MAT 169: Calculus III with Analytic Geometry

James V. Lambers

April 12, 2012

Contents

Chapter 1

Sequences and Series

1.1 Introduction

This course is the third course in the calculus sequence, following MAT 167 and MAT 168. Its purpose is to prepare students for more advanced mathematics courses, particularly those pertaining to multivariable calculus (MAT 280) and numerical analysis (MAT 460 and 461). The course will focus on three main areas, which we briefly discuss here.

1.1.1 Sequences and Series

Nearly all students have had to use a scientific calculator. Consider the following: how does a calculator efficiently evaluate many of its functions, such as sin, cos or exp, when its hardware is only able to perform the four basic arithmetic operations, addition, subtraction, multiplication and division?

The answer stems from the fact that it is generally not possible to evaluate such functions exactly; rather, one has to settle for an *approximation*. However, this is no problem, because a calculator or computer can only represent real numbers to limited accuracy anyway. To approximate a given function in a manner that is suitable for a calculator, we use *infinite series*, which is a sum of infinitely many terms. For example, we can write

$$\exp x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots = \sum_{k=0}^{\infty} \frac{1}{k!}x^k.$$

The last term in the above equation uses *sigma notation* to express a sum of infinitely many terms in a concise way, when those terms can be described by a pattern. The above series is called a *power series* because each of its

terms features a power of x. When we study infinite series, we will consider important questions such as

- When does an infinite series sum, or *converge*, to a finite number? We will see that in many cases, a sum of infinitely many terms does not converge at all, but rather continues growing, or *diverging*. For example, consider the two series

$$\sum_{n=1}^{\infty} \frac{1}{n}, \quad \sum_{n=1}^{\infty} \frac{1}{n^{1.01}}.$$

  Although the individual terms in the series may not differ by very much, especially for larger values of n, the first series does not converge, while the second one does.

- If we truncate a series after a given number of terms, how well does the sum of the retained terms approximate the sum of the entire series? This is particularly important when using series to evaluate functions such as those implemented in scientific calculators. For example, suppose we take the abovementioned series for $\exp x$ and compute only the first 20 terms. If $x = 0.1$, then the result is correct to at least 16 significant digits. However, if $x = 10$, then we only obtain two correct digits.

1.1.2 Vectors and the Geometry of Space

Next, we will become acquainted with three-dimensional space, in order to prepare you for later coursework in multivariable calculus. Our basic tools will be *vectors*, which can be used to represent either a position or direction in space. For example, if we represent three-dimensional space using *Cartesian* coordinates x, y and z, then the *origin* is the point with coordinates $x = 0$, $y = 0$ and $z = 0$, typically denoted by the ordered triple $(0, 0, 0)$. Then, the point in space $(1, 0, -2)$ has coordinates $x = 1$, $y = 0$ and $z = -2$, meaning that it is located 1 unit from the origin along the positive x-axis, 0 units from the origin along the y-axis, and 2 units from the origin along the *negative* z-axis. This is illustrated in Figure 1.1. We will use vectors to facilitate the description of, and operations on, lines and planes. This is particularly useful in computer graphics. Consider the problem of rendering a two-dimensional image, say for a frame in a film, of a collection of three-dimensional objects. To what point in 2-D space does any given point in 3-D space correspond? This question is answered by computing

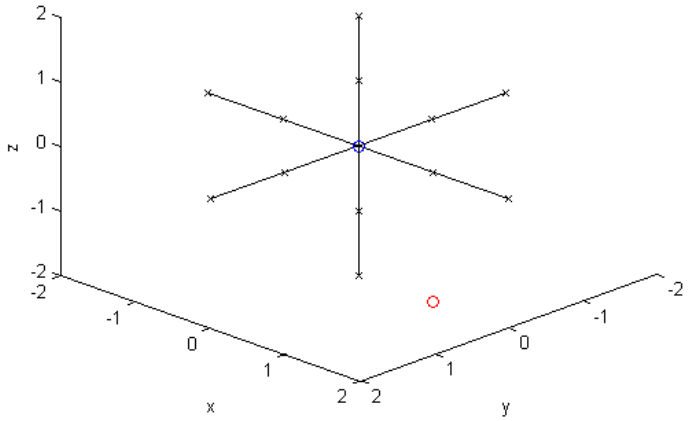

Figure 1.1: The origin in three-dimensional space, and the point $(1, 0, -2)$.

the *projection* of points in 3-D space onto a given plane in 2-D space that corresponds to the "screen". We will learn about an operation called the *dot product* that is used to compute projections. Another operation, called the *cross product*, is very useful for describing planes.

1.1.3 Parametric Equations and Polar Coordinates

Finally, will study curves and surfaces in space. Often, it is necessary to describe the position of an object as a function of time. Therefore, we will study curves that are described by *parametric equations*, where the parameter is usually time. For example, in combat, the military needs to track the position of enemy projectiles over time. Using radar data, they can then construct parametric equations for the projectile's position in 3-D space, and then differentiate the equations with respect to time in order to estimate its velocity and then its trajectory, so that it can be intercepted.

While we will work with Cartesian coordinates x, y and z for most of the course, we will find that it is often useful to represent curves or functions in the xy-plane using *polar coordinates* r and θ, where r represents *distance* from the origin, and θ represents the *angle* that the point makes with the origin and the positive x-axis. For example, the point $x = 0$, $y = 1$, which is

1 unit from the origin and makes a 90-degree angle with the origin and the positive x-axis, has polar coordinates $(1, \pi/2)$. Certain curves are far more easily described using polar coordinates. To see this, consider the equations

$$r = \sin 2t, \quad (x^2 + y^2)^{3/2} = 2xy.$$

These two equations describe the same curve, which resembles a four-leaf clover. We will learn how to compute areas of regions enclosed by parametric curves, and lengths of parametric curves, that are expressed in either Cartesian or polar coordinates.

1.2 Review of Calculus

We now review some basic concepts from the first two calculus courses before we begin our study of the third. Recall that these two courses focused on the following fundamental problems:

- computing the instantaneous rate of change of one quantity with respect to another, which is a *derivative*, and
- computing the total change in a function over some portion of its domain, which is a *definite integral*.

1.2.1 Limits

The basic problems of differential and integral calculus described in the previous paragraph can be solved by computing a sequence of approximations to the desired quantity and then determining what value, if any, the sequence of approximations approaches. This value is called a *limit* of the sequence. As a sequence is a function, we begin by defining, precisely, the concept of the limit of a function.

Definition *We write*

$$\lim_{x \to a} f(x) = L$$

if for any open interval I_1 containing L, there is some open interval I_2 containing a such that $f(x) \in I_1$ whenever $x \in I_2$, and $x \neq a$. We say that L is the **limit of** $f(x)$ **as** x **approaches** a.

We write

$$\lim_{x \to a^-} f(x) = L$$

if, for any open interval I_1 containing L, there is an open interval I_2 of the form (c, a), where $c < a$, such that $f(x) \in I_1$ whenever $x \in I_2$. We say that

L is the **limit of $f(x)$ as x approaches a from the left**, *or the* **left-hand limit of $f(x)$ as x approaches a.**

Similarly, we write

$$\lim_{x \to a^+} f(x) = L$$

if, for any open interval I_1 containing L, there is an open interval I_2 of the form (a, c), where $c > a$, such that $f(x) \in I_1$ whenever $x \in I_2$. We say that L is the **limit of $f(x)$ as x approaches a from the right**, *or the* **right-hand limit of $f(x)$ as x approaches a.**

We can make the definition of a limit a little more concrete by imposing sizes on the intervals I_1 and I_2, as long as the interval I_1 can still be of arbitrary size. It can be shown that the following definition is equivalent to the previous one.

Definition *We write*

$$\lim_{x \to a} f(x) = L$$

if, for any $\epsilon > 0$, there exists a number $\delta > 0$ such that $|f(x) - L| < \epsilon$ whenever $0 < |x - a| < \delta$.

Similar definitions can be given for the left-hand and right-hand limits.

Note that in either definition, the point $x = a$ is specifically excluded from consideration when requiring that $f(x)$ be close to L whenever x is close to a. This is because the concept of a limit is only intended to describe the behavior of $f(x)$ *near* $x = a$, as opposed to its behavior *at* $x = a$. Later in this section we discuss the case where the two distinct behaviors coincide.

1.2.2 Continuity

In many cases, the limit of a function $f(x)$ as x approached a could be obtained by simply computing $f(a)$. Intuitively, this indicates that f has to have a graph that is one continuous curve, because any "break" or "jump" in the graph at $x = a$ is caused by f approaching one value as x approaches a, only to actually assume a different value at a. This leads to the following precise definition of what it means for a function to be continuous at a given point.

Definition *(Continuity) We say that a function f is* **continuous at** *a if*

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

We also say that $f(x)$ has the **Direct Subsitution Property** *at $x = a$.*

We say that a function f is **continuous from the right** *at a if*

$$\lim_{x \to a^+} f(x) = f(a).$$

Similarly, we say that f is **continuous from the left** *at a if*

$$\lim_{x \to a^-} f(x) = f(a).$$

The preceding definition describes continuity at a single point. In describing where a function is continuous, the concept of continuity over an interval is useful, so we define this concept as well.

Definition *(Continuity on an Interval) We say that a function f is* **continuous on the interval** (a, b) *if f is continuous at every point in (a, b). Similarly, we say that f is continuous on*

1. *$[a, b)$ if f is continuous on (a, b), and continuous from the right at a.*
2. *$(a, b]$ if f is continuous on (a, b), and continuous from the left at b.*
3. *$[a, b]$ if f is continuous on (a, b), continuous from the right at a, and continuous from the left at b.*

1.2.3 Derivatives

The basic problem of differential calculus is computing the instantaneous rate of change of one quantity y with respect to another quantity x. For example, y may represent the position of an object and x may represent time, in which case the instantaneous rate of change of y with respect to x is interpreted as the velocity of the object.

When the two quantities x and y are related by an equation of the form $y = f(x)$, it is certainly convenient to describe the rate of change of y with respect to x in terms of the function f. Because the instantaneous rate of change is so commonplace, it is practical to assign a concise name and notation to it, which we do now.

Definition *(Derivative) The* **derivative** *of a function $f(x)$ at $x = a$, denoted by $f'(a)$, is*

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

provided that the above limit exists. When this limit exists, we say that f is **differentiable** *at a.*

Remark Given a function $f(x)$ that is differentiable at $x = a$, the following numbers are all equal:

- the derivative of f at $x = a$, $f'(a)$,
- the slope of the tangent line of f at the point $(a, f(a))$, and
- the instantaneous rate of change of $y = f(x)$ with respect to x at $x = a$.

This can be seen from the fact that all three numbers are defined in the same way. □

1.2.4 Riemann Sums and the Definite Integral

There are many cases in which some quantity is defined to be the product of two other quantities. For example, a rectangle of width w has uniform height h, and the area A of the rectangle is given by the formula $A = wh$. Unfortunately, in many applications, we cannot necessarily assume that certain quantities such as height are constant, and therefore formulas such as $A = wh$ cannot be used directly. However, they can be used indirectly to solve more general problems by employing the notation known as *integral calculus*.

Suppose we wish to compute the area of a shape that is not a rectangle. To simplify the discussion, we assume that the shape is bounded by the vertical lines $x = a$ and $x = b$, the x-axis, and the curve defined by some continuous function $y = f(x)$, where $f(x) \geq 0$ for $a \leq x \leq b$. Then, we can approximate this shape by n rectangles that have width $\Delta x = (b - a)/n$ and height $f(x_i)$, where $x_i = a + i\Delta x$, for $i = 0, \ldots, n$. We obtain the approximation

$$A \approx A_n = \sum_{i=1}^{n} f(x_i)\Delta x.$$

Intuitively, we can conclude that as $n \to \infty$, the approximate area A_n will converge to the exact area of the given region. This can be seen by observing that as n increases, the n rectangles defined above comprise a more accurate approximation of the region.

More generally, suppose that for each $n = 1, 2, \ldots$, we define the quantity R_n by choosing points $a = x_0 < x_1 < \cdots < x_n = b$, and computing the sum

$$R_n = \sum_{i=1}^{n} f(x_i^*)\Delta x_i, \quad \Delta x_i = x_i - x_{i-1}, \quad x_{i-1} \leq x_i^* \leq x_i.$$

The sum that defines R_n is known as a *Riemann sum*. Note that the interval $[a, b]$ need not be divided into subintervals of equal width, and that $f(x)$ can be evaluated at *arbitrary* points belonging to each subinterval.

If $f(x) \geq 0$ on $[a, b]$, then R_n converges to the area under the curve $y = f(x)$ as $n \to \infty$, provided that the widths of all of the subintervals $[x_{i-1}, x_i]$, for $i = 1, \ldots, n$, approach zero. This behavior is ensured if we require that

$$\lim_{n\to\infty} \delta(n) = 0, \quad \text{where} \quad \delta(n) = \max_{1\leq i\leq n} \Delta x_i.$$

This condition is necessary because if it does not hold, then, as $n \to \infty$, the region formed by the n rectangles will not converge to the region whose area we wish to compute. If f assumes negative values on $[a, b]$, then, under the same conditions on the widths of the subintervals, R_n converges to the *net* area between the graph of f and the x-axis, where area below the x-axis is counted negatively.

We define the *definite integral* of $f(x)$ from a to b by

$$\int_a^b f(x)\, dx = \lim_{n\to\infty} R_n,$$

where the sequence of Riemann sums $\{R_n\}_{n=1}^{\infty}$ is defined so that $\delta(n) \to 0$ as $n \to \infty$, as in the previous discussion. The function $f(x)$ is called the *integrand*, and the values a and b are the *lower* and *upper limits of integration*, respectively. The process of computing an integral is called *integration*.

1.2.5 Extreme Values

In many applications, it is necessary to determine where a given function attains its minimum or maximum value. For example, a business wishes to maximize profit, so it can construct a function that relates its profit to variables such as payroll or maintenance costs. We now consider the basic problem of finding a maximum or minimum value of a general function $f(x)$ that depends on a single independent variable x. First, we must precisely define what it means for a function to *have* a maximum or minimum value.

Definition *(Absolute extrema) A function f has a* **absolute maximum** *or* **global maximum** *at c if $f(c) \geq f(x)$ for all x in the domain of f. The number $f(c)$ is called the* **maximum value** *of f on its domain. Similarly, f has a* **absolute minimum** *or* **global minimum** *at c if $f(c) \leq f(x)$ for all*

x in the domain of f. The number $f(c)$ is then called the **minimum value** *of f on its domain. The maximum and minimum values of f are called the* **extreme values** *of f, and the absolute maximum and minimum are each called an* **extremum** *of f.*

Before computing the maximum or minimum value of a function, it is natural to ask whether it is possible to determine in advance whether a function even has a maximum or minimum, so that effort is not wasted in trying to solve a problem that has no solution. The following result is very helpful in answering this question.

