purchasing (outputs[a])												
	Processing sector						*Final demand*						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	A	*B*	*C*	*D*	*E*	*F*	*Gross inventory accumulation (+)*	*Exports to foreign countries*	*Government purchases*	*Gross private capital formation*	*Households*	*Total gross output*
Industry producing (inputs[b])												
Processing sector												
1. Industry *A*	10	15	1	2	5	6	2	5	1	3	14	64
2. Industry *B*	5	4	7	1	3	8	1	6	3	4	17	59
3. Industry *C*	7	2	8	1	5	3	2	3	1	3	5	40
4. Industry *D*	11	1	2	8	6	4	0	0	1	2	4	39
5. Industry *E*	4	0	1	14	3	2	1	2	1	3	9	40
6. Industry *F*	2	6	7	6	2	6	2	4	2	1	8	46
Payments sector												
7. Gross inventory depletion (−)	1	2	1	0	2	1	0	1	0	0	0	8
8. Imports	2	1	3	0	3	2	0	0	0	0	2	13
9. Payments to government	2	3	2	2	1	2	3	2	1	2	12	32
10. Depreciation allowances	1	2	1	0	1	0	0	0	0	0	0	5
11. Households	19	23	7	5	9	12	1	0	8	0	1	85
12. Total gross outlays	64	59	40	39	40	46	12	23	18	18	72	431

[a] Sales to industries and sectors along the top of the table from the industry listed in each row at the left of the table.

[b] Purchases from industries and sectors at the left of the table by the industry listed at the top of each column.

Source: The Elements of Input-Output Analysis, by William Miernyk. Copyright 1965 by Random House, Inc. Reprinted by permission of Random House, Inc.

We can use this table to see how one computes the technical coefficients of production that we discussed earlier. Each technical coefficient of production should show the number of units of input j required to produce one unit of output i. Thke next table consists of technical coefficients of production derived from the input-output matrix in the preceding table. Consider industry C: It receives a total of 40 units of inputs, one of which comes from the depletion of its own inventories. Seven of these 40 units come from industry F. Therefore the technical coefficient of production is $7/39 \approx 0.18$. This tells us that every unit of output produced by industry C requires 0.18 unit of the output of Industry F. (Note that we subtract inventory depletion from total gross outlay before computing the technical coefficient of production.)

		Inputs purchased from industries:					
		A	B	C	D	E	F
Outputs	A	0.16	0.26	0.03	0.05	0.13	0.13
pro-	B	0.08	0.07	0.18	0.03	0.08	0.18
duced	C	0.11	0.04	0.21	0.03	0.13	0.07
by	D	0.17	0.02	0.05	0.21	0.15	0.09
indus-	E	0.06	0.00	0.03	0.36	0.08	0.04
tries	F	0.03	0.11	0.18	0.15	0.05	0.13

Technical coefficients of production can be useful to a researcher or forecaster if they are based on up-to-date data that accurately portray the actual productive processes being surveyed. For example, one can use technical coefficients of production to determine the probable effects of a decrease in the output of steel on the output of cars, on apartment construction, and even on Christmas toys.

We can also trace the effects of public policies such as road building, increased defense expenditures, changes in international trade policy and the like by means of technical coefficients of production. The U. S. Department of Commerce has long maintained a sophisticated input-output model to assist it in predicting the consequences of a wide range of public and private actions. Regional development authorities and larger corporations have also made extensive use of input-output models.[5]

[5]See Miller and Blair [14] for a full development of the technique and for an overview of its applications. Also, a Web search of "input-output" analysis will yield a large number of useful sites, including many studies by the International Monetary Fund.

It is possible to derive input-output multipliers that show the total change in output that will occur as a result of a change in the output of one industry. According to the table above, industry D receives 0.36 units of output from industry E whenever industry D expands its output by one unit. Suppose that industry E's output increases initially by one unit; this initially causes industry D's output to rise by 0.36 units.

Observe in turn that when industry D's output rises by one unit, industry E's output rises by 0.15 unit. Thus a 0.36-unit increase in industry D's output has a feedback effect that increases industry E's output by $0.36 \times (0.15 = 0.05$ units. But this increase in industry E's output once again requires additional inputs from industry D, and so forth. The original expansion in industry E's output has set off a chain reaction of secondary increases in output. This multiplier process, which is like the "national income multiplier" of Keynesian models, can be quantified, so that we can isolate and analyze the final, terminal effects of a given action.

