1. A BRIEF SUMMARY OF CALCULUS

Calculus is one of the greatest intellectual achievements of humankind. It allows us to solve mathematical problems that cannot be solved by other means, and that in turn allows us to make predictions about the behavior of real-world systems that we could not otherwise make. But beyond its usefulness, calculus has an elegant beauty that leads mathematicians to view it as a work of art.

A full presentation of calculus requires the equivalent of four or more standard-length American college courses, although this is usually achieved through three one-semester courses. Because many of the most important mathematical methods in the life sciences are calculus-based, the reader needs to know some calculus in order to proceed. Fortunately, much of the material in the full development of calculus is not vital for the development of the these mathematical methods. The typical reader of this material has probably taken only the first course in calculus, and has therefore seen barely one third of the full presentation. For some readers, this study of calculus may be part of the distant past. We therefore begin with a very brief summary of the calculus background required for the mathematical work that follows. In subsequent chapters, the reader will be expected to have a solid understanding of the material of this chapter.

Isaac Newton was using calculus by 1666, but solid mathematical definitions of the core concepts of calculus were first published by Augustin Cauchy in 1840. Careful development of calculus theory is the subject of *analysis*, which is a very important branch of pure mathematics. According to mathematical dogma, nothing can be said to be true until it has been proven.[1] Hence, the mathematics community is almost unanimous in the view that all mathematical instruction should be based on definitions, theorems, and proofs. This is just the right attitude to take about many areas of mathematics. But it is harmful to the student of calculus. We will adopt a pragmatic definition of truth, in which something is true if it works. This was good enough for Newton, Gottfried Leibniz, and other developers of calculus, and it should be good enough for the calculus student as well. Those readers who want to see mathematical proofs can find them in any standard calculus text.

This chapter can be subdivided into four portions. The first portion consists of Section 1, which introduces the notion of continuous mathematical systems as contrasted with the more familiar discrete systems of mathematics. The central breakthrough of calculus is the method for developing notions of continuous mathematics from notions of discrete mathematics, using the limit process. The overall pattern that we use to create continuous mathematics from discrete mathematics is the focus of this first section. Sections 2 and 3 comprise the second portion of this chapter, which is devoted to the development of the two central concepts of calculus: the derivative and the definite integral. These concepts are defined as the solutions of geometry problems, and then the geometry problems are solved by the procedure introduced in Section 1. Sections 4 and 5, consisting of the principal computational techniques of calculus, is the third part of the chapter. The final part of the chapter treats the applications of the derivative and the definite integral in a very broad way.

After studying this chapter, you should be very comfortable with the concepts of the derivative and the definite integral and how they are applied to a variety of problem types and have an intuitive feel for continuous mathematics. You should be able to compute simple derivatives and definite integrals, but you do not need to be an expert at calculus computations.

[1] "A mathematician is someone who believes that one should not drive a car until after he has built one himself." (G. Ledder)

1.1 Continuity and Limits

The central concepts of calculus are continuity and limits. Unfortunately, these concepts are very difficult to define. Mathematics did not acquire a full understanding of continuity and limits until more than 100 years after the methods of calculus were discovered. This situation poses a pedagogical problem for anyone who would try to teach calculus. Do we begin with the mathematical definitions, because good definitions are an essential mathematical foundation or do we proceed without mathematical definitions, because the calculus methods can be developed without them?

Mathematicians as a group needed to be able to do calculus before they could create a sound mathematical foundation for the subject. It seems unreasonable to expect the individual student to do what the mathematical community could not so. Hence, we will make no attempt to give a sound mathematical treatment of continuity and limits. Instead, we will present these concepts through an intuitive approach.

Discrete and continuous quantities

There are two basic processes that we generally use to indicate quantity: counting and measuring, where we are using "measuring" to refer specifically to analog methods such as measuring the length of an object with a ruler. The important characteristic of measuring that distinguishes it from counting is that all values within some interval are possible measurements. One could, for example, cut a piece of string of length π by laying it on a circle of diameter 1 and then use the string to make a mark on a ruler to indicate the length π. We can't offer a procedure for measuring a length of e, but it is possible in principle that a ruler could have a mark that measures out any length shorter than the ruler.

Given this restricted use of the concept of measuring, quantities that are determined by counting are **discrete** and quantities that are determined by measuring are **continuous**.[1] Note that we are not defining the terms "discrete" and "continuous" here; we are just applying them to the methods used to indicate quantity.

Example 1.1.1
The number of individuals in a population is discrete because only non-negative integers are possible. The location (say the exact longitude and latitude) of each individual is continuous because all values are possible within the spatial domain occupied by the population. ◇

Perhaps the best way to understand the difference between discrete and continuous is with a thought experiment. Suppose we use computer software to plot a smooth graph; say we choose $y = x^2 + 3x$ with the viewing window $-3 < x < 3$ and $-4 < y < 20$. Any such graph is created by plotting such a large number of points that you can't see the space between the points. There is an appearance of a smooth curve simply because your eye can't resolve the marks into distinct points, but the actual data plotted in the graph is discrete. Figure 1.1.1a is a plot consisting of 300 points.
Now suppose we lock in the coordinates of the points used to create the graph, so that we will always be examining the same discrete data set. Next, we draw a rectangular box centered on

[1] In actual practice, measurements must be reported as discrete approximations, rounded off to the nearest integral number of some unit of measure. Note that we can make the unit as small as we like, measuring time in milliseconds, microseconds, or nanoseconds as needed. In practice, we can only go so far, but in principle there is no limit to how small a unit of time we can use.

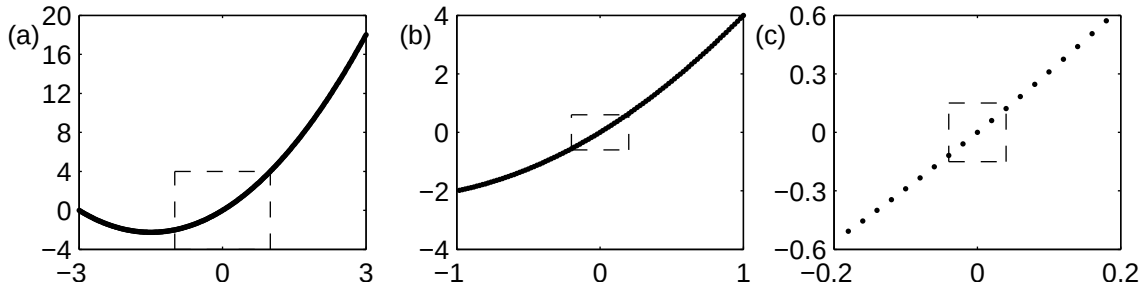

Figure 1.1.1: A data set consisting of points with x values separated by 0.02

a particular point, say the box with corners at $(-1, -4)$ and $(1, 4)$, which is centered around $(0, 0)$, which we assume is one of the points in the data set. Finally, we use the coordinates of the corners of the box as the dimensions of the viewing window for a new graph of the same physical size. We are "zooming in" on a portion of the graph. Not all the points in the original graph will show up in the smaller viewing window. The points will also be farther apart. In Figure 1.1.1b, there are 100 points and the plot still appears to be smooth. If we continue to zoom in on the same point, eventually the viewing window will be small enough that we see some space separating the discrete points. Figure 1.1.1c shows only 19 points, which are clearly distinct. Zooming in on the same point again and again, the graph will eventually be reduced to the single point in the center.

- When you repeatedly zoom in on discrete data, you eventually have a set of points that are far removed from each other, regardless of the actual distance between the points. Stated in a different way, the distance between discrete points becomes infinite if we report that distance in smaller and smaller units of measurement.

The plots in Figure 1.1.1 were made from a set of points. Suppose we use the formula and let the computer choose the points. When we zoom in on the point $(0, 0)$ without locking in the coordinates of the points being displayed, the plot rendered by the software will be just as smooth as the original plot. The software will choose points that are closer together if we make the viewing window smaller.

- When you repeatedly zoom in on continuous data, your new graph is as smooth as the old graph because more points are added in between the points of the previous graph. In principle, this process can be repeated infinitely many times without ever resolving the plot into isolated points.

The boundary between discrete and continuous can be blurry. Suppose we measure the masses of some samples of water. For the sake of the thought experiment, let's say that all of the water is made from hydrogen atoms of a single isotope and oxygen atoms of a single isotope.[2] No matter whether we measure the mass in milligrams, micrograms, or nanograms, we can always find a sample whose mass is between any two consecutive integer units. Mass appears to be continuous. However, our water is made of molecules, which are discrete. If we use "molecules of water" as our unit of mass, we can only measure integer quantities. So mass in this example is actually discrete. In practice, there is no harm in thinking of mass as a continuous quantity. The number of cells in a gram of heart tissue is far less than the number

[2]This gives all of the molecules exactly the same mass.

of molecules in a gram of water, but it still seems reasonable to think of the number of cells as continuous. Given that the mass of a mosquito is approximately 0.5 mg, there are about 2000 individuals in a gram of mosquitos. If we are doing an experiment with approximately a gram of mosquitos, it is not clear whether we could think of the number of mosquitos as a continuous quantity.