Theorem *(Extreme Value Theorem) If f is continuous on $[a, b]$, then f has an absolute maximum and an absolute minimum on $[a, b]$.*

Now that we can easily determine whether a function has a maximum or minimum on a closed interval $[a, b]$, we can develop an method for actually finding them. It turns out that it is easier to find points at which f attains a maximum or minimum value in a "local" sense, rather than a "global" sense. In other words, we can best find the absolute maximum or minimum of f by finding points at which f achieves a maximum or minimum with respect to "nearby" points, and then determine which of these points is the absolute maximum or minimum. The following definition makes this notion precise.

Definition *(Local extrema) A function f has a* **local maximum** *at c if $f(c) \geq f(x)$ for all x in an open interval containing c. Similarly, f has a* **local minimum** *at c if $f(c) \leq f(x)$ for all x in an open interval containing c. A local maximum or local minimum is also called a* **local extremum**.

At each point at which f has a local maximum, the function either has a horizontal tangent line, or no tangent line due to not being differentiable. It turns out that this is true in general, and a similar statement applies to local minima. To state the formal result, we first introduce the following definition, which will also be useful when describing a method for finding local extrema.

Definition*(Critical Number) A number c in the domain of a function f is a* **critical number** *of f if $f'(c) = 0$ or $f'(c)$ does not exist.*

The following result describes the relationship between critical numbers and local extrema.

Theorem *(Fermat's Theorem) If f has a local minimum or local maximum*

at c, then c is a critical number of f; that is, either $f'(c) = 0$ *or* $f'(c)$ *does not exist.*

This theorem suggests that the maximum or minimum value of a function $f(x)$ can be found by solving the equation $f'(x) = 0$.

1.2.6 The Mean Value Theorem

While the derivative describes the behavior of a function at a point, we often need to understand how the derivative influences a function's behavior on an interval. It is often necessary to approximate a function $f(x)$ by a function $g(x)$ using knowledge of $f(x)$ and its derivatives at various points. It is therefore natural to ask how well $g(x)$ approximates $f(x)$ away from these points.

The following result, a consequence of Fermat's Theorem, gives limited insight into the relationship between the behavior of a function on an interval and the value of its derivative at a point.

Theorem *(Rolle's Theorem) If f is continuous on a closed interval* $[a, b]$ *and is differentiable on the open interval* (a, b)*, and if* $f(a) = f(b)$*, then* $f'(c) = 0$ *for some number c in* (a, b)*.*

By applying Rolle's Theorem to a function f, then to its derivative f', its second derivative f'', and so on, we obtain the following more general result, which will be useful in analyzing the accuracy of methods for approximating functions by polynomials.

Theorem *(Generalized Rolle's Theorem) Let* $x_0, x_1, x_2, \ldots, x_n$ *be distinct points in an interval* $[a, b]$*. If f is n times differentiable on* (a, b)*, and if* $f(x_i) = 0$ *for* $i = 0, 1, 2, \ldots, n$*, then* $f^{(n)}(c) = 0$ *for some number c in* (a, b)*.*

A more fundamental consequence of Rolle's Theorem is the Mean Value Theorem itself, which we now state.

Theorem *(Mean Value Theorem) If f is continuous on a closed interval* $[a, b]$ *and is differentiable on the open interval* (a, b)*, then*

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

for some number c in (a, b)*.*

Remark The expression

$$\frac{f(b) - f(a)}{b - a}$$

is the slope of the secant line passing through the points $(a, f(a))$ and $(b, f(b))$. The Mean Value Theorem therefore states that under the given assumptions, the slope of this secant line is equal to the slope of the tangent line of f at the point $(c, f(c))$, where $c \in (a, b)$. $\square$

The Mean Value Theorem has the following practical interpretation: the average rate of change of $y = f(x)$ with respect to x on an interval $[a, b]$ is equal to the instantaneous rate of change y with respect to x at some point in (a, b).

1.2.7 The Mean Value Theorem for Integrals

Suppose that $f(x)$ is a continuous function on an interval $[a, b]$. Then, by the Fundamental Theorem of Calculus, $f(x)$ has an antiderivative $F(x)$ defined on $[a, b]$ such that $F'(x) = f(x)$. If we apply the Mean Value Theorem to $F(x)$, we obtain the following relationship between the integral of f over $[a, b]$ and the value of f at a point in (a, b).

Theorem *(Mean Value Theorem for Integrals) If f is continuous on $[a, b]$, then*

$$\int_a^b f(x)\,dx = f(c)(b - a)$$

for some c in (a, b).

In other words, f assumes its average value over $[a, b]$, defined by

$$f_{ave} = \frac{1}{b - a}\int_a^b f(x)\,dx,$$

at some point in $[a, b]$, just as the Mean Value Theorem states that the derivative of a function assumes its average value over an interval at some point in the interval.

The Mean Value Theorem for Integrals is also a special case of the following more general result.

Theorem *(Weighted Mean Value Theorem for Integrals) If f is continuous on $[a, b]$, and g is a function that is integrable on $[a, b]$ and does not change sign on $[a, b]$, then*

$$\int_a^b f(x)g(x)\,dx = f(c)\int_a^b g(x)\,dx$$

for some c in (a, b).

In the case where $g(x)$ is a function that is easy to antidifferentiate and $f(x)$ is not, this theorem can be used to obtain an estimate of the integral of $f(x)g(x)$ over an interval.

Example Let $f(x)$ be continuous on the interval $[a, b]$. Then, for any $x \in [a, b]$, by the Weighted Mean Value Theorem for Integrals, we have

$$\int_a^x f(s)(s-a)\,ds = f(c)\int_a^x (s-a)\,ds = f(c)\,\frac{(s-a)^2}{2}\bigg|_a^x = f(c)\frac{(x-a)^2}{2},$$

where $a < c < x$. It is important to note that we can apply the Weighted Mean Value Theorem because the function $g(x) = (x-a)$ does not change sign on $[a, b]$. □

1.3 Taylor's Theorem

In many cases, it is useful to approximate a given function $f(x)$ by a polynomial, because one can work much more easily with polynomials than with other types of functions. As such, it is necessary to have some insight into the accuracy of such an approximation. The following theorem, which is a consequence of the Weighted Mean Value Theorem for Integrals, provides this insight.

Theorem *(Taylor's Theorem) Let f be n times continuously differentiable on an interval $[a, b]$, and suppose that $f^{(n+1)}$ exists on $[a, b]$. Let $x_0 \in [a, b]$. Then, for any point $x \in [a, b]$,*

$$f(x) = P_n(x) + R_n(x),$$

where

$$\begin{aligned} P_n(x) &= \sum_{j=0}^{n} \frac{f^{(j)}(x_0)}{j!}(x-x_0)^j \\ &= f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2}f''(x_0)(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n \end{aligned}$$

and

$$R_n(x) = \int_{x_0}^{x} \frac{f^{(n+1)}(s)}{n!}(x-s)^n\,ds = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)^{n+1},$$

where $\xi(x)$ is between x_0 and x.

The polynomial $P_n(x)$ is the nth *Taylor polynomial* of f with *center* x_0, and the expression $R_n(x)$ is called the *Taylor remainder* of $P_n(x)$. When the center x_0 is zero, the nth Taylor polynomial is also known as the nth *Maclaurin polynomial*.

The final form of the remainder is obtained by applying the Mean Value Theorem for Integrals to the integral form. As $P_n(x)$ can be used to approximate $f(x)$, the remainder $R_n(x)$ is also referred to as the *truncation error* of $P_n(x)$. The accuracy of the approximation on an interval can be analyzed by using techniques for finding the extreme values of functions to bound the $(n+1)$-st derivative on the interval.

Because approximation of functions by polynomials is employed in the development and analysis of many techniques in numerical analysis, the usefulness of Taylor's Theorem cannot be overstated. In fact, it can be said that Taylor's Theorem is the Fundamental Theorem of Numerical Analysis, just as the theorem describing inverse relationship between derivatives and integrals is called the Fundamental Theorem of Calculus.

We conclude our discussion of Taylor's Theorem with some examples that illustrate how the nth-degree Taylor polynomial $P_n(x)$ and the remainder $R_n(x)$ can be computed for a given function $f(x)$.

Example If we set $n = 1$ in Taylor's Theorem, then we have

$$f(x) = P_1(x) + R_1(x)$$

where

$$P_1(x) = f(x_0) + f'(x_0)(x - x_0).$$

This polynomial is a linear function that describes the tangent line to the graph of f at the point $(x_0, f(x_0))$.

If we set $n = 0$ in the theorem, then we obtain

$$f(x) = P_0(x) + R_0(x),$$

where

$$P_0(x) = f(x_0)$$

and

$$R_0(x) = f'(\xi(x))(x - x_0),$$

where $\xi(x)$ lies between x_0 and x. If we use the integral form of the remainder,

$$R_n(x) = \int_{x_0}^{x} \frac{f^{(n+1)}(s)}{n!}(x - s)^n \, ds,$$

then we have

$$f(x) = f(x_0) + \int_{x_0}^{x} f'(s)\,ds,$$

which is equivalent to the Total Change Theorem and part of the Fundamental Theorem of Calculus. Using the Mean Value Theorem for integrals, we can see how the first form of the remainder can be obtained from the integral form. □

Example Let $f(x) = \sin x$. Then

$$f(x) = P_3(x) + R_3(x),$$

where

$$P_3(x) = x - \frac{x^3}{3!} = x - \frac{x^3}{6},$$

and

$$R_3(x) = \frac{1}{4!}x^4 \sin\xi(x) = \frac{1}{24}x^4 \sin\xi(x),$$

where $\xi(x)$ is between 0 and x. The polynomial $P_3(x)$ is the 3rd Maclaurin polynomial of $\sin x$, or the 3rd Taylor polynomial with center $x_0 = 0$.

If $x \in [-1, 1]$, then

$$|R_n(x)| = \left|\frac{1}{24}x^4 \sin\xi(x)\right| = \left|\frac{1}{24}\right| |x^4||\sin\xi(x)| \leq \frac{1}{24},$$

since $|\sin x| \leq 1$ for all x. This bound on $|R_n(x)|$ serves as an upper bound for the error in the approximation of $\sin x$ by $P_3(x)$ for $x \in [-1, 1]$. □

Example Let $f(x) = e^x$. Then

$$f(x) = P_2(x) + R_2(x),$$

where

$$P_2(x) = 1 + x + \frac{x^2}{2},$$

and

$$R_2(x) = \frac{x^3}{6}e^{\xi(x)},$$

where $\xi(x)$ is between 0 and x. The polynomial $P_2(x)$ is the 2nd Maclaurin polynomial of e^x, or the 2nd Taylor polynomial with center $x_0 = 0$.

If $x > 0$, then $R_2(x)$ can become quite large, whereas its magnitude is much smaller if $x < 0$. Therefore, one method of computing e^x using a

Maclaurin polynomial is to use the nth Maclaurin polynomial $P_n(x)$ of e^x when $x < 0$, where n is chosen sufficiently large so that $R_n(x)$ is small for the given value of x. If $x > 0$, then we instead compute e^{-x} using the nth Maclaurin polynomial for e^{-x}, which is given by

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

and then obtaining an approximation to e^x by taking the reciprocal of our computed value of e^{-x}. □

Example Let $f(x) = x^2$. Then, for any real number x_0,

$$f(x) = P_1(x) + R_1(x),$$

where

$$P_1(x) = x_0^2 + 2x_0(x - x_0) = 2x_0x - x_0^2,$$

and

$$R_1(x) = (x - x_0)^2.$$

Note that the remainder does not include a "mystery point" $\xi(x)$ since the 2nd derivative of x^2 is only a constant. The linear function $P_1(x)$ describes the tangent line to the graph of $f(x)$ at the point $(x_0, f(x_0))$. If $x_0 = 1$, then we have

$$P_1(x) = 2x - 1,$$

and

$$R_1(x) = (x - 1)^2.$$

We can see that near $x = 1$, $P_1(x)$ is a reasonable approximation to x^2, since the error in this approximation, given by $R_1(x)$, would be small in this case. □

1.4 Sequences

1.4.1 What is a Sequence?

A *sequence* is an ordered list of numbers. The ordering of the numbers in a sequence is indicated by an *index* that is associated with each number. Usually, the indices are taken from the *natural numbers*, which are the positive integers 1, 2, 3, ... We will consider *infinite* sequences such as

$$a_1, a_2, a_3, \ldots, a_n, \ldots.$$

Each number in a sequence is called a *term*. In the above example, a_1 is the first term, a_2 is the second term, and a_n is the nth term.

There are many ways to describe sequences, but they all have one thing in common: any description of a sequence should include formula for computing the nth term.

Example The sequence of all even natural numbers can be described by

$$a_n = 2n, \quad n \geq 1, \quad \text{or} \quad \{2n\}_{n=1}^{\infty}.$$

Similarly, the sequence of all odd natural numbers can be defined by

$$a_n = 2n - 1, \quad n \geq 1, \quad \text{or} \quad \{2n-1\}_{n=1}^{\infty}.$$

Note that in both cases, it is explicitly stated that the index of the first term is 1; this will usually be the case, but in some contexts, it is more intuitive to use a different number, such as 0, as the index of the first term. Note that the second form of each sequence does not specify a variable, such as a, to identify a given term. In this case, a different expression may be used, such as x or α. In the $\{\}$ notation, the indices can be omitted if they are already known. □

Example It is often convenient to define each term in a sequence in terms of preceding terms. For example, we can define

$$a_{n+1} = \frac{a_n}{2} + \frac{1}{a_n}, \quad n \geq 1,$$

provided that we include a definition of a_1 to start the sequence. In this case, the sequence begins as follows:

$$a_1 = 1, \quad a_2 = 1.5, \quad a_3 = 1.41\bar{6}.$$

What happens to a_n as n increases?

Sequences can also be defined in terms of more than one preceding element. The best-known sequence of this type is the *Fibonacci sequence*, defined by

$$f_1 = 1, \quad f_2 = 1, \quad f_n = f_{n-1} + f_{n-2}.$$

This sequence, whose terms increase rapidly, arose from the study of breeding of rabbits. Because the definition of f_n, for $n > 2$, involves three terms of the sequence, we say that the Fibonacci sequence obeys a *three-term recurrence relation*.

The Fibonacci sequence can be defined using a formula for f_n that is in *closed form*; that is, it does not refer to any previous terms. However, as this formula is

$$f_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n,$$

it is generally preferable to use the three-term recurrence relation to compute terms of the sequence, unless, for example, one wants to compute f_n for *only* a few large values of n. □

1.4.2 Why Do We Need Sequences?

In this chapter, our primary goal is to learn how to work with infinite *series*, which are *sums* of the terms of a sequence, for various applications as approximating functions in a manner that is feasible for calculators or computers. We will see that in order to define what it means to add infinitely many numbers, we need the concept of a sequence.

Sequences are also useful in other contexts that are unrelated to infinite series. Many methods for solving mathematical problems on a computer are *iterative* in nature; that is, they produce a sequence of approximate solutions to a given problem that, hopefully, are getting closer to the exact solution, in some sense. An example of this, which may be familiar to you, is *Newton's method* for finding the roots, or zeros, of a function.

1.4.3 Recognizing Sequences

The definitions of the sequences in the previous examples can be used to compute the terms of the sequence, but sometimes, it is necessary to go the other way, and obtain a definition of a sequence from some of its terms.

