Principles of Calculus Modeling

An Interactive Approach

Donald Kreider • Dwight Lahr

Principles of Calculus Modeling
An Interactive Approach

Donald Kreider
Dwight Lahr

Department of Mathematics
Dartmouth College
Hanover, NH 03755

donald.kreider@dartmouth.edu
dwight.lahr@dartmouth.edu

10 August 2002

Authors supported in part by grants from the National Science Foundation through its Mathematics Across the Curriculum Project at Dartmouth, and the Andrew W. Mellon Foundation. Cover drawing by Perico.

Contents

Preface

Why write another calculus book? After all, there are already many good ones from which to choose. Does the world really need a calculus book by K and L?

We asked ourselves these questions many times before deciding to go ahead and write yet another calculus book. That decision was based on our desire to develop a book that did four things that the others do not. We wanted a book that:

- Had the spirit of the way calculus was developed in the seventeenth century. Concepts evolved before becoming precise. Graphical, algebraic, and numerical approaches were combined to gain intuition and achieve results.
- Dealt with real applications with real data. This is never an easy task, and most so-called *real* data has been significantly cleaned up before it is analyzed. But we wanted to restore some of the emphasis on this aspect of the role of calculus in the world today, and put the techniques more in perspective.
- Used modern technology both to enhance an understanding of the subject and as a tool to implement some calculus procedures. We have learned a lot about computing and calculus over our years of teaching, and we wanted to see if we could capture those ideas in a book.
- Was lean. We did not want to produce another thick, heavy book, but one that still would have all the essential ideas of a first course in calculus.

Although we have achieved the leanness, there is actually more to the book than meets the eye because it is intended to be used with its companion Web site *klbooksite*. That Web site contains on-line materials that we consider to be an integral part of the book. For example, klbooksite contains a Web page for each section of the book. A typical section page contains:

- A summary of the section.
- What you should know after studying the material of the section.
- A list of the applets of the section.
- Worked out examples.
- A link to videos of problems being worked out.
- A quiz (with answers) that you can take to test your knowledge.
- A link to a printout (in PDF) of all the exercises of the section.
- A link to these same exercises in the on-line self-checking format.

There are also other supplementary materials that you might find useful such as sample exams.

The Web site klbooksite can be found at http://www.math.dartmouth.edu/~klbooksite/.

We envision students reading a section of the book, and then going to klbooksite for an overview of the section and links to the applets and homework.

Animations will require that computer users have installed Macromedia Flash player. Applets will require both PC and Mac users to view them with Microsoft Internet Explorer browser (where Mac users also must

install the accompanying MRJ package from Apple). Exercise printouts require Adobe Acrobat Reader to view them. And video clips need something like RealPlayer. All of these software tools are free and readily available on the Web. Go to klbooksite to find out more.

Students will want to know what prerequisite knowledge they should bring to the book. We assume familiarity with the Cartesian coordinate system in the plane, and with common geometric figures such as lines, parabolas, ellipses, and circles. We also assume knowledge of basic trigonometry, although we give an extensive introduction to the essential ideas. In the first chapter, we touch on all of the foregoing high school topics to standardize terminology and establish a point of view aimed at the study of calculus.

Many people have assisted in the effort of producing materials. First, we would like to express our appreciation to all the students who have been members of the Calculus I Development Team: garduate students Erin M. Boyer, Emily Dryden, Dominic W. Klyve, Rebecca Martel, Amanda J. Sheppard, Lee J. Stemkoski; and undergraduates Blythe Adler and Ohene K. Ohene-Adu.

We also would like to thank Edwin Gailits, Fuxing Hou, Jane Korey, Kim Rheinlander, Tom Shemanske, and Dorothy Wallace for their various contributions to the project.

Last, our special thanks go to Susan J. Diesel who managed the students, and contributed to the overall project in ways too many to enumerate.

–Donald Kreider and Dwight Lahr

Chapter 1

Modeling Discrete Data

1.1 Introduction to the Issues

Most of us were introduced to mathematics through counting. A typical everyday problem that we learned to solve is: If a school bus holds 50 children, how many buses are required to take three classes, of 30, 35, and 32 students each, on a school trip? Such problems taught us how to add, subtract, multiply, and divide. Now, we can readily calculate that 2 buses will be needed.

Then came algebra. Algebra taught us to work with symbols: If Jeff is half as old as his father, and the sum of their ages is 60, how old is Jeff? We all know what to do. We could let x be the age of the father, and y be the age of Jeff. Then the information leads to two equations that yield $x = 40$ and $y = 20$.

Now, we are going to begin the study of calculus. Calculus gives us the tools to answer questions about movement and change, while building on our knowledge of arithmetic and algebra. A typical example of the kind of question we will learn to answer is the following: Suppose we drop an object vertically, from rest, off the roof of a building. How long will it take for it to hit the sidewalk ten meters below?

Just as we have been trained in the arithmetic and algebra problems above, the first thing we must do is translate the problem into mathematical language. Because we are just finding our way, this is not so easy. But it appears that as the question now stands, we need some additional information. For example, it would be nice if we had a formula that would give us the time corresponding to any distance of fall. Even though this may seem like too much to ask for, we will see that this indeed will be one of the outcomes of our study of calculus. But to jump to that formula now would be like pulling a rabbit out of a hat. We want to understand how we get it, and where it comes from, so that we can apply the same or similar principles in other situations. Thus, we will take smaller steps.

Suppose for example that a similar experiment has been conducted in a lab where the distances (in meters) have been recorded every tenth of a second for one second, as in the table below. If we could extrapolate the data to 10 meters, then we would be able to answer the question about the object falling from the roof. So, how shall we proceed?

time (s)	distance (m)
0.10	0.049
0.20	0.196
0.30	0.441
0.40	0.784
0.50	1.225
0.60	1.764
0.70	2.401
0.80	3.136
0.90	3.969
1.00	4.900

A reasonable approach is to try to find a function that will model the data. That is, find a function $f(t)$

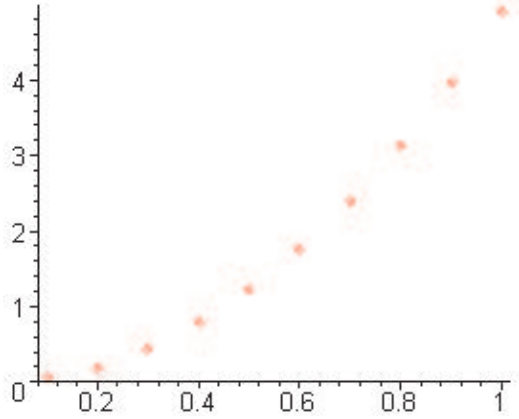
of time t whose values at the recorded times come close to matching the distances in the table. Then we can use that function to find the desired time, namely, by solving for t in the equation $f(t) = 10$. But how do we find such a function, and how do we know which is best if we find more than one?

Let's postpone the question of deciding which function is best among several choices. So far we don't even have one candidate; hence, let's deal with that problem first. Building on what we know, a good place to start our search is with the *elementary functions*. After all, these functions – polynomials, exponentials, logarithms, trigonometric functions – in addition to being familiar to us, have proved themselves throughout the years to be very valuable in just such situations.

1.1.1 Modeling with an Elementary Function

We will start with the least complicated elementary function and continue testing until, hopefully, we find one that fits the data fairly well. Eventually we will have to become more precise about concepts such as *fairly well*, but for now we will forge ahead just to see if the plan holds any promise at all.

The data certainly are not constant. Thus, we check to see if they are linear. That is, is there a function of the form $f(t) = at + b$ that fits the data in the table? Before going further and substituting points in an effort to find a and b, perhaps we should plot the data to see what their graph looks like. Here is a plot:



The plot clearly shows that the data do not fall on a (straight) line. We could verify this numerically from the table by showing that the slope changes as we move from one point to the next. Instead, in our search for an elementary function, we will trust our eyes and move on to quadratic functions; that is, functions of the form $f(t) = at^2$. (As a first step, we have decided not to use the most general form of a quadratic. If this does not yield good results, we will next try a more generate quadratic function whose graph is not symmetric about the origin.)

Substituting $t = .1$ and $at^2 = .049$ gives $a = 4.9$. Thus, the candidate is now $f(t) = 4.9t^2$ and we check it against other values in the table: $f(.2) = 4.9 \times .2^2 = 0.196$, $f(.3) = .441$, etc. Amazing! This function fits the data of the table exactly. (You might want to substitute the other values of the table to assure yourself of this fact.)

Now, we can complete our plan and answer the original question: The time it takes the object to fall from rest a distance of ten meters is found by solving the equation $4.9t^2 = 10$. Thus, the desired time is about 1.43 seconds.

1.1.2 Review of Modeling Issues So Far

This is a good place to review the key points of our work so far, before moving on to more complicated situations.

We began with a typical question about motion that calculus was developed to address. Rather than go right to the solution, we have approached the problem from the standpoint of someone who does not know calculus. Our aim is to identify those elements of the problem that will be crucial in our future development of calculus and its tools. The first key thing to note is that real-world problems very often involve discrete

data. This is true even though the underlying process may be continuous. When we translate the problem into mathematical terms, our first attempt usually involves some finite number of discrete measurements.

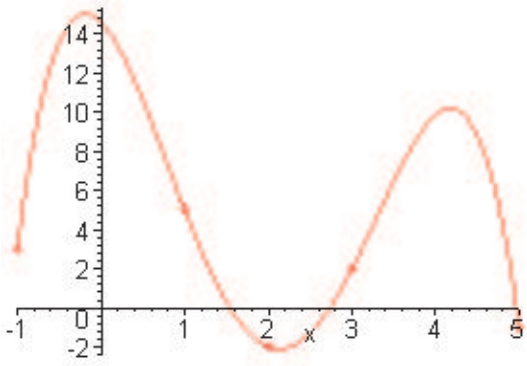
In the example we have considered of the object falling off a roof, we had the benefit of laboratory data. Someone had to go to the lab and conduct a more controlled experiment where it is possible to measure the distances that correspond to specific instants of elapsed time. We don't know the form of the function that lies behind the data. The best we can do is to observe and record positions at a finite number of times.

Applet: Falling Object **Try it!**

The next key thing to note is that there is usually an underlying function that fits the data and also takes on values at intermediary points. Furthermore, graphing the data is an important aid in identifying the function. So, we should expect functions and graphs to play major roles in our work modeling problems of motion. In fact, our mathematical modeling will take the form of asking the following question: Do we know of a mathematical function that fits the data and fills in intermediary points in a way that is consistent with the system we are observing?

Invariably, the answer to that question will be an elementary function, or a function related to an elementary function. We will be seeking a function by trial and error that comes as close as we can to hitting the data points.

You probably already know that given a finite number n of points, it is always possible to find a polynomial of degree at most $n-1$ that passes through them. However, this function usually is not satisfactory for modeling purposes because of its bahavior either between the data points or beyond them. For example suppose that we want a polynomial that passes through the points (-1, 3), (1, 5), (2, -2), (3, 1), and (5, -1). Since the general 4th degree polynomial $p(x) = ax^4 + bx^3 + cx^2 + dx + e$ has 5 unknown coefficients, we can expect to choose them in such a way that the 5 equations $p(-1) = 3$, $p(1) = 5$, $p(2) = -2$, $p(3) = 1$, and $p(5) = -1$ are satisfied. Solving these equations does indeed produce solutions for a, b, c, d, and e, and we graph the resulting polynomial $p(x)$ below.



In scientific applications it is rarely interesting to find a polynomial that passes *exactly* through a given set of data points. This is because the data are themselves usually approximations, and so finding a polynomial of degree n that passes through the n points is ascribing to the degree of the polynomial a meaning that has no scientific basis.

Applet: Falling Object **Try it!**

We will return to this issue in an example below involving census data. When we examined the lab data for the falling object, it was so precise that we were able to find a quadratic polynomial that fit the 10 points exactly. But suppose the points were merely close to the values of the quadratic and not exactly the same. We probably would still choose the quadratic over the 9th degree polynomial that passes exactly through the points. This is true because we have the sense that the quadratic works just as well for the intermediate points that correspond to the underlying continuous process. Hence, we want to emphasize this consideration when we choose a fitting function: It should work just as well at intermediate points, and at reasonably near points beyond the range of the data. The meanings of *just as well* and *reasonably near* will depend on the problem and the nature of the data we are attempting to model.

Moreover, there may be some characteristics of the physical system that inherently favor one function over another. We will see later when we develop some calculus ideas that this is the case for the falling object. We will be able to confirm that our choice of the quadratic function is correct on theoretical grounds.

1.1.3 Fitting a Curve to Data: The Method of Least Squares

Most data that comes from real-world situations is not fit exactly by any function of interest. This is primarily the case because the data themselves are only measured estimates. Take, for example, the US population census data (in thousands) for the years 1790 - 1850.

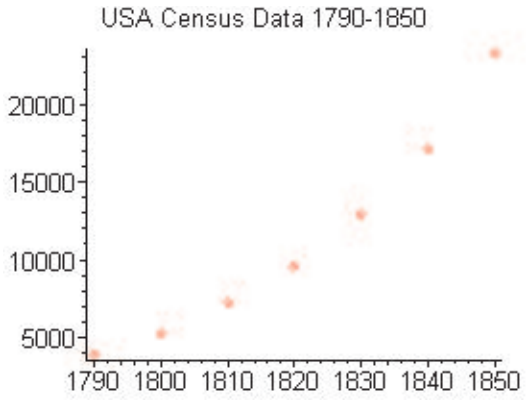
USA Census Data (thousands)	
Year	Population
1790	3929
1800	5297
1810	7224
1820	9618
1830	12901
1840	17120
1850	23261

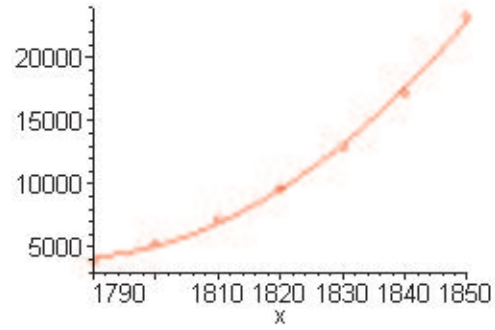
Let's fit a quadratic polynomial to the census data. But how do we do this? The most common approach, and the one implemented by computer algebra systems such as Maple, Mathematica, and Mathcad, is to use the method of Least Squares. In general, suppose we start with data points (x_1, c_1), (x_2, c_2), ..., (x_n, c_n) and we want to fit the data with the nth degree polynomial $y = p(x)$. Then we will define the Best Least Squares nth degree fit to be the polynomial of degree less than or equal to n that minimizes the sum of the individual squared errors, namely,

$$(p(x_1) - c_1)^2 + (p(x_2) - c_2)^2 + (p(x_3) - c_3)^2 + \ldots + (p(x_n) - c_n)^2 = \sum_{i=1}^{n} (p(x_i) - c_i)^2$$

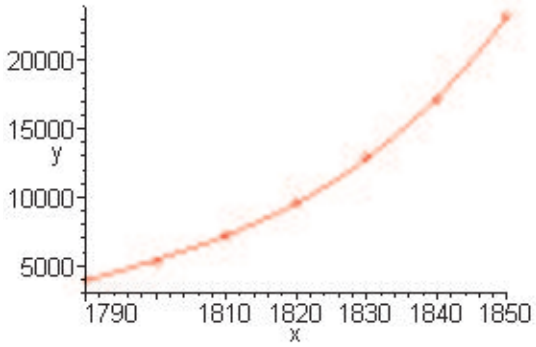
where we have used so-called sigma-notation to write the summation in a compact form. That is, the outcome of a computer implementation of the Least Squares method will be to determine the coefficients of the polynomial $p(x)$ that will minimize the sum of the squared errors. Squaring the errors does not allow the signed-errors to cancel one another in the summation; sqaring also has the effect of giving more weight to bigger errors than smaller ones.

Here are the results of using Maple to find the Least Squares quadratic fit of the census data. The quadratic polynomial is $q(x) = 4.417023810x^2 - 15766.11310x + .1407294507 \times 10^8$, and below we show its graph with the original data points.



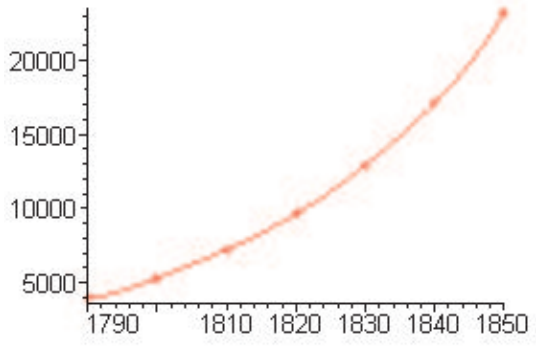


Visually, the fit looks pretty good. However, why stop with a quadratic? Next, let's fit a cubic polynomial to the data points. The Least Squares cubic fit is the polynomial $p(x) = .05097685185x^3 - 273.9165873x^2 + 490765.3753x - 293179528.2$; its graph also shows a good fit of the data.



Speaking again from a strictly visual perspective, the cubic fit actually looks *better* than the quadratic. In fact, we *know* that the cubic yields a smaller sum-of-squared-errors than the quadratic because in the Least Squares procedure, a quadratic function is always a possible outcome of fitting a 3rd degree polynomial to the data. Therefore, the sum of the squared errors for the cubic must be smaller than the sum of squared errors for the quadratic. In fact, given that we have formulas for the two polynomials, we can calculate directly the sum of squared errors for each. When we do this, we find them to be about 606852 for the quadratic and 47455 for the cubic. So, the cubic fit is considerably better according to this criterion.

But where does this leave us? First, we now have a mathematical criterion for deciding which of two different functions gives a better fit to a set of discrete data, namely, the one that has the smaller sum of squared errors. The essential consideration here is to be consistent with the criterion applied by the curve fitting computer algorithm we are using. Second, this criterion alone will not tell us which function to use to model the data. After all, with these 7 data points, we could always find the polynomial that pases through all of the points exactly and hence the sum of squared errors would be 0.



The 6th degree polynomial graphed above does just that: it hits every one of the 7 census points exactly. Moreover, it seems to give just as much of the proper feel at intermediate points as the cubic or the quadratic least squares fits. And yet we know that census data is inexact by its very nature; hence, it seems inappropriate to put the emphasis on obtaining an exact fit. Then, how do we decide what function to use to model the data? It is clear that we have to bring some other consideration to bear on the problem to decide among competing curves. For example, in the present case of the census data, it turns out that there are good theoretical reasons for the best fitting function not to be a polynomial at all, but rather an exponential. However, since we have not discussed exponential functions yet, we will have to postpone the completion of our analysis of the census data until we are familiar with exponential functions and their properties. We have reached a point in our discussion where we need to develop some new mathematical skills before we can continue further.

Applet: Least Squares Fitting **Try it!**

1.1.4 Our Agenda for This Chapter

We opened this section with an introduction to the kinds of questions that calculus can answer. We found that calculus gives us the tools to analyze motion and change. It does this by providing a system for studying and transforming functions. The functions are the tools whereby we represent moving and changing systems. The entire process starts with a set of discrete data and finding a function to model it.

Ultimately, calculus gives us the vocabulary to pose fundamental questions about a system, and the means to answer them. The elementary functions provide us with a robust library of functions which are sufficient for many of the purposes of calculus, enabling us to match the properties of the function with the observed behavior of the physical system. These *observables* are in the form of discrete data. Thus, the first step in modeling problems in calculus is to fit a (usually elementary, or related) function to the data.

The method of least squares is a powerful tool for determining the best fit. We normally do not seek a function that passes through all of the data points exactly because a set of data rarely describes an underlying system completely. The data are subject to experimental error and reflect other factors (such as friction in a moving system) that we do not choose to (or cannot) take into account. Thus, a function that passed through all of the data points would put too much emphasis on the data themselves and not the system. It is difficult to include all of the variables in a model because then the model can become too complicated to analyse. Thus, we look for both a mathematical reason and a systemic reason to choose our modeling function to be the simplest possible, making simplifying assumptions about our system along the way. The rest of the modeling involves studying the function and using the results to predict the behavior of the system.

We have raised many issues here, and have incidentally identified many gaps in our knowledge. In this section, we will start to fill them by:

1. Studying functions and their graphs.
2. Studying the elementary functions and their properties.
3. Getting more experience with modeling real discrete data.

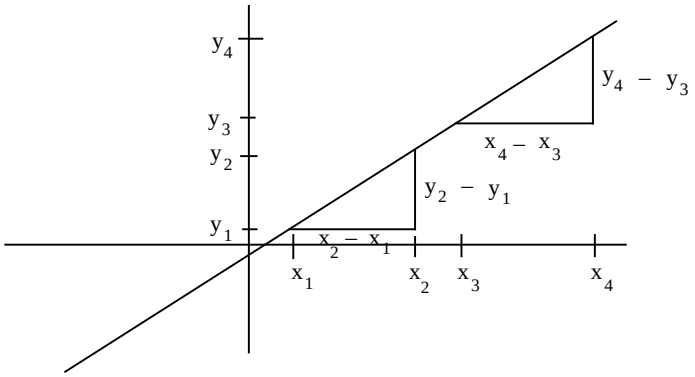
All too often, the study of calculus appears to consist of a long list of rules (the derivative of this, the integral of that). In actuality, calculus begins with rates of change of functions that model discrete data and describe elementary geometric shapes (e.g., lines, parabolas, circles). These modeling ideas are an indispensible part of the foundation of the subject, and form the rich and textured fabric that ties calculus to the moving and changing world around us.

Applet: Calculator: Values of Elementary Functions **Try it!**
Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

1.2 Lines in the Plane

We have just seen in Section 1.1 that when we want to find a mathematical expression to represent a set of points in the plane, it is natural to begin by considering if the points fall on a line. But what are the various mathematical expressions for a line and how do we recognize a line from its equation?

Let's start by considering first a line in the plane. We want to develop an equation in x and y that will represent it. To do this, we label four points (x_1, y_1), (x_2, y_2), (x_3, y_3), and (x_4, y_4) on the line.



Observe that the above triangles are similar because the corresponding angles are the same. Suppose now that given a right triangle whose hypotenuse has endpoints (x_1, y_1) and (x_2, y_2) we call $y_2 - y_1$ the *rise* and $x_2 - x_1$ the *run* where we form the differences in the same order (right point minus left, or vice versa, but in the same order for both the x and y coordinates). Then using this language, the similarity of the above triangles implies that the ratio of the rise to the run is the same for both these triangles:

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y_4 - y_3}{x_4 - x_3}$$

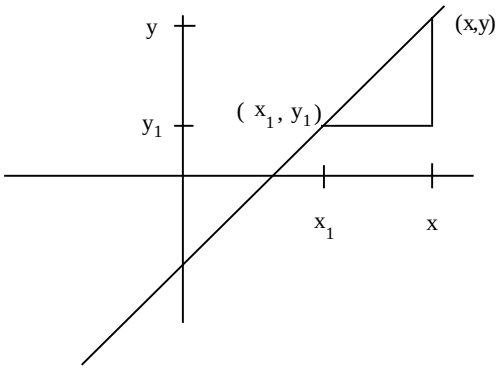
Thus, for any two different points on the line, the ratio of the rise to the run is the same constant m. Because this constant thereby characterizes the line, we give it a special name.

Definition 1: If a line passes through points (x_1, y_1) and (x_2, y_2), where $x_1 \neq x_2$, we call $m = \frac{y_2 - y_1}{x_2 - x_1}$ the *slope* of the line.

Note that slopes are not defined for vertical lines (which have equations $x = k$ where k is a constant), and that the slope of a horizontal line is equal to 0. Also, reversing the order of the points (left minus right, instead of right minus left) does not change the value of the slope:

$$\frac{y_1 - y_2}{x_1 - x_2} = \frac{-(y_2 - y_1)}{-(x_2 - x_1)} = \frac{y_2 - y_1}{x_2 - x_1}$$

Suppose now that we label two points on a line, (x, y) and (x_1, y_1), where we assume that the point (x_1, y_1) has known coordinates x_1 and y_1.



Then beginning with $\frac{y - y_1}{x - x_1} = m$, we can rewrite this equation to obtain an important form of equation of a line.

Point-Slope Form of Equation of a Line: Suppose that a line of slope m passes through the point (x_1, y_1). Then the line has equation $y - y_1 = m(x - x_1)$.

If, instead of assuming we know the value of the slope and the coordinates of a point on the line, we assume that we know the coordinates of two points, we can still write the equation in point-slope form. First we calculate the slope, and then use it together with the coordinates of either one of the points to write an equation of the line in the point-slope form. We will illustrate this in the examples below.

If we start with the point-slope form of equation, solve for y and distribute m over the parentheses on the right-hand side, we thereby obtain the form $y = mx + b$. (Following the line of reasoning here, $b = y_1 - mx_1$, but this is not so important.) Setting $x = 0$, we see that this means that the line passes through the point $(0, b)$; the point b on the $y-$axis is called the y-intercept. Conversely, if a line of slope m passes through the point $(0, b)$, then from the point-slope form we see that the line has equation $y - b = m(x - 0)$, or $y = mx + b$.

Slope-Intercept Form of Equation of a Line: Suppose that a line of slope m has y-intercept b. Then the line has equation $y = mx + b$.

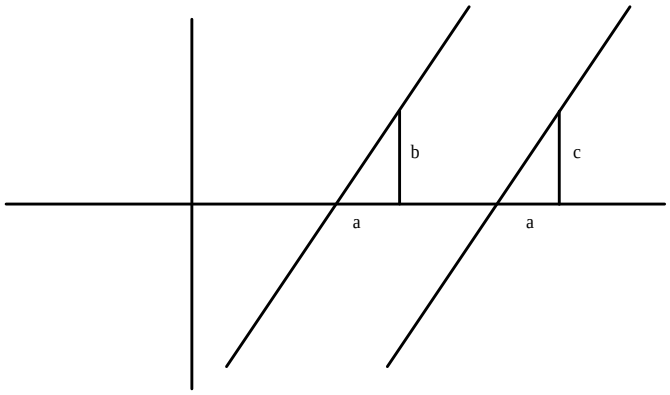
Finally, we see that by moving the terms involving x and y to one side of the equation and the constant terms to the other side, we can obtain an equation of the line in the so-called *general* form $ax + by = c$, where a, b, c are constants and the b here is not the same as the one above that we used to represent the y-intercept. Conversely, if we begin with an equation of the form $ax + by = c$, it is not too hard to see how to rewrite it to obtain, say, the slope-intercept form. (We will illustrate this in the examples below.) Hence, $ax + by = c$ is an equation of a line.

General Form of Equation of a Line: A line in the plane has an equation of the form $ax + by = c$ for some constants a, b, c; and conversely, an equation of the form $ax + by = c$, where a, b, and c are constants, represents a line in the plane.

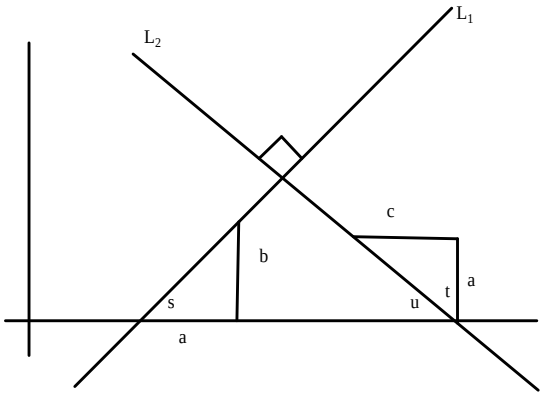
Example 1: To find an equation of the line through $(3, 4)$ and $(8, -6)$, we first find the slope of the line: $m = \frac{-6-4}{8-3} = -2$. Then from the point-slope form we obtain an equation $y - 4 = (-2)(x - 3)$.

Example 2: To find a general form of equation of the line through $(2, 7)$ and $(5, 6)$, we first find the slope: $m = \frac{6-7}{5-2} = -\frac{1}{3}$. Then we can arrive at a point-slope form: $y - 7 = -\frac{1}{3}(x - 2)$. And finally rewrite this equation in general form: $3y - 21 = 2 - x$, whence $x + 3y = 23$.

In closing our discussion of lines in the plane, we mention that two lines of slopes m_1 and m_2 are parallel if and only if $m_1 = m_2$; moreover, the lines are perpendicular if and only if $m_1 = -1/m_2$. These relationships can be discerned from the sketches that follow.



For example, if the two lines above are parallel, then the two triangles are congruent. Hence, $b = c$. Thus, the slopes of the two lines are the same, namely, b/a.



Now, suppose the two lines L_1 and L_2 above are perpendicular. Then, in degrees, $u + t = 90$ and $u + s = 90$ implies $s = t$. Thus, the two little triangles are congruent and $b = c$. Hence, the slope of one line is $m_1 = b/a$ and the slope of the other is $m_2 = -a/b$. That is, $m_1 m_2 = -1$ or $m_1 = -\frac{1}{m_2}$.

Example 3: Find an equation of the line through $(1, 2)$ and parallel to the line with equation $5x-7y = 2$. The slope of the latter line can be found by solving for y to obtain the slope-intercept form of the equation: $y = \frac{5}{7}x - \frac{2}{7}$. Thus, the line has slope $5/7$. So, in point-slope form, the equation of the line we seek is $y - 2 = (5/7)(x - 1)$.

Example 4: To find an equation of the line through $(1, 2)$ and perpendicular to the line with equation $5x-7y = 2$, we note from the previous example that the slope must be $m = -7/5$ (the negative reciprocal of the slope of the line whose equation is given). Hence, the equation in point-slope form is $y-2 = (-7/5)(x-1)$.

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

1.3 Functions and Their Graphs

At the heart of calculus lie two fundamental concepts—*function* and *limit*. From them are derived several additional basic concepts—*continuity*, *derivative*, and *integral*. It is the study of these several concepts and the mathematical techniques that accompany them that constitutes the subject of calculus. The applications of the concepts give calculus its power to model and understand dynamic systems. Understanding the concept of function thus becomes a first priority when studying calculus.

A function f is a correspondence between objects x in its domain and objects $f(x)$ in its range. For example the square function might be represented as $f(x) = x^2$, emphasizing the algebraic definition of the function. Or it might be represented as $f : x \to x^2$, emphasizing the correspondence or mapping between a number x and its square. The latter notation is very common in mathematics. In either case values of the function for given values of x are denoted by $f(x)$, as in $f(2) = 4$, $f(5) = 25$, or $f(-2.5) = 6.25$. In general, then:

Definition1: A function f on a set D into a set S is a rule that assigns a unique element $f(x)$ in S to each element x of D. The set D is called the *domain* of the function f and the subset $\{f(x) \in S : x \in D\}$ of S is called the *range* of f. It is common in this context to call x the *independent variable* because we assign its value, and y the *dependent variable* because we compute its value.

For instance, $f(x) = mx + b$, where m and b are real numbers, is a function. It takes any real number x and produces another real number $mx + b$. (If $m = 2$ and $b = 1$, then $f(1) = 3$.) Thus, the domain is the set of all real numbers; so is the range. The function f is a so-called *linear* function: the points $(x, f(x))$ lie on the line $y = mx + b$ in the plane.

Example 1: Let f be the function that maps each real number into its square. In this case the rule defining f may be given by an algebraic formula $f(x) = x^2$. The domain of f is the set of all real numbers. The range of f is the set of all non-negative real numbers.

Example 2: $g(x) = \sqrt{x}$, the function that maps real numbers into their real square roots. We take the domain of f to be the set of all non-negative real numbers. And then the range of f is also the set of all non-negative real numbers. (Why?)

Example 3: $f(x) = \sqrt{1 - x^2}$. A function is not completely defined, of course, until its domain is specified. In calculus it is conventional to assume that the domain is the set of all real numbers for which the value of the function is defined and is a real number. This would require that $0 \leq 1 - x^2$. Thus the domain of f is

$$\{x : 0 \leq 1 - x^2\} = \{x : x^2 \leq 1\} = \{x : -1 \leq x \leq 1\}$$

Example 4: Consider the function $y = 1/x$. This time we have specified the function, one that maps real numbers into their reciprocals, by writing an equation relating the independent variable x and the dependent variable y. We could also have written $x \to 1/x$ to indicate the mapping, or we could have explicitly named the function $h(x) = 1/x$. Suffice it to say that we have many ways to specify the rule that defines the function, and we will use the one that is most convenient in a given situation. To complete this example we note that $1/x$ is not defined when $x = 0$. Since we have not otherwise specified the domain, we follow the *calculus domain convention* (cf Example 3) and take the domain to be the set of all real numbers different from zero, i.e. the set $\{x : x \neq 0\}$. This set would also commonly be written as the union of two intervals: $(-\infty, 0) \cup (0, \infty)$. (Remember that parentheses in the interval notation indicate that the end point of the interval is *not* included in the set, and square brackets mean that the end point *is* included.)

Functions can be represented algebraically, numerically, or graphically. When, for example, we state Newton's law of gravitation in the form

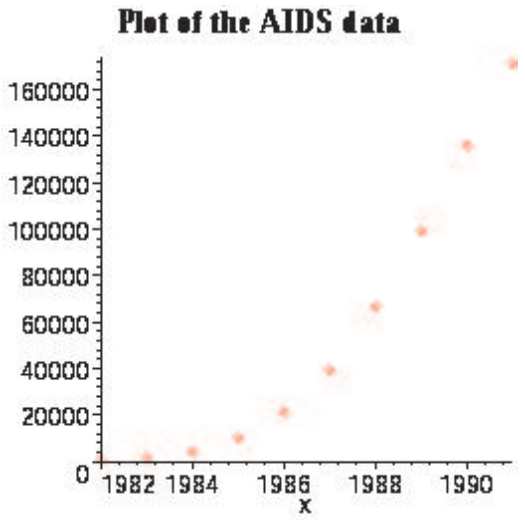
$$gravitationalforce = \frac{k}{r^2}$$

where k is a constant and r is the distance separating two bodies, we are giving an algebraic definition of a function. In this case the function expresses the inverse square law that relates the gravitational force to the distance r separating the two bodies. Often no such algebraic formula exists to express the functional

relationship between two quantities but instead a table of numerical data is given. For example the spread of the AIDS virus in the United States during the first ten years following its discovery in 1981 is given in a table published by the Center for Disease Control:

Year	No of AIDS cases
1982	295
1983	1374
1984	4293
1985	10211
1986	21278
1987	39353
1988	66290
1989	98910
1990	135614
1991	170851

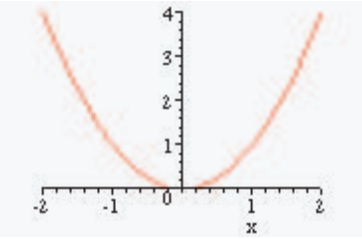
Such examples are common in science, where data is gathered in the laboratory or in the field. We can interpret this table as defining a function $AIDS(n)$ with finite domain, that gives the correspondence between the year n and the cumulative number of AIDS cases up to that year. Although the table is an efficient presentation of the AIDS function, it does not permit us to visualize its overall behavior. For that a graph of the data is most instructive.



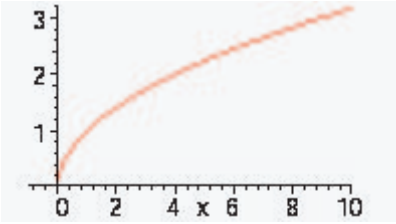
Alternatively, we can imagine a continuous function $f(x)$ such that $f(n) \approx AIDS(n)$ for each n and the graph of f runs smoothly between the data points. Such a function $f(x)$ is said to *model* the data. Finding a function that "fits" the data and that is also based on known scientific principles (in this example epidemiological principles) is an important problem in science. This is one of the most compelling reasons for studying the family of elementary functions of calculus—polynomials, rational functions, trigonometric functions, and exponential and logarithmic functions. Each of these classes of functions possesses its own unique and distinguishing properties, and choosing a function that models a given set of data involves matching those properties against the requirements of a particular scientific problem.

Graphing Functions: The graph of a function f often reveals its behavior more clearly than tabular or algebraic representations, thus familiarity with the graphs of selected basic functions is an important precursor to studying calculus. The graph of f is just the set of points $\{(x, y) : x \in \text{domain}, y = f(x)\}$. It is usually the geometrical picture of this set of points, plotted on an xy-coordinate system, that we have in mind when we use the term *graph*. For example the following are graphs with which you should be familiar:

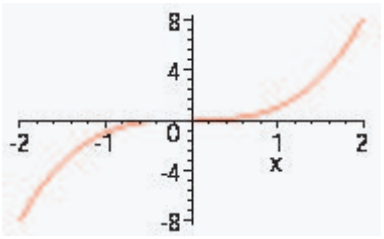
Example 5: $y = x^2$. The graph is the set of points $\{(x, x^2) : x \text{ is real}\}$. In drawing the picture we choose to plot the points in the range $-2 \leq x \leq 2$. We accept the fact that our geometrical picture of the graph does not cover the entire domain of the function. Normally we choose an x-range and y-range that reveal the most important behavior of the function.



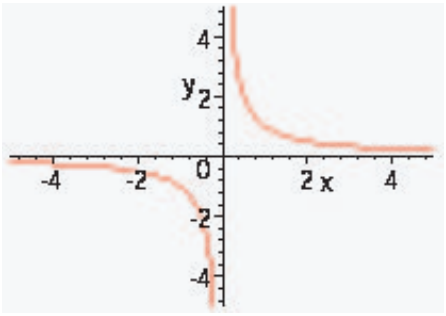
Example 6: $f(x) = \sqrt{x}$. In this example the domain of f is the interval $[0, \infty)$, and we choose the x-range of the plot to be $-1 \leq x \leq 10$.



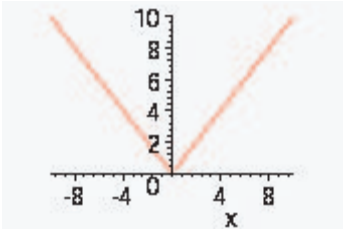
Example 7: $f(x) = x^3$. We choose the x-range $-2 \leq x \leq 2$.



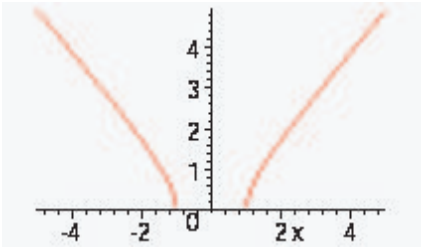
Example 8: Graph of $f(x) = \frac{1}{x}$.



Example 9: Graph of $f(x) = |x|$.



Example 10: Graph of $f(x) = \sqrt{x^2 - 1}$. Here we must have $0 \leq x^2 - 1$. Hence the domain of f is $(-\infty, -1] \cup [1, \infty)$.

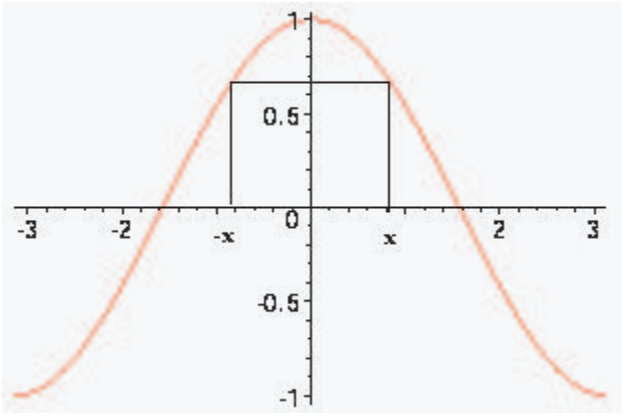


Graphs of functions may be drawn, albeit very tediously, by plotting many points and connecting them with a "smooth curve". Calculators or computers can be valuable tools in such pursuits. But the downside of drawing graphs in this way is that it obscures the most important properties of the functions and misdirects our attention to trivial computational details. A sketch of a graph, embodying the functions signature features, is often all that we need. Is the function positive or negative? Where does its graph cross the coordinate axes? Does it "shoot off to infinity"? Are there any gaps in its domain?

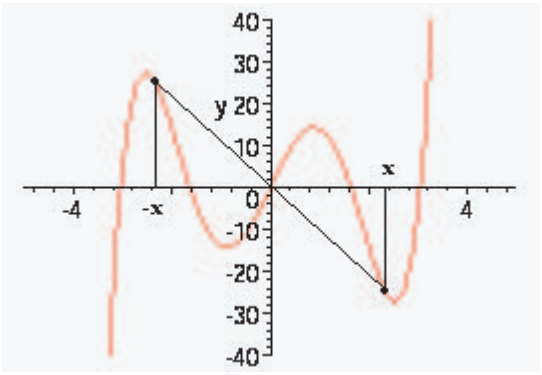
Fortunately, we will see that a little knowledge of graphs goes a long way. From the familiar graphs of a small number of basic functions we can recognize and sketch the graphs of many more functions that are related to them. We end this section with such an example—recognizing and exploiting symmetries in the graphs of functions.

Applet: Function Grapher **Try it!**

Even and Odd Functions; Symmetry and Reflections: In examples 1, 5 and 6, above, the graphs were symmetric about the y-axis. A point (x, y) lies on one of those graphs if and only if its mirror image $(-x, y)$ in the y-axis is also on the graph.



Examples 3 and 4 display a different kind of symmetry. A point (x, y) lies on the graph in Example 3 if and only if the point $(-x, -y)$ is also on the graph. The graph in this case is said to be "symmetric" about the origin.

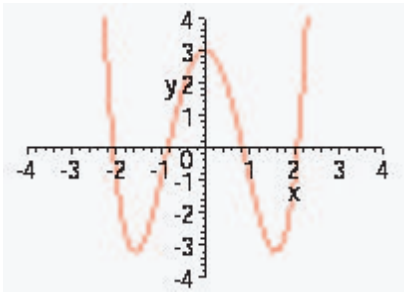


Definition 2: A function f is said to be an *even function* if $-x$ is in its domain whenever x is, and $f(-x) = f(x)$. Such a function is symmetric about the y-axis.

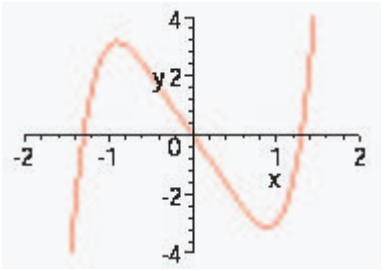
Definition 3: A function f is said to be an *odd function* if $-x$ is in its domain whenever x is, and $f(-x) = -f(x)$. Such a function is symmetric about the origin.

When a function is recognized as either an even or an odd function, its graph can be drawn exploiting the symmetry. Only half of the graph, for positive values of x, need be drawn. It can then be completed by reflecting in the y-axis (for an even function) or the origin (for an odd function).

Example 11: The function $f(x) = x^4 - 5x^2 + 3$ is an even function. Notice that replacing x by $-x$ does not change the function; i.e. $f(-x) = f(x)$. Its graph displays symmetry about the y-axis.



Example 12: On the other hand, the function $f(x) = 2x^5 - x^3 - 4x$ is an odd function. In this case it is clear that $f(-x) = -f(x)$. Only odd powers of x appear in the expression, thus if x is replaced by $-x$ the minus sign can be "factored out".



Notice that an odd function, if it is defined for $x = 0$, must have the value zero there. For then it would have to satisfy $f(-0) = -f(0)$, and this implies that $f(0) = 0$. Why?

Applet: Symmetry, Odd and Even Functions **Try it!**

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

1.4 Defining New Functions from Old

The notion of symmetry discussed in the previous section can be used to define new functions from old. For example the graphs of $y = \sqrt{-x}$ and $y = \sqrt{x}$ are mirror images of each other in the y-axis. Each has a domain that is only half of the real axis, hence cannot be an even function. However, replacing x by $-x$ in the latter has the effect of reflecting its graph in the y-axis, turning it into the graph of the former. More generally:

Theorem 1: Reflections in special lines: For an equation in x and y

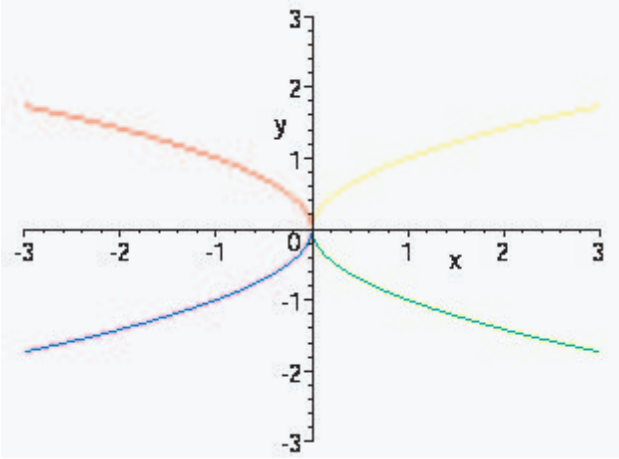
1. Replacing x by $-x$ corresponds to reflecting the graph of the equation in the y-axis.

2. Replacing y by $-y$ corresponds to reflecting the graph of the equation in the x-axis.

3. Replacing both x and y by their negatives corresponds to reflecting the graph of the equation in the origin.

4. Interchanging x and y in an equation corresponds to reflecting the graph of the equation in the line $y = x$.

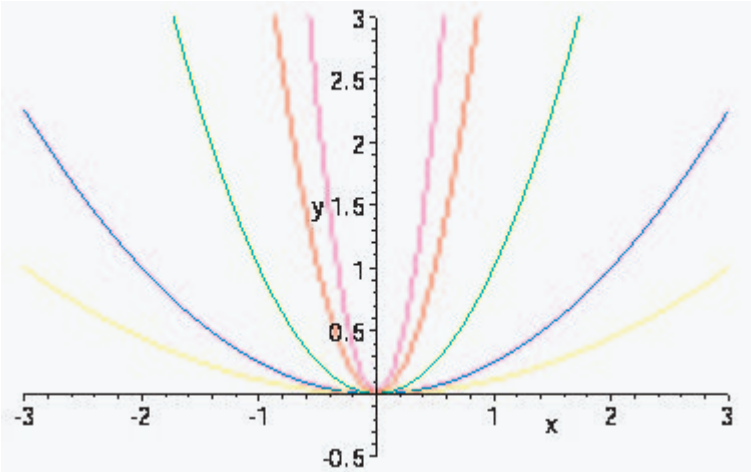
Example 1: The plot below shows the graphs of four functions: $\sqrt{x}$, $\sqrt{-x}$, $-\sqrt{x}$, and $-\sqrt{-x}$. Identify in the plot the graph of each of these functions by plotting a few points for each. To get started, verify that $\sqrt{x}$ is in the upper right-hand quadrant, and $-\sqrt{x}$ is in the lower right-hand quadrant.



Even though we often can define new functions from existing ones using symmetry, acquiring a sufficient repertoire of functions for solving all the problems that can arise may seem hopeless. Modeling real-life and scientific situations places strenuous demands upon our mathematical experience and knowledge. As we mentioned in Section 1.1, however, a small number of basic functions, most of which already occur in your precalculus study, can serve as *building-blocks* for a much wider and very useful class of functions, the *Elementary Functions*. It is on those building-blocks that we focus in this section. And it is with the class of elementary functions under our belts that we will begin our study of calculus.

In Section 1.2 we already noted some of the ways in which new functions can be defined from old. The function $f(x) = \sqrt{-x}$, for example, is closely related to the function $g(x) = \sqrt{x}$. We can think of it as obtained by reflecting (the graph of) $g(x)$ in the y-axis.

Scaling a graph: Another simple geometric transformation that yields new functions from old is a *stretch*—either parallel to the x-axis or parallel to the y-axis. For example compare the graphs of $y = x^2$ and $y = (cx)^2$, for various values of the constant c. When x is replaced by cx, horizontal distances are multiplied by the factor $1/c$. When $0 < c < 1$ this is seen geometrically as "stretching" the graph of x^2 parallel to the x-axis. And when $c > 1$ it appears as a "compression" of the graph.

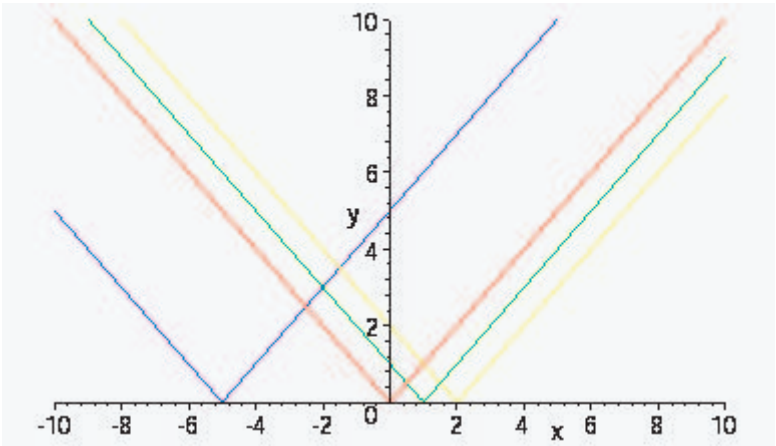


Identify in this plot the graphs of $(\frac{1}{3}x)^2$, $(\frac{1}{2}x)^2$, x^2, $(2x)^2$, and $(3x)^2$. Notice that we only need to be familiar with the one basic function x^2 in order to "understand" the other functions $(cx)^2$.

Theorem 2: Replacing x by cx in a function $y = f(x)$ results in a horizontal stretching or compression of the graph of f. When $0 < c < 1$ the graph is *elongated* horizontally by the factor $1/c$, and when $c > 1$ it is *compressed* horizontally by the factor $1/c$.

Applet: Stretching Graphs **Try it!**

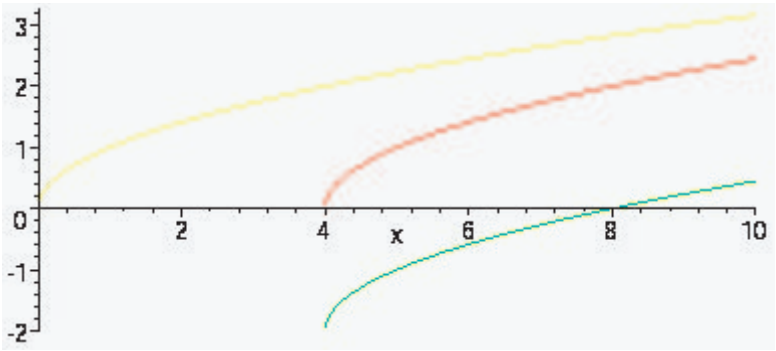
Shifting a graph: The graph of a function may be shifted a units horizontally by replacing x by $x - a$. In the following plot, for example, the graphs of $|x|$, $|x - 1|$, $|x - 2|$, and $|x + 5|$ are shown. Notice that replacing x by $x - 2$ shifts the graph of $|x|$ 2 units to the *right*. And replacing x by $x + 5$ shifts it 5 units to the *left*.



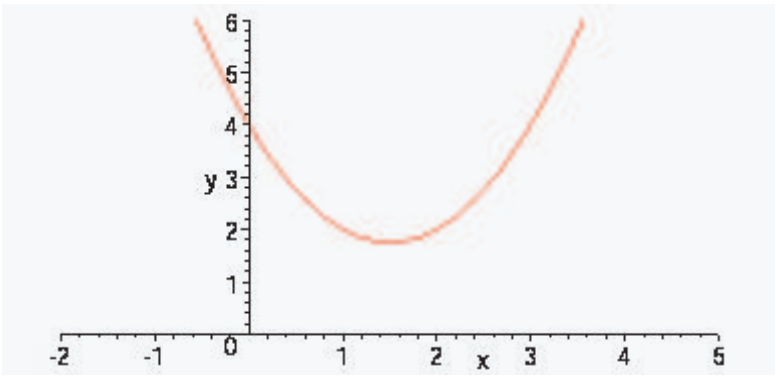
Theorem 3: Assume that the constant a is positive. To shift the graph of a function $f(x)$ to the right by a units, replace x by $x - a$. To shift it to the left by a units, replace x by $x + a$.

Vertical stretching and shifting may also be accomplished. The graph of $cf(x)$ is obtained by elongating the graph of $f(x)$ by the factor c in a direction parallel to the y-axis. And the graph of $f(x)+a$ is obtained by shifting the graph of $f(x)$ upward a units. We will not state these two additional geometrical transformations as separate theorems, but they will often be applied in graphing elementary functions.

Example 2: Draw the graph of the function $y = \sqrt{x - 4} - 2$. We compare this function with $\sqrt{x}$. Replacing x by $x - 4$ shifts the graph 4 units to the right. Then subtracting 2 shifts the graph 2 units *down*.



Example 3: Draw the graph of $y = x^2 - 3x + 4$. Completing the square in this quadratic expression, we write the equation in the form $y = (x - \frac{3}{2})^2 + \frac{7}{4}$. We then recognize the graph as that of x^2, shifted 3/2 units to the right and 7/4 units up.



Applet: Shifting Graphs **Try it!**
Applet: New Functions from Old **Try it!**
Applet: New Functions from Old Game **Try it!**

Arithmetical operations and Composition: Another way to build new functions from old is implicit in our use of algebraic expressions to define functions. The polynomial $P(x) = x^3 + 5x$ for example is constructed from the constant function 5 and the identity function x using addition and multiplication. (From 5 and x we obtain $5x$ by multiplication. $x^3 = x \cdot x \cdot x$ is also obtained from x by multiplication. And $P(x)$ is just the sum of these two functions. Similarly the rational function $R(x) = \frac{x^3+5x}{3x^2+1}$ is obtained from the constant functions 1, 3, 5, and the identity function x using division along with addition and multiplication. Finally, the function $f(x) = \sqrt{x^3 + 5x}$ is obtained by *composition* from the polynomial $P(x)$ and the square root function $\sqrt{x}$. We may think of composition as using the functions P and $\sqrt{}$ sequentially, first applying P to x, then applying $\sqrt{}$ to the result.

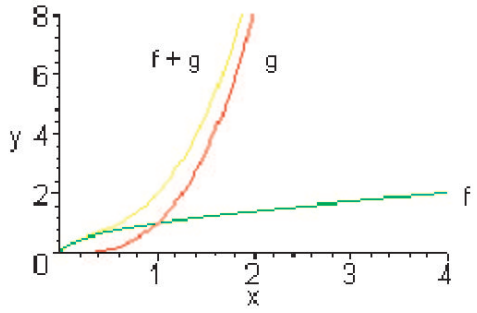
Such use of arithmetical operators and the operation of composition to define new functions is so prevalent in algebra that we rarely give them second notice. For the record, however, we record the building of new functions in these ways in formal theorems:

Definition 1: Let f and g be functions and let x be in the domain of both functions. Then the functions $f + g$, $f - g$, fg and f/g are defined by the rules:

1. $(f + g)(x) = f(x) + g(x)$
2. $(f - g)(x) = f(x) - g(x)$
3. $(fg)(x) = f(x) \cdot g(x)$
4. $(f/g)(x) = f(x)/g(x)$, when $g(x) \neq 0$

Example 4: As an illustration of Definition 1, let $f(x) = \sqrt{x}$ and $g(x) = x^3$. Then $(f+g)(x) = \sqrt{x}+x^3$ for all x such that $x \geq 0$. That is, $(f + g)(x)$ is defined for any nonnegative value of x to be the sum of the

square root of x and the cube of x. Hence, for example, $(f+g)(4) = 2+64 = 68$. As in the graph below, $f+g$ is a new function which we define pointwise to be the sum of the values of f and g.



Definition 2: Let f and g be functions, let x be in the domain of f, and $g(x)$ in the domain of f. Then the *composite function* $f \circ g$ is defined by the rule

$$(f \circ g)(x) = f(g(x)).$$

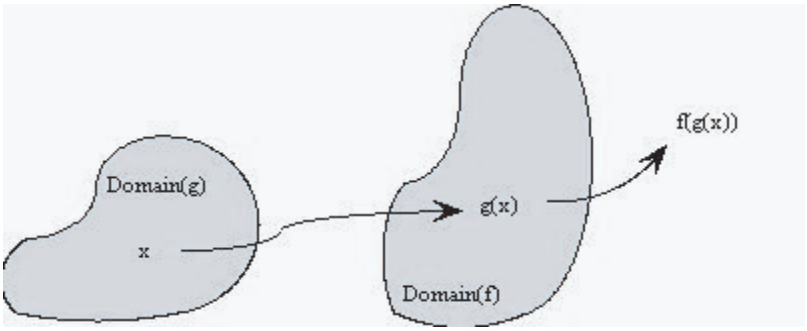
Using the language of mappings,

$$x \overset{g}{\longmapsto} g(x) \text{ and } g(x) \overset{f}{\longmapsto} f(g(x)).$$

We may think of the operation of *composition* as "joining the arrows", yielding

$$x \overset{f\circ g}{\longmapsto} f(g(x)).$$

The domain of $f \circ g$ is $\{x : g(x) \in \text{domain of } f\}$.



Example 5: Let $f(x) = x^2 - 4$ and $g(x) = x - 2$. Then

$$(f \circ g)(x) = f(g(x)) = g(x)^2 - 4 = (x-2)^2 - 4 = x^2 - 4x.$$

Example 6: Let $f(x) = 2 + \sqrt{x}$ and $g(x) = 4 - x^2$. Then

$$(f \circ g)(x) = 2 + \sqrt{g(x)} = 2 + \sqrt{4 - x^2}.$$

What is the domain of $f \circ g$? We must determine for which real numbers x is $g(x)$ in the domain of f. This requires that $g(x) \geq 0$, i.e. that $4 - x^2 \geq 0$. Thus $x^2 \leq 4$ or, finally, the domain of $f \circ g$ is $-2 \leq x \leq 2$.

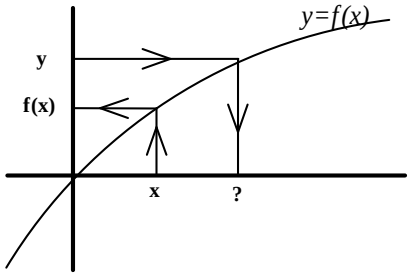
Applet: Arithmetical Operations on Functions **Try it!**

Inverse Functions: The final operation for building new functions from old is that of *taking inverses*. A familiar example is provided by the square root function $f(x) = \sqrt{x}$ and the square function $g(x) = x^2$. Their inverse relationship is exhibited (for $x > 0$) by the identities $\sqrt{x^2} = x$ and $(\sqrt{x})^2 = x$. In the form of equations we say that $y = x^2$ can be "solved for x" as $x = \sqrt{y}$; or, conversely, $y = \sqrt{x}$ can be "solved for x"

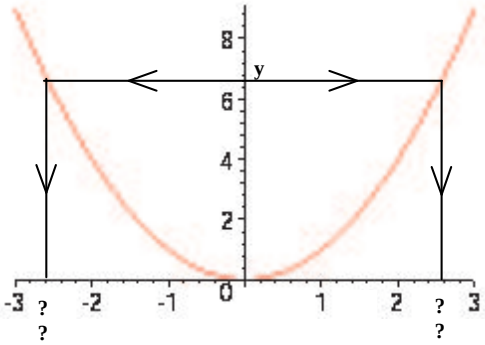
as $x = y^2$. The relationship can also be viewed as reversing the arrow in the mapping that defines either of these two functions: the mapping $x \longmapsto y = x^2$ has the *inverse* mapping $y \longmapsto x = \sqrt{y}$.

Alternatively, if one considers a function to be a set of ordered pairs (x, y), then the corresponding *inverse relation* is the set of ordered pairs (y, x). If this inverse relation is also a function, then we call it the *inverse function*.

Or, if one thinks of the graph of a function f, computing values of $f(x)$ amounts to starting at a point x on the x-axis, tracing a vertical line to the graph and then a horizontal line to the y-axis. Computing the inverse relation of f amounts to reversing the direction of the arrows (see below), starting with a value y on the y-axis, tracing a horizontal line to the graph and then a vertical line to the x-axis. We will denote the "landing point" on the x-axis as $f^{-1}(y)$ and if f^{-1} defined by this rule is a function, we will refer to it as the *inverse of* f.



We hasten to dispose of the slight problem that the inverse relation of f is not always a function. In our picture there was a unique value on the x-axis corrsponding to a given value on the y-axis. But this need not be the case in general. We need look no further than the square function with which we began the discussion of inverse functions. Starting with a given value of y on the y-axis, which of the reverse arrows is to be followed to define an inverse function (see the following plot)? The problem arises, of course, from the fact that there are multiple values of x on the x-axis that are mapped to the *same* value of y. Under such circumstances the inverse mapping (or relation) is not a well-defined function. And we say that the function f does not have an inverse.



The solution is to restrict the domain of f so that the ambiguity does not arise. For the square function $f(x) = x^2$ we can do this, for example, by restricting the domain to the interval $0 \leq x \leq \infty$. Then there is a one-to-one correspondence between points x on the *positive* x-axis and points $y = f(x)$ on the *positive* y-axis. In such a case we say that the function f is 1–1, and then the inverse mapping as we have described it is unique.

Definition 3: A function f is said to be 1–1 if $f(x_1) = f(x_2)$ implies that $x_1 = x_2$. In other words different values of x are mapped to different values of y. Such a function is also said to pass the *horizontal line test*, in the sense that every line parallel to the x-axis intersects the graph of f in at most one point.

Theorem 4: If f is a 1–1 function then it has an inverse function which we will denote by f^{-1}. (Caution: do not confuse this with $1/f$, the reciprocal of f.) The domain of f^{-1} is the *range* of f; and the range of f^{-1} is the *domain* of f. The functions f and f^{-1} satisfy

$$y = f^{-1}(x) \qquad \text{if and only if} \qquad f(y) = x.$$

Example 7: We used the function $f(x) = x^2$ above as an example of a function that is not 1–1 on its full domain but that *is* 1–1 when its domain is restricted suitably. On the positive x-axis it is 1–1, and its inverse is called (the positive) square root of x. (We could have chosen the negative real axis as the restricted domain of f, and then the inverse function would have been the *negative* square root.

Example 8: Show that the function $f(x) = 5x + 2$ is 1–1 and find f^{-1}. Since the graph of f is a straight line that is not horizontal, it clearly passes the *horizontal line test*; thus f is 1–1 on its entire domain $(-\infty, \infty)$. Algebraically, we see that

$$f(x_1) = f(x_2) \Rightarrow 5x_1 + 2 = 5x_2 + 2 \Rightarrow x_1 = x_2.$$

To find the inverse function we follow the steps:

Step 1 $y = 5x + 2$

Step 2 Solve for x: $x = \frac{y-2}{5}$

Step 3 Reverse the roles of x and y: $y = \frac{x-2}{5}$

Step 4 Thus: $f^{-1}(x) = \frac{x-2}{5}$

Example 9: Find the inverse function of $f(x) = \sqrt{3x-1}$. To see that f is 1–1 we notice that

$$\sqrt{3x_1 - 1} = \sqrt{3x_2 - 1} \Rightarrow 3x_1 - 1 = 3x_2 - 1 \Rightarrow x_1 = x_2.$$

Then, following the same steps as in Example 8 we have

Step 1 $y = \sqrt{3x-1}$

Step 2 Solve for x: $y^2 = 3x - 1$, so $x = \frac{y^2+1}{3}$

Step 3 Reverse the roles of x and y: $y = \frac{x^2+1}{3}$

Step 4 Thus: $f^{-1}(x) = \frac{x^2+1}{3}$

Applet: Inverse Functions **Try it!**

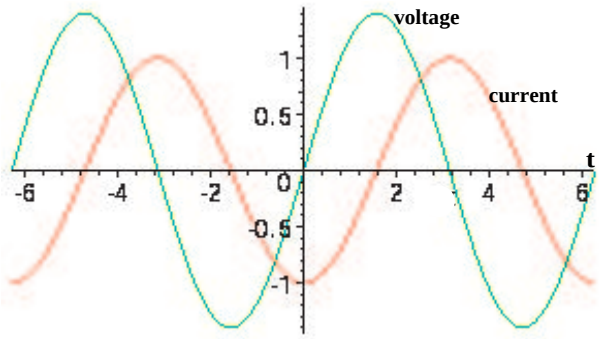
Summary: This completes our survey of building blocks for defining new functions from old. Through arithmetic operations, composition, and taking inverses, we will build up our stable of functions—the so-called Elementary Functions—that suffice for much of calculus. We need to add only the trigonometric functions and the exponential and logarithm functions as "simple starting functions" for our building exercise. In the next section we handle the trigonometric functions. And the exponential and logarithm functions will follow thereafter.

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

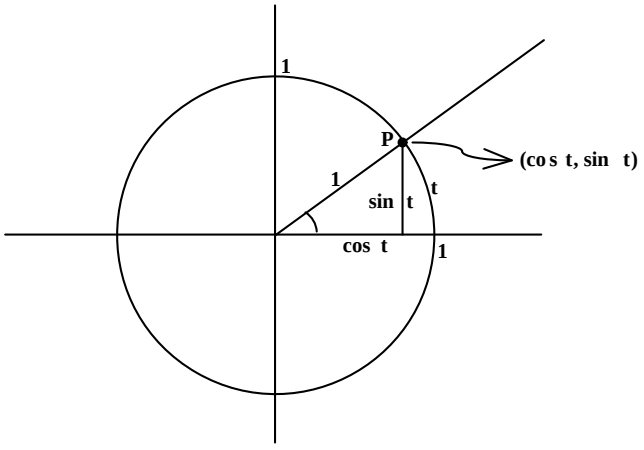
1.5 Trigonometric Functions

Modeling with Trigonometric Functions: You first met the trigonometric functions in algebra and trigonometry in high school. In a typical trigonometry course the functions $\sin x$, $\cos x$, and the other related functions $\tan x$, $\sec x$, etc., are defined as ratios of sides in right triangles. The focus is on measuring the sides and angles of triangles, hence the term *trigonometric functions*. Much of your attention was directed to applications in geometry growing out of this connection with right triangles as well as to identities that express relationships between the several trigonometric functions. In calculus the focus changes. The trigonometric functions are defined in terms of arclength on a unit circle, and the emphasis is on the periodic behavior of the trigonometric functions. It is their periodicity that leads to their most important applications in science—modeling phenomena that repeat as a function of time. Simple harmonic motion (sinusoidal motion), light and sound waves, electricity, gravitational waves in the universe, oscillations of the pendulum of a clock, oscillations of atomic crystals on which our most accurate time keeping is based—all these are periodic phenomena that are modeled mathematically by the trigonometric functions. The essential characteristics of any periodic motion are the *amplitude* (how big are the oscillations?) and *period* (how long is a single wave?). Sometimes the *phase* (how much has the wave been translated to the right or delayed in time?) is also important, as when one is comparing two related wave forms such as voltage and current—in the standard way of generating electricity the current lags the voltage by 90 degrees (the phase angle is $\pi/2$. Their graphs might look as follows:



Here the amplitude of the voltage is 1.4, the amplitude of the current is 1, both have period 2π, and the current lags the voltage by a phase angle of $\pi/2$. The *period* tells us how long a single wave is, and, especially when the independent variable is time, this is sometimes described instead in terms of *frequency* (how many oscillations occur in a unit distance or unit of time).

Definition 1: The Trigonometric Functions: In calculus we define the trigonometric functions in terms of arc length on a unit circle. Choose a point P on the circle (see the plot below), at a distance t from the positive x-axis measured counterclockwise along the circle. Then the trigonometric functions $\cos t$ and $\sin t$ are defined to be the coordinates of the point P.



The arclength t in the plot is a measure of the central angle that subtends the arc. It is called the *radian*

measure of the angle. An angle of 360° subtends the entire circle whose length is 2π, thus $360° = 2\pi$ radians, and $1° = \pi/180$ radians. Also $180° = \pi$ radians, $90° = \pi/2$ radians, and $60° = \pi/3$ radians We will consistently use radian measure of angles in calculus.