10.3 Questions and Problems

1. Given the following linear-programming problem.
   Maximize: $3 \cdot X_1 + 4 \cdot X_2$
   Subject to: $2 \cdot X_1 + X_2 \leq 12$
   $3 \cdot X_1 + 2 \cdot X_2 \leq 20$
   $X_1, X_2 \geq 0$

   (a) What are the decision variables in the problem?

   (b) Graph the problem, both by hand and using *Maxima*.

   (c) What quantities of X_1 and X_2 maximize the objective function subject to the two constraints? What is the value of the objective function given these values?

   (d) What if the nonnegativity constraint were removed? What difference would this make in the solution? Would such a solution be sensible? What happens when you remove these constraints in *Maxima*. How do interpret the error message?

   (e) Define and solve the dual problem.

2. Hashimi, Richmond, and Blaylock (HR&B) operate an accounting firm that performs tax audits and also completes tax returns. The trio has a good enough reputation that they can conduct as many audits and returns as they choose. In a typical week, HR& B devote 115 hours of time to performing audits and doing tax returns.

   They have decided that, of these 115 hours, 75 are production hours and 40 are review hours. Each time an audit is performed, 10 hours of production time and 4 hours of review time are used. Each time a tax return is completed, 3 hours of production time and 2 hours of review time are used. An audit is priced at $1000, while a tax return is priced at $400. HR&B wishes to maximize the revenues that they receive from performing audits and completing tax returns.

   (a) On average, how many audits and how many tax returns should HR&B conduct? How many staff hours are devoted to audits, and how many to tax returns? What is the average weekly income?

   (b) Select any other feasible solution, and demonstrate that it generates less total revenue than the solution identified in (a).

   (c) Define the dual to this maximization program and determine the shadow prices of production and review.

   (d) Based on (c) suppose that a retired accountant offers 10 hours of service and that this accountant has the same skills as the HR&B partners. If they hire the accountant, to which activity should they assign her? (The partners will still provide production and review hours as if the part-time accountant were not employed.)

   (e) If they hire the accountant and use her as indicated in (d), by how much does their weekly revenue change?

3. Tony and Jim open a lemonade stand on their front sidewalk. They can make ordinary lemonade, or they can make a Lemon Fizz for their customers. The ingredients per liter for each of these drinks are as follows:

Lemonade	*Lemon Fizz*
0.25 liters of sugar	0.75 liters of sugar
2 lemons	3 lemons
1 liter of water	1 liter of ginger ale

Water is free and available in any quantity required. A total of 4 liters of sugar, 25 lemons, and 5 liters of ginger ale are available. Tony and Jim believe that they can sell each 0.5-liter glass of lemonade for $0.75, while each 0.5-liter Lemon Fizz will sell for $1.25. They wish to maximize the revenue that they realize from the sale of lemonade and fizzes. How much of each drink should Tony and Jim sell? How much revenue will they earn?

Appendix A

Additional Review Questions

These review questions are selected from a final example in Professor Ostrosky's course at Illinois State University, "Introduction to Mathematical Economics," which closely parallels the material presented in this book.

1. Assume that consumption is a linear function of income. The marginal propensity to consume is 0.90. If income were zero, this function would imply dissaving of $180. What is the explicit consumption function? What is the multiplier for this problem if consumption is the only component of aggregate expenditures (C + I + G + X - M) that is affected by income?

2. Given the CES production function
   $Q = \gamma \cdot [\delta \cdot K^{-\rho} + (1 - \delta) \cdot L^{-\rho}]^{-\nu/\rho}$
   where Q = output, L = labor, K = capital, and $\gamma, \delta, \rho, and \nu$are parameters. Determine (a) dQ and (b) the degree of homogeneity.

3. Given the demand curve $q = a - b \cdot P$, where $a, b > 0$ are parameters, what is the price elasticity of demand at each intercept? Prove your answers using the formula for elasticity.

4. The production function for a firm's product is $Q = 12 \cdot L + 20 \cdot K - L^2 - 2 \cdot K^2$. The per-unit cost to the firm of L and K is $4 and $8 per unit, respectively.

   a. Derive the equations for the marginal products of L and K, MPL and MPK.

b. Use the condition $MPL/MPK = 20/40$ to determine the optimal relatiionship between the employment of these two resources.

c. Suppose that the firm wants the total cost of inputs to be \$88. Find the greatest output possible subject to this cost constraint, using the Lagrangian multiplier approach. Confirm that the result agrees with your solution in (b).

5. Assume that the demand per week for the NoFuzz Cable is 10,000 subscribers when the price is \$60 per unit, and 20,000 subscribers when the price is \$40.

   a. Determine the demand equation, assuming that it is linear.

   b. What is the elasticity at a price of \$60?

   c. What advice would you give to the Cable Company based on the information contained in this demand curve?