There is no completely deterministic way to derive a formula for the terms of a given sequence, but a useful strategy is to examine how each term differs from the previous one. This examination yields the following clues to a sequence's definition:

- Do the terms alternate in sign? If so, this suggests that the formula for the nth term should include a factor of $(-1)^n$, which is equal to 1 when n is even and -1 when n is odd.

- Do consecutive terms, or portions of them, have a constant ratio? Selected terms may be given as fractions, and it might be observed that

either the numerators or denominators change by being multiplied by a factor of, for example, 3, in which case the formula for the nth term likely includes a factor of 3^n.

- Do consecutive terms, or portions of them, differ by a constant amount? For example, it might be observed that each term is 4 more than the previous term, which suggests that the definition of the nth term should include $4n$.

Example Consider the sequence whose first few terms are

$$a_1 = 2, \quad a_2 = -\frac{8}{4}, \quad a_3 = \frac{32}{7}, \quad a_4 = -\frac{128}{10}.$$

The terms alternate in sign, with the even-numbered terms being negative, so the definition of a_n should include the factor $(-1)^{n+1}$, which is equal to -1 when n is even. If we rewrite $a_1 = 2/1$, we see that the denominators differ by 3, so we can express the denominator of a_n as $3n - 2$. Finally, the numerators are 2, 8, 32 and 128, which are equal to 2^1, 2^3, 2^5 and 2^7. That is, they are odd powers of 2, so use the first example above and conclude

$$a_n = (-1)^{n+1} \frac{2^{2n-1}}{3n-2}, \quad n \geq 1.$$

□

1.4.4 Limits of Sequences

There are two key questions that need to be asked about any given sequence:

1. As the index increases, do the terms in the sequence *converge* to a particular value?

2. If so, what is that value? If not, how does the sequence behave? That is, do its terms become infinitely large, or do they remain bounded but continually oscillate?

Before we can attempt to answer these questions, we need to precisely define what it means for the terms of a sequence to converge.

To help us to formulate an appropriate definition, we consider an example. Consider the sequence

$$a_n = \frac{n^2 + n}{n^2 + 2n + 1}.$$

If we rewrite this sequence as

$$\begin{aligned} a_n &= \frac{n^2+2n+1-(2n+1)+n}{n^2+2n+1} \\ &= \frac{n^2+2n+1}{n^2+2n+1} - \frac{n+1}{n^2+2n+1} \\ &= 1 - \frac{n+1}{n^2+2n+1} \\ &= 1 - \frac{n+1}{(n+1)^2} \\ &= 1 - \frac{1}{n+1}, \end{aligned}$$

we see that the terms in this sequence become closer to 1 as n increases, because the fraction $1/(n+1)$ decreases toward zero. In fact, by choosing n large enough, we can make $1/(n+1)$ as small as we want. Specifically, suppose we want $1/(n+1) < \epsilon$ for some $\epsilon > 0$. By solving this inequality for n, we find that we need only choose n so that $n > 1/\epsilon - 1$. We conclude that a_n converges to 1 as n tends to infinity.

In general, we say that a sequence $\{a_n\}_{n=1}^{\infty}$ *converges* to a *limit* L if, for any $\epsilon > 0$, we can find an index N such that $|a_n - L| < \epsilon$ whenever $n > N$. Informally, the sequence converges to L if we can make all of its terms, beyond some index, as close to L as we want. We write

$$\lim_{n\to\infty} a_n = L.$$

If the sequence $\{a_n\}$ converges to L, we say that it is a *convergent* sequence. Otherwise, we say that it is *divergent*.

Now, consider the sequence defined by

$$a_n = \frac{n^2+4n+3}{n+2}.$$

By rewriting this sequence as

$$\begin{aligned} a_n &= \frac{n^2+4n+4-4+3}{n+2} \\ &= \frac{n^2+4n+4}{n+2} - \frac{1}{n+2} \\ &= n+2 - \frac{1}{n+2}, \end{aligned}$$

we see that as n increases, a_n also increases. In fact, by choosing n large enough, we can make a_n as large as we want. Therefore, not only does this sequence diverge, but it tends toward infinity as n does.

In general, we say that

$$\lim_{n\to\infty} a_n = \infty$$

if, for any positive number M, there is an index N such that $a_n > M$ whenever $n > N$. That is, we can make all terms in the sequence, beyond some index, as large as we want. We also say that the sequence $\{a_n\}$ *diverges to* ∞.

1.4.5 Relation to Limits of Functions

Previously, we have learned much about how to compute limits of functions at infinity. It is therefore natural to ask whether the tools used to compute limits of functions can be applied to compute limits of sequences.

Thankfully, that is in fact the case. Specifically, if a function $f(x)$ satisfies

$$\lim_{x\to\infty} f(x) = L,$$

and we define the sequence $\{a_n\}_{n=1}^{\infty}$ by $a_n = f(n)$ where n is a positive integer, then we can conclude that

$$\lim_{n\to\infty} a_n = L.$$

It follows that for any sequence $\{a_n\}$ for which we have a formula to define each term a_n, we can apply, to a function defined using this same formula, all of the available techniques for computing limits of functions at infinity in order to compute the limit of the sequence, if it has one.

In particular, this result allows us to apply the *limit laws* to conclude that the limit of a sum, difference, product, or quotient of convergent sequences is equal to the sum, difference, product or quotient of their limits, respectively. Furthermore, raising the terms of a convergent sequence to a positive power raises its limit to that same power, if the terms are positive. We can also apply the *Squeeze Theorem* to show that if the terms of a sequence $\{a_n\}$ are bounded above and below by two convergent sequences that have the same limit, then $\{a_n\}$ converges to this limit as well.

We can also use techniques for computing limits at infinity of rational functions, as demonstrated in the following example.

Example Consider the sequence defined by

$$a_n = \frac{2n^3 + 3n^2 + 4n + 1}{n^3 + 5n^2 + 3n + 2}, \quad n \geq 1.$$

This sequence is obtained by taking the values of the function

$$f(x) = \frac{2x^3 + 3x^2 + 4x + 1}{x^3 + 5x^2 + 3x + 2}, \quad x \geq 0.$$

The limit of this function as $x \to \infty$ can be computed by dividing the numerator and denominator by the highest power of x, which is x^3. This yields

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} \frac{2 + \frac{3}{x} + \frac{4}{x^2} + \frac{1}{x^3}}{1 + \frac{5}{x} + \frac{3}{x^2} + \frac{2}{x^3}} = 2.$$

We conclude that $\lim_{n\to\infty} a_n = 2$ as well. □

We introduced the concept of an infinite sequence of numbers, and precisely defined what it means for such a sequence to converge to a limit, or approach infinity. Now, we will discuss various techniques, based on these definitions, for testing whether a sequence converges, and, if it does, finding its limit.

1.4.6 Testing Convergence of Sequences

We have learned that techniques for computing the limit of a *function* $f(x)$, as $x \to \infty$, can be applied to compute the limit of a *sequence* $\{a_n\}$ as $n \to \infty$, through the relation $a_n = f(n)$, where n is any index of a term in $\{a_n\}$. We applied this approach earlier in this section to a sequence $\{a_n\}$ in which a_n was a quotient of polynomials, and divided each by the highest power of the variable in order to compute the limit. We now illustrate the use of some other techniques for computing limits of functions at infinity in order to compute limits of sequences.

Example Consider the sequence

$$a_n = \sqrt{n^2 + 1} - \sqrt{n^2 - 1}, \quad n \geq 1.$$

To determine whether it converges, we compute the limit of the function

$$f(x) = \sqrt{x^2 + 1} - \sqrt{x^2 - 1}, \quad x \geq 1,$$

as $x \to \infty$. By multiplying and dividing $f(x)$ by its *conjugate*, we obtain

$$\begin{aligned} f(x) &= \left(\sqrt{x^2+1} - \sqrt{x^2-1}\right) \frac{\sqrt{x^2+1} + \sqrt{x^2-1}}{\sqrt{x^2+1} + \sqrt{x^2-1}} \\ &= \frac{(\sqrt{x^2+1} - \sqrt{x^2-1})(\sqrt{x^2+1} + \sqrt{x^2-1})}{\sqrt{x^2+1} + \sqrt{x^2-1}} \\ &= \frac{(x^2+1) - (x^2-1)}{\sqrt{x^2+1} + \sqrt{x^2-1}} \\ &= \frac{2}{\sqrt{x^2+1} + \sqrt{x^2-1}} \\ &= \frac{2}{|x| \left(\sqrt{1+\frac{1}{x^2}} + \sqrt{1-\frac{1}{x^2}}\right)}. \end{aligned}$$

As $x \to \infty$, the expressions under both radical signs approach 1, from which we conclude that

$$\lim_{n\to\infty} a_n = \lim_{x\to\infty} f(x) = 0.$$

□

Example The terms of the sequence

$$a_n = \frac{n^2}{e^n}, \quad n \geq 1,$$

are fractions in which both the numerator and denominator become infinite as $n \to \infty$. Because of the exponential, there is no "highest power of n" that we can divide both by in order to reveal the limit, but since the limit is the *indeterminate form* ∞/∞, we can use *l'Hospital's Rule* on the corresponding function $f(x) = x^2/e^x$. We have

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{n^2}{e^n} = \lim_{n\to\infty} \frac{2n}{e^n} = \lim_{n\to\infty} \frac{2}{e^n} = 0.$$

While we are actually applying l'Hospital's Rule to the function $f(x)$, we can carry out the steps on a_n instead, because of the equivalence of the limit of the sequence $a_n = f(n)$ and the limit of the function $f(x)$, as n and x tend to infinity. □

1.4.7 Alternating Sequences

An *alternating sequence* is a sequence in which the terms alternate signs. That is, for each n, a_{n+1} is the opposite sign of a_n. The presence of the alternating sign can make convergence analysis cumbersome, so it is desirable to be able to exclude it from consideration if possible.

Example The sequence

$$a_n = \frac{(-1)^n}{n+1}, \quad n \geq 0,$$

is an example of an alternating sequence. To determine its limit, we consider the related sequence

$$b_n = |a_n| = \frac{1}{n+1}, \quad n \geq 0,$$

since $(-1)^n = 1$ or -1 for any integer n, and therefore satisfies $|(-1)^n| = 1$. It is easy seen that

$$\lim_{n\to\infty} b_n = 0,$$

since the numerator is fixed at 1 and the denominator increases with n. Since $b_n \to 0$ as $n \to \infty$, and b_n is the *magnitudes* of a_n, it follows that $a_n \to 0$ as well. □

1.4.8 Growth Rates of Functions

When working with sequences, it is helpful to know the relative *growth rates* of functions such as polynomials or exponential functions, in order to quickly determine whether the terms of a sequence tend to zero, infinity, or a finite nonzero number.

Example The sequence

$$a_n = \frac{2^n}{n!}, \quad n \geq 0,$$

cannot be related to a function of x, because $n!$ is only defined when n is an integer. Instead, we will try to determine whether the terms are bounded by those of a simpler sequence, in which case we may be able to easily conclude that the limit is zero. We have, for $n \geq 1$,

$$a_n = \frac{2 \cdot 2 \cdot 2 \cdot \cdots \cdot 2}{1 \cdot 2 \cdot 3 \cdot \cdots \cdot n} = \left(\frac{2 \cdot 2 \cdot 2 \cdot \cdots \cdot 2}{1 \cdot 2 \cdot 3 \cdot \cdots \cdot (n-1)}\right)\frac{2}{n} = 2\left(\frac{2 \cdot 2 \cdot 2 \cdot \cdots \cdot 2}{3 \cdot 4 \cdot 5 \cdot \cdots \cdot (n-1)}\right)\frac{2}{n} < \frac{4}{n},$$

since the expression in parentheses must be less than 1. Because $1/n \to 0$ as $n \to \infty$, and multiplying by 4 does not change this, we conclude that $a_n \to 0$ as well. □

From the preceding example, we see that the factorial function $n!$ grows more rapidly than the exponential function 2^n, since otherwise the terms

of the sequence would not tend to zero as $n \to \infty$. The following list of categories of functions is ordered from most rapidly growing to least rapidly growing:

1. Factorial functions such as $n!$
2. Exponential functions such as e^n
3. Polynomial functions such as n^3
4. Logarithmic functions such as $\ln n$, where $n > 0$

Relationships such as these are often used in computer science, where they play a role in the analysis of running time or complexity of algorithms.

1.4.9 Geometric Sequences

The *geometric sequence* $\{r^n\}$, where r is any real number, is of particular interest because of its frequent appearance in infinite series that arise in a number of applications. It can exhibit three types of behavior, depending on the value of r.

- If $r = 1$, then $r^n = 1$ for any n, so the limit of $\{r^n\}$ is trivially equal to 1.
- If $|r| < 1$, then r^n decreases in magnitude as n increases, and therefore the limit is equal to 0.
- If $|r| > 1$, then r^n *increases* in magnitude, so that the terms become infinitely large. On the other hand, if $r = -1$, the terms oscillate endlessly between -1 and 1. In either case, the sequence is divergent.

1.4.10 Recursively Defined Sequences

When a sequence is defined using a formula that defines a_{n+1} in terms of a_n, it is possible to compute its limit L by keeping in mind that if the sequence $\{a_n\}$ converges, then $\{a_{n+1}\}$ converges as well, and has the same limit L. It follows that the formula that defines a_{n+1} in terms of a_n can be viewed as an equation that is satisfied by setting both a_n and a_{n+1} equal to L.

Example Consider the sequence

$$a_{n+1} = \frac{a_n}{2} + \frac{1}{a_n}, \quad a_1 = 1.$$

We assume that this sequence converges to a value L, and substitute it for a_n and a_{n+1} above, since both expressions converge to L if the sequence is convergent. We then have the equation

$$L = \frac{L}{2} + \frac{1}{L},$$

which is satisfied by either $\sqrt{2}$ or $-\sqrt{2}$. Since $a_1 = 1$, all terms of the sequence must be positive, so the limit is $\sqrt{2}$. It is interesting to note that the original sequence can be obtained by applying Newton's method to the function $f(x) = x^2 - 2$, in order to compute $\sqrt{2}$. □

1.4.11 Bounded and Monotonic Sequences

In some cases, even if it is not practical to compute the limit of a sequence, it is still helpful to know whether the sequence converges. For example, this kind of information is valuable when analyzing a method for solving an equation that computes a sequence of approximations, and it is only necessary to know whether this sequence is going to converge, since its convergence would imply successful solution of the equation. We now introduce some terminology in order to help us to classify sequences, which will then help us to quickly determine whether certain kinds of sequences converge.