Reference to the figure also tells us that our new definition of the trigonometric functions is consistent with the usual definitions as ratios of sides of a right triangle. In the small right triangle in the figure with angle t (radians), the hypotenuse of the triangle is 1, the adjacent side has length $\cos t$, and the opposite side has length $\sin t$. Thus, for example, the ratio of the adjacent side to the hypotenuse is $\frac{\cos t}{1}$, as it should be. The main thing to remember is that we are measuring angles in radians instead of degrees, so $\cos t$ and $\sin t$ are functions of the *real variable* t. There is thus no longer any reason to restrict our attention to angles less than 360°. We may measure any positive distance along the unit circle that we wish, and since the point P on the circle returns to its starting point when $t = 2\pi$, we note that the values of $\sin t$ and $\cos t$ repeat when we "wrap around" the circle, i.e. when t increases by a multiple of 2π. This gives us the fundamental *periodic* behavior of $\sin t$ and $\cos t$, the property that makes them so useful in modeling periodic phenomena.

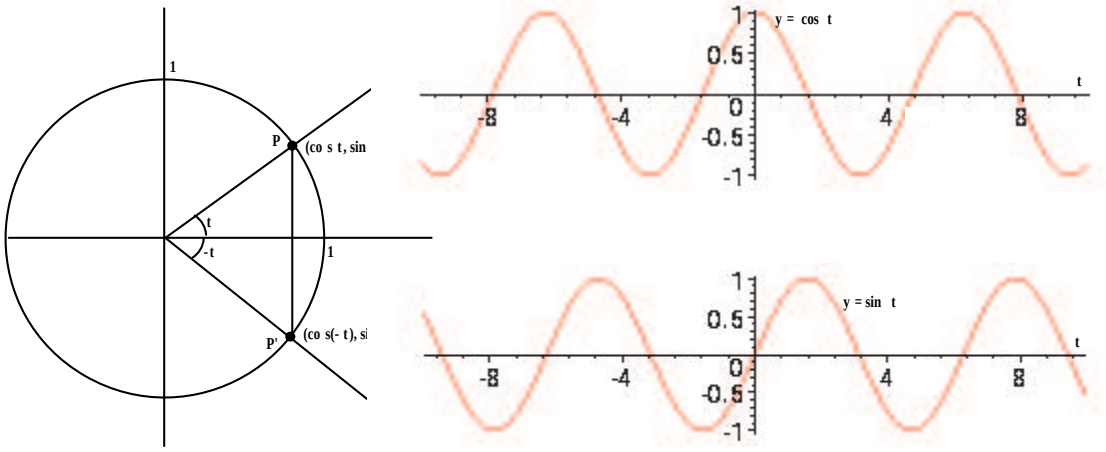
We can also allow t to take on negative values. These simply correspond to distances measured *clockwise* along the circle, beginning at the point $(1, 0)$. The sin and cos functions thus have domain $-\infty < t < \infty$.

Applet: Definitions of sin(x)and cos(x) **Try it!**

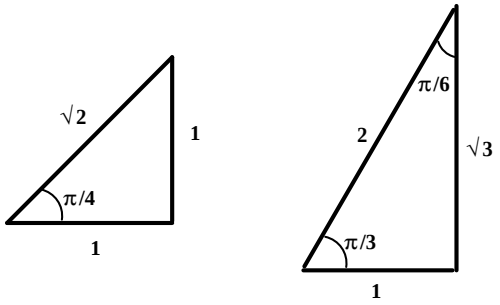
Theorem 1: The trigonometric functions sin and cos are defined for all real values of t, and are *periodic with period* 2π. I.e. they satisfy

$$\sin(t + n \cdot 2\pi) = \sin t \text{ for any real } t \text{ and any integer } n.$$

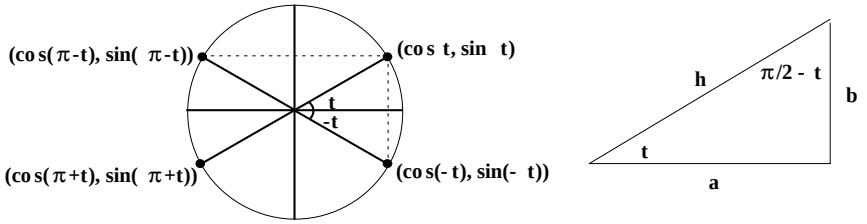
Many of the familiar properties of the trigonometric functions follow immediately from their definition. Since $(\sin t, \cos t)$ is a point on the unit circle it is clear that $\sin^2 t + \cos^2 t = 1$ for all values of t. (Note: we follow the usual convention of writing $\sin^2 t$ instead of the more cumbersome $(\sin t)^2$; however, when doing calculations, we do type the latter so that a computer can understand it.) Moreover it is also clear that $\cos(-t) = \cos t$ and $\sin(-t) = -\sin t$ (see the figure below), thus cos is an *even* function and sin is an *odd* function. These symmetries and the periodicity, along with special values such as $\cos 0 = 1$, $\cos \pi/2 = 0$, $\cos \pi = -1$, $\sin 0 = 0$, $\sin \pi/2 = 1$, etc., enable us to sketch their graphs:



Other special values of sin and cos can be read directly from the geometry of the unit circle. For example the angle $\pi/4$ (or 45°) determines an isosceles right triangle with hypotenuse 1; thus the legs of this triangle both have length $1/\sqrt{2}$. It follows that $\cos \pi/4 = \sin \pi/4 = 1/\sqrt{2}$. In making such computations it is often more convenient to use a reference triangle that is *similar* to the small one in the unit circle but with more convenient dimensions. In this example it is easier to refer to a triangle with legs of length 1 and hypotenuse of length $\sqrt{2}$ and to read off the values of the trigonometric functions using their traditional definitions as ratios of sides of the triangle. Reference triangles for several special angles are shown below. From these can you read off, for example, the values of $\sin \pi/6$, $\cos \pi/6$, $\sin \pi/3$, and $\cos \pi/3$?



The examples above dealt with angles in the first quadrant ($0 \le t \le \pi/2$). Angles in other quadrants can be handled using, again, the geometry of the unit circle. An angle t in the second quadrant, for example, can be "reflected" in the y-axis and compared with the angle $\pi - t$ in the first quadrant. Examining the figure below it is clear that $\cos(\pi - t) = -\cos t$ and $\sin(\pi - t) = \sin t$. In particular, therefore, $\cos(5\pi/6) =$ -$cos(\pi/6)$, and the value can then be read off from one of the reference triangles shown above. Angles in the third quadrant can be reflected in the origin and compared with an angle in the first quadrant. And angles in the fourth quadrant can be reflected in the x-axis. By considering only angles in the first quadrant, therefore, and using the reference triangle method for computing sin and cos of special angles, we can compute the values of these trigonometric functions for angles of any size.



We should mention, finally, the identities $\cos(\pi/2 - t) = \sin t$ and $\sin(\pi/2 - t) = \cos t$. The angles t and $\pi/2 - t$ are *complementary*. They "share" a single right triangle. Do you see that the ratio $\frac{a}{h}$ is $\cos t$ as well as $\sin(\pi/2 - t)$? As a general rule, it is not necessary to remember hundreds of unrelated facts and identities for the trigonometric functions. A little understanding of the geometry involved goes a long way.

Definition: Other trigonometric functions: For the record we define the other trigonometric functions that are often used. They are

$$\tan t = \frac{\sin t}{\cos t} \qquad \cot t = \frac{1}{\tan t} = \frac{\cos t}{\sin t}$$

$$\sec t = \frac{1}{\cos t} \qquad \csc t = \frac{1}{\sin t}$$

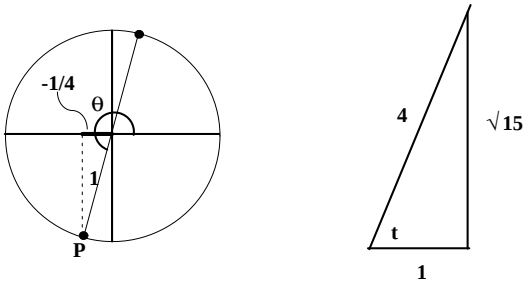
Notice that all of these additional trigonometric functions are defined in terms of sin and cos. And, in fact, the sine function also can be defined in terms of cos since, for angles in the first quadrant

$$\sin t = \sqrt{1 - \cos^2 t}.$$

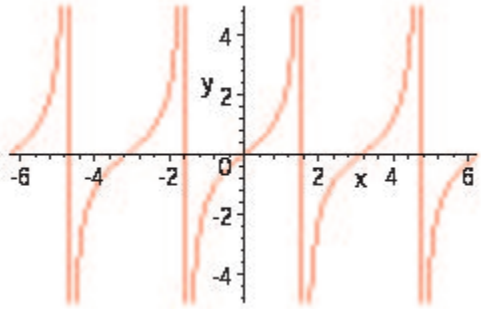
In a sense, then, most of the trigonometric functions are superfluous—only one of them is needed, and all the others can be defined in terms of that one. But this would obscure the rich set of identities relating the different trigonometric functions. It is this algebraic richness that can be exploited in solving many problems. We do not intend to pause here for an extensive survey of such identities, nor do we recommend memorizing endless lists of identities and properties. Rather, we have mentioned a few basic identities that follow more or less directly from the unit circle definition, and we prefer to "derive" more identities only as the need for them arises.

Example 1: Suppose θ is a third quadrant angle and $\cos\theta = -1/4$. Find $\tan\theta$. Refer to the figure below. We first observe that the sign of $\tan\theta$ is positive since both sin and cos are negative in the third quadrant. Then, reflecting the point P in the origin, we may use the more convenient reference triangle shown in the figure to compute the value of $\tan\theta$. Thus $\tan\theta = +\tan t = \sqrt{15}$. The values of the other

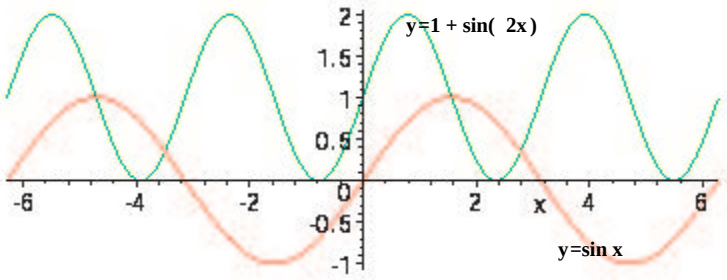
trigonometric functions can also be computed from this reference triangle, taking into account their signs in the third quadrant. So $\sin\theta = -\sqrt{15}/4$, $\sec\theta = -4$, etc..



Example 2: Sketch the graph of $\tan x$. Since $\tan x = \frac{\sin x}{\cos x}$, we notice that the x-intercepts of the graph occur where $\sin x = 0$, i.e. at the points $x = n\pi$ where n is an integer. Moreover $\tan x$ is undefined whenever $\cos x = 0$. Thus the graph has vertical asymptotes $x = \pi/2 + n\pi$, n an integer.



Example 3: Sketch the graph of $f(x) = 1 + \sin 2x$. We can visualize this graph as the graph of $\sin x$ compressed by the factor 2 (all horizontal distances multiplied by $1/2$) and shifted up 1 unit. Moreover since $\sin x$ has period 2π, the function $f(x)$ has period π. We show the graphs of both functions below.



Applet: Trigonometric Identities **Try it!**
Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

1.6 Exponential and Logarithm Functions

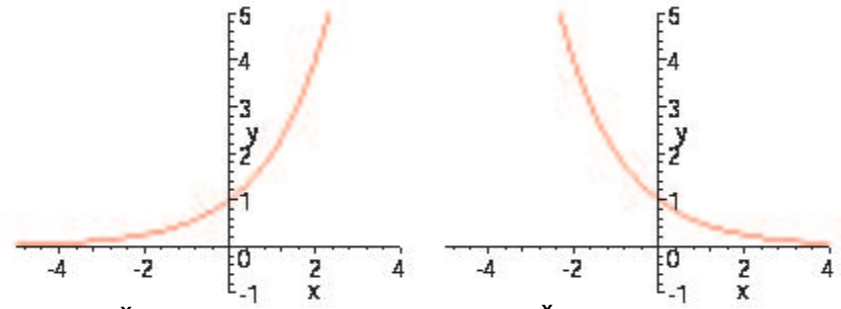
Throughout the preceding sections we have talked about the useful class of *elementary functions* but have not yet defined precisely what we mean by that term. We have cited polynomial functions, rational functions, and trigonometric functions as belonging to the class. And we have emphasized for each of these the unique properties that make them valuable in modeling real world problems. For example the polynomial and rational functions are the basis of nearly every algebraic expression we write. And the trigonometric functions are *periodic*, making them the functions of choice in modeling physical processes that repeat themselves.

In this section we add the *exponential functions* and *logarithmic functions* to the list. You will be happy to know that that is close to the end of the story. From the basic functions named above we obtain all the remaining elementary functions by applying arithmetical operations—addition, subtraction, multiplication, and division—and by using composition of functions and inverses of functions. Much of calculus is involved in studying the properties of the basic functions named above and in learning to use them in applications that rely upon their unique properties. The full class of elementary functions provides a rich source from which we can draw functions to represent a variety of real world objects and model their behavior.

Exponential functions make their most dramatic debut in population modeling. The population of Mexico, for example, increased at the rate of 2.6% per year during the 1980's. Beginning with a population of 67.38 million people in 1980, the population increased each year by a factor of 1.026. Thus in 1981 the population was $67.38(1.026)$ million, in 1982 it was $67.38(1.026)^2$ million, and in general it is $P(t) = 67.38(1.026)^t$ where t is the number of years that have elapsed since 1980. This is an *exponential function*. It is called an exponential function because the base is constant, in this case the constant 1.026, and the independent variable t is in the exponent. The base represents the *growth factor* by which the population increases each year. Notice that the growth *rate* of 2.6% corresponds to a growth *factor* of 1.026. A growth *rate* of r% per year corresponds to a growth *factor* of $1 + r/100$.

In general, for any positive constant a, we have an exponential function a^t. The key property of such functions is the constant ratio a between the values of the function in any two consecutive years.

Definition 1: Let a be a positive real number. Then $P(x) = Ba^x$ is called a general exponential function.



Because $a^0 = 1$, B is the value of the exponential function at 0: $P(0) = Ba^0 = B$.

When the growth factor $a > 1$ the graph is increasing. From our hint that exponential functions model population growth we would expect that. What is surprising is the rate at which they grow. Even exponential functions that begin their lives by growing slowly eventually go through the roof with gusto. This has disasterous implications for the world if constant growth rates are sustained. The term *exponential growth* refers to exponential functions with positive growth rates (growth factors greater than 1).

When the growth factor is less than 1 the graph is decreasing. This might model the balance in your bank account where as a result of inflation the value of your nest egg decreases each year. If the inflation rate is 3.5%, the growth rate of the value of your balance is −3.5% per year, corresponding to a growth factor of $1 + (-.035) = 1 - .035 = .965$. Clearly we should be referring instead to the "shrinkage rate" and the "shrinkage factor". The shape of the second curve above will explain why inflation can give you such a pain in your liver. The term *exponential decay* is an apt expression of the behavior of exponential functions with negative growth rates (growth factors between 0 and 1).

True Confessions We have behaved in our discussion as though we know what we mean by a^x. In fact

we have never given a complete definition. If x is an integer or a rational number we *have* defined the value of a^x. For example $a^0 = 1$. For a positive integer n the definition is $a^n = \underbrace{a \cdot a \cdot a \cdot \cdots \cdot a}_{n}$. For a positive rational number $r = m/n$ we define $a^r = \sqrt[n]{a^m}$. And for a negative integer or rational number we define $a^r = 1/a^{-r}$.

These are our definitions from the treatment of exponents in algebra. And we know, in addition, certain basic laws of exponents:

Laws of Exponents

If $a > 0$ and $b > 0$, and x and y are any real numbers, then

(i)	$a^0 = 1$	(ii)	$a^{x+y} = a^x a^y$
(iii)	$a^{-x} = \frac{1}{a^x}$	(iv)	$a^{x-y} = \frac{a^x}{a^y}$
(v)	$(a^x)^y = a^{xy}$	(vi)	$(ab)^x = a^x b^x$

But we have not defined a^x when x is not a rational number. For example we have not defined the quantity 2^π nor $3^{\sqrt{2}}$. We can hardly consider a^x to be a continuous function if it remains undefined for irrational numbers. In effect its graph is full of holes since the irrational numbers are densely mixed with the rational numbers.

Having confessed we will now sidestep this issue, important as it is, with the promise that it will be fixed later on when we have additional tools of calculus available. In the meantime we will "fill in the holes" by requiring that a^x be a continuous function. This would imply that the quantity $3^{\sqrt{2}}$ is defined, for example, (relying on the fact that $\sqrt{2}$ is the infinite non-repeating decimal 1.4142135623730950488...) as the limit of the sequence of values

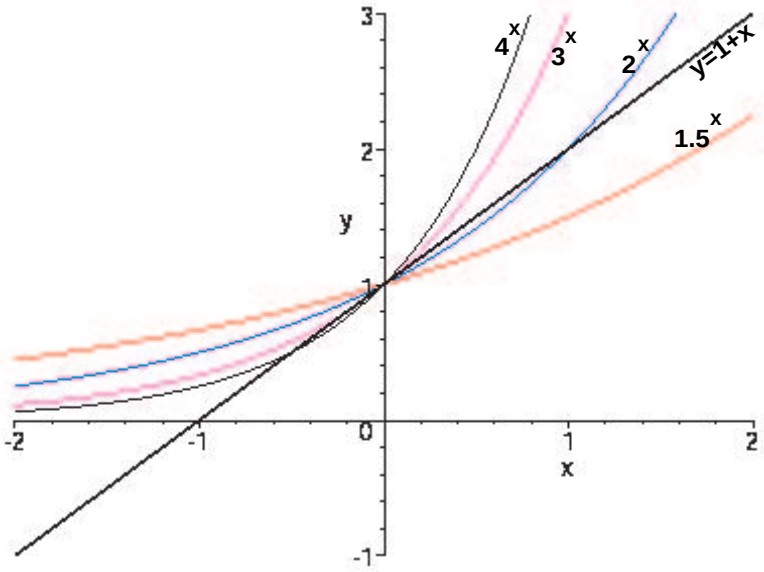
$$3^{1.4},\ 3^{1.41},\ 3^{1.414},\ 3^{1.4142},\ 3^{1.41421},\ \ldots$$

That such limits always exist needs proof, of course, and that will come. Until then we will, without embarrassment, assume that the exponential function a^x is defined and continuous for all real values of x and that all the usual laws of exponents hold. This will enable us to move on to the applications that make these functions so important.

Example 1: We can use the laws of exponents to ease our task when computing with exponentials. For example $2^{10} = (2^5)^2 = 32^2 = 1024$. And $2^{20} = (2^{10})^2 = 1024^2 = 1,048,576$.

Example 2: We can freely interchange exponential and "root" notation. For example $\sqrt{x} = x^{\frac{1}{2}}$, $x^{\frac{3}{5}} = \sqrt[5]{x^3}$.

Example 3: The graphs of functions a^x for diffent values of $a > 1$ are quite similar.



They differ in their slope at the point $(0, 1)$. (We are using the word *slope* to mean the slope of the tangent line at the point on the curve. We will go beyond the everyday useage we rely on here and make the

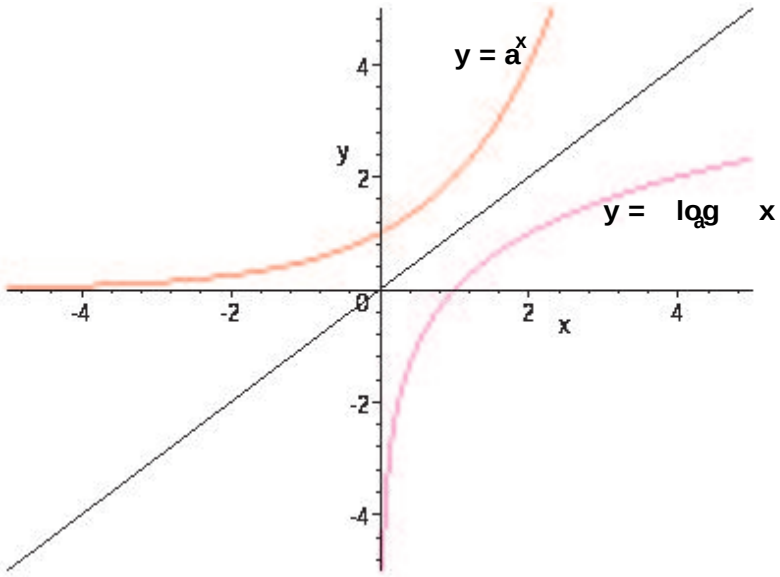
notion precise in Chapter 3. Being able to do so constitutes one of the triumphs of calculus.) We notice that 1.5^x and 2^x have slopes less than 1 (the graph of $y = 1 + x$ was included in the plot for comparison), while 3^x and 4^x have slopes greater than 1. Apparently the slope at $(0, 1)$ increases as a increases. Being curious, we can ask whether there is a value of a for which the slope of a^x at $(0, 1)$ is exactly equal to 1? We expect such a value between 2 and 3, and indeed there is! We will show that there is a number $e = 2.718281828459045...$, an infinite non-repeating decimal (and hence irrational) number, for which e^x has slope at $(0, 1)$ which is exactly 1. This will turn out to be the most "natural" base for an exponential function and will become the standard for calculus.

Applet: Comparing Exponential Functions **Try it!**

The Inverse of a^x: If $a > 1$ the function a^x is increasing on $(-\infty, \infty)$, and if $0 < a < 1$ it is decreasing on this interval. In either case it is a 1:1 function and so has an inverse.

Definition 2: The inverse of the general exponential function a^x, written as $\log_a x$, is called the *general logarithm function*. It is defined by the relations

$$y = a^x \Leftrightarrow x = \log_a y.$$



Many properties of $\log_a x$ follow immediately. Its graph is the reflection in the line $y = x$ of the graph of a^x. Its domain is $(0, \infty)$ since that is the range of a^x. Its range is $(-\infty, \infty)$ since that is the domain of a^x. The value of $\log_a 1 = 0$ since $a^0 = 1$. And the characteristic laws of logarithms hold:

Laws of Logarithms

If $a > 0$, $b > 0$, $a \neq 1$, and $b \neq 1$, then

(i) $\log_a 1 = 0$ (ii) $\log_a xy = \log_a x + \log_a y$

(iii) $\log_a \frac{1}{x} = -\log_a x$ (iv) $\log_a \frac{x}{y} = \log_a x - \log_a y$

(v) $\log_a x^y = y \log_a x$ (vi) $\log_a x = \frac{\log_b x}{\log_b a}$

All of these laws follow from the laws of exponents. We prove (ii) and (vi) as examples. For (ii), let $\log_a x = u$ and $\log_a y = v$. Then $\log_a x + \log_a y = u + v$. But $a^u = x$ and $a^v = y$, so $xy = a^u a^v = a^{u+v}$. Thus, finally, $\log_a xy = u + v = \log_a x + \log_a y$. For (vi), let $\log_a x = u$. Then

$$\begin{aligned} a^u = x &\Leftrightarrow \log_b a^u = \log_b x \\ &\Leftrightarrow u \log_b a = \log_b x \quad \text{(using (v))} \\ &\Leftrightarrow u = \frac{\log_b x}{\log_b a} \end{aligned}$$

The Number e: We have already pointed to the number $e \approx 2.718281828...$ that plays a special role for exponential functions. Namely the exponential function e^x crosses the y-axis at the point $(0, 1)$ with slope exactly equal to 1. For the moment we are taking that as our definition of e. (As promised, we will be returning later to give a precise definition of e^x as a continuous and diffentiable function.)

Definition 3: The *natural exponential function* e^x is that exponential function that crosses the y-axis with slope 1. Its inverse $\log_e x$ is called the *natural logarithm function* and is denoted more simply by $\ln x$.

The two functions e^x and $\ln x$ are the ones that occupy prime space in calculus. We will see that all other exponential and logarithm functions can be expressed in terms of these two, hence in a sense are redundant. We will also learn why they are termed "natural". It has to do with the fact that they have the simplest differentiation formulas among all exponential and logarithm functions. With e^x and $\ln x$ added to our repertoire of basic functions, we have also completed our definition of the class of *Elementary Functions* of calculus.

Of course e^x and $\ln x$, as special cases of the general exponential and logarithm functions, satisfy all the laws for exponents and logarithms listed above. But they are central enough in calculus that we state them again here.

Properties of e^x

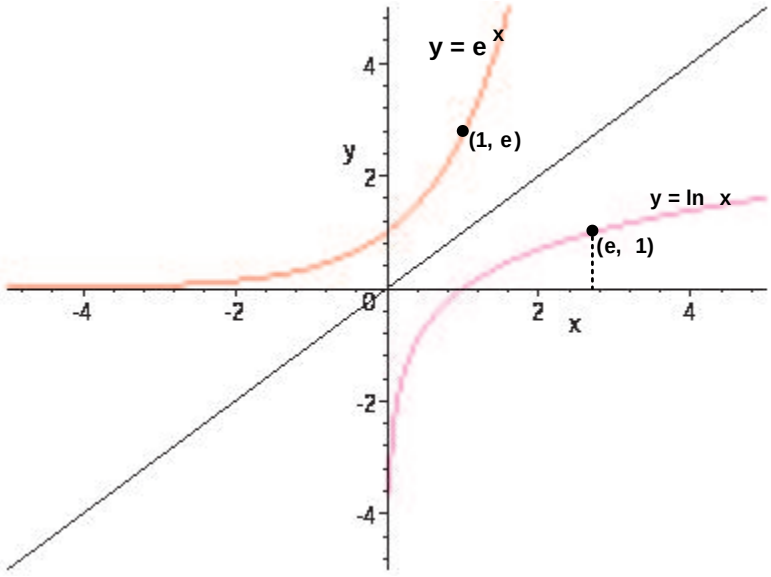
Domain of e^x is $(-\infty, \infty)$, and its Range is $(0, \infty)$

(i)	$e^0 = 1$	(ii)	$e^{x+y} = e^x e^y$
(iii)	$e^{-x} = \frac{1}{e^x}$	(iv)	$e^{x-y} = \frac{e^x}{e^y}$
(v)	$(e^x)^y = e^{xy}$	(vi)	$a^x = e^{x \ln a}$, $(a > 0)$

Properties of $\ln x$

Domain of $\ln x$ is $(0, \infty)$, and its Range is $(-\infty, \infty)$

(i)	$\ln 1 = 0$	(ii)	$\ln xy = \ln x + \ln y$
(iii)	$\ln \frac{1}{x} = -\ln x$	(iv)	$\ln \frac{x}{y} = \ln x - \ln y$
(v)	$\ln x^y = y \ln x$	(vi)	$\log_a x = \frac{\ln x}{\ln a}$



Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

1.7 Case Study: Modeling with Elementary Functions

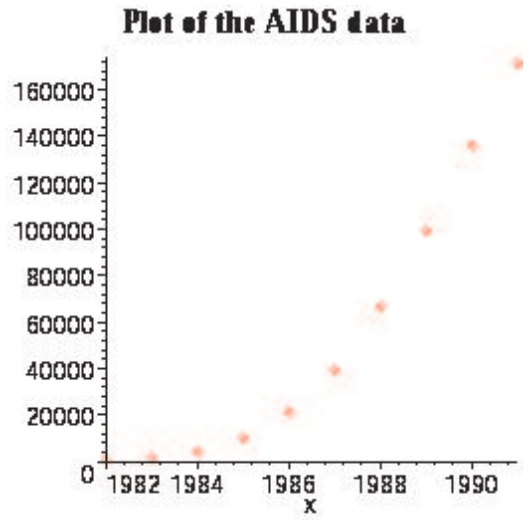
Animation: Modeling with Elementary Functions **To get you going on the Case Study!**

We close the first chapter on the elementary functions with a *Case Study in Calculus* (CSC). The purpose of a CSC is to consider a real application of calculus, with real data. In a CSC, we act like mathematicians working with scientists from other areas to answer questions in their fields. All of the answers involve using some of the calculus tools we are learning. One way to think of the CSCs is as extended homework problems. But they are more than that because they show the power of calculus at work in other disciplines. The concern we want to address in this CSC involves the AIDS data introduced earlier.

The spread of the AIDS virus in the United States during the first ten years following its discovery in 1981 is given in a table published by the Centers for Disease Control:

Year	No of AIDS cases
1982	295
1983	1374
1984	4293
1985	10211
1986	21278
1987	39353
1988	66290
1989	98910
1990	135614
1991	170851

The table defines a function $f(n)$, that gives the cumulative number of AIDS cases up to the year n. Here is a graph of the data.



Our interest is in finding a continuous function $f(x)$ whose graph approximates the data points and runs smoothly between them. Moreover, the function should be suitable for making future predictions. Such a function $f(x)$ is said to *model* the data. To obtain f, we have seen that mathematically a least squares fitting procedure is often the method of choice.

However, we want a function that both fits the points mathematically and is also based on known scientific principles. In the AIDS study, those principles are epidemiological. This is a good example of how mathematicians work with scientists in other fields to solve important problems.

But suppose we fit, for example, a cubic polynomial to the data points. And alternatively we fit an exponential function to the data points. The exponential appears appropriate because the graph of the AIDS data resembles the many examples of exponential population growth so often pictured in environmental publications. (We will study such population models ourselves later on.) How do we decide which fit is better?

There are of course two considerations. First, let's discuss the mathematical one. Because we are using least squares techniques to fit curves to the data, we are going to use the sum of the squares of the errors as a measure of the goodness of fit over the time period in question. That is, suppose we start with data points $(t_1, c_1), (t_2, c_2), (t_3, c_3), \ldots, (t_n, c_n)$ and we fit the data with the function $y = f(t)$. Then we will define the *sum of squared errors* over the period to be the sum of the individual squared errors at the data points, namely,

$$\sum_{i=1}^{n} (f(t_i) - c_i)^2$$

Moreover, we can use the sum of squared errors to develop a criterion for which fit is better.

Mathematical Criterion for the Best Fitting Function: One fitting function will be better than another if its sum of squared errors is smaller.

Squaring the errors does not allow the signed-errors to cancel one another in the summation; sqaring also has the effect of giving more weight to bigger errors than smaller ones.

Now that we have a mathematical criterion for deciding between fitting functions, we need to consider if there is a scientific one. Continuing with the AIDS example, one might be tempted to choose an exponential function because as we will see later, exponentials are associated with many population studies. However, it was reported in medical journals a few years ago that the epidemic does not seem to be following an exponential model. Indeed, it appears to be behaving very differently from other viral epidemics such as influenza epidemics. Rather, it appears to be best modeled by a cubic function, and several papers in the medical literature argue that this is to be expected for diseases like AIDS that have long latency periods and for which homogeneous mixing of the virus throughout the population is inhibited by variations in behavior

of different groups of people. Thus, from a scientific standpoint, we should seek a cubic function that best fits the data.

And this brings us to the objective of this CSC.

Objective: The purpose of this CSC is to fit both a cubic polynomial and an exponential to the AIDS data and compare the results mathematically. We want to determine if in this case, the mathematical criterion of *best fit* gives the same answer as the epidemiological one. We also want to make predictions about AIDS cases in the future.

Before we can meet the objective, we need to know how to fit an exponential function to a data set.

Fitting an Exponential by Least Squares: Most Computer Algebra Systems such as Maple or Mathematica cannot fit an exponential function to data directly. Instead, we must fit a line to the semi-log data. That is, given the (x, y) data, we fit a line to the $(x, \ln y)$ data. Suppose we carry out that process and obtain the fitting line $l(x) = mx + b$. Note that if $y = Ce^{kx}$, then $\ln y = \ln C + kx$. Thus, because we fit the semi-log data, we can determine the C and k of the exponential fitting function we seek: $\ln C = b$ and $k = m$. So, the exponential fitting function is

$$y = e^{mx+b} = \left(e^b\right) e^{mx}$$

Back to the CSC: There will be a CSC at the end of each section of the book. We have structured them all the same way, as reports that you will complete. The reports teach you how to analyse scientific modeling problems systematically. Every report will include a concluding written part because it is very important for mathematicians to be able to convey their findings and recommendations to a general audience. Below are the parts of the report that you will be completing.

Setup: The purpose of this part is for you to write a clear statement of the mathematics that will be applied to the problem. What is the nature of the information that is available? What are the steps that you should follow to analyze the information? What tools will you bring to bear in the analysis?

The setup phase in tackling a scientific problem is often called the *modeling phase*. Mathematical tools are chosen to apply to the problem. Often the setup involves deriving an equation or equations that express the physical aspects of the problem in mathematical language. It is these equations, then, that would be called the mathematical model for the physical problem. Once the physical problem has been thus expressed in mathematical terms one can apply the theories and techniques of mathematics toward finding solutions.

In our present AIDS study, you should indicate how to obtain a polynomial and an exponential fit of the data, and what steps you will take to carry out the objective.

Thinking and Exploring: Now it is time to think about the mathematics. Carry out the mathematical ideas that you outlined in the setup. What is the third degree best fitting polynomial? What is the best fitting exponential function? Which of these provides a better mathematical fit and why?

Applet: Fitting AIDS Data **Try it!**

In June, 2001 the Centers for Disease Control reported that there were 274,624 people in the US living with AIDS in 1998, 299,944 in 1999, and 322,865 in 2000. Would your model have predicted these numbers? What is the percentage error for each year?

Interpretation and Summary: Now that you have done the mathematics and explored the models, it is time to interpret and summarize the mathematical results in terms of the original objective.

Pretend that your synopsis is going to appear in the next issue of a magazine such as *Scientific American*. Include enough details so that a reader would learn what the major issues of the report are, and how you went about addressing them. What will you want to tell readers about your success with regard to the original stated objective of the investigation? Be sure to write in complete sentences using correct rules of standard English grammar. Make the write-up interesting and informative. Would the mathematical answers alone be sufficient to model the AIDS data? How good a predictor is your model of future AIDS cases? Is there further analysis that needs to be done?

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

Chapter 2

Modeling Rates of Change

2.1 Introduction to the Issues

One of the great breakthroughs of the seventeenth century was an understanding of motion. According to Aristotle, the force due to gravity affects the speed of an object, so that heavy objects fall faster than lighter ones. This belief was not challenged until the late sixteenth and early seventeenth centuries, when Galileo set out to establish through experiments the true effect of gravity on objects in free fall. The now famous story of Galileo dropping balls off the Leaning Tower of Pisa is probably apocryphal. We do know, however, that he conducted many experiments that involved him rolling balls down an incline plane, a ramp, where it is easier to measure speeds because the ball can be slowed down.

Suppose now that we turn to modern methods, and consider an object dropped vertically from rest near the surface of the earth. We can imagine, for example, that we drop a ball from the edge of a cliff. As the object falls, we record in a table, the distances it covers for a period of time. Here is such a table:

time (s)	distance (m)
0.10	0.049
0.20	0.196
0.30	0.441
0.40	0.784
0.50	1.225
0.60	1.764
0.70	2.401
0.80	3.136
0.90	3.969
1.00	4.900

The times are measured in seconds, and the distances in meters. How much can we learn about the distance function from this table of values? We faced similar questions in the first section of the book, and our approach there was to seek an elementary function that would model the data. In this section we are going to take a different tack by adopting the point of view of Newton, and study the speed and acceleration of the object. That is, we are going to use the characteristics of motion that are familiar to us all, namely, average speed and accelertion, to try and get information about the distance function.

2.1.1 Average Speed

If we are going to use the speed and acceleration of the falling object to get information about the distance it falls as a function of time, then a reasonable first step is to add two columns to the above table, one for speed and the other for acceleration. However, we immediately encounter a problem: How do we measure the speed at a given time? For example, if we drive from Philadelphia to New York City, a distance of 90 miles, in an hour and a half, then our average speed is $90/1.5 = 60$ miles per hour. But this tells us very

little about the speed we are going when we cross the state line into New York. This instantaneous speed is what a policeman's radar gun purports to measure. We will return to this issue later and see if indeed that is the case. For now, let us try to do the best we can with what we can calculate from the data, namely, average speeds.

An average speed is always computed over an interval of time. In fact, **average speed** is defined to be **change in distance divided by change in time**. Returning to the table of distances, where the times are recorded 0.1 seconds apart, we will add a column for the speeds. As an approximation to the instantaneous speed, the speed at time t will be the average speed over the interval $[t, t+0.1]$. (Even though we have chosen the interval of time to the right of t, we could just as easily have chosen the interval to the left.) We will also add a column of accelerations by deriving it from the column of speeds. That is, we will calculate the acceleration at t as the speed at $t+.1$ minus the speed at t, divided by 0.1; thus, it is the average acceleration over the interval $[t, t+.1]$ where we are using the average speeds to approximate the instantaneous speeds.

Derived Table of Speeds and Accelerations $[t, t+0.1]$

time t (s)	distance (m)	speed (m/s)	acc (m/s/s)
.100000	.049000	1.470000	9.800000
.200000	.196000	2.450000	9.800000
.300000	.441000	3.430000	9.800000
.400000	.784000	4.410000	9.800000
.500000	1.225000	5.390000	9.800000
.600000	1.764000	6.370000	9.800000
.700000	2.401000	7.350000	9.800000
.800000	3.136000	8.330000	9.800000
.900000	3.969000	9.310000	9.800000
1.000000	4.900000	10.290000	9.800000

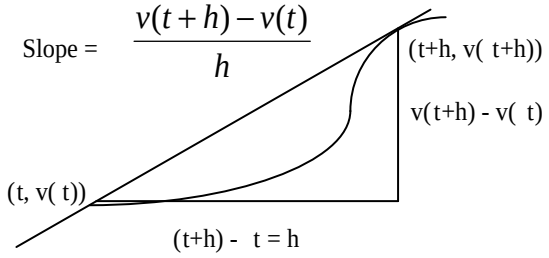
Before we investigate the implications of the derived table, let's go back to our everyday experiences with automobile travel. Speed is how fast we are traveling. It is the rate of change of distance with respect to time. Acceleration is the rate of change of speed with respect to time. We accelerate when we pull out to pass a car, and we decelerate when we approach a stop sign. When we accelerate, we feel our bodies being pushed against the seatback. The snap of the head we feel when we suddenly push down hard on the gas pedal is due to "jerk." Jerk is the rate of change of acceleration.

Now, we are ready to consider the derived table. Setting aside the fact that we are using averages as approximations, our approach has been to calculate the rate of change of distance, namely, the speed, and then the rate of change of speed, or acceleration. If we tried to go further and use the table to calculate the rate of change of acceleration, or jerk, we would not get very far: the accelerations are all constant. So, the jerk is zero.

Applet: Average Velocity **Try it!**

2.1.2 The Meaning of Constant Acceleration

Suppose the speed of a falling object is given by the function $v(t)$. Then the average acceleration over the interval $[t, t+h]$ is given by the quotient $\frac{v(t+h)-v(t)}{h}$; in the derived table, h = 0.1. Here is an important observation that connects this physical definition with algebra and geometry: geometrically, the quotient, which we will refer to as a *difference quotient*, is the slope of the line through the points $(t, v(t))$ and $(t+h, v(t+h))$ on the graph of the function v.



Thus, we can articulate the following question that is motivated by the constant accelerations in the derived table. Suppose we are given a finite number of points whose adjacent x-coordinates are h units apart. Suppose that the slope of the line through any two successive points is always the same; that is, these slopes are constant. Then how can we characterize the points geometrically?

From our work with the elementary functions, we should realize that all of the points must lie on a (straight) line. Hence, we have arrived at an important realization about what the derived table tells us. Assuming that the average speeds give good approximations to the instantaneous speeds, we can conclude that the speed function is linear.

2.1.3 Hypotheses and Open Questions

Let's take stock of where we are. We began with a table of values for the distance of a falling object as a function of time. Motivated by a desire to study this function by taking into account what we know about the characteristics of motion, we produced the derived table from calculations. These calculations of speed and acceleration in actuality were average and not intantaneous values. Even so, from the fact that the derived accelerations turned out to be constant, we have arrived at two hypotheses:

1. The acceleration of a falling object is constant as a function of time.
2. The speed of a falling object is linear as a function of time.

There are also at least two open questions that our work has raised about falling objects:

1. Can we find a description (i.e., a formula) for the distance function? This is really the question we started with.
2. How can we get better approximations to the instantaneous speeds?

Let's address the second question first. We have recorded the distances in our table 0.1 seconds apart, and have used the average speed over the interval $[t, t + 0.1]$ as an approximation to the speed at time t. Clearly, we could do better if we recorded the speeds 0.01 seconds apart, or even 0.001 seconds apart, etc. We can describe this process as *looking for a limiting value* of the average speeds as the interval h between successive times shrinks to zero. So, we should make a note for ourselves and return to this point later: For a given time t, it makes sense to define the instantaneous speed (or velocity) at t by way of this process. Symbolically, we write:

$$v(t) = \lim_{h \to 0} \frac{s(t+h) - s(t)}{h}$$

For the moment, let's return to the radar gun that we mentioned earlier. In reality, a radar gun does not measure the exact speed at an instant of time. Instead, it attempts to carry out the above limiting process by calculating the average speed of a moving car for a very small value of h, namely, the interval of time associated with one pulse of the radar gun. The limiting process that we have described symbolically demands more. That is, we want h, though non-zero, to get smaller and smaller without end; and we take the limiting value, if it exists, of the difference quotients (the average speeds) to be the instantaneous velocity. We will discuss this approach in more detail in the next section.

Now, for the first question. Given our hypotheses, we can find a representation for the distance function.

2.1.4 The Distance Function is Quadratic

We are assuming that the speed (or velocity) is linear, and that the initial speed is 0. Thus, the velocity has the form $v(t) = at$, for some constant a. We can derive the formula for distance from this assumption and two other facts. The first is a formal statement of what we already know, and the second is one that could have been established in Galileo's laboratory:

1. In the case of constant velocity, distance equals velocity times time.

2. In the interval of time from 0 to t, in the case of linear velocity the distance traveled is the same as if the object had traveled at a constant velocity equal to one-half the final velocity.

Now, putting all of this information together, we have

$$s(t) = v(t) \cdot \frac{t}{2} = \frac{at^2}{2}$$

That is, the distance function is quadratic. Also, from the first line of the table we can find the value of a: $a(.1^2)/2 = .049$ implies that $a = 9.8$ meters per second per second.

Note that a comparison with the table shows this to be the (constant) acceleration of the object, due to gravity no less.

2.1.5 Average Rate of Change

So far, we have started with a table of values, and generated a derived table. By studying the derived table, we have been able to learn a great deal of information about the underlying function that corresponds to the original table. Using the derived table has been very productive in this case.

Let's now work in the other direction. We will begin with an elementary function and then generate a derived table for it to see if anything important stands out. To compute the first derived, we need an analogous notion of average speed.

Definition: Given a function f, the *average rate of change* of f over an interval $[x, x+h]$ is

$$\frac{f(x+h) - f(x)}{h}$$

That is, if $y = f(x)$, the average rate of change of f is the change in f divided by the change in x over the interval. (The cahnge in f is $f(x+h) - f(x)$, and the change in x is $(x+h) - x = h$.) The average rate of change is also what we have called the *difference quotient* over the interval.

We will use the formula for the average rate of change of f over an interval $[x, x+h]$ to compute the first derived. The second derived will be the change in the first derived divided by h over that interval. We will use a small value of h, the distance between successive x-values, because we have agreed that this gives a better approximation to the instantaneous rate of change at a point. Note that in analogy with what we did in the case of motion, we are defining the *instantaneous rate of change* of a function at a point x to be the limit of the average rates of change over intervals $[x, x+h]$ as $h \to 0$. Let's carry out this process for the function $y = e^x$.

Derived Table for $y = e^x$

x	e^x	1st derived	2nd derived
.001000	1.001000	1.001501	1.003000
.002000	1.002002	1.002504	1.002000
.003000	1.003005	1.003506	1.004000
.004000	1.004008	1.004510	1.005000
.005000	1.005013	1.005515	1.006000
.006000	1.006018	1.006521	1.008000
.007000	1.007025	1.007529	1.007000
.008000	1.008032	1.008536	1.009000
.009000	1.009041	1.009545	1.010000
.010000	1.010050	1.010555	1.012000
.011000	1.011061	1.011567	1.011000
.012000	1.012072	1.012578	1.014000
.013000	1.013085	1.013592	1.014000
.014000	1.014098	1.014606	1.014000
.015000	1.015113	1.015620	1.017000
.016000	1.016129	1.016637	1.017000
.017000	1.017145	1.017654	1.019000
.018000	1.018163	1.018673	1.018000
.019000	1.019182	1.019691	1.021000
.020000	1.020201	1.020712	1.020000

The table is interesting indeed. A comparison of the columns for e^x and the 1st derived function would lead us to believe that the derived function of $y = e^x$ is $y = e^x$; that is, both the derived function and its derived function (the 2nd derived function) are the same as the original function. We can experiment further with this hypothesis by decreasing the value of h and observing the outcome. The same conclusion will pertain: It appears that the rate of change of the function $y = e^x$ at a point is the value of the function at that point.

This certainly seems special, doesn't it? We have used the word *special* before in connection with $y = e^x$. When we studied its properties, we found that it was special among all exponential functions because it has a slope of 1 at the origin. In our current investigation, this begs the question: What is the rate of change of $y = e^x$ at $x = 0$? Based on our work above, we know how to answer this question. At $x = 0$, we have to investigate the behavior of the difference quotients, the average rates of change,

$$\frac{e^{0+h} - e^0}{h}$$

for small values of h.

Difference Quotient for $y = e^x$ at $x = 0$

h	difference quotient
.100000000000000	1.051709180756480
.010000000000000	1.005016708416800
.001000000000000	1.000500166708000
.000100000000000	1.000050001670000
.000010000000000	1.000005000000000
.000001000000000	1.000000500000000
.000000100000000	1.000000050000000
.000000010000000	1.000000000000000
.000000001000000	1.000000000000000
.000000000100000	1.000000000000000

From the table, we are led to believe that this limit is 1. That is,

$$\lim_{h \to 0} \frac{e^h - 1}{h} = 1$$

This means that the limit of the difference quotients at $x = 0$ equals the slope of the tangent line to the graph of $y = e^x$ at $x = 0$. Coincidence? Certainly not. In fact, we will see that, for any point x, the instantaneous rate of change of a function at the point (what we have been calling the limit of the average rates of change, or of the difference quotients, at the point) is equal to the slope of the tangent line to the graph of the function at the point. And even more generally, this holds for each of the elementary functions. Truly amazing.

2.1.6 Our Agenda for This Chapter

Our attempt to find an explicit formula for the distance function of a falling object has raised many issues that require further investigation and work. On the other hand, our findings so far point to the high probability of success if we proceed along the following lines. In fact, this list will form our agenda for the rest of the chapter.

1. Develop a more explicit notion of limit.
2. Adopt the definition of *instantaneous rate of change at a point* as the limit of the average rates of change, or difference quotients, at the point as the length of the interval approaches zero.
3. Explore the geometric meaning of the definition of instantaneous rate of change at a point.
4. Apply the definition to each of the elementary functions to see if there are formula-like rules for calculating the instantaneous rate of change.
5. Use the definition of instantaneous rate of change and its consequences to obtain explicit functions for the position, velocity, and acceleration of a falling object.

We began this section with data recorded in a table, and used the data to calculate some related physical quantities. On the basis of our findings, we developed hypotheses and conjectures. Given that these hypotheses are true, we proved some mathematical results. The outcomes have contributed much to our understanding of the original data. Now, we have the task of going back and filling in the deatils in a systematic way, making definitions explicit, verifying the hypotheses, and producing rock-solid mathematical results. The above list is a blueprint for carrying out this task. Our study of the particular example of the falling object leads us to believe that there is a general theory lurking in the background. We will soon see that this indeed is the case, and that the theory will turn out to be *differential calculus*.

Applet: Derived Function **Try it!**
Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

2.2 The Legacy of Galileo, Newton, and Leibniz:

In the space age we take motion for granted—tracking earth-orbiting satellites, delivering supplies to the emerging international space station from a space shuttle, understanding the dynamics of a dying, collapsing star that is about to become a supernova, or even looking back in time to the birth of our universe in the "big bang". All of these phenomena relate to gravity and the wonderful way it governs our world. It boggles the minds of 21st century humans to remember that only 400 years ago the greatest minds among humans were only beginning to achieve a rudimentary understanding of gravity.

Galileo was interested in falling bodies. It is said that he spent his pasttime atop the leaning tower of Pisa, releasing objects to study how they fall to the ground. In that day it was widely believed that heavy objects fall to the ground faster than light objects. Rather than design experiments to test such a belief, however, the prevailing approach was to look to the classical Greek scientists for answers. It took a Galileo to move beyond the authority of the ancients and to forge a new scientific methodology—*observe nature, construct experiments to test what you observe, and construct theories that explain the observations.* This is the essence of the *scientific method* that has revolutionized our understanding of the natural world since Galileo's time.

Galileo learned from his experiments with falling objects that, neglecting the resistance of the atmosphere, objects of different weights fall at the same rate. Indeed he quantified this principle by discovering that the distance a body drops in time t is proportional to t^2. He was not able to explain why such a simple mathematical law should apply to falling bodies, however. It took Isaac Newton and the methods of calculus to do that. Gottfried Leibniz, co-inventor of calculus, took a slightly different point of view but also studied rates of change in a general setting.

In modern language we would say that an object, falling under the influence of gravity, will have constant *acceleration* of $9.8m/sec^2$, that consequently the *velocity* of the object is a linear function of time t and the distance the body has fallen is a quadratic function of t. Newton was able, using his new tools of calculus, to explain why falling bodies behave in this way. His *laws of motion* and of *universal gravitation* drew under one simple mathematical theory Newton's laws of falling bodies, Kepler's laws of planetary motion, the motion of a simple pendulum, and virtually every other instance of dynamic motion observed in the universe. The explosion of scientific knowledge of the past 400 years is due largely to the direction set by Galileo, Newton, and Leibniz.

Newton looked at Galileo's work on falling bodies and generalized the questions to any moving object. His question: How do we *find* the velocity of a moving object at time t? Newton was astute enough to realize that the more fundamental question was "What in fact do we mean by *velocity* of the object at the instant of time t"? We know how to find the *average velocity* of an object during a time interval $[t_1, t_2]$:

Definition 1: The average velocity during a time interval is the distance traveled divided by the elapsed time, i.e.

$$\text{Average velocity over } [t_1, t_2] = \frac{\text{distance traveled}}{t_2 - t_1}$$

Example 1: A car traveling from Boston to Hanover in three hours, a distance of 135 miles, had an *average velocity* of $135/3 = 45$ miles per hour. The traffic officer that stopped the car in the wilds of New Hampshire and issued a speeding ticket (92 miles per hour in a 65 mile per hour zone) was not thinking of average velocity, however. He said "you were traveling 92 miles per hour at the *instant* my radar gun caught you". The police officer was of the same mind as Isaac Newton in believing one can talk about the *instantaneous velocity* at a time t.

In fact the velocity measured by the officer was, in reality, also an average velocity, albeit measured over a very short time interval. The elapsed time in this case was the few milliseconds between two pulses from the radar gun, and the distance traveled was the few centimeters the car moved during this time. Although it thus begs the question of *instantaneous* velocity, it contains the seeds of a precise definition. Namely, to find the *instantaneous* velocity of an object at time t one can obtain, instead, the *average* velocity during a very short time interval about time t. As the length of the time interval gets shorter and shorter the average velocity *approaches* the instantaneous velocity. In short, the instantaneous velocity is the limiting value of average velocities computed over shorter and shorter time intervals. Making the notion of limit precise is the business of calculus.

Example 2: Suppose that an object is moving vertically so that its position at time t is given by $y = 4.9t^2$. Then the average velocity during the time interval $[1, 2]$ is

$$\text{ave vel on } [1,2] = \frac{y(2) - y(1)}{2-1} = \frac{4.9(2)^2 - 4.9(1)^2}{1} = 14.7 \text{ m/sec}$$

What is the (instantaneous) velocity at time $t = 1$? Taking a "short" time interval $[1, 1+h]$ we compute the average velocity

$$\begin{aligned} \text{ave vel on } [1,1+h] &= \frac{(4.9)(1+h)^2 - (4.9)(1)}{(1+h)-1} \\ &= \frac{4.9(1+2h+h^2-1)}{h} \\ &= 4.9\frac{2h+h^2}{h} \\ &= 4.9(2+h),\ h \neq 0 \end{aligned}$$

Letting $h \to 0$ the last expression for average velocity evidently approaches the value $2(4.9) = 9.8$ m/sec. In the next section we will formalize the process of taking limits, but for the moment our method is to simplify the given expression algebraically until the limiting value is clear. Note that the first three expressions, above, are meaningless ("0/0") when we substitute $h = 0$. The last expression, however, does have a value when $h = 0$, and this is also the value of the limit that we seek. The key to this is the notion of *continuity* which we take up in the next section. [Notice, by the way, that the limiting value is the same whether we approach the point $y = 1$ from the right or from the left, i.e. whether we allow h to approach 0 through positive or negative values.]

Definition 2: Let $x(t)$ be a function that gives the position at time t of an object moving on the x-axis. Then

$$\text{Ave vel}[t_1, t_2] = \frac{x(t_2) - x(t_1)}{t_2 - t_1}$$

$$\text{Velocity}(t) = \lim_{h \to 0} \frac{x(t+h) - x(t)}{h}$$

The quotient in the last expression is called a *difference quotient.*

Example 3: Suppose a particle moves along the x-axis with its position at time t given by $x(t) = t^2 - 4t + 1$. Initially the particle was at the point $x(0) = 1$. Find its average velocities during the intervals $[1, 2]$, $[2, 3]$, and $[1, 3]$. And find its (instantaneous) velocity $v(t)$ at time t.

$$\text{Ave vel}[1,2] = \frac{(2^2 - 4\cdot 2 + 1) - (1 - 4 + 1)}{2-1} = \frac{-3-(-2)}{1} = -1$$

$$\text{Ave vel}[2,3] = \frac{(3^2 - 4\cdot 3 + 1) - (2^2 - 4\cdot 2 + 1)}{3-2} = \frac{-2-(-3)}{1} = 1$$

$$\text{Ave vel}[1,3] = \frac{(3^2 - 4\cdot 3 + 1) - (1 - 4 + 1)}{3-1} = \frac{-2-(-2)}{2} = 0$$

$$\begin{aligned} v(t) &= \lim_{h \to 0} \frac{x(t+h) - x(t)}{h} \\ &= \lim_{h \to 0} \frac{[(t+h)^2 - 4(t+h) + 1] - [t^2 - 4t + 1]}{h} \\ &= \lim_{h \to 0} \frac{t^2 + 2th + h^2 - 4t - 4h + 1 - t^2 + 4t - 1}{h} \\ &= \lim_{h \to 0} \frac{2ht + h^2 - 4h}{h} = \lim_{h \to 0} (2t + h - 4) = 2t - 4 \end{aligned}$$

Having found a general formula for $v(t)$ we can analyze the motion of the particle quite easily. At the initial time $t = 0$ the velocity was $v(0) = -4$, meaning that the particle was moving to the left with a speed of 4 units per second, and its position was $x(0) = 1$. It continued moving to the left until $t = 2$ at which time its velocity was $v(2) = 0$ and its position was $x(2) = -3$. From that time onward its velocity is positive so it is moving to the right.

Summary: In this section we began with the common notion of *average velocity*, the ratio of distance traveled to elapsed time, and moved to a definition of *instantaneous velocity*. The latter depended on the notion of "taking a limit". Newton and Leibniz were among the first to recognize the fundamental importance of limits and to develop methods for working with them. The collection of such techniques is what has become known as *calculus*, and it is the reason that Newton and Leibniz are remembered as the co-inventors of calculus. In the subsequent sections we will retrace their steps.

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

2.3 Limits of Functions

The notion of *instantaneous velocity* was defined in terms of a *limit* of average velocities over shorter and shorter time intervals. For a particle moving on the x-axis, with position $x(t)$ at time t, we introduced the notation

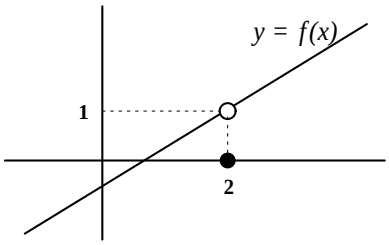
$$v(t) = \lim_{h \to 0} \frac{x(t+h) - x(t)}{h}$$

to formalize this process. In our first computations we proceeded informally in evaluating several limits, confident that our intuition would not lead us astray. Indeed, that is not a bad approach. Newton operated informally throughout his work, revolutionizing the understanding of dynamic motion. It was not until nearly 150 years later that it was found necessary to revisit the foundations of calculus and to provide precise definitions for its underlying notions such as limit.

Definition 1: We say that a function f approaches the limit L as x approaches a, written $\lim_{x \to a} f(x) = L$, if we can make $f(x)$ as close to L as we please by taking x sufficiently close to a.

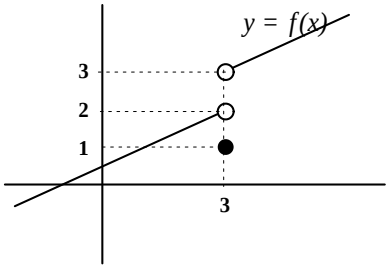
We hasten to note that this "definition" serves to introduce notation and language for talking about limits, not to resolve the deeper questions of what we mean by phrases such as "close to L" or what is meant by "sufficiently close". Suffice it to say that precise meanings can be given to these phrases, and that we may safely follow our intuitions in interpreting and working with them. From time to time we will point out some of the subtleties as necessary.

Example 1: Let f be the function whose graph is shown. The values of f lie on a straight line except that f has a different value at $x = 2$.



The graph is not continuous at $x = 2$ because a single point has been "moved" to $(2, 0)$. Notice that $\lim_{x \to 2} = 1$, the value that $f(x)$ is "trying to achieve" at $x = 2$, although its actual value is $f(2) = 0$. The fact that the *limit* of f and the *value* of f differ is the essence of the "discontinuity of f" at that point.

Example 2: Again, consider the function $y = f(x)$ whose graph is shown.



In this case the limit of f as $x \to 3$ does not exist. As the point $x = 3$ is approached from the right the function "tries to achieve" the value 3. From the left it "tries to achieve" the value 2. It is customary to introduce the notion of *right-hand limit* and *left-hand limit* in such cases and to write $\lim_{x \to 3^+} f(x) = 3$ and $\lim_{x \to 3^-} f(x) = 2$. The actual value of $f(3)$ is irrelevant to the consideration of whether a limit does or does not exist. In fact $f(3) = 1$, differing from both the right-hand limit and the left-hand limit. Again, the function is "discontinuous" at $x = 3$.

Theorem 1: The limit of f as $x \to a$ exists if and only if both the right-hand and left-hand limits exist and have the same value. I.e.

$$\lim_{x \to a} f(x) = L \Longleftrightarrow \lim_{x \to a^-} f(x) = L \text{ and } \lim_{x \to a^+} f(x) = L$$

Example 3: $\lim_{x \to 1} \frac{x-1}{x+1} = \frac{0}{2} = 0$. In this case the function has the value 0 at the point $x = 1$, and we may thus calculate the limit by simply evaluating the function. In doing this we are implicitly using the fact

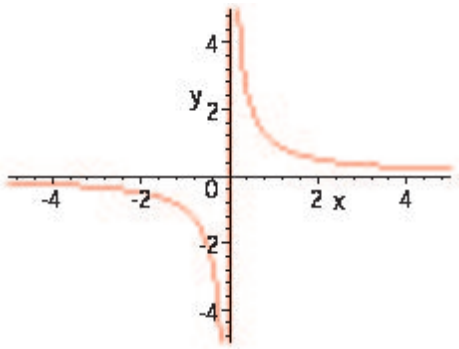
that the function is *continuous* at the point. I.e. the value that the function is "trying to achieve" at the point coincides with its actual value at the point. This characterization of "continuity" is made precise in the next sections.

Example 4: Calculate the limit $\lim_{x\to 2}(x^2-4)/(x-2)$. Here we cannot substitute the value $x=2$ into the expression, for this would yield the meaningless form "0/0". Our approach is to simplify the expression so as to make its behavior more obvious:

$$\begin{aligned}\lim_{x\to 2}\frac{x^2-4}{x-2} &= \lim_{x\to 2}\frac{(x-2)(x+2)}{x-2}\\ &= \lim_{x\to 2}(x+2)=4\end{aligned}$$

What we have done through algebraic simplification is reveal the "offending" factor $x-2$ that let to the "0/0" form. After cancelling it we can evaluate the limit by substituting $x=2$. The technique is simple. Its justification is a bit deeper. The cancellation of the factor $x-2$ is legitimate only when $x\neq 2$ since division by zero is not defined. Thus the two functions $(x^2-4)/(x-2)$ and $x+2$ are equal for all values of x *except* for $x=2$, and this means that they have the same limit at $x=2$. But the function $x+2$ is *continuous* at $x=2$ (as we will see in the following sections) and so its limit as $x\to 2$ agrees with its value there. The mystery of using the value of a function when computing its limit is thus resolved.

Example 5: Consider $\lim_{x\to 0}(1/x)$. Here the limit does not exist (DNE). The values of the function increase without bound as x approaches 0 from the right (we say that $\frac{1}{x}\to\infty$), and they decrease without bound as x approaches 0 from the left (we say that $\frac{1}{x}\to-\infty$). In particular they do not approach any real number L.



We will write $\lim_{x\to 0^+}(1/x)=\infty$ in this case even though the right-hand limit DNE (∞ is not a real number). Similarly we write $\lim_{x\to 0^-}(1/x)=-\infty$.

Example 6: Let $f(x)=(x+3)/(x-5)$ and consider $\lim_{x\to 5} f(x)$. The analysis of this example is essentially the same as that of Example 5. From the right we have $\lim_{x\to 5^+} f(x)=+\infty$ and from the left $\lim_{x\to 5^-} f(x)=-\infty$. Thus the given limit DNE.

Many of the examples above involved calculation of $\lim_{x\to a} f(x)$ where $f(x)$ is a rational function, i.e. a quotient of two polynomials. Our general line of attack for such problems was first to try substituting $x=a$. If this yields a meaningful real value L, then the limit is L. On the other hand if it yields one of the meaningless forms "0/0" or "a/0" ($a\neq 0$) we must look further. In the "a/0" case the limit does not exist nor do the right-hand or left-hand limits. But it still may be possible to classify these limits as $\pm\infty$. In the "0/0" case the limit may or may not exist. We proceed by rewriting the expression algebraically, cancelling the factor $x-a$ if possible, and then study the resulting simpler expression.

Theorem 2: If $\lim_{x\to a} f(x)=A$ and $\lim_{x\to a} g(x)=B$ both exist, then

1. $\lim_{x\to a}(f(x)+g(x))=\lim_{x\to a} f(x)+\lim_{x\to a} g(x)=A+B$
2. $\lim_{x\to a}(f(x)-g(x))=\lim_{x\to a} f(x)-\lim_{x\to a} g(x)=A-B$
3. $\lim_{x\to a}(f(x)g(x))=\lim_{x\to a} f(x)\cdot\lim_{x\to a} g(x)=A\cdot B$
4. $\lim_{x\to a}(f(x)/g(x))=\lim_{x\to a} f(x)/\lim_{x\to a} g(x)=A/B\ \ (B\neq 0)$

Example 7: Find $\lim_{x\to 1}(x^2+3x-1)/(x^3+4x+6)$. First we apply part 4 of Theorem 2, which states that the limit of a quotient is the quotient of the limits if each limit exists. Then we apply parts 1 and 3:

$$\begin{aligned}\lim_{x\to 1}\frac{x^2+3x-1}{x^3+4x+6} &= \frac{\lim_{x\to 1}(x^2+3x-1)}{\lim_{x\to 1}(x^3+4x+6)} && \text{(Thm.2, Part 4)}\\ &= \frac{1^2+3\cdot 1-1}{1^3+4\cdot 1+6} && \text{(Thm.2, Parts 1 and 3)}\\ &= \frac{3}{11}\end{aligned}$$

In this case the substitution of $x=1$ is completely meaningful, yielding the value of the limit. (We remark again that this is implicitly using the continuity of the rational function at the point $x=1$.)

Example 8: Consider $\lim_{x\to 0}\frac{|x|}{x}$. Since $|x|$ is defined piecewise ($|x|=x,\ x\geq 0$, and $|x|=-x,\ x<0$), we consider separately the right-hand and left-hand limits. For $x>0$ we have

$$\lim_{x\to 0^+}\frac{|x|}{x}=\lim_{x\to 0^+}\frac{x}{x}=1.$$

And for $x<0$ we have

$$\lim_{x\to 0^-}\frac{|x|}{x}=\lim_{x\to 0^-}\frac{-x}{x}=-1.$$

Since the right-hand and left-hand limits have different values, the limit does not exist.

Example 9: Let $f(x)=|x-1|/(x^2-1)$ and consider $\lim_{x\to 1}f(x)$. Again we consider the cases $x<1$ and $x>1$ separately. For $x<1$

$$\lim_{x\to 1^-}\frac{|x-1|}{x^2-1}=\lim_{x\to 1^-}\frac{-(x-1)}{x^2-1}=\lim_{x\to 1^-}\frac{-1}{x+1}=-\frac{1}{2},$$

and for $x>1$

$$\lim_{x\to 1^-}\frac{|x-1|}{x^2-1}=\lim_{x\to 1^-}\frac{x-1}{x^2-1}=\lim_{x\to 1^-}\frac{1}{x+1}=\frac{1}{2},$$

Since the right-hand and left-hand limits have different values, the limit DNE.

Example 10: Let $f(x)=1/x$. Let us compute $\lim_{h\to 0}[f(x+h)-f(x)]/h$.

$$\begin{aligned}\lim_{h\to 0}\frac{f(x+h)-f(x)}{h} &= \lim_{h\to 0}\frac{\frac{1}{x+h}-\frac{1}{x}}{h}\\ &= \lim_{h\to 0}\frac{\frac{x-(x+h)}{x(x+h)}}{h}\\ &= \lim_{h\to 0}\frac{\frac{-h}{x(x+h)}}{h}\\ &= \lim_{h\to 0}\frac{-1}{x(x+h)}\\ &= -\frac{1}{x^2}\end{aligned}$$

The notation $\lim_{x\to a}f(x)=L$ means that L is a real number and that the values of $f(x)$ approach L arbitrarily closely as the values of x approach a. One way to make this precise would be to invent a game: if you name a positive number ϵ, no matter how small, a second player Dr. Delta will try to respond with a positive number δ that is small enough to guarantee that $|f(x)-L|<\epsilon$ whenever $0<|x-a|<\delta$. If Dr. Delta has a winning strategy for this game, i.e. if he can win the game no matter how cleverly (how small) you choose your number ϵ, then the limit of $f(x)$ as $x\to a$ exists and is L.

The game described above is the famous (perhaps infamous) δ–ϵ definition of limit. Let's examine an example of the game in operation:

Example 11: Let $f(x) = x^2$ and let $L = 4$. Let us "prove" that $\lim_{x\to 2} f(x) = L$. Playing the game, suppose you name the positive number $\epsilon = 0.0001$. Dr. Delta will respond by naming $\delta = 0.000033$. Aha, he wins this game. For whenever $|x - 2| < 0.000033$ we see that

$$|f(x) - L| = |x^2 - 4| = |x - 2| \cdot |x + 2| \tag{2.1}$$

$$< 0.000033 \cdot 2.000033 = 0.000066001089 < 0.00001. \tag{2.2}$$

Sorry, he's gotcha! But you might then come back with a still smaller number $\epsilon = 10^{-100}$. No problem! Dr. Delta will just respond by choosing δ to be one-third of whatever number you name. For if $\delta = \frac{1}{3}\epsilon$, then whenever $|x - 2| < \delta$ we see that

$$|f(x) - L| = |x^2 - 4| = |x - 2| \cdot |x + 2| \tag{2.3}$$

$$< \frac{1}{3}\epsilon \cdot (2 + \frac{1}{3}\epsilon) < \frac{1}{3}\epsilon \cdot 3 = \epsilon \tag{2.4}$$

In computing the inequalities we used the fact that $|2 + \frac{1}{3}\epsilon| < 3$. This is certainly true if your value of ϵ is small. If you were impetuous enough to choose a large value ($\epsilon > 1/3$, for example) you would just be making Dr. Delta's life easier. What the example shows is that he has a winning strategy for playing the game. He just gives the function $\delta = \frac{1}{3}\epsilon$, and then he can go home and leave you to play the game by yourself. This means that Dr. Delta has "proved" that $\lim_{x\to a} f(x) = L$.