In actual fact, many quantities either depend on mass or energy, both of which are fundamentally discrete. Hence, we could choose to use discrete mathematics for almost everything. And yet the previous paragraph talks as though we should prefer to think of quantities as continuous rather than discrete. If continuous quantities are so much harder to understand, why should we prefer to use them? The answer is that models based on continuous mathematics are often simpler and easier to study than models based on discrete mathematics. This will only become clear in Chapter 4.

The limit

The essential tool of continuous mathematics is the *limit*.

Example 1.1.2
Consider the function defined by

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

The graph of this function, seen in Figure 1.1.2, gives the appearance of a continuous quantity. However, a quick calculation shows that the formula cannot assign a value to $f(1)$. The function is continuous except for this one value of x. But now suppose we zoom in on the point $(1, 2)$. No matter how many times we zoom in, we can never see the hole in the graph. Based on our examination of the graph, it appears that $(1, 2)$ is as much a part of the curve as the point $(0, 3)$. What do we make of the observations that the function is not continuous in terms of algebraic calculation, but its graph cannot be distinguished from that of a continuous function? The answer is the difference between algebra and calculus. In algebra, the calculation is what counts. In calculus, it is the graph that matters.

How do we indicate that the point $(1, 2)$ ought to be part of the graph even though it cannot be determined from the formula for the function? We say that the *limit* of $f(x)$ as x approaches 1 is 2, and we write

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

◇

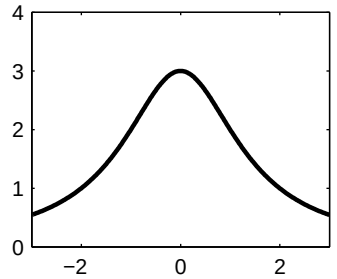

Figure 1.1.2: The function f of Example 1.1.2, a graph with an invisible hole

Example 1.1.2 suggests an intuitive statement about the meaning of *limit*.

- $\lim_{x \to a} f(x) = L$ means that no matter how many times we zoom in on the point (a, L), the graph always appears to include the point (a, L), regardless of whether $f(a) = L$ or not.

We obtained the conclusion $\lim_{x \to 1} f(x) = 2$ in Example 1.1.2 using graphical methods. We could have obtained the same conclusion using algebra.

Example 1.1.3
Observe that

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

except for $x = 1$. The functions

$$f(x) = \frac{6x - 6}{x^3 - x^2 + 2x - 2}, \qquad g(x) = \frac{6}{x^2+2}$$

are identical at all points except $x = 1$. Since a single point missing from a graph cannot be seen,

$$\lim_{x \to 1} f(x) = \lim_{x \to 1} g(x) = 2,$$

even though $f(1) \neq g(1)$. ◇

Mathematicians say that f and g of Example 1.1.3 are *almost equal*, which means that all their limit values are equal even though the function values are not always equal.[3] In continuous models, we are able to treat almost equal functions as exactly equal because we use limit values rather than algebraic values.

Summary

Thinking of functions in terms of their graphs, rather than the formulas used to calculate points on the graph, allows us to ignore holes in a graph caused by isolated input values for which the formula doesn't work. In these cases, we can define the (graphical) function value as a limit. These ideas are used in calculus to develop the central concepts of the derivative and the definite integral. In the next two sections, we will develop these concepts using the same overall plan:

1. State a general problem about functions that can be solved algebraically only for the simplest cases, and define the calculus concept to be the as-yet-unknown solution of that general problem.

2. Develop a scheme to obtain approximate solutions of the general problem. The scheme must include a refinement parameter that can be adjusted to make the approximation more and more accurate, up to an extreme parameter value for which the approximation formula does not work.

3. Think of the approximation result as a function of the refinement parameter. Determine a limit value for the approximation at the extreme value of the parameter by zooming in. This gives the desired solution on the principle that holes in a graph can be ignored in continuous models.

[3] Limit values are all equal for functions that differ only at one or more isolated points.

1.2 The Derivative

In Section 1.1, we developed some basic ideas and presented the procedure used to develop the key concepts of calculus. We now apply that procedure to the first of these concepts, the derivative.

The tangent slope problem

1. State a general problem about functions that can be solved algebraically only for the simplest cases, and define the calculus concept to be the as-yet-unknown solution of that general problem.

PROBLEM
Find the function $f'(x)$ whose value is the slope of the line tangent to $y = f(x)$ at any point where the tangent line is well defined.

DEFINITION
The **derivative** of the function $f(x)$ is the function $f'(x)$ whose value is the slope of the line tangent to $y = f(x)$.

Note that we are giving a semantic definition of the word "derivative" rather than a mathematical definition of the mathematical object called the derivative. We must discover the mathematical definition of the derivative by completing the three-step program for concept development in calculus. The problem in this case meets the requirements for step one. It is general, in that we can apply it to any function whose graph has well-defined tangents. It can be solved algebraically for the elementary case of a linear function. It cannot be solved algebraically for any curve other than the arc of a circle.

An approximation scheme

2. Develop a scheme to obtain approximate solutions of the general problem. The scheme must include a refinement parameter that can be adjusted to make the approximation more and more accurate, up to an extreme parameter value for which the approximation formula does not work.

Algebra provides a simple formula for determining the slope of a line, provided that two points on the line are known. What makes the tangent slope problem difficult is that we only know one point on the line. Our procedure does not require an algebraic solution–only an algebraic approximation. If we choose two points on the curve, we can connect them with a straight line and determine its slope. A line defined in this manner is a *secant* line rather than a tangent line. We can use the secant slope as an estimate of the tangent slope.

Example 1.2.1
Consider the function $f(x) = 3 - (x-2)^2$. Figure 1.2.1 shows this function, along with the tangent line at the point $x = 1$ and the secant line connecting that point with the point $x = 2$. The slope of the secant line is easy to compute:

$$m_{\text{sec}} = \frac{\Delta y}{\Delta x} = \frac{f(2) - f(1)}{2 - 1} = \frac{3 - 2}{1} = 1,$$

$$m_{\text{sec}} = \frac{\Delta y}{\Delta x} = \frac{f(2) - f(1.5)}{2 - 1.5} = \frac{2.75 - 2}{1} = 0.75.$$

Both secant slopes approximate the tangent slope. Neither approximation is very accurate, but we can make better approximations by moving the second point closer to the first point. ◇

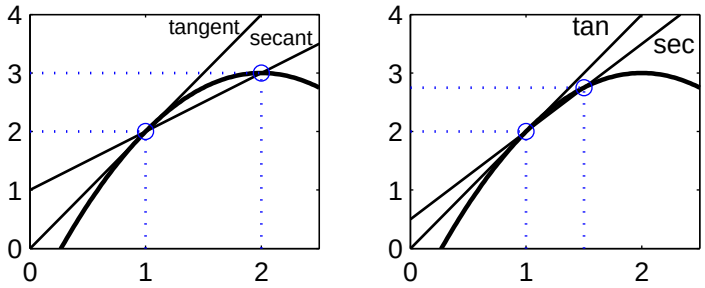

Figure 1.2.1: The function f of Example 1.2.1, along with a tangent and two secants

Solution of the problem

3. Think of the approximation result as a function of the refinement parameter. Determine a limit value for the approximation at the extreme value of the parameter by zooming in. This gives the desired solution, on the principle that holes in a graph can be ignored in continuous models.

Our approximation scheme needs to be very clear about which secant line is being used. To that end, observe that a secant line is determined by the choices of the two x coordinates of the points. The first point is the one where we have drawn the tangent line, so we use x to denote that point. For the second point, we could use the x coordinate as well; instead, it is more convenient to define h to be the horizontal distance Δx between the points. We can then prescribe a specific secant line by choosing values of x and h. Because the secant slope depends on these two quantities, we can write a general formula for the secant slope as a function of x and h.

APPROXIMATION
The slope of the tangent line to the graph $y = f(x)$ at x can be approximated by

$$m_{\text{sec}}(x, h) = \frac{\Delta y}{\Delta x} = \frac{f(x + h) - f(x)}{h}, \tag{1.2.1}$$

becoming as accurate as desired by making h small enough.

Example 1.2.2
Applying our new notation to the function $f(x) = 3 - (x-2)^2$ (Example 1.2.1), we have

$$m_{\text{sec}}(1,1) = 1, \qquad m_{\text{sec}}(1,0.5) = 0.75.$$

In general, we have a family of secant slopes that approximate $f'(1)$, our notation for the tangent slope at $x = 1$:

$$m_{\text{sec}}(1,h) = \frac{[3-(1+h-2)^2] - [3-(1-2)^2]}{h} = \frac{[3-(h-1)^2]-2}{h}.$$

◇

It is important to be very clear about the distinction between the quantities x and h. The former is the independent variable of the functions f and f' and represents the point where the tangent line is to be drawn. The latter is a parameter that measures the coarseness of the approximation. The smaller a value we choose for h, the more accurate an approximation to the tangent slope we get from the secant slope. The value $h = 0$ is the extreme value referred to in the procedure. We can't take $h = 0$, because then we would have only one point and no secant; however, we can get as close to this extreme value as we wish. The stage is set for us to use our zooming procedure to solve the tangent slope problem.