- We say that a sequence is $\{a_n\}_{n=1}^{\infty}$ is *non-decreasing* if $a_{n+1} \geq a_n$ for all $n \geq 1$. Similarly, we say that $\{a_n\}_{n=1}^{\infty}$ is *non-increasing* if $a_{n+1} \leq a_n$ for all $n \geq 1$.
- We say that a sequence is $\{a_n\}_{n=1}^{\infty}$ is *increasing* if $a_{n+1} > a_n$ for all $n \geq 1$. Similarly, we say that $\{a_n\}_{n=1}^{\infty}$ is *decreasing* if $a_{n+1} < a_n$ for all $n \geq 1$.
- A sequence that is either non-increasing or non-decreasing is said to be *monotonic*. A sequence that is increasing or decreasing is also monotonic. Note that an increasing sequence is also non-decreasing, but not the other way around.
- A sequence $\{a_n\}_{n=1}^{\infty}$ is *bounded above* if there is a number M such that $a_n \leq M$ for all $n \geq 1$. On the other hand, if there is a number m such that $a_n \geq m$ for all $n \geq 1$, we say that the sequence is *bounded below*.
- A sequence that is both bounded above and bounded below is said to be *bounded*.

The reason why these terms are helpful is because of the *Monotonic Sequence Theorem*, which states that any sequence that is both bounded and monotonic is convergent. This is because any set of numbers that is bounded above must have a *least upper bound* L, also known as a *supremum*, and if the sequence is increasing, its terms have no choice but to continually increase toward L. After all, if they do not approach L, then L is not the least upper bound, while if they exceed L, then L is not an upper bound at all. Similar reasoning applies to the case of a decreasing sequence and its *greatest lower bound*, or *infinum*.

Because the notions of monotonicity and boundedness are useful in determining whether a sequence converges, it is also useful to be able to determine whether a sequence is monotonic.

Example The sequence defined by $a_n = 1$, for $n \geq 1$, is an example of a monotonic sequence, because it is non-increasing *and* non-decreasing. It is also bounded above and below, so it is convergent, trivially so, to the limit 1. □

Example The sequence defined by $a_n = 1/2^n$, for $n \geq 0$, is a decreasing, and thus monotonic, sequence. It is bounded above, by 1, and below, by 0. Therefore it converges, and its limit is 0. □

Example The sequence defined by $a_n = \log n$, for $n \geq 1$, is a monotonic sequence, because it is increasing. It is bounded below by 0, but it is not bounded above, and is divergent. The same properties hold for the sequence $a_n = 2^n$. □

Example The sequence defined by $a_n = (-1)^n$, for $n \geq 0$, is not monotonic, because its terms oscillate continually between -1 and 1. It is bounded below, by -1, and above, by 1, but it is divergent. Similar properties hold for the sequence defined by $a_n = \sin n$. □

Example Consider the sequence defined by

$$a_n = 1 - 2^{-n}, \quad n \geq 0.$$

To determine whether this sequence is increasing or decreasing, we examine the difference between two terms and try to ascertain whether this difference is always positive or always negative. We have

$$a_{n+1} - a_n = (1 - 2^{-(n+1)}) - (1 - 2^{-n}) = 2^{-n} - 2^{-(n+1)} = 2^{-n}(1 - 2^{-1}) = \frac{2^{-n}}{2} = 2^{-(n+1)}.$$

This difference is always positive, so we conclude that $\{a_n\}$ is an increasing sequence. □

When the terms of a sequence are fractions, it is helpful to cross-multiply to determine monotonicity.

Example Consider the sequence defined by

$$a_n = \frac{n+1}{n+2}, \quad n \geq 1.$$

We will show that this sequence is increasing. To accomplish this, we must show that $a_{n+1} > a_n$, which is equivalent to

$$\frac{(n+1)+1}{(n+1)+2} > \frac{n+1}{n+2},$$

or

$$\frac{n+2}{n+3} > \frac{n+1}{n+2}.$$

Cross-multiplying, we obtain

$$(n+2)^2 > (n+1)(n+3),$$

which, upon expansion, yields

$$n^2 + 4n + 4 > n^2 + 4n + 3,$$

which is true for all n. Therefore, the sequence is in fact increasing. □

Example The sequence

$$a_n = (-1)^n, \quad n \geq 0,$$

has terms that alternate between 1 and -1. That is, $a_0 = 1, a_1 = -1, a_2 = 1, a_3 = -1$, and so on. Now, suppose we want to construct a formula for a sequence whose terms alternate signs in pairs. That is,

$$a_0 = a_1 = 1, \quad a_2 = a_3 = -1, \quad a_4 = a_5 = 1,$$

and so on. We present two approaches to this.

The first approach makes use of the *floor function* $\lfloor x \rfloor$. For any real number x, $\lfloor x \rfloor$ is defined to be the greatest integer that is less than or equal to x. For example,

$$\lfloor 1 \rfloor = 1, \quad \lfloor \pi \rfloor = 3, \quad \lfloor 14.9 \rfloor = 14.$$

The floor function is also known as the *greatest integer function*, and is an example of a *step function*, since its graph, shown in Figure 1.2, consists of line segments that are arranged like steps.

To define our sequence, we use the floor function to obtain the sequence of exponents for -1,

$$b_0 = 0, \quad b_1 = 0, \quad b_2 = 1, \quad b_3 = 1, \quad b_4 = 2, \quad b_5 = 2,$$

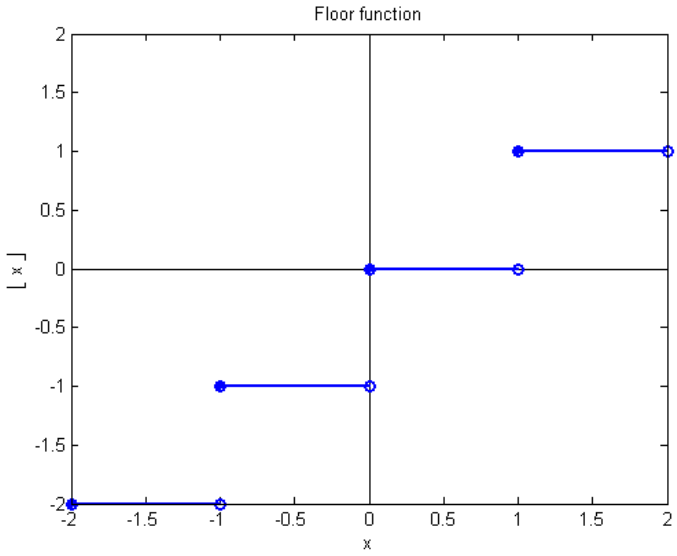

Figure 1.2: The graph of the floor function $\lfloor x \rfloor$, for $-2 \leq x \leq 2$.

and so on. This is accomplished by setting $b_n = \lfloor x/2 \rfloor$, which, for odd x, results in a fraction, equal to 0.5, which is eliminated by taking the floor, or *rounding down*. By using b_n as the exponent to -1, we obtain the sequence

$$a_n = (-1)^{b_n} = (-1)^{\lfloor x/2 \rfloor},$$

which has the desired terms that alternate between 1 and -1 in pairs. In general, to obtain a sequence that alternates between 1 and -1 every n terms, we can use the sequence with terms $(-1)^{\lfloor x/n \rfloor}$.

An alternative approach uses trigonometric functions. Figure 1.3 shows the graph of $\sin x$ for $0 \leq x \leq 4\pi$. The circles on the graph indicate the points corresponding to

$$x = \frac{\pi}{4} + \frac{k\pi}{2}, \quad k = 0, 1, \ldots, 7.$$

Note that the values of $\sin x$ at these points alternates between $\pm\sqrt{2}/2$ in pairs. Therefore, we can define the sequence a_n by

$$a_n = \sqrt{2} \sin\left(\frac{\pi}{4} + \frac{n\pi}{2}\right).$$

□

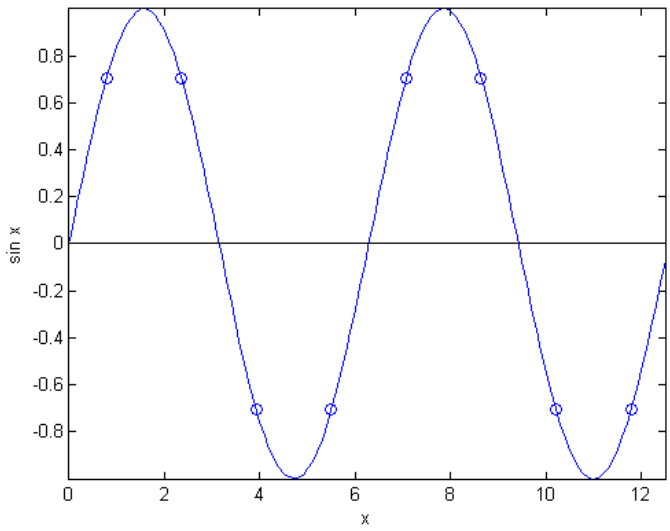

Figure 1.3: Graph of $\sin x$ for $0 \leq x \leq 4\pi$.

1.4.12 Summary

- Sequences are ordered lists of numbers, typically indexed by integers. Each number in the sequence is called a term.

- Sequences can be defined by a formula that specifies each term as a function of the index. In some cases, though, it is more convenient to use a recurrence relation that defines each term in terms of one or more previous terms. Any definition should indicate the index of the first term, which is usually 1.

- If the formula for the terms of a sequence is not known, it can sometimes be inferred from the first few terms of the sequence by examination of how these terms change as the index increases. Terms that alternate in sign suggest the use of $(-1)^n$, where n is the index. If the terms are fractions, it is wise to examine the numerator and denominator separately.

- A sequence converges to a value, called its limit, if all of the terms beyond some index can be made arbitrarily close to the limit.

- A sequence that does not converge to any value is divergent. A divergent sequence may have terms that tend to infinity as the index does.

- The limit of a convergent sequence can be computed by using techniques for computing limits of functions at infinity, applied to the function obtained from the formula for each term of the sequence. These techniques include:

  - the limit laws, such as the law that the limit of a sum is the sum of the limits
  - the Squeeze Theorem
  - the technique of dividing the numerator and denominator by the highest power of x, if both are polynomials of x

- Any techniques that can be used to compute the limit of a function $f(x)$ as $x \to \infty$ can also be applied to compute limits of sequences, when each term a_n of the sequence can be defined in terms of the value of a function $f(n)$ at the term's index n. These techniques include, among others, l'Hospital's Rule, and algebraic techniques such as multiplying and dividing by the conjugate or the highest power of n.

- The behavior of an alternating sequence can sometimes be studied more easily by examining the absolute value of its terms, thus filtering out the alternation and isolating the terms' magnitude.

- When the terms of a sequence are defined to be a fraction, it is helpful to consider the relative growth rates of functions to determine whether the terms converge to zero or tend to infinity. As a rule, exponential functions grow more rapidly than polynomials, which grow more rapidly than logarithmic functions.

- A particularly useful sequence is the geometric sequence r^n, for a real number r. This sequence converges to 1 if $r = 1$, to 0 if $|r| < 1$, and diverges otherwise.

- If a sequence $\{a_n\}$ is defined recursively, with a_{n+1} defined in terms of a_n, the limit can be computed by setting both equal to an unknown value L and solving for it. If there is more than one solution, the initial term can be used to determine which is the limit.

- A sequence $\{a_n\}$ is increasing if its terms increase as n increases, and decreasing if its terms decrease. A monotonic sequence is either increasing or decreasing. A sequence can be shown to be monotonic by comparing consecutive terms directly, using algebraic techniques such as cross-multiplying.
- A sequence is bounded above if its terms never exceed a given number M, and bounded below if its terms are never exceeded by a given number m. A sequence that is bounded above and below is called bounded.
- According to the Monotonic Sequence Theorem, a bounded monotonic sequence is convergent.

1.5 Series

1.5.1 What is a Series?

An *infinte series*, usually referred to simply as a *series*, is an sum of all of the terms of an infinite sequence. Specifically, let $\{a_n\}_{n=1}^{\infty}$ be a sequence. Then we can define a series to be the sum of the terms of $\{a_n\}$,

$$a_1 + a_2 + a_3 + \cdots + a_n + \cdots.$$

We refer to the terms of $\{a_n\}$ as the *terms* of the series.

Writing a series in this manner is cumbersome, so we instead use *sigma notation* to write this series as

$$\sum_{n=1}^{\infty} a_n.$$

The expression below the upper case Greek letter sigma indicates what name is given to the index (in this case, n), and its initial value (in this case, 1). The expression above the sigma indicates the final value of the index, or ∞ for an infinite sum.

Note that this means sigma notation can be used to represent finite sums as well.In other words, if we wrote, for example, "10" above the sigma instead of ∞, then we would be specifying that only the first 10 terms of $\{a_n\}$ should be added. That is,

$$\sum_{n=1}^{10} a_n = a_1 + a_2 + a_3 + \cdots + a_{10}.$$

For either a finite or infinite sum, the expression to the right of the sigma is what is to be added, for each value of the index. This means that if n is the name assigned to the index, then *every* occurrence of n within the summed expression is to be replaced with each value of the index, and the resulting terms are summed.

Example The finite sum

$$\sum_{n=0}^{4} \frac{2^n}{n+1}$$

is evaluated as follows:

$$\begin{aligned}
\sum_{n=0}^{4} \frac{2^n}{n+1} &= \frac{2^0}{0+1} + \frac{2^1}{1+1} + \frac{2^2}{2+1} + \frac{2^3}{3+1} + \frac{2^4}{4+1} \\
&= 1 + 1 + \frac{4}{3} + 2 + \frac{16}{5} \\
&= \frac{128}{15}.
\end{aligned}$$

□

We now need to define what it means to compute the sum of infinitely many terms. For this concept, sequences play their most important role. Given an infinite sequence $\{a_n\}_{n=1}^{\infty}$ of terms to be summed, we define a sequence of *partial sums*, denoted by $\{s_n\}_{n=1}^{\infty}$, as follows:

$$\begin{aligned}
s_1 &= a_1 \\
s_2 &= s_1 + a_2 \\
&= a_1 + a_2 \\
s_3 &= s_2 + a_3 \\
&= a_1 + a_2 + a_3 \\
&\vdots \\
s_n &= s_{n-1} + a_n \\
&= a_1 + a_2 + \cdots + a_{n-1} + a_n.
\end{aligned}$$

Then, we can view the series as a sequence of partial sums. This leads to the following definitions.

A series

$$\sum_{n=1}^{\infty} a_n$$

is said to be *convergent* if its sequence of partial sums, $\{s_n\}_{n=1}^{\infty}$, is convergent. The limit of $\{s_n\}$, if it exists, is called the *sum* of the series. If the sequence of partial sums is divergent, then we say that the series is *divergent*.

Example Consider the series

$$\sum_{n=0}^{\infty} \frac{1}{2^n}.$$

The sequence of partial sums is

$$\begin{aligned}
s_0 &= 1 \\
s_1 &= s_0 + \frac{1}{2} \\
&= \frac{3}{2} \\
s_2 &= s_1 + \frac{1}{4} \\
&= \frac{7}{4} \\
s_3 &= s_2 + \frac{1}{8} \\
&= \frac{15}{8} \\
&\vdots \\
s_n &= \frac{2^{n+1} - 1}{2^n} \\
&= \frac{2^n(2 - 2^{-n})}{2^n} \\
&= 2 - \frac{1}{2^n}.
\end{aligned}$$

The partial sums converge to 2, so we say that 2 is the sum of the series. □

1.5.2 Why Do We Need Series?

Series are applied throughout mathematics, as well as physics, computer science, and various branches of engineering. They are particularly useful for describing functions or solutions of equations using sums of "simple" functions such as polynomials or basic trigonometric functions. They are also useful for analyzing the performance of numerical methods for solving equations.