6. Blinko Company is the sole producer of artificial lightning bugs. Management has determined that the company's total revenue function is $TR = 100 \cdot Q - Q^3$. Determine the point elasticity of demand for artificial lightning bugs when $Q = 5$.

7. Assume the following national-income model:

$Y = C + I + G$	
$C = a_0 + a_1 \cdot (Y - T)$	$a_0 > 0$ and $0 < a_1 < 1$
$I = b_0 + b_1 \cdot Y + b_2 \cdot i$	$b_0 > 0, b_1 > 0, b_2 > 0$
$G = G_0$	a constant
$i = i_0$	the interest rate, a constant
$T = t_0 + t_1 \cdot Y$	$t_0 > 0, 0 < t_1 < 1$

   a. List all parameters, endogenous variables, and exogenous variables.

   b. Solve for the equilibrium income level using matrix algebra.

8. A firm's total cost function is $TC = x^2/4 + 3 \cdot x + 400$, where x is the number of units produced. At what level of output will average cost per unit be a minimum?

9. Find the extreme value(s) of $W = -x^3 + 3 \cdot x \cdot z + 2 \cdot y - y^2 - 3 \cdot z^2$.

10. How much of goods x, y, and z should a person consume so as to maximize utility, where the utility function is given by $U = 10 \cdot x \cdot y \cdot z$, and where the price of x, is \$1, the price of y is \$2, the price of z is \$4, and the available budget is \$120?

11. Solve the following set of equations using inverse matrix algebra:
$x_1 + 2 \cdot x_3 + x_4 = 4$
$x_1 - x_2 + 2 \cdot x_4 = 12$
$2 \cdot x_1 + x_2 + x_4 = 12$
$x_1 + 2 \cdot x_2 + x_3 + x_4 = 12$

12. Suppose that A and B are the only two firms in the market selling the same product (we say that they are duopolists). The industry's inverse demand function for the product is $P = 92 - q_A - q_B$, where q_A and q_B denote the output produced and sold by A and B, respectively. For A the cost function is $C_A = 10 \cdot q_A$ and for B it is $C_B = q_B^2/4$. Suppose that the firms enter into an agreement on output and price control by jointly acting as a monopoly.

    a. Express profit as a function of q_A and q_B, and determine how they should allocate output so as to maximize the profit of the monopoly.

    b. Determine the price that they should charge and their joint profit level.

    b. Determine the price that they should charge and their joint profit level.

    c Confirm that, at the values of Q, q_B, and q_B that you have determined, the two firms' marginal cost levels are the same.

13. Given the Cobb-Douglas production function $Q = A \cdot K^{\alpha} L^{\beta}$, where A, α, β are parameters, show that the expansion path of a firm is equal to a *ray* ($K = c \cdot L$, where c is a constant), implying that the optimal input ratio should be the same at all output levels. Remember that the expansion path of a firm describes the least-cost combinations required to produce varying levels of Q. Assume that the cost of production is given by $C = P_K \cdot K + P_L \cdot L$, where P_K and P_L are constants (*i. e.*,they are not affected by the firm's employment levels of K and L.

14. The Sweet-Tooth Candy Company produces two delectable varieties of candy, A and B, for which the average costs of production are constant at \$7.00 and \$8.00 per kilogram, respectively. The quantities q_A and q_B (in kilograms) of A and B that can be sold each week are given by the joint-demand functions

$$q_A = 24 \cdot (P_B - P_A) \text{ and } q_B = 20 \cdot (30 + P_A - 2 \cdot P_B),$$

where P_A and P_B are the selling prices (in dollars per kilogram) of A and B, respectively. Determine the selling prices that will maximize Sweet Tooth's profit.

15. Suppose that a price-searching firm is practicing price discrimination by selling the same product in two separate markets at different prices. Let q_A be the number of units sold in market A, where the demand function is $P_A = f(q_A)$, and let q_B be the number of units sold in market B, where the demand function is $P_B = g(q_B)$. Assuming that all units are produced at one plant and that transportation and marketing costs are the same in both markets, let the cost function for producing $q = q_A + q_B$ units be $C = C(q)$. [Hint: Keep in mind that total revenue from market A is solely a function of q_A, and total revenue from market B is solely a function of q_B.] Set up and determine only the first-order conditions for the monopolist to maximize profits with respect to outputs q_A and q_B. Interpret your results.