Ooh! Having done that example we will never do it again, at least not in this book. Newton and Leibniz, the co-inventors of calculus, never gave such proofs. Suffice it to say that should the need ever arise we could drag the δ–ϵ argument from the shelf to settle a difficult limit case. But for the large part we will rely upon our intuition about limits, and this will enable us to concentrate on the techniques of calculus that will give us so much power in solving problems.

Applet: Limits of Functions **Try it!**

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

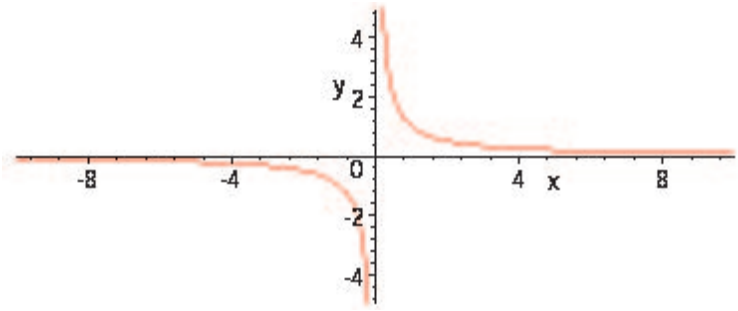
2.4 Limits at Infinity and Infinite Limits

It may be argued that the notion of *limit* is the most fundamental in calculus— indeed, calculus begins with the study of functions and limits. In many respects it is an intuitive concept. The idea of one object "approaching" another, even approaching it "arbitrarily closely" seems natural enough. Our language is full of words that express exactly such actions. And in mathematics the idea of the value of x approaching some real number L, and coming as close to it as we please, does not seem to be stretching the ideas of everyday speech. It is surprising, then, that there are such subtle consequences of the concept of limit and that it took literally thousands of years to "get it right".

In the definition of $\lim_{x\to a} f(x)$, a is a real number. It would be natural to relax this to include the cases $x \to \infty$ and $x \to -\infty$. This means that the value of x increases beyond all bounds that you might name ($x \to \infty$) or decreases below all (negative) bounds that you might name.

Definition 1: $\lim_{x\to\infty} f(x) = L$ means that the value of $f(x)$ approaches L as the value of x approaches $+\infty$. This means that $f(x)$ can be made as close to L as we please by taking the value of x sufficiently large. Similarly, $\lim_{x\to-\infty} f(x) = L$ means that $f(x)$ can be made as close to L as we please by taking the value of x sufficiently small (in the negative direction).

Example 2: $\lim_{x\to\infty}(1/x) = 0$. This is simply the observation that by taking x sufficiently large we can make its reciprocal as close to zero as we please. Similary $\lim_{x\to-\infty}(1/x) = 0$.



The line $y = 0$ is approached by the graph of $y = 1/x$ as $x \to \infty$ and also as $x \to -\infty$. It is called a *horizontal asymptote* of the graph. Also the vertical line $x = 0$ is approached by the graph as $x \to 0$. It is called a *vertical asymptote* of the graph. Finding such horizontal and vertical asymptotes for a graph aids in sketching the graph.

Example 3: $\lim_{x\to\infty} \frac{x}{x^3+2} = 0$. This limit is often said to be of the "$\frac{\infty}{\infty}$" form because both the numerator and denominator approach ∞ as x increases without bound. It is not immediately evident what the quotient does as $x \to \infty$, so we try to rewrite it in a more transparent way. Dividing the numerator and denominator by the highest power of x in the denominator often helps:

$$\lim_{x\to\infty} \frac{x}{x^3+2} = \lim_{x\to\infty} \frac{\frac{1}{x^2}}{1+\frac{2}{x^3}} = \frac{0}{1} = 0$$

After dividing the numerator and denominator by x^3 the resulting expression is again a quotient, but this time not of the "$\frac{\infty}{\infty}$" form. Both the numerator and denominator have finite limits, so we are able to evaluate the limit of the quotient as the quotient of the limits when both exist.

Example 4: Evaluate the limit $\lim_{x\to\infty}(x^4 - x^2 + 2)/(x^3 + 3)$. Again this is an "$\frac{\infty}{\infty}$" form, so we try dividing numerator and denominator by the hightest power of x in the denominator:

$$\lim_{x\to\infty} \frac{x^4 - x^2 + 2}{x^3 + 3} = \lim_{x\to\infty} \frac{x - \frac{1}{x} + \frac{2}{x^2}}{1 + \frac{3}{x^3}} = \infty$$

In the rewritten expression it is clear that the numerator approaches ∞ while the denominator approaches 1. The quotient therefore increases without bound.

Example 5: Evaluate the limit $\lim_{x\to\infty}(5x^3 + 6x^2 + x + 1)/(3x^3 + x^2 - x + 7)$. Dividing numerator and denominator by the highest power in the denominator, we have

$$\lim_{x\to\infty} \frac{5x^3 + 6x^2 + x + 1}{3x^3 + x^2 - x + 7} = \lim_{x\to\infty} \frac{5 + \frac{6}{x} + \frac{1}{x^2} + \frac{1}{x^3}}{3 + \frac{1}{x} - \frac{1}{x^2} + \frac{7}{x^3}} = \frac{5}{3}$$

The examples involve rational functions $R(x) = P(x)/Q(x)$, i.e. quotients of polynomials. Three typical situations are illustrated. In Example 3 the degree of the numerator is less than the degree of the denominator, and the limit is 0. In Example 4 the degree of the numerator is greater than the degree of the denominator, and the limit is ∞. In Example 5 the degrees of the numerator and denominator are the same, and the limit is the quotient of the coefficients of the highest power terms. These three cases are often codified as rules:

Dominant Term Rule: For the limit $\lim_{x\to\infty} P(x)/Q(x)$, where $P(x)$ is a polynomial of degree n and $Q(x)$ is a polynomial of degree m,

1. If $n < m$, the limit is 0,

2. If $n > m$, the limit is $\pm\infty$,

3. If $n = m$, the limit is the quotient of the coefficients of the highest powers.

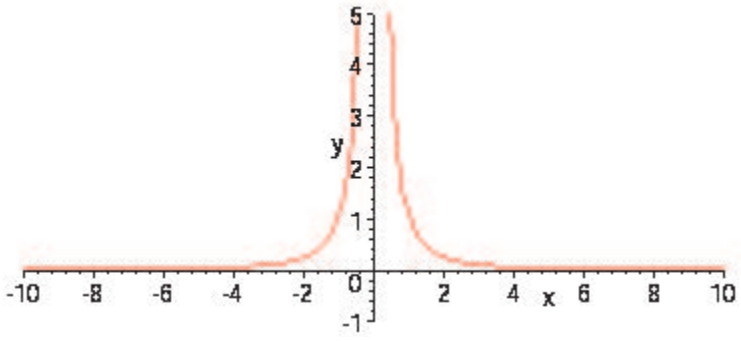
Our advice is to ignore this rule as just so much clutter. Memorizing more rules just obscures the technique illustrated in the three examples. The technique applies to more than just limits of rational functions, hence warrants your attention.

Example 6: As an example involving a non-rational function, evaluate $\lim_{x\to\infty} x/\sqrt{3x^2+2}$. In this example, thinking in the dominant term style, we suspect that the denominator will behave very much like the function $\sqrt{3x^2} = \sqrt{3}x$. Thus we guess that the limit is $1/\sqrt{3}$. This is indeed the case as the following computation shows:

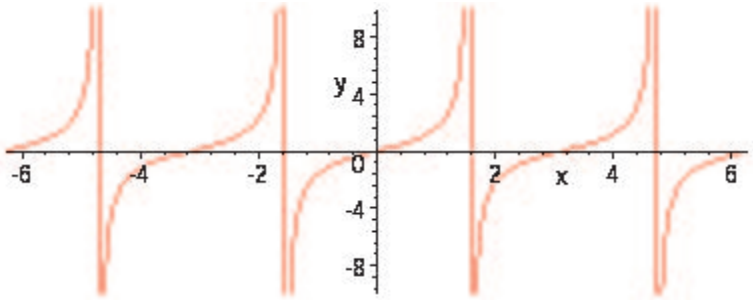
$$\begin{aligned}\lim_{x\to\infty} \frac{x}{\sqrt{3x^2+2}} &= \lim_{x\to\infty} \frac{x}{\sqrt{x^2(3+\frac{2}{x^2})}} \\ &= \lim_{x\to\infty} \frac{1}{\sqrt{3+\frac{2}{x^2}}} = \frac{1}{\sqrt{3}}\end{aligned}$$

Note that the *Dominant Term Rule* does not apply directly to this example, but the technique underlying it does. Before concluding this section, we give a few examples of infinite limits:

Example 7: Evaluate $\lim_{x\to 0} 1/x^2$. The limit does not exist, of course, since it is of the form "$\frac{1}{0}$". But let us analyse the right-hand and left-hand limits at 0. Clearly $\lim_{x\to 0^+}(1/x^2) = \infty$ as does the left-hand limit (the function is an even function). In this case the right-hand and left-hand limits do not differ, so we can also write $\lim_{x\to 0}(1/x^2) = \infty$. Although the limit DNE, the notation signals additional information about how the function $1/x^2$ behaves in the vicinity of 0. The y-axis is a vertical asymptote, and the x-axis is a horizontal asymptote.



Example 8: $\lim_{x\to\pi/2} \tan x = \lim_{x\to\pi/2} \frac{\sin x}{\cos x}$ does not exist (it is of the form "$\frac{1}{0}$"). In this case $\lim_{x\to\pi/2^+} \tan x = -\infty$ and $\lim_{x\to\pi/2^-} \tan x = \infty$ (see the graph below). The lines $x = \pi/2 + n\pi$, n any integer, are vertical asymptotes.



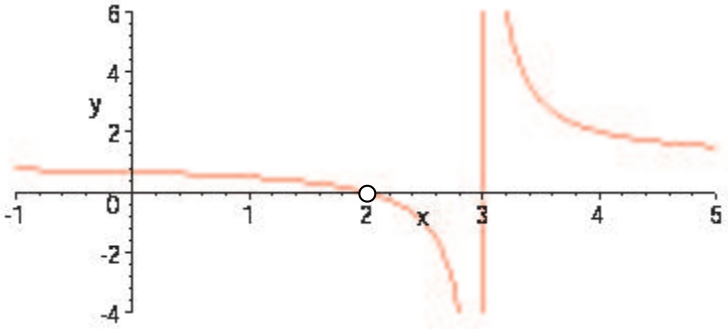
Example 9: Find the horizontal and vertical asymptotes, if any, of

$$f(x) = \frac{x-2}{(x^2-5x+6)} + 1, \quad (x \neq 2, x \neq 3)$$

and then sketch the graph. We note, first, that the denominator factors; thus

$$f(x) = \frac{x-2}{(x^2-5x+6)} + 1 = \frac{x-2}{(x-2)(x-3)} + 1 = \frac{1}{x-3} + 1 = \frac{x-2}{x-3}$$

From the last expression we see that the line $x = 3$ is a vertical asymptote and that the right-hand limit is ∞ and the left-hand limit is $-\infty$. We notice, also, that the function $(x-2)/(x-3)$ vanishes at $x = 2$. Thus the graph appears to cross the x-axis at $x = 2$ except that this point is not in the domain of the original function. The point is missing from the graph.



Summary: In this section we extended our notion of limit to include points $x = a$ where the limit does not exist but where we could use the notation $\pm\infty$ to provide additional information about the behavior of the function. We also extended the limit definition to include limits as $x \to \pm\infty$. With limits thus defined, and with our skill honed in computing some of them, we now turn in the next sections to the several ideas of calculus that depend on the notion of limit—continuity, derivative, and integral.

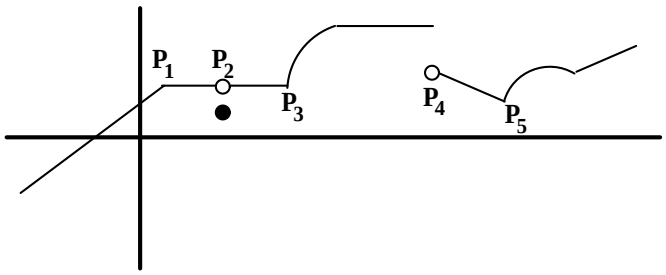
Applet: Limits of Functions **Try it!**

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

2.5 Continuity

We have used the term *continuity* in connection with many of the functions we have introduced. It is, roughly, synonymous with "unbroken"—by and large the elementary functions of calculus have graphs that can be drawn "without lifting the pencil from the paper". Some functions have breaks in their graphs, as does for example the function $f(x) = x/|x|$ that takes on the value 1 for $x > 0$ and the value -1 for $x < 0$. Drawing its graph requires "lifting the pencil" at the point $x = 0$ to jump over the break. Similarly the function whose graph is represented below is not continuous at the points P_2 and P_4. At P_2 the value of the function is "out of place", and at P_4 the graph takes a "jump". On the other hand at all other points the graph (and hence the function) is continuous. The points P_1, P_3 and P_5 are not breaks in the graph. The graph turns a sharp corner at these points, but the pencil need not be lifted.



Except in a few circumstances our intuition about continuity is quite sound. In a very precise sense we shall see that the Elementary Functions of calculus are, indeed, continuous at nearly every point and that the few places where they are not continuous are easily recognized. For example the rational function defined in Example 9 of Section 2.4 is continuous except at the two points $x = 2, 3$ where division by zero is excluded. And the function $\tan x = \sin x/\cos x$ is continuous everywhere except the points $x = \pi/2 + n \cdot \pi$ where the cosine function vanishes (cf. Example 2 in Section 1.4).

In preparation for the next definition, we will say that an *interior point* of a set of real numbers is a point that can be enclosed in an open interval that is contained in the set.

Definition 1: A function is *continuous* at an interior point c of its domain if $\lim_{x\to c} f(x) = f(c)$. If it is not continuous there, i.e. if either the limit does not exist or is not equal to $f(c)$ we will say that the function is *discontinuous* at c.

Note that we are requiring that

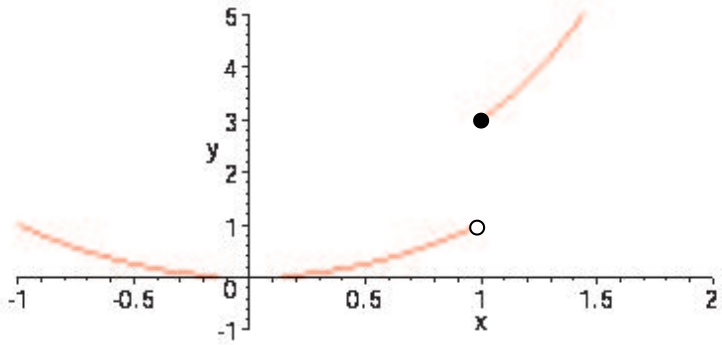
1. The function f is defined at the point $x = c$,
2. The point $x = c$ is an interior point of the domain of f,
3. $\lim_{x\to c} f(x)$ exists, call it L, and
4. $L = f(c)$.

It is the notion of limit, once again, that enables us to capture precisely the idea of continuity. The existence of the limit of $f(x)$ as $x \to c$ says, in effect, that no matter how powerful a microscope we might use to examine the graph of $f(x)$ in the vicinity of $x = c$, we would never see a "break" in the graph. The right-hand and left-hand limits both exist at c and have the same value, and moreover that value coincides with the value of $f(c)$. The plot shown above illustrates a function that has two places where it is not continuous. At P_2 the limit exists (the right-hand and left-hand limits exist and are equal) but is not equal to the value of the function. At P_4 the limit does not exist (the left-hand limit does exist and is the value of the function, but the right-hand limit has a different value). In both cases we can characterize the graph as "broken" and the function as not continuous. We will see below that we can say further, that the discontinuity at P_2 is *removable* (in a sense it is less serious: we need only redefine the value of the function at that one point). And we can say that the function is *left continuous* at P_4 because the left-hand limit exists and coincides with the value of the function.

Example 1: Is the function

$$f(x) = \begin{cases} x^2 & x < 1 \\ x^3 + 2 & 1 \le x \end{cases}$$

continuous at $x = 1$? We certainly notice that the graph of f is broken at $x = 1$ (note the graph below). Note that $f(1) = 3$. $\lim_{x\to 1^+} f(x) = \lim_{x\to 1^+}(x^3 + 2) = 3$. And $\lim_{x\to 1^-} f(x) = \lim_{x\to 1^-}(x^2) = 1$. The limit as $x \to 1$ does not exist since the right-hand and left-hand limits differ, thus the function is not continuous at $x = 1$. We could, however, say that the function is *right continuous* at $x = 1$ since the right-hand limit is equal to the function value (cf Definition 2, below).



Definition 2: A function f is *right continuous* at a point c if it is defined on an interval $[c, d]$ lying to the right of c and if $\lim_{x\to c^+} f(x) = f(c)$. Similarly it is *left continuous* at c if it is defined on an interval $[d, c]$ lying to the left of c and if $\lim_{x\to c^-} f(x) = f(c)$.

In Definition 1, the notion of continuity was defined only at *interior* points of the domain. Definition 2 enables us to speak as well of the continuity of a function at endpoints of its domain. For example we often define functions with restricted domains as with $f(x) = x^2$, $1 \le x \le 2$. And we certainly wish to assert that this function is continuous on the closed interval $[1, 2]$. The following definition accomplishes this.

Definition 3: A function f is continuous at a point $x = c$ if c is in the domain of f and:

1. If $x = c$ is an interior point of the domain of f, then $\lim_{x\to c} f(x) = f(c)$.

2. If $x = c$ is not an interior point of the domain but is an endpoint of the domain, then f must be right or left continuous at $x = c$, as appropriate. (Note: Such a point c that is not an interior point but is an endpoint of the domain will correspond to some interval of the form $[c, b)$ or $(b, c]$ that is in the domain.)

Thus, from Definition 3 we can conclude that a function f is continuous on a closed interval $[a, b]$ if it is continuous at every interior point of the interval, is right continuous at a, and is left continuous at b.

Here are several definitions that govern how we speak about the notion of continuity.

Definition 4: A function f is said to be a *continuous function* if it is continuous at every point of its domain.

Definition 5: A *point of discontinuity* of a function f is a point in the domain of f at which the function is not continuous.

Caution: Note that, according to these definitions, a function such as $f(x) = 1/x$ is a continuous function, in spite of the fact that its graph consists of two pieces. At every point where the function *is* defined it is continuous. At the single point $x = 0$ where it is *not* defined continuity is not an issue. In particular we cannot call $x = 0$ a discontinuity of f because it is not in the domain of f. (To say that f *is not continuous at* $x = 0$ is not the same as saying $x = 0$ *is a discontinuity of* f. The latter requires that 0 be in the domain of f whereas the former says nothing about whether 0 is in the domain of f.) This is a potential source of confusion, but we will be very consistent in this use of language.

In beginning this section we pointed out that most of the functions we meet in calculus and in applications are continuous wherever they are defined. In particular all polynomials, rational functions, trigonometric functions, the absolute value function, and the exponential and logarithm functions are continuous. Moreover all functions built from these using the operations of addition, subtraction, multiplication, division, composition, and taking inverses are also continuous.

Example 2: All of the following functions are continuous everywhere on their domains:

$$5x^3 - 2x + 1 \qquad \frac{3x+1}{x^2-5} \qquad \sqrt{x+2} \qquad \sin 2\pi x$$

$$(x+\tan x)^{\frac{1}{3}} \qquad \frac{\cos x}{1+\tan x} \qquad |x^2-3|$$

Example 3: The rational function $f(x)=\frac{x^2-4}{x-2}$ is a continuous function. The only point not in its domain is $x=2$. We notice, however, that $\lim_{x\to 2}=4$ exists. It is as though the function "wants to be continuous" at this point with the value 4. If we oblige by extending the definition of the function we obtain its *continuous extension*

$$F(x)=\begin{cases} f(x) & \text{if } x \text{ is in the domain of } f \\ 4 & \text{if } x=2 \end{cases}$$

We notice that $F(x)$ is none other than the function $x+2$ obtained by factoring the numerator of $f(x)$ and simplifying.

Example 4: The function

$$f(x)=\begin{cases} \sin x & x\neq \pi/3 \\ 0 & x=\pi/3 \end{cases}$$

is discontinuous at $\pi/3$. This discontinuity is removable, however, by redefining the value of f at $\pi/3$. The "proper" value would be $\lim_{x\to\pi/3} f(x)=\lim_{x\to\pi/3}\sin x=\sin\pi/3=1/2$.

Definition 6: If c is a discontinuity of a function f, and if $\lim_{x\to c} f(x)=L$ exists, then c is called a *removable discontinuity*. The discontinuity is removed by defining $f(c)=L$.

Definition 7: If f is not defined at c but $\lim_{x\to c} f(x)=L$ exists, then f has a *continuous extension* to $x=c$ by defining $f(c)=L$.

The situations in Definitions 6 and 7 are very closely related. They differ only in that $f(c)$ is defined in the first case (but is the "wrong" value) and undefined in the second. In both cases, since the limit exists and has the value L, the lack of continuity at the point is cured by simply giving f the "correct" value L. This is not an idle observation. Our very useful technique for computing limits of the form "$\frac{0}{0}$" by performing some sort of algebraic simplification amounts to finding the continuous extension of a function.

Example 5: Find $\lim_{x\to -2}(x^3+8)/(x+2)$. Solution: rewrite the fraction as

$$\frac{x^3+8}{x+2}=\frac{(x+2)(x^2+2x+4)}{x+2}=x^2+2x+4 \qquad \text{if } x\neq -2$$

The polynomial x^2+2x+4 is thus seen to be the *continuous extension* of the given rational function to $x=-2$. The limit of the rational function as $x\to -2$ is the same as the limit of the polynomial as $x\to -2$. And the limit of the polynomial is the same as its *value* at $x=-2$ since it is continuous there. Thus the value of the limit is $(-2)^2+2(-2)+4=4$. This is what lay behind our rather peculiar technique of first simplifying the rational expression and then "plugging in" the value $x=-2$ to get the limit.

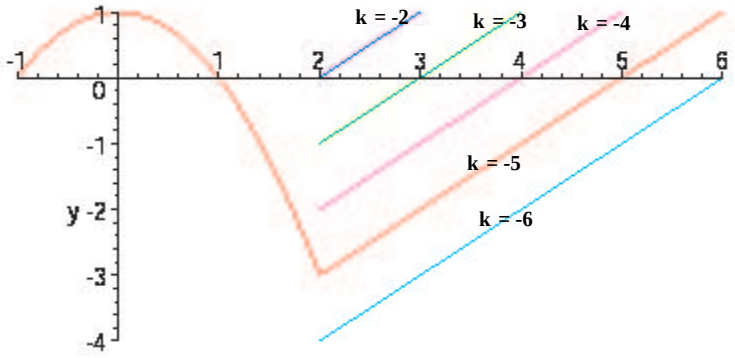
Example 6: Suppose that $f(x)$ is defined piecewise as

$$f(x)=\begin{cases} -x^2+1 & x<2 \\ x+k & x>2 \end{cases}$$

Let us find a value of the constant k such that f has a continuous extension to $x=2$. We see that $\lim_{x\to 2^+} f(x)=\lim_{x\to 2^+}(x+k)=2+k$, whereas $\lim_{x\to 2^-} f(x)=\lim_{x\to 2^-}(-x^2+1)=-3$. Thus we must have $2+k=-3$, or $k=-5$, for continuity. The desired continuous extension is then

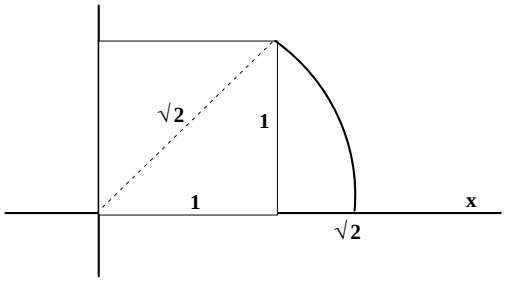
$$F(x)=\begin{cases} -x^2+1 & x<2 \\ x-5 & x\geq 2 \end{cases}$$

The function f to the left of $x=2$ is defined to be the quadratic polynomial $y=-x^2+1$. To the right it is defined to be a straight line in the family of lines $y=x+k$, k a constant. Determining k so that the right and left hand limits are equal amounts to picking out the particular straight line that meets the parabola at the point $(2,-3)$, as shown in the graph below.

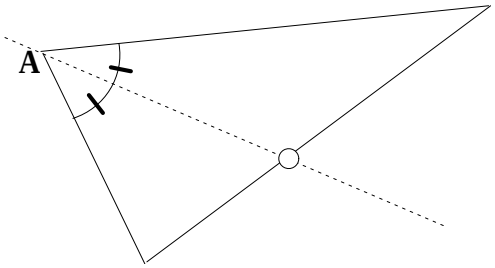


Continuity ... so what? Why such fuss over the notion of continuity? It seems like a perfectly natural property of any self-respecting function. It turns out that nearly all functions that anyone bothers with are continuous. And the only time discontinuity seems to raise its depraved head is when we take the trouble to define a function piecewise in some strange way.

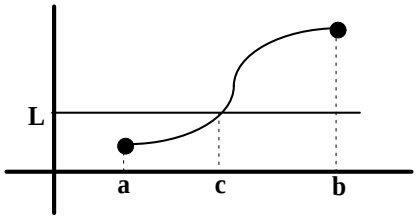
But it was not always so obvious. Pythagorus went bananas when he first noticed that the square root of 2 is not a rational number. That meant to him that a point on the x-axis could easily be constructed geometrically (see the figure below) that could not be "measured" by any number that is a quotient of integers. To the Pythagoreans, who did not believe in irrational numbers, this was a disturbing defect in nature. At issue was the *continuity* of the real line and the lack of richness in their number system for labeling every point on the line. It took humans more than 2000 years to arrive at our present-day notion of *real numbers*, filling in the gaps in the number system so that numbers correspond one-to-one to points on the geometrical line.



And Euclid, no slouch of a mathematician, stumbled over the same issue of continuity. In a number of proofs in his monumental axiomatic treatment of geometry he made the assumption that the bisector of any angle in a triangle must intersect the opposite side of the triangle. How could it possibly miss? (See the figure below.) But it was again thousands of years later when it was definitely shown that Euclid could not possibly have proved this apparently simple fact from his axioms. Again it was the notion of *continuity* that had raised its head. Euclid's axioms were not strong enough to establish the continuity of the number line. The following theorem, stated in modern language, is what enables us today to avoid Euclid's mistake.



The Intermediate Value Theorem: If a function f is continuous on a closed interval $[a, b]$, and if $f(a) < L < f(b)$ (or $f(a) > L > f(b)$), then there exists a point c in the interval $[a, b]$ such that $f(c) = L$.



Can we be forgiven for asking again "How can it possibly miss?" If the graph of f begins below the line $y = L$ and ends up above it, how can it possibly avoid passing through the line? At issue is continuity! The same issue over which Euclid stumbled. The proof of the theorem is subtle. It must show, in effect, that our system of real numbers today is finally rich enough to avoid "gaps" or "jumps" that would allow the graph to get from one side of the line $y = L$ to the other side of it without actually intersecting it. This is what continuity prevents. Within the notion of continuity lurks some two millennia of effort to understand geometry and our number system. This is what we are taking for granted when we say, from our vantage point in the 21st century, that "continuity is a simple and intuitive notion". It is ... but ....

Example 7: Show that the equation $x^5 - 3x + 1 = 0$ has a solution in the interval $[0, 1]$. It is worth remarking that we have no algebraic method for solving this equation in exact form, similar to the quadratic formula for solving a second degree equation for example. But consider the function $f(x) = x^5 - 3x + 1$. It is a continuous function, hence continuous on the interval $[0, 1]$. And clearly $f(0) = 1$ and $f(1) = -1$. The number $L = 0$ lies between 1 and -1, hence by the Intermediate Value Theorem the graph of f must intersect the line $L = 0$. I.e. there is a point c in the interval where $f(c) = 0$. So there is, indeed, a solution in the interval. Although we have no formula to express the solution, we know that one exists. Thus we can seek to find numerical approximations for the solution. This, indeed, is exactly what a computer program such as Maple does when it "solves" this equation.

Example 8: Does the equation $1/x = 0$ have a solution? Encouraged by Example 8 we may notice that the function $f(x) = 1/x$ has positive values when $x > 0$ and negative values when $x < 0$. In particular $f(-1) = -1$ and $f(1) = 1$. Can we conclude, then that there is a solution in the interval $[-1, 1]$? In this case the answer is "no"! The function is not continuous on the interval $[-1, 1]$, hence the Intermediate Value Theorem does not apply. And of course we should have known better. There is no real number whose reciprocal is zero!

Summary: Continuity is a simple and intuitive notion. It is precisely defined in terms of the notion of limit. We have seen that most functions of calculus are continuous, and that knowledge of this fact enables us to "turn the tables" and use continuity in the calculation of limits. Finally, the Intermediate Value Theorem (IVT) captures the essence of continuity—a continuous curve "has no gaps". It cannot go from one side of a straight line to the other without intersecting the line. Euclid would have danced with joy! And we can use the IVT as a tool in solving equations.

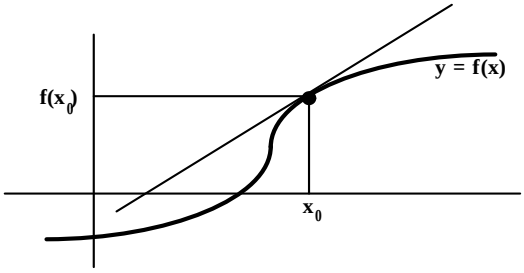
Applet: Continuity of Functions **Try it!**

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

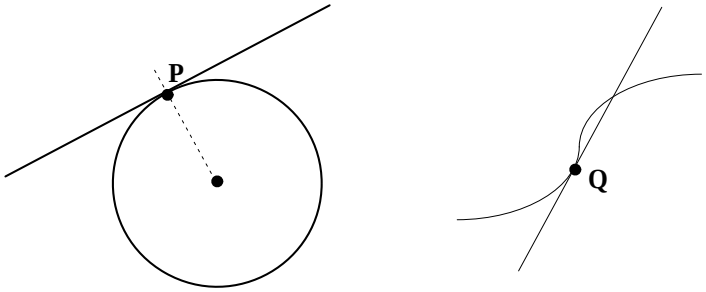
2.6 Tangent Lines and Their Slopes

The Tangent Line Problem Given a function $y = f(x)$ defined in an open interval and a point x_0 in the interval, define the tangent line at the point $(x_0, f(x_0))$ on the graph of f.

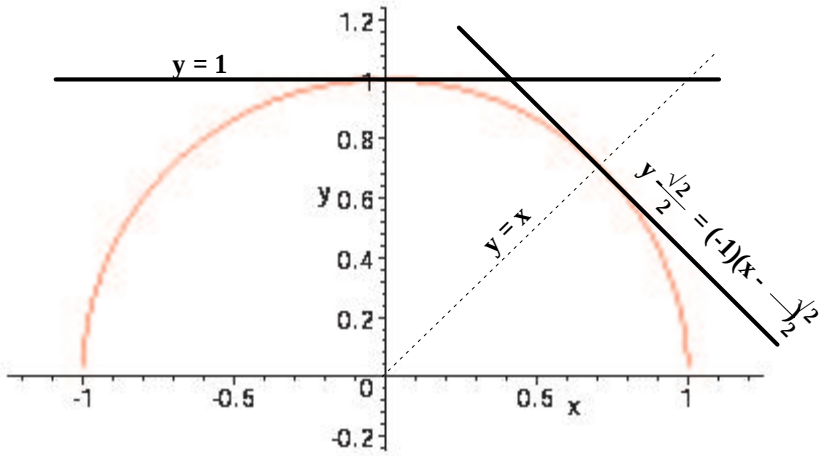


With this problem we begin our study of calculus. Indeed one can say that there are two fundamental problems in calculus—the tangent line problem, and the problem of calculating areas. The first problem leads to the study of the *derivative* and *differential calculus*. The second problem leads to *integral calculus*.

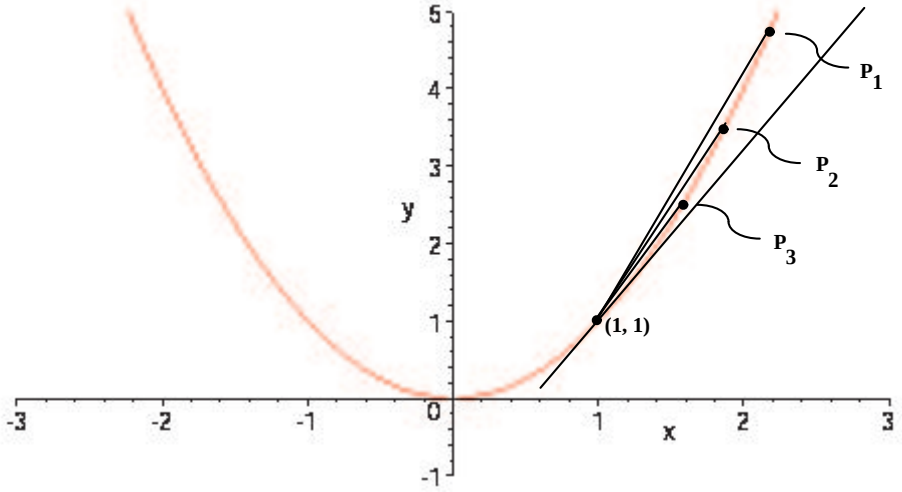
Now the problem of finding the tangent line to a curve has already arisen in geometry. There the tangent to a circle is defined as a line that intersects the circle in exactly one point P. It "touches" the circle at the point P and is perpendicular to a diameter of the circle through P. This definition suffices for the very special case of a circle, but for a more general curve it is not a satisfactory definition. We neither know how to construct a line perpendicular to the curve nor is it the case that a tangent line intersects the curve in only one point (see the figure below where the tangent line at Q intersects the curve more than once). Somehow the manner in which a tangent line intersects a curve is a very local phenomenon at the point of tangency Q—what the line does further from the point Q does not matter. It is no surprise, therefore, that a satisfactory definition of tangent line will involve the notion of limit.



Example 1: Find the equations of the tangent lines to the graph of $f(x) = \sqrt{1-x^2}$ at the points $(0,1)$ and $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$. Let us note, first of all, that the graph of f is a semi-circle and that the given points are, indeed, on the circle. Thus we can use our special knowledge of circles to solve the problem. The tangent line at $(0,1)$ is perpendicular to the y-axis, hence has the equation $y = 1$. And the tangent line at $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$ is perpendicular to the line $y = x$, hence has slope $m = -1$. Using the point-slope form $y - y_0 = m(x - x_0)$ for the equation of a line through (x_0, y_0) with slope m, we can write the equation of the second tangent line as $y - \frac{\sqrt{2}}{2} = (-1)(x - \frac{\sqrt{2}}{2})$.



Example 2: Let $f(x) = x^2$. Find the equation of the tangent line to the graph of f at the point $(1,1)$. This time we know nothing special about the



geometry of the curve, so we adopt a different procedure. Let us choose several points P_1, P_2, and P_3 on the curve and draw the secant lines from these points to the given point $(1,1)$. (See the figure.) It seems reasonable to assume that these secant lines approximate the tangent line at $(1,1)$, the approximation being better as the point P_i approaches $(1,1)$. Thus we might assume that the slopes of the secant lines approach the slope of the tangent line in the limit as P_i approaches $(1,1)$. This would give us a procedure for calculating the slope of the tangent line: namely let $P = (1+h, f(1+h))$ be a point on the graph near $(1,1)$, calculate the slope of the secant line from P to $(1,1)$, and take the limit as $h \to 0$. Carrying this out

we obtain

$$\begin{aligned}\text{Slope of tangent line} &= \lim_{h\to 0}\frac{f(1+h)-f(1)}{h}\\ &= \lim_{h\to 0}\frac{(1+h)^2-1}{h}\\ &= \lim_{h\to 0}\frac{2h+h^2}{h} = \lim_{h\to 0}(2+h) = 2\end{aligned}$$

This procedure is the basis of our formal definition:

Definition 1: Given a function f and a point x_0 in its domain, the slope of the tangent line at the point $(x_0, f(x_0))$ on the graph of f is

$$\lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h}$$

if this limit exists. The quotient in this limit (the slope $(f(x_0+h)-f(x_0))/h$ of the secant line) is called the *difference quotient*.

Note that if we let $x = x_0 + h$, then $h = x - x_0$ and we can write the above limit in the equivalent form

$$\lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0}$$

Example 3: Given $f(x) = \sqrt{x}$, find the equation of the tangent line at $x = 4$. We are seeking the line through the point $(4, 2)$ with slope given by the limit of the difference quotient (Definition 1). Thus the slope is

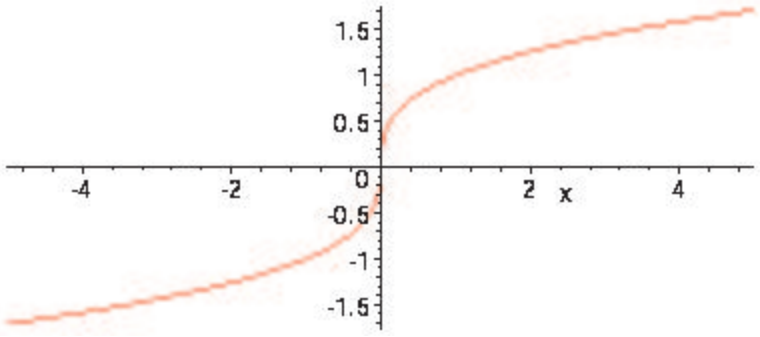
$$\begin{aligned}\text{Slope of tangent line} &= \lim_{h\to 0}\frac{f(4+h)-f(4)}{h}\\ &= \lim_{h\to 0}\frac{\sqrt{4+h}-2}{h}\\ &= \lim_{h\to 0}\frac{\sqrt{4+h}-2}{h}\cdot\frac{\sqrt{4+h}+2}{\sqrt{4+h}+2}\\ &= \lim_{h\to 0}\frac{(4+h)-4}{h(\sqrt{4+h}+2)}\\ &= \lim_{h\to 0}\frac{1}{\sqrt{4+h}+2} = 1/4\end{aligned}$$

The slope of the tangent line is $1/4$, so its equation is $y - 2 = \frac{1}{4}(x-4)$. (Note the common trick employed to "simplify" the difference quotient—multiplying numerator and denominator by an expression chosen to rationalize the numerator.)

Example 4: Apply the method of Definition 1 to find the tangent line to the graph of $f(x) = x^{1/3}$ at $x = 0$. In this case the tangent line would be $y - 0 = m(x-0)$, or $y = mx$, where

$$m = \lim_{h\to 0}\frac{(0+h)^{1/3}-0^{1/3}}{h} = \lim_{h\to 0}\frac{1}{h^{2/3}} = \infty$$

However the limit does not exist, thus the slope of the tangent line does not exist. In this case the tangent line does exist but it is vertical. Its equation is thus $x = 0$ (which is not of the form $y = mx$).



Example 5: Let f be the piecewise defined function

$$f(x) = \begin{cases} 2 - x^2 & x \leq 1 \\ x^3 & x > 1 \end{cases}$$

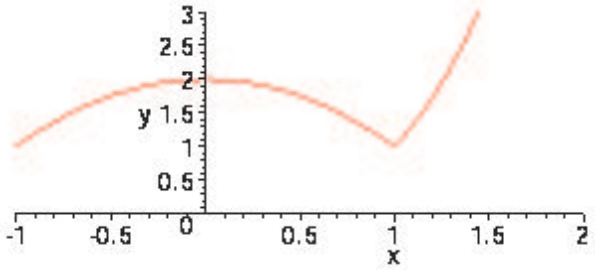
Is the function continuous, and does it have a tangent line at $x = 1$? We first notice that the function is continuous at $x = 1$ because the right-hand limit, left-hand limit, and $f(1)$, all equal 1. We must also treat the difference quotient in two cases depending on whether h is positive or negative. The limit of the difference quotient from the left is

$$\lim_{h \to 0^-} \frac{[2 - (1+h)^2] - 1}{h} = \lim_{h \to 0^-} \frac{-2h - h^2}{h} = \lim_{h \to 0^-} (-2 - h) = -2$$

and from the right the limit is

$$\lim_{h \to 0^+} \frac{(1+h)^3 - 1}{h} = \lim_{h \to 0^+} \frac{3h + 3h^2 + h^3}{h} = \lim_{h \to 0^+} (3 + 3h + h^2) = 3$$

The right and left limits are not equal, hence the slope is undefined. In this case there is no unique tangent line—the graph has a sharp corner at $(1, 1)$.



Applet: Secant and Tangent Lines **Try it!**
Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

2.7 The Derivative

The tangent line problem has been solved. Given a function f and a point x_0 in its domain, the tangent line to the graph of f at the point $(x_0, f(x_0))$ is given by $y = f(x_0) + m(x - x_0)$, where $m = \lim_{h\to 0}(f(x_0 + h) - f(x_0))/h$, provided the limit exists. The quotients $(f(x_0 + h) - f(x_0))/h$ represent the average rate of change of f over the interval $[x_0, x_0 + h]$.

The method is quite general, the point x_0 being *any* point in the domain of f. Thus a simple change in our point of view allows us to focus on the *rule* whereby the slope m is computed from any point x in the domain of f. The new function defined by this rule is called the *derivative function*, or simply the *derivative of* f:

Definition 1: The *derivative* of a function f is a new function defined by

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

The domain of f' is the set of points x where this limit exists, i.e. the set of points where the graph of f has a tangent line. The equation of the tangent line at the point $(x_0, f(x_0))$ is $y = f(x_0) + f'(x_0)(x - x_0)$.

We will say that a function f is *differentiable* at a point $x = a$ if the derivative function f' exists at a.

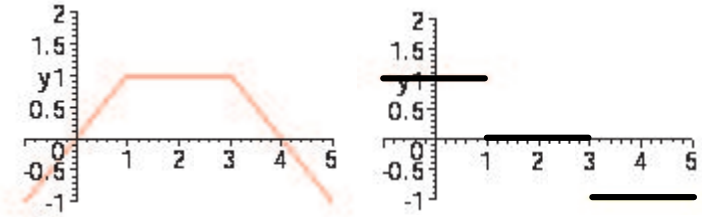
Example 1: Suppose we consider the piecewise defined function

$$f(x) = \begin{cases} x & x \le 1 \\ 1 & 1 < x < 3 \\ -x + 4 & 3 \le x \end{cases}$$

Let us find the derivative function f'. This is an especially simple case, the graph of f consisting of three pieces of straight lines. Since $f'(x)$ represents the slope of the graph at x (i.e. the slope of the tangent line to the graph), we know immediately that

$$f'(x) = \begin{cases} 1 & x < 1 \\ 0 & 1 < x < 3 \\ -1 & 3 < x \end{cases}$$

At $x = 1$ and $x = 3$ the derivative is not defined; therefore, f is not differentiable at those points. The graph has sharp "corners" at those two points. At $x = 1$ the slope "jumps" from 1 on the left to 0 on the right. And at $x = 3$ it jumps from 0 on the left to -1 on the right. The derivative function f' is not continuous at those points.



Examples of some commonly encountered functions follow:

Example 2: $f(x) = k$, where k is a constant. Then the graph of f is a horizontal straight line, with slope zero at every point. Thus $f'(x) = 0$ for all x. We notice that this is exactly the result we obtain using Definition 1:

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{k - k}{h} = \lim_{h\to 0} \frac{0}{h} = 0$$

Example 3: (Linear function) $f(x) = ax + b$, a, b constants. Then $f'(x) = a$, since the graph is a straight line with slope a. Again, using Definition 1 we obtain the same result:

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{[a(x+h) + b] - [ax + b]}{h} = \lim_{h\to 0} \frac{ah}{h} = a$$

Example 4: (The derivative of x^2) For $f(x) = x^2$, we have

$$\lim_{h\to 0} \frac{(x+h)^2 - x^2}{h} = \lim_{h\to 0} \frac{2xh + h^2}{h} = \lim_{h\to 0} (2x + h) = 2x.$$

Example 5: (The derivative of x^3) For $f(x) = x^3$, we have

$$\lim_{h\to 0} \frac{(x+h)^3 - x^3}{h} = \lim_{h\to 0} \frac{3x^2h + 3xh^2 + h^3}{h} = \lim_{h\to 0} (3x^2 + 3xh + h^2) = 3x^2.$$

Example 6: (The derivative of $1/x$) For $f(x) = 1/x$, we have

$$\lim_{h\to 0} \frac{\frac{1}{x+h} - \frac{1}{x}}{h} = \lim_{h\to 0} \frac{x - (x+h)}{hx(x+h)} = \lim_{h\to 0} \frac{-1}{x(x+h)} = \frac{-1}{x^2}.$$

Example 7: (The derivative of $\sqrt{x}$) For $f(x) = \sqrt{x}$, we have

$$\begin{aligned} \lim_{h\to 0} \frac{\sqrt{x+h} - \sqrt{x}}{h} &= \lim_{h\to 0} \frac{\sqrt{x+h} - \sqrt{x}}{h} \cdot \frac{\sqrt{x+h} + \sqrt{x}}{\sqrt{x+h} + \sqrt{x}} \\ &= \lim_{h\to 0} \frac{x+h-x}{h(\sqrt{x+h} + \sqrt{x})} = \frac{1}{2\sqrt{x}}. \end{aligned}$$

Several of the previous examples are special cases of the following general rule:

Theorem 1 (The Power Rule): Suppose that $f(x) = x^r$, where r is any real number. Then $f'(x) = rx^{r-1}$.

It is immediately clear that Examples 4 and 5 verify this rule. In Example 6, if we write $1/x = x^{-1}$, we see that the rule also applies and yields the result $(-1)x^{-2} = -1/x^2$. And in Example 7, writing $\sqrt{x} = x^{1/2}$, the rule yields $(1/2)x^{-1/2}$, which is the same result we obtained using the "rationalization" trick.

We will verify a number of special cases of the Power Rule as we develop appropriate tools, and eventually, when we have precisely defined the general power function x^r, we will prove the rule in its most general form.

Example 8: Find an equation of the tangent line to the graph of $f(x) = x^{4/3}$ at the point where $x = 1$. We know that the desired equation can be written as $y = f(1) + f'(1)(x-1) = 1 + f'(1)(x-1)$, thus we need only compute $f'(1)$. From the Power Rule we have $f'(x) = (4/3)x^{4/3-1} = (4/3)x^{1/3}$. So, $f'(1) = 4/3$, and the equation of the tangent line is $y = 1 + (4/3)(x-1)$.

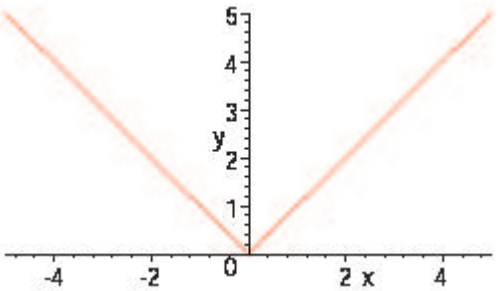
Example 9: Find the derivative of $f(x) = |x|$. We notice immediately that $f'(x) = -1$ when $x < 0$, and that $f'(x) = 1$ when $x > 0$. What happens at $x = 0$? It should be clear from the graph of f, which has a sharp corner at $x = 0$, that $f'(0)$ is not defined. Indeed

$$\lim_{h\to 0^+} \frac{|0+h| - 0}{h} = \lim_{h\to 0^+} \frac{h}{h} = 1$$

and

$$\lim_{h\to 0^-} \frac{|0+h| - 0}{h} = \lim_{h\to 0^+} \frac{-h}{h} = -1.$$

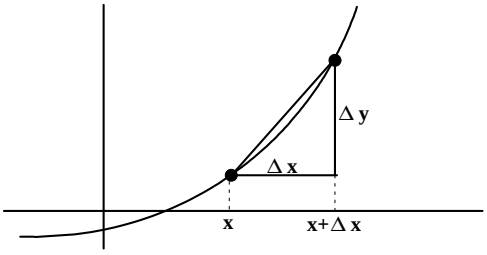
Since the right-hand and left-hand limits of the difference quotient are not the same the limit does not exist. I.e. $f'(0)$ is not defined, and hence there is no tangent line to the graph at the point $(0, 0)$.



Notation for the Derivative: Thus far we have denoted the derivative of f as f'. Corresponding to our many ways of writing functions, for example the function $f(x) = x^2$ may be written as $y = x^2$, x^2, or simply y, we can also write the derivative of f in a number of ways:

$$y' = D_x y = D_x x^2 = \frac{dy}{dx} = \frac{d}{dx} f(x) = f'(x)$$

All of these expressions can be read as "take the derivative of the function f". The notation $\frac{dy}{dx}$ is especially interesting. it was introduced by Leibniz to suggest that the derivative is a limit of a difference quotient.



Writing Δx instead of h to denote a "small" change in x, and $\Delta y = f(x + \Delta x) - f(x)$ to denote the corresponding change in y, then the difference quotient (slope of the secant line) is $\frac{\Delta y}{\Delta x}$. As Δx becomes small the quotient $\frac{\Delta y}{\Delta x}$ becomes a better and better approximation for the slope of the graph at $(x, f(x))$, and in the limit as $\Delta x \to 0$ approaches the value of $f'(x)$.

The notation $f'(x)$ lends itself easily to denoting the value of the derivative at a particular point $x = x_0$, namely as $f'(x_0)$. When using the notation $\frac{dy}{dx}$ for the derivative, we will customarily denote its value at x_0 by $\left.\frac{dy}{dx}\right|_{x=x_0}$.

Example 10: For the function $y = f(x) = 1/x$, find the slope of its tangent line at $x = 2$. Compare it with the average rate of change over the interval $[2, 3]$. Using the notations introduced above we can write

$$f'(2) = \left.\frac{dy}{dx}\right|_{x=2} = \left.\frac{d}{dx}\left(\frac{1}{x}\right)\right|_{x=2} = \left.\frac{d}{dx}(x^{-1})\right|_{x=2} = \left.\left(-\frac{1}{x^2}\right)\right|_{x=2} = -\frac{1}{2^2} = -\frac{1}{4}$$

The notation of Leibniz is very convenient and flexible, and we will make heavy use of it. The average rate of change, or the slope of the secant line, on the interval $[2, 3]$ is $\frac{\Delta y}{\Delta x}$ or $(1/3 - 1/2)/(3 - 2) = -1/6$.

Higher Order Derivatives: We can extend our notation to the case of repeated differentiation. When we differentiate a function $f(x)$ we obtain a new function $f'(x)$. The derivative is again a candidate for differentiation, and we call its derivative *the second derivative of* $f(x)$. So long as the derivatives exist we can continue this process to obtain a succession of higher derivatives. There are a variety of notations for higher derivatives just as there are many alternative notations for y'. Depending on the circumstances we may denote the second derivative, for example, as one of the following:

$$y'' = f''(x) = \frac{d^2y}{dx^2} = \frac{d}{dx}\frac{d}{dx} f(x) = \frac{d^2}{dx^2} f(x) = D_x^2 y = D_x^2 f(x).$$

Third derivatives may be denoted by

$$y''' = f'''(x) = \frac{d^3y}{dx^3} = \frac{d^3}{dx^3} f(x) = D_x^3 y = D_x^3 f(x),$$

and in general, the n^{th} derivative, where n is a positive integer, may be denoted by

$$y^{(n)} = f^{(n)}(x) = \frac{d^n y}{dx^n} = \frac{d^n}{dx^n} f(x) = D_x^n y = D_x^n f(x),$$

Just as the first derivative has many interpretations that lead to applications—rate of change, slope of a curve, velocity—so do higher derivatives. The second derivative, for example, can represent acceleration—the rate of change of velocity. And we will see, in the next chapter, that it has an important geometrical

interpretation related to graphs—concavity, or curvature. Third and fourth derivatives occur in describing engineering problems related to the elastic bending of beams.

Summary: The tangent line problem led us to define the slope of a curve at a point in terms of the limit of slopes of secant lines. This culminated in the notion of *derivative of a function* $y = f(x)$ — a new function $y' = f'(x)$ that gives the slope at any point x. We have illustrated the computation of derivatives graphically (in simple cases), through the direct application of the limit definition (Definition 1), and through the use of rules that can greatly speed the computation of derivatives (e.g. the Power Rule). In the next sections we will greatly expand our repertoire of differentiation rules to provide powerful ways of quickly computing derivatives. And, armed with such power, we will tackle a variety of applications of the derivative to problems in science, economics, and other fields.

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

2.8 Differentiation Rules

The Power Rule is an example of a *differentiation rule*. For functions of the form x^r, where r is a constant real number, we can simply write down the derivative rather than go through a long computation of the limit of a difference quotient. Developing a repertoire of such basic rules, and gaining skill in using them, is a large part of what calculus is about. Indeed, calculus may be described as the study of elementary functions, a few of their basic properties such as *continuity* and *differentiability*, and a toolbox of computational techniques for computing derivatives (and later integrals). It is the latter computational ingredient that most students recall with such pleasure as they reflect back on learning calculus. And it is skill with those techniques that enables one to apply calculus to a large variety of applications.

Building the Toolbox We begin with an observation. It is possible for a function f to be *continuous* at a point a and not be *differentiable* at a. Our favorite example is the absolute value function $|x|$ which is continuous at $x = 0$ but whose derivative does not exist there. The graph of $|x|$ has a sharp corner at the point $(0, 0)$ (cf. Example 9 in Section 1.3). Continuity says only that the graph has no "gaps" or "jumps", whereas differentiability says something more—the graph not only is not "broken" at the point but it is in fact "smooth". It has no "corner". The following theorem formalizes this important fact:

Theorem 1: If $f'(a)$ exists, then f is continuous at a.

The proof is not difficult. To show that f is continuous at a we must show that $\lim_{x\to a} f(x) = f(a)$, or that $\lim_{h\to 0} f(a+h) = f(a)$. We will accomplish this by showing that $\lim_{h\to 0}(f(a+h) - f(a)) = 0$:

$$\begin{aligned}\lim_{h\to 0}(f(a+h) - f(a)) &= \lim_{h\to 0}\frac{f(a+h)-f(a)}{h}\cdot h \\ &= \lim_{h\to 0}\frac{f(a+h)-f(a)}{h}\cdot\lim_{h\to 0} h = f'(a)\cdot 0 = 0\end{aligned}$$

We have used in the proof only elementary properties of limits concerning sums or products of two functions. Notice that the crucial step was in isolating the difference quotient $(f(a+h) - f(a))/h$, whose limit, $f'(a)$, exists by our assumption.

The theorem confirms our intuition that *differentiability* is a stronger notion than *continuity*. A function can be continuous without being differentiable, but it cannot be differentiable without also being continuous. A function whose derivative exists at every point of an interval is not only continuous, it is *smooth*, i.e. it has no sharp corners.

We now proceed to develop differentiation rules. Cognizant of the way functions are built from a small number of simple functions using algebraic operations and composition, we examine how differentiation regards these operations. In the theorems that follow we assume that f and g are functions whose derivatives f' and g' exist.

Theorem 2: Suppose $y = f(x)$ is a function that has derivative f'. Then, $(cf)' = cf'$, where c is a constant. Or in Leibniz's notation $\frac{d}{dx}(cf(x)) = c\cdot\frac{d}{dx}f(x)$.

A proof simply uses the corresponding property of limits:

$$\frac{d}{dx}(cf(x)) = \lim_{h\to 0}\frac{cf(x+h)-cf(x)}{h} = c\cdot\lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = c\cdot f'(x).$$

Just as constants can be moved outside a limit, so they can be moved outside the operation of differentiation.

Example 1: $(3x^2)' = 3(x^2)' = 3\cdot 2x = 6x$. In Leibniz's notation $\frac{d}{dx}(3x^2) = 3\cdot\frac{d}{dx}(x^2) = 3\cdot 2x = 6x$.

Theorem 3: If f and g are functions with derivatives f' and g', respectively, then $(f+g)' = f' + g'$. In words, the derivative of a sum is the sum of the derivatives.

Again, this follows immediately from the corresponding property of limits:

$$\begin{aligned}\frac{d}{dx}(f(x)+g(x)) &= \lim_{h\to 0}\frac{[f(x+h)+g(x+h)]-[f(x)+g(x)]}{h} \\ &= \lim_{h\to 0}\frac{f(x+h)-f(x)}{h} + \lim_{h\to 0}\frac{g(x+h)-g(x)}{h} = f'(x)+g'(x)\end{aligned}$$

In fact it follows that the derivative of any number of terms is the sum of the derivatives of each term. For example $(f+g+h)' = ((f+g)+h)' = (f+g)' + h' = f' + g' + h'$.

Example 2: Theorems 2 and 3 taken together enable us to differentiate any polynomial. For example

$$\begin{aligned}\frac{d}{dx}(3x^2+2x+7) &= \frac{d}{dx}(3x^2)+\frac{d}{dx}(2x)+\frac{d}{dx}(7)= \\ &= 3\cdot\frac{d}{dx}(x^2)+2\cdot\frac{d}{dx}(x)+\frac{d}{dx}(7) \\ &= 3\cdot 2x+2\cdot 1+0=6x+2.\end{aligned}$$

And, similarly

$$\frac{d}{dx}(x+\sqrt{x})=1+\frac{d}{dx}x^{1/2}=1+(1/2)x^{-1/2}=1+\frac{1}{2\sqrt{x}}.$$

Theorem 4 (The Product Rule): If f and g are functions with derivatives f' and g', respectively, then $(fg)'=fg'+gf'$. In words, "the derivative of a product is the first factor times the derivative of the second, plus the second factor times the derivative of the first".

It is, in fact, useful to learn to state the theorem in words. Comparing a given example to the mathematical statement is prone to error, whereas carrying out the necessary computations while reciting the rule is a convenient skill to learn. We look at several examples of the rule in action and then provide a proof.

Example 3: Find $f'(x)$ in two ways, given $f(x)=(5x+3)(x+2)$. The first way, of course, might be to multiply out the given expression and then differentiate the resulting polynomial: $[(5x+3)(x+2)]'=(5x^2+13x+6)'=10x+13$. Using the product rule we get

$$\begin{aligned}[(5x+3)(x+2)]' &= (5x+3)(x+2)'+(x+2)(5x+3)' \\ &= (5x+3)\cdot 1+(x+2)\cdot 5 \\ &= 5x+3+5x+10 \\ &= 10x+13.\end{aligned}$$

Example 4: If $y=\sqrt{x}(x^2+2)$, find $\frac{dy}{dx}$. Using the product rule, this time carrying out the computations as we recite the rule:

$$\frac{dy}{dx}=\sqrt{x}\cdot 2x+(x^2+2)\frac{1}{2\sqrt{x}}.$$

Students always ask the question "Must I simplify this?" The answer is "yes" if you know what you want to do with the derivative or what other results it needs to be compared with. This is normally the case. However, note that there is no unique "simplest" form. It definitely does depend on what use you intend to make of the result. A reasonable simplification in this case might be $y'=2x^{3/2}+(1/2)x^{3/2}+(1/2)x^{-1/2}=(5/2)x^{3/2}+(1/2)x^{-1/2}$.

Theorem 5 (The Reciprocal Rule): Suppose f has derivative f'. Then for any x such that $f(x)\neq 0$, $\left(\frac{1}{f}\right)'(x)=-\frac{f(x)'}{f(x)^2}$. That is, $\left(\frac{1}{f}\right)'=-\frac{f'}{f^2}$.

Again, this follows from the limit definition:

$$\begin{aligned}\left(\frac{1}{f}\right)'(x) &= \lim_{h\to 0}\frac{\left(\frac{1}{f}\right)(x+h)-\left(\frac{1}{f}\right)(x)}{h} \\ &= \lim_{h\to 0}\frac{\frac{1}{f(x+h)}-\frac{1}{f(x)}}{h} \\ &= \lim_{h\to 0}\frac{f(x)-f(x+h)}{hf(x)f(x+h)} \\ &= -\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}\frac{1}{f(x+h)}\frac{1}{f(x)} \\ &= -\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}\lim_{h\to 0}\frac{1}{f(x+h)}\lim_{h\to 0}\frac{1}{f(x)} \\ &= -f'(x)\frac{1}{f(x)}\frac{1}{f(x)}=-\frac{f'(x)}{f(x)^2}\end{aligned}$$

In evaluating the three limits, we recognized the first as the definition of $f'(x)$. In the second we used the continuity of f at x (Theorem 1). And the third was independent of h.

Example 5: Let $f(x) = \frac{1}{x^2+1}$. Then $f'(x) = -\frac{2x}{(x^2+1)^2}$. Again we wrote this down while we recited the rule "minus the derivative of the denominator divided by the square of the denominator".

Theorem 6 (The Quotient Rule): Suppose f and g have derivatives f' and g', respectively. Then for any x such that $g(x) \neq 0$, $\left(\frac{f}{g}\right)'(x) = \frac{g(x)f(x)' - f(x)g(x)'}{g(x)^2}$. That is, $\left(\frac{f}{g}\right)' = \frac{gf' - fg'}{g^2}$. In words, "the derivative of a quotient is the denominator times the derivative of the numerator minus the numerator times the derivative of the denominator all divided by the denominator squared".

The quotient rule is really just the product and reciprocal rules combined, for

$$\left(\frac{f}{g}\right)' = \left(f \cdot \frac{1}{g}\right)' = f \cdot \left(-\frac{g'}{g^2}\right) + \frac{1}{g} \cdot f' = \frac{gf' - fg'}{g^2}.$$

Example 6: $f(x) = \frac{x+1}{x+2}$. Then, writing as we recite the rule:

$$f'(x) = \frac{(x+2)(1) - (x+1)(1)}{(x+2)^2} = \frac{x+2-x-1}{(x+2)^2} = \frac{1}{(x+2)^2}.$$

Example 7: $f(x) = \frac{1+\sqrt{x}}{x^2+3x+2}$. Then

$$f'(x) = \frac{(x^2+3x+2)\frac{1}{2\sqrt{x}} - (1+\sqrt{x})(2x+3)}{(x^2+3x+2)^2}.$$

It would try one's patience to obtain this result using the limit definition instead of the quotient rule. Should we simplify it? No, unless we know what use is to be made of it.

Example 8: For $f(x) = \frac{1}{x} = x^{-1}$, find the derivative three ways, using the power rule, the reciprocal rule, and the quotient rule:

Power Rule: $f'(x) = (-1)x^{-2} = -\frac{1}{x^2}$

Reciprocal Rule: $f'(x) = -\frac{1}{x^2}$

Quotient Rule: $f'(x) = \frac{x \cdot 0 - 1 \cdot 1}{x^2} = -\frac{1}{x^2}$

The collection of rules that we now have enable us to write down the derivatives of a remarkable variety of functions, knowning only the derivatives of a few basic functions. There is one situation not covered by our rules, however, namely how do we deal with the composition of functions? How would we differentiate $\sqrt{x^2+1}$, for example? We will add one final rule to our arsenal to handle functions that are built up by the operation of composition, the so-called *chain rule*. It is perhaps the most important differentiation rule of all.

Theorem 7 (The Chain Rule): Let $(f \circ g)(x) = f(g(x))$ be the function defined from f and g by composition. Assume that g is differentiable at the point x and that f is differentiable at the point $g(x)$. Then the composite function $f \circ g$ is differentiable at the point x, and

$$(f \circ g)'(x) = [f(g(x))]' = f'(g(x))g'(x).$$

Leibniz' notation gives us a useful alternative way to write the chain rule. If we define $u = g(x)$, we can write the composition as the "chain" of functions $y = f(u)$, where $u = g(x)$. Then the chain rule takes the form

$$\frac{dy}{dx} = f'(u)\frac{du}{dx} = \left.\frac{dy}{du}\right|_{u=g(x)} \cdot \frac{du}{dx}, \qquad \text{or simply } \frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

if we remember that the first factor $\frac{dy}{du}$ is evaluated at $u = g(x)$.

Example 9: For the function $\sqrt{x^2+1}$ we would have, applying the first statement of the chain rule,

$$\left[\sqrt{x^2+1}\right]' = \frac{1}{2\sqrt{x^2+1}} \cdot 2x$$

To apply the second form of the rule we write $y = \sqrt{u}$, where $u = x^2 + 1$; then we have

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} = \frac{1}{2\sqrt{u}} \frac{du}{dx} = \frac{1}{2\sqrt{u}} \cdot 2x = \frac{1}{2\sqrt{x^2+1}} \cdot 2x = \frac{x}{\sqrt{x^2+1}}.$$

Leibniz' notation really comes into full bloom in writing the chain rule in the second form, above. Remembering that derivatives measure *rate of change*, we interpret $\frac{dy}{du}$ as measuring how much faster y changes than u, and $\frac{du}{dx}$ as measuring how much faster u changes than x. Thus it seems perfectly natural that $\frac{dy}{dx}$ should be the product of these two derivatives, measuring how much faster y changes than x. (If y changes twice as fast as u, and u changes three times as fast as x, then y is changing six times as fast as x.)

Before coming back to a proof of the chain rule we consider a few more examples that illustrate the ease of its use in practice.

Example 10: Differentiate $y = (x^2 + 2)^{10}$. Writing this as $y = u^{10}$, where $u = x^2 + 2$, we have

$$\frac{dy}{dx} = \frac{dy}{du} \frac{du}{dx} = 10u^9 \cdot 2x = 20x(x^2 + 2)^9.$$

As with nearly any rule, after gaining some facility with its use one can find shortcuts. Instead of explicitly substituting $u = g(x)$ one simply "thinks it" instead. To different $f(g(x))$, then, one "thinks" of the inside function $g(x)$ as an indivisible *glob*, and recites "take the derivative of f with respect to *glob* and then multiply by the derivative of *glob* with respect to x". In this way the derivatives of many composite functions may be written down directly as one recites the rule.

Example 11: Differentiate $f(x) = (1 + 3\sqrt{x})^{35}$. Thinking of $1 + 3\sqrt{x}$ as the *glob* in this case, we think, and write, immediately

$$f'(x) = 35(glob)^{34} \cdot \frac{d}{dx}(glob) = 35(1 + 3\sqrt{x})^{34} \cdot 3\frac{1}{2\sqrt{x}}$$

(Of course, we normally don't reveal the "*glob*" part outside the family.)

Example 12: For $f(x) = \left(\frac{x+1}{x^2+1}\right)^3$, we have

$$f'(x) = 3\left(\frac{x+1}{x^2+1}\right)^2 \cdot \frac{(x^2+1)(1) - (x+1)(2x)}{(x^2+1)^2}.$$

Simplify? Sure, go ahead.