1.5.3 Geometric Series

The series in the preceding example is a *geometric series*. The general form of a geometric series is

$$\sum_{n=0}^{\infty} ar^n,$$

where r is called the *common ratio* of the series, because each term in the series, for $n \geq 1$, is obtained by multiplying the previous term by r. This type of series arises in a variety of applications, such as the analysis of numerical methods for solving linear or differential equations.

We now try to determine whether a geometric series converges, and if so, compute its limit. To do this, we examine the sequence of partial sums. We have

$$\begin{aligned} s_{n+1} &= a + ar + ar^2 + \cdots + ar^n + ar^{n+1}, \\ rs_n &= ar + ar^2 + \cdots + ar^n + ar^{n+1}, \end{aligned}$$

which yields the relation $s_{n+1} = a + rs_n$. We also have $s_{n+1} = s_n + ar^n$, by the definition of a partial sum. Equating these, and rearranging, yields

$$a(1 - r^n) = s_n(1 - r),$$

which, for $r \neq 1$, leads to a *closed-form* representation of the nth partial sum,

$$s_n = a\frac{1 - r^n}{1 - r}.$$

We can now determine convergence of the geometric series:

- If $r = 1$, the nth partial sum is $s_n = an$, and therefore the series diverges.
- If $r = -1$, the numerator in s_n oscillates between 0 and $2a$, so the series again diverges.
- If $|r| > 1$, then r^n diverges, so due to its presence in the numerator of s_n, the series diverges.
- Finally, if $|r| < 1$, then $r^n \to 0$, and the series converges to

$$\lim_{n\to\infty} s_n = \frac{a}{1 - r}.$$

We now consider several examples of geometric series.

Example The series

$$\sum_{n=1}^{\infty} \frac{e^n}{10^{n-1}}$$

can be viewed as a geometric series, but we must be careful, because a geometric series uses an initial index of zero, while the initial index for this series is 1. Therefore, we must first rewrite the series to use an initial index of zero before determining the values of a and r.

Since we wish to subtract 1 from the initial index, we must compensate by replacing n by $n + 1$ throughout the expression to be summed. This yields the equivalent series

$$\sum_{n=0}^{\infty} \frac{e^{n+1}}{10^n} = e \sum_{n=0}^{\infty} \frac{e^n}{10^n}.$$

This is a geometric series with $a = e$ and $r = e/10$. Since $e \approx 2.718281828$, we have $|r| < 1$, and therefore the series converges to

$$\frac{a}{1-r} = \frac{e}{1-\frac{e}{10}} = \frac{10e}{10-e}.$$

□

Example Consider the series

$$\sum_{n=0}^{\infty} 2^{-2n} 3^n.$$

Using the laws of exponents, we rewrite this series as

$$\sum_{n=0}^{\infty} (2^2)^{-n} 3^n = \sum_{n=0}^{\infty} \frac{3^n}{4^n}.$$

It follows that this is a geometric series with $a = 1$ and $r = 3/4$, which converges to

$$\frac{a}{1-r} = \frac{1}{1-\frac{3}{4}} = 4.$$

□

Example The series

$$\sum_{n=0}^{\infty} \frac{(x-2)^n}{2^n}$$

is an example of a *power series*, since the terms are constants times powers of $(x-2)$. We will see much more of power series, as they are very useful for approximating functions in a way that is practical for implementation on a calculator or computer. This particular power series is also a geometic series with $a = 1$ and $r = (x-2)/2$. Therefore, it converges if $|(x-2)/2| < 1$, which is true if $-2 < x - 2 < 2$, or $0 < x < 4$. □

Example Geometric series can be used to convert repeating decimals into fractions. Consider the repeating decimal $0.\overline{142857}$. This can be written as the infinite series

$$\frac{142857}{10^6} + \frac{142857}{10^{12}} + \frac{142857}{10^{18}} + \cdots + \frac{142857}{10^{6n}},$$

which is a geometric series with $a = 142857/10^6$ and $r = 10^{-6}$, where the 6 arises due to the fact that the sequence of repeating digits has 6 terms: 1, 4, 2, 8, 5 and 7. This is a convergent geometric series, and its limit is

$$\frac{a}{1-r} = \frac{142857}{10^6(1-10^{-6})} = \frac{142857}{10^6-1} = \frac{142857}{999999} = \frac{1}{7}.$$

□

Example Consider the geometric series

$$\sum_{n=0}^{\infty} \frac{1}{2^n},$$

for which $a = 1$ and $r = \frac{1}{2}$. The sum of the first 10 terms is given by

$$s_9 = 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^9} = \frac{1 - \frac{1}{2^{10}}}{1 - \frac{1}{2}} = \frac{1023}{512} \approx 2.$$

Because $|r| < 1$, this series converges, and to the sum $\frac{1}{1-r} = 2$. On the other hand, changing r to 2 yields the 10th partial sum $\frac{1-2^{10}}{1-2} = 1023$. This series is divergent. □

Earlier in this section, we defined the concept of an infinite series, and what it means for a series to converge to a finite sum, or to diverge. We also worked with one particular type of series, a geometric series, for which it is particularly easy to determine whether it converges, and to compute its limit when it does exist. Now, we consider other types of series and investigate their behavior.

1.5.4 Telescoping Series

Consider the series

$$\sum_{n=1}^{\infty} \frac{1}{n} - \frac{1}{n+1}.$$

If we write out the first few terms, we obtain

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{1}{n} - \frac{1}{n+1} &= \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{4} - \frac{1}{5}\right) + \cdots \\ &= 1 + \left(\frac{1}{2} - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{3}\right) + \left(\frac{1}{4} - \frac{1}{4}\right) + \cdots. \end{aligned}$$

We see that nearly all of the fractions cancel one another, which reveals the partial sum

$$s_n = \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \cdots + \left(\frac{1}{n} - \frac{1}{n+1}\right) = 1 - \frac{1}{n+1}.$$

Because this sequence of partial sums converges, the series converges, to 1. This is an example of a *telescoping series*. It turns out that many series have this property, even though it is not immediately obvious.

Example The series

$$\sum_{n=1}^{\infty} \frac{1}{n(n+2)}$$

is also a telescoping series. To see this, we compute the *partial fraction decomposition* of each term. This decomposition has the form

$$\frac{1}{n(n+2)} = \frac{A}{n} + \frac{B}{n+2}.$$

To compute A and B, we multipy both sides by the common denominator $n(n+2)$ and obtain $1 = A(n+2) + Bn$. Substituting $n = 0$ yields $A = 1/2$, and substituting $n = -2$ yields $B = -1/2$. The series is now

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{1}{n(n+2)} &= \frac{1}{2}\left(\sum_{n=1}^{\infty} \frac{1}{n} - \frac{1}{n+2}\right) \\ &= \frac{1}{2}\left[\left(1 - \frac{1}{3}\right) + \left(\frac{1}{2} - \frac{1}{4}\right) + \left(\frac{1}{3} - \frac{1}{5}\right) + \left(\frac{1}{4} - \frac{1}{6}\right) + \cdots\right] \\ &= \frac{1}{2}\left[1 + \frac{1}{2} - \left(\frac{1}{3} - \frac{1}{3}\right) - \left(\frac{1}{4} - \frac{1}{4}\right) - \frac{1}{5} - \frac{1}{6} \cdots\right]. \end{aligned}$$

It can be seen from the first four terms above that the nth partial sum is

$$s_n = \frac{1}{2}\left[1 + \frac{1}{2} - \frac{1}{n+1} - \frac{1}{n+2}\right],$$

which converges to the limit 3/4. □

Not all telescoping series converge. It is essential to examine the sequence of partial sums.

1.5.5 Harmonic Series

One series of interest is actually a divergent one, the *harmonic series*

$$\sum_{n=1}^{\infty} \frac{1}{n}.$$

This series is the best-known example of a series that diverges even though the sequence of its terms converges to zero.

To see that it diverges, we can use the fact that the terms of this series are at least as large as those of the series

$$\sum_{n=1}^{\infty} \frac{1}{2^{\lceil \log_2 n \rceil}},$$

where, for each n, the nth term is 1 divided by the smallest power of 2 that is greater than or equal to n. It can be shown that this series diverges by examining the sequence of partial sums directly, and since $\sum 1/n$ has terms that are at least as large, it must diverge as well.

Although this series diverges, its terms are quite close to that of a series that converges. In fact, the series

$$\sum_{n=1}^{\infty} \frac{1}{n^{1+\epsilon}},$$

for any $\epsilon > 0$, is convergent.

1.5.6 Basic Convergence Tests

Because summing a series requires adding infinitely many numbers, it makes sense, intuitively, that these numbers must get smaller as the index $n \to \infty$, if there is to be any hope that the sum will converge to a finite number.

This is in fact the case: if a series converges, the sequence of its terms must converge to zero.

However, the *converse* is not true: if a series has terms that converge to zero, it does not necessarily converge. The harmonic series, above, is an example of a divergent series whose terms converge to zero. Instead, we can use the *contrapositive* statement to arrive at a condition for *divergence*, rather than convergence: if the terms of a series do *not* converge to zero, then it diverges.

We will learn about several tests that can be used to prove that a series converges, but for now, we note that certain simple combinations or modifications of convergent series are also convergent. Specifically, if

$$\sum_{n=1}^{\infty} a_n \quad \text{and} \quad \sum_{n=1}^{\infty} b_n$$

are convergent series, with limits S_a and S_b, respectively, then the series

$$\sum_{n=1}^{\infty} a_n + b_n, \quad \sum_{n=1}^{\infty} a_n - b_n, \quad \sum_{n=1}^{\infty} ca_n,$$

where c is a constant, are also convergent, with limits

$$\sum_{n=1}^{\infty} a_n + b_n = S_a + S_b, \quad \sum_{n=1}^{\infty} a_n - b_n = S_a - S_b, \quad \sum_{n=1}^{\infty} ca_n = cS_a.$$

Example Using the result of previous examples, we have

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{1}{2^n} + \frac{1}{n(n+2)} &= \sum_{n=1}^{\infty} \frac{1}{2^n} + \sum_{n=1}^{\infty} \frac{1}{n(n+2)} \\
&= \sum_{n=0}^{\infty} \frac{1}{2^{n+1}} + \frac{1}{2} \sum_{n=1}^{\infty} \frac{1}{n} - \frac{1}{n+2} \\
&= \frac{1}{2} \frac{1}{1-\frac{1}{2}} + \frac{3}{4} \\
&= \frac{7}{4}.
\end{aligned}$$

1.5.7 Summary

- An infinite series, or simply series, is the sum of the terms of a sequence. Sigma notation provides a concise way of describing a series,

using only its initial index, final index (or ∞), and definition of each term.

- The partial sum of a series is the sum of its first n terms, for any value of the index n. A series converges if the sequence of its partial sums converges; otherwise, it diverges.
- A geometric series is any series whose terms are of the form ar^n, for $n \geq 0$. The number r is called the common ratio. If $|r| < 1$, the series converges to $\frac{a}{1-r}$; otherwise, it diverges.
- A telescoping series is a series in which all but a finite number of terms cancel. When a series has terms that are rational functions, a partial fraction decomposition can be used to determine whether the series is in fact a telescoping series.
- The harmonic series, with terms $1/n$, is an example of a series whose terms converge to zero, but is still divergent.
- If a series converges, then its terms must converge to zero, but the converse is not necessarily true: a series whose terms converge to zero may still diverge. On the other hand, if the terms of a series do not converge to zero, then the series must diverge.
- Adding or subtracting convergent series yields a convergent series, whose sum is obtained by adding or subtracting the sums of the individual series. Similarly, multiplying the terms of a convergent series by a constant multiplies its sum by the same constant.

1.6 Convergence Tests

1.6.1 The Integral Test

Previously, we have defined the sum of a convergent infinite series

$$\sum_{n=1}^{\infty} a_n$$

to be the limit of the sequence of partial sums $\{s_n\}_{n=1}^{\infty}$, defined by

$$s_k = a_1 + a_2 + \cdots + a_k = \sum_{n=1}^{k} a_n.$$

In other words, if S is the sum of the series, then

$$S = \lim_{k\to\infty} s_k = \lim_{k\to\infty} \sum_{n=1}^{k} a_n.$$

Figure 1.4 illustrates the partial sum s_8 for the series

$$\sum_{n=1}^{\infty} n^{-3/2}. \tag{1.1}$$

It can be seen from the figure that this partial sum can be viewed as the sum of the areas of rectangles, each with width 1 and height $n^{-3/2}$, for $n = 1, 2, \ldots, 8$. If we exclude the leftmost rectangle, then the sum of the

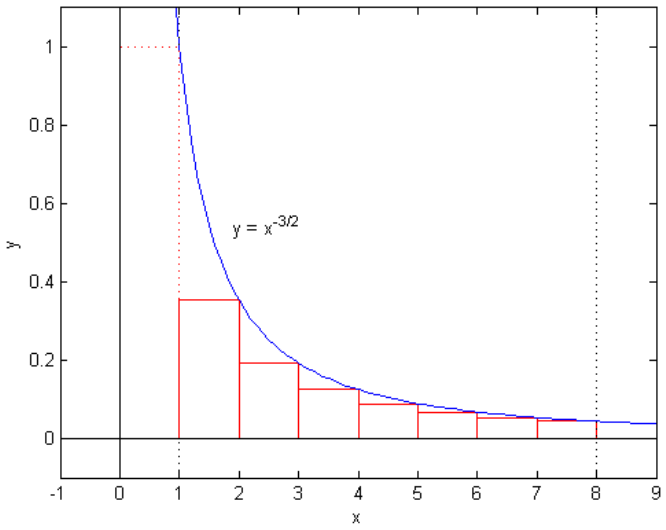

Figure 1.4: Partial sum s_8 of the series (1.1). The leftmost rectangle, shown with dotted edges, is excluded from the approximation of the area given by the integral (1.2). All other rectangles, shown with solid edges, represent the partial sum $\tilde{s}_7$ of the modified series (1.4).

areas of the remaining rectangles can be viewed as an approximation of the area of the region bounded by the curves $y = x^{-3/2}$, $y = 0$, $x = 1$, and $x = 8$.

The exact value of this area is given by the integral

$$\int_1^8 x^{-3/2}\,dx, \tag{1.2}$$

which is defined to be the limit of a sequence of Riemann sums $\{R_m\}_{m=1}^{\infty}$. Each Riemann sum approximates the area of this region by the sum of the areas of m rectangles, each of width $\Delta x = (8-1)/m$. Specifically, we have

$$R_m = \sum_{n=1}^{m} n^{-3/2}\frac{7}{m}. \tag{1.3}$$

From the definitions of the original series (1.1) and the Riemann sum (1.3), we can see that the partial sum shown in Figure 1.4, s_8, is equal to $R_7 + 1$, since 1 is the area of the leftmost, excluded rectangle. Furthermore, the partial sum $\tilde{s}_8$ of the modified series

$$\sum_{n=2}^{\infty} n^{-3/2}, \tag{1.4}$$

defined by

$$\tilde{s}_8 = 2^{-3/2} + \cdots + 8^{-3/2} = \sum_{n=2}^{8} n^{-3/2},$$

is a *lower bound* for the exact area given by the integral.