Example 1.2.3
Given the function $f(x) = 3-(x-2)^2$ and the point $x = 1$, we have already obtained the secant slope formula

$$m_{\text{sec}}(1,h) = \frac{[3-(h-1)^2]-2}{h}.$$

This formula does not assign a value to $h = 0$; however, if $h \neq 0$, we have

$$m_{\text{sec}}(1,h) = \frac{3-(h^2-2h+1)-2}{h} = \frac{2h-h^2}{h} = 2-h.$$

The graph of this function appears in Figure 1.2.2. ◇

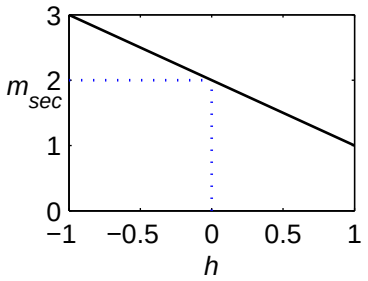

Figure 1.2.2: The function $m_{\text{sec}}(1,h)$ of Example 1.2.3

The plot in Figure 1.2.2 has a hole in it at the point $(0,2)$. But as we saw in Section 1.1, we cannot see a hole in a graph, no matter how many times we zoom in. If we allow the limit value to represent the secant slope function at $h = 0$, we have

$$f'(1) = \lim_{h\to 0} m_{\text{sec}}(1,h) = 2.$$

There was nothing special about the point $x = 1$ in Examples 1.2.1 through 1.2.3.

SOLUTION
The slope of the line tangent to $y = f(x)$ at any point x where there is a unique tangent is given by

$$f'(x) = \lim_{h \to 0} m_{\text{sec}}(x, h), \quad \text{where} \quad m_{\text{sec}}(x, h) = \frac{\Delta y}{\Delta x} = \frac{f(x+h) - f(x)}{h}. \tag{1.2.2}$$

Example 1.2.4
Let $f(x) = 3 - (x-2)^2 = 3 - (x^2 - 4x + 4) = -x^2 + 4x - 1$. For arbitrary x, and $h \neq 0$, we have

$$\begin{aligned} m_{\text{sec}}(x, h) &= \frac{[-(x+h)^2 + 4(x+h) - 1] - [-x^2 + 4x - 1]}{h} \\ &= \frac{[-(x^2 + 2xh + h^2) + 4x + 4h - 1] - [-x^2 + 4x - 1]}{h} \\ &= \frac{-2xh - h^2 + 4h}{h} = -2x - h + 4. \end{aligned}$$

Then

$$f'(x) = \lim_{h \to 0}(-2x - h + 4) = -2x + 4.$$

In particular, $f'(x) = 2$, as we had already determined. ◇

Summary

The crucial point of this section is the definition at the beginning:

- The **derivative** $f'(x)$ is the function that represents the slope of the tangent to the graph of f at x.

We will use this concept in applications in Section 1.7. In the remainder of these notes, much of our work with continuous models will be based on the concept of the derivative as stated here.

1.3 The Definite Integral

In Section 1.2, we developed the key concept of the derivative by using the three-step procedure outlined in Section 1.1. We now apply this procedure to the second key concept of calculus: the definite integral.

The area problem

1. State a general problem about functions that can be solved algebraically only for the simplest cases, and define the calculus concept to be the as-yet-unknown solution of that general problem.

PROBLEM
Find the area A bounded by the nonnegative function $f(x)$, the x axis, and the vertical lines $x = a$ and $x = b$, with $a < b$.

DEFINITION
The **definite integral** of the function $f(x)$ over the interval $a \le x \le b$ is the area described in the previous statement of the problem. This quantity is denoted $\int_a^b f(x)\,dx$, which is read "the integral of f from a to b." The function f is called the **integrand** and the numbers a and b are called the **limits of integration**.

As with the derivative, this is a semantic definition of the term "definite integral." We must discover the mathematical definition of the definite integral by completing the three-step concept development program. The area problem can be applied to any function whose graph is moderately smooth. It can be solved algebraically for the elementary case of a linear function.

An approximation scheme

2. Develop a scheme to obtain approximate solutions of the general problem. The scheme must include a refinement parameter that can be adjusted to make the approximation more and more accurate, up to an extreme parameter value for which the approximation formula does not work.

We can use the formula for area of a rectangle to approximate the area under a curve. While we won't necessarily have a good approximation, we can use more rectangles to make the approximation better.

Example 1.3.1
Consider the function $f(x) = 1 + x^2$ between $x = 0$ and $x = 1$. Figure 1.3.1 shows this function, along with approximations using two and four rectangles. The calculated area clearly depends on the number of rectangles used. If $A(n)$ is the area approximation obtained with n rectangles, then we have

$$A(2) = f(.5) \cdot 0.5 + f(1) \cdot .5 = 1.25 \cdot .5 + 2 \cdot .5 = 1.625,$$

$$A(5) = f(.2)\cdot.2+f(.4)\cdot.2+f(.6)\cdot.2+f(.8)\cdot.2+f(1)\cdot.2 = 1.04\cdot.2+1.16\cdot.2+1.36\cdot.2+1.64\cdot.2+2\cdot.2 = 1.44.$$

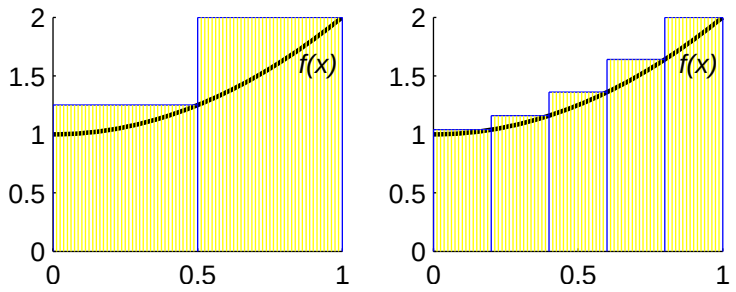

Figure 1.3.1: The function $1+x^2$, along with area approximations using two and five rectangles

◇

Notice that we have followed the practice in Example 1.3.1 of using the x values at the *right* of each interval to determine the height of the approximating rectangle. We could just as well have chosen the left boundary of each interval, or we could have obtained a better approximation using the midpoint of each interval. In the use of the area problem to define the definite integral, and of these is equally good. The right endpoint is more convenient because it makes for the simplest general approximation formula.

Solution of the problem

3. Think of the approximation result as a function of the refinement parameter. Determine a limit value for the approximation at the extreme value of the parameter by zooming in. This gives the desired solution, on the principle that holes in a graph can be ignored in continuous models.

The notation we need for the approximation scheme has already been introduced in Example 1.3.1. The refinement parameter is n, the number of rectangles, and the extreme value of the parameter is infinity.

Example 1.3.2
Consider the function $f(x) = 1 + x^2$ between $x = 0$ and $x = 1$. With n intervals, the width of each interval is $1/n$. The x values on the right side of each interval are $1/n, 2/n, \ldots, 1$. The rectangle heights are obtained by substituting these x values into the function f:

$$1 + \frac{1}{n^2}, \quad 1 + \frac{4}{n^2}, \quad 1 + \frac{9}{n^2}, \quad \ldots, \quad 1 + 1^2.$$

Thus,

$$A(n) = \left(1 + \frac{1}{n^2}\right)\frac{1}{n} + \left(1 + \frac{4}{n^2}\right)\frac{1}{n} + \left(1 + \frac{9}{n^2}\right)\frac{1}{n} + \cdots + (2)\frac{1}{n} = \sum_{k=1}^{n}\left(1 + \frac{k^2}{n^2}\right)\frac{1}{n}.$$

◇

In determining the general formulas for the approximation scheme, we need to keep in mind that the left and right boundaries of the area are not necessarily 0 and 1. In general, if the integration interval is $[a, b]$, then the width of each interval is $\Delta x = (b - a)/n$. The x value at the right of the first interval is $x_1 = a + \Delta x$, and each successive x value is another Δx to the right: $x_k = a + k\Delta x$. We have the following result:

APPROXIMATION
The area under $f(x)$ on $a \le x \le b$ can be approximated by

$$\Delta x = \frac{b-a}{n}, \qquad x_k = a + k\Delta x, \qquad A(n) = \Delta x \sum_{k=1}^{n} f(x_k), \tag{1.3.1}$$

becoming as accurate as desired by making n large enough. Note that the factor Δx can be removed from the summation because it is not a function of k.