Let us prove the Chain Rule: Assume that $y = f(g(x))$, that g is differentiable at x_0, and f is differentiable at $g(x_0)$. Then we must show that $f(g(x))$ is differentiable at x_0 and that $(f \circ g)'(x_0) = f'(g(x_0))g'(x_0)$. As usual we begin with the limit definition of the derivative at x_0:

$$(f \circ g)'(x_0) = \lim_{x \to x_0} \frac{f(g(x)) - f(g(x_0))}{x - x_0},$$

where we must show that the limit exists and has the given value. Can we rewrite the difference quotient in a more transparent form? A naive (and not completely correct) first step might be to multiply and divide by $g(x) - g(x_0)$ as follows:

$$\frac{f(g(x)) - f(g(x_0))}{x - x_0} = \frac{f(g(x)) - f(g(x_0))}{g(x) - g(x_0)} \cdot \frac{g(x) - g(x_0)}{x - x_0}$$

The second factor we immediately recognize to be the difference quotient for g, whose limit as $x \to x_0$ is $g'(x_0)$. And, feeling on a roll, we notice that the first factor is also a difference quotient of sorts—it is the slope of a secant line to the graph of f at the point $g(x_0)$. As such, when $x \to x_0$, $g(x) \to g(x_0)$ (g is continuous at x_0), and the quotient should approach the slope $f'(g(x_0))$ of the graph of f. This would yield exactly the expression $f'(g(x_0))g'(x_0)$ that we want.

What, if anything, is wrong with the above "proof"? The one sticky point is that we multiplied *and divided* by $g(x) - g(x_0)$, and this would be a problem if we were dividing by zero. Could $g(x) - g(x_0) = 0$ for

values of x different from x_0? And could this happen for values of x arbitrarily close to x_0? The bad news is that it *can*, even though such circumstances are rare. But it takes only *one* exception to render our proof invalid. The good news is that we can fix the problem by taking a slightly closer look at the argument we gave above.

We begin with our assumption that f is differentiable at the point $u_0 = g(x_0)$, i.e. that $\lim_{u\to u_0}(f(u) - f(u_0))/(u - u_0)$ exists and is equal to $f'(u_0)$. Let us introduce the function

$$Q(u) = \begin{cases} \frac{f(u)-f(u_0)}{u-u_0} & \text{if } u \neq 0 \\ \\ f'(u_0) & \text{if } u = u_0 \end{cases}$$

and notice that it is simply the continuous extension of the difference quotient to the point u_0. I.e. $\lim_{u\to u_0} Q(u) = f'(u_0) = Q(u_0)$. Notice also that $f(u)(u - u_0) = f(u) - f(u_0)$. (If $u \neq u_0$ this is immediate from the definition of Q, and if $u = u_0$ it is obvious since both sides of the equation are zero.) In particular, $f(g(x)) - f(g(x_0)) = f(u) - f(u_0) = Q(u)(u - u_0) = Q(g(x))(g(x) - g(x_0))$, and we can return to our initial argument:

$$\begin{aligned} (f \circ g)'(x_0) &= \lim_{x\to x_0} \frac{f(g(x)) - f(g(x_0))}{x - x_0} \\ &= \lim_{x\to x_0} \frac{Q(g(x))(g(x) - g(x_0))}{x - x_0} \\ &= \lim_{x\to x_0} Q(g(x)) \lim_{x\to x_0} \frac{g(x) - g(x_0)}{x - x_0} \\ &= Q(g(x_0))g'(x_0) \\ &= Q(u_0)g'(x_0) = f'(u_0)g'(x_0) = f'(g(x_0))g'(x_0) \end{aligned}$$

In concluding that $\lim_{x\to x_0} Q(g(x)) = Q(g(x_0))$ we made key use of the continuity of g at x_0 and Q at $u_0 = g(x_0)$.

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

2.9 Derivatives of the Trigonometric Functions

The trigonometric functions are of fundamental importance in modeling periodic phenomena—light and sound waves, oscillating crystals, time-keeping devices, and a myriad of similar periodic motions. Their derivatives measure the velocity, frequency, and energy of such physical systems and, as a result, occur in differential equations that describe such systems. In this section we develop the essential differentiation rules for sine, cosine, and other trigonometric functions.

We begin with the function $\sin x$. Resorting to the limit definition of derivative:

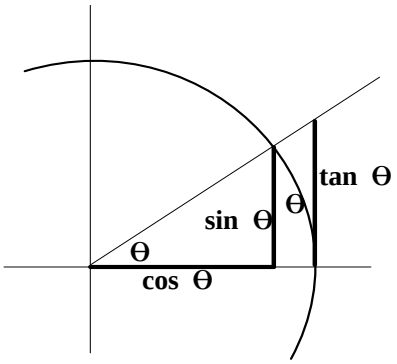
$$\begin{aligned}\frac{d}{dx}\sin x &= \lim_{h\to 0}\frac{\sin(x+h)-\sin x}{h}\\ &= \lim_{h\to 0}\frac{\sin x\cos h+\cos x\sin h-\sin x}{h}\\ &= \lim_{h\to 0}(\sin x)\frac{\cos h-1}{h}+\lim_{h\to 0}(\cos x)\frac{\sin h}{h}\\ &= (\sin x)\lim_{h\to 0}\frac{\cos h-1}{h}+(\cos x)\lim_{h\to 0}\frac{\sin h}{h}\end{aligned}$$

The computation will be complete when we evaluate the two important limits in the final line.

Theorem 1: $\lim_{\theta\to 0}\dfrac{\sin\theta}{\theta}=1$, and $\lim_{\theta\to 0}\dfrac{\cos\theta-1}{\theta}=0$.

Proof: The first of these limits is easily made convincing by calculating the value of $\sin\theta/\theta$ for some small values of θ.

θ	$\sin\theta/\theta$
10^{-1}	.99833416646828152307
10^{-2}	.99998333341666646825
10^{-3}	.99999983333334166667
10^{-4}	.99999999833333333417
10^{-5}	.99999999998333333333
10^{-6}	.99999999999983333333
10^{-7}	.99999999999999833333
10^{-8}	.99999999999999998333
10^{-9}	.99999999999999999983
10^{-10}	1.00000000000000000000



But no amount of numerical computation constitutes a *proof*. For that we refer to the definition of the trigonometric functions in terms of the unit circle. Refer to the figure, above, in which the angle θ, measured in radians, is the length of the arc subtended on the circle by the central angle θ. The values of $\sin\theta$, $\cos\theta$, and $\tan\theta$ are then the lengths of the labeled line segments. From the figure we see that $\sin\theta\le\theta\le\tan\theta$, or dividing the inequalities by $\sin\theta$, $1\le\theta/\sin\theta\le 1/\cos\theta$. Finally, taking reciprocals we have $1\ge\sin\theta/\theta\ge\cos\theta$. From this we conclude, finally, that $\lim_{\theta\to 0}(\sin\theta)/\theta=1$ since it is "sandwiched" between 1 and $\cos\theta$, both of which have the limit 1 as $\theta\to 0$. We should note that the argument just given, based on the figure, assumed that $\theta>0$. Thus, strictly speaking, we have just shown that $\lim_{\theta\to 0^+}(\sin\theta)/\theta=1$. The left-hand limit must, of course, be the same since $(\sin\theta)/\theta$ is an even function.

The second of the two limits in the theorem can be obtained from the first. Using known trigonometric identities we have

$$\begin{aligned}
\lim_{\theta\to 0}\frac{\cos\theta-1}{\theta} &= \lim_{\theta\to 0}\frac{\cos\theta-1}{\theta}\frac{\cos\theta+1}{\cos\theta+1}\\
&= \lim_{\theta\to 0}\frac{\cos^2\theta-1}{\theta(\cos\theta+1)} = \lim_{\theta\to 0}\frac{-\sin^2\theta}{\theta(\cos\theta+1)}\\
&= -\lim_{\theta\to 0}\sin\theta\cdot\frac{\sin\theta}{\theta}\cdot\frac{1}{\cos\theta+1}\\
&= -\lim_{\theta\to 0}\sin\theta\cdot\lim_{\theta\to 0}\frac{\sin\theta}{\theta}\cdot\lim_{\theta\to 0}\frac{1}{\cos\theta+1}\\
&= -0\cdot 1\cdot\frac{1}{2}=0
\end{aligned}$$

Example 1: Evaluate $\lim_{x\to 0}\frac{\sin 3x}{x}$. We rewrite the expression so as to recognize the limit of Theorem 1:

$$\begin{aligned}
\lim_{x\to 0}\frac{\sin 3x}{x} &= \lim_{x\to 0}3\cdot\frac{\sin 3x}{3x} =\\
&= 3\cdot\lim_{x\to 0}\frac{\sin 3x}{3x} = 3\cdot\lim_{u\to 0}\frac{\sin u}{u} = 3\cdot 1 = 3
\end{aligned}$$

Example 2: The function $f(x)=\sin x/x$ is defined and continuous at every point except $x=0$. However it has a continuous extension to $x=0$ since the $\lim_{x\to 0} f(x)=1$ exists:

$$F(x)=\begin{cases}\frac{\sin x}{x} & \text{if } x\neq 0\\ \\ 1 & \text{if } x=0\end{cases}$$

Applet: Limit of sin(x)/x as x approaches 0 **Try it!**

With the two limits of Theorem 1 evaluated, we now know that the derivative of $\sin x$ is $\sin x\cdot 0+\cos x\cdot 1$, which we record in the following theorem.

Theorem 2: $\frac{d}{dx}\sin x=\cos x$, and $\frac{d}{dx}\cos x=-\sin x$.

The first is proved. As for $\frac{d}{dx}\cos x$, we can use the trigonometric identity of complementary angles to deduce this as follows:

$$\frac{d}{dx}\cos x=\frac{d}{dx}\sin(\frac{\pi}{2}-x)=\cos(\frac{\pi}{2}-x)\cdot(-1)=-\sin x.$$

Example 3: Differentiate $\sin 2x$, $\sin(x^2+1/x)$, and $\cos(3x+\sqrt{x})$. Using the chain rule in these three examples we have

$$\frac{d}{dx}\sin 2x=\cos 2x\cdot 2=2\cos 2x$$

$$\frac{d}{dx}\sin(x^2+\frac{1}{x})=\cos(x^2+\frac{1}{x})\cdot(2x-\frac{1}{x^2})=(2x-\frac{1}{x^2})\cos(x^2+\frac{1}{x})$$

$$\frac{d}{dx}\cos(3x+\sqrt{x})=-\sin(3x+\sqrt{x})\cdot(3+\frac{1}{2\sqrt{x}})$$

Example 4: Differentiate $y=\sin x\cos x$ and $y=\sin^2(\cos(x^2+2))$. The first function is a product, thus $y'=(\sin x)(-\sin x)+(\cos x)(\cos x)=\cos^2 x-\sin^2 x$. The second function requires repeated uses of the

chain rule:

$$\begin{aligned}
\frac{dy}{dx} &= 2\sin(\cos(x^2+2))\cdot\frac{d}{dx}\sin(\cos(x^2+2)) \\
&= 2\sin(\cos(x^2+2))\cos(\cos(x^2+2))\cdot\frac{d}{dx}\cos(x^2+2) \\
&= 2\sin(\cos(x^2+2))\cos(\cos(x^2+2))(-1)\sin(x^2+2)\cdot\frac{d}{dx}(x^2+2) \\
&= 2\sin(\cos(x^2+2))\cos(\cos(x^2+2))(-1)\sin(x^2+2)\cdot(2x) \\
&= -4x\sin(\cos(x^2+2))\cos(\cos(x^2+2))\sin(x^2+2)
\end{aligned}$$

Differentiation rules for the remaining trigonometric functions are obtained from those of $\sin x$ and $\cos x$:

Theorem 3: The derivatives of $\tan x$, $\cot x$, $\sec x$, and $\csc x$ are:

$$\begin{aligned}
\frac{d}{dx}\tan x &= \sec^2 x \\
\frac{d}{dx}\cot x &= -\csc^2 x \\
\frac{d}{dx}\sec x &= \sec x\tan x \\
\frac{d}{dx}\csc x &= -\csc x\cot x
\end{aligned}$$

The proofs of these rules are by direct calculation, using the product, quotient and power rules:

$$\frac{d}{dx}\tan x = \frac{d}{dx}\frac{\sin x}{\cos x} = \frac{(\cos x)(\cos x)-(\sin x)(-\sin x)}{\cos^2 x} = \frac{1}{\cos^2 x} = \sec^2 x$$

$$\frac{d}{dx}\cot x = \frac{d}{dx}\frac{\cos x}{\sin x} = \frac{(\sin x)(-\sin x)-(\cos x)(\cos x)}{\sin^2 x} = \frac{-1}{\sin^2 x} = -\csc^2 x$$

$$\frac{d}{dx}\sec x = \frac{d}{dx}\frac{1}{\cos x} = -\frac{(-\sin x)}{\cos^2 x} = \frac{1}{\cos x}\cdot\frac{\sin x}{\cos x} = \sec x\tan x$$

$$\frac{d}{dx}\csc x = \frac{d}{dx}\frac{1}{\sin x} = -\frac{\cos x}{\sin^2 x} = -\frac{1}{\sin x}\cdot\frac{\cos x}{\sin x} = -\csc x\cot x$$

As always the most common applications of these rules is in combination with the chain rule and all the other differentiation rules.

Example 5: $\dfrac{d}{dx}\sec^2 x = 2\sec x\cdot\sec x\tan x = 2\sec^2 x\tan x.$

Example 6: Compute the derivative of $y = (\sin^3(\tan^2(2x)))^4$. This example requires six applications of the chain rule:

$$\begin{aligned}
y' &= \frac{d}{dx}(\sin^3(\tan^2(2x)))^4 \\
&= 4(\sin^3(\tan^2(2x)))^3 \frac{d}{dx}\sin^3(\tan^2(2x)) \\
&= 4(\sin^3(\tan^2(2x)))^3 \cdot 3\sin^2(\tan^2(2x))\frac{d}{dx}\sin(\tan^2(2x)) \\
&= 4(\sin^3(\tan^2(2x)))^3(3)\sin^2(\tan^2(2x)) \cdot \cos(\tan^2(2x))\frac{d}{dx}\tan^2(2x) \\
&= 4(\sin^3(\tan^2(2x)))^3(3)\sin^2(\tan^2(2x))\cos(\tan^2(2x)) \cdot 2\tan(2x)\frac{d}{dx}\tan(2x) \\
&= 4(\sin^3(\tan^2(2x)))^3(3)\sin^2(\tan^2(2x))\cos(\tan^2(2x))(2)\tan(2x) \cdot \sec^2(2x)\frac{d}{dx}(2x) \\
&= 4(\sin^3(\tan^2(2x)))^3(3)\sin^2(\tan^2(2x))\cos(\tan^2(2x))(2)\tan(2x)\sec^2(2x) \cdot 2 \\
&= 48(\sin^3(\tan^2(2x)))^3\sin^2(\tan^2(2x))\cos(\tan^2(2x))\tan(2x)\sec^2(2x)
\end{aligned}$$

Were you able to keep those six applications of the chain rule straight? It would be understandble if you couldn't. In complicated situations like this it is often helpful to explicitly write the chain of functions that go into the composite function:

$$y = u^4,\ u = v^3,\ v = \sin w,\ w = t^2,\ t = \tan s,\ s = 2x$$

Then, according to the chain rule,

$$\begin{aligned}
\frac{dy}{dx} &= \frac{dy}{du}\cdot\frac{du}{dv}\cdot\frac{dv}{dw}\cdot\frac{dw}{dt}\cdot\frac{dt}{ds}\cdot\frac{ds}{dx} \\
&= 4u^3 \cdot 3v^2 \cdot \cos w \cdot 2t \cdot \sec^2 s \cdot 2
\end{aligned}$$

and, if we now substitute the values of all the intermediate variables back into the final expression, we obtain the same result as before. In simpler examples we had encouraged "thinking the intermediate functions" rather than explicitly writing them down. But in complex cases such as the present one that may leave us befuddled.

Summary: In this section we introduced the basic differentiation formulas for each of the trigonometric functions. Only the derivative of $\sin x$ was computed directly from the limit definition. All the others then followed by using our general differentiation rules, and likewise we can handle through the general rules any more complicated expressions that involve trigonometric functions. It is now time to move on to important applications of differentiation and its important place in modeling real-world problems.

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

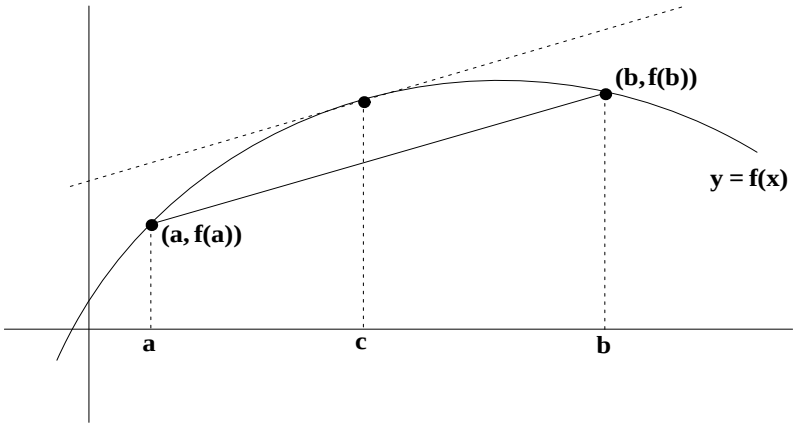
2.10 The Mean Value Theorem

The derivative of a function is a powerful tool for analyzing the function's behavior. If $f'(x_0)$ exists at a point x_0, for example, then we not only know that the function is continuous there but also that its graph has a tangent line at $(x_0, f(x_0))$. We have characterized this fact by saying that the graph is "smooth" at the point, i.e. it does not have a "corner". Another way of saying this is to say that the function can be approximated by the linear function $f(x_0) + f'(x_0)(x - x_0)$ in the immediate vicinity of the point. We sometimes call this kind of information about the function "local information" because it tells us how the function behaves at a single point.

We can also ask questions about a function's "global behavior". For example does the function *increase* throughout some interval, or does it *decrease*? Does its graph *rise* as we move from left to right on the x-axis, or does it *fall*? We also expect that the derivative will give us this kind of global information. It does—and the main tool for extending the derivative's influence from individual points to an entire interval is the *Mean Value Theorem*.

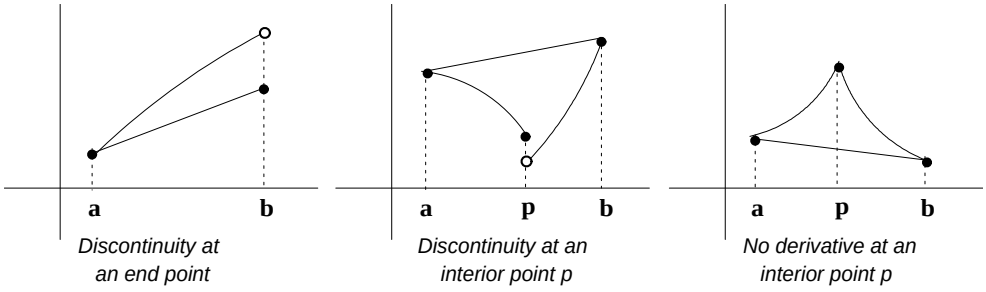
Theorem 1 (The Mean Value Theorem): Suppose that f is defined and continuous on a closed interval $[a, b]$, and suppose that f' exists on the open interval (a, b). Then there exists a point c in (a, b) such that

$$\frac{f(b) - f(a)}{b - a} = f'(c).$$



We will provide a proof presently, but first let us be sure to understand just what the theorem says and how it provides the "global" view that we promised. Taken at face value, especially in the presence of the figure, it seems rather obvious. It says, in effect, that there is a place in the interval (a, b) where the tangent line to the graph is parallel to the secant line connecting the end points $(a, f(a))$ and $(b, f(b))$. (Just imagine sliding the secant line upward until it becomes tangent to the graph!) Or, in the language of rates of change, if your *average speed* from the initial to the final toll booths on the New Jersey Turnpike was 75 miles per hour (remember that they stamp the times on the ticket), then there was at least one moment along the way when your instantaneous speed was *exactly* 75 miles per hour. If you try to convince the officer that you were always driving below the speed limit you will lose—the officer, having taken calculus in college, will recite to you the Mean Value Theorem.

Of course the hypotheses of the theorem are essential. If the function is not continuous (it has a jump somewhere) or not differentiable (its graph turns a sharp corner somewhere) then there need be no place where the secant and tangent lines are parallel. (Look at the following figures.)



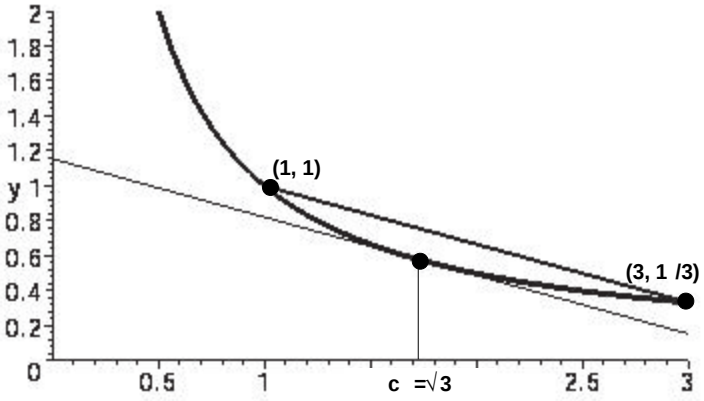
Applet: Mean Value Theorem **Try it!**

Example 1: Consider the function $f(x) = |x|$ on $[-1, 1]$. The Mean Value Theorem does not apply because the derivative is not defined at $x = 0$. Indeed $(|1| - |-1|)/(1 - (-1)) = 0$, and there is clearly no value of c for which $f'(c) = 0$.

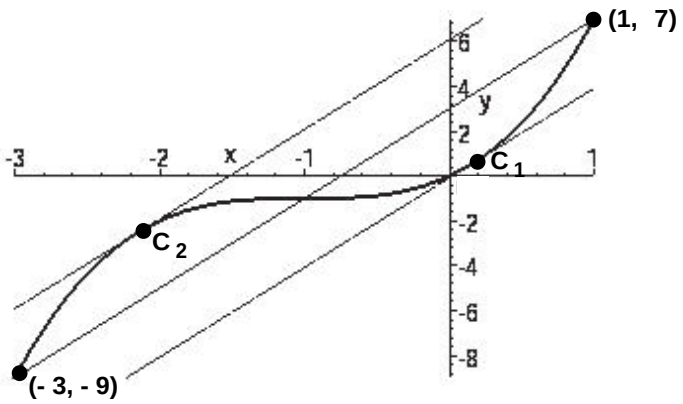
Example 2: Under what circumstances does the Mean Value Theorem apply to the function $f(x) = 1/x$? The only point to avoid is $x = 0$ where the function and its derivative are not defined. Thus we can apply the theorem on any interval $[a, b]$ that does not contain $x = 0$. It does not apply, for example, to the interval $-1 \le x \le 1$. It does apply to to the interval $[1, 3]$. Thus there is a point c in the open interval $(1, 3)$ such that $f'(c) = -1/c^2$ is equal to

$$\frac{f(b) - f(a)}{b - a} = \frac{\frac{1}{3} - \frac{1}{1}}{3 - 1} = -\frac{1}{3}.$$

Solving $-1/c^2 = -1/3$ we have $c = \sqrt{3}$.



Example 3: Verify the conclusion of the Mean Value Theorem for the function $f(x) = (x + 1)^3 - 1$ on the interval $[-3, 1]$. First we note that f is continuous on the closed interval $[-3, 1]$ and its derivative $f'(x) = 3(x + 1)^2$ is defined in the open interval $(-3, 1)$, thus the Mean Value Theorem applies. The slope of the secant line is $(f(1) - f(-3))/(1 - (-3)) = 4$, thus there is at least one point c in the interval where $f'(c) = 4$. Solving $3(c + 1)^2 = 4$ we in fact find that there are two such points $c = -1 + (2/3)\sqrt{3}$ and $c = -1 - (2/3)\sqrt{3}$. The two tangent lines that are parallel to the secant line are plotted in the figure.



Let us suppose, now, that a function f is defined on an interval I and that $f'(x) > 0$ for every $x \in I$. This means that the slope of the graph of f is positive everywhere in the interval I. Then the function is *increasing* throughout the interval; i.e. its graph is *rising*. Let us formalize the language we are using here and then prove the statement.

First of all, let us remember that an interval I is the set of real numbers lying between a and b, where a and b are real numbers or $\pm\infty$. The end points a and b may or may not be included in the interval I (of course $\pm\infty$ cannot be included since they are not real numbers). An interval can be *closed*, e.g. $[-5, 5]$. It can be *open*, e.g. $(2, 10)$, $(0, \infty)$, or $(-\infty, \infty)$. Or it can be "half open", e.g. $(3, 5]$, $[0, \infty)$, or $(-\infty, 2]$.

Definition 1: Suppose that f is defined on an interval I, and let x_1 and x_2 denote points in I:

1. f is *increasing on* I if $f(x_1) < f(x_2)$ whenever $x_1 < x_2$
2. f is *decreasing on* I if $f(x_1) > f(x_2)$ whenever $x_1 < x_2$
3. f is *nondecreasing on* I if $f(x_1) \leq f(x_2)$ whenever $x_1 < x_2$
4. f is *nonincreasing on* I if $f(x_1) \geq f(x_2)$ whenever $x_1 < x_2$

Of course "increasing" and "nondecreasing" are very closely related. If a function f is *increasing* its graph is rising (from left to right) whereas if it is *nondecreasing* its graph is generally rising but may have "level plateaus".

Theorem 2: Let I be an interval and let J be the open interval consisting of I minus its endpoints (if any). Suppose that f is continuous on I and differentiable on J. Then

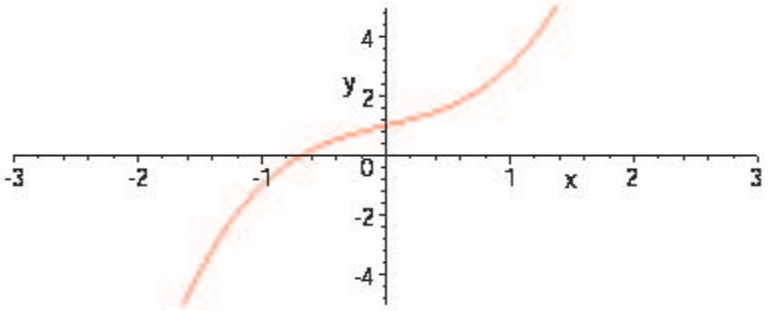
1. If $f'(x) > 0$ for every $x \in J$, then f is *increasing* on I.
2. If $f'(x) < 0$ for every $x \in J$, then f is *decreasing* on I.
3. If $f'(x) \geq 0$ for every $x \in J$, then f is *nondecreasing* on I.
4. If $f'(x) \leq 0$ for every $x \in J$, then f is *nonincreasing* on I.

For a proof we notice that if x_1 and x_2 are any two points in I, $x_1 < x_2$, then the conditions of the Mean Value Theorem are met for the interval $[x_1, x_2]$. Thus there is a point c between x_1 and x_2 such that

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} = f'(c).$$

Then $f(x_2) - f(x_1) = (x_2 - x_1)f'(c)$ and, since $x_2 - x_1$ is positive, the sign of $f(x_2) - f(x_1)$ is the same as that of $f'(c)$. All four statements of the theorem follow immediately.

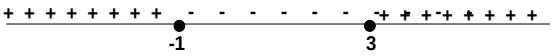
Example 4: On what interval is the function $f(x) = x^3 + x + 1$ increasing (decreasing)? We need to examine the derivative $f'(x) = 3x^2 + 1$. We notice that, in fact, the derivative is positive for all values of x, hence f is increasing on $(-\infty, \infty)$.



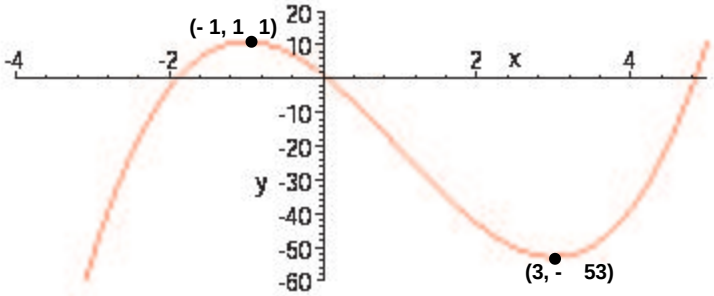
Example 5: Find the intervals on which the function $f(x) = 2x^3 - 6x^2 - 18x + 1$ is increasing and those on which it is decreasing. We examine the derivative

$$f'(x) = 6x^2 - 12x - 18 = 6(x-3)(x+1).$$

It is easy to see that the derivative changes sign at the points $x = 3,\ -1$. In the intervals $(-\infty, -1)$, $(-1, 3)$, and $(3, \infty)$, we record its sign on the following "sign graph":



The function, then, is increasing on $(-\infty, -1)$ and $(3, \infty)$, and it is decreasing on $(-1, 3)$. Computing the values of f at just a few strategic points enables us to sketch a rather accurate graph. The points $(-1, 11)$ and $(3, -53)$ where the derivative changes sign are most useful. The graph "rises" to the first of these points, "falls" between the two points, and "rises" again to the right of the second point. Note these behaviors in the graph below:



Let us return, finally, to a proof of the Mean Value Theorem. Like the Intermediate Value Theorem that we discussed earlier, it depends very much on the deep underlying properties of the real number system, namely the *continuity* of the real line. As such, a correct proof eluded many people in the past. Even Gauss, one of the brightest mathematicians in the eighteenth century, gave a proof of this theorem that subsequently turned out to be deficient. Such historical events make us cautious today, even in the presence of a theorem with such strong intuitive geometric appeal as the Mean Value Theorem.

A correct proof depends, as we noted above, on continuity properties of the real numbers. One such property is that a continuous function takes on a maximum and a minimum value on a closed interval. We state this very fundamental fact as a theorem, and we do not give a proof since it is buried so deeply in the foundations of our number system.

Theorem 2: If f is continuous on a closed interval $[a, b]$, then there is a point c_1 in the interval where f assumes its *maximum* value, i.e. $f(x) \leq f(c_1)$ for every x in $[a, b]$, and a point c_2 where f assumes its *minimum* value, i.e. $f(x) \geq f(c_2)$ for every x in $[a, b]$.

Finding the point (or points) where f assumes its maximum and minimum values is an important application of calculus. It sometimes goes under the name "Optimization Theory". The following theorem helps us solve the *maximum-minimum* problem:

Theorem 3: If f is defined in an open interval (a, b) and achieves a maximum (or minimum) value at a point $c \in [a, b]$ where $f'(c)$ exists, then $f'(c) = 0$.

Let us prove this in the case of a maximum value. The proof for a minimum value is similar. If $f(c)$ is a

maximum value and $f'(c)$ exists, then $f(x) \le f(c)$ for all x in (a, b). Then, if $a < x < c$

$$\frac{f(x) - f(c)}{x - c} \ge 0, \qquad \text{so} \quad f'(c) = \lim_{x \to c^-} \frac{f(x) - f(c)}{x - c} \ge 0,$$

and, if $c < x < b$

$$\frac{f(x) - f(c)}{x - c} \le 0, \qquad \text{so} \quad f'(c) = \lim_{x \to c^+} \frac{f(x) - f(c)}{x - c} \le 0.$$

It thus follows that $f'(c) = 0$.

It is Theorem 3 that helps us find where a continuous function assumes its maximum and minimum values in a closed interval. The only possible locations for the maximum and minimum values are the so-called *critical points* where the derivative exists and is zero or where the derivative does not exist, or points that are not interior points of the interval (i.e. end points). In a typical application this narrows the search to a finite number of points that we can calculate.

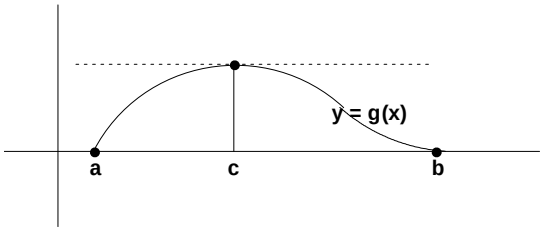
Example 6: Returning to the function $f(x) = 2x^3 - 6x^2 - 18x + 1$ of Example 5, let us find the points in the interval $[-4, 4]$ where the function assumes its maximum and minimum values. We know that the maximum and minimum values must occur at a critical point or an end point of the interval. The derivative is defined at every point of the interval, thus the only critical points are where $f'(x) = 0$. These are the points $x = -1, 3$. Let us make a small table of the function values at these two points and at the end points of the interval:

x	$f(x)$
-1	11
3	-53
-4	-151
4	-39

These are the only candidates for the maximum or minimum value of $f(x)$ on the interval $[-4, 4]$. Thus the maximum value is 11, occurring at the point $(-1, 11)$. And the minimum value is -151, occurring at the point $(-4, -151)$, one of the end points of the interval. The point $(3, -53)$ is of geometrical interest. The function has a *relative minimum* at this point, although the *absolute* minimum occurs elsewhere, at the left end point of the interval. We will return to these notions again in the context of curve sketching and other applications of the derivative.

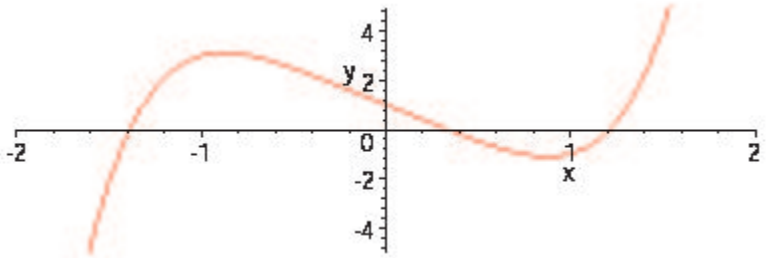
A special case of the Mean Value Theorem is *Rolle's Theorem*. We prove it first and then use it to prove the Mean Value Theorem.

Rolle's Theorem Suppose that the function g is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b). If $g(a) = 0$ and $g(b) = 0$ then there exists a point c in the open interval (a, b) where $g'(c) = 0$.



The proof follows easily from Theorem 3. Assuming that the function g is not identically zero on the interval it must assume either its maximum or minimum value at an interior point c. At such a point $g'(c) = 0$.

Example 7: Use Rolle's Theorem to show that the equation $x^5 - 3x + 1 = 0$ has exactly three real roots. Letting $f(x) = x^5 - 3x + 1$, we have $f'(x) = 5x^4 - 3$. The derivative f' is defined for all real values of x, hence the only critical points of f are where $f'(x) = 0$. There are only two such points, $x = \pm\sqrt[4]{3/5}$. We conclude, therefore, that there can be at most three real roots of the given equation since by Rolle's Theorem there must be a zero of the derivative between every pair of roots. That there are at least three real roots may be deduced from the fact that $f(x)$ is continuous and changes sign at least three times (note the graph below).



Proof of the Mean Value Theorem Let f satisfy the requirements of the mean value theorem. At a point x in the interval $[a, b]$ let $g(x)$ be the vertical distance between the graph of f and the chord connecting the points $[a, f(a)]$ and $[b, f(b)]$ (see the graph accompanying the statement of the Mean Value Theorem). Then the function $g(x)$ is defined by

$$g(x) = f(x) - \left(f(b) + \frac{f(b) - f(a)}{b - a}(x - a) \right),$$

and we notice that the function g satisfies the requirements of Rolle's Theorem, i.e. g is continuous in $[a, b]$, differentiable in (a, b), and $g(a) = g(b) = 0$. Thus there is a point c in the open interval (a, b) where $g'(c) = 0$. Computing $g'(x)$ we have

$$g'(x) = f'(x) - \frac{f(b) - f(a)}{b - a}.$$

Substituting $x = c$ we have $f'(c) - (f(b) - f(a))/(b - a) = 0$, and this is the conclusion of the Mean Value Theorem.

Summary: We can learn much about a function f by studying its derivative. For example we can determine where f is *increasing* and where it is *decreasing*. And we can discover where it assumes its *maximum* and *minimum* values. The key to the relationship between such global properties of f and the behavior of its derivative f' is the Mean Value Theorem. It will often arise in similar circumstances—when we need to connect *local* and *global* behavior of a function. We will turn shortly to the problem of finding "anti derivatives" of functions and to the surprising connection of them to the second major problem of calculus—defining and finding the area of geometrical regions determined by graphs of functions. The Mean Value Theorem will again turn out to play a star role in those new investigations.

Exercises: Problems **Check what you have learned!**

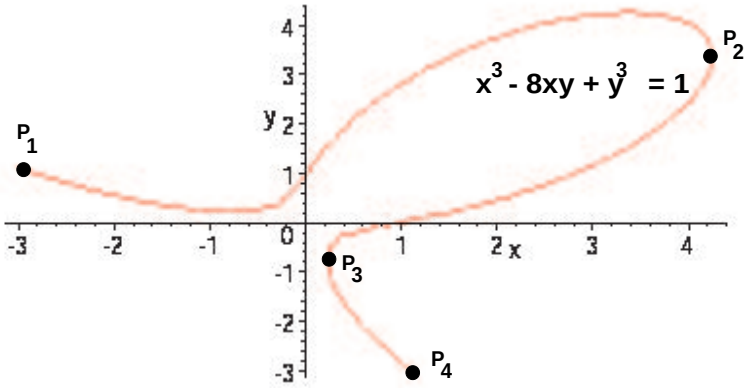
Videos: Tutorial Solutions **See problems worked out!**

2.11 Implicit Differentiation

Derivatives are a powerful tool for studying curves in the xy-plane. For any curve that is the graph of a function $y = f(x)$ the derivative y' gives us information about slope, maxima and minima, and general tendencies such as *increasing* or *decreasing*.

Many curves, however, are not the graphs of functions. A circle of radius 1, for example, does not pass the "vertical line test" and hence is not the graph of a function. It is, however, the graph of the *equation* $x^2 + y^2 = 1$. And we have seen that this equation defines *two* functions $y = \sqrt{1-x^2}$ and $y = -\sqrt{1-x^2}$ whose graphs are the upper and lower semi-circles that make up the circle. The circle can thus be broken into two pieces each of which can be studied using derivatives. In this example we were able to solve the given equation for y. But we will not always be so lucky.

The equation $x^3 - 8xy + y^3 = 1$, for example, resists our most clever efforts to explicitly solve for y as a function of x. Indeed, as we can see from the graph of the equation plotted below, the curve can be broken into three pieces each of which passes the vertical line test (the piece from P_1 to P_2, the piece from P_2 to P_3, and the piece from P_3 to P_4). It would seem, therefore, that the given equation determines *implicitly* three functions of x; let us call them $y = f_1(x)$, $y = f_2(x)$, and $y = f_3(x)$. We do not have *explicit* formulas for these three functions, however, and thus our techniques developed so far are of little help. We will see how to overcome this difficulty using a very important technique called *implicit differentiation*.



The general setting for our discussion of implicitly defined functions is an equation $F(x, y) = 0$, where F is an expression containing the two variables x and y. The graph of the equation is, in general, a curve, but as we have seen it is not usually the graph of a function since it does not satisfy the vertical line test. A function $f(x)$ is said to be *implicitly defined* by the equation if $F(x, f(x)) = 0$ on some interval I. Our wish is to find the derivative of $f(x)$ without explicitly solving the equation (which we normally will not be able to do).

Example 1: The functions $\sqrt{1-x^2}$ and $-\sqrt{1-x^2}$ are implicitly defined by the equation $x^2 + y^2 = 1$. For example, substituting the first function into the equation we have

$$x^2 + \left(\sqrt{1-x^2}\right)^2 = x^2 + (1-x^2) = 1.$$

Thus the function is a "solution" of the equation.

The chain rule is the key to finding the derivative of an implicitly defined function. We will illustrate this in examples:

Example 2: Consider one of the functions $f(x)$ defined implicitly by the equation $x^2 + y^2 = 1$. Then $f(x)$ satisfies the equation $x^2 + (f(x))^2 = 1$. Differentiating the equation, and using the chain rule, we have

$$2x + 2f(x)f'(x) = 0,$$

and solving the resulting equation for $f'(x)$ we have

$$f'(x) = -\frac{x}{f(x)}.$$

A formula for the derivative has thus been obtained, albeit in terms of the function $f(x)$ itself. This is natural in light of the fact that there is more than one function defined implicitly by the equation $x^2 + y^2 = 1$. If $f(x) = \sqrt{1-x^2}$ our formula for the derivative gives $f'(x) = -x/f(x) = -x/\sqrt{1-x^2}$. And if $f(x) = -\sqrt{1-x^2}$ the formula gives $f'(x) = -x/f(x) = -x/\left(-\sqrt{1-x^2}\right) = x/\sqrt{1-x^2}$. Notice that these results agree with our previous methods for computing these derivatives.

Example 3: We repeat Example 2, simplifying our notation. Given the equation $x^2 + y^2 = 1$, we *think* of the functions $y = f(x)$ implicitly defined by the equation. Since y is a function of x, we can differentiate the equation, using the chain rule, obtaining

$$2x + 2y\frac{dy}{dx} = 0.$$

Solving for $\frac{dy}{dx}$ we then obtain

$$\frac{dy}{dx} = -\frac{x}{y}.$$

We have thus again obtained a formula for $\frac{dy}{dx}$ in terms of the function y itself. And, of course, if we were lucky enough to know a formula for y in terms of x we could substitute it for y at this point. But even in its present form, without eliminating y, we can use the formula for computing the slope of the graph or for other applications. For example the slope of the graph (in this case the circle) at the point $(0,1)$ is $\left.\frac{dy}{dx}\right|_{(0,1)} = -\frac{0}{1} = 0$. And the slope at the point $(-\frac{1}{2}, -\frac{\sqrt{3}}{2})$ is

$$\left.\frac{dy}{dx}\right|_{(-\frac{1}{2},-\frac{\sqrt{3}}{2})} = -\left(\frac{-\frac{1}{2}}{-\frac{\sqrt{3}}{2}}\right) = -\frac{1}{\sqrt{3}}.$$

Example 4: Use implicit differentiation to find the equation of the tangent line to the graph of $xy^2 + x^2y - 6 = 0$ at the point $(1,2)$. (Note, incidently, that this point does indeed lie on the graph .) The equation of the tangent line is $y = 2 + f'(1)(x-1)$, thus we need only find the slope at the point $(1,2)$. Differentiating the equation implicitly, using the product and chain rules, we obtain

$$1 \cdot y^2 + x \cdot 2y\frac{dy}{dx} + 2x \cdot y + x^2 \cdot \frac{dy}{dx} - 0 = 0,$$

and solving for $\frac{dy}{dx}$ we have

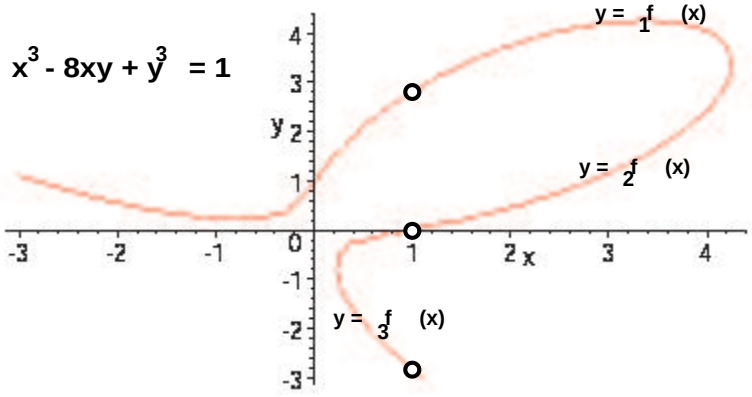
$$\frac{dy}{dx} = \frac{-y^2 - 2xy}{2xy + x^2} = -\frac{y^2 + 2xy}{x^2 + 2xy}.$$

At the point $(1,2)$ the value of the derivative is

$$\left.\frac{dy}{dx}\right|_{(1,2)} = -\frac{4+4}{1+4} = -\frac{8}{5},$$

and hence the tangent line has the equation $y = 2 - (8/5)(x-1)$.

Example 5: Return to the equation $x^3 - 8xy + y^3 = 1$ with which we begin this section.



Although we cannot explicitly solve this equation for y in terms of x, we can find the derivative (of any one of the three functions defined implicitly by the equation). Differentiating implicitly

$$3x^2 - 8y - 8xy' + 3y^2y' = 0,$$

and, solving for y', we have $y' = (8y - 3x^2)/(3y^2 - 8x)$. We can use this formula to find the slope at any point on the curve. Let us try, for example, to find the points on the curve for which $x = 1$. For this we solve the equation $1 - 8y + y^3 = 1$, i.e. $y^3 - 8y = 0$. This equation has three roots, $y = 0$, $\sqrt{8}$, $-\sqrt{8}$. Thus the three points $(1, 0)$, $(1, \sqrt{8})$ and $(1, -\sqrt{8})$ are on the curve. The slope at $(1, 0)$ is

$$y'|_{(1,0)} = \frac{8 \cdot 0 - 3 \cdot 1^2}{3 \cdot 0^2 - 8 \cdot 1} = \frac{3}{8}.$$

Similarly, at $(1, \sqrt{8})$ the slope is

$$y'|_{(1,\sqrt{8})} = \frac{8 \cdot \sqrt{8} - 3 \cdot 1^2}{3 \cdot \left(\sqrt{8}\right)^2 - 8 \cdot 1} = \frac{8\sqrt{8} - 3}{16} \approx 1.22671$$

And, finally, the slope at the point $(1, -\sqrt{8})$ is

$$y'|_{(1,-\sqrt{8})} = \frac{8 \cdot (-\sqrt{8}) - 3 \cdot 1^2}{3 \cdot \left(-\sqrt{8}\right)^2 - 8 \cdot 1} = \frac{-8\sqrt{8} - 3}{16} \approx -1.60171$$

We have, in fact, computed the derivatives $f_1(1)$, $f_2(1)$ and $f_3(1)$. The functions f_1 and f_2 are *increasing* in the vicinity of $x = 1$, and the function f_3 is decreasing. And we learned this in spite of the fact that we have no explicit formulas for the functions. Such is the value of implicit differentiation!

Example 7: Suppose a differentiable function f has an inverse f^{-1}. Find the derivative of f^{-1}. To solve this problem, let $y = f^{-1}(x)$. Then because f and f^{-1} are inverse functions, $f(y) = x$. Hence, differentiating implicitly we have: $f'(y)\frac{dy}{dx} = 1$. Now, solving for $\frac{dy}{dx}$ we get:

$$\frac{dy}{dx} = \frac{1}{f'(y)}$$

$$\left[f^{-1}(x)\right]' = \frac{1}{f'(f^{-1}(x))}$$

The above two formulas express the derivative of f^{-1} in terms of the derivative of f.

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

2.12 Derivatives of Exponential and Logarithm Functions

We began our study of calculus with a review of polynomials and trigonometric functions. Now, we have added the exponential and logarithmic functions to the list of functions that we are calling *elementary functions*. Their properties also should be familiar to us now. The elementary functions are all essential because they are very important in modeling real-world situations.

For purposes of modeling, in addition to the functions themselves, we also need to be familiar with their derivatives. As we have seen, it is relatively straightforward to develop a formula for the derivative of a polynomial using algebra. The derivative of the sine was a bit more complicated because we had to analyze and compute a particular limit. Once done, the derivatives of the other trig functions followed rather readily. The case of the exponential function is similar. Once we have computed a particular limit, the derivatives of the exponential and logarithmic functions will follow in a straightforward mannner by calculating the limits of the difference quotients directly from the definition of the derivative.

The Derivative of $y = e^x$: Let $y = e^x$. Then

$$\begin{aligned}\frac{dy}{dx} &= \lim_{h\to 0}\frac{e^{x+h}-e^x}{h}\\ &= \lim_{h\to 0}\frac{e^x(e^h-1)}{h}\\ &= e^x\left(\lim_{h\to 0}\frac{e^h-1}{h}\right)\end{aligned}$$

Therefore, the derivative will exist if the limit $\lim_{h\to 0}\frac{e^h-1}{h}$ exists. But we have seen this limit before in two different contexts. In Section 2.1.4 we interpreted it to be the limit of the difference quotients of $y = e^x$ at the point $x = 0$. We now know that in fact this is precisely the definition of the derivative of $y = e^x$ at $x = 0$, which furthermore is the slope of the tangent line to the graph of the function there. From a table of values, we estimated this limit to be equal to 1. We now can return to the earlier context in Section 1.5 in which we defined the number e to be the base of that exponential function whose tangent line to its graph at the point $(0, 1)$ is of slope 1. When we did that, we were not able to make precise the notion of the tangent line at a point on an arbitrary curve. The derivative allows us to overcome that obstacle. The value of the limit $\lim_{h\to 0}\frac{e^h-1}{h}$ is the derivative of the exponential function $y = e^x$ at $x = 0$, which equals the slope of the tangent line at $(0, 1)$ of the graph, or 1. Thus, we now can complete the calculation of the derivative of $y = e^x$ for any x:

$$\frac{dy}{dx} = e^x\left(\lim_{h\to 0}\frac{e^h-1}{h}\right) = e^x\cdot 1 = e^x$$

Applet: Limits of Functions **Try it!**

We now give as a theorem the chain rule form of this result.

Theorem 1: Let u be a function of x. Then

$$\tfrac{d}{dx}e^u = e^u\tfrac{du}{dx}$$

In words, the theorem tells us that the derivative of an exponential function is produced by multiplying the function by the derivative of the exponent.

Example 1: If $y = e^{17x}$, then using the theorem we get $\frac{dy}{dx} = 17e^{17x}$.

Example 2: Let $y = e^{\sin x}$. Then $y' = (\cos x)e^{\sin x}$.

The Derivative of $y = \ln x$: Now that we know the derivative of the exponential function, we can find

the derivative of its inverse $y = \ln x$ by implicit differentiation:

$$\begin{aligned} y = \ln x &\Leftrightarrow e^y = x \\ e^y \cdot \frac{dy}{dx} &= 1 \\ \frac{dy}{dx} &= \frac{1}{e^y} \\ &= \frac{1}{x} \end{aligned}$$

where in the last equality we use the fact that $e^y = x$.

So, we have found a function whose derivative is $1/x$, namely, $\ln x$. But we have to be somewhat careful here. The natural log is only defined for positive values of x. Suppose x is negative. Then what?

If $x < 0$, then $|x| = -x$ and $\frac{d}{dx}\ln(-x) = \frac{1}{-x} \cdot -1$ by the chain rule. Thus, $\frac{d}{dx}\ln(-x) = \frac{1}{x}$. Hence, we have shown that the following theorem is true.

Theorem 2: Let u be a function of x. Then

$$\frac{d}{dx}\ln|u| = \frac{1}{u}\frac{du}{dx}$$

Example 3: By the chain rule, the derivative of $y = \ln 7x$ is $y' = \frac{1}{7x} \cdot 7 = \frac{1}{x}$.

Example 4: If $y = \ln x^2$, then $y' = \frac{1}{x^2}(2x) = \frac{2}{x}$. We get the same answer if we use a property of logarithms to simplify first: $y = \ln x^2 = 2\ln x$; so, $y' = 2\frac{1}{x}$.

The Calculus Standards: e^x and $\ln x$

In calculus we rely on $y = e^x$ as the standard exponential function, and on $y = \ln x$ as the standard logarithm. They have the simplest differentiation formulas. Moreover, as we have seen, we can obtain any other general exponential or logarithm as follows, $a > 0$:

$$a^x = e^{x \ln a}$$

$$\log_a x = \frac{\ln x}{\ln a}$$

Example 5: Let $y = 2^x$. Then to find its derivative we rewrite the function as $y = e^{x \ln 2}$ and use the chain rule: $y' = e^{x \ln 2} \ln 2 = \ln 2\ 2^x$. Note that $\ln 2$ is just a constant.

Example 6: Even though $y = x^x$ does not have a constant base, we proceed to rewrite it as $y = x^x = e^{\ln x^x} = e^{x\ \ln x}$. Then, using the chain rule, we get $y' = e^{x\ \ln x}\ \left(\ln x + x\ \frac{1}{x}\right)$. Hence, $y' = x^x\,(\ln x + 1)$.

The Equation $y' = ky$

Suppose y is a function of x and satisfies the equation $y' = ky$ where k is a constant. Hence, the equation tells us that y is a constant multiple of of its own derivative. If $k = 1$, then $y = e^x$ has this property and thus solves the equation. In fact, the chain rule leads us to conclude that $y = e^{kx}$ solves the equation for any k. That is, the derivative of $y = e^{kx}$ is $ke^{kx} = ky$.

The equation $y' = ky$, because it includes a derivative, is an example of a *differential equation*. Differential equations are very important in calculus especially because of their centrality in modeling real-world phenomena. In fact, this particular differential equation is very important in the study of many populations, a topic that we will take up later.

The method we have used to solve the equation is called *guess-and-check*: we guessed a solution and then showed that it satisfied the equation. We have not shown that this is the *only* solution. Can you think of others? We will return to the question of uniqueness of solutions when we continue the study of differential equations shortly.

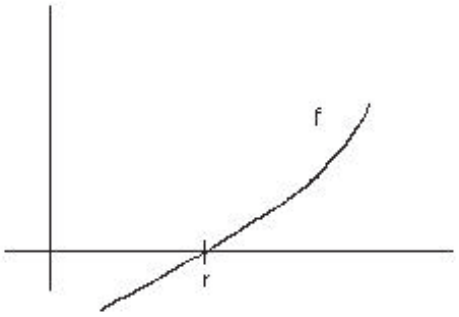
Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

2.13 Newton's Method

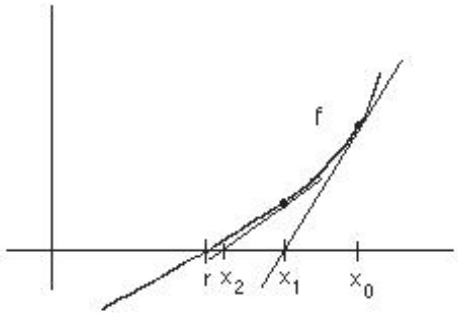
The tangent line at a point on the graph of a differentiable function can give a great deal of information about the function. It is this fact that makes the concept of the derivative so powerful, and has maintained its importance since its introduction in the seventeenth century.

For example, suppose we want to find a *root* of the equation $f(x) = 0$; that is, a number r such that $f(r) = 0$. The number r is called a *zero* of f.



We will describe a procedure, called *Newton's Method*, to find the root using tangent lines. The method is very beautiful in that it is easy to explain, and works very well in many circumstances. The idea is to use points where certain tangent lines intersect the x-axis to get close to the root.

To be specific, assume that f is differentiable. Choose a starting value x_0 near r on the x-axis. Then the tangent line at $(x_0, f(x_0))$ in many cases will intersect the x-axis in a point x_1 closer to r. Next, we repeat what we just did. That is, we draw a new tangent line at $(x_1, f(x_1))$ and hope that the point x_2 where it intersects the x-axis will be even closer to r. It most often is. Thus, we continue to repeat, defining a sequence $x_0, x_1, x_2, x_3, \ldots, x_n, \ldots$ such that $x_n \to r$.



To implement the procedure, we need an expression for x_n. Note that an equation of the tangent line at $(x_0, f(x_0))$ is $y = f(x_0) + f'(x_0)(x - x_0)$. So, with $y = 0$, we find x_1 by solving the equation:

$$\begin{aligned} 0 &= f(x_0) + f'(x_0)(x_1 - x_0) \\ f'(x_0)x_1 &= -f(x_0) + x_0 f'(x_0) \\ x_1 &= x_0 - \frac{f(x_0)}{f'(x_0)} \end{aligned}$$

Thus, we have found a formula for x_1 in terms of x_0 and functions of x_0.

But there is nothing special here about x_0 and x_1. Given x_n, we can determine x_{n+1} in a similar way. That is,

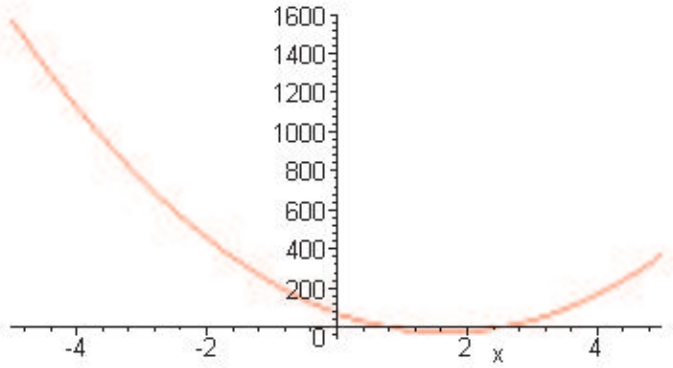
$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad n = 0, 1, 2, \ldots$$

The above procedure with this formula is known as *Newton's Method*.

Example 1: Use Newton's Method to find the zeros of $f(x) = 12(3x^2 - 10x + 6)$. We calculate the derivative: $f'(x) = 12(6x - 10)$. Thus for $n \geq 0$,

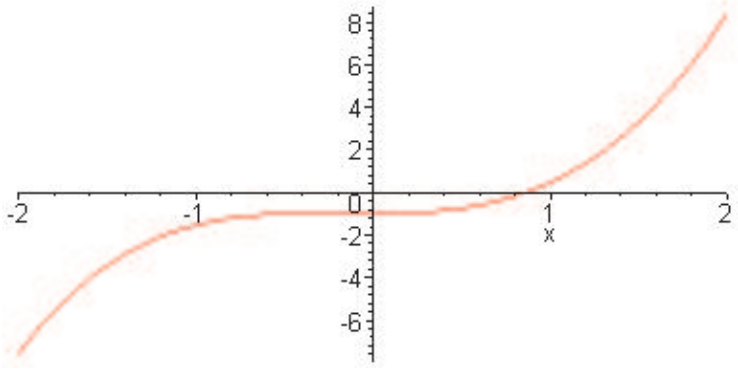
$$\begin{aligned} x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \\ &= x_n - \frac{12(3x_n^2 - 10x_n + 6)}{12(6x_n - 10)} \end{aligned}$$

Below is a sketch of the graph.



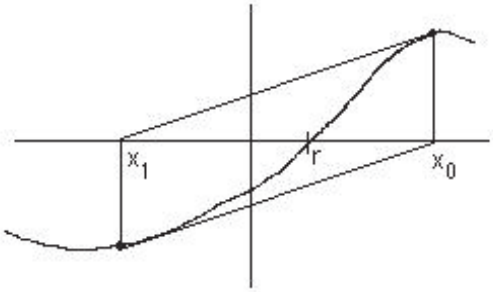
The equation has two roots. If we let $x_0 = 3$, then using a computer or programmable calculator we get approximately $x_1 = 2.625000000$, $x_2 = 2.551630435$, $x_3 = 2.548589014$, $x_4 = 2.548583771$. In fact, $x_{10} = 2.548583770$. If, on the other hand, we let $x_0 = 0.5$, then Newton's method will give an approximation to the other root: $x_1 = .7500000000$, $x_2 = .7840909091$, $x_3 = .7847493172$, $x_4 = .7847495630$; $x_{10} = .7847495630$.

Example 2: To solve the equation $x^3 = \cos x$, we let $f(x) = x^3 - \cos x$. Then $f'(x) = 3x^2 + \sin x$. Here is a sketch of the function that we can use to define a starting value x_0.



If we let $x_0 = 0.5$, we get approximately $x_1 = 1.112141637$, $x_2 = .9096726937$, $x_3 = .8672638182$, $x_4 = .8654771353$, $x_5 = .8654740331$. Also, $x_{10} = .8654740331$.

Newton's Method can fail as seen, for example, in the following sketch. However, it is a surprisingly powerful technique for the simplicity of the idea.



Applet: Newtons Method **Try it!**
Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

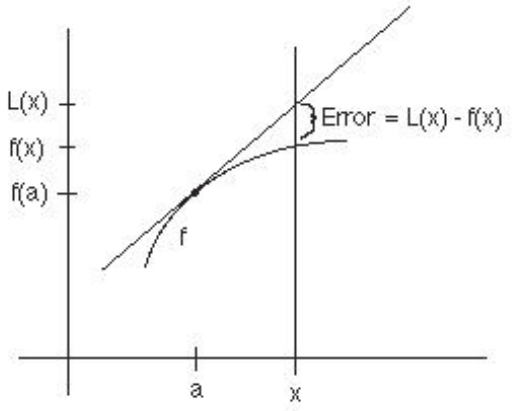
2.14 Linear Approximations

One of the most important ideas associated with the tangent line at a point on the graph of a function is that it provides a local linearization of the function. That is, no matter how complicated the graph of a differentiable function, no matter the difficulty of computing function values, near a point where the derivative exists, we can approximate the function by the tangent line. For this reason we have the following definition.

Definition: Let f be a differentiable function. Then the linearization (or linear approximation) of f about $x = a$ is the function $L(x)$ defined by

$$L(x) = f(a) + f'(a)(x - a)$$

Note that L is the linear function whose graph is the tangent line at $(a, f(a))$. In particular, instead of values on the graph of f near $(a, f(a))$, we use values on the tangent line, which may be relatively more straightforward to compute. Of course, there usually will be some error because the tangent line is different from the graph of the function. However, $\lim_{x\to a}(L(x) - f(x)) = f(a) - f(a) = 0$ implies that that $L(x) \approx f(x)$ for values of x near a.



Example 1: Find the linearization of $f(x) = \sqrt{x}$ about $x = 9$. Use it to approximate $\sqrt{8}$. In this example, $f'(x) = \frac{1}{2\sqrt{x}}$. Thus,

$$\begin{aligned} L(x) &= \sqrt{9} + \frac{1}{2\sqrt{9}}(x - 9) \\ &= 3 + \frac{1}{6}(x - 9) \\ L(8) &= 3 + \frac{1}{6}(8 - 9) \\ &= \frac{17}{6} \end{aligned}$$

Example 2: Use an appropriate linerization to find an approximate value of $\cos(36\,\text{deg})$. We change to radians: $36\,\text{deg} = (36/180)\pi = \pi/5$ radians. Thus, to approximate $\cos(\pi/5)$ we let $f(x) = \cos x$ and $a = \pi/6$. Hence,

$$\begin{aligned} L(x) &= \cos\frac{\pi}{6} + \left(-\sin\frac{\pi}{6}\right)\left(x - \frac{\pi}{6}\right) \\ &= \frac{\sqrt{3}}{2} - \frac{1}{2}\left(x - \frac{\pi}{6}\right) \\ L(\frac{\pi}{5}) &= \frac{\sqrt{3}}{2} - \frac{1}{2}\left(\frac{\pi}{5} - \frac{\pi}{6}\right) \end{aligned}$$

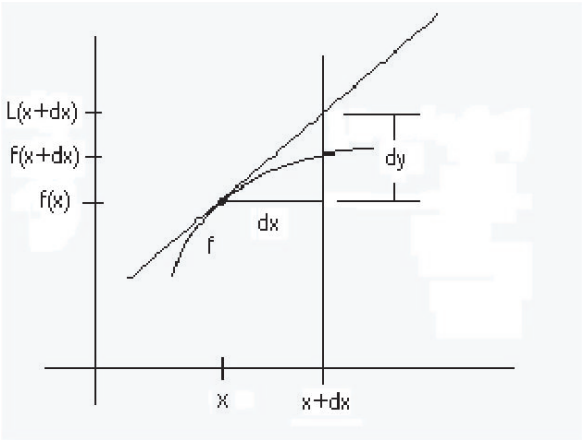
The concept of linearization is especially important because it is this notion of the derivative as the *best linear approximation* that generalizes to functions of more than one variable. The linearization is *best* in the sense that not only does $\lim_{x\to a}(L(x) - f(x)) = 0$, but also

$$\lim_{x\to a} \frac{L(x) - f(x)}{x - a} = 0$$

That is, $L(x) \to f(x)$ faster than $x \to a$. We will close by showing that the above limit is indeed zero:

$$\begin{aligned}
\lim_{x\to a} \frac{L(x) - f(x)}{x - a} &= \lim_{x\to a} \frac{f(a) + f'(a)(x - a) - f(x)}{x - a} \\
&= \lim_{x\to a} \left(f'(a) - \frac{f(x) - f(a)}{x - a} \right) \\
&= f'(a) - f'(a) \\
&= 0
\end{aligned}$$

Differentials: Consider now a slight modification of the sketch with which we started the section.



The new sketch shows the tangent line to the graph of a differentiable function $y = f(x)$ at the point (x, y). Now let dx be an increment in x, and let dy be the resulting vertical rise of the tangent line as shown. Then the ratio of dy to dx equals the slope of the tangent line, $f'(x)$.

$$\frac{dy}{dx} = f'(x)$$

Note that in this equation, dy/dx is the ratio of two numbers, not a notation for the derivative.

Until now, dx and dy could not be separated from the symbol $\frac{dy}{dx}$. When we separate them in the way we just did, we call dx the *differential of* x and dy the *differential of* y. Formally, given a differentiable function $y = f(x)$, we treat dx as an independent variable, and define the dependent variable dy according to the formula $dy = f'(x)dx$. That is, dy is obtained by the formal multiplication of $f'(x)$ and dx.

For example, if $y = x^2$, then $dy = 2x\,dx$. Or if $y = \ln x$, then $dy = \frac{1}{x}dx$.

Leibniz had the idea that the notation of calculus should facilitate its use. The new meanings that we have given to the symbols dy and dx do just that, but we will not experience their full power until we put them to use when we learn to integrate in Chapter 3.

Relating differentials to the linear approximation $L(x)$ that we discussed above, note that if $\Delta x = dx$ is a change in x at the point (x, y), then the corresponding change Δy in y is $\Delta y = f(x + dx) - f(x)$, while $dy = L(x + dx) - L(x)$. Since $L(x) = f(x)$, we have that $dy = L(x + dx) - f(x)$. Thus, in general Δy and dy are different and the absolute error in using the tangent line approximation near (x, y) is $|\Delta y - dy| = |f(x + dx) - L(x + dx)|$. Hence, we can think of the differential dy as an approximation to Δy.

Applet: Best Linear Approximation **Try it!**
Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

2.15 Antiderivatives and Initial Value Problems

We have described the problem of differentiation as "half of calculus". The problem of finding the slope of a curve led to a general definition of derivative and was followed by development of general techniques for finding derivatives of given functions. This was the substance of *differential calculus*. Interpretations of the derivative as *rate of change*, *slope of a graph*, and *velocity of a moving object* point to a great variety of applications that characterize differential calculus as the science of dynamic behavior.

Humanity is prone to walk backwards. We like to run videos in reverse. We like to undo our mistakes. And for every procedure known to humans we want to know what happens if we reverse the procedure.

So far we have operated differentiation only in the forward direction. But now we ask the reverse question—if we are given the derivative of a function f can we find an *antiderivative*; i.e. can we find a function F such that $F'(x) = f(x)$? And what interpretations and applications would follow from reversing the process of differentiation?

Definition 1: An *antiderivative* of a function f on an interval I is another function F such that $F'(x) = f(x)$ for all $x \in I$.

Example 1: Find an antiderivative of $f(x) = 2x$. From our knowledge of differentiation techniques we immediately think of x^2. Can we find others? Yes. Indeed we can write down many more: $x^2 + 1$, $x^2 - 3$, $x^2 + 72$. And in fact $f(x) = x^2 + C$ is an antiderivative of f for any constant C. This is the end of our search, for we will show that any two antiderivatives differ only by a constant.

If $F(x)$ and $G(x)$ are antiderivatives of $f(x)$ on an interval I, then

$$\frac{d}{dx}(G(x) - F(x)) = G'(x) - F'(x) = f(x) - f(x) = 0$$

for every $x \in I$. Thus $G-F$ is a differentiable function whose derivative is identically zero on I. The following theorem shows that any such function must be a constant C; i.e. $G(x) - F(x) = C$, or $G(x) = F(x) + C$ for all $x \in I$.

Theorem 1: Suppose that h is differentiable in an interval I and $h'(x) = 0$ for all $x \in I$. Then h is a constant function; i.e. $h(x) = C$ for all $x \in I$, where C is a constant.