This last point suggests that if we extend the interval of integration to $[1, \infty)$, and find that the *improper integral*

$$\int_1^{\infty} x^{-3/2}\,dx = \lim_{k\to\infty}\int_1^k x^{-3/2}\,dx$$

exists and is finite, then the modified series (1.4) must converge, because each partial sum of the modified series is bounded above by the "partial integral" on the interval $[1, k]$. If this sequence of integrals converges, it follows that the sequence of modified partial sums must converge, and therefore the modified series must be convergent. Because the original series (1.1) and modified series (1.4) only differ by the inclusion of the first term $1^{-3/2} = 1$, we conclude that the original series must be convergent as well.

Evaluating the integral (1.2), we obtain

$$\int_1^{\infty} x^{-3/2}\,dx \quad = \quad \lim_{k\to\infty}\int_1^k x^{-3/2}\,dx$$

$$\begin{aligned} &= \lim_{k\to\infty} -2x^{-1/2}\Big|_1^k \\ &= \lim_{k\to\infty} 2 - 2k^{-1/2} \\ &= 2. \end{aligned}$$

It follows that the sum of the modified series (1.4) is less than 2, and therefore the sum of the original series (1.1) must be less than 3. Unfortunately, it is not possible to analytically compute the exact value of the sum, although it can be approximated numerically. Nonetheless, we at least know that in some circumstances, we can use integrals to determine whether a series converges, as this example illustrates.

Integrals can also be used to determine that a series is divergent. Suppose that the integral

$$\int_1^\infty f(x)\,dx$$

is divergent, meaning that it is not equal to any finite value. By examining Figure 1.5, we can see that just as the integral from 1 to k of $f(x)$ is an *upper* bound on the partial sum $\tilde{s}_k$, that *excludes* the first term, this integral is also a *lower* bound on the partial sum s_k, that *includes* the first term. Therefore, if the sequence of integrals from 1 to k diverges as $k \to \infty$, it follows that the sequence of partial sums $\{s_k\}_{k=1}^\infty$, consisting of values that are greater than these integrals, must diverge as well.

We summarize this discussion by stating the *Integral Test*: Let the sequence $\{a_n\}_{n=1}^\infty$ be defined by $a_n = f(n)$, for $n \geq 1$, where $f(x)$ is a continuous, positive, *decreasing* function defined on $[1, \infty)$. If the improper integral

$$\int_1^\infty f(x)\,dx = \lim_{k\to\infty} \int_1^k f(x)\,dx$$

converges to a finite value F, then the series

$$\sum_{n=1}^\infty a_n$$

is convergent. Furthermore, if S is the sum of this series, then $S < F + a_1$. On the other hand, if the integral diverges, then the series diverges.

The reason why it is important that $f(x)$ is positive and decreasing is that it must be ensured that the partial sums of the modified series (with the first term excluded) are a *lower bound* for the value of the integral, because otherwise, no conclusive statement can be made about the convergence of

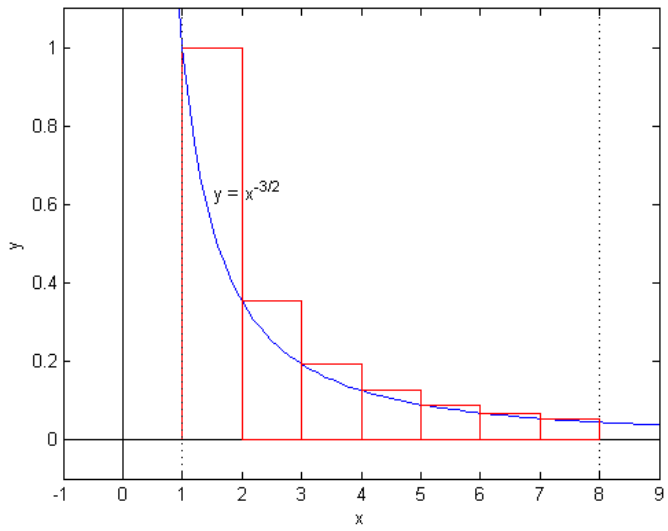

Figure 1.5: Partial sum s_7 of the series (1.1).

the series based on the existence of the integral. It should be mentioned that it is not absolutely necessary for $f(x)$ to be positive and decreasing on the *entire* interval $[1, \infty)$. It is sufficient that there exist some $c \geq 1$ such that $f(x)$ is positive decreasing for all $x > c$. This is because the convergence of a series is not affected by the behavior of a finite number of terms.

Example Consider the *p-series*

$$\sum_{n=1}^{\infty} \frac{1}{n^p},$$

where $p > 0$. The function $f(x) = 1/x^p$, for $x \geq 1$, is a continuous, positive, decreasing function, so we can apply the Integral Test to determine whether this series converges. We have

$$\int_1^{\infty} \frac{1}{x^p}\, dx = \lim_{k\to\infty} \frac{1}{1-p} x^{1-p}\Big|_1^k = \lim_{k\to\infty} \frac{1}{1-p}(k^{1-p} - 1).$$

In order for the integral to converge, we must have $p > 1$. We conclude that the p-series converges if $p > 1$, and diverges if $p \leq 1$. □

Example Consider the series

$$\sum_{n=0}^{\infty} \frac{1}{1+n^2}.$$

We attempt to determine whether this series converges using the Integral Test, by evaluating the integral

$$\int_0^{\infty} \frac{1}{1+x^2}\,dx.$$

Note that while the Integral Test, as stated previously, uses 1 for both the initial index and the lower limit of integration, it can also be applied with different values for these quantities, as long as they are consistent. We have

$$\int_0^{\infty} \frac{1}{1+x^2}\,dx = \lim_{k\to\infty} \tan^{-1} x\Big|_0^k = \lim_{k\to\infty} (\tan^{-1} k - \tan^{-1} 0) = \frac{\pi}{2}.$$

It follows from the convergence of this integral that the series is convergent. □

1.6.2 The Comparison Test

The integral test indicates convergence or divergence of a series through comparison with the integral of a related function, but for many series, this is not feasible because the integral cannot be easily evaluated. A practical alternative is to compare the series to another series that is known to converge or diverge, since there are already certain types of series whose behavior is known.

Suppose we have two series

$$\sum_{n=1}^{\infty} a_n \quad \text{and} \quad \sum_{n=1}^{\infty} b_n,$$

whose terms are non-negative, that satisfy $0 \leq a_n \leq b_n$ for all $n \geq 1$. We denote the partial sums of these series by p_k and q_k, respectively:

$$p_k = \sum_{n=1}^{k} a_n, \quad q_k = \sum_{n=1}^{k} b_n.$$

Suppose that the series $\sum b_n$ is convergent to a sum S_b. Because of the relationship between a_n and b_n, and the fact that both sets of terms are

non-negative, we must have $0 \leq p_k \leq q_k$ for $k \geq 1$. Furthermore, the sequences $\{p_k\}$ and $\{q_k\}$ are both increasing.

A convergent sequence is bounded, so $\{q_k\}$, which converges to S_b, is bounded. In fact, because it is also an increasing sequence, it must be bounded above by S_b. Since $p_k \leq q_k$, for all $k \geq 1$, p_k cannot exceed S_b either. Therefore, $\{p_k\}$ is also a bounded increasing sequence, so by the Monotonic Convergence Theorem, it is convergent. We conclude that $\sum a_n$ converges.

Now, suppose that we have the opposite relationship, $a_n \geq b_n$, and that $\sum b_n$ *diverges*. Following similar reasoning as before, we have $p_k \geq q_k$, and both sequences are still increasing. If $\{p_k\}$ converges, it must be bounded, which means $\{q_k\}$ is also bounded. But that would imply that $\{q_k\}$, and therefore $\sum b_n$, converge, which is a contradiction. Therefore, $\{p_k\}$ diverges, from which we conclude that $\sum a_n$ diverges.

We summarize this discussion by stating the *Comparison Test*:

- if $0 \leq a_n \leq b_n$ for $n \geq 1$, and $\sum b_n$ converges, then $\sum a_n$ also converges.
- If $0 \leq b_n \leq a_n$ for $n \geq 1$, and $\sum b_n$ diverges, then $\sum a_n$ also diverges.

The Comparison Test is illustrated in Figure 1.6(a).

Example Consider the series

$$\sum_{n=0}^{\infty} \frac{1}{2^n}, \quad \sum_{n=0}^{\infty} \frac{1}{2^n + 1}.$$

The first series is known to converge, as it is a geometric series with $a = 1$ and $r = 1/2$. The second series, due to the addition of 1 in the denominator, has terms that are less than the corresponding terms of the first series. It follows that this series is also convergent. □

There are two scenarios for which the Comparison Test is useless: $a_n \leq b_n$ and $\sum b_n$ diverges, and $b_n \leq a_n$ and $\sum b_n$ converges. In both cases, we cannot make a definitive statement about whether $\sum a_n$ converges or diverges. However, there are pairs of series that fall into these categories whose terms behave similarly, so it would be nice to be able to determine the convergence or divergence of the series with "simpler" terms, and then use the result to draw a conclusion about the behavior of the other series.

We have previously stated that if $\sum a_n$ converges, then $\sum ca_n$ also converges, where c is any constant. Similarly, if $\sum a_n$ diverges, then $\sum ca_n$ also

diverges, if $c \neq 0$. However, we have also stated that the behavior of a finite number of terms does not affect the convergence or divergence of a series. These facts suggest that if a series $\sum b_n$ has terms that are not exactly multiples of the terms of another series $\sum a_n$, but approach being multiples in the limit as $n \to \infty$, the two series should have the same behavior.

This is in fact the case, which leads to the *Limit Comparison Test*: suppose that two series $\sum a_n$ and $\sum b_n$ have terms that satisfy

$$\lim_{n\to\infty} \frac{a_n}{b_n} = c,$$

where c is a nonzero, finite number. Then, $\sum a_n$ and $\sum b_n$ both converge, or both diverge. This test is illustrated in Figure 1.6(b), in which the divergence of the series $\sum b_n$, with the larger terms, can be used to establish the divergence of the series $\sum a_n$, which is not possible using the Comparison Test.

Example Consider the series

$$\sum_{n=1}^{\infty} \frac{n^2+1}{n^3+3n^2+2n}, \quad \sum_{n=1}^{\infty} \frac{1}{n}.$$

The second series is divergent, but we cannot use the Comparison Test directly on these two series, because $n^3 + 3n^2 + 2n > n(n^2 + 1)$ for $n \geq 1$, which means that the terms of the first series are *less* than those of the second. We could try scaling the second series by a constant $c > 1$ and then using the Comparison Test, but since the Comparison Test requires that a relationship apply to *all* terms of both series, finding an appropriate c can be tedious. It is easier to apply the Limit Comparison Test, which yields

$$\lim_{n\to\infty} \frac{\frac{1}{n}}{\frac{n^2+1}{n^3+3n^2+2n}} = \frac{n^3+3n^2+2n}{n(n^2+1)} = \frac{n^3}{n^3}\frac{1+\frac{3}{n}+\frac{2}{n^2}}{1+\frac{1}{n^2}} = 1,$$

from which we can conclude that the first series also diverges. □

1.7 Other Convergence Tests

In this section, we develop additional tests that, for many series, will enable us to quickly determine whether a given series converges or diverges. Although these new tests, like the Integral and Comparison Tests, can only tell us whether a series converges, as opposed to helping us compute its limit, they do offer us one advantage that the previous tests do not: they are applicable to series whose terms are not necessarily positive.

1.7.1 The Alternating Series Test

An *alternating series* is a series whose terms alternate signs, so that two consecutive terms always have opposite signs.

Example The series

$$\sum_{n=0}^{\infty} \left(-\frac{1}{2}\right)^n,$$

in addition to being a geometric series with $a = 1$ and $r = -\frac{1}{2}$, is an alternating series whose first few terms are

$$a_0 = 1, \quad a_1 = -\frac{1}{2}, \quad a_2 = \frac{1}{4}, \quad a_3 = -\frac{1}{8}.$$

□

Any alternating series has terms of the form $a_n = (-1)^n b_n$, where $b_n = |a_n| > 0$. In the preceding example, $b_n = 1/2^n$.

Suppose that we have an alternating series with terms $\{(-1)^n b_n\}_{n=0}^{\infty}$, for which $b_n > 0$ for $n \geq 0$, such that the terms are non-increasing in magnitude: $b_{n+1} \leq b_n$ for $n \geq 0$. While one would normally try to establish convergence or divergence by examining the sequence of partial sums, because of the alternating signs of the terms, we will instead examine *alternating* partial sums.

Specifically, consider the sequence of even-numbered partial sums:

$$\begin{aligned} s_{2n} &= b_0 - b_1 + b_2 + \cdots + b_{2n} \\ &= (b_0 - b_1) + (b_2 - b_3) + \cdots + b_{2n}. \end{aligned}$$

Because $b_n \geq b_{n+1}$, each quantity in parentheses is a non-negative number, which means $s_{2n} \geq (b_0 - b_1)$ for $n \geq 0$. That is, the sequence of even-numbered partial sums is bounded below. A similar grouping of terms can be used to show that $s_{2n} \leq b_0$ for $n \geq 0$, so this sequence is also bounded above. In other words, it is bounded.

On the other hand, by the definition of a partial sum, we have

$$s_{2n} = s_{2n-2} - b_{2n-1} + b_{2n} = s_{2n-2} - (b_{2n-1} - b_{2n}) \leq s_{2n-2},$$

which shows that this sequence of partial sums is also non-increasing. That is, this sequencec is monotonic. It follows from the Monotonic Sequence Theorem that the sequence is convergent. A similar procedure can be used

to show that not only does the sequence of *odd-numbered* partial sums converge, but it converges to the same limit as the even-numbered partial sums, *provided that the sequence of terms converges to zero*, as must be the case for any convergent series. We conclude that the sequence of *all* partial sums converges, so the alternating series is convergent.

This leads to the *Alternating Series Test*: if the alternating series

$$\sum_{n=0}^{\infty}(-1)^n b_n = b_0 - b_1 + b_2 - b_3 + \cdots,$$

where $b_n > 0$ for $n \geq 0$, satisfies these conditions:

$$b_{n+1} \leq b_n, \quad n \geq 0, \quad \text{and} \quad \lim_{n\to\infty} b_n = 0,$$

then the series is convergent.

Example Consider the alternating series

$$\sum_{n=1}^{\infty}(-1)^n \frac{n}{n^2+1}.$$

The terms in the series converge to zero as $n \to \infty$. Furthermore, by differentiating the function

$$f(x) = \frac{n}{n^2+1}$$

with respect to x, we can confirm that this function satisfies $f'(1) = 0$, and $f'(x) < 0$ for $x > 1$. It follows that this function is non-increasing for $x \geq 1$, and therefore the terms of the series are non-increasing. We conclude that the series passes the Alternating Series Test, and converges. □

1.7.2 Estimating Error in Alternating Series

When computing the sum of a convergent series numerically, it is desirable to know how many terms are required in order to approximate the sum to within a given level of accuracy. For general series, it is difficult to estimate the error incurred by truncating the series after a given number of terms, although for some series, a variant of the Integral Test may be used. For alternating series, however, it is particularly simple to estimate this error, if the series satisfies the Alternating Series Test.