Example 1.3.3
Suppose we want to determine the area underneath one positive portion of the sine function. The equation $\sin x = 0$ has solutions $x = 0, \pm\pi, \pm 2\pi, \ldots$, and the function is positive on the interval $[0, \pi]$. With n subdivisions, we have

$$\Delta x = \frac{\pi}{n}, \qquad x_k = \frac{k\pi}{n}, \qquad A(n) = \frac{\pi}{n} \sum_{k=1}^{n} \sin\left(\frac{k\pi}{n}\right).$$

The approximations with $n = 2$ and $n = 6$ are illustrated in Figure 1.3.2. Specifically,

$$A(2) = \frac{\pi}{2}\left[\sin\left(\frac{\pi}{2}\right) + \sin(\pi)\right] = \frac{\pi}{2} \approx 1.571,$$

$$A(6) = \frac{\pi}{6}\left[\sin\left(\frac{\pi}{6}\right) + \sin\left(\frac{2\pi}{6}\right) + \sin\left(\frac{3\pi}{6}\right) + \sin\left(\frac{4\pi}{6}\right) + \sin\left(\frac{5\pi}{6}\right) + \sin(\pi)\right] = \frac{\pi}{6}(2+\sqrt{3}) \approx 1.954.$$

Note that for any n, the last rectangle has height 0 because $\sin \pi = 0$. ◇

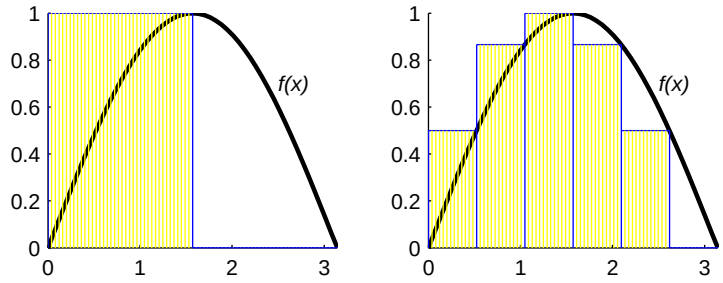

Figure 1.3.2: The function $\sin x$, along with area approximations using two and six rectangles

There are some technical difficulties involved in the definite integral that we did not encounter in the derivative. For the moment, we can simply write down the solution of the area problem as a limit.

SOLUTION
The area under the graph of $y = f(x) \ge 0$ on the interval $a \le x \le b$ is

$$\int_a^b f(x)\,dx = \lim_{n\to\infty}\left[\Delta x \sum_{k=1}^{n} f(x_k)\right], \quad \text{where} \quad \Delta x = \frac{b-a}{n}, \quad x_k = a + k\Delta x. \tag{1.3.2}$$

We have restricted consideration to nonnegative integrands. This has been done to make the area-integral connection easier to visualize. However, the restriction $f \ge 0$ is not necessary

for the definition and calculation of definite integrals. If the function is negative, then the summation will be negative, and we will have a negative result for the definite integral. The integral still represents the area between the curve and the x axis, provided that we associate negative areas with regions below the x axis.

The limit as $n \to \infty$

In our previous discussion of limits, we have worked with functions that were continuous except at the point of interest and we associated the limit value with the y value of the hole in the graph. Obviously this will not work if we are looking at the limit as an integer n increases to infinity. The area approximations for $\int_0^1 (1+x^2)\,dx$ and $\int_0^\pi \sin x\,dx$ are plotted in Figure 1.3.2. Note that both functions have an asymptote,[1] and this is the limit value that we seek. As a practical matter, the details of computing the limits that define definite integrals need not concern us. We will develop the best methods for integral computation in Section 1.5. In that section, we will see that the correct values are

$$\int_0^1 (1+x^2)\,dx = \frac{4}{3}, \qquad \int_0^\pi \sin x\,dx = 2.$$

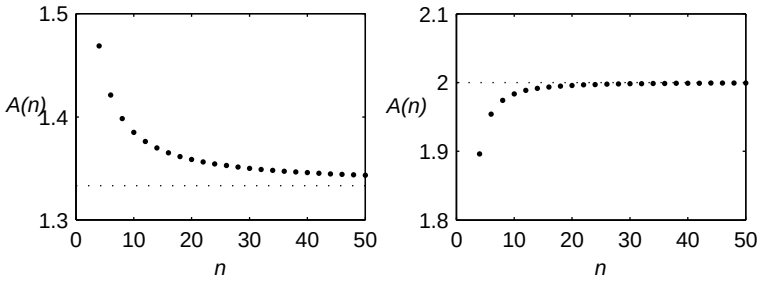

Figure 1.3.3: The approximations $A(n)$ for Examples 1.3.2 and 1.3.3

Summary

The crucial point of this section is the connection between the definite integral and area:

- The **definite integral** $\int_a^b f(x)\,dx$ is the area between the graph of $y = f(x)$ and the x axis over the interval $a \le x \le b$, with positive f corresponding to areas above the x axis and negative f corresponding to areas below the x axis.

We will use this concept in applications in Section 1.8. Continuous models in probability, which we will see in Chapter 3, also utilize the definite integral.

[1] We are using the term here because its meaning is clear from the context. Technically, the term should be restricted to continuous functions.

1.4 Computing Derivatives

In solving the tangent slope problem of Section 1.2, we obtained a formula, Equation 1.2.2, that provides a mathematical definition of the derivative. In principle, this formula could be used to compute derivatives; in practice, there is a much better way to do this. As a starting point, we need

- derivative formulas for a small number of basic functions, and
- general rules for reducing differentiation problems to the basic formulas.

Elementary derivative formulas

Table 1.4.1 summarizes the elementary derivative formulas that we need for our differentiation scheme. These formulas can be obtained with varying amounts of difficulty. The first five

$f(x)$	$x^p \quad [p \neq 0]$	e^{ax}	$\ln x$	$\sin ax$	$\cos ax$	$\arctan \dfrac{x}{a}$	$\arcsin \dfrac{x}{a}$
$f'(x)$	px^{p-1}	ae^{ax}	$\dfrac{1}{x}$	$a\cos ax$	$-a\sin ax$	$\dfrac{a}{a^2+x^2}$	$\dfrac{1}{\sqrt{a^2-x^2}}$

Table 1.4.1: Elementary derivative formulas

come from a combination of Equation 1.2.2 with selected limit values that are not difficult to determine from a graph, while the latter two come from application of the basic rules given below.

Example 1.4.1
Let $f(x) = e^{ax}$. Then the secant slope is

$$m_{\text{sec}}(x,h) = \frac{e^{ax+ah} - e^{ax}}{h} = \frac{e^{ax}e^{ah} - e^{ax}}{h} = e^{ax}\frac{e^{ah}-1}{h} = ae^{ax}\frac{e^{ah}-1}{ah}.$$

Thus,

$$f'(x) = \left[\lim_{h\to 0}\frac{e^{ah}-1}{ah}\right] ae^{ax} = \left[\lim_{y\to 0}\frac{e^{y}-1}{y}\right] ae^{ax}.$$

Using our graphical zooming technique, we can assert

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

hence,

$$f'(x) = ae^{ax}.$$

$\diamond$

General derivative rules

The task of symbolic differentiation is greatly facilitated by the existence of general rules that allow us to reduce differentiation of various algebraic structures to differentiation of elementary functions.

Let f and g be differentiable functions and let a and b be real constants.

Linearity rules

$$[af(x) + bg(x)]' = af'(x) + bg'(x). \tag{1.4.1}$$

Product rule

$$[f(x)g(x)]' = f'(x)g(x) + f(x)g'(x). \tag{1.4.2}$$

Quotient rule

$$\left[\frac{f(x)}{g(x)}\right]' = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}, \qquad g \neq 0. \tag{1.4.3}$$

Chain rule

$$[f(g(x))]' = f'(g(x))g'(x). \tag{1.4.4}$$

Example 1.4.2

$$\left[3x^2 + \sqrt{x} + 2^x\right]' = 3(x^2)' + (x^{1/2})' + (e^{x\ln 2})' = 3 \cdot 2x + \frac{1}{2}x^{-1/2} + (\ln 2)e^{x\ln 2} = 6x + \frac{1}{2\sqrt{x}} + (\ln 2)2^x.$$

$$\left[x^2 \cos 3x\right]' = (x^2)' \cos 3x + x^2(\cos 3x)' = 2x\cos 3x - 3x^2 \sin 3x.$$

$$\left[\frac{x}{a+x}\right]' = \frac{(x)'(a+x) - x(a+x)'}{(a+x)^2} = \frac{(a+x) - x \cdot 1}{(a+x)^2} = \frac{a}{(a+x)^2}.$$

$$[\sin(\ln x)]' = \cos(\ln x)(\ln x)' = \frac{\cos(\ln x)}{x}.$$

◇

Our small list of elementary formulas and general rules allow us to compute derivatives for most any function we will encounter in biology. Computer algebra systems, such as Maple and Mathematica, can also compute derivatives. For this reason, it is not crucial for life science students to master the mechanics of differentiation. It is useful to be able to compute derivatives for relatively simple cases, such as those in Example 1.4.2.