The proof is an immediate consequence of the mean value theorem. For if a and b are any two points in I, there is a point $c \in I$ where $h(b) - h(a) = (b - a)h'(c)$. But $h'(c) = 0$, so $h(b) = h(a)$.

This settles the antiderivative problem. If $F(x)$ is one antiderivative of $f(x)$, then any other antiderivative must be of the form $F(x) + C$, where C is a constant. We refer to $F(x) + C$ as the *general antiderivative* and denote it by

$$\int f(x)dx$$

which is called the *indefinite integral of* f. Note that the indefinite integral is nothing but the general antiderivative of f, i.e.

$$\int f(x)dx = F(x) + C$$

where F is one antiderivative of f, and C is an arbitrary constant.

Example 2: In the language of indefinite integrals, the result of Example 1 is just the statement

$$\int 2xdx = x^2 + C.$$

Example 3: Each of our differentiation formulas has a companion *integral* formula. For example

$$\int x^r dx = \frac{x^{r+1}}{r+1} + C$$
$$\int \cos x \, dx = \sin x + C$$
$$\int \sin x \, dx = -\cos x + C$$
$$\int \sec^2 x \, dx = \tan x + C$$
$$\int e^x \, dx = e^x + C$$
$$\int \frac{1}{x} \, dx = \ln|x| + C$$

Each of these formulas is verified by simply differentiating the right hand side.

Note too that just one step removed from these basic integrals are integrals such as

$$\int \cos 3x \, dx = \frac{1}{3}\sin 3x + C$$
$$\int \sin 5x \, dx = -\frac{1}{5}\cos 5x + C$$
$$\int e^{7x} \, dx = \frac{1}{7}e^{7x} + C$$

Here the formulas amount to undoing instances of the chain rule involving constants, and can be verified as usual by differentiating the right hand sides.

The following theorem gives a useful property of indefinite integrals. Just as for derivatives, the indefinite integral of a sum of functions is the sum of the indefinite integrals of the terms.

Theorem 2: Suppose the functions f and g both have antiderivatives on the interval I. Then for any constant a, the function $af + g$ has an antiderivative on I and

$$\int (af + g)(x) \, dx = a \int f(x) \, dx + \int g(x) \, dx$$

The proof follows from the fact that if F is an antiderivative of f (i.e., $F' = f$), and G is an antiderivative of g (i.e., $G' = g$), then $aF + G$ is an antiderivative of $af + g$. We verify this by differentiation: $(aF + G)' = aF' + G' = af + g$.

Example 4: Theorem 2 allows us to find the indefinite integral of any sum of functions whose indefinite integrals we already know, for instance those from the list of Example 3. A typical example is:

$$\int (6x^5 + 4\cos(x) - \frac{1}{x}) \, dx = 6\frac{x^6}{6} + 4\sin(x) - \ln|x| + C$$

That the answer is correct can be verified by differentiating the right hand side, thereby obtaining the function under the integral sign.

Differential Equations: Finding an antiderivative of f can be thought of as solving the equation $\frac{dy}{dx} = f(x)$ for the unknown function y. Such equations that involve one or more derivatives of an unknown function are called *differential equations*. They are of fundamental importance in mathematical modeling. Solving a differential equation means finding a function $f(x)$ that satisfies the equation identically when substituted for the unknown function y.

Example 5: Solve the differential equation $y' = 2x + \sin x$. This is just the antiderivative problem, thus the *general solution* is $y = x^2 - \cos x + C$. Checking the solution means substituting it for y in the differential equation. We see that it does indeed satisfy the equation:

$$\begin{aligned}\frac{d}{dx}(x^2 - \cos x + C) &= \frac{d}{dx}(x^2) - \frac{d}{dx}(\cos x) + \frac{d}{dx}(C) \\ &= 2x - (-\sin x) + 0 \\ &= 2x + \sin x.\end{aligned}$$

Example 6: Solve the *second-order* differential equation

$$\frac{d^2y}{dx^2} + y = 0.$$

The equation is called *second-order* because it involves the second derivative of the unknown function. We do not yet have techniques for solving this equation, but we can easily verify that it has many solutions. For example $\sin x$ is a solution as can be seen by substituting into the equation. [If $y = \sin x$, then $y' = \cos x$ and $y'' = -\sin x$. Hence $y'' + y = 0$.] Other solutions are: $\cos x$, $3\sin x$, $-5\cos x$, $2\sin x - 3\cos x$. Indeed it turns out that the general solution of this differential equation is $y = C_1 \sin x + C_2 \cos x$, where C_1 and C_2 are independent arbitrary constants. We expect that the number of arbitrary constants in the general solution of a differential equation is the *order* of the equation, i.e. the highest order of a derivative that appears in the equation.

In most applications of differential equations the problem at hand will provide additional conditions that enable us to determine values for the arbitrary constants. Then we seek a *particular solution* that satisfies not only the differential equation but also the additional conditions.

Definition 2: An *initial-value problem* is a differential equation together with enough additional conditions to specify the constants of integration that appear in the general solution. The *particular solution* of the problem is then a function that satisfies both the differential equation and also the additional conditions.

The term *initial-value problem* comes from the fact that in many applications of differential equations the independent variable is time t, and the additional conditions specify the state of the system at some initial time, say $t = 0$.

Example 7: Solve the initial value problem $\frac{dx}{dt} = 2t + \sin t$ subject to the initial condition $x(0) = 0$. This problem might be modelling, for example, the motion of an object moving on the x-axis, with velocity $\frac{dx}{dt}$ at time t given by the differential equation. Initially, at time $t = 0$, the object was at the origin. We have already obtained the general solution $x(t) = t^2 - \cos t + C$ of this equation in Example 5. (We have only changed the names of the independent and dependent variables.) The initial conditon enables us to determine a particular value of the constant C. Substituting $t = 0$ into the general solution we obtain

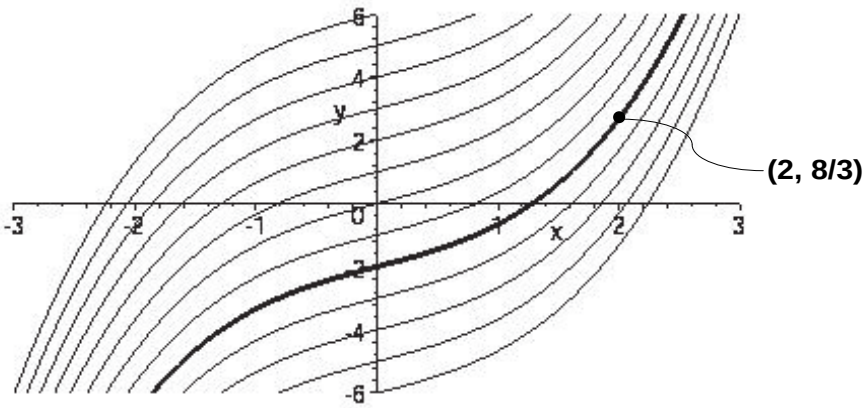
$$0 = 0^2 - \cos 0 + C = 0 - 1 + C = -1 + C.$$

Thus we must have $C = 1$ in order to satisfy the initial condition. And we then obtain the *particular solution* $x = t^2 - \cos t + 1$. Knowing the position of the object at the initial time $t = 0$ and its velocity at any time t, the solution gives us the position of the object at all future times.

Example 8: Solve the differential equation $y' = x^2 + 1$ subject to the additional condition $y(2) = 8/3$. This is again just an antiderivative problem. We solve the equation by "integrating" both sides, thus:

$$\int \frac{dy}{dx}\, dx = \int (x^2 + 1) dx$$

The integral on the left-hand side is y since the indefinite integral is just antidifferentiation. And the integral on the right is $(1/3)x^3 + x + C$. The general solution is thus $y = (1/3)x^3 + x + C$. (It is only necessary to add the arbitrary constant on one side since otherwise we can combine them into a single constant.) Applying the initial condition we must have $8/3 = (1/3)2^3 + 2 + C$, and this yields $C = -2$. The desired particular solution is $y = (1/3)x^3 + x - 2$.



Notice that the general solution is a family of curves that differ only by a vertical translation. The plot shows members of the family for values of C ranging from -6 to 6. The geometric significance of the initial condition $y(2) = 8/3$ is apparent—it "picks out" from the family of curves the particular member of the family that passes through the point $(2, 8/3)$. This is the member corresponding to $C = -2$.

Example 9: Solve the initial-value problem

$$y'' = \cos x, \quad y'\left(\frac{\pi}{2}\right) = 2, \quad y\left(\frac{\pi}{2}\right) = \pi.$$

This time the differential equation is of order two, and two initial conditions are given. Initial-value problems that specify the values of a function and its derivatives at a single point are very common. For a second-order equation, for example, this often comes about by specifying the initial position and velocity (momentum) of an object. For the example at hand we solve the problem by performing two integrations. Integrating both sides of the equation we have

$$\int y''dx = \int \cos x \, dx,$$

or $y' = \sin x + C_1$, where C_1 is an arbitrary constant. Integrating again

$$\int y'dx = \int (\sin x + C_1)dx$$

we obtain the general solution $y = -\cos x + C_1 x + C_2$, where C_2 is a second arbitrary constant. Finally, we obtain the desired particular solution by applying the initial conditions: setting $x = \frac{\pi}{2}$ and the values of y and y' to π and 2 respectively:

$$2 = \sin\frac{\pi}{2} + C_1$$

$$\pi = -\cos\frac{\pi}{2} + C_1\left(\frac{\pi}{2}\right) + C_2$$

Solving these two equations for the constants C_1 and C_2 we find that $C_1 = 1$ and $C_2 = \pi/2$. Thus, finally, $y = -\cos x + x + \pi/2$.

Summary: Differential equations are at the heart of modelling motion in dynamic systems. They provide the language in which we can describe the state of a physical system. For example an object in motion may be described in terms of its position, velocity, and acceleration. An equation relating these properties is thus an equation involving a function and its first and second derivatives. Initial-value problems have additional conditions that allow a *particular solution* to be picked out from the *general solution*. Not surprisingly we will see that differential equations occupy a great deal of our attention from now on.

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

2.16 Velocity and Acceleration

In a previous section we defined the notions of *average velocity* and *instantaneous velocity*. For an object traveling on a straight line, whose position at time t is given by the function $x = x(t)$,

$$\text{Average velocity on the interval } [t, t+h] \ = \frac{x(t+h) - x(t)}{h}, \text{ and}$$

$$\text{Instantaneous velocity } v(t) \text{ at time } t \ = \lim_{h \to 0} \frac{x(t+h) - x(t)}{h}.$$

Recall also from our previous discussion that the acceleration $a(t)$ of the object at time t is defined similarly from the velocity function:

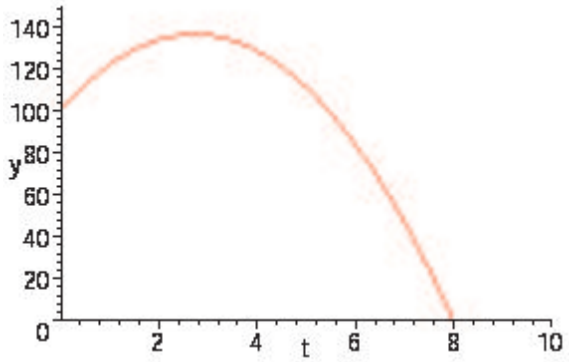
$$a(t) = v'(t) = x''(t).$$

Example 1: An object dropped from a cliff has acceleration $a = -9.8$ m/sec^2 under the influence of gravity. Let its height at time t be given by $s(t)$. Then its motion is described by the initial-value problem

$$\frac{d^2s}{dt^2} = -9.8, \quad s(0) = s_0, \quad s'(0) = v_0$$

where s_0 is the height of the cliff and v_0 is the object's initial velocity. We solved an initial-value problem similar to this one in the previous section. Integrating both sides of the differential equation we get $s'(t) = -9.8t + C_0$, and a second integration gives $s(t) = -4.9t^2 + C_0t + C_1$, where C_0 and C_1 are arbitrary constants. The initial conditions enable us to evaluate the constants. From $s'(0) = v_0$ we obtain $-9/8 \cdot 0 + C_0 = v_0$, and hence $C_0 = v_0$. And from $s(0) = s_0$ we then obtain $-4.9 \cdot 0^2 + v_0 \cdot 0 + C_1 = s_0$, and hence $C_1 = s_0$. The particular solution that describes the motion of the object is thus $s = -4.9t^2 + v_0t + s_0$.

Example 2: Let us apply the result of Example 1 to a particular case. Suppose that a baseball is thrown upward from the roof of a 100 meter high building. It hits the street below eight seconds later. What was the initial velocity of the baseball, and how high did it rise above the street before beginning its descent? Letting $y(t)$ be the height of the baseball above the street, and assuming that it was thrown at time $t = 0$, we know from Example 1 that $y(t) = -4.9t^2 + v_0t + 100$. When $t = 8$ the baseball hits the street,



i.e. $s(8) = 0$. Thus from $0 = -4.9 \cdot 8^2 + v_0 \cdot 8 + 100$ we find that the initial velocity was 26.7 meters/sec^2. Finally, the highest point in the baseball's trajectory, the point when $s(t)$ achieves its maximum value, is when $v(t) = s'(t) = -9.8t + v_0 = 0$. This occurs when $t_{\max} = v_0/9.8 = 26.7/9.8 = 2.72$ seconds (approximately). At this time the baseball is

$$s(t_{\max}) = -4.9(t_{\max})^2 + v_0(t_{\max}) + 100 = 136.37$$

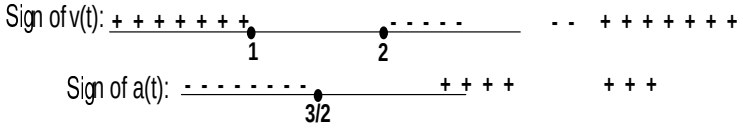
meters above the street.

Example 3: Suppose an object moves on the x-axis so that its position at time t is given by $x(t) = 2t^3 - 9t^2 + 12t + 6$, $-\infty < t < \infty$. When is the object at rest? When is it moving to the right? When is it speeding up? When is it slowing down? These questions are answered by studying $v(t) = x'(t)$ and $a(t) = x''(t)$:

$$v(t) = x'(t) = 6t^2 - 18t + 12 = 6(t-1)(t-2)$$

$$a(t) = x''(t) = 12t - 18 = 12\left(t - \frac{3}{2}\right)$$

The object is at rest when $v(t) = 0$. This occurs when $t = 1$ and when $t = 2$. The position of the object at these times is $x(1) = 11$ and $x(2) = 10$. The object is moving to the right when $v(t) > 0$ and to the left when $v(t) < 0$. A sign graph for $v(t)$, below,



shows that the object is moving right when $t \in (-\infty, 1) \cup (2, \infty)$, and it is moving left in the interval $(1, 2)$. We must be careful when considering when it is speeding up or slowing down. The *velocity* $v(t)$ can be positive or negative; the *speed* of the object is $|v(t)|$. If the object is moving to the right it is speeding up when $a(t) > 0$, but if the object is moving left it is speeding up when $a(t) < 0$ (i.e. the velocity is negative and is becoming more negative). More simply, it is speeding up when $v(t)$ and $a(t)$ have the same sign and slowing down when they have opposite signs. From the sign graphs, above, we can easily compare the signs and see that the object is speeding up on $(1, \frac{3}{2}) \cup (2, \infty)$, and slowing down on $(-\infty, 1) \cup (\frac{3}{2}, 2)$.

Example 4: An object is traveling along the x-axis and its position at time t is given by $x = f(t) = 3\sin\frac{\pi}{2}t$ meters to the right of the origin. At time $t = 1$ describe its position, velocity and acceleration. These three quantities are just the values of $f(1)$, $f'(1)$ and $f''(1)$. Since $x' = f'(t) = 3\frac{\pi}{2}\cos\frac{\pi}{2}t$ and $x'' = f''(t) = -3\left(\frac{\pi}{2}\right)^2\sin\frac{\pi}{2}t$, we can state that the position of the object is $3\sin\left(\frac{\pi}{2}\cdot 1\right) = 3$; i.e. it is 3 meters to the right of the origin. Its velocity is $3\frac{\pi}{2}\cos\left(\frac{\pi}{2}\cdot 1\right) = 0$ meters/second; i.e. it is momentarily at rest. And its acceleration is $-3\left(\frac{\pi}{2}\right)^2\sin\left(\frac{\pi}{2}\cdot 1\right) = -3\left(\frac{\pi}{2}\right)^2 = -7.4$ meters/second^2; i.e. it is "speeding up toward the left".

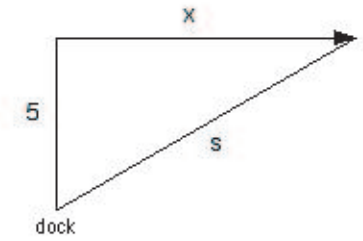
Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

2.17 Related Rates

One of the applications of mathematical modeling with calculus involves related-rates word problems. Suppose we have an equation that involves two or more quantities that are changing as functions of time. Then differentiating the equation implicitly with respect to time gives an equation that involves the rates of change of these quantities. By relating the rates in this way, we often can answer interesting questions about the model that we use to specify the original problem.

Example 1: Suppose a ship is traveling at a speed of 28 knots (nautical miles per hour) in a line perpendicular to a dock on the shore as in the sketch below. How fast is the distance between it and the dock increasing 15 minutes after it passes directly opposite the dock five miles away?



Consider the sketch that describes the situation. Here x is the distance traveled by the ship; thus, we are *given* that

$$\frac{dx}{dt} = 28 \text{ nautical miles per hour}$$

and we are asked to *find*

$$\frac{ds}{dt} \text{ when } t = \tfrac{1}{4} \text{ hour.}$$

We begin by writing an equation that involves x and s:

$$s^2 = x^2 + 25$$

Then we differentiate with respect to t:

$$2s\frac{ds}{dt} = 2x\frac{dx}{dt}$$

Hence, solving for $\frac{ds}{dt}$ we get:

$$\frac{ds}{dt} = \frac{x}{s}\frac{dx}{dt}$$

Now, we can *answer the question*: When $t = 1/4$, $x = (28)(1/4) = 7$. Thus, from the Pythagorean Theorem, $s = \sqrt{25 + 49} = \sqrt{74}$. Hence,

$$\frac{ds}{dt} = \frac{7}{\sqrt{74}}(28) \approx 22.78 \text{ knots}$$

Example 2: How fast is the area of a square changing when the side is of length 10 cm and is increasing at a rate of 4 cm/s? To answer the question we begin with $A = s^2$ where s is the length of the side and A is the area. Differentiating with respect to t we get:

$$\frac{dA}{dt} = 2s\frac{ds}{dt}$$

Thus, when $s = 10$, $\frac{dA}{dt} = (2)(10(4) = 80$ square centimeters per second.

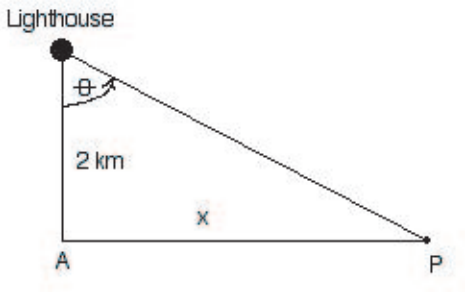
These examples illustrate a general procedure for solving Related Rates problems:

1. Begin by making a sketch whenever you can.
2. Define symbols and write down what is given.
3. Write an equation that links the variables.

4. Differentiate the equation implicitly with respect to time to get an equation that links (that is, relates) the rates.

5. Answer the question by substituting in specific values that are given.

Example 3: A lighthouse is located on a mound of rock out in the ocean 2 km from the nearest point A on a straight shoreline. If the lamp rotates at a rate of 3 revolutions per minute, how fast is the lighted spot P on the shoreline moving along the shoreline when it is 4 km from point A? We start by making a sketch.



With reference to the sketch, we have:

$$\tan\theta = \frac{x}{2}$$

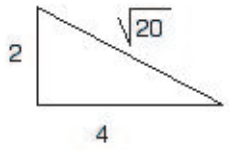
$$\sec^2\theta\,\frac{d\theta}{dt} = \frac{1}{2}\frac{dx}{dt}$$

$$\frac{dx}{dt} = 2\sec^2\theta\,\frac{d\theta}{dt}$$

Now, with reference to the given information, we have to change *revolutions* to *radians* so that we can use the differentiation formulas above: 1 revolution per minute equals 2π radians per minute. Thus,

$$\frac{d\theta}{dt} = 6\pi \text{radians per minute}$$

And from the triangle below, we see that when $x = 4$, $\sec\theta = \sqrt{20}/2$.



Therefore, we can answer the question: when $x = 4$, we have

$$\frac{dx}{dt} = 2\left(\frac{\sqrt{20}}{2}\right)^2 6 = 60\pi \text{ km/min}$$

Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

2.18 Case Study: Torricelli's Law

Animation: Torricelli's Law **To get you going on the Case Study!**

We close section 2 with a *Case Study in Calculus* (CSC) that is an extended example of modeling rates of change. The purpose of a CSC is to consider a real application of calculus, with real data. The question we want to answer in the present section is straightforward enough:

Objective: To determine how long it would take a tank of given dimensions to empty its liquid contents through a bottom outlet hole.

In this CSC, we will play the role of a mathematician who works with a group of physicists. We will take note of certain physical *laws*, described to us by physicists, and use these laws to set up a differential equation that can be solved using methods that we have learned.

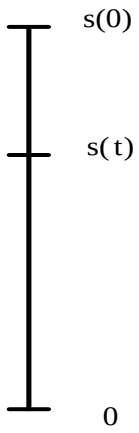
To a physicist a law is a statement for which there is solid empirical evidence of its truth. That is, the statement has never been known to be false. This test of truth, close to the legal standard of *beyond a shadow of a doubt*, is not good enough for mathematicians. Mathematicians require a valid deductive argument to *prove* the truth of a statement.

However, in the real world, compromises must be made. Our compromise as mathematicians will be to accept laws of physicists in order to set up our model, and then to apply mathematical analysis to derive an answer. Of course, we actually will have come full circle because interpreting the mathematical answer as it applies in a given situation may not be so crisp and clear-cut. The world of real data is messy, and approximations and simplifications have to be made. Keep these distinctions in mind as you go through the CSC. And by all means, have fun putting calculus to work.

Background: Falling Objects

Let m be the mass of a falling object, and let g be the acceleration due to gravity. Then $g \approx 9.8$ meters per second squared, or $g \approx 32.2$ feet per second squared.

Let $s(t)$, $s(t) \geq 0$ for all t, be the position of the object above the ground at time t according to the following coordinate axis:



If a is the acceleration and $v(t)$ is the velocity of the object at time t, then we have:

$$\begin{aligned} a &= -g \\ v &= -gt \\ s &= -\frac{gt^2}{2} + s_0 \end{aligned}$$

Thus, we can find the final velocity v_f at the final time t_f:

$$\begin{aligned} \frac{gt_f^2}{2} &= s_0 \\ t_f &= \sqrt{\frac{2}{g}s_0} \\ v_f &= -\sqrt{2gs_0} \end{aligned}$$

We can also write the Initial Value Problem that gives the velocity $v = \frac{ds}{dt}$ as a function of s:

$$\frac{ds}{dt} = -\sqrt{2gs}, \qquad s(0) = s_0$$

We have made the substitution $v(t) = \frac{ds}{dt}$, the rate of change of position with respect to time. The velocity is negative because the object is moving toward the origin on our positive scale. The differential equation involves the derivative of an unknown function $s(t)$. The equation $s(0) = s_0$ defines the *initial condition* and tells us that the starting position of the object is the greatest point on the coordinate line we have chosen. The differential equation and the initial condition is what we will need for the exploration of Torricelli's Law below.

The CSC: Torricelli's Law

Evangelista Torricelli (1608-1647) was an Italian physicist and mathematician who was a disciple of Galileo. He also served as Galileo's secretary, and is credited with discovering the following principle.

Torricelli's Law: Water in an open tank will flow out through a small hole in the bottom with the velocity it would acquire in falling freely from the water level to the hole.

This law is not at all obvious. Presumably, Torricelli discovered it through a combination of studying empirical data and demonstrating great physical insight. Here is one plausible explanation, although a complete derivation of the law requires a good understanding of hydrostatics. Imagine the water as a collection of tiny balls undergoing elastic collision. If we consider a vertical chain of balls, the kinetic energy of a falling ball will be completely transferred to the next one. Thus, the new initial velocity of the next ball equals the final velocity of the last, and so on down the line. When the last ball gets to the outlet hole, it will carry the same kinetic energy as if the top ball had fallen all the way down. But no matter how he arrived at it, Torricelli was indeed correct: today his law is a well-established scientific fact.

We can make Torricelli's Law more specific by introducing some notation. Suppose a cylindrical tank with cross-sectional area A has an outlet hole in the bottom. Further, suppose that $h(t)$ is the height of water above the outlet at time t, a is the area of the outlet hole, and $V(t)$ is the remaining fluid volume at time t.

Consider now the change in volume of water in the tank from time t to time $t + \Delta t$. This equals the amount of water that flows out through the outlet hole in time $[t, t + \Delta t]$. If we think of this water as filling a small cylindrical tube whose top is the outlet hole of cross-sectional area a, then the height of the tube is the velocity of a drop of water (if it were constant) times the time. By Torricelli, the initial velocity of a drop is $\sqrt{2gh}$ (the final velocity from the background analysis of falling objects above). Thus, the volume of the tube in the interval $[t, t + \Delta t]$ is approximately $a\sqrt{2gh}\ \Delta t$. Equating the change of volume of the tank with the volume of water in this tube and using this approximation for the latter, we get

$$\begin{aligned} V(t) - V(t + \Delta t) &\approx a\sqrt{2gh}\ \Delta t \\ \frac{V(t) - V(t + \Delta t)}{\Delta t} &\approx a\sqrt{2gh} \\ \frac{V(t + \Delta t) - V(t)}{\Delta t} &\approx -a\sqrt{2gh} \\ \lim_{\Delta t \to 0} \frac{V(t + \Delta t) - V(t)}{\Delta t} &= -a\sqrt{2gh} \\ \frac{dV}{dt} &= -a\sqrt{2gh} \end{aligned}$$

Now, we also have that the volume V of remaining water at time t equals the cross-sectional area A of the tank times the height h of water at time t. Thus, the rate of change of volume with respect to time is the cross-sectional area of the tank times the rate of change of the height of water with respect to time. Putting these comments together with the above equation for $\frac{dV}{dt}$ will give a differential equation involving $\frac{dh}{dt}$ and h, but we will leave that for the setup part of the CSC.

Thus far, we have collected information relevant to answering the question posed at the beginning. Let's restate the objective a bit more precisely, and then make an inventory of that information.

The Objective of the CSC: To determine, from the background information above, how long it would take a cylindrical tank of given dimensions to empty through a bottom outlet hole of known diameter.

A possible scenario for needing to solve this problem is that you are a consultant for an oil company. The company wants to know how long it will take to empty its oil storage tank into its tanker trucks.

Nature of the information that is available: Dimensions of a tank, Torricelli's Law, Equation of Motion of a falling object.

All of this information appears above. We have to make sense of it. To do so, we will bring to bear our analytical skills, and complete a number of steps using some of the mathematical tools we have learned. Here are the steps.

Steps to follow to analyze the information: First, model the physical situation as a differential equation. Next, solve the differential equation. And finally, relate the solution to the question posed in the objective.

It is always a good idea to list the tools that we will use: derivatives, antiderivatives, our brains. The CSC is an extended application of calculus. It is *extended* in the sense that we will have to complete several stages to arrive at a solution. So, as always, thinking is important. In fact, the less routine a problem, the more important thinking becomes. But don't worry. A major purpose of the CSC is to learn to think clearly about such problems. Toward that end, we have structured the steps of the analysis in the form of a report that you will complete. The main sections of the report are as follows (see homework).

Setup: This is what is called the *modeling phase*. The model is often a differential equation. In this section, we derive the equation to be solved, being careful to be clear about the reasoning– scientific or mathematical. We do not solve the differential equation here, but in the next section.

Thinking and Exploring: Now it is time to think about the mathematics. We will suggest some activities for you to carry out to develop an understanding of the differential equation and its relationship to the solution of the barrel problem stated in the objective above.

Applet: Function Grapher **Try it!**

Interpretation and Summary: Now that you have modeled the barrel problem, and thought about the model and derived mathematical facts, it is time to interpret and summarize the mathematical results in terms of the original objective. This is a very important part of the report because mathematicians often have to explain their work to non-mathematicians, and be convincing about the proposed course of action. Pretend that your synopsis is going to appear in the next issue of a magazine such as *Scientific American*. Include enough details so that a reader would learn what the major issues of the report are, and how you went about addressing them. What will you want to tell readers about your success with regard to the original stated objective of the investigation? You should take care to write in complete sentences using correct rules of standard English grammar. Make the report interesting, compelling. Ask yourself: Would a friend enjoy reading it? Would an oil company executive follow your advice? Both answers should be yes.

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

Chapter 3

Modeling with Differential Equations

3.1 Introduction to the Issues

A differential equation is an equation involving derivatives and functions. In the last section, we began with a table of values for the rate of change of a function, and we wanted to know how to use it to get information about the function. Now, we are going to assume that we have a complete description of the derivative of a function in the form of an equation that it satisfies, and we are going to address the question: What is the function? That is, how can we obtain f from f' ?

This question is not an issue of idle speculation. Many physical and biological systems can be modeled with differential equations. The main reason is because often it is relatively easy to measure the amount of something that is present at a given time, and then how the amount changes as the system goes from one state to another. For example, we have already discussed the empirical observation that, at any time, the rate of increase of a large rabbit population is proportional to the number of rabbits at that time. If we let $y(t)$ be the number of rabbits at time t, then this observation can be rewritten as the differential equation $\frac{dy}{dt} = ky$, where k is a constant of proportionality. If we let $y(0)$ be the number of rabbits at the beginning of the observation period, then in mathematical terms we say that we have a differential equation and an accompanying initial condition.

In this section we will take up the solution of the general form of this problem. That is, given a differential equation and an initial condition, we want to find the function that satisfies it. To say that f satisfies a differential equation means probably exactly what you think it does; namely, that when f is substituted into the differential equation, the left-hand-side equals the right-hand-side. But this begs the question, how do we find such an f? Many books are devoted to answering this question. A typical one might have a title such as *1001 Methods for Solving Differential Equations*. Our object here is not to become experts in finding solutions to 1001 different kinds of differential equations, but to explore some of the general approaches that always merit consideration.

3.1.1 Solution by Inspection

It may seem a bit silly, but the first thing to do when examining a differential equation is to try to guess a solution. Once we have a candidate, we then have to verify that our hunch is correct by showing that indeed the function does satisfy the equation. Thus, this method is referred to as *guess-and-check*.

For example, the differential equation

$$\frac{dy}{dx} = ky$$

says that the derivative of the function is a constant times the (same) function. So, we ask ourselves: Is there an elementary function whose derivative is a constant times itself? You might ask, why are we restricting to the elementary functions? This is a good question, and its answer gives one of the reasons

that these functions are so important. It turns out that, as we found in the first section, the elementary functions have shown themselves to be extremely valuable in modeling. They come up over and over again. Thus, they are always a good place to start, especially if we are guessing.

But back to answering the question. An exponential function looks like a good candidate. In fact, the derivative of e^{kt} equals ke^{kt}. So, indeed, the derivative of this function is k times the function, and we have solved the equation. See how powerful the guess-and-check method is?

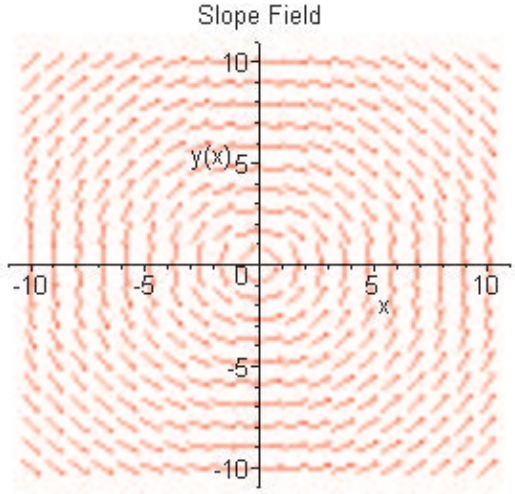
3.1.2 Slope Fields

In the guess-and-check method, we are considering the equation from a formulaic point of view. That is, we ask ourselves if we can think of a function whose derivative-formula has the desired relationship to the function in the equation. We could also think about the equation from the perspective of slopes. For example, the differential equation

$$\frac{dy}{dx} = -\frac{x}{y}$$

tells us that at every point (x, y) of the plane, the graph of the solution function has slope $-x/y$. For example, at the point (1, 1), the slope of the tangent line to the solution curve passing through that point is -1/1 = -1. Or at (2, -1), the slope of the solution curve passing through it is -2/-1 = 2. This suggests plotting short tangent lines at points of the plane that are sufficiently close together. Then, starting at a given point, draw in the curve that follows the tangent lines as one moves away from that point. We then can try to recognize, if possible, this solution curve as the graph of a function we know. In the worse case, we have approximate values for the solution function at specific values of x.

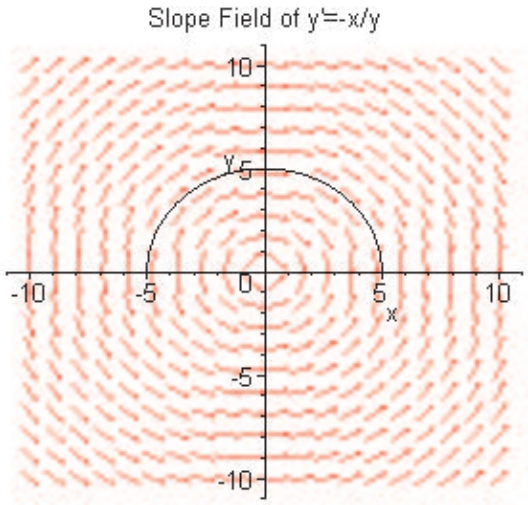
A plot of the tangent lines is called the *slope field* of the differential equation. Here is a slope field for the equation we are considering.



The slope field gives a family of particular solutions. From a starting point, say (-5, 0), it looks like the curve that follows the slopes (or tangent lines) is a semicircle centered at the origin. Thus, although not at first apparent as a guess, the slope field suggests that a solution function might be $y(x) = \sqrt{a^2 - x^2}$, where the curve passes through the point $(a, 0)$. Hence, we substitute this function into the differential equation and check it:

$$\frac{dy}{dx} = \frac{1}{2\sqrt{a^2 - x^2}}(-2x) = -\frac{x}{y}$$

Eureka! It checks. Thus, the slope field has given us a candidate to apply our guess-and-check method to. Here is a plot on the slope field of this solution curve in the particular case that it goes through the point (-5, 0).



The slope field raises the issue of how close together we should plot the points. We will return to slope fields later when we will consider what to do if we are unable to guess a solution function whose graph follows the slopes. The technique is called Euler's Method; it is a numerical technique that results almost immediately from what we have just done. The distance between points will be a crucial issue in assuring ourselves of a good approximation to the particular solution we seek. But we will postpone that discussion for now, and continue to concentrate on outlining general approaches to solving a differential equation. In the rest of the Chapter, we will pursue the details and ramifications that we identify as we go along.

Applet: Slope Field **Try it!**

3.1.3 An Analytical Tool: Separation of Variables

In the last section, we considered Initial Value Problems (IVPs). They involved differential equations that can be put in the form

$$\frac{dy}{dx} = g(x); y(a) = y_a$$

That is, we can solve explicitly for the derivative as a function of x. Hence the general solution is found by integrating both sides of the equation and using the initial condition to determine the particular solution. The IVP

$$\frac{dy}{dx} = x^2 + x + 1; y(0) = 2$$

is such an example. The particular solution is

$$y = \frac{x^3}{3} + \frac{x^2}{2} + x + 2$$

Thus, there is no need to guess; we have a systematic procedure for solving these IVPs.

Differential equations that are just one step removed from these are those that we call *separable*. They are of the form

$$\frac{dy}{dx} = g(x) \cdot h(y)$$

To understand why these equations are called *separable*, we reinterpret them using differentials. Given a function $y = f(x)$, we have defined the differential dy of y according to the formula $dy = f'(x)dx$. Thus, the ratio of dy over dx equals $f'(x)$.

Now, viewing the equation in terms of differentials, a separable equation is one in which we can put all of the y's and dy's (as products) on one side of the equation and all of the x's and dx's (as products) on the other. That is, the variables can be separated to obtain the equation

$$\frac{dy}{h(y)} = g(x) \cdot dx$$

Then, we can integrate both sides of the separated equation, and solve to find a general solution. We will show later in the chapter why the method works, but for now we will confine ourselves to showing how it works. Let's begin by illustrating with an example we have already considered. Let

$$\frac{dy}{dx} = -\frac{x}{y}$$

Separating the variables, integrating, and solving, we get:

$$\begin{aligned}
y \cdot dy &= -x \cdot dx && \text{(separate the variables)} \\
\int y\, dy &= \int -x\, dx && \text{(integrate both sides)} \\
\tfrac{y^2}{2} &= -\tfrac{x^2}{2} + C && \text{(find indefinite integrals)} \\
y^2 &= -x^2 + C_1 && \text{(rename constant; } C_1 = 2C\text{)} \\
y = \sqrt{C_1 - x^2} &\text{ or } y = -\sqrt{C_1 - x^2} && \text{(solve for } y\text{)}
\end{aligned}$$

Thus, we see that the solution y to the differential equation satisfies the relationship $x^2 + y^2 = C_1$, where C_1 is a constant; in other words, the points lie on a circle centered at the origin. As soon as we know a point through which the circle passes, we can give a particular solution. When we exhibited above the slope field of this equation, we wanted the solution that passed through (-5, 0). Note that as we could have observed from the slope field, there are two solutions: $y = \sqrt{25 - x^2}$ and $y = -\sqrt{25 - x^2}$, the top and bottom halves of the circle.

We can also use the method of Separation of Variables to solve the differential equation $dy/dx = ky$ that we solved above by guess-and-check. Separating the variables, integrating, and solving, we get

$$\begin{aligned}
\frac{dy}{y} &= k dx \\
\int \frac{1}{y}\, dy &= \int k\, dx \\
\ln|y| &= kx + C
\end{aligned}$$

Thus, $y = e^{kx+C}$ or $y = -e^{kx+C}$. Hence, $y = e^C e^{kx}$ or $y = -e^C e^{kx}$. Thus, $y = y_0 e^{kx}$, where y_0 is the initial value of the function y.

The method of Separation of Variables is an important analytical tool. Given that the differential equation is in a separable form, the method allows us to approach its solution in a systematic way. It also allows us to put the equation in a form where we can turn to tables or numerical methods to evaluate the integrals if we do not recognize any antiderivatives. But not even every simple-looking equation is separable. For example, consider the equation

$$\frac{dy}{dx} = x - y$$

This equation is not separable. It also is not so obvious what the solution is. Therefore, short of using our generic book *1001 Methods for Solving Differential Equations*, or a computer algebra system to solve the equation, our approach would be to generate its slope field and/or use Euler's method to approximate the solution curve corresponding to a given initial condition.

Applet: Slope Field **Try it!**

3.1.4 Existence and Uniqueness of Solutions of Initial Value Problems

While we do not want to say much about this subject, we do want to say enough to give assurances that, in most cases we will meet, a solution to an IVP *will exist* and *will be unique*. We will divide the subsection into two parts, first stating the main results, and then outlining additional details for those who would like to see where to look in the advanced literature for even more information. In either case, the material is interesting but comes close to going beyond the scope of this text.

Consider the IVP $y' = F(x, y)$, $y(a) = b$, where $F(x, y)$ is continuous in a domain D that is an open region of the xy-plane and (a, b) is a point in D.

As we have said before, a solution of the IVP is a function $f(x)$ that satisfies the equations; i.e. $f'(x) = F(x, f(x))$ and $f(a) = b$.

In the cases we will meet, the existence of a solution will not be an issue because our techniques will allow us to find one. However, an important result in the theory of differential equations is Peano's Existence Theorem, which states that under the conditions above, there is always at least one solution of the IVP, and any such solution is differentiable. (The theorem actually says more, namely, that any such solution can be extended in both directions to the limits of the region D. In fact, there are maximal and minimal solutions $fMax(x)$ and $fMin(x)$ such that all other solutions lie between them, and the "bundle" of such solutions completely covers the part of the region lying between their graphs. But we will not attempt to explain here exactly what any of this parenthetical comment means.)

Now we come to the main issue about which we need assurances: uniqueness. Even though our techniques will produce a solution, how do we know that it is the only one? In particular, how do we know that there is not another solution different from the one that we get from applying the method of separation of variables to a separable equation?

Here is the answer. In our case we consider a differential equation $y' = g(x)h(y)$. We assume g and h are continuous in a region D containing the initial point (a, b). Then there is always a solution (Peano), and if g' and h' are continuous, the solution is unique. Thus, the solution found using the method of separation of variables is unique if g' and h' satisfy these conditions.

The example $y' = y$ is instructive. (We solved this as a separable differential equation $y' = ky$ above with $k = 1$, $y \neq 0$.) Here $g(x) = 1$ and $h(y) = y$. Thus, the continuity and differentiability conditions are met because $g'(x) = 0$ and $h'(y) = 1$ imply that g, h, g', and h' are all continuous. And indeed there is a unique solution through any point (a, b). If the initial point is on the x-axis, however, the unique solution is $y = 0$, and this is not found by the method of separation of variables (one cannot divide by zero). Of course the fact that the method of separation of variables did not "find" the solution $y = 0$ has nothing to do with existence. The solution exists and is unique. It is only that the application of the separation of variables method was not valid in the case $y = 0$.

Whew! For most of us, this is all we need to know about the existence and uniqueness of solutions of a differential equation. However, we will now conclude our discussion with a few ideas that can point the way to further investigations of the subject at a subsequent point in your study of calculus.

To have uniqueness of the IVP $y' = F(x, y)$, $y(a) = b$, where $F(x, y)$ is continuous in a domain D, it is sufficient that $F(x, y)$ satisfy a Lipshitz condition in D, i.e. that there exist a constant M such that $|F(x, y_1) - F(x, y_2)| \leq M|y_1 - y_2|$. (Note: In a subsequent course on multivariable mathematics, you will learn that it is actually sufficient that $F(x, y)$ have continuous partial derivatives in the region D, and that the continuity of the partial derivatives in any closed bounded subregion of D implies a Lipschitz condition in that subregion.)

Example of non-uniqueness: The differential equation $y' = 3y^{\frac{2}{3}}$ has the "general solution" $y = (x+C)^3$. Just check that this is true by differentiation and substitution: $y = (x+C)^3$ implies $y' = 3(x+C)^2$, and $3y^{\frac{2}{3}} = 3((x + C)^3)^{\frac{2}{3}} = 3(x + C)^2$. But note that $y = 0$ is also a solution not of the general form. Of

course if one specifies a domain D that does not touch the x-axis, then one has uniqueness within this domain. The function $3y^{\frac{2}{3}}$ has a continuous derivative $2y^{-\frac{1}{3}}$ when bounded away from the x-axis. For example $D = \{(x, y)|y > 0\}$ is such a domain.

3.1.5 Our Agenda for This Chapter

The prevalence of differential equations and their role in applications is that they are the mathematical language used to describe many biological and physical systems. We have seen that differential equations arise quite naturally in modeling systems that involve measurable quantities and their related rates of change as the system moves over time from one state to another. Hence, we cannot overemphasize the importance of differential equations in real-world applications. The world changes; differential equations describe how. Therefore, in this section we will be studying first approaches to modeling with and solving differential equations. This will involve applications, and both analytical and numerical techniques. Because we have already gotten started, we will take up where we left off by building on and extending the topics considered in this section. In particular, here is an agenda for the rest of the chapter as it grows out of our investigations so far.

1. Revisit the population model $\frac{dy}{dx} = ky$ and study further examples of growth ($k > 0$), and of decay ($k < 0$).
2. Revisit Separation of Variables and investigate why the method works.
3. Revisit slope fields and introduce Euler's Method.
4. Discuss analytical and numerical methods for studying more general population models.

Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

3.2 Exponential Growth and Decay

Derivatives are functions that measure rates of change. A rate of change can be a powerful tool for expressing quantitatively a qualitative description. In the process, we build what we call a model. For example, we know that a nation's population grows or declines depending on the birth and death rates. Can we say something more basic about growing populations? For example, does it make sense to say that at any time t, the rate of change of the size of a growing population is proportional to its size? Probably. In fact, this is most certainly true in many many cases. And look how easy it is to turn this qualitative statement about population growth into an equation.

Just let $y(t)$ be the size of the population at time t. Then the statement becomes:

$$\frac{dy}{dt} = ky; y(0) = y_0$$

where k is a constant of proportionality and $y(0)$ is the size of the population at $t = 0$.

This is another example of a differential equation, that is, an equation that includes derivatives. We have turned it into an *Initial Value Problem* by specifying an initial condition. That is, a particular point $(0, y(0))$ that the solution passes through (in this case, at time$t = 0$). The resulting differential equation becomes a model of the population. Of course, we then have to test the predictions of this model against actual data for this population that we can get from, say, census tables. If the model is a good one, then we can use it to predict the size of the population in the future.

Note that if $k > 0$, then the population is growing, and if $k < 0$, then the population is decreasing. In the next theorem, we solve the differential equation.

Theorem 1: The IVP $\frac{dy}{dt} = ky$, $y(0) = y_0$, k constant, has unique solution $y = y_0\, e^{kt}$.

The proof is a nice mixture of calculus and algebra. For, if $y \neq 0$, we can rewrite the equation, integrate both sides, and solve for y:

$$\begin{aligned} \frac{1}{y}\,\frac{dy}{dt} &= k \\ \int \frac{1}{y}\,\frac{dy}{dt}\,dt &= \int k\,dt \end{aligned}$$

The integral on the left should be read as a reversal of the chain rule because $\frac{d}{dt}\ln|y| = \frac{1}{y}\,\frac{dy}{dt}$. Thus, we can integrate both sides of the integral equation, then exponentiate both sides, and solve for y:

$$\begin{aligned} \ln|y| &= kt + C \\ |y| &= e^{kt+C} \\ &= e^{kt}\, e^{C} \\ y &= \pm e^{C}\, ekt \\ &= Be^{kt} \end{aligned}$$

where $B = \pm e^C$ is just a constant. Next, we use the initial condition $y(0) = y_0$ to find B: $y_0 = y(0) = Be^0 = B \cdot 1 = B$. Thus, the solution is $y = y_0\, e^{kt}$, and the proof is complete.

In applied problems, we go straight to the solution and do not repeat its derivation. Hence, once we have described the model and written down the solution to the IVP from the theorem, the problems involve only algebra.

Example 1: Suppose a bacteria culture grows at a rate proportional to the number of cells present. If the culture contains 700 cells initially and 900 after 12 hours, how many will be present after 24 hours? To solve this problem, we note first that because the growth is proportional to the number of cells present, then if we let $y(t)$ be the number of cells present at time t, we know from the theorem that $y(t) = y_0e^{kt}$ where

$y_0 = 700$. So, the problem at hand is to find k from the given information:

$$\begin{aligned} y(12) &= 700e^{12k} \\ 900 &= 700e^{12k} \\ \frac{900}{700} &= e^{12k} \\ \ln\left(\frac{900}{700}\right) &= 12k \\ k &= \frac{\ln 900 - \ln 700}{12} \end{aligned}$$

We can approximate this number using a calculator: $k \approx .0209$. Then we can answer the question: After 24 hours there will be approximately $y(24) = 700e^{.0209 \cdot 24} \approx 1156$ cells.

Doubling Time and Half-Life: In an exponential growth model, the *doubling time* is the length of time required for the population to double. In a decay model, the *half-life* is the length of time required for the population to be reduced to half its size. A characteristic of exponential models is that these numbers are independent of the point in time from which the measurement begins.

Example 2: A radioactive substance that decays according to an exponential model has a half-life of 600 years. What percentage of an original sample is left after 10 years? Once again, our assumption implies that the amount present at time t is given by $y(t) = y_0 e^{kt}$. We use the information about the half-life to find k:

$$\begin{aligned} \frac{y_0}{2} &= y_0 e^{600k} \\ \frac{1}{2} &= e^{600k} \\ \ln\left(\frac{1}{2}\right) &= 600k \\ k &= \frac{\ln 1 - \ln 2}{600} \\ k &= \frac{-\ln 2}{600} \\ k &\approx -0.001155 \end{aligned}$$

Now, we can answer the question: $\frac{y(10)}{y(0)} = e^{10k} = e^{10 \cdot (-.001155)} \approx 0.9885$, or 98.85 percent.

Newton's Law of Cooling: This law states that a hot object introduced into an environment maintained at a fixed cooler temperature will cool at a rate proportional to the difference between its own temperature and that of the surrounding environment. That is, if $y(t)$ is the temperature of the object t units of time after it is introduced into a medium at fixed temperature T_m, we have

$$\frac{dy}{dt} = k\,(y - T_m)\,; y(0) = y_0$$

where k is a constant.

Example 3: Suppose a metal object at 112 degrees Fahrenheit is removed from boiling water and placed on a plate in a room maintained at 68 degrees F. Suppose the object cools to 90 degrees in 5 minutes. How long will it take to cool to 80 degrees? Note that in this problem, $T_m = 68$, and $y_0 = 112$. We will return to the problem after we solve the differential equation.

We need to solve the above differential equation to find $y(t)$. Proceeding as before, we have

$$\begin{aligned} \frac{1}{y - T_m}\,\frac{dy}{dt} &= k \\ \int \frac{1}{y - T_m}\,\frac{dy}{dt}\,dt &= \int k\,dt \end{aligned}$$

This looks familiar. In fact, because the integrand on the left is the derivative of $\ln|y - T_m|$, we know from following our previous steps that the solution is $y - T_m = Be^{kt}$. Thus, substituting $y(0) = y_0$ yields $B = y_0 - T_m$ and we get the solution

$$y - T_m = (y_0 - T_m)\, e^{kt}$$

Example 3 (continued): We have that $y - 68 = (112 - 68)e^{kt}$. Thus,from the given information, we find k: $90 - 68 = 44e^{5k}$, so $22/44 = e^{5k}$, and $k = (\ln 22 - \ln 44)/5$, whence $k \approx -.1386294362$. Now, we can answer the question: We want to know at what time t the temperature of the metal object is 80 degrees. That is, we solve for t in the equation $80 - 68 = 44e^{-.1386294362t}$: $12/44 = e^{-.1386294362t}$, or $t = (\ln 44 - \ln 12)/.1386294362$. So, the object will reach 80 degrees approximately 9.37 minutes after it is removed from the water.

Applet: Calculator: Values of Elementary Functions **Try it!**
Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

3.3 Separable Differential Equations

We have already seen that the differential equation $\frac{dy}{dx} = ky$, where k is a constant, has solution $y = y_0 e^{kx}$. We have solved this equation in three ways: by guess-and-check in Section 3.1, and by algebraic manipulation and integration in Section 3.2. The differential equation, representing exponential growth or decay, is also an example of a *separable* differential equation, which we solved as such in Section 3.1.

As introduced in Section 3.1, a first-order differential equation in x and y is called *separable* if it is of the form

$$\frac{dy}{dx} = g(x)h(y)$$

where $y = f(x)$. That is, when the equation is written in terms of differentials, the x's and dx's can be put on one side of the equation and the y's and dy's on the other in such a way that we can solve the equation by integrating both sides:

$$\frac{1}{h(y)}\, dy = g(x)\, dx$$

$$\int \frac{1}{h(y)}\, dy = \int g(x)\, dx$$

This procedure to solve the differential equation is called the *method of separation of variables*.

Example 1: As a review, let's again solve the equation $\frac{dy}{dx} = ky$ by the method of separation of variables. The method begins by rewriting the equation using differentials. First, we separate the y's and dy's from the x's and dx's, and then we integrate both sides of the rewritten equation, and solve for y:

$$\frac{1}{y}\, dy = k\, dx$$

$$\int \frac{1}{y}\, dy = \int k\, dx$$

$$\ln|y| = kx + C$$

From this point on, we do exactly what we did before: we solve for y by exponentiating both sides:

$$|y| = e^{kx+C}$$

$$y = \pm e^C e^{kx} = y_0 e^{kx}$$

Justification for the Method of Separation of Variables: But why is the method of separation of variables valid? After all, on the left side of the separated equation we are integrating with respect to y, and on the right side with respect to x. Using differentials facilitates the method and is a reflection of the genius of Leibniz who believed that the notation should be chosen to motivate the correct answer. However, we have just described a subtlety that we don't want to slide over. The method does indeed give the correct answer, but we must prove it. *Proof by notation* will not siffice.

In fact, we need to show that given the equation

$$\frac{dy}{dx} = g(x)h(y)$$

the antiderivative of $\frac{1}{h(y)}$ as a function of y equals the antiderivative of $g(x)$ as a function of x.

The function $y = f(x)$ is a solution of the above equation implies that

$$\begin{aligned} f'(x) &= g(x)h(f(x)) \\ \frac{f'(x)}{h(f(x))} &= g(x) \end{aligned}$$

Let $H(y)$ be any antiderivative of $1/h(y)$; so $H'(y) = 1/h(y)$. Then applying the chain rule yields

$$\begin{aligned} \frac{d}{dx}H(f(x)) &= H'(f(x))f'(x) \\ &= f'(x)\frac{1}{h(f(x))} \\ &= g(x) \end{aligned}$$

Thus, the solution $y = f(x)$ satisfies the equation

$$H(f(x)) = \int g(x)\,dx$$

However, this is just the result of the method of separation of variables, which is to rewrite the differential equation as

$$\frac{1}{h(y)}\,dy = g(x)\,dx$$

and to integrate both sides (the left side with respect to y and the right with respect to x) obtaining an equation of the form

$$H(y) = \int g(x)\,dx$$

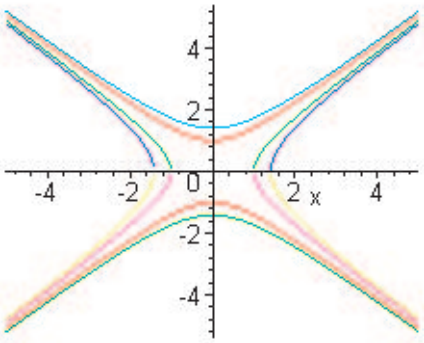
Then this equation implicitly defines the solution $y = f(x)$, as desired.

More Examples of the Method of Separation of Variables: In the rest of the section, we will consider additional examples of solving separable differential equations.

Example 2: We can use the method of separation of variables to solve the differential equation $\frac{dy}{dx} = \frac{x}{y}$.

$$\begin{aligned} y\,dy &= x\,dx \\ \int y\,dy &= \int x\,dx \\ \frac{y^2}{2} &= \frac{x^2}{2} + C \\ y^2 - x^2 &= C_1 \end{aligned}$$

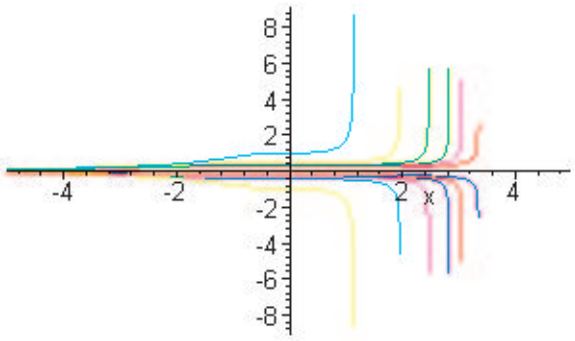
The solution curves are hyperbolas. We can't really go any further unless we knew, say, a point that the solution curve passed through.



Example 3: Solve the IVP $\frac{dy}{dx} = x^2y^3; y(3) = 1$. Separating the variables and integrating, we get:

$$\begin{aligned} \frac{1}{y^3}\,dy &= x^2\,dx \\ \int \frac{1}{y^3}\,dy &= \int x^2\,dx \\ -\frac{1}{2y^2} &= \frac{x^3}{3} + C \end{aligned}$$

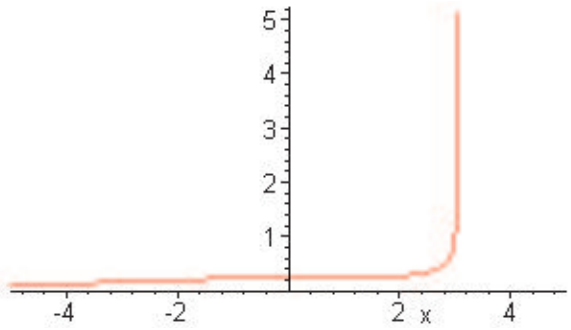
Here are some solution curves:



From $y(3) = 1$, we find the particular solution:

$$\begin{aligned} -\frac{1}{2} &= \frac{27}{3} + C \\ C &= -\frac{19}{2} \\ -\frac{1}{2y^2} &= \frac{x^3}{3} - \frac{19}{2} \\ y^2 &= \frac{1}{19 - \frac{x^3}{3}} \\ y &= \frac{1}{\sqrt{19 - \frac{x^3}{3}}} \end{aligned}$$

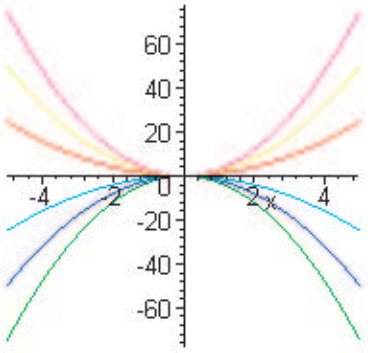
Note that we know that y is the positive square root because we have the initial condition $y(3) = 1$. Here is the particular solution:



Example 4: Solve $\frac{dy}{dx} = \frac{2y}{x}$.

$$\begin{aligned}
\frac{1}{y}\,dy &= \frac{2}{x}\,dx \\
\int \frac{1}{y}\,dy &= \int \frac{2}{x}\,dx \\
\ln|y| &= 2\ln|x| + C \\
\ln|y| &= \ln|x^2| + C \\
|y| &= x^2 e^C \\
y &= C_1 x^2
\end{aligned}$$

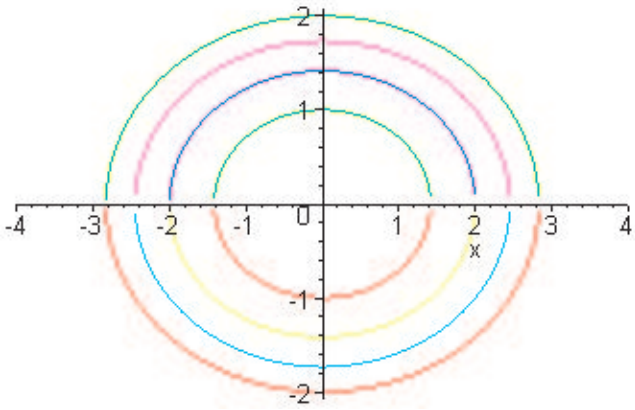
The solution curves are a family of parabolas.



Example 5: Solve $\frac{dy}{dx} = -\frac{x}{2y}$.

$$\begin{aligned}
2y\,dy &= -x\,dx \\
\int 2y\,dy &= -\int x\,dx \\
y^2 &= -\frac{x^2}{2} + C \\
2y^2 + x^2 &= C_1
\end{aligned}$$

The solutions are a family of ellipses:



Example 6: We can also solve Torricelli's equation by the method of separation of variables. We found in Section 2.18 that the equation is of the form $y' = k\sqrt{y}$, where k is a constant. Then we have:

$$\begin{aligned} y^{-\frac{1}{2}}\,dy &= k\,dx \\ \int y^{-\frac{1}{2}}\,dy &= \int k\,dx \\ 2y^{1/2} &= kx + C \\ y^{1/2} &= \frac{1}{2}kx + C_1 \\ y &= \left(\frac{1}{2}kx + C_1\right)^2 \end{aligned}$$

This is the form of the general solution that we explored in the case study of the previous section.

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

3.4 Slope Fields and Euler's Method

In this section we are going to study the geometric information that we get from a differential equation that gives an explicit formula for the derivative. Our intent is to use that information to find a solution of the equation. Consider the differential equation

$$\frac{dy}{dx} = F(x, y); y(x_0) = y_0$$

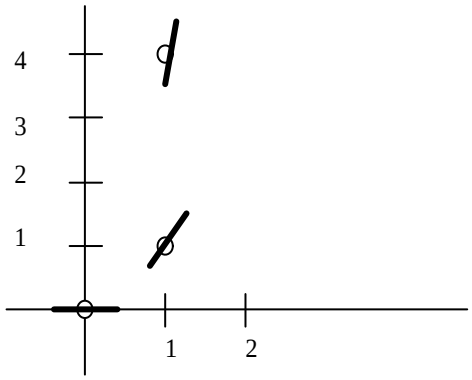
where $F(x, y)$ is a given function of x and y. (For example, $F(x, y)$ might be $8\sqrt{y}$, or it might be $x - y$.) The problem is actually stated in the form of an Initial Value Problem (IVP). We are looking for the particular solution of the equation that passes through the point (x_0, y_0). Assume now that we do not know the solution $y(x)$ and let us interpret what the equation tells us about tangent lines.

The equation says that at any point (x, y) in the plane we can compute the slope $\frac{dy}{dx}$ of the tangent line through that point. That is, at each point (x, y) in the plane, we can draw a short straight line whose slope is $F(x, y)$ from the differential equation. The resulting two-dimensional plot of tangent lines is called the *slope field* or *direction field* of the differential equation.

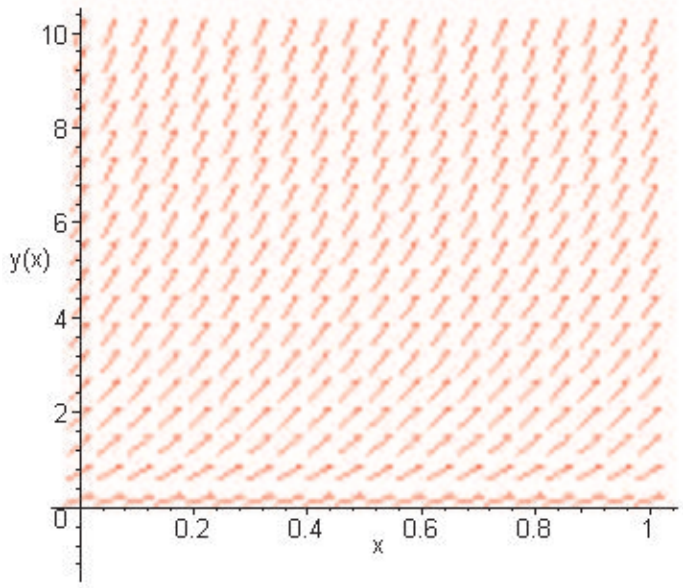
Slope fields are important because sometimes we can guess the shape of the solution curve by sketching a curve that satisfies the given initial condition and follows the slopes of the slope field.

Example 1: Let $F(x, y) = 8\sqrt{y}$. Then we can make a table of points (x, y) and corresponding slopes given by the differential equation $\frac{dy}{dx} = F(x, y)$.

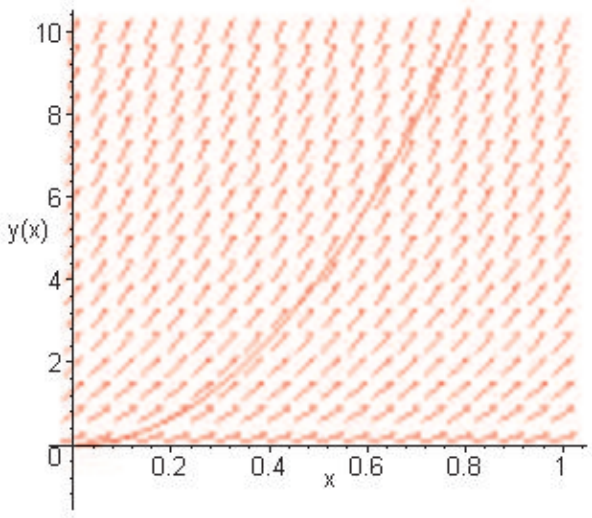
Point (x, y)	Slope $F(x, y)$
$(0, 0)$	0
$(1, 1)$	8
$(1, 4)$	16



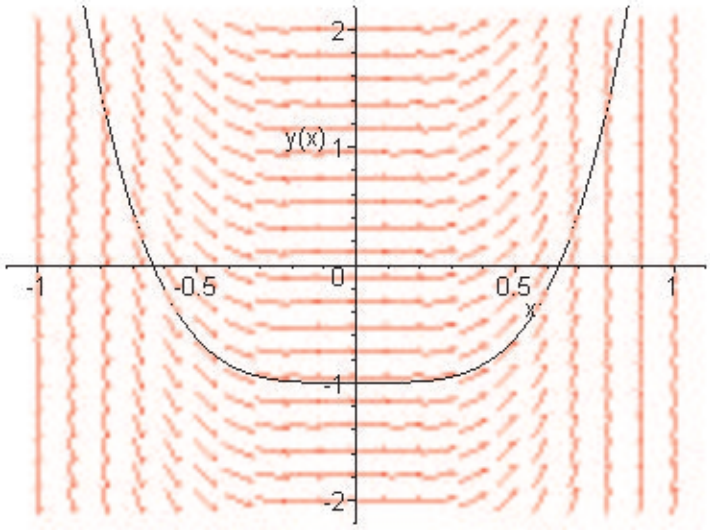
We see from the above example that this can be an exhaustive procedure to carry out by hand. However, it is an excellent task for a computer. Here is a computer display of a slope field for the same equation $\frac{dy}{dx} = 8\sqrt{y}$.



Can you guess the shape of the solution curve that passes through $(0, 0)$? Put your pencil at $(0, 0)$ and see if you can sketch in a curve that *follows the slopes*. That is, when you are done, each line that touches your curve should look like the tangent line at that point.



Example 2: Here is another example of using slope fields to visualize solutions of differential equations. We will consider the differential equation $\frac{dy}{dx} = 24x^3$. The slope field shows a plot of slopes for this equation. The particular solution plotted here is seen to be the curve that passes through the initial point $(0, -1)$ and follows the slope field. (By the way, you can solve the equation analytically. What is the general solution?)



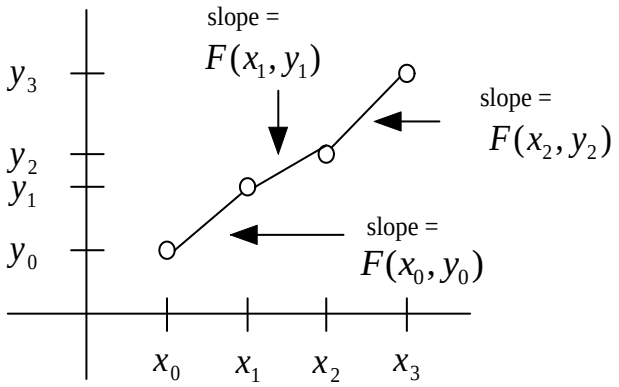
The Next Step: Euler's Method

The above examples suggest a simple way to approximate the desired particular solution numerically. Since the differential equation determines the slope at each point (x, y) of a particular curve, we can approximate a nearby point (x_1, y_1) on the curve by following the tangent line. The resulting procedure is called *Euler's Method*.

Assume that the following IVP is given:

$$\frac{dy}{dx} = F(x, y); P_0 = (x_0, y_0)$$

The method consists of starting at the initial point $P_0 = (x_0, y_0)$, specifying an increment Δx, and plotting a sequence of line segments joined end to end. The slope of each segment is the value of the derivative at the initial point of the segment. We then use the polygonal path as an approximation to the graph of the solution curve through P_0.