Consider the general alternating series used to develop the Alternating Series Test. We established that the sequence of even-numbered partial sums

is non-increasing. Similarly, the sequence of odd-numbered partial sums is non-decreasing. Since both sequences converge to the same limit, which is the sum s of the series, it follows that s lies between any two consecutive partial sums s_n and s_{n+1}, for some $n \geq 0$. Therefore,

$$|s - s_n| \leq |s_{n+1} - s_n|.$$

However, $s_{n+1} - s_n = b_{n+1}$, the next term in the series. We conclude that the error in the nth partial sum is bounded above by b_{n+1}.

We have just proved the *Alternating Series Estimation Theorem*: If an alternating series

$$\sum_{n=0}^{\infty}(-1)^n b_n,$$

where $b_n > 0$ for $n \geq 0$, satisfies these conditions:

$$b_{n+1} \leq b_n, \quad n \geq 0, \quad \text{and} \quad \lim_{n\to\infty} b_n = 0,$$

then

$$|s - s_n| \leq b_{n+1},$$

where s is the sum of the series and s_n is the nth partial sum.

Example Consider the convergent alternating series

$$\sum_{n=0}^{\infty}\left(-\frac{1}{3}\right)^n = \sum_{n=0}^{\infty}(-1)^n\frac{1}{3^n}.$$

We wish to approximate the sum s of this series with a partial sum s_n that includes enough terms so that $|s - s_n| \leq 0.001$. By the Alternating Series Estimation Theorem, we must choose n so that $1/3^{n+1} \leq 0.001$. Rearranging, we obtain the condition $1000 \leq 3^{n+1}$, or, by taking the natural logarithm of both sides,

$$n \geq \frac{\ln 1000}{\ln 3} - 1 = \frac{3\ln 10}{\ln 3} - 1 \approx 5.29.$$

Therefore, we must use the first 7 terms (from $n = 0$ to $n = 6$) to approximate the sum.

In this case, we can confirm that using this many terms is necessary and sufficient, since we can compute the sum of the series. This is a geometric series with $a = 1$ and $r = -1/3$, so the sum is $1/(1 - (-1/3)) = 3/4$. Using the first 7 terms yields an error of approximately 0.000343, while using only the first 6 terms yields an error of approximately 0.00103. □

1.7.3 Absolute Convergence

Previous tests for convergence have been applicable to series whose terms are all positive, or that alternate in sign. However, many series do not fall into either category. For such series, we can try to determine convergence by examining a new series obtained by taking the absolute values of each term, and applying one of the previous tests, such as the Integral Test or Comparison Test, to this modified series.

Specifically, given a series

$$\sum_{n=1}^{\infty} a_n,$$

suppose that we are able to determine that the modified series $\sum |a_n|$ is convergent. We then say that the original series is *absolutely convergent*. However, we wish to determine whether the original series is convergent. Because $0 \leq a_n + |a_n| \leq 2|a_n|$, and $\sum 2|a_n|$ is convergent, we can employ the Comparison Test to conclude that $\sum(a_n + |a_n|)$ is also convergent. However, the difference of two convergent series is also convergent, from which it follows that the series

$$\sum_{n=1}^{\infty} |a_n| - \sum_{n=1}^{\infty}(|a_n| + a_n) = \sum_{n=1}^{\infty} a_n$$

is convergent.

This leads to the *Absolute Convergence Test*: an absolutely convergent series is convergent.

Example The series

$$\sum_{n=0}^{\infty} \frac{\cos n\pi x}{n^2+1}$$

is an example of a *Fourier series*, which is useful for approximating functions, as well as working with data from signals and images. For any given x, the terms of this series will vary in sign, with a pattern that depends on x. However, if we consider the series obtained by taking the absolute value of the terms, we have

$$\sum_{n=0}^{\infty} \left| \frac{\cos n\pi x}{n^2+1} \right| \leq \sum_{n=0} \frac{1}{n^2+1} \leq 1 + \sum_{n=1}^{\infty} \frac{1}{n^2},$$

because $|\cos n\pi x| \leq 1$ regardless of x. Using two applications of the Comparison Test, based on the relationships above, we can conclude that the original series converges. □

It is important to note that the *converse* of the Absolute Convergence Test is not true: a convergent series is not necessarily absolutely convergent.

Example The alternating series

$$\sum_{n=1}^{\infty}(-1)^n\frac{1}{n}$$

is convergent, by the Alternating Series Test, but taking the absolute values of the terms yields the *harmonic series*, which, being a p-series with $p = 1$, is divergent. □

A series that is convergent, but not absolutely convergent, is said to be *conditionally convergent*.

1.7.4 The Ratio Test

Many of the convergence tests that we have seen impose a condition on *all* of the terms of a series, but this is unnecessarily restrictive, because convergence of a series is not affected by a finite number of terms. Rather, it is the behavior of the terms as the index $n \to \infty$ that is most deterministic of the behavior of the entire series.

For example, suppose that a series $\sum a_n$ satisfies the condition

$$\left|\frac{a_{n+1}}{a_n}\right| = L,$$

for some number $L < 1$. It follows that as $n \to \infty$, the terms $\{a_n\}$ behave approximately like a geometric series with a common ratio r that is less than one in absolute value, because the ratio of the magnitudes of consecutive terms is less than one, in the limit. Because such a geometric series is convergent, and this series involves the absolute values of the terms, this suggests that the series is absolutely convergent.

This is in fact the case, and leads to one of the most useful tests for convergence, the *Ratio Test:* let

$$L = \lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right|,$$

assuming this limit exists. If $L < 1$, then the series is absolutely convergent. If $L > 1$, or the limit does not exist, then the series is divergent. If $L = 1$, the test is inconclusive, and another test must be used instead.

This test is particularly useful for series whose terms involve powers of n, exponential functions, or factorials, as this example shows.

Example Applying the Ratio Test to the series

$$\sum_{n=0}^{\infty} \frac{n!}{2^n(n+1)^2}$$

yields

$$\begin{aligned}
\lim_{n\to\infty} \left| \frac{(n+1)!}{2^{n+1}(n+2)^2} \Big/ \frac{n!}{2^n(n+1)^2} \right| &= \lim_{n\to\infty} \left| \frac{(n+1)!}{n!} \frac{2^n}{2^{n+1}} \frac{(n+1)^2}{(n+2)^2} \right| \\
&= \lim_{n\to\infty} \left| \frac{n+1}{2} \left(\frac{n+1}{n+2} \right)^2 \right| \\
&= \infty
\end{aligned}$$

and therefore the series diverges. □

1.7.5 The Root Test

A similar test to the Ratio Test is the Root Test, which is convenient for series whose terms are raised to the nth power, where n is the index. The basic idea behind this test is that because a geometric series $\sum r^n$ converges when $|r| < 1$, if the terms of a more general series $\sum a_n$ converge, in absolute value, to an expression of the form L^n where $L < 1$, this series should converge as well. If this is the case, then $L = \sqrt[n]{|a_n|}$.

With this idea in mind, the *Root Test* is as follows: let

$$L = \lim_{n\to\infty} \sqrt[n]{|a_n|},$$

assuming this limit exists. If $L < 1$, then the series is absolutely convergent. If $L > 1$, or the limit does not exist, then the series is divergent. If $L = 1$, the test is inconclusive, and another test must be used instead.

Example Applying the Root Test to the series

$$\sum_{n=0}^{\infty} \left(\frac{1}{\pi} \tan^{-1} n \right)^n$$

yields

$$\lim_{n\to\infty} \sqrt[n]{\left| \left(\frac{1}{\pi} \tan^{-1} n \right)^n \right|} = \lim_{n\to\infty} \frac{1}{\pi} \tan^{-1} n = \frac{1}{\pi} \frac{\pi}{2} = \frac{1}{2} < 1,$$

which indicates that the series converges absolutely. □

1.7.6 Summary

- The Integral Test can be used to determine whether a series converges. If the terms $\{a_n\}_{n=1}^{\infty}$ of a series are defined to be the values of a continuous, positive, decreasing function $f(n)$ where n is each positive integer, and if $f(x)$ is integrable on the interval $[1, \infty)$, then the series converges, and the sum of the series, excluding the first term a_1, is bounded above by the value of the integral.

- The Integral Test can also be used to test for the divergence of a series. If the integral of $f(x)$ from 1 to ∞ diverges, then the series whose terms are defined by $a_n = f(n)$, for $n \geq 1$, also diverges.

- The function $f(x)$ need not be positive and decreasing for the entire domain of integration, but it needs to be positive and decreasing for all $x > c$, for some c in the domain.

- The initial index of the series and the lower limit of integration do not have to be equal to 1, but the same value should be used for both.

- Using the Integral Test, it can be shown that the p-series, whose terms are given by $a_n = n^{-p}$, converges if $p > 1$ and diverges if $p \leq 1$.

- The Comparison Test states that a series whose terms are dominated by those a convergent series is also convergent, and that a series whose terms dominate those of a divergent series is also divergent. In both cases, the terms of both series must be non-negative.

- The Limit Comparison Test states that if two series have terms whose ratio converges to a nonzero, finite limit, then the series have the same behavior; that is, they both converge or both diverge. This test can be useful in cases for which the Comparison Test is inconclusive.

- The Alternating Series Test states that an alternating series is convergent if its terms are non-increasing in magnitude, and converge to zero.

- If an alternating series passes the Alternating Series Test, then any partial sum of the series deviates from the overall sum by no more than the next term in the series.

- A series is absolutely convergent if the new series obtained by taking the absolute values of the terms of the original series is convergent.

- A series that is absolutely convergent is also convergent, but not necessarily other way around. A convergent series that fails to be absolutely convergent is conditionally convergent.

- The Ratio Test is one of the most useful tests for determining whether a series converges. If the ratio $|a_{n+1}/a_n|$ converges to a limit that is less than 1, then $\sum a_n$ converges absolutely. If the limit is larger than 1, or does not exist, then the series diverges. Otherwise, the test is inconclusive.

- The Root Test determines convergence based on the limit of the nth root of the absolute value of the nth term, as $n \to \infty$. If the limit less than 1, the series converges absolutely. If the limit is greater than 1, or nonexistent, the series diverges. Otherwise, the test is inconclusive.

1.8 Power Series

Now that we have learned about how to test sequences and series for convergence, and, in some cases, compute their limits when they do converge, we are ready to take our next step toward our primary goal: to be able to approximate functions to within a given degree of accuracy using "simple" functions that can readily be evaluated using simple arithmetic operations, the only ones a computer can perform. In this next step, we consider series that depend on a variable x, like the functions we wish to approximate.

1.8.1 What is a Power Series?

A *power series* is a series of the form

$$\sum_{n=0}^{\infty} c_n(x-a)^n,$$

where the constants c_n, for $n \geq 0$, are called the *coefficients*, and the number a is called the *center*.

When a value is substituted for the variable x, the power series reduces to a series of constant terms that can then be tested for convergence, using any of the tests that we have previously discussed. The result of such a test, convergence or divergence, depends on x.

1.8.2 Convergence of Power Series

We now illustrate, through some examples, how tests for the convergence of general series can be used to determine the values of x for which a power series converges.

Example Consider the power series

$$\sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

Applying the Ratio Test, for any *fixed* x, yields

$$\lim_{n\to\infty} \left| \frac{x^{n+1}}{(n+1)!} \Big/ \frac{x^n}{n!} \right| = \lim_{n\to\infty} \left| \frac{x^{n+1}}{x^n} \frac{n!}{(n+1)!} \right| = \lim_{n\to\infty} \left| \frac{x}{n+1} \right| = 0,$$

since no matter how large $|x|$ is, eventually n will exceed it. Therefore, this power series is absolutely convergent for all x. □

Example The power series

$$\sum_{n=0}^{\infty} n^n x^n = \sum_{n=0}^{\infty} (nx)^n$$

diverges when $x \neq 0$, because, by the Ratio Test,

$$\lim_{n\to\infty} \left| \frac{((n+1)x)^{n+1}}{(nx)^n} \right| = \lim_{n\to\infty} \left| \left(1 + \frac{1}{n}\right)^n (n+1)x \right| = \lim_{n\to\infty} |e(n+1)x| = \infty.$$

On the other hand, if $x = 0$, then all terms vanish except for $n = 0$, and the series (trivially) converges to 1. □

Example Consider the power series

$$\sum_{n=1}^{\infty} \frac{x^n}{n}.$$

Applying the Ratio Test yields

$$\lim_{n\to\infty} \left| \frac{x^{n+1}}{n+1} \Big/ \frac{x^n}{n} \right| = \lim_{n\to\infty} \left| \frac{x^{n+1}}{x^n} \frac{n}{n+1} \right| = |x|.$$

It follows that the power series converges when $|x| < 1$, and diverges when $|x| > 1$. When $x = 1$, the series diverges, because it is the harmonic series, but when $x = -1$, it converges, by the Alternating Series Test. □

1.8.3 The Radius of Convergence

In the preceding examples, we encountered three different scenarios regarding convergence of a power series:

- The power series converges for all x
- The power series converges only for $x = a$, where a is the center of the series
- The power series converges when $|x - a| < R$ for some number R, and diverges for $|x - a| > R$.

It can be shown that for any power series, these are the *only* possible outcomes of a test for convergence. In other words, there exists an interval, called the *interval of convergence*, such that for x in this interval, the power series, when evaluated at x, is convergent.

The interval may be the entire real line $(-\infty, \infty)$, a single point $x = a$, or a finite interval of the form $|x - a| < R$, or $a - R < x < a + R$. In this case, the number R is called the *radius of convergence*. The series diverges for $|x - a| > R$, but when $|x - a| = R$, either convergence or divergence is possible. We say that when the interval of convergence is the entire real line, $R = \infty$, and if it is the single point $x = a$, then $R = 0$.

1.8.4 Representing Functions as Power Series

We have learned that a convergent power series

$$\sum_{n=0}^{\infty} c_n(x - x_0)^n \tag{1.5}$$

is a function of x, whose domain is the series' interval of convergence. Given such a power series, it is very helpful to know which function it represents, because that knowledge can provide an easier approach to evaluating or analyzing the function than working with the power series directly.

However, the opposite task is also very useful: given a function $f(x)$, find the corresponding power series. We will see that in some cases, the power series can be the only practical way to evaluate the function, so knowledge of its power series is vital. One useful technique for obtaining the power series of a function is to relate the function to the sum of a convergent geometric series

$$\sum_{n=0}^{\infty} ar^n = \frac{a}{1 - r},$$

as the following examples show.

Example Let

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

If we rewrite $f(x)$ to fit the form $a/(1-r)$, for some a and r, we obtain

$$f(x) = \frac{2}{1-(-2x)},$$

which, when compared to the form $a/(1-r)$, yields $a = 2$ and $r = -2x$. It follows that $f(x)$ is the sum of the geometric series

$$\sum_{n=0}^{\infty} 2(-2x)^n,$$

provided that $|-2x| < 1$, or $|x| < 1/2$. □

The common ratio r need not be a linear function of x, as the following example shows.