The chain rule in terms of variables

The chain rule is of vital importance in many applications of mathematics, even given a choice to use computer algebra systems to compute derivatives. The reason is that the chain rule is often used to change variables in a differential equation model. We will see this use of the chain rule in that context, but for now we focus on the understanding of the chain rule that will be needed at that time. We begin with a physical example that illustrates how the chain rule is much more than a mere computational tool.

Example 1.4.3
A metal rod is placed so that one end is in a campfire and the other end is resting on the cold ground outside of the fire. An insect is walking along the metal rod away from the fire at a constant speed. Suppose the temperature along the metal rod is given by $T(x)$ and the position of the insect is given by $x = s(t)$. The variable x serves as the independent variable for the temperature profile and the dependent variable for the motion of the insect. This notation is not completely consistent, because in $T(x)$ we are using x to simultaneously represent all locations on the rod, while in $x = s(t)$ we are using x to represent the unique location of the insect at any particular time. Adding to the confusion, we might be interested in the temperature experienced by the insect. The insect's temperature is changing in time because of the insect's motion; hence, if we use u to denote the insect's temperature, then we should have u be determined by a function of t. This temperature u is determined by evaluating T at the point x that gives the insect's location; thus,

$$u = T(s(t)).$$

We can use this equation to write a relationship between the derivatives of u and T. Applying Equation 1.4.4, we have

$$u'(t) = T'(s(t))s'(t).$$

In this equation, $u'(t)$ is the rate of change of the insect's temperature with respect to time, $T'(x)$ is the rate of change of the temperature with respect to the distance at any point on the metal rod, $T'(s(t))$ is the rate of change of the temperature with respect to the distance at the insect's location, and $u'(t)$ is the speed of the insect's motion. ◇

This is a good point at which to introduce an alternative notation for the derivative. Our standard notation, in which f' is the derivative of f, is fine as long as there is no difficulty in knowing what variable is the independent variable. The equation $u'(t) = T'(s(t))s'(t)$ can be confusing because the prime symbol refers to the time derivatives of u and s and the spatial derivative of T. An alternative is to include the independent variable in the notation by using $\frac{dT}{dx}$ instead of T' and $\frac{du}{dt}$ instead of u'. We then have

$$\frac{du}{dt}(t) = \frac{dT}{dx}(s(t))\frac{ds}{dt}(t).$$

This notation can also be confusing. It is common in mathematical modeling to address this confusion by abandoning the standard function notation of mathematics. Carefully reread the text of Example 1.4.3, and you will see that the five quantities, T, x, s, t, and u are really of two types: t, x, and u are variables, while T and s are functions that are used to indicate the relationships among the variables. In terms of variables, we have three simultaneously changing quantities, with time t as the independent variable and the insect's location x and temperature u as the dependent variables. The dependent variables have rates of change denoted du/dt and dx/dt, and the derivative du/dx denotes the rate of change of the insect's temperature with respect to the insect's position. These rates of change are related by the chain rule, which then appears as

$$\frac{du}{dt} = \frac{du}{dx}\,\frac{dx}{dt}.$$

While less mathematically precise, this last form is much more elegant. It looks like an equation about multiplication of fractions, which is particularly memorable even if not precisely correct.

Chain rule

If u depends on x and x depends on t, then

$$\frac{du}{dt} = \frac{du}{dx}\,\frac{dx}{dt}. \tag{1.4.5}$$

Change of variables in derivatives

A given quantity can be measured in a variety of units. In a laboratory setting, one chooses units that are convenient for measurement, given the experiment at hand. Time might be measured in years for an experiment on progression of AIDS, months for certain cancer survival experiments, weeks for human embryo development, days for antibiotic therapy experiments, hours for bacteria growth in a culture, or minutes for blood clotting experiments, to name a few examples. In mathematical models, time is measured in units that are tied to the rates of crucial processes rather than familiar measurement units.

Example 1.4.4
Suppose we are studying the natural growth of an invasive species in a lake. For the sake of a thought experiment, suppose that the population is known to double in approximately 3 years. Let P be the population and t the time measured in years. A model of the population growth would assume that the growth is deterministic rather than random, and with a doubling time of exactly 3 years. A discrete model is given by the equation

$$P(t+3) = 2P(t).$$

A continuous model is given by the equation

$$\frac{dP}{dt} = kP,$$

where k is an as-yet-unknown constant and we have assumed that the rate of population growth is proportional to the current population. This model will be discussed in detail in Chapter 4. For now, let us merely observe that the formula

$$P = Ae^{kt}$$

has the desired property for any constant value of A, since $(Ae^{kt})' = A \cdot ke^{kt} = kAe^{kt} = kP$, where the prime symbol refers to derivatives with respect to time. Now we can use the problem details to determine k. We still must have $P(t+3) = 2P(t)$ for our continuous model; hence,

$$Ae^{k(t+3)} = 2Ae^{kt}.$$

Dividing both sides of the equation by Ae^{kt} yields

$$e^{3k} = 2,$$

from which we obtain $3k = \ln 2$ or

$$k = \frac{\ln 2}{3} \approx 0.231.$$

Returning to our model and solution, we can at this point ask ourselves what the model and solution would be if we measured time in some unit other than years. One possibility is "doubling times," a unit based on the population growth rather than convenience of measurement. Suppose s is the time measured in terms of doubling times. Then $s = t/3$ because every three years corresponds to one doubling time. In terms of s, the population is

$$P = Ae^{k(3s)} = A\left(e^{3k}\right)^s = A\,2^s,$$

where we have used the equation $e^{3k} = 2$ that we determined earlier. The use of s removes the quantity k from the solution, which makes for a simpler result.

Another possibility, less obvious but ultimately more convenient, is to measure time in multiples of $(1/k) \approx 4.328$ years. In other words, we define a new time variable by $z = kt$, so that $t = 1/k$ corresponds to $z = 1$. In terms of z, the population is

$$P = Ae^{k(z/k)} = Ae^z.$$

As with the replacement of t by s, replacing t by z removes the constant k. This change has the additional advantage of simplifying the differential equation. By differentiating $P = Ae^z$, we can see that the original model changes from

$$\frac{dP}{dt} = kP$$

to

$$\frac{dP}{dz} = P.$$

Using z rather than t eliminates k from both the original model and the solution of the model. There are important mathematical advantages gained from this change. $\diamond$

Example 1.4.4 required a long-winded discussion to arrive at a general statement about mathematical models:

- Choosing units based on quantities that arise naturally in a model rather than units based on convenience of measurement makes models simpler.

The connection with the computation of derivatives is that most mathematical models do not have simple solution formulas. The change of variables usually must be made without knowledge of the solution. Returning to Example 1.4.4, how could we have changed the model from the equation $dP/dt = kP$ to $dP/dz = P$ without already knowing the formula $P = Ae^z$? The answer is supplied by the chain rule. We have an equation that contains dP/dt and we want an equation that contains dP/dz instead. These quantities are related by the chain rule and the equation $z = kt$ that defines z. We have

$$\frac{dP}{dt} = \frac{dP}{dz}\frac{dz}{dt} = k\frac{dP}{dz}.$$

Substituting this relation into $dP/dt = kP$ yields

$$k\frac{dP}{dz} = kP,$$

which reduces to

$$\frac{dP}{dz} = P.$$

The process of changing an equation so that quantities with units convenient for measurement are replaced by quantities with units that arise naturally in the model is called **nondimensionalization**.

1.5 Computing Definite Integrals

Equation 1.3.2 is an excellent mathematical definition of the definite integral, but it is almost useless as a means of computing definite integrals. As a general rule, it is harder to compute definite integrals than derivatives. We developed a complete set of rules in Section 1.4 for obtaining derivative formulas for any function made from elementary components. No such scheme is possible for definite integrals. Instead, we will develop (1) an exact method that only works some of the time and (2) improved formulas for integral approximation.

The Fundamental Theorem of Calculus

So far, we have treated the derivative and the definite integral as two important calculus concepts whose only relationship is that they are defined to give the solution of a geometry problem. In fact, the connection between derivative and definite integral is so intimate that the formal statement of this connection is called the Fundamental Theorem of Calculus.

The Fundamental Theorem is not difficult to derive; calculus textbooks generally include a derivation. In keeping with our emphasis on understanding and use of ideas, we are instead going to develop the Fundamental Theorem by solving the same problem twice from two different points of view. The result will be a statement of the Fundamental Theorem that is intuitively clear, indicates how the theorem is used to compute definite integrals, and explains why most definite integrals cannot be computed in this way.

Example 1.5.1
Suppose you begin a drive on an Interstate highway. You travel in the direction of increasing mileposts, and at some point you stop. How much change has there been in your position?