Note that the x-coordinates of the points are equally spaced Δx units apart. So, for each n, $x_{n+1} = x_n + \Delta x$. It turns out that we can write a formula for y-coordinates as well.

Theorem 1: Given the Initial Value Problem $\frac{dy}{dx} = F(x, y); P_0 = (x_0, y_0)$, and Δx specified, then the endpoints of the line segments that make up the polygonal path in Euler's Method are

$$\begin{aligned} x_{n+1} &= x_n + \Delta x \\ y_{n+1} &= y_n + \Delta x F(x_n, y_n) \end{aligned}$$

where $n = 0, 1, 2, 3, \ldots$.

The proof is not too difficult. For the nth line segment has equation $y_{n+1} - y_n = F(x_n, y_n)(x_{n+1} - x_n)$. Thus, $y_{n+1} = y_n + F(x_n, y_n)(x_{n+1} - x_n) = y_n + F(x_n, y_n)\Delta x$.

Example 3: Let $\frac{dy}{dx} = x - y; y(0) = 1$. On the interval $[0, 1]$ approximate $y(1)$ with two steps of size $1/2$. Here $F(x, y) = x - y$ and $\Delta x = 1/2$. Thus, $y_1 = y_0 + \Delta x F(x_0, y_0) = 1 + (1/2)(-1) = 1/2$; and $y_2 = y_1 + \Delta x F(x_1, y_1) = 1/2 + (1/2)(0) = 1/2$. Therefore $y(1) \approx 1/2$.

The simplicity of the Euler's Method idea is deceiving. The theorem tells us to start at the initial point and step along successively computing the endpoints of the line segments of the polygonal path. The method and its numerical cousins turns out to be one of the most useful and powerful techniques for exploring solutions of differential equations when exact solutions are too difficult or impossible to obtain. The fact that it may be tedious to generate the points by hand is irrelevant. All we need to do is to call in a Computer Algebra System (e.g., Maple, Mathematica, Derive) or an applet for reinforcements. Then we can decrease the size of Δx and get a better approximation of a desired value of y with very little trouble.

Applet: Euler's Method **Try it!**

Exercises: Problems **Check what you have learned!**

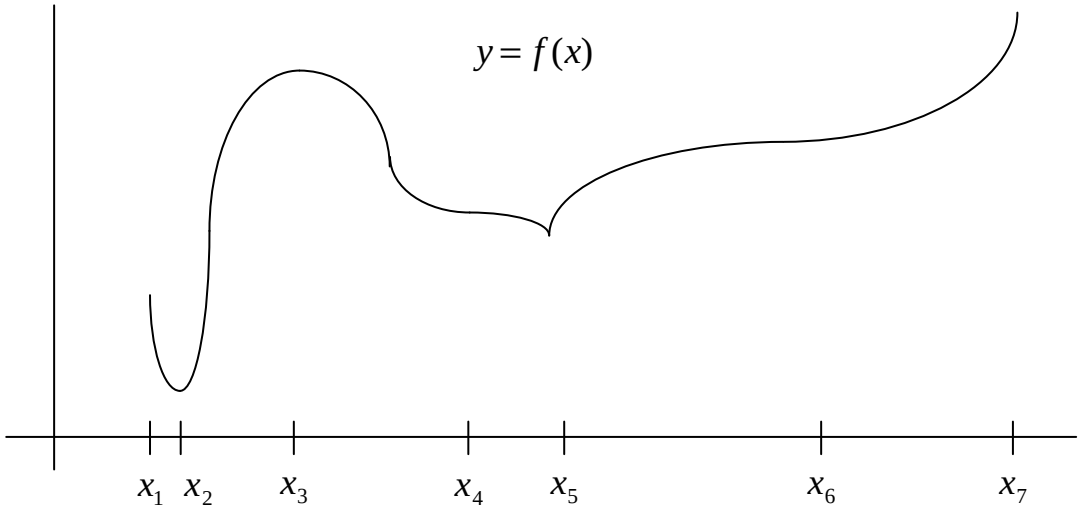
Videos: Tutorial Solutions **See problems worked out!**

3.5 Issues in Curve Sketching

One of the most useful applications of the derivative is in curve sketching. Roughly speaking, given a function defined by a formula, we want to produce its sketch. As we shall see, the first and second derivative are excellent tools for this purpose.

The First Derivative and Extreme Values

Here is the graph of a function:



The function is defined on the interval $[x_1, x_7]$. In what follows, we want to develop the language to describe the high points and the low points of the graph, as well as the general shape. We will start with a discussion of the various kinds of extreme values. Don't be put off by the number of definitions in this section. They give us a vocabulary with which to discuss the concepts, and need to be recorded for easy reference. However, once we get to the examples at the end of the section, you will see that the analyses flow very smoothly.

Definition 1: The function f has an absolute maximum value $f(x_0)$ at x_0 in its domain if $f(x) \leq f(x_0)$ for all x in the domain. The function f has an absolute minimum value $f(x_0)$ at x_0 in its domain if $f(x_0) \leq f(x)$ for all x in the domain.

Example 1: In the graph above, the absolute maximum value occurs at x_7 and the absolute minimum value at x_2. The absolute max and absolute min values are the greatest and least values taken on by the function throughout its domain.

We can see from the graph that the absolute maximum and minimum values do not tell the entire story. There are also local maximum values that correspond to the peaks of the graph, and local minimum values that correspond to the valleys. The concept of *local* is described in terms of neighborhoods.

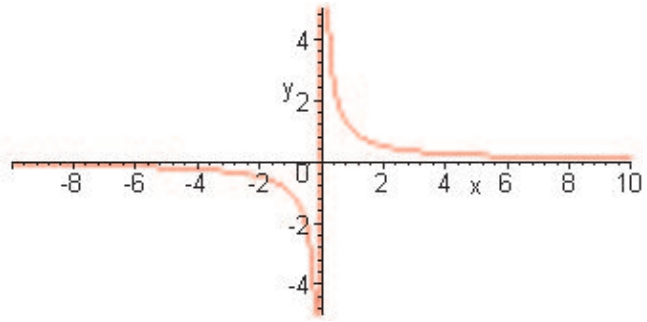
Definition 2: A neighborhood of the point $x = x_0$ is an open interval containing x_0.

Definition 3: The function f has a local maximum value $f(x_0)$ at x_0 in its domain if $f(x) \leq f(x_0)$ for all x in the domain of f in some neighborhood of x_0. The function f has a local minimum value $f(x_0)$ at x_0 in its domain if $f(x_0) \leq f(x)$ for all x in the domain of f in some neighborhood of x_0.

Example 2: With reference to the graph of the function above, there are local maximum values at x_1, x_3, x_7, and local minimum values at x_2 and x_5. Note that the open interval that defines the local neighborhood need not be wholly contained in the domain of the function; this is true in this example at the enpoints of the domain. In fact, we can define an endpoint in this way.

Definition 4: An endpoint of the domain is a point of the domain that does not belong to an open interval contained entirely in the domain.

Example 3: The function $f(x) = 1/x$ has no absolute maximum value, and no absolute minimum value. Nor does it have any local maximum or minimum values. The domain has no endpoints.



Critical Points

When it exists, the derivative can be used to detect points where local maxima and minima occur.

Definition 5: The point x is a critical point if x is in the domain of f and $f'(x) = 0$.

Definition 6: The point x is a singular point if x is in the domain of f but $f'(x)$ is not defined.

The following theorem is a consequence of the Mean Value Theorem. Although we will not prove it, the theorem is extremely important because it will lead to a procedure for finding extreme values.

Theorem 1: Suppose the function f is defined on an interval I and has a local maximum (or local minimum) value at $x = x_0$ in I. Then x_0 must be one of the following:

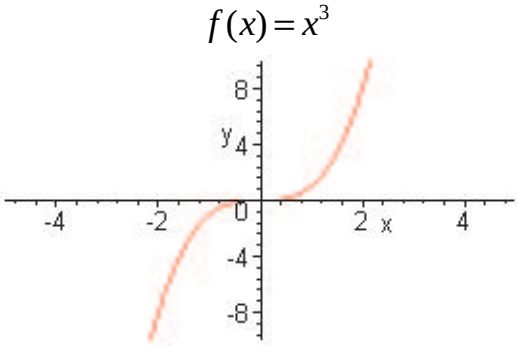
1. x_0 is a critical point of I if $f'(x_0)$ exists [i.e, if $f'(x_0)$ exists, then $f'(x_0) = 0$.]; or
2. x_0 is a singular point of f; or
3. x_0 is an endpoint point of I.

Example 4: Back to the first graph above. The theorem gives a clasification of the points where the local maximum and minimum values occur. Note that $f'(x) = 0$ for $x = x_2, x_3$. There is a singular point at x_5 because the derivative does not exist (the limits of the difference quotients from the left and right are not equal, just as with a corner point). And, of course, there are the endpoints.

Our real objective is to start without a graph of a function, and use the first derivative (and the second derivative) to graph it. Thus, we have to be very careful about what our theorems say.

For example, the theorem does not say that if $f'(x) = 0$, then the function has a local maximum or minimum value at x. Instead, it tells us that the solutions of $f'(x) = 0$ are candidates only. At x_6 in our sketch it looks as though the derivative is zero (horizontal tangent line), but there is neither a local max nor local min there.

Example 5: If $f(x) = x^3$, then $f'(x) = 3x^2$. Hence, $f'(0) = 0$. But the function does not have either a local maximum or local minimum value at $x = 0$.



A continuous function on a closed bounded interval will always have an absolute maximum and an absolute minimum value. The proof of this theorem is beyond the scope of our work and lies in the more advanced subject of real analysis. However, it is important, because together with what we already know about local extrema, it gives us a procedure for finding absolute extreme values on a closed bounded interval.

To find the extreme values (absolute maximum and absolute minimum values) on a closed bounded interval:

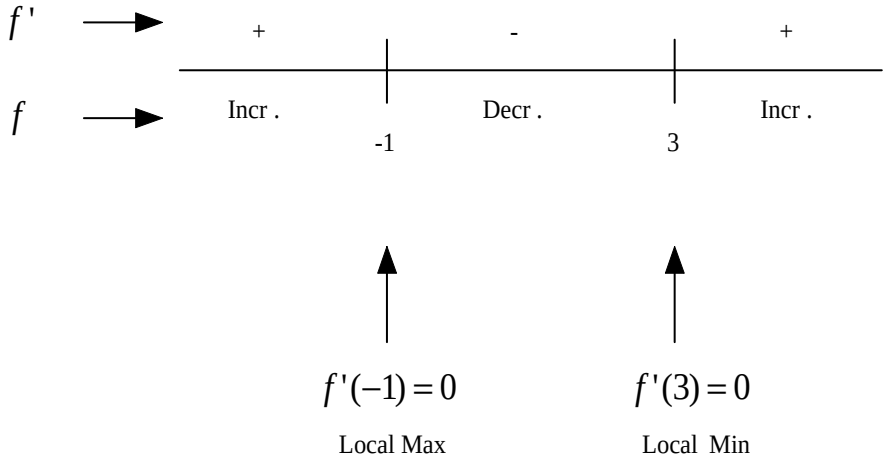
1. Find the critical points (i.e., solve for x in $f'(x) = 0$).

2. Find the singular points (i.e., points x for which $f'(x)$ is not defined).

3. Test the points in 1 and 2, and test the endpoints. The maximum and minimum values will be among them.

First Derivative Test for Local Max/Min: In Example 5 we saw that $f'(x_0) = 0$ does not imply that f has a local max or local min value at x_0. However, suppose that x_0 is an interior point of an interval and $f'(x_0) = 0$. Then if on a neighborhood of x_0, we have that f' is positive to the left of x_0 and negative to the right, then this means that f is increasing as we approach x_0 from the left and decreasing as we continue past x_0 to the right. Hence, the function f has a local maximum at the point x_0. This is the so-called first-derivative test. (Analogously, iff' is negative to the left and positive to the right of x_0, then f has a local minimum at x_0.

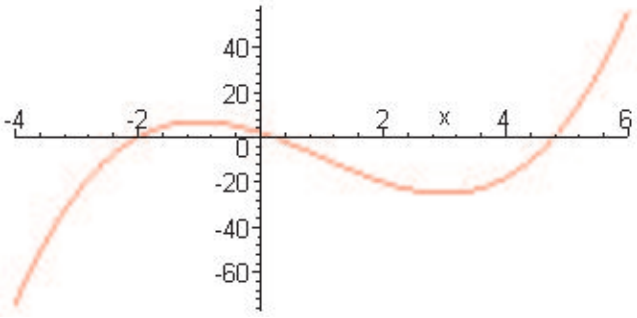
Example 6: Find the maximum and minimum (both local and absolute) values of the function $f(x) = x^3 - 3x^2 - 9x + 2$ on the interval $[-2, 2]$. To solve the problem, we note first that there are no singular points because the function is a polynomial. Hence, we will proceed to find the critical points and to determine which ones give local maxima and minima. If we display a sign table for f', we will be aided in using the first derivative test. Here is the derivative: $f'(x) = 3x^2 - 6x - 9 = 3(x - 3)(x + 1)$.



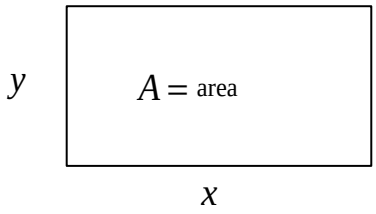
Note that because the function is increasing and then decreasing as it passes through the critical point at $x = -1$, we have a local maximum value there. Analogously, we have a local minimum value at $x = 3$. To find the maximum and minimum values of f on the interval, we make a table of values containing the critical points and the endpoints. Note that there is only one critical point in the interval.

x	$f(x)$
-2	0
-1	7
2	-20

So, the maximum value is 7 and the minimum value is -20 on the interval $[-2, 2]$. Also note that $f(3) = -25$ would be the minimum value on any interval that included 3. [Note: Here is a graph of the function that you can produce on your calculator. We still need some information about the shape of the graph to do as well by hand. We will solve that problem next by using the 2nd derivative.]



Example 7: (Optimization) What dimensions maximize the area of a rectangle of fixed perimeter 1 meter? To solve this problem, let's assign the variables as in the following sketch:

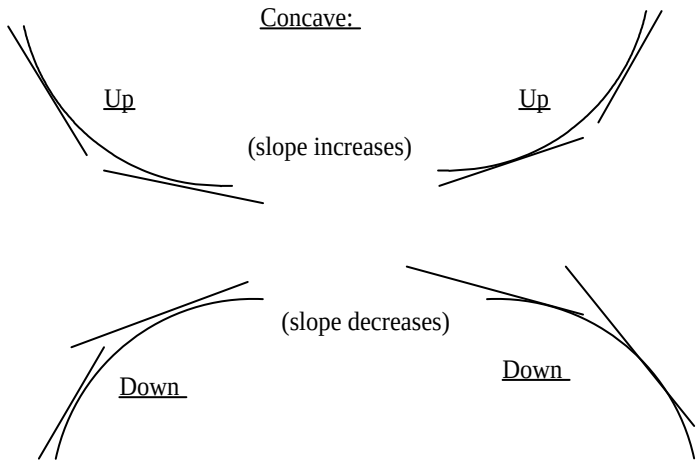


Then $2x + 2y = 1$ implies $y = 1/2 - x$. Thus, $A = xy = x(1/2 - x) = x/2 - x^2$. Now, we calculate the derivative, set it equal to 0, and solve: $A' = 1/2 - 2x$; so, $1/2 - 2x = 0$ gives $x = 1/4$. If we argue that x can be any value in the interval $[0, 1]$, with both endpoints representing degenerate rectangles of no width or no height, and hence of 0 area, then the area is a maximum when $x = 1/4$. That is, when the rectangle is a square, $1/4$ meter on a side.

Applet: Curve Sketching: Increasing/Decreasing **Try it!**

The Second Derivative: Concavity and Inflection Points

Suppose $y = f(x)$ is a given function. If f' is an increasing function on an open interval, then the slope of the tangent line to the graph of f is increasing as we move from left to right. The graph of f thus bends upward, and we call it *concave up*.



Definition 7: The function f is concave up on an open interval if f' exists there and is increasing. Similarly, f is concave down on an open interval if f' exists and is decreasing there.

If you try putting together pieces of the graphs from the above sketch, you can see that at a point where $f'(x) = 0$, the concavity can either remain the same as you move from left to right or it can switch from up to down, or from down to up. We give a special name to the latter situation.

Definition 8: The function f has an inflection point at the point x_0 if $f'(x_0)$ exists and the concacity switches at x_0 from up to down or down to up.

From the fact that a positive second drivative on an interval implies that the first derivative is increasing, we can use f'' to test for concavity as in the next theorem.

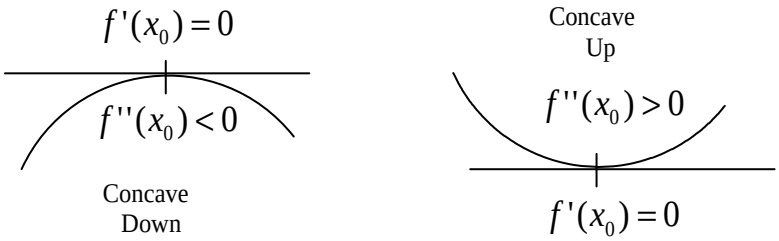
Theorem 2: (2nd Derivative Test for Concavity)

1. If $f''(x) > 0$ on an interval, then f is concave up on the interval.
2. If $f''(x) < 0$ on an interval, then f is concave down on the interval.
3. If f has an inflection point at x_0 and $f''(x_0)$ exists, then $f''(x_0) = 0$.

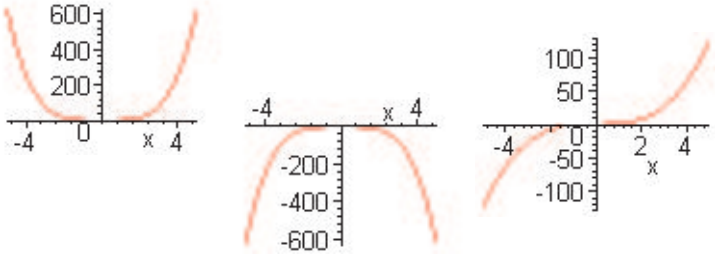
Closely related is the so-called *Second Derivative Test* for local max/min.

Theorem 3: (2nd Derivative Test for Local Max/Min.)

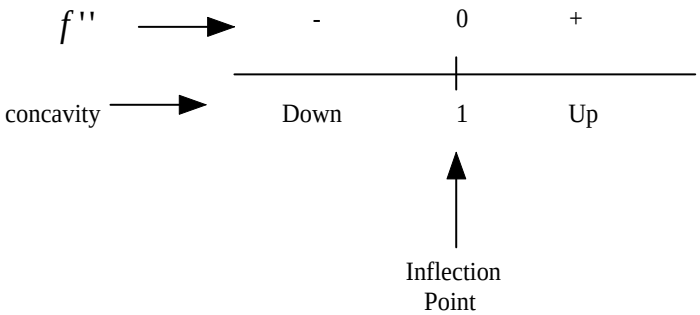
1. If $f'(x_0) = 0$ and $f''(x_0) < 0$ then f has a local maximum at $x = x_0$.
2. If $f'(x_0) = 0$ and $f''(x_0) > 0$ then f has a local minimum at $x = x_0$.
3. If $f'(x_0) = 0$ and $f''(x_0) = 0$ then no conclusion can be drawn. (f may have a local minimum, a local maximum, or an inflection point.)



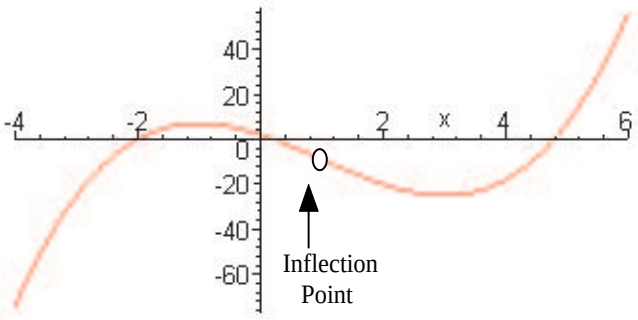
Example 8: To illustrate part 3 of the theorem, note that all three of the following functions satisfy the conditions $f'(x_0) = 0$ and $f''(x_0) = 0$ at $x_0 = 0$: $f(x) = x^4$, $f(x) = -x^4$, $f(x) = x^3$. But x^4 has a minimum at 0; $-x^4$ has a maximum at 0; and x^3 has neither at 0.



Example 9: (Example 6 continued) Find the intervals over which the function $f(x) = x^3 - 3x^2 - 9x + 2$ is concave up, and those where it is concave down. Also, find all points of inflection (if any). We need the second derivative: $f'(x) = 3x^2 - 6x - 9$; $f''(x) = 6x - 6 = 6(x - 1)$. Note that $f''(x) = 0$ implies that $x = 1$ is a candidate for an inflection point. We make a sign table for f''.



Hence, because the concavity switches at $x = 1$, this is indeed an inflection point. Likewise, the graph is concave down on the interval $(-\infty, 1)$ and concave up on the interval $(1, \infty)$.



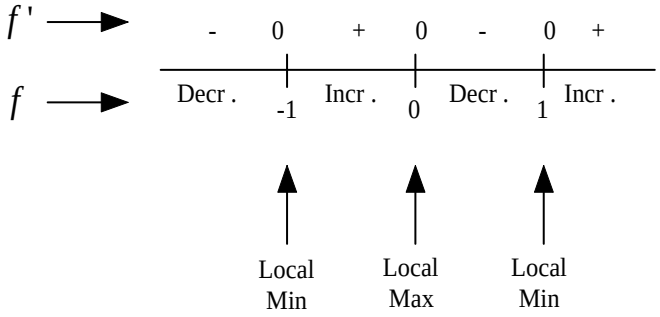
Applet: Curve Sketching: Concavity **Try it!**

Curve Sketching with y' and y'': Putting it all together

We now have enough techniques in hand to sketch the graph of a function using the first and second derivative. This is the goal of the section, and we have finally reached it. Observe in the following examples how straightforwardly the analysis proceeds. But without the foregoing vocabulary and essential ideas, we would not know how to begin. Now we do.

Example 10: Sketch a graph of the function $f(x) = x^4 - 2x^2 - 3$ using f' and f''. We start by computing the first and second derivatives: $f'(x) = 4x^3 - 4x = 4x(x^2 - 1) = 4x(x - 1)(x + 1)$; $f''(x) = 12x^2 - 4 = 4(3x^2 - 1)$.

Increasing/Decreasing: we make a sign table for f'.



Concavity: we make a sign table for f''.

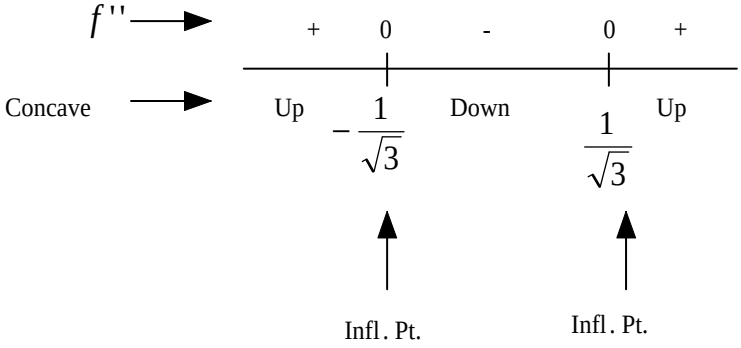
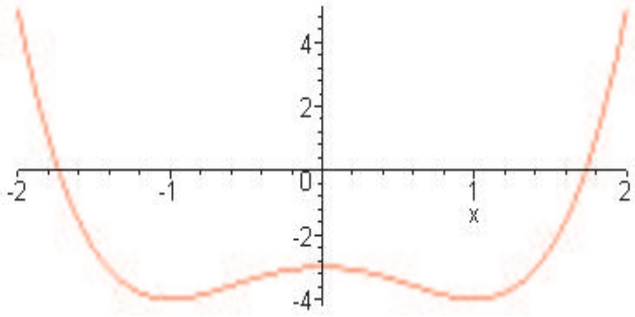


Table of Values:

x	$f(x)$
-1	-4
$-1/\sqrt{3}$	-32/9
0	-3
$1/\sqrt{3}$	-32/9
1	-4

Sketch: (I.e., assemble the above information into a sketch.)



Example 11: Sketch the rational function $f(x) = (x^2 - 1)/(x^2 - 4)$. Find all vertical and horizontal asymptotes.

Zeros: $f(x) = 0 \Rightarrow x^2 - 1 = 0$, or $x = \pm 1$.

Vertical Asymptotes: $f(x) = \frac{(x-1)(x+1)}{(x-2)(x+2)}$ implies vertical asymptotes $x = 2$ and $x = -2$. The limit from the left at $x = 2$ is $-\infty$ and the limit from the right is ∞. The function is symmetric about the y-axis; so the limit from the left at -2 equals ∞ and the limit from the right at -2 is $-\infty$.

Horizontal Asymptotes: $\lim_{x\to\infty} f(x) = 1$. By symmetry, $\lim_{x\to-\infty} f(x) = 1$. Thus, $y = 1$ is the horizontal asymptote.

Increasing/Decreasing:

$$f'(x) = \frac{(x^2-4)2x - (x^2-1)2x}{(x^2-4)^2} = \frac{2x(-3)}{(x^2-4)^2} = -\frac{6x}{(x^2-4)^2}$$

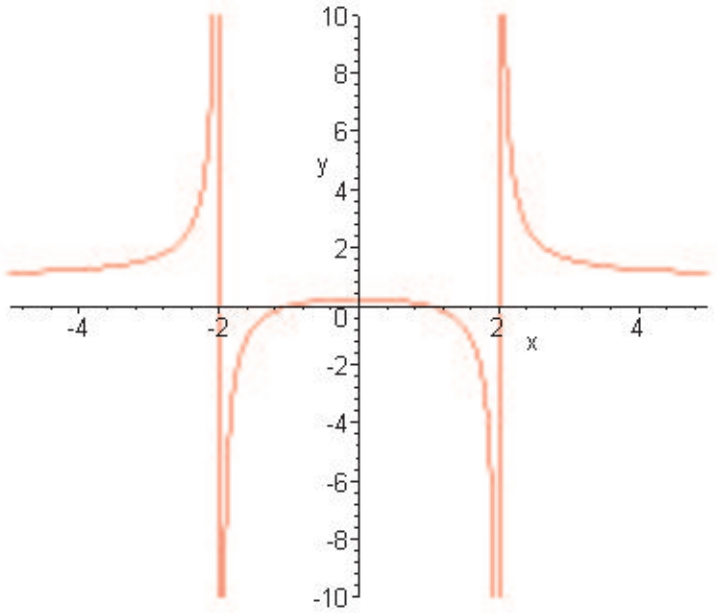
Now, $f'(x) > 0 \Rightarrow -6x > 0$, or $x < 0$. Thus, f is increasing on the interval $(-\infty, 0)$. Similarly, f is decreasing on the interval $(0, \infty)$.

Concavity:

$$\begin{aligned} f''(x) &= \frac{(x^2-4)^2(-6) - (-6x)2(x^2-4)2x}{(x^2-4)^4} \\ &= \frac{(-6)(x^2-4)(x^2-4-4x^2)}{(x^2-4)^4} \\ &= \frac{6(x^2-4)(3x^2+4)}{(x^2-4)^4} \\ &= \frac{6(3x^2+4)}{(x^2-4)^3} \end{aligned}$$

Now, $f''(x) = 0$ has no solutions. Hence, there are no candidates for points of inflection. Note that the numerator of $f''(x)$ is always positive. Thus, the sign of f'' comes from the denominator. A check shows that $f''(x) > 0$ on the intervals $(-\infty, -2)$ and $(2, \infty)$ and hence f is concave up there, and $f''(x) < 0$ on the interval $(-2, 1)$ implying that f is concave down there. Moreover, -2 and 2 are not inflection points because they do not belong to the domain of f.

Here is a sketch:



Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

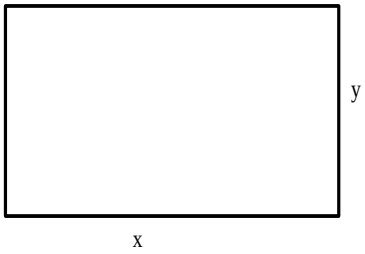
3.6 Optimization

Another application of mathematical modeling with calculus involves word problems that seek the largest or smallest value of a function on an interval. This class of problems is called *optimization* problems. For example, suppose we want to know the dimensions of a rectangle of fixed perimeter, say 1 meter, that maximizes the area. We solved this problem in the last section as an example of optimization and found that the answer is a square, $\frac{1}{4}$ meter on a side. We can outline the steps of a general procedure to follow to solve such problems, but the best way to learn is through practice.

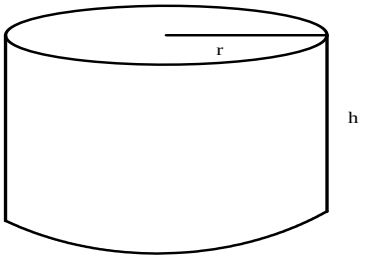
General procedure for solving optimization problems:

1. Begin by making a sketch whenever you can.
2. Define symbols and write down what is given, what is to be found.
3. Write equations that link the variables.
4. Rewrite the quantity to be maximized, or minimized, as a function of a single variable.
5. Take note of the domain over which the optimization is to occur.
6. Treat the function to be maximized, or minimized, much as you would in a sketching problem and find the extreme value(s).

Example 1: Suppose a rectangle has a fixed area of 9 square meters. Find the dimensions that minimize the perimeter. Let x and y be the lengths of the sides, as in the sketch below. Then the area is $9 = xy$, and the perimeter P is given by $P = 2x + 2y$. We want to minimize the perimeter, but it is a function of two variables. However, we can solve for y in the area equation to get $y = \frac{9}{x}$ and substitute it into the perimeter equation. This will give us P as a function of the single variable x. Here are the results: $P = 2x + \frac{18}{x}$, where $x > 0$. Taking the derivative of P and setting it equal to 0, we get: $P'(x) = 2 - \frac{18}{x^2}$; hence $P'(x) = 0$ implies $2x^2 = 18$ or $x = 3$. Using the second derivative test we see that $P''(x) = \frac{36}{x^3}$ which is positive, so $x = 3$ is indeed a minimum (local but also absolute). Thus, the perimeter will be a minimum when the rectangle is a square, 3 meters on a side.



Example 2: Find the dimensions of the 1-liter cylindrical can that can be made from the least amount of tin. This problem is of interest to can-makers who would like to minimize the cost of raw materials. However, if the solution turns out to be too tall and skinny a can, its shape might not be appropriate for the intended use. So, the can actually used in practice may cost a bit more to make than the minimal one. But let's solve the problem and see what the dimensions of the minimal can are. We shall assume that the can is a perfect cylinder without seams. The volume of the can is 1000 cubic centimeters. If r is the radius of the can and h is the height, then $\pi r^2 h = 1000$.



Moreover, the surface area of the can is the sum of the areas of the top, bottom, and side. We can calculate the area of the side by thinking of cutting it open along a line perpendicular to the top and bottom, and laying the side out on a flat surface; it is then a rectangle with one edge of length h and the other edge of length equal to the circumference of the top, namely, $2\pi r$. Thus, the side has surface area $2\pi rh$ and the top and bottom each have area πr^2. So, the total surface area of the can is $A = 2\pi rh + 2\pi r^2$.

The problem is to minimize the surface area A. But as it stands, A is a function of two variables; we need to eliminate one of them. We use the equation $\pi r^2 h = 1000$ to solve for h: $h = 1000/(\pi r^2)$. Substituting into the expression for A yields:

$$A = 2\pi r\left(\frac{1000}{\pi r^2}\right) + 2\pi r^2 = \frac{2000}{r} + 2\pi r^2$$

The domain of A is the positive real numbers. Taking the derivative of A and setting it equal to 0 yields:

$$A' = -\frac{2000}{r^2} + 4\pi r = 0$$

$$-2000 + 4\pi r^3 = 0$$

$$\pi r^3 = 500$$

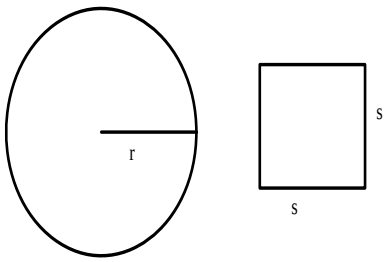
$$r = \sqrt[3]{\frac{500}{\pi}}$$

Using the Second Derivative Test, we see that $A'' = \frac{4000}{r^3} + 4\pi$; hence,

$$A''\left(\sqrt[3]{\frac{500}{\pi}}\right) > 0$$

and so A has a minimum for that value of r. Note that $r = \sqrt[3]{\frac{500}{\pi}} \approx 5.42\ \ cm$ and $h = 1000/(\pi r^2) \approx 10.84\ \ cm$ means that the can will look approximately square in profile.

Example 3: Cut a wire of length L into two pieces and bend one piece to make a square and the other to make a circle. How should you cut the wire so that the sum of the areas of the square and the circle is a minimum? To solve the problem, begin with a sketch:



Then the sum of the areas of the square and circle is $A = s^2 + \pi r^2$. And the sum of the perimeters is L: $L = 4s + 2\pi r$. Next, we substitute for s in A so that A becomes a function of the single variable r: $4s = L - 2\pi r$, or $s = (L - 2\pi r)/4$. Thus,

$$A = \left(\frac{L - 2\pi r}{4}\right)^2 + \pi r^2$$

Now, $r \geq 0$ and $L - 2\pi r \geq 0$ imply that $0 \leq r \leq \frac{L}{2\pi}$; this is the domain of A. Differentiating, we get

$$A' = 2\left(\frac{L - 2\pi r}{4}\right)\left(-\frac{2\pi}{4}\right) + 2\pi r$$

$$A' = -\frac{\pi}{4}(L - 2\pi r) + 2\pi r$$

$$A' = -\frac{\pi}{4}L + 2\pi r\left(1 + \frac{\pi}{4}\right)$$

Setting A' equal to 0, we get

$$r = \frac{\frac{\pi}{4}L}{2\pi\left(1 + \frac{\pi}{4}\right)}$$

$$r = \frac{L}{8\left(\frac{4+\pi}{4}\right)}$$

$$r = \frac{L}{2(4+\pi)}$$

Now, $A(r)$ is defined on the closed interval $\left[0, \frac{L}{2\pi}\right]$. So, to determine the extreme values of A, we need only compute the values of A at the critical point and at the endpoints, and compare them. Note that $r = 0$ corresponds to all of the wire being used for the square, and $r = \frac{L}{2\pi}$ corresponds to all of the wire being used for the circle. With $A(r) = \left(\frac{L-2\pi r}{4}\right)^2 + \pi r^2$ we have:

$$A(0) = \left(\frac{L}{4}\right)^2 = \frac{L^2}{4 \cdot 4}$$

$$A\left(\frac{L}{2\pi}\right) = 0 + \pi\left(\frac{L}{2\pi}\right)^2 = \frac{L^2}{4\pi}$$

$$A\left(\frac{L}{2(4+\pi)}\right) = \left(\frac{L - \frac{\pi L}{4+\pi}}{4}\right)^2 + \pi\left(\frac{L^2}{4(4+\pi)^2}\right) = \frac{L^2}{4(4+\pi)}$$

Thus, $\pi < 4 < 4 + \pi$ shows that

$$A\left(\frac{L}{2\pi}\right) > A(0) > A\left(\frac{L}{2(4+\pi)}\right)$$

Hence, A is a maximum when all of the wire is used for the circle, and A is a minimum when $r = \frac{L}{2(4+\pi)}$. Thus, to minimize the sum of the areas, cut the wire $2\pi r = \frac{\pi L}{4+\pi}$ units from one end.

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

3.7 Case Study: Population Modeling

Animation: Population Modeling **To get you going on the Case Study!**

In a previous section we have discussed how to set up scientific models using differential equations. We have also discussed methods of solving the equations either exactly or numerically. Now that we have some experience with exact techniques such as separation of variables and approximate techniques such as Euler's method, we are going to consider a *Case Study in Calculus* (CSC) that will require us to find both exact and numerical solutions.

As we have said before, the purpose of a CSC is to consider a real application of calculus, with real data. In this section, we will study two approaches to modeling populations in order to:

Objective: Predict the size of the US population well into the 21^{st} century.

In a CSC, we adopt the viewpoint of a mathematician who works in a setting where he or she is expected to find answers to real, everyday questions like the one above. In this particular CSC, a possible scenario is that you are a mathematician working for the Federal Census Bureau. Because the nation needs to know population trends so that it can better predict and plan for future impact on resources, your task is to provide population estimates.

An important purpose of this CSC is to give first-hand experience doing mathematics in an experimental setting. As such, you should keep in mind the **Scientific Method** which in our context takes the form:

1. Translate real-world problems into mathematical models.
2. Subject the models to mathematical analysis and prediction.
3. Draw conclusions from the models.
4. Test the conclusions in the laboratory and compare the results with the original real-world data.
5. Revise the model as necessary and repeat the above steps until the model is a reliable predictor of real-world observations.

There are two approaches below, called *Malthus* and *Verhulst*, modeling the U.S. population based on two different assumptions. The experimental and analytical activities are very parallel, carrying out the same set of investigations, but with the two different models. For each model, there will be a Setup, and a Thinking and Exploring step to complete. But there will be only one final written summary that combines and compares the conclusive interpretations for the two approaches.

The Malthus and Verhulst Models

The Malthus model for growth of a population assumes an ideal environment. Resources are unlimited, disease is constrained, and individuals are happy. The population increases at a rate proportional to the number of individuals present.

In reality populations do not exhibit such unrestricted growth since the density of individuals eventually rises until conditions for living are no longer satisfactory and competition with other individuals in the same population or with other organisms reduces fertility and longevity of the population. Although the initial stages of population growth seem to be exponential, the growth curve eventually levels out and approaches a horizontal asymptote that represents the maximum carrying capacity of the environment. Thus, the Verhulst model, that takes into account the effects of a limiting environment, is a more realistic model.

The CSC will study each of these models in turn and compare them for the US census data. We will use only the populations recorded in the census of 1790 and of 1990.

USA Census (in millions)	
Year	Population
1790	3.9
1990	250

The Malthus Model: Exponential Growth

In this model, we assume uninhibited exponential growth of the population. Thus, starting with a population of 3.9 million in 1790, we have the Initial Value Problem

$$\frac{dQ}{dt} = kQ, \qquad Q(0) = 3.9$$

An explicit statement of our goal is as follows:

Objective: Using actual U.S. population census data, find the growth constant k of the U.S. population during the period 1790 to 1990, assuming exponential growth.

A major purpose of the CSC is to learn to think clearly about such applied problems. As an aid, we have structured the steps of the analysis in the form of a report that you will complete. The main sections of the report are as follows (see homework). We will reach our objective in two different ways, which we describe in the Setup.

Setup: This is what is called the *modeling phase* where we set up the differential equation. We have already done that above. We have written the initial-value problem that models the U.S. population beginning in the year 1790, assuming exponential growth. The initial condition reflects the actual U.S. population (in millions) in the year 1790, which for the sake of this investigation we take to be 3.9 million persons.

In the Thinking and Exploring part below, you will be using Euler's method for solving the IVP we just described. You will carry out a trial and error experiment using several values of the growth constant k seeking a value that leads to a projected population in 1990 that is close to 250 million. Once the experimental process is complete, you then will solve the IVP explicitly by the method of separation of variables and compare this value of k with the experimental one.

We have described what kind of mathematical facts you will be looking for, and the methods that we expect you to use. You are not to carry out the investigation as part of your setup since this is part of Thinking and Exploring below.

Thinking and Exploring: First, use Euler's method to find k experimentally so that the population in 1990 is 250 million people. Then use the result to predict the population in 2100. Finally, solve the IVP by the method of separation of variables and find k explicitly.

Applet: Euler Population Predictions **Try it!**

The Verhulst Model: Limited Exponential Growth

Various alternative models of population growth have been suggested. In most of them the growth rate of the population is assumed to be dependent on population size rather than constant. One especially well-known model was proposed by Verhulst in 1838 to describe the growth of human populations. The same model was independently used by Pearl and Reed (1920) to describe the growth of the U.S. population. The Verhulst model assumes that the growth rate declines, from a value k when conditions are very favorable, to the value 0 when the population has increased to the maximum value M that the environment can support. Specifically, it replaces the growth constant k by the expression

$$k\frac{M - Q(t)}{M}$$

Note that when $Q(t)$ is small (relative to M) this expression is close to k, and as $Q(t)$ approaches M it becomes close to 0. This leads to the differential equation

$$\frac{dQ}{dt} = k\frac{M - Q}{M}Q$$

The factor $\frac{M-Q}{M}$ that has a value between 0 and 1 is sometimes called the *unrealized potential for population growth.* When Q is small it has a value close to 1, and the growth of the population is essentially exponential.

As Q approaches its asymptotic limiting value, however, the factor $\frac{M-Q}{M}$ is close to zero, and the population grows ever more slowly.

Here is the specific objective we will be pursuing during the Verhulst part of the CSC, followed by the setup, and Thinking and Exploring.

Objective: The U.S. population cannot sustain exponential growth indefinitely. The Malthus model gives unrealistic projections of the population over the next century. We would like to use the Verhulst model instead to make such projections. Some demographers and environmental analysts have proposed $M = 750$ million as the asymptotic limit of the U.S. population. Accepting this value, we would like to find the constant k in the Verhulst equation that models the growth of the population from 1790 to 1990, and then project the model over the next several centuries.

Setup: The differential equation is given above. We also need to assume that $Q(0) = 3.9$ million, and $M = 750$ million, the maximum value of the population ($0 \leq Q(t) \leq M$). The rest of the setup is the same as the setup for Malthus. That is, under Thinking and Exploring below, you will be using Euler's method to find k experimentally so that the population in 1990 is 250 million, and then you will be solving the differential euqation explicitly (by separation of variables) and comparing the two values of k.

Thinking and Exploring: Repeat the steps that you carried out for Malthus under Thinking and Exploring. That is, first implement Euler's Method for the Verhulst equation, and carry out the actual experimentation of running the program with different values of the constant k. Your objective, as before, is to match the known U.S. population of 250 million in the year 1990. Next, using this value of k, project the U.S. population, using the Verhulst model, in the year 2100, the year 2200, and for a few centuries beyond that. Finally, show how the same problem can be handled algebraically, using the mathematical solution of the IVP (by separation of variables). Project the population in the future year using the algebraic methods. How do they compare with the values obtained experimentally using Euler's method?

Note: In the process of solving the Verhulst differential equation by separation of variables, you will need the decomposition

$$\frac{M}{(M-Q)Q} = \frac{1}{M-Q} + \frac{1}{Q}$$

Verify for yourself that this equation is correct. The right-hand sum is easy to rewrite as the fraction on the left-hand side of the equation. But it is not so obvious that the fraction on the left can be written as the sum on the right if that sum is not known. To decompose the fraction on the left into the sum on the right is the subject of a method called *partial fractions*. It is a general method of integration that is taken up in a second course in calculus. We need only this instance of its use here. In fact, given the decomposition we need not concern ourselves further where it comes from because we can verify independently that the equation is correct. Just show that the right-hand side equals the left.

Combined Comparative Interpretation and Summary

This part of the report will be comparative in nature. Now that you have derived mathematical facts, based on the Malthus and Verhulst models, it is time to interpret and summarize the mathematical results in terms of the original objective to project the U.S. population into the future, based on the two models, and to learn as much as possible from the exercise that might guide public policy in the coming decades.

The written summary is a very important part of the report because mathematicians often have to explain their work to non-mathematicians, and be convincing about the proposed course of action. Pretend that your synopsis is going to appear in the next issue of a magazine such as *Scientific American*. Include enough details so that a reader would learn what the major issues of the report are, and how you went about addressing them. You should take care to write in complete sentences using correct rules of standard English grammar.

You will probably want to comment on how realistic the Malthus and Verhulst models seem to be. Explain why you think the models do, or do not, realistically predict the U.S. population very far into the future. What environmental or social factors might influence the growth of the population, diminishing the usefulness of the Malthus model? What advice do you have for public policy makers? Make the report

interesting, compelling. Ask yourself: Would a policy maker be able to understand it, and feel compelled to follow its recommendations?

Exercises: Problems **Check what you have learned!**

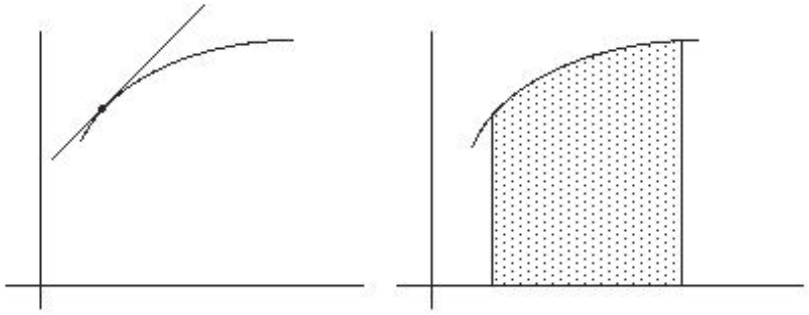
Videos: Tutorial Solutions **See problems worked out!**

Chapter 4

Modeling Accumulations

4.1 Introduction to the Issues

Traditionally, the purpose of calculus is twofold: to find the slope of a curve at a point; and to find the area lying under a curve and above an interval of the x-axis.



We have already dealt with the first problem. Its solution leads to the definition of the derivative. The derivative of a function at a point is then the slope of the tangent line to the graph of the function at the point. Moreover, the original issue of studying the slope of a curve gets transformed into the much more general issue of defining the rate of change of a function. What began as a somewhat restrictive investigation and set of concerns explodes into a set of tools for addressing very general problems in dynamic settings limited only by one's imagination.

The second concern of calculus, that of finding the area under a curve, will also turn out to have very far reaching consequences. But before discussing the generalities, we will begin with some examples of the area problem itself. In this way, we will become familiar with what is at issue, and we will be able to establish an agenda for future work.

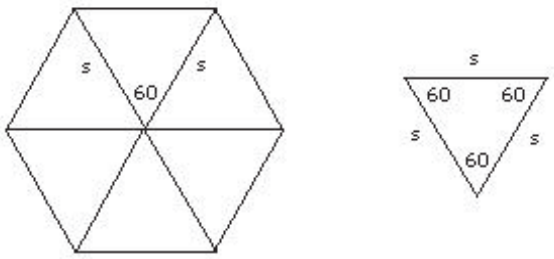
4.1.1 The Area of a Circle

We all know the formula A for the area of a circle: $A = \pi r^2$, where r is the radius and π is the irrational number whose decimal expansion to 20 decimal places begins 3.14159265358979323846. But have you ever stopped to wonder what this all means? That is, what exactly *is* the area of a circle? What is its *definition*?

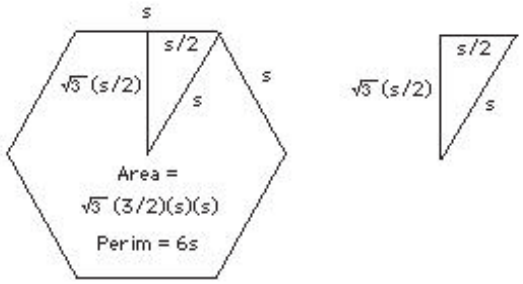
The last question is not easy to answer, is it? Think about the question a bit just to see where you come out. For now, we will postpone an answer and turn to a question we can answer fairly straightforwardly, namely, how do we compute the area of a circle? For circles of radius 1, this is equivalent to asking, how do we compute π?

Consider a circle whose radius is of length one, a so-called *unit circle*. Our task is to compute its area. Archimedes faced this same problem centuries ago, and his methods are still valid today. The trick is to

approximate the area of the circle by that of a geometric figure whose area can be calculated from a simple formula (not the kind that involves a number like π). A hexagon turns out to be an excellent starting place for these purposes.



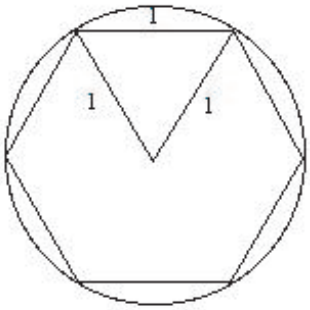
Pictured above is a regular hexagon. Notice that it is composed of six congruent isoseles triangles, each with a 60 (= 360/6) degree central angle. Thus, the base angles of each triangle are also 60 degrees, and the third side has the same length as the other two. Hence, we can find the height of each isoseles triangle and its area using the relationship between the lengths of the legs of a 30-60-90 degree right triangle. With reference to the sketch below, the area of one of the isoseles triangles is $\frac{1}{2}s\frac{\sqrt{3}s}{2}$.



Thus, the area of the hexagon is six times as large, or

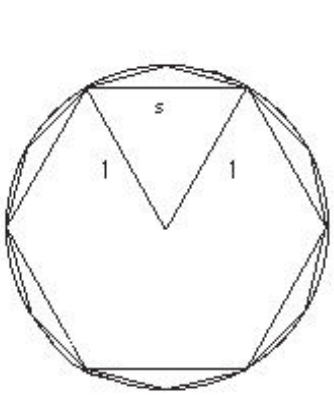
$$\frac{3}{2}\sqrt{3}s^2$$

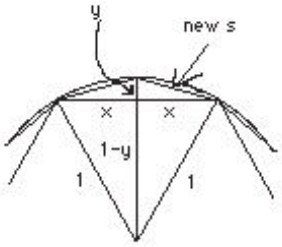
Let's not lose sight of our objective: We want to find the area of a unit circle, and hence the value of π. As a first approximation, we will use the area of an inscribed hexagon. That is, with $s = 1$, the formula we have just derived tells us that the area of the hexagon is $\frac{3\sqrt{3}}{2}$, or approximately 2.598.



This does not give a very good approximation to the area of the circle. We surely can do better, but how? We could replace the hexagon with more completely filling shapes whose areas we can still calculate. One of the best ways to do this is to double the number of edges of the hexagon, thereby obtaining a regular 12-gon; and then to continue doubling repeatedly to obtain in succession a 24-gon, a 48-gon, a 96-gon, and

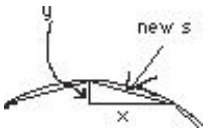
so on. This certainly makes sense from the viewpoint of filling the area of the circle. It also yields figures whose areas can be calculated readily from one stage to the next.





Let's calculate the area of the 12-gon. In the sketch above, $x = s/2$ and from the Pythagorean Theorem we find that $x^2 + (1-y)^2 = 1$. Solving for y in just a few steps yields $(1-y)^2 = 1-x^2$, $1-y = \sqrt{1-x^2}$, or $y = 1 - \sqrt{1-x^2}$. Also, $(new\ s)^2 = x^2 + y^2$, or $new\ s = \sqrt{x^2+y^2}$. Thus, the area of the 12-gon equals the area of the hexagon plus 12 times the area of a little triangle:

$$\frac{3}{2}\sqrt{3} + 12\frac{xy}{2}$$



In fact, we can double the number of sides from 12 to 24 by repeating the steps here with *new s* replacing s and x_1 and y_1 replacing x and y, respectively. The area of the 24-gon is then

$$\frac{3}{2}\sqrt{3} + 12\frac{xy}{2} + 24\frac{x_1 y_1}{2}$$

Summarizing, if we begin with a hexagon inscribed in a circle of radius 1, and obtain a sequence of regular polygons by doubling the number n of sides, then at each stage we obtain a new approximation to the area of the circle by completing the following five steps:

1. $x = s/2$ [To begin $s = 1$.]
2. $y = 1 - \sqrt{1-x^2}$
3. $new\ s = \sqrt{x^2+y^2}$
4. $new\ n = 2n$ [To begin $n = 6$.]
5. $new\ A = A + (new\ n)\frac{xy}{2}$ [To begin $A = \frac{3}{2}\sqrt{3}$.]

Here is a table showing the results of 10 doublings.

Areas of Regular Polygons	
sides	area
6	2.598076
12	3.000000
24	3.105829
48	3.132629
96	3.139350
192	3.141032
384	3.141452
768	3.141558
1536	3.141584
3072	3.141590
6144	3.141592

Thus, we get an approximation to the area of a unit circle, and hence an approximation to π. The best approximation in the table comes from a regular inscribed polygon of 6144 sides. Clearly, we could do even better by doubling the number of sides to 12288, 24576, etc.

Applet: Approximating Areas: Inscribed Polygons **Try it!**

4.1.2 What is the Area of a Circle?

Returning to the earlier question that we postponed during our calculations, do we now know what the area of a unit circle is? You may answer: Of course, it is π. But π is the *value* of the area of a unit circle. We seem to be in the position of knowing the value of something without knowing how to define that something. Note that we don't have this problem for the regular polygons that we have used. Beginning by defining the area of a triangle to be one-half the product of its height and the length of its base, we can define the area of a regular polygon to be the sums of the areas of its component triangles. This is precisely how we calculated the areas of the triangles and polygons above.

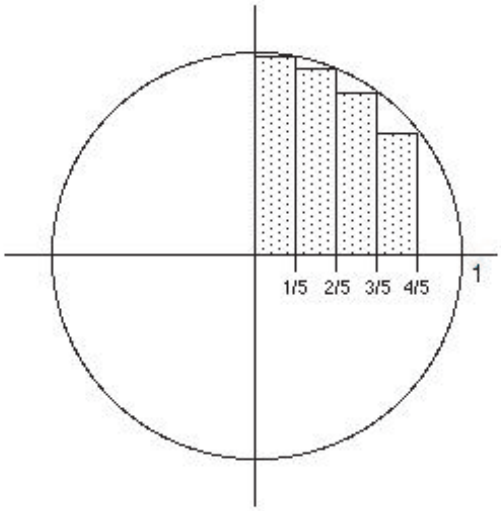
The definition of the area of a circle is not so simple. From what we have done so far, it seems to involve a more elaborate process. In fact, from our work it is reasonable to *define* the area of a unit circle to be the limit of the areas of the inscribed regular polygons that come from starting with a hexagon and doubling the number of sides at each successive stage.

Hold on now, you may say. Do you mean that the area of a circle is tied to a hexagon and the polygons that come from a doubling process? Said another way: What is so special about these geometrical figures? Why not use any collection of shapes that are contained in the circle and fill it in the limit?

4.1.3 Another Calculation of the Area of a Circle

If the questions in the last paragraph could have come from you, you certainly would be justified in your skepticism. Our decision to use an inscribed hexagon and a doubling process was based primarily on two factors: the polygons appear to fill the circle very rapidly; and their areas are relatively easy to calculate. However, we will see that other approaches are just as appealing, perhaps even more so.

For example, suppose we consider approximating the area of a quarter-circle with rectangles, as shown in the sketch. We divide the interval [0, 1] into n subintervals of equal length $h = 1/n$; in the sketch, $h = 1/5$. The circle is the graph of the function $f(x) = \sqrt{1 - x^2}$, where x is between 0 and 1, and the upper right-hand corner of each rectangle lies on the circle. Thus, the area of the rectangle on the subinterval, say, $[3/n, 4/n]$ is $(1/n)f(4/n)$. We then add the areas of the rectangles and use this as an approximation to the area of the quarter-circle. Note that the rectangle on the last subinterval is of zero height, and hence its area is 0.



In the sketch, there are 4 rectangles, and the sum of their areas is $(1/5)[f(1/5)+f(2/5)+f(3/5)+f(4/5)]$. We can increase the number of rectangles, and calculate the value of the sum of the areas of the rectangles for subintervals of shorter and shorter length; that is, for rectangles of narrower and narrower width. As the rectangles get narrower, they come closer to filling the quarter-circle. We will multiply the sums by 4 to obtain an approximation to the area of the full circle. Here are some results.

Approximating Area of Unit Circle with Rectangles	
rectangles	sum of areas times 4
5	2.637049
500	3.137487
1000	3.139555
2000	3.140580
5000	3.141189

With just five rectangles we obtain a rather crude approximation of the area, but as the number of rectangles increases we begin to see the familiar value of π emerge. Certainly the polygon approximations deliver a better approximation to the area with fewer computational steps. But it appears that we could calculate the area of a circle as accurately as we please using a sufficient number of rectangles.

Applet: Approximating Area: Using Rectangles **Try it!**

Although the two methods used to approximate the area of a circle differ, it still seems clear that we could define the area in terms of a limiting process. Whether the limiting process is that of calculating the areas of inscribed polygons with an ever increasing number of sides, or calculating the sum of areas of inscribed rectangles as the width of the rectangles approach zero, the area of the circle can be defined in terms of a circle-filling limiting process involving simple geometrical figures. For calculating purposes it is important only that the areas of the geometrical figures are known and easy to calculate. This is certainly true for the rectangles. Their areas are readily computed because their heights are obtained immediately from the function whose graph is the circle. Let's make note of two central features we have identified in calculating areas and come back to them later: the first is *limit*, the second is *function*.

4.1.4 The Method of Accumulations

We started with the problem of finding the area under a curve and above an interval. Generalizing from the two approaches we have taken to finding the area of a circle, a fruitful approach seems to be to accumulate small pieces that approximate the area, and pass to the limit. In the limit, the area is filled, and hopefully we can then evaluate it as the limit of the values of the areas of the pieces. The process of *passing to the limit* not only provides a calculational tool, but it gives a way to *define* what is meant by the *area under the curve*. Even in the case of a figure as familiar as a circle, this in and of itself is a worthwhile accomplishment.

But area is not the only thing that lends itself to what we will call the *method of accumulations*. For example, suppose instead of the area of a circle we wanted to find the circumference.

4.1.5 The Circumference of a Circle

Archimedes showed that π is between 223/71 and 22/7. He did this by calculating the perimeters of the 96-sided regular polygons inscribed in, and circumscribed about, a circle. Because the circumference c of the circle lies between these two perimeters, and because $c = 2\pi r$, we get an estimate for π.

We can take Archimedes' hint and use the method of accumulations to find the circumference of a unit circle from the perimeters of inscribed polygons. In fact, we have already done most of the work when we found the area by starting with an inscribed hexagon and successively doubling the number of sides. Using our previous notation, *new* n is the number of sides of the polygon at the next stage, and *new* s is the length of a side. Thus, the improved estimate of the circumference is (*new* s)(*new* n), and of π is (*new* s)(*new* n)/2. Here are the results for 10 doublings.

Circumferences of Regular Polygons

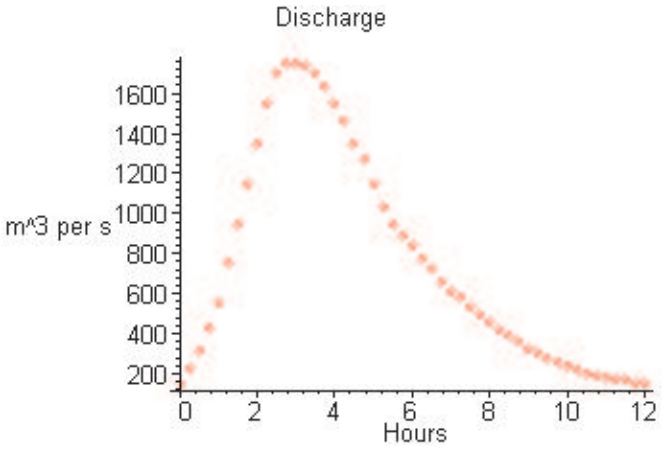
sides	perimeter	π
6	6.000000	3.000000
12	6.211656	3.105829
24	6.265260	3.132629
48	6.278700	3.139350
96	6.282066	3.141032
192	6.282906	3.141452
384	6.283116	3.141558
768	6.283170	3.141584
1536	6.283182	3.141590
3072	6.283182	3.141592
6144	6.283182	3.141593

So, by adding up the lengths of the sides of each polygon and letting the length of the sides get uniformly shorter, we get a progressively better approximation of the circumference of the circle. Again, the quantity we want to calculate (or define) is approached through a limiting process.

4.1.6 The Volume of Water in a River

Even though our examples thus far have been geometric in nature, the method of accumulations is really quite general. Its use often arises in the study of real-world applications. To illustrate this point, let us consider an example from the everyday world of flood forecasting. Suppose we measure over time the flow rate of a river stream. That is, we record at a finite number of times, the number of cubic meters of water that pass a fixed point during one second. Earth Scientists call the volume of water per unit time the *discharge* of a stream; mathematicians would prefer a term like *flow rate*. This is just one example of many in which specialists from different areas use different terminology for the same concept, and may even give different meanings to the same terms. However, as long as we understand both languages, there should be no problem talking to both groups. Since we are describing a problem in earth sciences, we will use their terminology. This also happens to be the terminology used to record the data that you might want to look up in a book or on the World Wide Web for a river of interest.

Prior to a rain storm, the stream will be flowing at some background level of discharge known as *base flow*. However, following a period of heavy precipitation, the rain falling in the watershed drains into the stream, and the discharge increases over time. The discharge of a stream does not rise immediately with the onset of precipitation, rather it takes time to flow across the watershed and into the stream. If the discharge of the stream exceeds the carrying capacity of the channel, the stream overflows its banks and floods. Here is an example of a typical table of discharge data for a 12-hour period. First the graph.

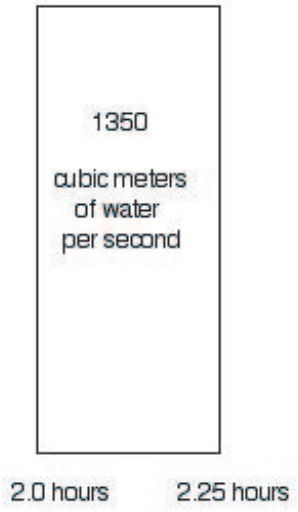


Then the table.

Discharge of a River Stream					
hours	m^3/s	hours	m^3/s	hours	m^3/s
0	150	4	1550	8	460
0.25	230	4.25	1460	8.25	423
0.5	310	4.5	1350	8.5	390
0.75	430	4.75	1270	8.75	365
1	550	5	1150	9	325
1.25	750	5.25	1030	9.25	300
1.5	950	5.5	950	9.5	280
1.75	1150	5.75	892	9.75	260
2	1350	6	837	10	233
2.25	1550	6.25	770	10.25	220
2.5	1700	6.5	725	10.5	199
2.75	1745	6.75	658	10.75	188
3	1750	7	610	11	180
3.25	1740	7.25	579	11.25	175
3.5	1700	7.5	535	11.5	168
3.75	1630	7.75	500	11.75	155
				12	150

There are many questions that earth scientists and regional planners may want to ask about the data and the stream. We will take these up in a more complete form later after we have some additional mathematical tools with which to work. For now, we will limit ourselves to answering one question: What is the volume of water that flowed past the fixed point in the stream during the 12 hours of recorded data?

To answer the question, we will first assume that the discharge is constant over each of the subintervals of time. The sketch below shows the situation for the interval of time from 2 to 2.25 hours, where 1350 cubic meters per second is taken from the table.



The volume of water that flows past the fixed point during this time interval is 1350 (cubic meters per second) times 3600 (seconds per hour) times 0.25 (hours); or, 1,215,000 cubic meters. To get the total volume over the 12 hours, we first find the volume over each subinterval and then, once again, add them together. Here is what we get.

Volume Over Each Subinterval					
subint	m^3	subint	m^3	subint	m^3
0	135000	4	1395000	8	414000
0.25	207000	4.25	1314000	8.25	380700
0.5	279000	4.5	1215000	8.5	351000
0.75	387000	4.75	1143000	8.75	328500
1	495000	5	1035000	9	292500
1.25	675000	5.25	927000	9.25	270000
1.5	855000	5.5	855000	9.5	252000
1.75	1035000	5.75	802800	9.75	234000
2	1215000	6	753300	10	209700
2.25	1395000	6.25	693000	10.25	198000
2.5	1530000	6.5	652500	10.5	179100
2.75	1570500	6.75	592200	10.75	169200
3	1575000	7	549000	11	162000
3.25	1566000	7.25	521100	11.25	157500
3.5	1530000	7.5	481500	11.5	151200
3.75	1467000	7.75	450000	11.75	139500
				12	135000

The sum of the volumes over the subintervals yields the total volume during the twelve hours: 33.1848 million cubic meters of water have passed the given point.

Once again, as we found in our previous examples, taking measurements closer together should yield a more accurate approximation of the total volume. On the other hand, given the nature of the problem, this probably is not necessary. After all, recording the data 15 minutes apart seems demanding enough on earth science personnel as it is.

We want to keep in mind how the example of river flooding differs from the ones we have looked at heretofore. First, it involves real data. Earth scientists really do collect discharge data and they really do use it to make predictions about flooding. Second, the sum, over a period of time, of rates-times-time quantities yields an approximation to the total thing (in this case volume) whose rate of change we have measured. The latter is a key point that we will return to later.

Applet: Accumulation: River Flow **Try it!**
Applet: Accumulation: Distance Traveled **Try it!**

4.1.7 Our Agenda for This Chapter

Calculus has two general lines of development:

1. slope → rate of change → derivative
2. area → method of accumulations → integration

We have discussed the derivative in the last section. In the present section we study the integral. At this point we only know that integration has something to do with the method of accumulation. So, as we plan our agenda for the Chapter, we will begin with a precise definition of the integral and proceed from there. In actuality, there are two but related notions of integral, the so-called *definite integral* and the *indefinite integral* that we already have defined as the *general antiderivative* of a function. Here is what we need to do to understand both integrals.

1. Develop an explicit definition of the definite integral.
2. Study the theoretical properties of integrals; in particular, relate the definite integral to area and accumulation.
3. Develop algebraic rules for finding integrals.
4. Develop numerical techniques for evaluating definite integrals.
5. Surprise: Discover the relationship between derivatives and integrals.

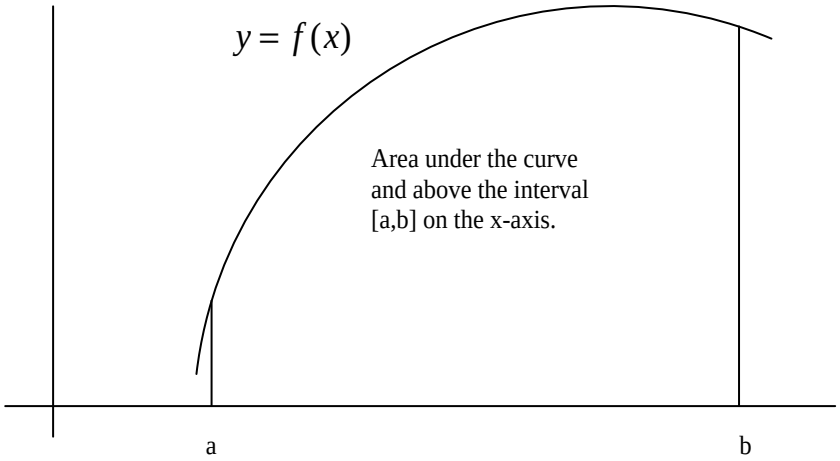
It is hard to motivate the last point in the list until we get more experience with integrals. However, Leibniz is the one who first drew special attention to the connection. It is so important that we call it *The Fundamental Theorem of Calculus*. This theorem makes a beautiful observation that unifies the conceptual framework of the study of calculus by relating derivatives and integrals to each other. Terrific!

Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

4.2 The Definite Integral

Thus far, we have discussed the *Tangent Line Problem*. Its solution led to the definition of the derivative and to the rich array of applications that we have been studying. Now, we are ready to state the other fundamental problem with which calculus deals: *The Area Problem*.

The Area Problem: Find the area of the region in the xy-plane lying above the interval $[a, b]$ on the x-axis and under the graph of the nonnegative continuous function $y = f(x)$.