16. Given that

$$A = \begin{bmatrix} 1 & -1 & 2 \\ 0 & 3 & 4 \end{bmatrix}, B = \begin{bmatrix} 44 & 0 & -3 \\ -1 & -2 & 3 \end{bmatrix}$$

and

$$C = \begin{bmatrix} 2 & -3 & 0 & 1 \\ 5 & -1 & -4 & 2 \\ -1 & 0 & 0 & 3 \end{bmatrix}, D = \begin{bmatrix} 2 \\ -1 \\ 3 \end{bmatrix}$$

find $A + B$, $3 \cdot A - 4 \cdot B$, $A \cdot D$, $D' \cdot D$, $D \cdot D'$.
Show the following:
(A+B)'=A'+B' and $(A + B) \cdot D = A \cdot D + B \cdot D$.

17. Evaluate the determinant $\begin{vmatrix} 2 & 3 & -1 & 2 & 1 \\ 0 & 1 & -1 & 1 & 2 \\ 0 & 0 & -1 & 2 & 3 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 2 & 5 \end{vmatrix}$.

18. Find the inverse for the matrix $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$.

19. Solve the following equations using matrix algebra.
$x_1 + x_2 + x_3 = 0$, $x_1 + 3 \cdot x_3 = 1$ and $2 \cdot x_2 + 2 \cdot x_1 = 0$

20. Find the extreme value(s) of $Z = 2 \cdot x_1^2 + x_1 \cdot x_2 + 4 \cdot x_2^2 + x_1 \cdot x_3 + x_3^2 + 2$.

21. Determine the values of x_1, x_2, and x_3 that maximize or minimize the function $Z = x_1^2 + x_2^2 + 7 \cdot x_3^2 - 2 \cdot x_1 \cdot x_3 + 10$ subject to the constraint $x_1 + 2 \cdot x_2 + 3 \cdot x_3 = 0$. Evaluate Z at this set of x_i values.

Bibliography

[1] Bassi LJ (1976) The Diet Problem Revisited. American Economist, 20: 35–39.

[2] Bishop RL (1968) The Effects of Specific and Ad Valorem Taxes, Quarterly Journal of Economics, 82:198–218.

[3] Bradley SP, Hax AC and Magnanti TL (1977). Applied Mathematical Programming. Addison-Wesley, New York. Available at: http://web.mit.edu/15.053/www/

[4] Chiang AC (2016). Fundamental Methods of Mathematical Economics, 3rd ed. McGraw-Hill, New York.

[5] Cyrenne P (2014). Salary Inequality, Team Success and the Superstar Effect. Available at: ftp:// ftp.repec.org/opt/ReDIF/RePEc/win/winwop/2014-02.pdf.

[6] Garicano L and Rossi-Hansberg L (2015). Knowledge-Based Hierarchies: Using Organizations to Understand the Economy. Available at: https://www.princeton.edu/ erossi/KBH.pdf.

[7] Hammock MR and Mixon JW (2013). Microeconomic Theory and Computation. Springer, New York.

[8] Jerison D. (2006)Single Variable Calculus. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

[9] Leydold J and M Petry (2011) Introduction to *Maxima* for Economics. Available at http://statmath.wu.ac.at/ leydold/maxima/.

[10] McAfee RP (2007). Introduction to Economic Analysis. Available at http://www.mcafee.cc/Introecon/IEA2007.pdf.

[11] Mankiw GN (2016). Macroeconomics, 9th ed. Macmilan, New York.

[12] Meglicki Z (2001). *Maxima*, *Maple*, and *Mathematica*: the Summary. http://beige.ucs.indiana.edu/ P573/node35.html.

[13] Miller DE (2012) Using *wxMaxima* for Basic Set Operations. Available at http:// andrejv.github.io/wxmaxima/help.html.

[14] Miller RE, Blair PD (2009). Input-Output Analysis: Foundations and Extensions. Cambridge, Cambridge University Press.

[15] Mitra-Kahn BH (2008) Debunking the Myths of Computable General Equilibrium Models. Available at http://econpapers.repec.org/paper/epacepawp/2008-1.htm

[16] Ostrosky AL, Koch JV (1979). Introduction to Mathematical Economics. Waveland Press Inc, Prospect Heights IL.

[17] Perlis S (2012). Theory of Matrices. Literary Licensing, Whitefish MT.

[18] Solow, RM (1956). A Contribution to the Theory of Economic Growth. Quarterly Journal of Economics, 70: 65–94. doi:10.2307/1884513. JSTOR 1884513. Pdf.

[19] Stigler GJ (1945) The Cost of Subsistence, Journal of Farm Economics, 27:303–314.

Index