Let t be time and x be distance as measured by mileposts. Your location is a function of time. If you start at time a and finish at time b, then you have traveled from milepost $x(a)$ to milepost $x(b)$. The total change in your position is the difference between these mileposts:

$$\Delta x = x(b) - x(a).$$

$\diamond$

Example 1.5.2
Suppose you cannot see the mileposts on the road. Instead, your car has a device that records your speed v as a function of time t. How much change has there been in your position?

Suppose we partition the time interval $a \le t \le b$ into n equal subintervals. Each subinterval is of length $\Delta t = (b-a)/n$. The first interval is from time a to time $a + \Delta t$, the second is from time $a + \Delta t$ to time $a + 2\Delta t$, and so on. Now for each of these subintervals, we can approximate the distance traveled. Assuming that your speed at time $a + k\Delta t$ is approximately correct for the entire k^{th} interval that ends at that time, your distance during that interval is given by

$$(\Delta x)_k \approx v(a + k\Delta t)\,\Delta t.$$

Summing over all n subintervals, we have

$$\Delta x \approx \sum_{k=1}^{n} v(a + k\Delta t)\,\Delta t = \Delta t \sum_{k=1}^{n} v(a + k\Delta t).$$

This approximation becomes exactly correct in the limit as the number of subintervals goes to infinity. By Equation 1.3.2, we have

$$\Delta x = \lim_{n\to\infty}\left[\Delta t \sum_{k=1}^{n} v(a+k\Delta t)\right] = \int_a^b v(t)\,dt.$$

◇

Examples 1.5.1 and 1.5.2 solved the same problem using different pieces of information. The solutions have to be the same, and so we can combine them together to obtain the formula

$$\int_a^b v(t)\,dt = x(b) - x(a). \tag{1.5.1}$$

We derived Equation 1.5.1 from a specific scenario involving a car driving on a highway with variables indicating the mileposts and the velocity. But what details were really necessary for the mathematical result? The answer is that the narrative in which the mathematics is couched is not important. All that matters are the relationships among the quantities. The primary requirement is that the quantities x and v have to have the right mathematical relationship; the quantity represented by v must be the rate of change of the quantity represented by x. Rewriting Equation 1.5.1, with a little extra attention to mathematical issues, brings us to the Fundamental Theorem:

Theorem 1.5.1 (Fundamental Theorem of Calculus) *Suppose f' is piecewise continuous on the interval $[a, b]$. Then*

$$\int_a^b f'(t)\,dt = f(b) - f(a). \tag{1.5.2}$$

The first statement is a technical detail of mathematical importance. Piecewise continuity is the minimal degree of smoothness required of the derivative for the integral to make sense. A function is **piecewise continuous** on an interval $a \le t \le b$ if (1) it is continuous except for at most finitely many points in the interval and (2) is bounded as t approaches any of these points of discontinuity. Generally, any function that we are likely to encounter will meet the first of these requirements, but we must always be careful to check the second requirement. Several familiar functions, such as $1/x$, have isolated points of discontinuity that prevent us from being able to apply the Fundamental Theorem.

Computing definite integrals with the Fundamental Theorem

The Fundamental Theorem is useful for computing definite integrals, provided that the integrand of the definite integral can be written as the derivative of some known function. As examples, consider the integrals that we examined in Section 1.3.

Example 1.5.3
Note the derivative formulas

$$\left[x + \frac{1}{3}x^3\right]' = 1 + x^2, \qquad (-\cos x)' = \sin x.$$

By the Fundamental Theorem,

$$\int_0^1 (1+x^2)\,dx = \int_0^1 \left[x + \frac{1}{3}x^3\right]' dx = \left[1 + \frac{1}{3}1^3\right] - \left[0 + \frac{1}{3}0^3\right] = \frac{4}{3},$$

$$\int_0^\pi \sin x\,dx = \int_0^\pi (-\cos x)'\,dx = (-\cos\pi) - (-\cos 0) = (1) - (-1) = 2.$$

◇

This sounds easy, but there is a catch. We can only compute $\int_a^b F(x)\,dx$ by the Fundamental Theorem if we can find a function $f(x)$ for which $f' = F$. Such a function f is called an **antiderivative** of F. Each of the elementary derivative rules in Table 1.4.1 can easily be converted to a suitable antiderivative rule; these rules are summarized in Table 1.5.1.

$f'(x)$	$x^r \quad [r \neq -1]$	e^{ax}	$\frac{1}{x}$	$\cos ax$	$\sin ax$	$\frac{1}{a^2+x^2}$	$\frac{1}{\sqrt{a^2-x^2}}$
$f(x)$	$\frac{x^{r+1}}{r+1}$	$\frac{1}{a}\,e^{ax}$	$\ln\lvert x\rvert$	$\frac{1}{a}\,\sin ax$	$-\frac{1}{a}\,\cos ax$	$\frac{1}{a}\arctan\frac{x}{a}$	$\arcsin\frac{x}{a}$

Table 1.5.1: Elementary antiderivative formulas

The antiderivative formulas can sometimes be used even if the requirements of Theorem 1.5.1 are not quite met.

Example 1.5.4

$$\int_0^\infty e^{-ax}\,dx = \int_0^\infty \left[-\frac{1}{a}\,e^{-ax}\right]'\,dx = \lim_{b\to\infty}\left[-\frac{1}{a}\,e^{-ab}\right] - \left[-\frac{1}{a}\,e^0\right] = \frac{1}{a}, \quad a > 0.$$

◇

Example 1.5.4 shows that the definite integral can sometimes exist even if the area it represents is unbounded in the x direction. It is also possible for the definite integral to exist when the area it represents is bounded in the y direction.

Example 1.5.5

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx = \int_0^1 (2\sqrt{x})'\,dx = 2\sqrt{1} - 2\sqrt{0} = 2.$$

◇

Substitution

The antidifferentiation technique called *substitution* corresponds to the chain rule of differentiation. If we integrate Equation 1.4.4 and apply the Fundamental Theorem, we obtain the equation

$$\int_a^b f'(g(x))g'(x)\,dx = \int_a^b [f(g(x))]'\,dx = f(g(b)) - f(g(a)).$$

Now let w be the independent variable of the function f. Thus, w is the variable that is associated with the function $g(x)$. The last equation then becomes

$$\int_a^b f'(w)\frac{dw}{dx}\,dx = f(g(b)) - f(g(a)) = \int_{g(a)}^{g(b)} f'(w)\,dw = \int_{w(a)}^{w(b)} f'(w)\,dw,$$

where the second equality follows from the Fundamental Theorem and the last from the association of w with g. Finally, we define $F(w) = f'(w)$ and obtain the substitution rule:

Theorem 1.5.2 (Substitution rule) *If $w'(x)$ is continuous on $[a, b]$ and F is piecewise continuous on an interval containing $w(a)$ and $w(b)$, then*

$$\int_a^b F(w)\frac{dw}{dx}\,dx = \int_{w(a)}^{w(b)} F(w)\,dw.$$

Theorem 1.5.2 prescribes a method for changing variables in a definite integral; thus, it is analogous to the method for changing variables in derivatives that we developed in Section 1.4.

Example 1.5.6
To compute $\int_0^2 xe^{x^2}\,dx$, we let $w = x^2$. Then we have

$$\int_0^2 xe^{x^2}\,dx = \int_0^2 \frac{e^{x^2}}{2}\cdot 2x\,dx = \int_0^4 \frac{e^w}{2}\,dw = \frac{1}{2}\int_0^4 (e^w)'\,dw = \frac{e^4-1}{2}.$$

Note that the integration limits changed in the second equality from $0 \le x \le 2$ to $0 \le w \le 4$, which follows from the equation $w = x^2$. Also, the function $F(w)$ (in this case $e^w/2$) is whatever remains of the integrand after $dw/dx = 2x$ is attached to dx. ◇

Numerical approximation of integrals

Equation 1.3.2 was only one of several possible approximation formulas for the definite integral. We obtained that approximation by using the right endpoint of each subinterval to determine a value of the integrand, which we then used to indicate the height of an approximation rectangle. We could just as easily have chosen the left endpoint or the midpoint. We therefore have three possible approximations using rectangles. We can also average the left and right approximations, yielding an improved approximation called the trapezoidal rule. Here is a summary of the geometric approximations for the definite integral:

APPROXIMATION
Let $\Delta x = \dfrac{b-a}{n}$ and $x_k = a + k\Delta x$. Then we have

$$\int_a^b f(x)\,dx \approx \text{RIGHT}(n) = \Delta x \sum_{k=1}^{n} f(x_k),$$

$$\int_a^b f(x)\,dx \approx \text{LEFT}(n) = \Delta x \sum_{k=1}^{n} f(x_{k-1}) = \Delta x \sum_{k=0}^{n-1} f(x_k),$$

$$\int_a^b f(x)\,dx \approx \text{MID}(n) = \Delta x \sum_{k=1}^{n} f(x_{k-1/2}),$$

and

$$\int_a^b f(x)\,dx \approx \text{TRAP}(n) = \frac{\text{LEFT}(n) + \text{RIGHT}(n)}{2} = \frac{\Delta x}{2}\left[f(a) + 2\sum_{k=1}^{n-1} f(x_k) + f(b)\right].$$