The problem seems approachable enough. That is, we are all familiar with areas and know how to calculate them for some basic geometric figures. In fact, before we go very far let's list three assumptions about areas that we can all can agree to.

Assumptions about Areas:

1. Area is a nonnegative number.

2. The area of a rectangle is its length times its width.

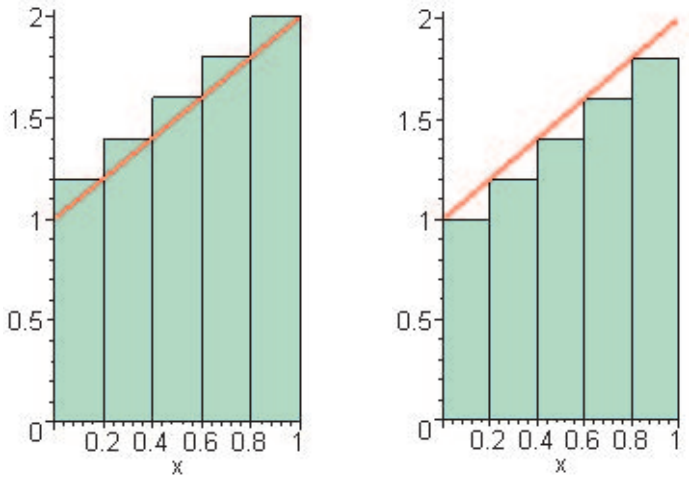
3. Area is additive. That is, if a region is completely divided into a finite number of non-overlapping subregions, then the area of the region is the sum of the areas of the subregions.

We probably do not need to say much about these assumptions. We have all computed the area of a square and a rectangle, and have subdivided regions into smaller regions whose areas we then added willy-nilly as the situation required to find the original area.

Returning to the Area Problem, because rectangles are convenient, our approach will be to use them to approximate the area of the region in question. We will start by considering an example and introducing some terminology.

Upper and Lower Sums; the Method of Exhaustion

Example 1: Suppose we want to use rectangles to approximate the area under the graph of $y = x + 1$ on the interval $[0, 1]$. Here are two possible ways to do it, as illustrated in the sketches.



In both sketches, we use five rectangles, but in the left picture, the area of the rectangle on each subinterval exceeds the area under the graph, while in the right picture the area of each rectangle is less than that of the corresponding subregion. The triangles at the top of each rectangle represent the amount by which we go over or fall short of the area of the region under the graph. We will call the sum of the areas of the rectangles in the left picture an *Upper Riemann Sum*, and the sum of the areas of the rectangles in the right picture a *Lower Riemann Sum*. Let's calculate these quantities.

Each rectangle is of width 0.2. In the Upper Sum, the height of each rectangle is f evaluated at the right endpoint of the subinterval; in the Lower Sum the heights are f evaluated at the left endpoint of the subinterval). Upper Sum $= .2f(.2)+.2f(.4)+.2f(.6)+.2f(.8)+.2f(1) = \frac{8}{5}$. Lower Sum $= .2f(0)+.2f(.2)+.2f(.4)+.2f(.6)+.2f(.8) = \frac{7}{5}$. Now, in the example we are considering, the region is the sum of a rectangle and a triangle. So, we know that the exact area is $1+\frac{1}{2} = \frac{3}{2}$. And of course, $\frac{3}{2} = \frac{15}{10}$ is between $\frac{7}{5} = \frac{14}{10}$ and $\frac{8}{5} = \frac{16}{10}$.

Note that we can get a better approximation to the area under the graph by using rectangles of smaller width. For example, if we double the number of recatngles to 10 so that the width of each rectangle is 0.1, then the Upper Sum $= \frac{31}{20}$ and Lower Sum $= \frac{29}{20}$.

The process of increasing the number of rectangles to improve the approximation to the area whose value we seek is reminiscent of the *Greek Method of Exhaustion*. The inventors of calculus asked: Instead of stopping with a finite number of rectangles, why not take the limit of the sum of the areas of the rectangles as their widths approach 0? This should yield the exact value of the area, if (as always, an important proviso) the limit exists. Why not, indeed!

Let n stand for the number of rectangles, U for the Upper Riemann Sum, and L for the Lower Riemann Sum. Here are some values for the same example we have been discussing. You can compare these approximations with the exact value of 1.5:

n	U	L
100	1.505000000	1.495000000
150	1.503333333	1.496666667
200	1.502500000	1.497500000
300	1.501666667	1.498333333
500	1.501000000	1.499000000

General Procedure for finding the Area Under a Curve and Above an Interval: The above example suggests the following procedure for calculating the area under a curve.

1. Let $y = f(x)$ be given and defined on an interval $[a, b]$. Subdivide the interval $[a, b]$ into n subintervals. Label the endpoints of the subintervals $a = x_0 \leq x_1 \leq x_2 \leq x_3 \cdots \leq x_n = b$. Define $P = \{x_0, x_1, x_3, \ldots, x_n\}$ to be a *partition* of $[a, b]$.

2. Let $\Delta x_i = x_i - x_{i-1}$ be the width of the i^{th} subinterval, $1 \leq i \leq n$.

3. Form the Upper Riemann Sum $U(P, f)$: the height of each rectangle is the *maximum* value M_i of the function on that i^{th} subinterval.

$$U(P, f) = M_1\Delta x_1 + M_2\Delta x_2 + M_3\Delta x_3 + \cdots + M_n\Delta x_n$$

4. Form the Lower Riemann Sum $L(P, f)$: the height of each rectangle is the *minimum* value m_i of the function on that i^{th} subinterval.

$$L(P, f) = m_1\Delta x_1 + m_2\Delta x_2 + m_3\Delta x_3 + \cdots + m_n\Delta x_n$$

5. Take the limit as $n \to \infty$ and the maximum $\Delta x_i \to 0$.

We have that $L(P, f) \leq$ Area $\leq U(P, f)$. So, if the limit of the Upper Riemann Sums and the limit of the Lower Riemann Sums approach a common value, **this number is defined** to be the **area** under the curve and above the interval $[a, b]$.

Sigma Notation

From our discussion of the example above, we seem to have defined a working procedure to find the area of a region lying above an interval of the x-axis and under the graph of a function. But before going further, the process can be facilitated by introducing some useful notation.

Sigma Notation: If m and n are integers with $m \leq n$, and if f is a function defined on the integers from m to n, then the symbol $\sum_{i=m}^{n} f(i)$, called sigma notation, is defined to be $f(m) + f(m+1) + f(m+2) + \cdots + f(n)$.

So, sigma notation is just a way of writing the sum in a compact form. (The word *sigma* comes from the Greek letter Σ.)

Example 2: Here are three examples:

1. $\sum_{i=1}^{n} i = 1 + 2 + 3 + 4 \cdots + n$
2. $\sum_{i=1}^{n} i^2 = 1^2 + 2^2 + 3^2 + 4^2 \cdots + n^2$
3. $\sum_{i=1}^{n} 1 = \underbrace{1 + 1 + 1 + 1 \cdots + 1}_{ntimes}$

Note that we can evaluate the sums in the above example by simply adding the numbers.

Example 3: If we add 1 to itself n times, the sum is n. So, $\sum_{i=1}^{n} 1 = n$. For instance, $\sum_{i=1}^{5} 1 = 1 + 1 + 1 + 1 + 1 = 5$.

Also,

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

For instance, $\sum_{i=1}^{5} i = 1 + 2 + 3 + 4 + 5 = 15 = 5 \cdot 6/2$. This sum is often referred to as the *Gauss sum* because when he was a young school boy, the mathematical genius Gauss was able to solve the problem of adding the first 100 numbers for his teacher in lightning speed. Here is probably the way he did it:

1	2	3	4	5	$\cdots$	98	99	100
100	99	98	97	96	$\cdots$	3	2	1
101	101	101	101	101	$\cdots$	101	101	101

That is, if you write the numbers first from 1 to 100, then in reverse order from 100 to 1, and add them, you get 100 times 101. But this is twice the answer you want, so you must divide by 2. Hence, the answer that you want is $(100 \cdot 101)/2 = 5050$. Pretty clever! Notice that there is nothing special about 100. We could prove the result we have stated for n in an analogous way, but we won't bother with the details here.

We will simply state the result for the sum of the squares of the first n numbers:

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

Example 4:

$$\begin{aligned}\sum_{i=1}^{10}(3i^2+2i+1) &= 3\sum_{i=1}^{10} i^2 + 2\sum_{i=1}^{10} i + \sum_{i=1}^{10} 1 \\ &= \frac{3 \cdot 10 \cdot 11 \cdot 21}{6} + \frac{2 \cdot 10 \cdot 11}{2} + 10 \\ &= 1275\end{aligned}$$

The Area Problem Revisited

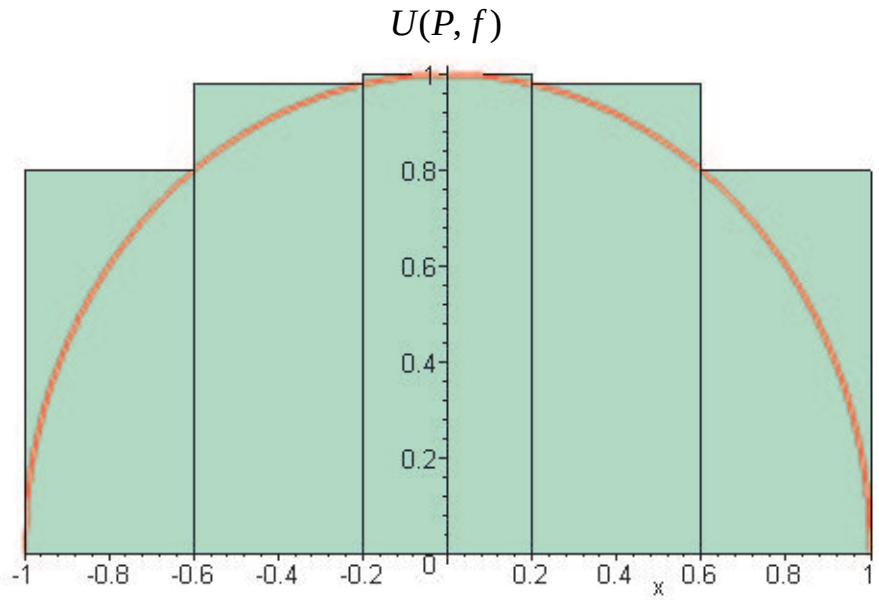
So far as an example we have considered a region whose top boundary is a line. And based on that example, we have outlined some fairly general procedures. Let's approximate the area of another region that we recognize just to be sure that we are going in the right direction. We will also make use of the terminology we have introduced. Note that in sigma notation the Upper and Lower Riemann sums can be stated compactly as

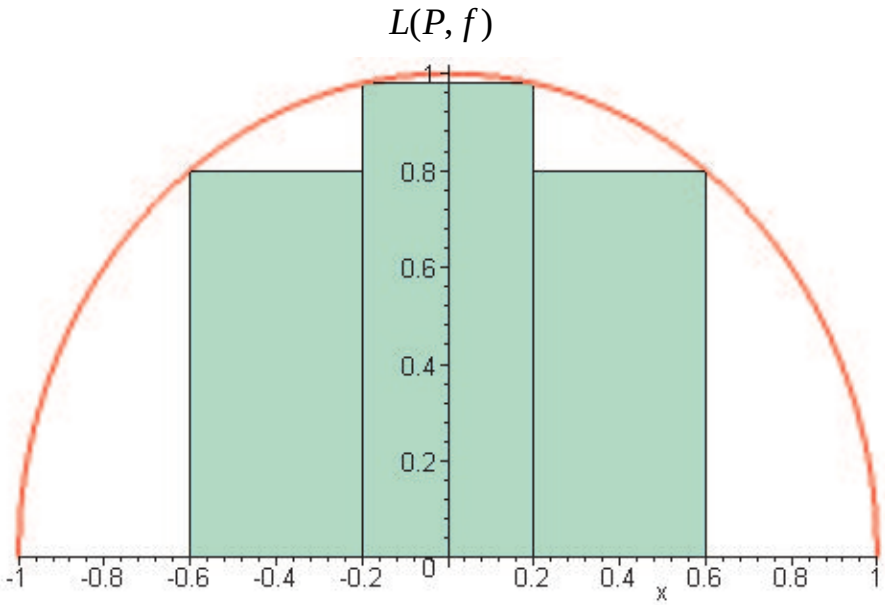
$$U(P,f) = \sum_{i=1}^{n} M_i \Delta x_i$$

$$L(P,f) = \sum_{i=1}^{n} m_i \Delta x_i$$

where M_i and m_i are, respectively, the maximum and minimum values of f on the i^{th} subinterval $[x_{i-1}, x_i]$, $1 \leq i \leq n$.

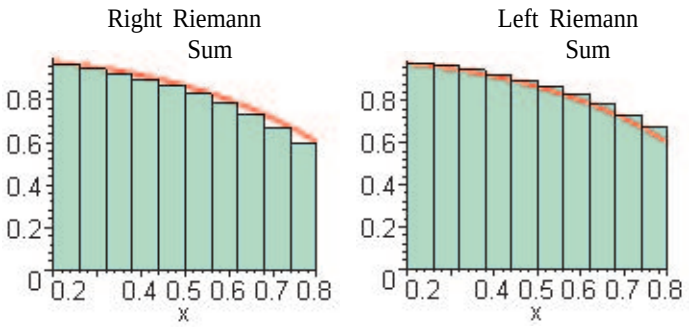
Example 5: Let $f(x) = \sqrt{1-x^2}$ on the interval [-1,1]. Using 5 subintervals of equal width, find $U(P,f)$ and $L(P,f)$. To solve this problem, first note that the region is bounded by a semicircle and the x-axis. In this example, we are given that the widths of the rectangles are all the same, namely, $\Delta x = 2/5 = 0.4$. To form the Upper Riemann Sum, we use the maximum height of the rectangle on each subinterval. So, we get that $U(P,f) = \Delta x(f(-.6) + f(-.2) + f(0) + f(.2) + f(.6)) \approx .4(4.559591794) \approx 1.823836718$. The Lower Riemann Sum uses the minimum height of the rectangle on each subinterval. Thus, $L(P,f) = \Delta x(f(-.6) + f(-.2) + f(.6)) \approx .4(2.579795897) \approx 1.031918359$. The actual value of the area is $(\pi \cdot 1^2)/2$ which is approximately 1.570796327. Once again, we would expect that as we let $\Delta x \to 0$, we would get a better and better approximation of the area under the graph.





There are two other Riemann Sums that are convenient to use because their formulas do not depend on the characteristics of the function. Given a partition P of $[a, b]$, $P = \{a = x_0, x_1, x_3, \ldots, x_n = b\}$, and $\Delta x_i = x_i - x_{i-1}$ the width of the i^{th} subinterval, $1 \leq i \leq n$; let f be defined on $[a, b]$. Then the **Right Riemann Sum** is $\sum_{i=1}^{n} f(x_i)\Delta x_i$ and the **Left Riemann Sum** is $\sum_{i=0}^{n-1} f(x_i)\Delta x_i$

The left Riemann Sum uses the left endpoint of each subinterval to determine the height of the rectangle on that subinterval, while the Right Riemann Sum uses the right endpoint. Note that if f is decreasing on $[a, b]$, then $U(P, f)$ is a Left Riemann Sum, and $L(P, f)$ is a Right Riemann Sum. Similar comments apply to increasing functions.



Example 6: Approximate the area of the region under the graph of the function $f(x) = \sqrt{1-x^2}$ on the interval [0,1] using a Left and a Right Riemann Sum first with 5 rectangles of equal width, then with 100. Note that because the function that describes the quarter-circle is decreasing on [0,1], the question asks us to find the Upper and Lower Riemann Sums for 5, and then 100 subintervals of equal width. A big advantage to the left and right Riemann sums is that their formulas are easily programmed into a programmable calculator or a computer. In this example, in the case of 5 rectangles, $x_i = 0 + i/5, 0 \le i \le 5$, and we want to find the Left Riemann Sum

$$\frac{1}{5}\sum_{i=0}^{4}\sqrt{1-x_i^2} = \frac{1}{5}\sum_{i=0}^{4}\sqrt{1-\frac{i^2}{25}}$$

and then the Right Riemann Sum

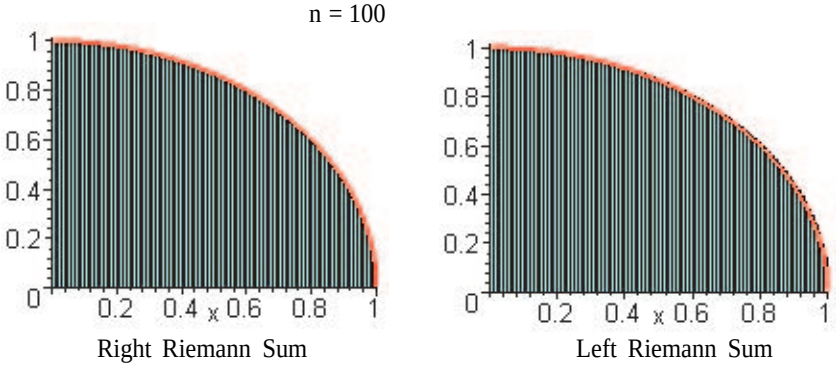
$$\frac{1}{5}\sum_{i=1}^{5}\sqrt{1-x_i^2} = \frac{1}{5}\sum_{i=1}^{5}\sqrt{1-\frac{i^2}{25}}$$

These evaluate to Left Riemann Sum $\approx .8592622072$ and Right Riemann Sum $\approx .6592622072$. The actual value is $\pi/4 \approx .7853981635$.

With 100 subintervals, we can get even closer to $\pi/4$ by evaluating the sums

$$\frac{1}{100}\sum_{i=0}^{99}\sqrt{1-x_i^2} = \frac{1}{100}\sum_{i=0}^{99}\sqrt{1-\frac{i^2}{100^2}} \approx .7901042579$$

$$\frac{1}{100}\sum_{i=1}^{100}\sqrt{1-x_i^2} = \frac{1}{100}\sum_{i=1}^{100}\sqrt{1-\frac{i^2}{100^2}} \approx .7801042577$$



Applet: Riemann Sums **Try it!**

The Definite Integral

Believe it or not, we almost have the definition of the definite integral in hand. We will state it formally so that we can refer to it conveniently as needed.

Definition: Let P be a partition of the interval $[a, b]$, $P = \{x_0, x_1, x_2, \ldots, x_n\}$ with $a = x_0 \leq x_1 \leq x_2 \leq \cdots \leq x_n = b$. Let $\Delta x_i = x_i - x_{i-1}$ be the width of the i^{th} subinterval, $1 \leq i \leq n$. Let f be a function defined on $[a, b]$. Next form the Upper Riemann Sum $U(P, f)$ where the height of the rectangle on each subinterval is the maximum value of f on that subinterval; and form the Lower Riemann Sum $L(P, f)$, where the height of the rectangle on each subinterval is the minimum height of f on that subinterval. Then we say that f is Riemann integrable on $[a, b]$ if there exists a number Φ such that $L(P, f) \leq \Phi \leq U(P, f)$ for all partitions of $[a, b]$. We write the number Φ as

$$\Phi = \int_a^b f(x)dx$$

and call it the definite integral of f over $[a, b]$.

The integral symbol is a stylized Greek sigma Σ from the summation notation we introduced above. The x is a so-called *dummy* variable in that it merely tells us the variable with respect to which we are integrating; hence, we could equally well write $\int_a^b f(t)dt$ or $\int_a^b f(r)dr$.

The definition looks a bit awkward to verify. However, there are two important theorems that come to our aid from advanced analysis, and which we rely on in practice.

Theorem 1: If f is Riemann integrable on $[a, b]$, then

$$\int_a^b f(x)dx = \lim_{\substack{n\to\infty \\ ||P||\to 0}} \sum_{i=1}^{n} f(c_i)\Delta x_i$$

where c_i is any point in the subinterval $[x_{i-1}, x_i]$, and $||P||$ is the maximum length of the Δx_i.

So, the Upper Riemann Sum, the Lower Riemann Sum, the Left Riemann Sum, and the Right Riemann Sum are all special cases of the sum in the above limit where we choose the points c_i in very particular ways. (That is, where f is a maximum, or a minimum, or the left endpoint, or the right endpoint, respectively.) In the examples, we usually take the subintervals to be of equal length, so as $n \to \infty$, the length of each subinterval automatically goes to 0.

The above theorem is much more than a theoretical result. We will see that we use it extensively in applications as a guide in setting up a mathematical model connected with the problem. But more about that later. The theorem below allows us to work effectively with the integral because most of the functions in which we will be interested are continuous or piecewise continuous.

Theorem 2: If f is continuous on $[a, b]$, then f is Riemann integrable on $[a, b]$.

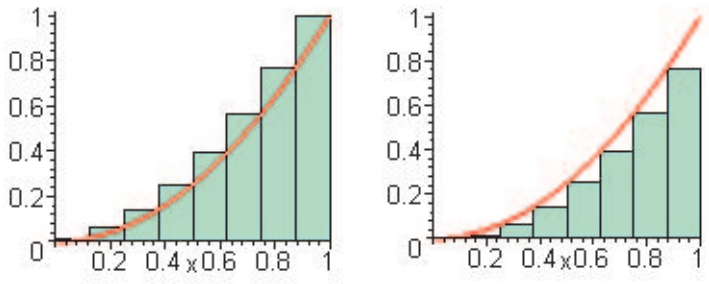
This theorem tells us that for continuous functions, we can use the limit of any convenient Riemann sums to evaluate the integral.

Example 7: Use an Upper Riemann Sum and a Lower Riemann Sum, first with 8, then with 100 subintervals of equal length to approximate the area under the graph of $y = f(x) = x^2$ on the interval [0, 1].

First with 8 subintervals:

$$U(P, f) = \frac{1}{8}\sum_{i=1}^{8} \frac{i^2}{64} \approx .3984375$$

$$L(P, f) = \frac{1}{8}\sum_{i=0}^{7} \frac{i^2}{64} \approx .2734375$$



Then with 100 subintervals:

$$U(P,f) = \frac{1}{100}\sum_{i=1}^{8} \frac{i^2}{10000} = .33835$$

$$L(P,f) = \frac{1}{100}\sum_{i=0}^{7} \frac{i^2}{10000} = .32835$$

Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

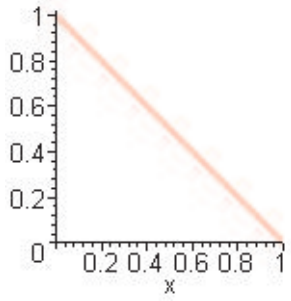
4.3 Properties of the Definite Integral

In the last section, we saw that if f is a nonnegative function on $[a, b]$, then the definite integral $\int_a^b f(x)dx$ is the area of the region under the graph of f and above the interval $[a, b]$. In fact, for most functions the definite integral *defines* the area under the graph. Before we consider some examples, let's give an obvious property of definite integrals but one worth noting.

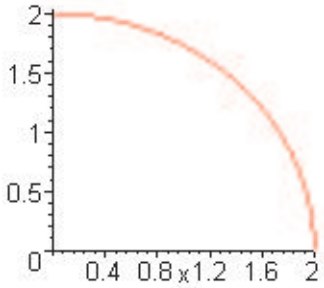
Property 1: $\int_a^a f(x)dx = 0$.

That is, if all of the Δx_i's are equal to 0, then the definite integral is 0. Now for some examples.

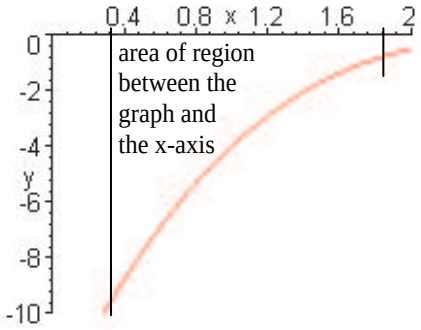
Example 1: Find $\int_0^1 (1 - x)dx$. From a sketch of the region, we see that the area is that of a right triangle whose legs are of length 1. Hence, the value of the integral is 1/2.



Example 2: Find $\int_0^2 \sqrt{4 - x^2}\, dx$. The area is that of a quarter-circle of radius 2 as shown in the sketch. Hence, from the area formula for a circle, we see that the value of the integral is $(\pi \cdot 2^2)/4 = \pi$.



If you look back at the definition of the definite integral, you will see that there is no requirement that f be a nonegative function. In fact, suppose f is continuous (and hence the definite integral exists) but f is strictly negative on $[a, b]$. Then in forming the the sums $\sum_{i=1}^n f(c_i)\Delta x_i$, note that $f(c_i)$ will be negative for all i and hence the sum will be negative. In fact, we can see that the definite integral will be the negative of the area of the region between the interval $[a, b]$ on the x-axis and the graph of f. (**Remember: areas are always nonnegative, but an integral may be negative.**) We have just proved a property of the definite integral.



Property 2: If f is integrable and $f(x) \geq 0$ on $[a, b]$, then $\int_a^b f(x)\, dx$ equals the area of the region under the graph of f and above the interval $[a, b]$. If $f(x) \leq 0$ on $[a, b]$, then $\int_a^b f(x)\, dx$ equals the negative of the

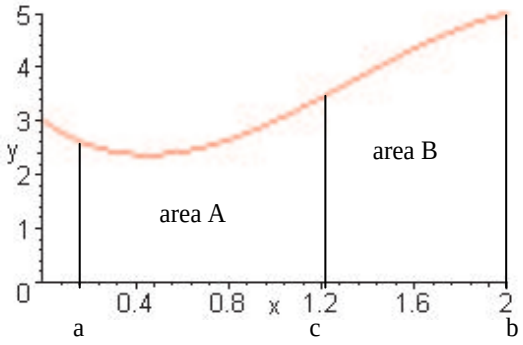
area of the region between the interval $[a, b]$ and the graph of f.

What happens if we integrate from right to left instead of from left to right? You guessed it.

Definition 1: $\int_b^a f(x)\,dx = -\int_a^b f(x)\,dx$.

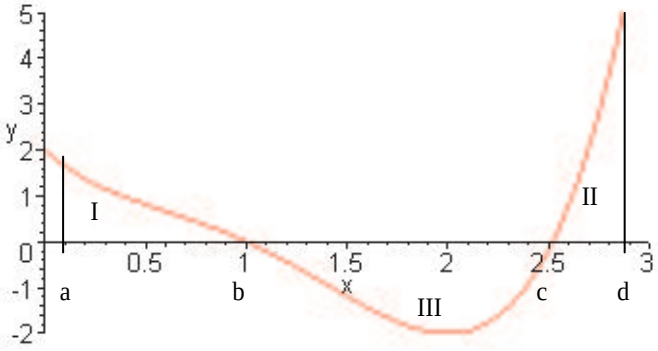
Because areas are additive, the following property also makes sense as seen in the sketch. That is, the total area over the interval $[a, b]$ is the sum of area A and area B. However, because of the above definition, the equation remains true even if c is not between a and b, or if b is less than a.

Property 3: $\int_a^c f(x)\,dx + \int_c^b f(x)\,dx = \int_a^b f(x)\,dx$.



Consider now a function that is both positive and negative on $[a, b]$. Then because of Property 3 and Property 2, the value of the definite integral is the sum of the areas of the regions between $[a, b]$ and the graph of f *above* the x-axis, minus the sum of the areas of the regions between $[a, b]$ and the graph of f *below* the x-axis. Expressed more compactly, the definite integral is the sum of the areas above minus the sum of the areas below. (Conclusion: whereas area is always nonnegative, the definite integral may be positive, negative, or zero.)

Example 3: In the sketch below, if I, II, and III represent the areas (all positive numbers) of the depicted regions, then $\int_a^d f(x)\,dx$ = I+II-III.

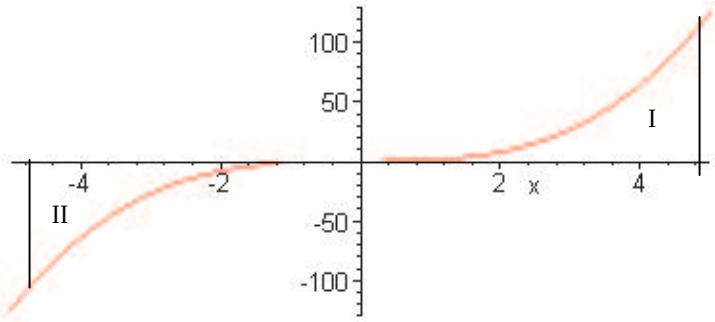


Integrals also add in the vertical direction. To understand this, note that for the finite sums that are involved in the definition of the integral, $\sum_{i=1}^n (Af(x_i) + Bg(x_i)) = \sum_{i=1}^n Af(x_i) + \sum_{i=1}^n Bg(x_i) = A\sum_{i=1}^n f(x_i) + B\sum_{i=1}^n g(x_i)$. As we pass to the limit in defining the integral, we need to use the fact that the limit of the sums is the sum of the limits. That is, we need to know that each limit exists. Hence we must assume that f and g are integrable to obtain the following result.

Property 4: If f and g are integrable on $[a, b]$, then $\int_a^b (Af(x) + Bg(x))\,dx = A\int_a^b f(x)\,dx + B\int_a^b g(x)\,dx$ for any constants A and B.

For some functions we can use symmetry to rewrite the integral in a simpler form. For instance, we have seen that odd functions are symmetric about the origin, and even functions are symmetric about the y-axis. Thus, we have the following two properties.

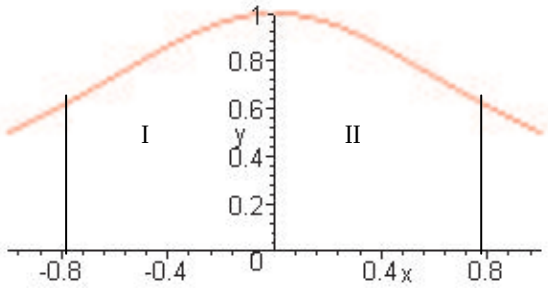
Property 5: If f is an odd function, then $\int_{-a}^a f(x)\,dx = 0$. That is, the definite integral of an odd function over a symmetric interval is zero.



area I = area II

In the case of even functions, the areas above and below the axis on a symmetric interval don't cancel, of course. However, we do have the following.

Property 6: If f is an even function, then $\int_{-a}^{a} f(x)\,dx = 2\int_{0}^{a} f(x)\,dx$.



area I = area II

Example 4: Let the function f be defined piecewise by

$$f(x) = \begin{cases} x & \text{if } x < 0 \\ \sqrt{-x^2 + 2x} & 0 \le x \le 2 \\ x - 2 & \text{otherwise} \end{cases}$$

Then from what we know about the areas of triangles and circles, we have

$$\int_{-1}^{3} f(x)\,dx = -\frac{1}{2} + \frac{\pi}{2} + \frac{1}{2} = \frac{\pi}{2}$$

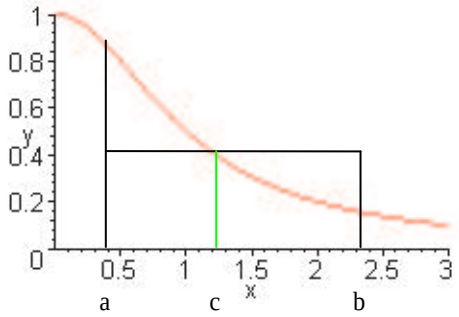
Notice that $\int_{-1}^{0} f(x)\,dx = -\int_{2}^{3} f(x)\,dx$.

Mean Value Theorem for Definite Integrals

The next property is is a bit more subtle than the ones we have met so far. And yet it is easy to describe.

Consider the following sketch. It seems clear that we can find a rectangle with base $[a, b]$ whose area equals the area of the region under the graph of f and above $[a, b]$. We just draw the top edge of a rectangle and adjust its height until we find the rectangle for which the areas are equal. That is, the *extra* area of the

rectangle equals the *omitted* area under the graph. The Mean Value Theorem for definite integrals says that this is always possible under not too restrictive conditions.



Theorem: (MVT for Definite Integrals) Let f be continuous on the interval $[a, b]$. Then there exists c in $[a, b]$ such that $\int_a^b f(x)\,dx = (b-a)f(c)$.

Note that on the right hand side of the equation we have the formula for the area of a rectangle of width $(b-a)$ and height $f(c)$. We think of $f(c)$ as the average value of f on the interval. Now, we can use the MVT formula to define *average value* precisely.

Definition 2: The average value of a continuous function on the interval $[a, b]$ is $\frac{1}{b-a}\int_a^b f(x)\,dx$.

As soon as we have some relatively straightforward ways to compute definite integrals, we can return to this formula and make use of it.

Applet: Mean Value Theorem for Integrals **Try it!**

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

4.4 The Fundamental Theorem of Calculus

We are about to discuss a theorem that relates derivatives and definite integrals. It is so important in the study of calculus that it is called the *Fundamental Theorem of Calculus*. It also gives us a practical way to evaluate many definite integrals without resorting to the limit definition. The theorem has two main parts that we will state separately as Part I and Part II.

Fundamental Theorem of Calculus (Part I-antiderivative): Suppose that f is a continuous function on the interval I containing the point a. Define the function F on I by the integral formula

$$F(x) = \int_a^x f(t)\,dt$$

Then F is differentiable on I and $F'(x) = f(x)$. That is, F is an antiderivative of f on I.

Fundamental Theorem of Calculus (Part II-evaluation): If $G(X)$ is any antiderivative of f on I (that is, $G'(x) = f(x)$ on I), then for any b in I,

$$\int_a^b f(x)\,dx = G(b) - G(a)$$

This theorem is truly remarkable. Leibniz seems to have been the first one to recognize its generality and significance. Let's look at some examples so that we can gain a better understanding of what the theorem says, and then we will outline a proof.

Example 1: To compute $\int_0^1 (x+1)dx$, we need only find an antiderivative of $x+1$, namely, $x^2/2+x$. Then we evaluate this antiderivative at 1 and subtract its value at 0. Thus, $\int_0^1 (x+1)\,dx = (1/2+1) - (0) = 3/2$.

We normally use a vertical bar to indicate evaluation of the antiderivative at the endpoints of the interval. That is,

$$G(x)\big|_a^b = G(b) - G(a)$$

Example 2: $\int_0^1 x^2\,dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3} - 0 = \frac{1}{3}$.

Example 3: $\int_0^{\pi/4} \sin x\,dx = -\cos x\big|_0^{\pi/4} = -1/\sqrt{2} - (-1) = 1 - 1/\sqrt{2}$.

Example 4: $\int_0^{\pi/4} \sec^2 x\,dx = \tan x\big|_0^{\pi/4} = 1 - 0 = 1$.

We can also illustrate Part I of the Fundamental Theorem.

Example 5: $\frac{d}{dx}\int_1^x t^2\,dt = x^2$.

Example 6: $\frac{d}{dx}\int_1^{x^2} t^3\,dt = (x^2)^3 \cdot 2x$ where we first have used the Fundamental Theorem and then the chain rule to complete the calculation of the derivative.

Example 7: Consider $\frac{d}{dx}\int_{x^2}^{x^3} e^{-t^2}\,dt$. We first have to put the integral in the correct form so that we can use the Fundamental Theorem:

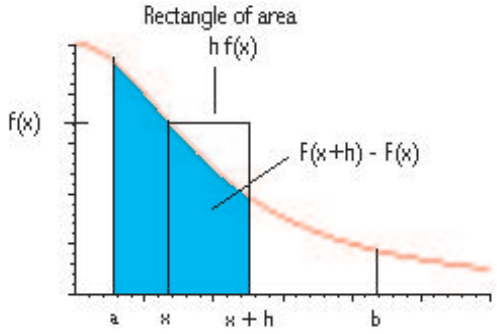
$$\begin{aligned}
\frac{d}{dx}\int_{x^2}^{x^3} e^{-t^2}\,dt &= \\
&= \frac{d}{dx}\left(\int_{x^2}^{0} e^{-t^2}\,dt + \int_0^{x^3} e^{-t^2}\,dt\right) \\
&= \frac{d}{dx}\left(-\int_0^{x^2} e^{-t^2}\,dt + \int_0^{x^3} e^{-t^2}\,dt\right) \\
&= -e^{-x^4}(2x) + e^{-x^6}(3x^2)
\end{aligned}$$

Now that we have gained some experience with the Fundamental Theorem through examples, let's look at a sketch of a proof in a special case.

Proof of the Fundamental Theorem (Part I): Fix x in I. Given that $F(x) = \int_a^x f(t)\,dt$, we need to evaluate the limit

$$\lim_{h\to 0} \frac{F(x+h) - F(x)}{h}$$

But look at the sketch below. Notice that $F(x)$ is the area under the graph of f and above the interval $[a, x]$, while $F(x+h)$ is the area under the graph of f and above the interval $[a, x+h]$. Thus, $F(x+h) - F(x)$ is the area under the graph of f and above the interval $[x, x+h]$.



But for small values of h, this area is approximately equal to the area of the rectangle of height $f(x)$ on the same base; its area is length times width, or $h \cdot f(x)$. Thus, for small h, the difference quotient is approximately equal to $\frac{hf(x)}{h} = f(x)$. In other words,

$$F'(x) = \lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = f(x)$$

thereby completing the proof of Part I.

Proof of the Fundamental Theorem (Part II): From Part I, we have that $F(x) = \int_a^x f(t)\,dt$ is an antiderivative of f. If G is another antiderivative, then we know from a previous result that they must differ by a constant. That is, $G(x) = F(x) + C$. Now, we know that $F(a) = \int_a^a f(t)\,dt = 0$. Thus, we can determine the value of C: $G(a) = F(a) + C = 0 + C = C$. Hence, $G(x) = F(x) + G(a)$, or $F(x) = G(x) - G(a)$. So, if b is any point in I, we have $G(b) - G(a) = F(b) = \int_a^b f(t)\,dt$, which is what we wanted to prove.

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

4.5 Techniques of Integration

In this section we are going to introduce the first approaches to evaluating an indefinite integral whose integrand does not have an immediate antiderivative. We begin with a list of integrals we should recognize.

$$\begin{aligned}
\int u^r\,du &= \frac{u^{r+1}}{r+1}+C, r\neq -1\\
\int \frac{1}{u}\,du &= \ln|u|+C\\
\int \sin u\,du &= -\cos u+C\\
\int \cos u\,du &= \sin u+C\\
\int \sec^2 u\,du &= \tan u+C\\
\int \sec u\tan u\,du &= \sec u+C\\
\int e^u\,du &= e^u+C
\end{aligned}$$

We can readily verify an equation by differentiating the right hand side and showing that we get the integrand on the left hand side. But what if an integral is not quite in the exact form that we require? For example, $\int e^{5x}\,dx$. What do we do then? Is there a systematic method that can minimize trial and error?

The Method of Substitution

We have already used differentials as an aid to integration when we discussed separable differential equations. In the present section, we will see that differentials continue to be a very useful technique for solving integrals. So, our first example will serve as a reminder of how to calculate them.

Example 1: If $y=x^3$, then $dy=3x^2\,dx$. Or if $y=\sin 4x$, then $dy=4\cos 4x\,dx$.

Reversing the Chain Rule: If $u=g(x)$ is a function of x, and f is a function of u, then the chain rule tells us that

$$(f(g(x)))'=f'(g(x))g'(x)$$

Thus, integrating the right hand side reverses the chain rule and we get

$$\int f'(g(x))g'(x)\,dx=f(g(x))+C$$

Now, we can rewrite the above integral by substituting into it $u=g(x)$ and the differential $du=g'(x)\,dx$. When we make these two substitutions we get

$$\int f'(u)\,du=f(u)+C$$

This last formula, combined with the use of differentials, constitutes the *Method of Substitution*.

Example 2: To find $\int e^{7x}\,dx$ by substitution, we look at the list of integrals at the beginning of the section and see that the *target* that it appears we should aim for is $\int e^u\,du$. Thus, we let $u=7x$. Then we calculate $du=7dx$; thus, $dx=\frac{du}{7}$. Substituting into the integral, we get the integral we were aiming for:

$$\int e^{7x}\,dx = \int \frac{e^u}{7}\,du = \frac{e^u}{7} + C = \frac{e^{7x}}{7} + C$$

We have already learned to solve the above integral by inspection. Indeed, we were doing nothing more than reversing the chain rule in an simple case. The next example is also one we have learned to do by inspection but which we can do formally by substitution.

Example 3: $\int \sin 2x\,dx$. Let $u = 2x$; then $du = 2dx$. So, substitution yields

$$\int \sin 2x\,dx = \int \frac{\sin u}{2}\,du = -\frac{\cos u}{2} + C = \frac{\cos 2x}{2} + C$$

Example 4: $\int \frac{x}{x^2+1}\,dx$. Let $u = x^2 + 1$; then $du = 2x\,dx$. So

$$\int \frac{x}{x^2+1}\,dx = \int \frac{1}{2u}\,du = \frac{1}{2}\ln|x^2+1| + C$$

Example 5: $\int \frac{x^2+1}{x}\,dx$. Be careful. This is not a substitution integral. That is, it is not an integral that requires substitution. We simplify the quotient to obtain two terms each of which we can integrate:

$$\int \frac{x^2+1}{x}\,dx = \int \left(x + \frac{1}{x}\right)\,dx = \frac{x^2}{2} + \ln|x| + C$$

Example 6: $\int \frac{x^2+1}{x^3+3x+2}\,dx$. This integral will yield to substitution: let $u = x^3 + 3x + 2$. Then $du = (3x^2 + 3)\,dx$. Thus, we substitute to get

$$\int \frac{x^2+1}{x^3+3x+2}\,dx = \frac{1}{3}\int \frac{1}{u}\,du = \frac{1}{3}\ln|u| + C = \frac{1}{3}\ln|x^3+3x+2| + C$$

Example 7: $\int \frac{\ln x}{x}\,dx$. Let $u = \ln x$; then $du = \frac{dx}{x}$. Thus,

$$\int \frac{\ln x}{x}\,dx = \int u\,du = \frac{u^2}{2} + C = \frac{(\ln x)^2}{2} + C$$

Thus far, we have not used the Method of Substitution with a definite integral. We will do so now.

Example 7 (continued):

$$\int_e^{e^2} \frac{\ln x}{x}\,dx$$

This is our first example of a definite integral requiring substitution. There are two basic ways to solve it: either we change the variable from x to u and change the limits of integration as well; or we leave the limits of integration unchanged and switch back from u to x.

Method 1: change the limits from x to u.

x	$u = \ln x$
e	1
e^2	2

$$\int_e^{e^2} \frac{\ln x}{x}\,dx = \int_1^2 u\,du = \left.\frac{u^2}{2}\right|_1^2 = \frac{2^2}{2} - \frac{1^2}{2} = \frac{3}{2}$$

Method 2: change the variable back to x and retain the original limits.

$$\int_e^{e^2} \frac{\ln x}{x}\,dx = \left.\frac{(\ln x)^2}{2}\right|_e^{e^2} = \frac{(\ln e^2)^2}{2} - \frac{(\ln e)^2}{2} = \frac{2^2}{2} - \frac{1^2}{2} = \frac{3}{2}$$

Example 8:

$$\int_0^{\pi/4} \tan x\,dx$$

We replace $\tan x$ by $\tan x = \frac{\sin x}{\cos x}$ and then use the substituiton $u = \cos x$, from which $du = -\sin x\,dx$. So, with a change of the limits of integration

x	$u = \cos x$
0	1
$\pi/4$	$1/\sqrt{2}$

the integral becomes:

$$\int_0^{\pi/4} \tan x\,dx = \int_0^{\pi/4} \frac{\sin x}{\cos x}\,dx = \int_1^{1/\sqrt{2}} -\frac{1}{u}\,du = -\ln|u||_1^{1/\sqrt{2}} = \frac{\ln 2}{2}$$

Integration by Parts

We saw above that really the Method of Substitution consists of reversing the chain rule. Another technique of integration that is often useful involves an undoing of the product rule. For, suppose that u and v are functions of x. Then starting with the product rule we get

$$\frac{d}{dx}(uv) = u\frac{dv}{dx} + v\frac{du}{dx}$$

$$u\frac{dv}{dx} = \frac{d}{dx}(uv) - v\frac{du}{dx}$$

$$\int u\frac{dv}{dx}\,dx = uv - \int v\frac{du}{dx}\,dx$$

Rewriting the last equation in terms of differentials yields

$$\int u\,dv = uv - \int v\,du$$

This is the so-called *integration by parts formula*. In practice, we can often use it to transform an integral that appears intractible into one whose integrand has an antiderivative we recognize.

Note that we first have to choose u; then dv is that part of the integrand that remains. In general, we must be able to differentiate u and integrate dv in order to use the rest of the parts formula. Also, we want an integrand that is simpler than the one with which we started. These simple observations should guide us in assigning u.

Let's consider some examples.

Example 1: Consider $\int xe^x\,dx$. Substitution does not appear to work. So, we try the only other technique we know, namely, parts. If we let $u = x$, then $dv = e^x\,dx$. Next, we find du and v, and use the parts formula.

$u = x$	$dv = e^x\,dx$
$du = dx$	$v = e^x$

$$\int xe^x\,dx = xe^x - \int e^x\,dx = xe^x - e^x + C$$

Example 2: Given $\int \ln x\,dx$, we let $u = \ln x$ and then $dv = dx$. And we proceed:

$u = \ln x$	$dv = dx$
$du = \frac{1}{x}\,dx$	$v = x$

$$\int \ln x\,dx = x\ln x - \int dx = x\ln x - x + C$$

Example 3: To find $\int_1^e \ln x\,dx$, we simply follow through in each term of the parts formula with the evaluation at the endpoints of the interval. Referring to the previous example,

$$\int_1^e \ln x\,dx = x\ln x|_1^e - \int_1^e dx = e - e + 1 = 1$$

Example 4: Given $\int x^2 \sin x\,dx$, use parts letting $u = x^2$ and $dv = \sin x\,dx$. Then

$u = x^2$	$dv = \sin x\,dx$
$du = 2x\,dx$	$v = -\cos x$

$$\int x^2 \sin x\,dx = -x^2\cos x + 2\int x\cos x\,dx$$

Now, we use parts again to evaluate the new integral.

$u = x$	$dv = \cos x\,dx$
$du = dx$	$v = \sin x$

$$\int x^2 \sin x\,dx = -x^2\cos x + 2(x\sin x + \cos x) + C$$

Example 5: We use parts to evaluate $\int e^x \sin x\,dx$:

$u = e^x$	$dv = \sin x\,dx$
$du = e^x\,dx$	$v = -\cos x$

$$\int e^x \sin x \, dx = -e^x \cos x + \int e^x \cos x \, dx$$

Using parts again:

$u = e^x$ $du = e^x \, dx$	$dv = \cos x \, dx$ $v = \sin x$

$$\int e^x \sin x \, dx = -e^x \cos x + e^x \sin x - \int e^x \sin x \, dx$$

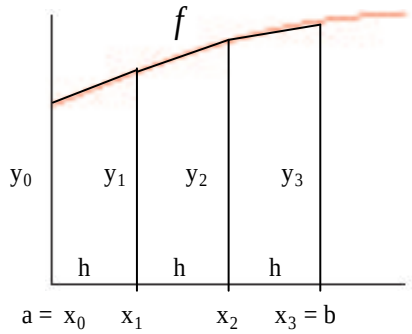
Now, we have an equation that has the unknown integral on both sides. Thus, we can solve for it to get

$$\int e^x \sin x \, dx = \frac{1}{2}\left(-e^x \cos x + e^x \sin x\right) + C$$

Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

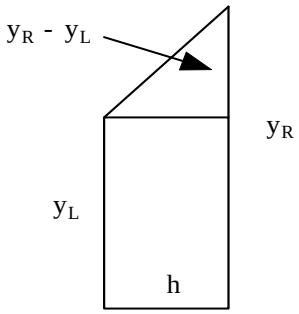
4.6 Trapezoid Rule

Many applications of calculus involve definite integrals. If we can find an antiderivative for the integrand, then we can evaluate the integral fairly easily. When we cannot, we turn to numerical methods. The numerical method we will discuss here is called the *Trapezoid Rule*. Although we often can carry out the calculations by hand, the method is most effective with the use of a computer or programmable calculator. But at the moment let's not concern ourselves with these details. We will describe the method first, and then consider ways to implement it.



The general idea is to use trapezoids instead of rectangles to approximate the area under the graph of a function. A trapezoid looks like a rectangle except that it has a slanted line for a top. Working on the interval $[a, b]$, we subdivide it into n subintervals of equal width $h = (b - a)/n$. This gives rise to the partition $a = x_0 \leq x_1 \leq x_2 \leq \cdots \leq x_n = b$, where for each j, $x_j = a + jh$, $0 \leq j \leq n$. Moreover, we let $y_j = f(x_j), 0 \leq j \leq n$. That is, the vertical edges go from the x-axis to the graph of f. Consult the sketch above where we have shown a finite number of subintervals.

If we are going to use trapezoids instead of rectangles as our basic area elements, then we have to have a formula for the area of a trapezoid.



With reference to the sketch above, the area of a trapezoid consists of the area of the rectangle plus the area of the triangle, or $hy_L + (h/2)(y_R - y_L) = h(y_L + y_R)/2$. So, the area is h times the average of the lengths of the two vertical edges.

Now, we return to the original problem of finding the definite integral of a function f defined on the interval $[a, b]$. We define the *Trapezoid Rule* as follows.

Definition: The n-subinterval trapezoid approximation to $\int_a^b f(x)\,dx$ is given by

$$\begin{aligned} T_n &= \frac{h}{2}\left(y_0 + 2y_1 + 2y_2 + 2y_3 + \cdots + 2y_{n-1} + y_n\right) \\ &= \frac{h}{2}\left(y_0 + y_n + 2\sum_{j=1}^{n-1} y_j\right) \end{aligned}$$

To see where the formula comes from, let's carry out the process of adding the areas of the trapezoids. Refer to the original sketch, and use the formula we derived for the area of a trapezoid. Note that when we add the areas of the trapezoids starting on the left, the area of the first , second, and third are:

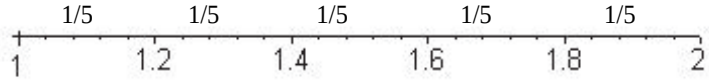
$$\frac{h}{2}(y_0 + y_1)$$

$$\frac{h}{2}(y_1 + y_2)$$

$$\frac{h}{2}(y_2 + y_3)$$

So, y_0 and y_3, the first and the last, each appear once; and all the other y_j's appear exactly twice. We can see from this example that there will be a similar pattern no matter the number of trapezoids: The first and the last vertical edge appears once, and all other vertical edges appear two times when we sum the areas of the trapezoids. This is exactly what the Trapezoid Rule entails in the formula above.

Example 1: Find T_5 for $\int_1^2 \frac{1}{x}\,dx$. We can readily determine that f(x) = 1/x, $h = 1/5$ (so $h/2 = 1/10$), and $x_j = 1 + j/5, 0 \le j \le 5$.



So,

$$T_5 = \frac{1}{10}\left(1 + \frac{1}{2} + 2\left(\frac{5}{6} + \frac{5}{7} + \frac{5}{8} + \frac{5}{9}\right)\right) \approx .0696$$

Example 2: Find T_5 for $\int_0^1 \sqrt{1 - x^2}\,dx$. That is, we are going to approximate one-quarter of the area of a circle of radius 1. The exact answer is $\pi/4$, or approximately .7853981635. Note that $h = 1/5$, $y_0 = 1$ and $y_5 = 0$. Thus,

$$T_5 = \frac{1}{10}\left(1 + 2\sum_{j=1}^{4}\sqrt{1 - \frac{j^2}{25}}\right)$$

or about .7592622072.

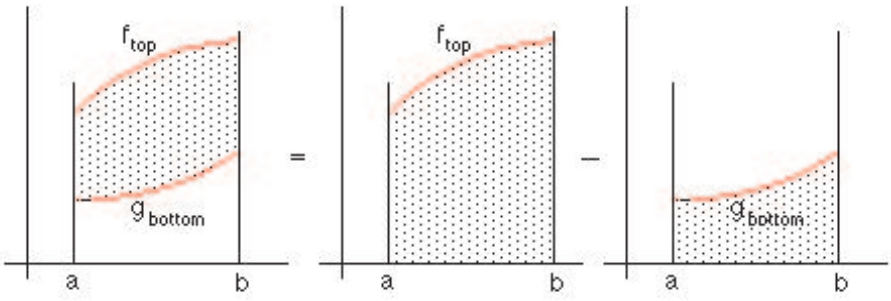
Applet: Numerical Integration **Try it!**

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

4.7 Areas Between Curves

We know that if f is a continuous nonnegative function on the interval $[a, b]$, then $\int_a^b f(x)\,dx$ is the area under the graph of f and above the interval. We are going to extend this notion a bit by considering how to find the area between two functions. To be specific, suppose we are given two continuous functions, f_{top} and g_{bottom} defined on the interval $[a, b]$, with $g_{bottom}(x) \leq f_{top}(x)$ for all x in the interval. How do we find the area bounded by the two functions over that interval?



We have used the notation f_{top} and g_{bottom} for obvious reasons. However, we want to caution you that all of the subsequent analysis really does assume that f lies above or is equal to g at every point throughout the interval. So, you want to be sure in problem-solving that you have verified that this is the case before using the formula that we will develop next.

The area of the region between the two curves and above the interval $[a, b]$ equals the area of the region under the graph of f_{top} on that interval minus the area of the region under the graph of g_{bottom} on the same interval. Thus, the area of the region between the two curves equals

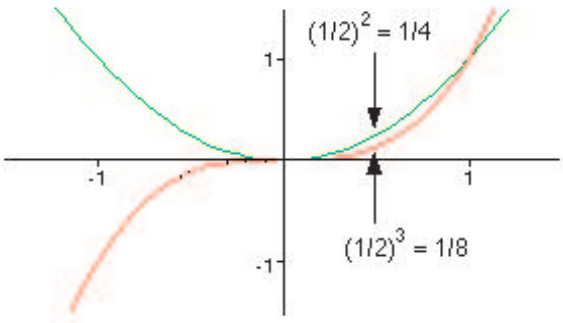
$$\int_a^b f_{top}(x)\,dx - \int_a^b g_{bottom}(x)\,dx = \int_a^b \left(f_{top}(x) - g_{bottom}(x)\right)\,dx$$

The last integral is normally the form in which we express the area between the two curves. But remember that to apply the formula, we have to know which curve is on top and which is on bottom, and we have to be certain that that relationship is maintained throughout the interval. Note though that the vertical placement of the curves does not matter. For example, both of them could lie below the x-axis, or one above and one below, and the formula would still hold.

Example 1: Find the area of the region between the graphs of $y = x^2$ and $y = x^3$ for $0 \leq x \leq 1$. Solving for the point(s) of intersection, we find that the curves intersect at $x = 0, 1$:

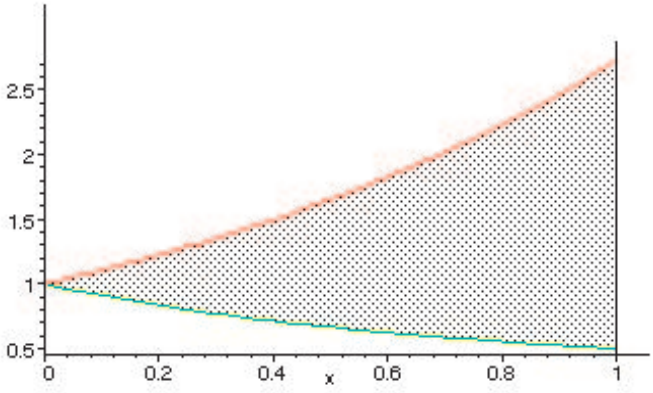
$$\begin{aligned} x^2 &= x^3 \\ x^2(x-1) &= 0 \end{aligned}$$

implies $x = 0, x = 1$. Here is what a quick sketch by hand might look like:



To determine which curve is on top, we plug in a convenient value of x to find that $y = x^2$ is on top throughout the interval $[0, 1]$. Thus, we can use the formula above: The area of the region between the two curves equals $\int_0^1 \left(x^2 - x^3\right)\, dx = \left(x^3/3 - x^4/4\right)\big|_0^1 = 1/3 - 1/4 = 1/12$.

Example 2: Find the area of the region between $y = e^x$ and $y = \frac{1}{1+x}$ on the interval $[0, 1]$. The graph is shown.



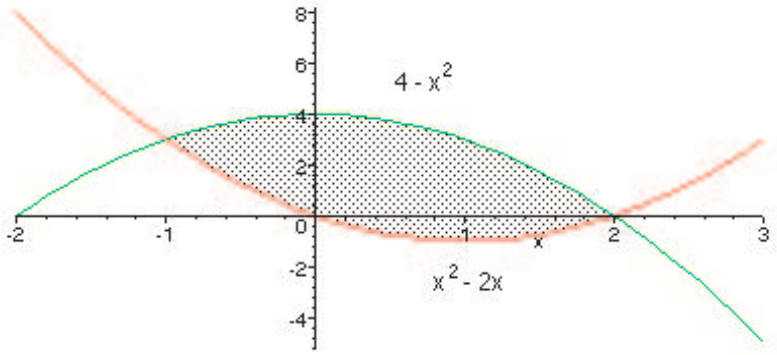
Because the curves intersect at $x = 0$, and $y = e^x$ is increasing while $y = 1/(1 + x)$ is decreasing, we know that e^x is on top. Thus, the area of the region between the curves equals

$$\begin{aligned}\int_0^1 \left(e^x - \frac{1}{1+x}\right)\, dx &= \left(e^x - \ln|1+x|\right)\big|_0^1 \\ &= e - \ln 2 - e^0 + \ln 1 \\ &= e - \ln 2 - 1\end{aligned}$$

Example 3: Find the area of the region bounded by $y = x^2 - 2x$ and $y = 4 - x^2$. To solve this problem, we need a sketch so that we can determine which function is on top over which intervals. We will begin by determining the points of intersection.

$$\begin{aligned}x^2 - 2x &= 4 - x^2 \\ 2x^2 - 2x - 4 &= 0 \\ 2(x^2 - x - 2) &= 0 \\ 2(x-2)(x+1) &= 0\end{aligned}$$

So, $x = 2$ or $x = -1$. Both functions are quadratic polynomials, so their graphs are parabolas, one opening up and the other down. We should recognize from the sign on x^2 which curve is on top, or we can test a value of x to find out.



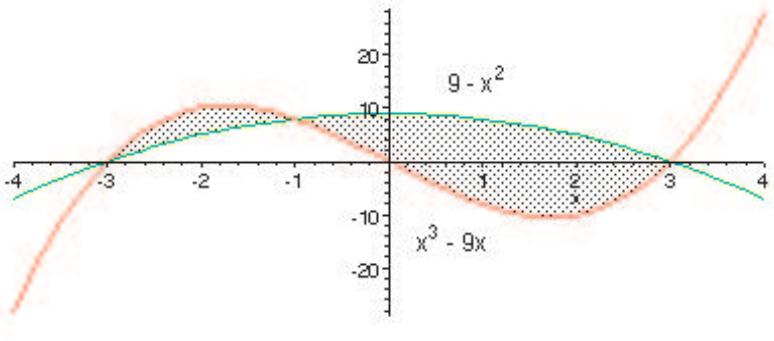
Thus, the area of the region can be gotten by applying our integral formula to obtain:

$$\begin{aligned}\int_{-1}^{2}(4-x^2-(x^2-2x))\,dx &= \left(4x-\frac{2x^3}{3}+x^2\right)\Big|_{-1}^{2}\\ &= 8-\frac{16}{3}+4-\left(-4+\frac{2}{3}+1\right)\\ &= 9\end{aligned}$$

Example 4: Find the area of the region bounded by the two curves $y=x^3-9x$ and $y=9-x^2$. [Hint: You probably will need to know that $x+1$ is a factor of x^3+x^2-9x-9.] Let's find the points of intersection:

$$\begin{aligned}x^3-9x &= 9-x^2\\ x^3+x^2-9x-9 &= 0\\ (x+1)(x^2-9) &= 0\\ (x+1)(x-3)(x+3) &= 0\end{aligned}$$

Note that to obtain the next to the last equation above, we used the hint and divided x^3+x^2-9x-9 by $x+1$. A sketch of the graphs follows:

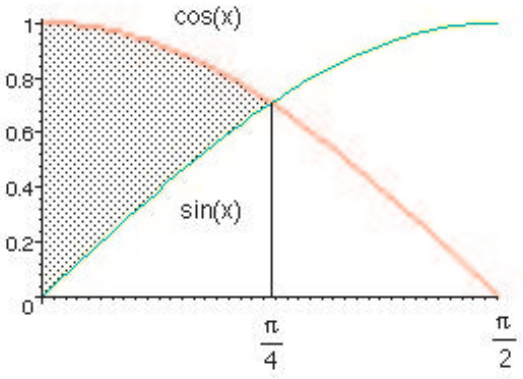


Thus, the area can be gotten by applying our integral formula twice– once to the interval $[-3,-1]$ where x^3-9x is on top, and once to the interval $[-1,3]$ where $9-x^2$ is on top– and adding the results together:

$$\int_{-3}^{-1}(x^3-9x-(9-x^2))\,dx+\int_{-1}^{3}(9-x^2-(x^3-9x))\,dx$$

Routine calculation gives the answer $\frac{148}{3}$. Try it for yourself and verify this result.

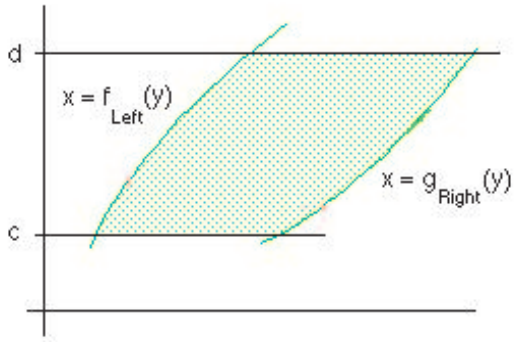
Example 5: Find the area between $\sin x$ and $\cos x$ on $[0,\pi/4]$. Here is a sketch:



So, the area is

$$\begin{aligned}\int_0^{\pi/4} (\cos x - \sin x)\, dx &= (\sin x + \cos x)\big|_0^{\pi/4} \\ &= \sin(\pi/4) + \cos(\pi/4) - (\sin 0 + \cos 0) \\ &= \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} - (0+1) \\ &= \sqrt{2} - 1\end{aligned}$$

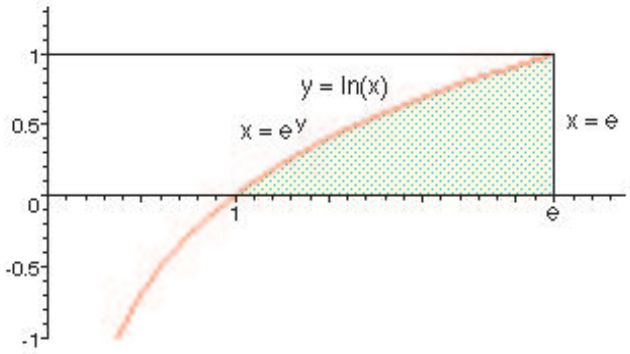
Functions of y: Thus far, we have only considered functions of x. We could just as well consider two functions of y, say, $x = f_{Left}(y)$ and $x = g_{Right}(y)$ defined on the interval $[c, d]$ on the y-axis as in the sketch below:



Then the area between the graphs can be found by subdividing the interval $[c, d]$ on the y-axis, and using horizontal rectangular area elements. In that case, we get that the area between the two curves is the definite integral

$$\int_c^d (g_{Right}(y) - f_{Left}(y))\ dy$$

Example 6: Find the area under the graph of $y = \ln x$ and above the interval $[1, e]$ on the x-axis.



We know that integrating with respect to x yields the definite integral $\int_1^e \ln x\, dx$. However, suppose we do not know (or remember, or want to investigate) an antiderivative for $\ln x$. Then we can try to solve this problem by integrating instead with respect to y. The functions are $x = e$ on the right, $x = e^y$ on the left, over the interval from $y = 0$ to $y = 1$:

$$\int_0^1 (e - e^y)\, dy = (ey - e^y)\big|_0^1 = e - e + 1 = 1$$

Thus, our problem is solved. Alternatively, we could go back to where we began. It turns out that by application of the Integration by Parts formula to the integral $\int \ln x\,dx$, $x\ln x - x$ is an antiderivative of $\ln x$. (Just calculate the derivative to verify this result.) Thus, we can evaluate the original integral with respect to x:

$$\int_1^e \ln x\,dx = (x\ln x - x)|_1^e = e - e - (0-1) = 1$$

As we should, we get the same answer for the area of the region integrating with respect to either x or y.
Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

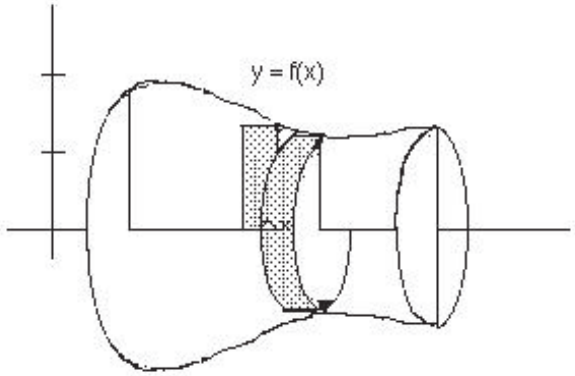
4.8 Volumes of Solids of Revolution

Integrals find application in many modeling situations involving continuous variables such as area or volume. They allow us to model physical entities that can be described through a process of adding up, or accumulating, smaller infinitesimal parts. In what follows, we will illustrate by discussing the very powerful *Riemann Sum* approach to modeling with integrals.

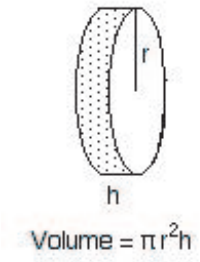
Riemann Sum modeling is based on describing real-world phenomena through continuous functions on closed intervals. In particular, Riemann Sums provide a method of analysis that proceeds by dividing an interval into finitely many small subintervals; developing on each individual subinterval a function-related formula that works there; summing these individual contributions; and passing to the limit as the length of the largest subinterval goes to zero. At the end of this process, the resulting integral computes, or even defines, the quantity that was desired at the outset.

We have already used this approach to describe the area under the graph of a continuous function and above a closed interval. Now we are going to use it to find the volume of a so-called *solid of revolution.*

We begin with a plane region R bounded by the non-negative function $y = f(x)$, $y = 0$, $x = a$, and $x = b$. We then rotate this region about the x-axis.

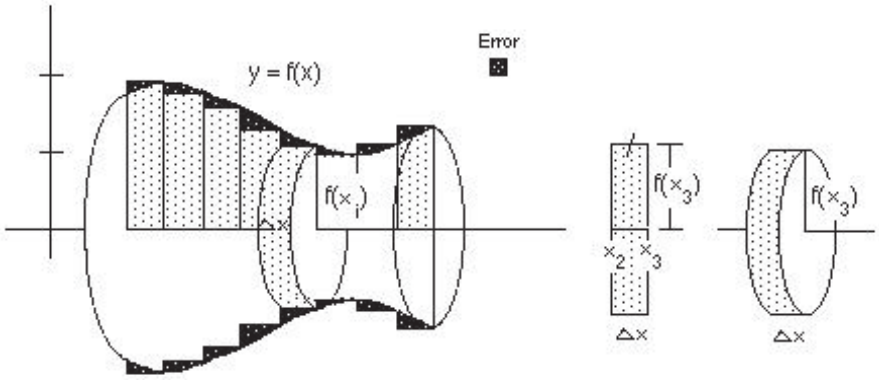


The resulting three-dimensional solid is called a *solid of revolution.* Note that its cross-sectional area in a plane perpendicular to the x-axis at x is a circular disk of radius $f(x)$. Also note that when we rotate a rectangular area element, we get a circular disk. We are going to use the volumes of these disks to approximate the volume of the solid of revolution. A right circular disk of radius r and width h has volume $\pi r^2 h$ (the area of the circular base times the width).