Example 1.5.7

In Example 1.3.1, we approximated $\int_0^1 (1+x^2)\,dx$ using 5 rectangles and the right endpoints. Our full set of approximations with $n = 5$ are

$$\text{RIGHT}(5) = .2[f(.2) + f(.4) + f(.6) + f(.8) + f(1)] = .2(1.04 + 1.16 + 1.36 + 1.64 + 2) = 1.44,$$

$$\text{LEFT}(5) = .2[f(0) + f(.2) + f(.4) + f(.6) + f(.8)] = .2(1 + 1.04 + 1.16 + 1.36 + 1.64) = 1.24,$$

$$\text{MID}(5) = .2[f(.1) + f(.3) + f(.5) + f(.7) + f(.9)] = .2(1.01 + 1.09 + 1.25 + 1.49 + 1.81) = 1.33,$$

$$\text{TRAP}(5) = \frac{\text{LEFT}(n) + \text{RIGHT}(n)}{2} = 1.34.$$

The MID and TRAP approximations for $n = 2$ are illustrated in Figure 1.5.1. Notice that we can illustrate the average of LEFT and RIGHT on any interval by using a trapezoid in which the top line connects the two points on the graph corresponding to the interval endpoints. This is the reason why the average of LEFT and RIGHT is called the trapezoidal rule. ◇

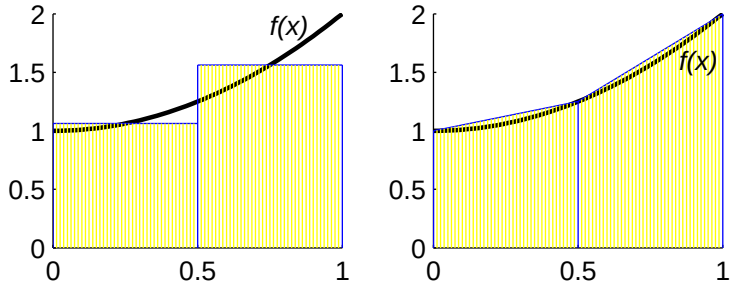

Figure 1.5.1: The midpoint and trapezoidal approximations of $\int_0^1 (1+x^2)\,dx$ with $n = 2$

Careful examination of Figure 1.5.1 shows that the trapezoidal approximation for this integral is too large and the midpoint approximation is too small. We saw in Example 1.5.3 that the correct result is 4/3, which is between the two approximations. We can achieve the correct answer if we use a weighted average of (2*MID+TRAP)/3. This is more than just a coincidence. With the help of Taylor series, an advanced differentiation topic, it can be shown that this 2 to 1 weighted average always gives the correct answer when the integrand is a quadratic function. It can be further shown that the 2 to 1 weighted average has several highly desirable properties that make it the ideal formula for numerical approximation of definite integrals.

DEFINITION

Let $\Delta x = \dfrac{b-a}{n}$ and $x_k = a + k\Delta x$. Then the **Simpson's rule** approximation with n subdivisions is

$$\int_a^b f(x)\,dx \approx \Delta x \left(\frac{2}{3}\sum_{k=1}^{n} f(x_{k-1/2}) + \frac{1}{3}\sum_{k=1}^{n-1} f(x_k) + \frac{1}{6}[f(a) + f(b)] \right). \tag{1.5.3}$$

Simpson's rule is the most commonly used technique for numerical approximation of definite integrals. Most calculators and computer programs use an adaptive[1] implementation of Simpson's rule.

[1] An *adaptive* routine is one in which additional formulas are used to estimate the amount of error in each subdivision; parts of the domain where the error is large are further subdivided until the error is below some desired bound.

Summary

The most powerful tool for computing definite integrals is the Fundamental Theorem of Calculus, which reduces the definite integral problem to a problem of finding antiderivatives. While we can compute *derivatives* for almost any function we do not have anything like a complete set of rules for finding *antiderivatives* for a broad class of functions. One problem is that determining the appropriate rule to use for finding an antiderivative is often difficult. A further problem is that not all functions even have an elementary antiderivative.

Example 1.5.8
We cannot compute

$$\int_0^1 \sin x^2\,dx.$$

The problem is *not* that we aren't clever enough to find a function f for which $f'(x) = \sin x^2$, but rather that *there is no function with the desired derivative.*

One might expect that

$$\int_0^1 \sin\sqrt{x}\,dx$$

would be even worse. However, with $w = \sqrt{x}$, and noting that $0 \le x \le 1$ is then equivalent to $0 \le w \le 1$, we have

$$\int_0^1 \sin\sqrt{x}\,dx = \int_0^1 2\sqrt{x}(\sin\sqrt{x})\frac{dx}{2\sqrt{x}} = 2\int_0^1 w\sin w\,dw.$$

From here, we observe that

$$(\sin w - w\cos w)' = \cos w - (\cos w - w\sin w) = w\sin w,$$

so

$$\int_0^1 \sin\sqrt{x}\,dx = 2\int_0^1 w\sin w\,dw = 2(\sin 1 - \cos 1) - 2(\sin 0 - 0) = 2(\sin 1 - \cos 1).$$

◇

To compute definite integrals without an elementary antiderivative formula, we have to settle for numerical approximation. Simpson's rule (Equation 1.5.3) provides a convenient and reasonably accurate means of numerical approximation.

1.6 Applications of the Derivative and Definite Integral

The derivative, f', is the rate of change of f, as represented by the slopes of the lines tangent to the curve $y = f(x)$. this geometric relationship is what makes the derivative useful. Any statement about tangents of a graph can be recast as statements about derivatives, which can then be used to calculate solutions to problems.

Local Behavior

You may have noticed that the process of zooming in on a point of a graph leads to the graph becoming closer to a straight line. this straight line is tangent to the curve. We can use the derivative to find an equation for this tangent line. In general, slope of a line is determined by $m = \Delta y/\Delta x$, where Δx and Δy are the differences between the coordinates of any two points on the line. In the case of a tangent line, we have only one point, but we can find the slope from the derivative instead. Let (x_0, y_0) be the point on the curve where the tangent is located, and let (x, y) be some other point on the tangent line. Using the two formulas for slope, we have

$$\frac{y - y_0}{x - x_0} = m_{\tan} = f'(x_0).$$

Solving for y yields a formula for the equation of the tangent line.

Theorem 1.6.1 *If f' exists at x_0, then the equation for the line tangent to the curve $y = f(x)$ at the point $x = x_0$ is*

$$y = f(x_0) + f'(x_0)(x - x_0). \tag{1.6.1}$$

Example 1.6.1
To find the equation of the line tangent to

$$y = f(x) = \frac{x}{1 + x}$$

at the point $x = 1$, we note that

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

Thus, $f(1) = 1/2$ and $f'(1) = 1/4$. The tangent line has the equation

$$y = \frac{1}{2} + \frac{1}{4}(x - 1).$$

$\diamond$

The tangent line is useful primarily because it serves as an approximation to the function when x is very close to x_0. For this reason, it is best to write the equation for the tangent line as in Equation 1.6.1.

Tangent line approximations are most useful in the context of a more complicated problem. This will be important in our study of differential equations in chapters 4 and 7.

Local Extrema

Suppose we are interested in locating peaks and valleys on graphs. Technically, peaks and valleys indicate *extrema*:

DEFINITION
A point x_0 is a **local maximum** of a function f if $f(x) \leq f(x_0)$ for all x in some small interval $(x_0 - \delta, x_0 + \delta)$. Similarly, x_0 is a **local minimum** of a function f if $f(x) \geq f(x_0)$ for all x in some small interval $(x_0 - \delta, x_0 + \delta)$. Local maxima and local minima are collectively known as **local extrema**.

In mathematical terms, it is often easier to prove a negative statement than it is to prove a positive statement.

Theorem 1.6.2 *If $f'(x_0) \neq 0$, then x_0 is not a local extrema.*

The proof of Theorem 1.6.2 follows immediately from Theorem 1.6.1. The tangent line to a function, which is given by Theorem 1.6.1, indicates the local behavior of the function. If the tangent slope is non-zero, then every interval around x_0 includes some points with both larger and smaller function values than $f(x_0)$. This rules out the possibility of x_0 being either a maximum or a minimum. As a consequence of Theorem 1.6.2, we can see that local extrema can occur only at points where $f' = 0$ or where there is no tangent line. A point that could be a local extremum by Theorem 1.6.2 is called a **critical point**. In the typical case of a function that has a tangent line at every point, extrema can only occur when $f' = 0$. There are a variety of formal tests for determining whether a given critical point is a local maximum, a local minimum, or neither. The reader should consult a full calculus text for details.