Summary of the Riemann Sum Volume of Revolution Method: In light of the description above of the Riemann Sum method to compute volumes of solids of revolution, we can summarize the general procedure that we will apply in different situations. These steps capture the essence of the modeling approach using Riemann Sums to find these volumes.

Begin with a continuous non-negative function f on a closed interval $[a, b]$. Revolve the graph of f around the x-axis to obtain a so-called *solid of revolution.* The problem is to compute its volume. To do this, proceed as follows:



1. Divide the interval $[a, b]$ into n subintervals of equal length $\Delta x = (b - a)/n$. Call the points of the subdivision $a = x_0 \leq x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_{n-1} \leq x_n = b$, where $x_i = a + i\Delta x$ for each i.
2. Erect on each subinterval $[x_{i-1}, x_i]$ a rectangle of height $f(x_i)$; then revolve this rectangle about the x-axis to generate a circular disk of radius $f(x_i)$ and width Δx. The volume of the disk is thus $\pi f(x_i)^2 \Delta x$.
3. The sum of the volumes of these disks provides an approximation to the volume of the solid of revolution:

$$\sum_{i=1}^{n} \pi f(x_i)^2 \Delta x$$

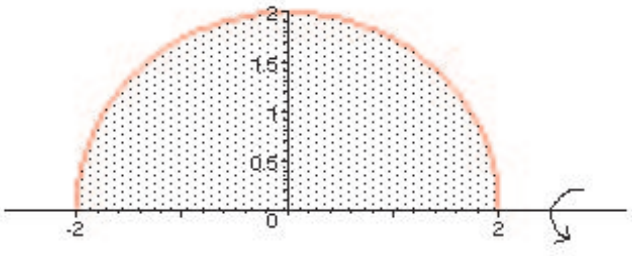
4. Taking the limit as $\Delta x \to 0$, the above approximation approaches the volume V of the solid of revolution. Moreover, we also recognize it as a limit of Riemann Sums that converge to the definite integral:

$$V = \lim_{\Delta x \to 0} \sum_{i=1}^{n} \pi f(x_i)^2 \Delta x = \int_a^b \pi f(x)^2 \, dx$$

5. Finally we use the integral formula to compute the volume V of the solid of revolution.

$$V = \pi \int_a^b f(x)^2 \, dx$$

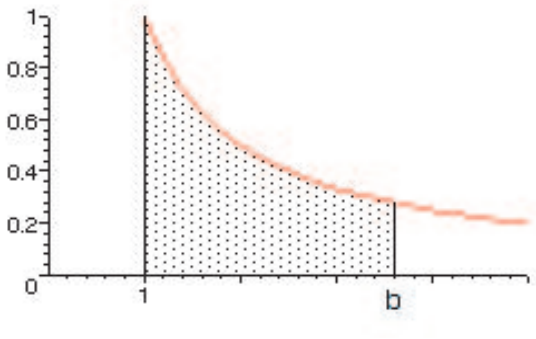
Example 1: Find the volume of a ball of radius 2. To solve this problem, we begin with a plane region which, when revolved about the x-axis, will generate the ball. To this end, we let $y = \sqrt{4 - x^2}$ define the upper boundary of a semicircle on the interval $[-2, 2]$.



Then using the formula from above, the volume of the ball is

$$V = 2\pi \int_0^2 (\sqrt{4 - x^2})^2 \, dx = 2\pi \int_0^2 (4 - x^2) \, dx = 2\pi \frac{16}{3}$$

Example 2: Revolve the region bounded on top by $y = 1/x$ and above the interval $[1, b]$, where $b > 1$, about the x-axis as pictured.



Then the volume is

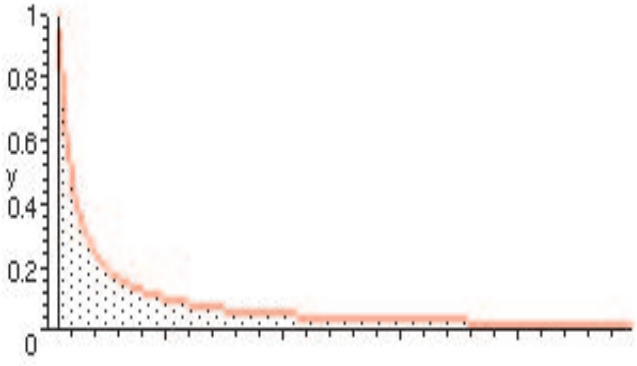
$$V = \pi \int_1^b x^{-2} \, dx = \pi \left(1 - \frac{1}{b}\right)$$

Thus, as $b \to \infty$, the volume of the 3-D pointed horn approaches the value π.

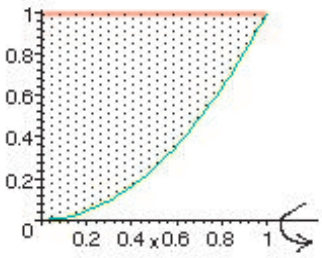
On the other hand, if we compute the area of the plane region above the interval $[1, b]$ and under the graph of $y = 1/x$, we get:

$$A = \int_1^b \frac{1}{x} \, dx = \ln b$$

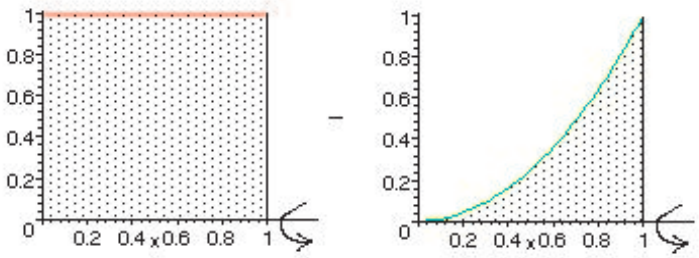
Thus, the area approaches ∞ as $b \to \infty$. So, the solid of revolution has finite volume whereas the planar figure that generated it has infinite area. Curious, eh?



Example 3: Revolve the region bounded by $y = x^2$ and $y = 1$ and $x \geq 0$ about the x-axis and compute the volume of the resulting solid. This solid will hold water if we turn it on its side.



We think of generating it by revolving the two plane regions shown and subtracting the 3-D results:

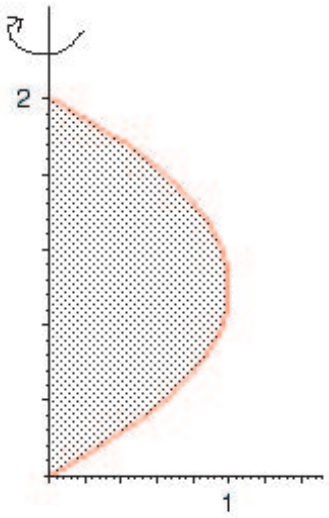


The formula we have developed for volume applies to each of these two situations. Thus, the volume of the solid we seek is:

$$V = \pi \int_0^1 1^2 \, dx - \pi \int_0^1 \left(x^2\right)^2 \, dx = \pi \int_0^1 (1 - x^4) \, dx = \frac{4}{5}\pi$$

Horizontal Rectangular Elements: We can also revolve a region about the y-axis using horizontal rectangular elements.

Example 4: Revolve the region bounded by $x = 0$ and $x = 2y - y^2$ about the y-axis and compute the volume of the resulting solid.



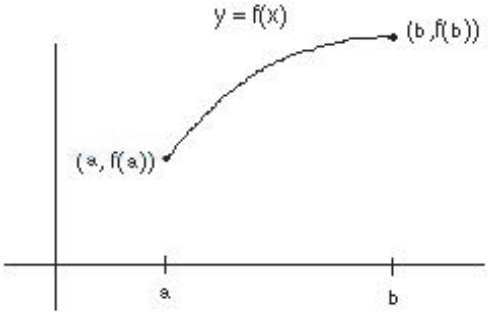
The volume is

$$V = \pi \int_0^2 \left(2y - y^2\right)^2 \, dy = \pi \int_0^2 (4y^2 - 4y^3 + y^4) \, dy = \frac{16}{15}\pi$$

Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

4.9 Arc Length

We have seen that in applications integrals are very powerful modeling tools through the use of Riemann Sums. As another illustration of the Riemann Sum modeling method, consider the problem of computing the length of a curve in the plane. To be explicit, we will assume that $y = f(x)$ is a continuous function defined on the interval $[a, b]$ and that $f'(x)$ exists at every point of the interval. Then our task is to determine the length of the graph of f from the point $(a, f(a))$ to the point $(b, f(b))$.

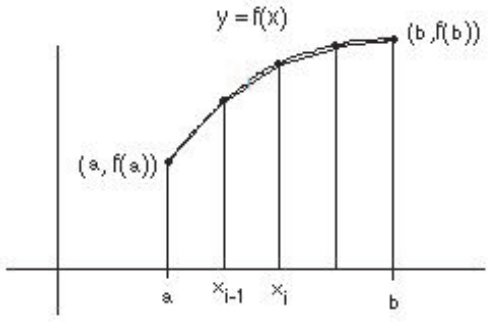


If the graph were a straight line, we could use the formula–that comes from the Pythagorean theorem–for the distance between two points to find the length of the line connecting them. If the graph is not a straight line, then it makes sense to do what we probably would have done in grade school, namely, use a finite number of straight line segments to approximate the length of the curve. We then can use the Riemann Sum approach to add the lengths, and pass to the limit as all lengths go to zero. This procedure should allow us to compute the length exactly. Let's outline the specific steps.

Summary of the Riemann Sum Method for Arc Length: Here are the steps in the modeling process of using Riemann Sums to find the arc length of a curve in the plane:

1. Divide the interval $[a, b]$ into n subintervals of equal length $\Delta x = (b - a)/n$. Call the points of the subdivision $a = x_0 \leq x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_{n-1} \leq x_n = b$, where $x_i = a + i\Delta x$ for each i.

2. On each subinterval $[x_{i-1}, x_i]$ connect the points $(x_{i-1}, f(x_{i-1})$ and $(x_i, f(x_i)$ on the graph of f with straight lines.



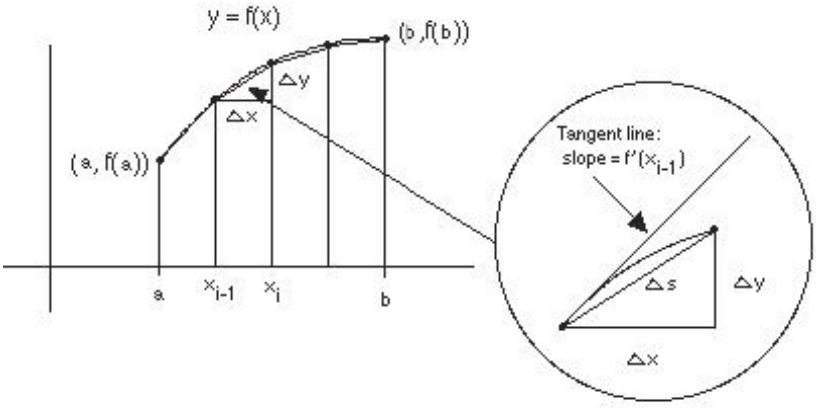
Now, let $\Delta x_i = x_i - x_{i-1}$ and $\Delta y = f(x_i) - f(x_{i-1})$. Then the length of the straight-line segment connecting the two points is:

$$(\Delta s)^2 = (\Delta x)^2 + (\Delta y)^2$$

$$\Delta s = \sqrt{(\Delta x)^2 + (\Delta y)^2}$$

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

But for small Δx, the slope $\frac{\Delta y}{\Delta x}$ of the line is approximately equal to the slope of the tangent line at the left-hand point, which equals $f'(x_{i-1})$.



Thus, we can replace Δs by the approximate value $\Delta s \approx \Delta x\sqrt{1+[f'(x_{i-1})]^2}$. We will use this approximate value for the length of the straight-line segment on each subinterval.

3. The sum of the approximate lengths of these line segments provides an approximation to the length of the curve:

$$\sum_{i=1}^{n}\sqrt{1+[f'(x_{i-1})]^2}\Delta x$$

4. Taking the limit as $\Delta x \to 0$, the above approximation approaches the length of the curve. Moreover, we recognize it as a limit of left-hand Riemann Sums. Thus, the limit is the definite integral

$$L = \lim_{\Delta x\to 0}\sum_{i=1}^{n}\sqrt{1+[f'(x_{i-1})]^2}\Delta x = \int_a^b \sqrt{1+[f'(x)]^2}\,dx$$

5. Now, use the integral formula to compute the length L of the graph of f between $x = a$ and $x = b$.

$$L = \int_a^b \sqrt{1+[f'(x)]^2}\,dx$$

The formula is called the *arc length* formula.

In many cases the arc length formula leads to an integrand for which we do not know an antiderivative and hence cannot apply the Fundamental Theorem of Calculus. In those situations, we still can find an accurate expression for the length of the graph by evaluating the definite integral using the Trapezoid Rule (or some other numerical method). In this age of calculators and computers, the Trapezoid Rule presents no practical obstacle, and hence the arc length formula is all we need to calculate the length of the graph of a function over an interval.

Example 1: To find the length of the arc $y = x^{3/2}$, from $x = 0$ to $x = 1$, we use the Arc Length formula:

$$\begin{aligned} L &= \int_0^1 \sqrt{1 + \left[\frac{3}{2}x^{1/2}\right]^2}\, dx \\ &= \int_0^1 \sqrt{1 + \left[\frac{9}{4}x\right]}\, dx \end{aligned}$$

This is a substitution integral with $u = 1 + \frac{9}{4}x$; then $du = (9/4)dx$. Also, changing the limits of integration, $u(0) = 1$ and $u(1) = 13/4$. So,

$$\begin{aligned} L &= \frac{4}{9}\int_1^{13/4} \sqrt{u}\, du \\ &= \frac{8}{27}\, u^{3/2}\Big|_1^{13/4} \\ &= \frac{13^{3/2} - 8}{27} \end{aligned}$$

Example 2: Find the length of the curve $y = x^4 + \frac{1}{32x^2}$ from $x = 1$ to $x = 2$. We calculate y' which we need for the arc length formula:

$$y' = 4x^3 - \frac{2}{32x^3} = 4x^3 - \frac{1}{16x^3}$$

$$\begin{aligned} L &= \int_1^2 \sqrt{1 + \left[4x^3 - \frac{1}{16x^3}\right]^2}\, dx \\ &= \int_1^2 \sqrt{1 + 16x^6 - \frac{8}{16} + \frac{1}{256x^6}}\, dx \\ &= \int_1^2 \sqrt{\frac{8}{16} + 16x^6 + \frac{1}{256x^6}}\, dx \\ &= \int_1^2 \sqrt{\left(4x^3 + \frac{1}{16x^3}\right)^2}\, dx \\ &= \int_1^2 \left(4x^3 + \frac{1}{16x^3}\right)\, dx \\ &= \left(x^4 - \frac{1}{32x^2}\right)\Big|_1^2 \\ &= 15 + \frac{3}{128} \end{aligned}$$

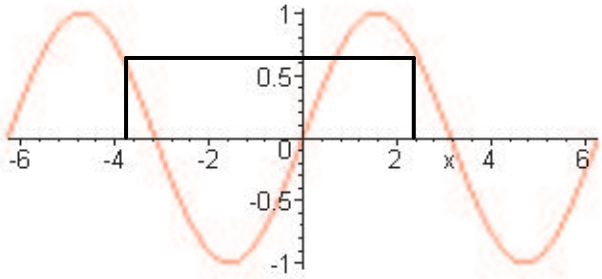
Applet: Numerical Integration **Try it!**
Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

4.10 Inverse Trigonometric Functions

We will introduce inverse functions for the sine, cosine, and tangent. In defining them, we will point out the issues that must be considered in defining the inverse of any periodic function. Then we will go on to find the derivative of the inverse sine and the inverse tangent. Their companion integration formulas will give us two new integrals that we will subsequently recognize how to calculate.

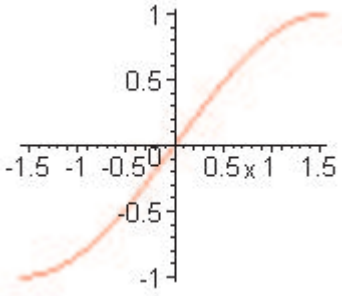
The Arcsine

We would like to define the inverse function of the sine. However, because the sine is periodic, it is not one-to-one and the graph of the sine function fails the horizontal line test. Hence the sine does not have an inverse unless we restrict its domain.



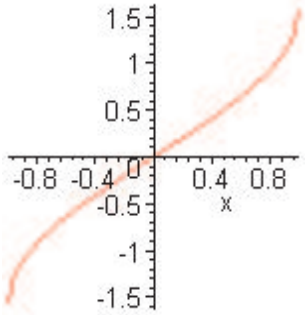
sine function
is not one-to-one

By convention we restrict the domain of the sine to the interval $[-\pi/2, \pi/2]$ where it is one-to-one of course. And we call its inverse on this restricted domain the *arcsine* function or the *inverse sine* function.



sine on restricted domain

Here is a graph of $y = \arcsin x$.



arcsine (inverse sine) function

We will formalize in a definition what we have just described.

Definition 1: Let $-1 \le x \le 1$. Then $y = \arcsin x$ if and only if $\sin y = x$ and $-\pi/2 \le y \le \pi/2$.

That is, we read $y = \arcsin x$ as: *y is the angle (in radians) between $-\pi/2$ and $\pi/2$ whose sine is equal to x.*

Example 1: Here are some values of the arcsine function:

1. $\arcsin 0 = 0$
2. $\arcsin 1 = \pi/2$
3. $\arcsin(-1) = -\pi/2$
4. $\arcsin(-1/\sqrt{2}) = -\pi/4$
5. $\arcsin(1/2) = \pi/6$
6. $\arcsin 5$ is not defined because 5 is not in the range of the sine.

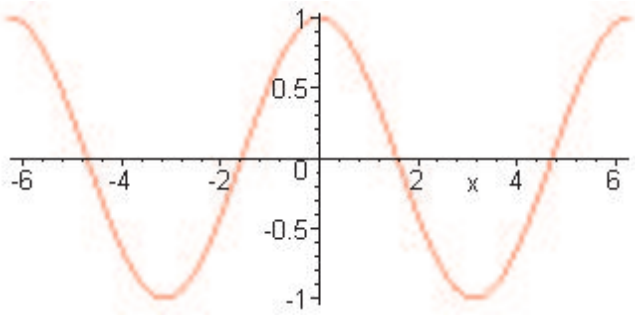
Example 2: We want to keep in mind that the sine and arcsine functions have an inverse function relationship but on a restricted domain:

1. $\arcsin(\sin(\pi/3)) = \pi/3$
2. $\arcsin(\sin(3\pi/4)) = \pi/4$. Note that the answer is not $3\pi/4$ because $3\pi/4$ is outside the domain $[-\pi/2, \pi/2]$ to which we restricted the sine.

Notation: We have been writing the inverse sine function as $y = \arcsin x$. There is an alternative notation that can be used interchangeably: $y = \sin^{-1} x$. Just be careful not to interpret this to mean the reciprocal of the sine.

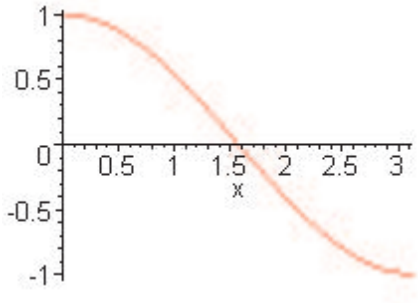
The Arccosine

Like the sine, the cosine function also fails to be invertible unless we restrict its domain. Corresponding to each value of y in the range of the cosine is an infinite number of x-values. We show part of the graph below.



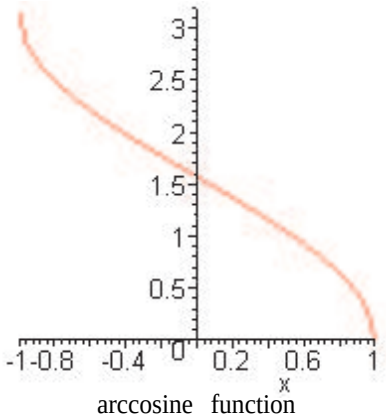
cosine function is not one-to-one

As before, we need to restrict the domain to an interval where the function is one-to-one. In the case of the cosine, the agreed upon convention is to restrict the domain to the interval $[0, \pi]$.



cosine on restricted domain

Now, we can define the inverse function by swapping domain and range and reversing the action of the cosine. That is, if the cosine maps x to y, then the arccosine maps y to x.



arccosine function

Let's summarize what we have done by collecting the information in a definition.

Definition 2: Let $-1 \leq x \leq 1$. Then $y = \arccos x$ if and only if $\cos y = x$ and $0 \leq y \leq \pi$.

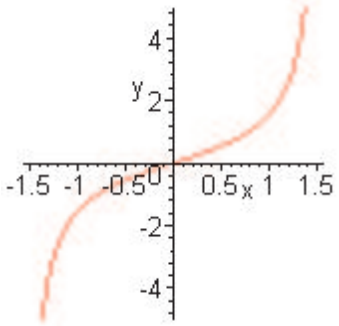
We read $y = \arccos x$ as: *y is the angle (in radians) between 0 and π whose cosine is equal to x.* As in the case of the sine, instead of $y = \arccos x$ we can just as well write $y = \cos^{-1} x$.

Example 3: Some values of the inverse cosine are:

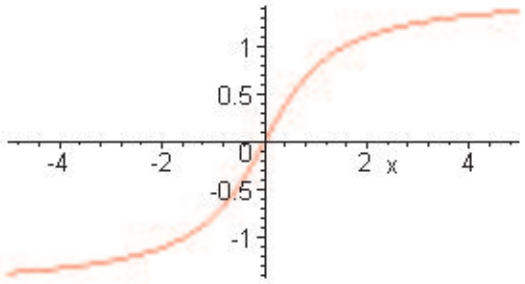
1. $\arccos 1 = 0$
2. $\arccos(-1) = \pi$
3. $\arccos 0 = \pi/2$
4. $\arccos(-1/2) = 2\pi/3$

Check them for yourself, remembering the way in which we restricted the domain of the cosine.

The Arctangent



Even though the tangent function is not one-to-one on its domain, it is one-to-one on the branch that passes through the origin. By convention, we use this branch to define the inverse. That is, to define the inverse function for the tangent, we restrict the domain to the open interval $(-\pi/2, \pi/2)$. Then the tangent is one-to-one and we can define the arctangent function accordingly.



inverse tangent

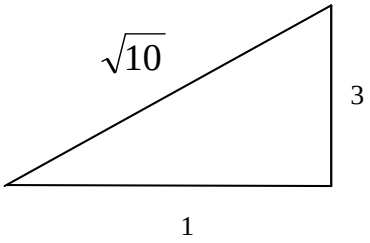
Definition 3: Let $-\infty < x < \infty$. Then $y = \arctan x$ if and only if $\tan y = x$ and $-\pi/2 < y < \pi/2$.

So, according to the definition, we read $y = \arctan x$ as: *y is the angle (in radians) strictly between $-\pi/2$ and $\pi/2$ whose tangent is equal to x.*

Example 4: Here are a few values of the arctangent:

1. $\arctan 0 = 0$
2. $\arctan 1 = \pi/4$
3. $\arctan(-1) = -\pi/4$
4. $\arctan(\tan(3\pi/4)) = \arctan(-1) = -\pi/4$

Example 5: To find $\cos(\arctan 3)$, we make a triangle from the information $y = \arctan 3$ and use the Pythagorean Theorem to complete it. Thus, we see that the cosine of the angle (and hence the answer to the problem) is $1/\sqrt{10}$.



Derivative of the Arcsine and the Arctangent

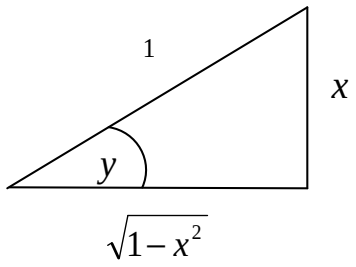
Arcsine: Now that we have defined inverse functions for some of the trigonometric functions, we will find their derivatives. In particular, we will discover some new antiderivatives that come up frequently in integration problems.

Theorem 1: Let $y = \arcsin x$. Then $\frac{dy}{dx} = \frac{1}{\sqrt{1-x^2}}$.

The proof starts with the defining relationship between the sine and arcsine functions: $y = \arcsin x \Leftrightarrow \sin y = x$ and $-\pi/2 \leq y \leq \pi/2$. Now, beginning with the right hand side and differentiating implicitly, we get:

$$\begin{aligned} \sin y &= x \\ \cos y \, \frac{dy}{dx} &= 1 \\ \frac{dy}{dx} &= \frac{1}{\cos y} \end{aligned}$$

Now, make a triangle using the relationship $\sin y = x$ to determine two of the sides, and apply the Pythagorean Theorem to find the third.



Thus, $\cos y = \sqrt{1-x^2}$. [It may appear that we are assuming that y is a first-quadrant angle. But because the cosine is positive in the first and fourth quadrants (that is, for $-\pi/2 \le y \le \pi/2$), this expression for the cosine is indeed correct for all values of y under consideration.] Hence, we have the desired result, namely, $\frac{dy}{dx} = \frac{1}{\sqrt{1-x^2}}$.

As usual, this theorem has a chain-rule form, and a companion integral formula. For, if u is a function of x, then

$$\frac{d}{dx}\arcsin u = \frac{1}{\sqrt{1-u^2}}\frac{du}{dx}$$

$$\int \frac{1}{\sqrt{1-u^2}}\,du = \arcsin u + C$$

Example 6: If $y = x\arcsin x$, then from the product rule we have $y' = \arcsin x + \frac{x}{\sqrt{1-x^2}}$.

Example 7: If $y = e^{\arcsin x}$, then $y' = e^{\arcsin x}\frac{1}{\sqrt{1-x^2}}$.

Example 8: Find $\int \frac{1}{\sqrt{a^2-x^2}}\,dx$, where a is a constant, by calculating the derivative of $\arcsin\frac{x}{a}$. Following the instructions and using the chain rule, we get:

$$\begin{aligned}\frac{d}{dx}\arcsin\frac{x}{a} &= \frac{1}{\sqrt{1-(x/a)^2}}\,\frac{1}{a}\\ &= \frac{a}{\sqrt{a^2-x^2}}\frac{1}{a}\\ &= \frac{1}{\sqrt{a^2-x^2}}\end{aligned}$$

Therefore, we can solve the integral given in the Example:

$$\int \frac{1}{\sqrt{a^2-x^2}}\,dx = \arcsin\frac{x}{a} + C$$

Example 9: Find $\int \frac{1}{\sqrt{3-x^2}}\,dx$. From the previous Example, the answer is

$$\int \frac{1}{\sqrt{3-x^2}}\,dx = \arcsin\frac{x}{\sqrt{3}} + C$$

Example 10: We can use parts to solve $\int \arcsin x\,dx$.

$u = \arcsin x$ $du = \frac{1}{\sqrt{1-x^2}}\,dx$	$dv = dx$ $v = x$

$$\int \arcsin x\,dx = x\arcsin x - \int \frac{x}{\sqrt{1-x^2}}\,dx$$

The new integral can be solved by substitution with $u = 1 - x^2$, and hence $du = -2x\,dx$.

$$\int \frac{x}{\sqrt{1-x^2}}\,dx = -\frac{1}{2}\int \frac{-2x}{\sqrt{1-x^2}}\,dx = -(1-x^2)^{1/2} + C$$

Thus,

$$\int \arcsin x\,dx = x\arcsin x + (1-x^2)^{1/2} + C$$

Arccosine: The derivative of the arccosine does not help us deal with integrals because we can use the arcsine in integration problems:

$$\frac{d}{dx}\arccos x = -\frac{1}{\sqrt{1-x^2}}$$

Instead of proving that result, we will go on to a proof of the derivative of the arctangent function. In spirit, all of these proofs are the same.

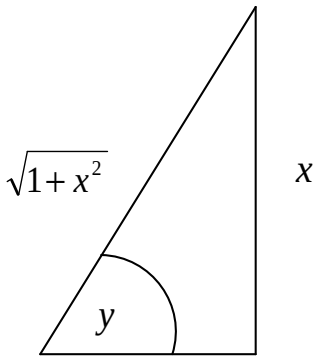
Arctangent: The arctangent function is defined through the relationship $y = \arctan x \Leftrightarrow \tan y = x$ and $-\pi/2 < y < \pi/2$. As we did in proving the derivative of arcsine, we will begin with the right hand side and differentiate implicitly.

Theorem 2: Let $y = \arctan x$. Then $\frac{dy}{dx} = \frac{1}{1+x^2}$.

The outline of the proof is the same as that for the derivative of the arcsine.

$$\begin{aligned} \tan y &= x \\ \sec^2 y\,\frac{dy}{dx} &= 1 \\ \frac{dy}{dx} &= \frac{1}{\sec^2 y} \end{aligned}$$

We now draw a triangle using the relationship $\tan y = x$ to determine two sides, and use the Pythagorean Theorem for the third.



Thus, $\sec y = \frac{1}{\cos y} = \sqrt{1+x^2}$. Hence $\frac{dy}{dx} = \frac{1}{1+x^2}$, and the proof is complete.

The chain-rule form of the theorem and the companion integral formula follow: If u is a function of x, then

$$\frac{d}{dx}\arctan u = \frac{1}{1+u^2}\frac{du}{dx}$$

$$\int \frac{1}{1+u^2}\,du = \arctan u + C$$

Example 11: If $y = \arctan x^2$, then by the chain rule we have $y' = \frac{1}{1+x^4}2x$.
Example 12: If $y = \arctan e^x$, then $y' = \frac{1}{1+e^{2x}}e^x$
Example 13: If a is constant, then the chain rule yields:

$$\frac{d}{dx}\arctan\frac{x}{a} = \frac{1}{1+(\frac{x}{a})^2}\frac{1}{a} = \frac{a}{a^2+x^2}$$

Thus

$$\int \frac{1}{a^2+x^2}\,dx = \frac{1}{a}\,\arctan\frac{x}{a} + C$$

Example 14: From the previous Example, we get that $\int \frac{1}{5+x^2}\,dx = \frac{1}{\sqrt{5}}\,\arctan\frac{x}{\sqrt{5}} + C$.
Applet: Calculator: Values of Elementary Functions **Try it!**
Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**

4.11 Case Study: Flood Watch

Animation: Flood Watch **To get you going on the Case Study!**

In this section, we have learned that if we are given the continuous derivative f' of a function on the interval $[a, b]$, then one version of the Fundamental Theorem of Calculus states that

$$\int_a^b f'(x)\,dx = f(b) - f(a)$$

This formula is very useful if we can find an antiderivative for f'; that is, if we have an explicit representation of f. However, in an applied setting, we may not have a formula for f and thus cannot use the fundamental theorem directly. The derivative $f'(x)$ may be given at only a finite number of points of the interval, and that may be all that is known. That and the fact that $f'(x)$ is a continuous rate of change of a function f. In these cases, we want to keep in mind that we can approximate the integral using a Riemann sum. This is the procedure that we will follow in the CSC of this section where the derivative is a non-negative function on an interval. We will use the Riemann sum

$$\sum_{i=1}^{n} f'(x_i)\Delta x_i$$

as an approximation to the area under the graph of the continuous function f'.

As usual, the purpose of a CSC is to consider a real application of calculus, with real data. In this section, we will work with our earth scientist colleagues to study problems related to river flooding . Without being too specific about our objective at this time, we will state it in general terms so that we have an idea where we are going.

Objective: From data collected ten years apart, to determine if the Gorge River has an increased or decreased likelihood of flooding.

In order to make the objective more concrete, we are going to have to learn how earth scientists collect and assemble information about rivers. We will play the role of mathematicians working with geologists to provide the mathematical analyses that regional and urban planners need for the management of resources having an impact on watersheds and the like. We have to be careful, though, because we will find that mathematicians and earth scientists sometimes use the same terminology for different concepts. Because this is a CSC and we want to get practice applying mathematics to a different field, we will use the earth-science terms. But be careful to translate them into their mathematical form so that you know what calculus tools to apply.

In the next paragraph, we describe all of the terms we will be using. Often the physical units (for example, volume of water per unit time) will help with the language conversions from earth sciences to mathematics.

River Flooding: Background

The flow of water on the surface of the earth in rivers and streams has an important impact on humans. Many of these water bodies provide important resources as navigation routes and sources of fresh water for residential and industrial use. The erosion force of the streams is responsible for the formation of many of our land forms both by removal as well as deposition of material. Perhaps their most dramatic impact is flooding, when the discharge (volume of water per unit time) exceeds the normal carrying capacity of the channel, and the water spills out onto the adjacent flood plain. In an ideal world, flood plains would be reserved for rivers and perhaps agriculture which rejuvenates from fresh silt and natural mineral fertilizers left by the river. However, the transportation and recreational value of rivers has lured many human settlements

The CSC is adapted from Skinner, B. J. and Porter, S. C., 1962, *The Dynamic Earth, An Introduction to Physical Geology*, 2nd ed., Wiley, 570 pp.; and from Leopold, L. B., 1968, *Hydrology for Urban Land Planning*, a guidebook on the hydrologic effects of urban land use, U.S. Geological Survey Circular 554. We also owe special thanks to Professor Dick Birnie, Department of Earth Sciences, Dartmouth College, for collaborating on this work.

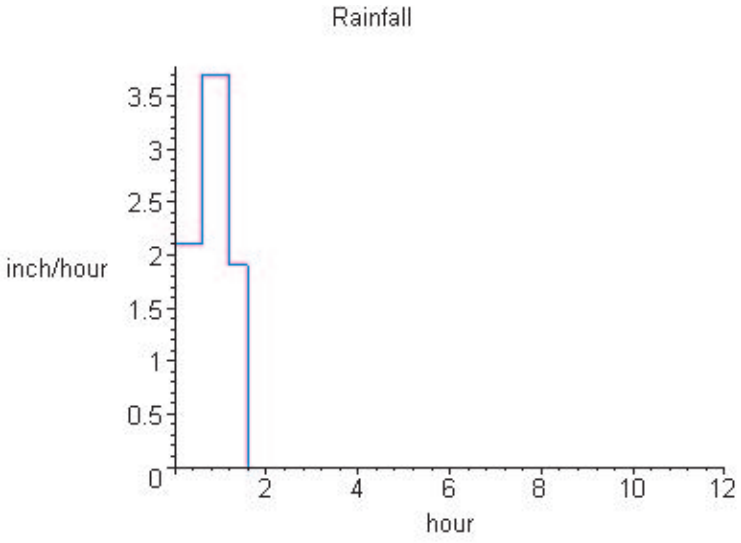
to river banks. Over time, many of these settlements have grown into major metropolitan centers such as St. Louis and New Orleans on the Mississippi River. When a river floods, the human impact is disastrous. People who live and work along a river should not be surprised by a flood. Floods are regular and expected events; they have happened in the past and they will continue to happen in the future.

Earth Scientists make continuous observations of the discharge (volume of water per unit time) of a stream. These observations lead to an understanding of the flow characteristics of the river and its watershed (area that drains into the river) and allow predictions of future behavior. This concept repeats itself throughout Earth Science: observations (measurements) are made, an understanding (perhaps a mathematical model) is derived by analyzing (tabulating, graphing, statistically processing) the observations; and then predictions are extrapolated from the model.

The fundamental way to display the observations of stream discharge is a *hydrograph*. A hydrograph has two components: the amount of rainfall (depth per unit time) and discharge (volume of water per unit time) (see the example below). Prior to a rain storm, the stream will be flowing at some background level of discharge known as base flow. However, following a period of heavy precipitation, the rain falling in the watershed drains into the stream, and the discharge increases over time. The plot below shows the amount of discharge over a 12 hour period following an intense rainstorm lasting about 96 minutes. The discharge increases rapidly to a peak about 3 hours after the storm and then gradually drops for the next 9 hours. The discharge of a stream does not rise immediately with the onset of precipitation, rather it takes time to flow across the watershed and into the stream. The difference in time between the *centroid* of the precipitation and the centroid of the discharge (runoff) is known as the *lag time* and that depends on the size and makeup of the watershed. Neglecting absorption by the ground, the area of the watershed times the depth of the rainfall equals the volume of discharge above the baseflow. If the total volume of discharge of the stream exceeds the carrying capacity of the channel, the stream overflows its banks and floods.

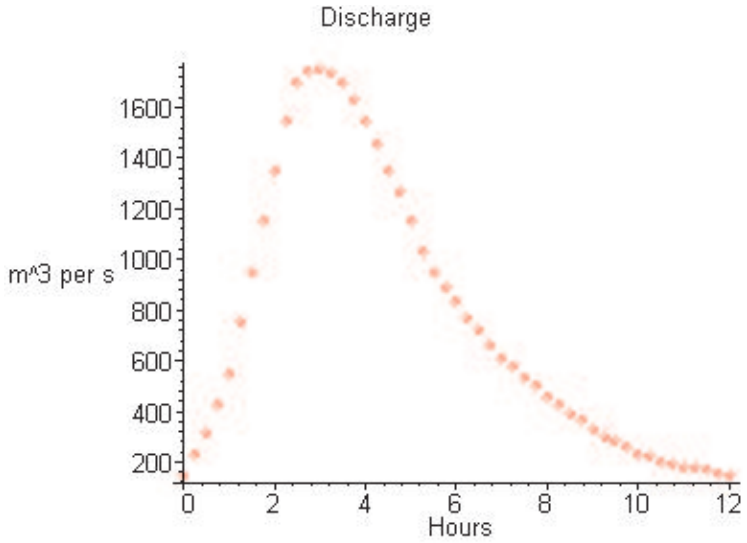
Example of a Hydrograph

A hydrograph consists of two parts: a plot of rainfall, and a plot of discharge. Here is a set of possible data giving the discharge of a specific stream caused by an intense amount of rainfall. First, the rainfall plot in inches per hour:



The rainfall is 2.1 $\frac{in}{hr}$ for the first 0.6 hours, then it is 3.7 $\frac{in}{hr}$ for the next 0.6 hours, and finally it is 1.9 $\frac{in}{hr}$ for 0.4 hours.

Next, the data representing the discharge D of a stream at specific instants of time beginning at the onset of the intense downpour of rain given above are as follows, where the discharge is measured in cubic meters per second:



Notice that although the discharge rises rather quickly, it drops slowly back to the base flow level of the stream.

To find the lag time, we need to compute the centroids of the plots of rainfall and discharge. The **centroid** of a plane region is the point (x_0, y_0) on which the region, thought of as a thin plate, would balance horizontally. To find the point, we use the formulas for the coordinates.

Theorem: Suppose the region of area A is defined as lying between two curves $y = f(x)$ and $y = g(x)$ where $f(x) \geq g(x)$ and $a \leq x \leq b$. Then its centroid is located at the point (x_0, y_0) given by

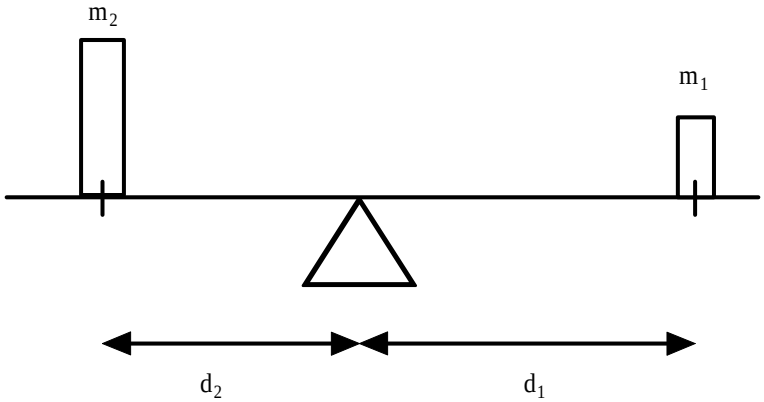
$$x_0 = \frac{1}{A}\int_a^b x(f(x) - g(x))\,dx$$

$$y_0 = \frac{1}{A}\int_a^b \frac{f(x)^2 - g(x)^2}{2}\,dx$$

Because the proof is instructive for our discussion of the hydrograph, we will justify this formula before proceeding further with the rainfall data.

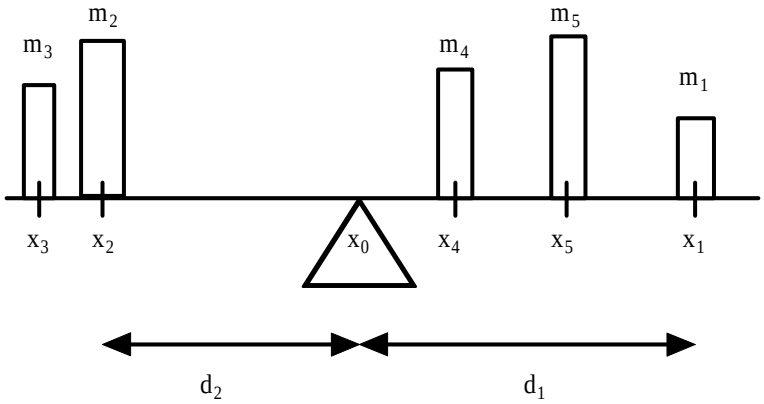
Derivation of Formulas for the Centroid

Suppose two masses m_1 and m_2 are placed on a see-saw, at distances d_1 and d_2, respectively, from the pivot-point. Then how far do we have to move one mass relative to the other so that the see-saw will balance?



This is an old problem that has a very simple solution: the see-saw will balance when $m_1 d_1 = m_2 d_2$. For example, if m_2 is large relative to m_1, then d_2 must be small relative to d_1. We all have experienced the

effects of this equation: the larger person has to sit closer to the pivot point to balance the smaller person sitting at the opposite end of the see-saw.



Suppose now that we consider n masses $m_1, m_2, m_3, \ldots, m_n$ located on a see-saw at the points $x_1, x_2, x_3, \ldots, x_n$, and suppose further that the pivot point is located at x_0. Then generalizing from the case of two masses, it turns out that balance is achieved when

$$\sum_{k=1}^{n} m_k(x_k - x_0) = 0$$

Note that if $x_k > x_0$, then $x_k - x_0$ is positive and is the distance of m_k from the pivot-point, while if $x_k < x_0$, then $x_k - x_0$ is negative and is the negative of the distance of m_k from the pivot-point. So, this equation is indeed a generalization of the one we started with involving only two masses. If we solve for x_0 in this equation, then we get that

$$x_0 \sum_{k=1}^{n} m_k = \sum_{k=1}^{n} m_k x_k$$

Calling $\sum_{k=1}^{n} m_k$ the *total mass*, we then have that

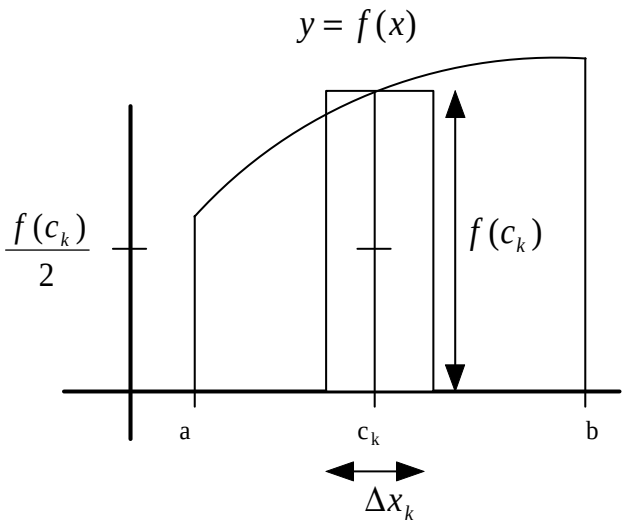
$$x_0 = \frac{\sum_{k=1}^{n} m_k x_k}{\text{total mass}}$$

By analogy with masses, if the rectangles pictured above are simply two-dimensional regions, we can write a similar formula for the coordinate of the pivot-point using areas:

$$x_0 = \frac{\sum_{k=1}^{n} x_k A_k}{\text{total area}}$$

where A_k is the area of the k^{th} rectangle. The product $x_k A_k$ is called the *moment* of the k^{th} rectangle.

Suppose now that we start with a function $y = f(x)$ defined on an interval $[a, b]$, and as usual we partition the interval into n subintervals with the points $x_0 < x_1 < x_2 < \cdots < x_n$. Let c_k be the midpoint of the k^{th} subinterval. Then $c_k A_k = c_k f(c_k) \Delta x_k$ is the moment of the k^{th} rectangle with respect to the y-axis.



We next recognize that the sum of the moments of the n rectangles is a Riemann sum whose limit is the definite integral (assumed to exist) of the function over the interval $[a, b]$:

$$\lim_{n\to\infty} \sum_{k=1}^{n} c_k f(c_k) \Delta x_k = \int_a^b x f(x)\, dx$$

In a similar manner, we can compute the moment of the k^{th} rectangle with respect to the x-axis. Because the y-coordinate of the center of the rectangle is $\frac{f(c_k)}{2}$, the moment with respect to the x-axis is $\frac{f(c_k)}{2} f(c_k)\Delta x_k$, that is, the y-coordinate of the center times the area of the rectangle. Once again, we recognize the sum of these moments as a Riemann sum, whose limit is a definite integral:

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{f(c_k)}{2} f(c_k) \Delta x_k = \frac{1}{2} \int_a^b f(x)^2\, dx$$

Finally, the point (x_0, y_0) whose coordinates satisfy

$$Ax_0 = \int_a^b x f(x)\, dx$$

$$Ay_0 = \frac{1}{2} \int_a^b f(x)^2\, dx$$

where A is the area under the graph of f, is called the *centroid* of the region. It is the balance point of the area of the region.

Back to Example of a Hydrograph

We had gotten to the point in our discussion of the hydrograph where we were about to compute the centroids of the rainfall and discharge data.

The rainfall plot is composed of three rectangles. We refer to the proof above of the centroid formula to find the x-coordinate of the centroid. The proof shows that we can calculate it as the sum of the midpoint of a rectangle times its area (the height of the rectangle times the width of the rectangle), all divided by the area of the rainfall plot.

The area of the rainfall data is

$$A = .6(2.1) + .6(3.7) + .4(1.9) = 4.24$$

Hence, we find that the coordinates of the centroid of the rainfall data are:

$$x_0 = \frac{1}{A}\left(.3(2.1)(.6) + .9(3.7)(.6) + 1.4(1.9)(.4)\right) \approx .81132$$

$$y_0 = \frac{1}{2A}\left((2.1^2)(.6) + (3.7^2)(.6) + (1.9^2)(.4)\right) \approx 1.45094$$

Applet: Flood Watch **Try it!**

To find the centroid of the discharge data, which is given in a convenient computational form such as an applet or a Maple worksheet, we use a Riemann sum approximation to calculate the integrals in question. First, letting $y_{\min}$ ($= 150$ in this example) be the base level of the flow, the (approximate) area A under the discharge graph is obtained as the sum

$$A = 3600\sum_{k=1}^{n}(f(c_k) - y_{\min})\Delta x_k \approx 7418$$

where we have shown explicitly the multiplier 3600 $\frac{sec}{hr}$ that is needed to make consistent the units on the vertical and horizontal axes.

Then, the x and y coordinates of the centroid of the discharge data due to rainfall are:

$$x_0 = \frac{1}{A}\sum_{k=1}^{n} c_k(f(c_k) - y_{\min})\Delta x_k \approx 4.294$$

$$y_0 = \frac{1}{2A}\sum_{k=1}^{n}(f(c_k)^2 - y_{\min}^2)\Delta x_k \approx 691.232$$

Note that x_0 is in hours, and y_0 in $\frac{m^3}{sec}$. The lag time is the difference of the x-values of the centroids of the discharge and rainfall data, or approximately 3.4827 hours. Moreover, if we wanted a measure of the intensity of the storm, we could compute the difference in the y-values of the centroids.

The CSC: Flood Watch

Now that we have considered an example of a hydrograph, and discussed the relevant mathematics, we are ready to state the objective, setup, and thinking and exploring issues for the CSC.

Objective: We are given two hydrographs for the Gorge River that have been recorded ten years apart. The overall objective is to analyze, compare, and interpret the two hydrographs, thereby reaching a conclusion about the increased or decreased likelihood of the Gorge River flooding.

Setup: We will be plotting the rainfall and discharge data, and assuming that the discharge represents a sample of measurements drawn from a continuous function. We will be using Riemann sums to calculate approximations to the area under the curve, and to the centroid of the discharge data. We will keep track of the various units of measurement to be certain that we have a consistent set.

Thinking and Exploring: We will be studying each hydrograph, and then comparing our findings for the two. The rainfall is measured in units of inches per hour, the discharge in units of cubic meters per second, and the horizontal time line in hours.

Some of the questions we will consider under Thinking and Exploring are: What is the average base flow? What is the time at which the discharge is maximum? What is the maximum discharge? What is the

discharge due to base flow over the 12 hour observation time? What is the discharge due to the rain storm over the 12 hour observation time? What is the total discharge over the 12 hour observation time? About how long did the rain storm last? What is the center of mass of the precipitation? What is the center of mass of the discharge? What is the lag time? What is the area of the watershed?

Applet: Flood Watch **Try it!**

As ususal, the last step in a CSC is to write a summary of the investigations.

Interpretation and Summary: After studying the two hydrographs of the Gorge River, and thinking about the mathematical tools, it is time to interpret and summarize the mathematical results in terms of the original objective.

Pretend that your synopsis is going to appear in the next issue of a magazine such as *Scientific American*. Include enough details so that a reader would learn what the major issues of the report are, and how you went about addressing them. What will you want to tell readers about your success with regard to the original stated objective of the investigation? Be sure to write in complete sentences using correct rules of standard English grammar.

To get you started, here are two questions that definitely need to be answered:

1. What changes in the watershed might have occurred in the 10 years between the hydrographs to account for the different pattern of the more recent hydrograph relative to the earlier one?

2. How might these changes influence the likelihood of the Gorge River flooding?

What advice do you have for public policy makers? Make the report interesting, compelling. Ask yourself: Would a policy maker be able to understand it, and feel compelled to follow its recommendations?

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

Chapter 5

Culminating Experience

5.1 Case Study: Sleuthing Galileo

Animation: Sleuthing Galileo **To get you going on the Case Study!**

Now that we have completed our first course in calculus, we should be able to take on a project that will put our knowledge to the test. The present CSC does just that. Constituting a real application of calculus with real data, the CSC asks us to consider some experimental data that Galileo collected in his laboratory. We will adopt a modern point of view and try to make sense of it using calculus. [The CSC is adapted from *Teaching Statistics with Data of Historic Significance: Galileo's Gravity and Motion Experiments*, by David A. Dickey and J. Tim Arnold, Journal of Statistics Education v.3, n.1 (1995).]

As we know by now, the big breakthrough of the seventeenth century was an understanding of motion. We have seen how powerful the concepts of rate of change, derivative, and integral are in the field of study that emerged. And although there are many applications of calculus throughout the sciences today, the beginnings of the subject can be traced back to Galileo (1564-1642) and his rolling ball experiments at the end of the sixteenth century.

Galileo was interested in solving practical problems, many of them coming from the field of gunnery and ballistics. To do this, he had to have a good understanding of the behavior of cannon balls in flight. Therefore, Galileo conducted many experiments to simulate in his laboratory the field conditions of a projectile of the day. We already have discussed a Galileo-like experiment in Section 2. There we observed, just as Galileo discovered, that the acceleration g due to gravity is constant. With modern methods of measurement we now know that g is about 9.8 meters per second per second or 32.2 feet per second per second.

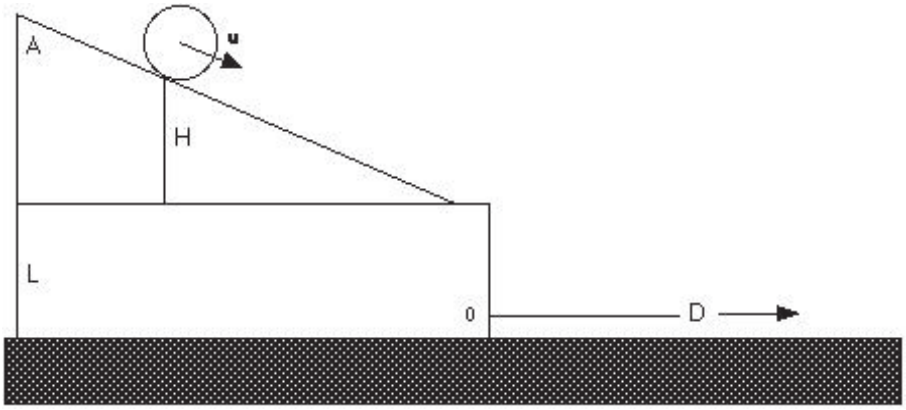
When we approached this question using average velocities, we saw that the experimental error cahracteristic of typical data would not have allowed us to conclude the constancy of the acceleration. However, Galileo did not have calculus. In the spirit of Section 1, he appears to have reached his conclusion by modeling his data with an elementary function. Instead, we are going to analyze his experiment using calculus and see if we can interpret the data that he has recorded.

We will be acting as modern day detectives attempting to reconstruct what went on in his lab and what motivated his thinking. We will be analyzing the problem, making conjectures, drawing conclusions, and giving possible interpretations of his data. We will be trying to get inside Galileo's head and think some of his own thoughts. Because he initiated many of the ideas of calculus as they relate to motion, his thinking and ours should go along similar lines.

Galileo was a 31 year old lecturer at the University of Padua when he began experimenting and writing a paper on the path of projectiles. He was 42 years old when he reported on these experimental studies and gave an explicit mathematical formulation of the motion of falling objects.

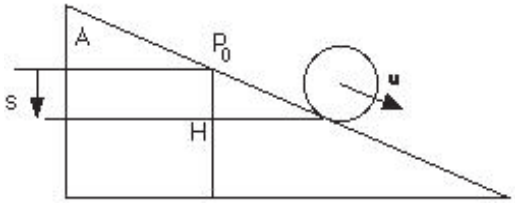
In the experiment we are going to consider, Galileo used an inclined ramp that ended in a narrow horizontal shelf at the lower end of the ramp. A groove was etched into the ramp, so that a ball released on the ramp would roll down following the groove all the way to the shelf. When it hit the shelf, which we will take to be the edge of the table on which the ramp sits, the ball would roll horizontally to the end before leaving the edge and falling to the floor some distance below. The velocity of the ball, as it rolls down the

ramp, has two independent components: one horizontal, the other vertical. When the ball hits the table, the velocity of the ball in the vertical direction is set to zero; the velocity of the ball in the horizontal direction is equal to the horizontal velocity with which it leaves the ramp.



One of Galileo's findings, later stated by Newton as his *First Law of Motion*, is as follows: An object without outside influence continues to move indefinitely with the same speed and direction that it has originally. Thus, because there are no influences on the ball in the horizontal direction, the horizontal speed is constant from the time the ball leaves the ramp until it hits the floor.

When the ball leaves the table, the distance it travels before hitting the floor depends on the height at which it is released to roll down the ramp. We can see this from the table below of data that Galileo recorded. (You will also explain this observation later on theoretical grounds, so start thinking now about why it is true.) Let H be the release height of the ball above the table, and let D be the horizontal distance traveled along the floor, measured from a point on the floor directly below the edge of the table.



Then Galileo recorded five measurements of D and of H as follows, where the units of measurement are punti (*points* in Italian):

Release Height H Above Table	Horizontal Distance D Traveled
1000	1500
828	1340
800	1328
600	1172
300	800

Using the tools of calculus we have learned this term, we will develop a model of the ball's motion and then try to interpret Galileo's data in light of this model. Here is a statement of the specific objective.

Objective: Your task is to make mathematical sense of Galileo's data, and to formulate some ideas about how this particular experiment might have contributed to a calculation of g, the acceleration due to gravity. In particular, you will be

1. modeling the rolling ball experiment starting from a conservation of energy argument,
2. verifying that Galileo's experimental data are consistent with the results of the theory, and
3. discussing any inconsistencies and/or open questions, and in particular the role of time measurements in the determination of g.

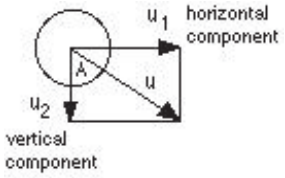
As is true of all Case Studies, the last step will be to write a report summarizing your findings. The structure of the CSC will provide a guide to the important issues that need to be considered.

Setup: Your Setup description can, and should, be brief, but this is where you describe the kinds of mathematical facts you are looking for and the methods you expect to use. In this CSC, because of the open-endedness of the investigation, it might be a good idea first to complete the *Thinking and Exploring* activities outlined below. Then you should be in a position to understand the mathematical issues and techniques, and be able to describe them here in the *Setup* part.

Thinking and Exploring: To help structure your thinking, we have divided this part of the CSC into four parts.

Applet: Sleuthing Galileo **Try it!**

Part I: The Mathematical Model: Let s be the vertical drop of the ball whose velocity down the ramp is u.



Then the gain in kinetic energy $\frac{mu^2}{2}$ equals the loss of potential energy mgs, where m is the mass of the ball and g is the acceleration due to gravity. Here are the steps you need to complete and understand to develop the model:

The ball on the ramp:

1. Show that $u = \sqrt{2gs}$.
2. Show that the terminal velocity u_T of the ball at the bottom of the ramp is $u_T = \sqrt{2gH}$.
3. Show that the horizontal and vertical components of velocity, u_1 and u_2, respectively, are given by trigonometry as $u_1 = u \sin A$ and $u_2 = u \cos A$.
4. Show that the terminal horizontal and vertical components of velocity at the bottom of the ramp are $u_{1T} = \sqrt{2gH} \sin A$ and $u_{2T} = \sqrt{2gH} \cos A$.

The ball in free flight:

1. Once the ball leaves the ramp, its horizontal velocity is constant, equalling the terminal horizontal velocity on the ramp u_{1T}. Explain.
2. The horizontal distance D can be written as $D = u_{1T} t_f$ where t_f is the length of time the ball is in free-flight. Explain.
3. The time t_f is the length of time it takes the ball to drop the vertical distance L from the table to the floor. Explain.
4. Show that $t_f = \sqrt{\frac{2L}{g}}$.

D as a function of H:

1. Show that $D = 2\sqrt{LH} \sin A$.

Part II: Galileo's Data: How do we verify that Galileo's data are a result of the model we just developed when we don't know the ramp angle A or the table height L? Here is where some detective work is required, and we will turn to a computing tool such as Maple to assist us.

1. Show that in our model, we can equally well regard H as a quadratic function of D: $H = \frac{D^2}{4L \sin^2 A}$.
2. Show that $\ln H = 2 \ln D - \ln k$ where $k = 4L \sin^2 A$.

3. Explain: With a computing tool such as Maple, if we fit a line $y = ax + b$ to the $(\ln D, \ln H)$ data, then assuming that H is a quadratic function of D in the data, a should be 2.

4. Continuing, if the data are consistent with the model, then show that $k = e^{-b}$ and $A = \arcsin\sqrt{\frac{k}{4L}}$.

5. Go ahead and fit a line to the natural log of the data. Do you believe from the result that the data exhibit H as a quadratic function of D? Explain.

6. Sensitivity Analysis: The parameter ϵ in the Maple program allows us to adjust the value of H in the data. That is, we fit a line to the $(\ln D, \ln H - \epsilon)$ data. Explain what ϵ represents physically. How do different values of ϵ affect a? With $L = 1100$, experiment to find a value of ϵ that yields a value of a as close to 2 as you can. What is the corresponding angle A of the ramp? Give a plausible explanation for the value of a you found for the original data ($\epsilon = 0$).

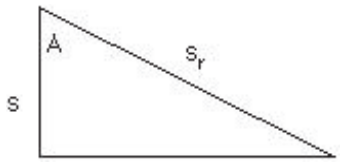
7. Speculating About A and L: What are some reasonable choices of A and L? Try to put yourself in Galileo's position when you think about this question. Review the physical situation (the ramp on the table, the ball rolling down the ramp, etc.) and the kind of flight patterns he hopes to see. In the Maple program, start by choosing a value for L and calculating the resulting angle A. Are there pairs A and L that are more plausible than others? Explain and give examples. Also, explain how A is being calculated.

Part III: The Total Travel Time and g: Although we have used both time and the gravitational acceleration g in developing our model, note that neither of them appears in the functional relationship between H and D. In Galileo's day, it was certainly easier to measure distances than times, but if we want to calculate g, then the issue of time measurements cannot be avoided. Why? The purpose of this part is to describe a procedure for determining g using our same ramp-table-floor setup.

Suppose we have a stopwatch that we start at the instant we release the ball on the ramp, and stop when we see/hear it hit the floor. Then we will show that we can determine g. Here are the steps.

1. The total time of travel is the time t_r on the ramp plus the time t_f of free-flight. (We are ignoring the time it takes the ball to roll off the edge of the table by making the assumption that it is insignificant. You should consider whether this assumption is warranted.) Recall the formula for t_f developed earlier. What is it?

2. To find t_r, look at the sketch below. We call s_r the distance the ball rolls on the ramp corresponding to a vertical drop of s. Then show that $s = s_r \cos A$.



3. Next show that $u = \sqrt{2gs_r \cos A}$.

4. Show that the solution to the above differential equation (with $s_r(0) = 0$) is $s_r = \frac{1}{2}gt^2 \cos A$.

5. Show that $s(t) = \frac{1}{2}gt^2 \cos^2 A$.

6. Show that $t_r = \frac{1}{\cos A}\sqrt{\frac{2H}{g}}$.

7. Explain why the formula for t_r makes sense by discussing its meaning as A varies, $0 \le A < \frac{\pi}{2}$.

8. Show that the total time of travel of the ball from beginning to end is $T = \frac{1}{\cos A}\sqrt{\frac{2H}{g}} + \sqrt{\frac{2L}{g}}$.

9. If $H = L$, find a formula for g. Use the Maple program to verify your formula for several equal values of H and L. Describe a specific experiment to determine the value of g.

10. Assuming g is a constant, what happens to the total time of travel as $A \to \frac{\pi}{2}$? as $A \to 0$? Discuss both the time of travel on the ramp and the time of free flight.

11. Do you think it is plausible that Galileo measured g in the way you described in number 9 above? Explain.

Part IV: The Path of the Ball: It would be nice to have a graphical representation of the path of the ball. After all, a graph of D vs. H does not seem all that illuminating from the viewpoint of what we see with our eyes. When we watch the ball travel down the ramp and then fall in free-flight to the floor, it is the path of the ball that our senses try to capture. The path of the ball describes the position of the ball at any instant of time.

The following steps will describe the x and y coordinates of the ball as a function of time, that is, the collection of points $(x(t), y(t))$, $0 \le t \le t_r + t_f$.

On the ramp:

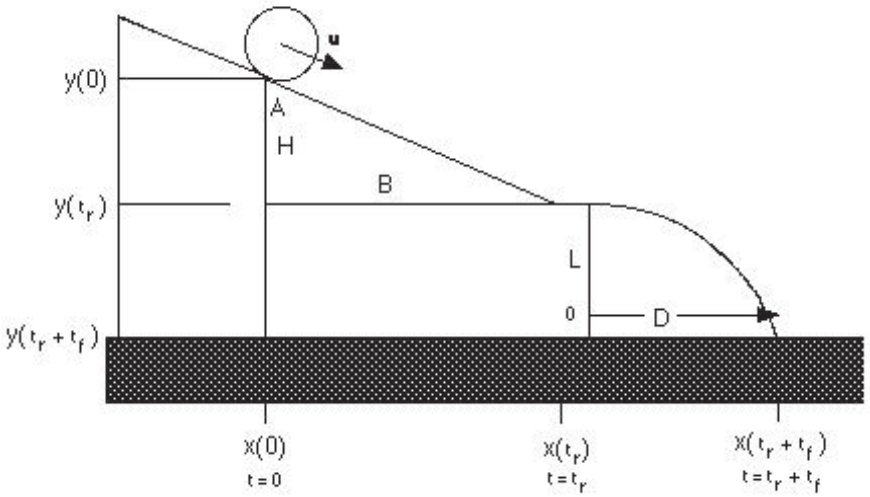
1. We found earlier that $s_r = \frac{1}{2}gt^2 \cos A$. Show that $u = gt \cos A$.

2. Show that $u_1 = \frac{1}{2}gt \sin 2A$ and that $u_2 = gt \cos^2 A$.

3. Show that $x(t) = \frac{1}{4}gt^2 \sin 2A$ and that $y(t) = \frac{1}{2}gt^2 \cos^2 A$.

4. Show that $x(t)^2 + y(t)^2 = s_r(t)^2$.

Overall: (See next sketch)



1. Explain why the following piecewise representations of $x(t)$ and $y(t)$ make sense:

$$x(t) = \begin{cases} \frac{1}{4}gt^2 \sin 2A & \text{if } t \le t_r \\ \sqrt{2gH}(t - t_r) \sin A + B & t > t_r \end{cases}$$

$$y(t) = \begin{cases} H + L - \frac{1}{2}gt^2 \cos^2 A & \text{if } t \le t_r \\ L - \frac{1}{2}g\,(t - t_r)^2 & t > t_r \end{cases}$$

2. Choose and fix your preferred values of A and L. Use the Maple program to plot the path of the ball for the 5 values of H in Galileo's data. Compare the resulting values of D in the plots and the data.

3. For the same values of A and L you just chose, fix a value of H. Plot the path for values of g equal to 10, 20, 100, 1000. What do you find? Explain and justify.

As you should have come to expect, the last step in a CSC is to write a summary of the investigations.

Interpretation and Summary: Summarize the entire investigation. Pretend that your synopsis is going to appear in the next issue of a magazine such as *Scientific American*. Include enough details so that a reader would learn what the major issues of the study are, and how you went about addressing them, especially in relation to the stated goals of the Objective. Be sure to write in complete sentences using correct rules of standard English grammar (no Italian or Latin, please).

Exercises: Problems **Check what you have learned!**

Videos: Tutorial Solutions **See problems worked out!**

Appendix A

List of Applets

Applets listed in particular sections have functionality appropriate to that material. Whereas tools are more generally useful and may apply in several different sections.

Section 1.1
- Falling Object
- Least Squares Fitting

Section 1.2
- Symmetry: Odd and Even Functions

Section 1.3
- Stretching Graphs
- Shifting Graphs
- New Functions from Old
- New Functions from Old Game
- Arithmetical Operations on Functions
- Inverse Functions

Section 1.4
- Definitions of sin(x) and cos(x)
- Trigonometric Identities

Section 1.5
- Comparing Exponential Functions

Section 1.6
- Fitting AIDS Data

Section 2.1
- Average Velocity
- Derived Function

Section 2.3
- Limits of Functions

Section 2.5
- Continuity of Functions

Section 2.6
- Secant and Tangent Lines

Section 2.9
- Limit of sin(x)/x as x approaches 0

Section 2.10
- Mean Value Theorem

Section 2.13
- Newton's Method

Section 2.4
- Best Linear Approximation

Section 3.1

- Slope Field

Section 3.4
- Euler's Method

Section 3.5
- Curve Sketching: Increasing/Decreasing
- Curve Sketching: Concavity

Section 3.6
- Euler Population Predictions

Section 4.1
- Approximating Areas: Inscribed Polygons
- Approximating Area: Using Rectangles
- Accumulation: River Flow
- Accumulation: Distance Traveled

Section 4.3
- Mean Value Theorem for Integrals

Section 4.11
- Flood Watch

Section 5.1
- Sleuthing Galileo

Tools
- Function Grapher
- Riemann Sums
- Numerical Integration
- Calculator: Values of Elementary Functions