Example 1.6.2
Consider the function $f(x) = x^4 - 8x^3 + 18x^2 - 16x$, for which $f'(x) = 4x^3 - 24x^2 + 36x - 16 = 4(x^3 - 6x^2 + 9x - 4)$. Critical points are the solutions of

$$x^3 - 6x^2 + 9x - 4 = 0.$$

Normally we must resort to numerical approximation to solve a cubic equation, but observe that $x = 1$ is a solution of this equation. This means that $x - 1$ is a factor. If we assume

$$x^3 - 6x^2 + 9x - 4 = (x - 1)(x^2 + ax + 4),$$

where the terms x^2 and 4 in the second factor are necessary to obtain x^3 and -4 in the product, then

$$x^3 - 6x^2 + 9x - 4 = x^3 + (a - 1)x^2 + (4 - a)x - 4,$$

from which we obtain $a = -5$. Thus, the equation for the critical points becomes

$$0 = x^3 - 6x^2 + 9x - 4 = (x - 1)(x^2 - 5x + 4) = (x - 1)[(x - 4)(x - 1)] = (x - 1)^2(x - 4).$$

Thus, the critical points are $x = 1$ and $x = 4$. Moreover, we have

$$f'(x) = 4(x^3 - 6x^2 + 9x - 4) = 4(x - 1)^2(x - 4).$$

The function f is increasing for $x > 4$ and decreasing for $x < 4$ (except for the critical point $x = 1$). The function changes from decreasing to increasing at $x = 4$; hence, $x = 4$ is a local minimum. The point $x = 1$ is not a local extremum. A graph of this function appears in Figure 1.6.1. $\diamond$

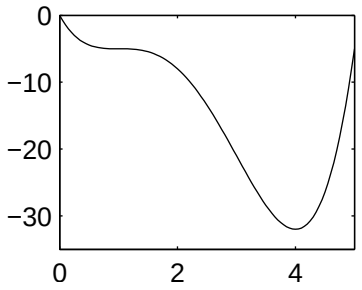

Figure 1.6.1: The graph of $y = x^4 - 8x^3 + 18x^2 - 16x$

Accumulation

Just as applications of the derivative can be cast in terms of tangent slope, applications of the definite integral can be cast in terms of areas. However, it is conceptually simpler to derive applications from the Fundamental Theorem (Theorem 1.5.1). Suppose $f(t)$ is any quantity that accumulates over time. It could be the size of an organism, the population of a community,[1] or the amount of oil released in an oil spill, for example. From a mathematical point of view, all of these applications and any other quantities that accumulate in time are the same. In the form

$$\Delta f = f(b) - f(a) = \int_a^b f'(t)\,dt, \tag{1.6.2}$$

the Fundamental Theorem indicates that the total accumulation of a quantity $f(t)$ over a time interval $a \le t \le b$ is the integral of the rate of change of the quantity. This is important because it is often easier to find a rate of change than it is to find a quantity itself. Viewed as an algebra equation, Equation 1.6.2 says that either $f(a)$ or $f(b)$ can be determined from the other, provided that f' is known on the interval between a and b.

It is often helpful to set up integrals using intuition rather than merely copying Equation 1.6.2.

Example 1.6.3
Suppose a certain organism becomes sexually mature at an average age of a_0 and then produces offspring at a rate $r(a)$ up to a maximum age of a_M. How many offspring are produced by the average individual that survives to age a_M?

To answer this question intuitively, suppose we start a clock when an average individual is born. (This means that age a and time t are interchangeable.) No offspring begin to accumulate until the time reaches a_0. Now think of a as an arbitrary fixed age between a_0 and a_M and think of da as a vanishingly small amount of time. Because da is vanishingly small, the function r takes on the constant value $r(a)$ over the time interval $[a, a + da]$. The amount of offspring production that occurs during this interval, on the average, is then $r(a)\,da$. The total amount of offspring production is the sum of these infinitesimal amounts over all values of a in the interval $[a_0, a_M]$. Interpreting the summation of infinitely many infinitesimal values over an interval as integration,[2] we then write the sum of the

[1] We can think of population decrease as negative accumulation.

[2] Mathematical purists often object to this "vanishingly small" conception, preferring limits of Riemann sums. Our chosen conceptualization is based on an alternative development of calculus, called *nonstandard analysis*. It is less formal and more intuitive than derivation with Riemann sums.

quantities $r(a)\,da$ as

$$\text{total } = \int_{a_0}^{a_M} r(a)\,da.$$

◇

Aggregation

The term *accumulation* has been used to denote the process of building up a quantity by adding amounts over time. In a similar manner, we use the term *aggregation* to indicate the process of building up a quantity by adding amounts over space. We consider here only the case of one-dimensional aggregation, such as aggregation along a narrow line or aggregation through a region with density variation in only one direction. Aggregation is a little more difficult to understand than accumulation because there is no natural flow of space to correspond to the natural flow of time.

Let x be the symbol representing the spatial coordinate over which the aggregation is to occur, and suppose we are interested in the region $a \le x \le b$. Now let $\bar{x}$ be a particular value of x in the region. Then we define $f(\bar{x})$ to be the amount of some quantity in the region $a \le x \le \bar{x}$. We have defined f so that $f(a) = 0$ and the total amount of the quantity in the region is $f(b)$. By the Fundamental Theorem, we have

$$\text{total } = f(b) = f(b) - f(a) = \int_a^b f'(x)\,dx. \tag{1.6.3}$$

The integral is essentially a sum of products of the form $f'(x)\,dx$. These products must have the same dimensions as f. Thus, if f is a mass, then f' must be mass per unit length; if f is population size, then f' must be population per unit length. Whatever quantity f is, we can calculate it by integrating the amount of f per unit length. The integrand f' is the **linear density** of f.

Example 1.6.4
Suppose grains of sand are distributed along a narrow line of length 10, with linear density $1000e^{-0.1x}$ grains per unit length. Then the number of grains located at a point along the line is $f'(x)\,dx = 1000e^{-0.1x}\,dx$. The total number of grains is

$$\int_0^{10} 1000e^{-0.1x}\,dx = -10000\int_0^{10}(-0.1e^{-0.1x})\,dx = -10000\int_0^{10}[e^{-0.1x}]'\,dx$$

$$= -10000[e^{-1} - 1] = 10000\,\frac{e-1}{e}.$$

◇

Volume and Average

Two other applications of integration are worth noting. First, the principle of accumulation can be applied to the problem of determining the volume of an object if we just think of the cross-sectional area perpendicular to the x axis as the linear density of volume.

Example 1.6.5
Let C be the curve given by $y = \sqrt{x}$ from $x = 0$ to $x = 1$. If we rotate the curve C around the x axis, we create a bullet-shaped region. At any particular point x, the cross-section of the region is a circle of radius $\sqrt{x}$. Thus, the area of the cross-section is $A(x) = \pi(\sqrt{x})^2 = \pi x$. The volume corresponding to the point x is the product of the cross-sectional area with the thickness dx, or $A(x)\,dx$. The total volume is the sum of these thin volumes, or

$$V = \int_0^1 A(x)\,dx = \int_0^1 \pi x\,dx = \frac{\pi}{2}\int_0^1 2x\,dx = \frac{\pi}{2}\int_0^1 (x^2)'\,dx = \frac{\pi}{2}(1-0) = \frac{\pi}{2}.$$

◇

Second, we can use integration to determine the average value of a function over an interval, by making use of a clever trick. We demonstrate with an example.

Example 1.6.6
Let $x(t)$ be a function that indicates the coordinates of an object moving along a straight line as a function of time and let $v(t) = x'(t)$ be the velocity of the object. On the time interval $[a, b]$, the object moves a distance $x(b) - x(a)$ in $b - a$ units of time. The average velocity $\bar{v}$ is the ratio of total distance to total time. Thus,

$$\bar{v} = \frac{x(b) - x(a)}{b - a} = \frac{1}{b-a}[x(b) - x(a)] = \frac{1}{b-a}\int_a^b x'(t)\,dt = \frac{1}{b-a}\int_a^b v(t)\,dt.$$

◇

There was nothing special about the velocity function in Example 1.6.6. In general, the average value of a function $f(x)$ on the interval $[a, b]$ is

$$\bar{f} = \frac{1}{b-a}\int_a^b f(x)\,dx. \tag{1.6.4}$$

Summary

Applications of the derivative are all tied, directly or indirectly, to its representing the slope of the tangent to $y = f(x)$. Applications of the definite integral can be thought of as stemming from the Fundamental Theorem of Calculus. In particular,

$$\text{total accumulation of stuff over } a \le t \le b = \int_a^b [\text{stuff per unit time}]\,dt$$

and

$$\text{total aggregation of stuff over } a \le x \le b = \int_a^b [\text{stuff per unit length}]\,dx.$$

These schematic equations point out that accumulation and aggregation are really the same thing, just with different symbols for the independent variables.