Example Let

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

Trying to fit $f(x)$ into the form $a/(1-r)$ requires a 1 in the denominator, which can be obtained by dividing the numerator and denominator by 4:

$$\frac{1}{(x-2)^2+4} = \frac{\frac{1}{4}}{\frac{(x-2)^2}{4}+1} = \frac{\frac{1}{4}}{1-\left[-\left(\frac{x-2}{2}\right)^2\right]}.$$

It follows that $a = 1/4$ and $r = -((x-2)/2)^2$, resulting in the power series

$$\sum_{n=0}^{\infty} \frac{1}{4}\left[-\left(\frac{x-2}{2}\right)^2\right]^n = \sum_{n=0}^{\infty} \frac{1}{4}(-1)^n\left(\frac{x-2}{2}\right)^{2n} = \sum_{n=0}^{\infty} \frac{(-1)^n}{4^{n+1}}(x-2)^{2n},$$

which converges when $|-((x-2)/2)^2| < 1$. This inequality holds if $|(x-2)/2| < 1$, or $|x-2| < 2$. That is, $f(x)$ is defined by this power series when $0 < x < 4$. Note that in terms of the general form of a power series in equation (1.5), $x_0 = 2$, while $c_n = 0$ whenever n is odd, because only even powers of $(x-2)$ are included. On the other hand, when n is even, $c_n = (-1)^{n/2}/4^{n/2+1}$. □

Example A function does not have to have a constant numerator in order to be defined using a geometric power series. Consider

$$f(x) = \frac{2x}{2-x}.$$

This can be rewritten as

$$\frac{x}{2-x} = \frac{2x}{2}\frac{1}{1-\frac{x}{2}} = x\frac{1}{1-\frac{x}{2}}.$$

The fraction is the sum of a geometric series with $a = 1$ and $r = x/2$. It follows that

$$f(x) = x\sum_{n=0}^{\infty}\left(\frac{x}{2}\right)^n,$$

provided $|r| = |x/2| < 1$, or $-2 < x < 2$.

However, this representation of $f(x)$ does not fit the usual form of a power series given by equation (1.5). To obtain this form, we can rewrite $f(x)$ again as follows:

$$f(x) = 2\frac{x}{2}\sum_{n=0}^{\infty}\left(\frac{x}{2}\right)^n = 2\sum_{n=0}^{\infty}\left(\frac{x}{2}\right)^{n+1} = 2\sum_{n=1}^{\infty}\left(\frac{x}{2}\right)^n = 2\sum_{n=1}^{\infty}\frac{1}{2^n}x^n = \sum_{n=1}^{\infty}\frac{1}{2^{n-1}}x^n,$$

which does fit the form in equation (1.5), with $c_0 = 0$ and $c_n = 1/2^{n-1}$ for $n \geq 1$. □

This last example shows that if $f(x)$ can be represented by a power series, then so can $x^p f(x)$ for any positive integer p, simply by shifting the indices of the coefficients $\{c_n\}_{n=0}^{\infty}$ up by p.

1.8.5 Differentiation and Integration of Power Series

We have previously learned how to compute power series representations of certain functions, by relating them to geometric series. We can obtain power series representation for a wider variety of functions by exploiting the fact that a convergent power series can be differentiated, or integrated, term-by-term to obtain a new power series that has the *same radius of convergence* as the original power series. The new power series is a representation of the derivative, or antiderivative, of the function that is represented by the original power series.

This is particularly useful when we have a function $f(x)$ for which we do not know how to obtain a power series representation directly. If its

derivative $f'(x)$, or its antiderivative $\int f(x)\,dx$, is a function for which a power series representation can easily be computed, such as the examples from earlier in this section, then we can integrate, or differentiate, this power series term-by-term to obtain a power series for $f(x)$.

Example The function

$$f(x) = \frac{4}{(2-x)^2}$$

is the derivative of the function

$$g(x) = \frac{2x}{2-x},$$

which, from earlier in this section, has the power series representation

$$\frac{2x}{2-x} = \sum_{n=1}^{\infty} \frac{1}{2^{n-1}} x^n.$$

This series converges when $-2 < x < 2$. To obtain a power series representation of $f(x)$, we differentiate this series term-by-term to obtain

$$\frac{4}{(2-x)^2} = \sum_{n=1}^{\infty} \frac{1}{2^{n-1}} n x^{n-1} = \sum_{n=0}^{\infty} \frac{(n+1)}{2^n} x^n,$$

which also converges when $-2 < x < 2$. □

Example The function

$$f(x) = \frac{1}{2} \tan^{-1} \frac{x-2}{2}$$

has the derivative

$$f'(x) = \frac{1}{(x-2)^2 + 4}.$$

From earlier in this section, this function has the power series

$$\frac{1}{(x-2)^2+4} = \sum_{n=0}^{\infty} \frac{(-1)^n}{4^{n+1}} (x-2)^{2n},$$

whose interval of convergence is $0 < x < 4$. Integrating this series term-by-term yields

$$\frac{1}{2} \tan^{-1} \frac{x-2}{2} = \sum_{n=0}^{\infty} \frac{(-1)^n}{4^{n+1}} \int (x-2)^{2n}\,dx = \sum_{n=0}^{\infty} \frac{(-1)^n}{4^{n+1}} \frac{(x-2)^{2n+1}}{2n+1} + C.$$

To determine the value of C, we substitute $x = 2$ into the above equation. This causes all terms in the series to vanish. We also have $f(2) = 0$, which yields $C = 0$. □

Example Consider the definite integral

$$\int_0^1 \frac{1}{1+x^4}\,dx.$$

Attempting to evaluate this integral using partial fraction decomposition is not possible without introducing complex numbers. Instead, we express the integrand as a (geometric) power series:

$$\frac{1}{1+x^4} = \frac{1}{1-(-x^4)} = \sum_{n=0}^{\infty}(-x^4)^n = \sum_{n=0}^{\infty}(-1)^n x^{4n}.$$

This power series has an interval of convergence of $-1 < x < 1$, which contains the interval of integration $(0, 1)$. Integrating the power series term-by-term from 0 to 1 yields

$$\int_0^1 \frac{1}{1+x^4}\,dx = \int_0^1 \sum_{n=0}^{\infty}(-1)^n x^{4n}\,dx = \sum_{n=0}^{\infty}(-1)^n \int_0^1 x^{4n}\,dx = \sum_{n=0}^{\infty}(-1)^n \left.\frac{x^{4n+1}}{4n+1}\right|_0^1 = \sum_{n=0}^{\infty}\frac{(-1)^n}{4n+1}.$$

This is an alternating series, which, by the Alternating Series Test, converges since, for all $n \geq 0$,

$$\frac{1}{4n+1} \geq 0, \quad \lim_{n\to\infty}\frac{1}{4n+1} = 0, \quad \text{and} \quad \frac{1}{4(n+1)+1} < \frac{1}{4n+1}.$$

Using the Alternating Series Estimation Theorem, we can evaluate this integral numerically, to any degree of accuracy we wish, by choosing n large enough so that $1/(4n+1)$ is sufficiently small. □

1.8.6 Summary

- A power series is a series whose terms are constants, called coefficients, times non-negative powers of $(x - a)$ for some constant a, which is called the center of the power series.

- The convergence or divergence of power series can be determined using the same tests that are used for other series, except that the behavior depends on x.

- A power series either converges only at the center, converges for all x, or converges within the interval $|x - a| < R$, where R is the radius of convergence, while diverging for $|x - a| > R$. In the latter scenario, when $|x - a| = R$, the power series can either converge or diverge.

- The interval of convergence contains all values of x for which the power series converges.

- A function $f(x)$ can sometimes be represented as a power series by viewing it as the sum of a convergent geometric series. The key is to correctly identify what function serves as the common ratio of the series, which may require algebraic manipulation of $f(x)$.

- The common ratio, which must appear in the denominator of $f(x)$, must be of the form $[c(x - x_0)]^q$, where c and x_0 are constants, and q is a positive integer.

- If $f(x)$ can be represented by a geometric power series, then so can $cf(x)$, for any constant c, and so can $x^p f(x)$, for any positive integer p.

- Given such a function $f(x)$, the corresponding power series converges if the common ratio is less than 1 in absolute value; that is, $c|x-x_0| < 1$. The radius of convergence is $1/|c|$.

- A power series representation of a function $f(x)$ can be differentiated term-by-term to obtain a power series representation of its derivative $f'(x)$. The interval of convergence of the differentiated series is the same as that of the original series.

- A power series representation of a function $f(x)$ can be anti-differentiated term-by-term to obtain a power series representation of its anti-derivative $\int f(x)\,dx$. The value of the constant of integration, C, can be determined by substituting the center of the power series for x. The interval of convergence of the anti-differentiated series is the same as that of the original series.

- A power series representation of a function $f(x)$ can be integrated term-by-term from a to b to obtain a series representation of the definite integral $\int_a^b f(x)\,dx$, provided that the interval (a, b) lies within the interval of convergence of the power series that represents $f(x)$.

1.9 Taylor and Maclaurin Series

We have learned how to construct power series representations of certain functions by relating them to geometric series, either directly, or indirectly through differentiation or integration. However, these techniques are not applicable to most functions. Therefore, we now consider the problem of computing the coefficients c_n, $n = 0, 1, \ldots$, of a power series

$$f(x) = \sum_{n=0}^{\infty} c_n (x - x_0)^n$$

that represents a more general function $f(x)$, at least for x near the center x_0.

To that end, we first note that the partial sum of such a power series,

$$T_n(x) = \sum_{j=0}^{n} c_j (x - x_0)^j,$$

is a polynomial of degree n. Furthermore, by substituting $x = x_0$, we find that

$$T_n(x_0) = c_0, \quad T_n'(x_0) = c_1, \quad T_n''(x_0) = 2c_2,$$

and, in general,

$$\frac{d^j}{dx^j}[T_n(x)]\bigg|_{x=x_0} = T_n^{(j)}(x_0) = j!c_j.$$

This suggests that a function $f(x)$, that has infinitely many derivatives at x_0, can be represented by the power series

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n,$$

for all x within the interval of convergence of the series. This is in fact the case, and we call this series the *Taylor series of f centered at x_0*. When $x_0 = 0$, this Taylor series is also known as the *Maclaurin series of f*.

Example The Maclaurin series for $f(x) = e^x$ is given by

$$e^x = 1 + x + \frac{x^2}{2} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!},$$

since $f^{(n)}(0) = e^0 = 1$ for all n. We have seen that this series converges for all x, so e^x can be represented by this power series for all x as well. □

While Taylor series have many theoretical uses, they not directly useful for practical applications, because most functions are not infinitely differentiable, and, of course, any actual computation can only include a finite number of terms. Therefore, it is highly desirable to know how accurately a function can be represented by its *truncated* Taylor series. That is, for a given x within the interval of convergence of the Taylor series, our goal is to estimate the error $R_n(x) = f(x) - T_n(x)$, where $T_n(x)$, the nth partial sum of the Taylor series of f, is also known as the nth-degree *Taylor polynomial of f centered at x_0*. The error, $R_n(x)$, is called the *Taylor remainder*.

We first consider the simplest Taylor polynomial, which is $T_0(x) = f(x_0)$. To obtain the error $R_0(x)$, we apply the Fundamental Theorem of Calculus, and obtain

$$f(x) = f(x_0) + \int_{x_0}^{x} f'(s)\,ds.$$

That is,

$$R_0(x) = \int_{x_0}^{x} f'(s)\,ds.$$

Unfortunately, expressing the remainder as an integral does little to help us to estimate its magnitude. However, we can apply the Mean Value Theorem for Integrals to obtain

$$f(x) = f(x_0) + f'(\xi)(x - x_0),$$

where ξ is between x_0 and x.

Now, we consider the first-degree Taylor polynomial

$$T_1(x) = f(x_0) + f'(x_0)(x - x_0).$$

This graph of this function is the tangent line of f at x_0. To obtain the error $R_1(x) = f(x) - T_1(x)$, we apply the Fundamental Theorem of Calculus twice, to represent both f *and its derivative* in terms of their values at x_0, plus an integral of their respective derivatives. We obtain

$$\begin{aligned}
f(x) &= f(x_0) + \int_{x_0}^{x} \left[f'(x_0) + \int_{x_0}^{s} f''(z)\,dz \right] ds \\
&= f(x_0) + \int_{x_0}^{x} f'(x_0)\,ds + \int_{x_0}^{x}\int_{x_0}^{s} f''(z)\,dz\,ds \\
&= f(x_0) + f'(x_0)(x - x_0) + \int_{x_0}^{x}\int_{z}^{x} f''(z)\,ds\,dz \\
&= T_1(x) + \int_{x_0}^{x} f''(z) \int_{z}^{x} ds\,dz
\end{aligned}$$

$$= T_1(x) + \int_{x_0}^{x} f''(z)(x-z)\,dz.$$

That is,

$$R_1(x) = f(x) - T_1(x) = \int_{x_0}^{x} f''(z)(x-z)\,dz.$$

To obtain a more useful representation of the remainder than this integral form, we use the fact that $x - z$ does not change sign on the interval of integration from x_0 to x. This allows us to apply the Weighted Mean Value Theorem for Integrals, which states that

$$\int_a^b f(x)g(x)\,dx = f(c)\int_a^b g(x)\,dx,$$

where c lies in $[a, b]$, and $g(x)$ does not change sign on (a, b). We then obtain

$$\begin{aligned} f(x) &= T_1(x) + f''(\xi)\int_{x_0}^{x} x - z\,dz \\ &= T_1(x) + f''(\xi)\int_0^{x-x_0} u\,du \\ &= T_1(x) + f''(\xi)\frac{(x-x_0)^2}{2}, \end{aligned}$$

where ξ lies between x_0 and x. That is,

$$R_1(x) = f''(\xi)\frac{(x-x_0)^2}{2}.$$

Example Let $f(x) = \sin x$. Then, the Taylor polynomial $T_1(x)$ is

$$T_1(x) = \sin 0 + (\cos 0)(x - 0) = x.$$

The remainder is given by

$$R_1(x) = -\frac{\sin\xi}{2}x^2.$$

Since $|\sin x| \leq 1$ for all x, it follows that if $|x| \leq 10^{-2}$, then the error in approximating $\sin x$ by x is at most

$$|R_1(x)| \leq \frac{(10^{-2})^2}{2} = \frac{10^{-4}}{2} = 5 \times 10^{-5}.$$

□

We now consider the case of an nth-degree Taylor polynomial. We will need the result

$$\int_{x_0}^{x}\int_{x_0}^{s_1}\int_{x_0}^{s_2}\cdots\int_{x_0}^{s_{n-1}} ds_n\cdots ds_3\,ds_2\,ds_1 = \frac{(x-x_0)^n}{n!},$$

previously used for $n = 1$ and $n = 2$. This formula can be proven for an arbitrary positive integer n using an inductive argument. We have

$$\begin{aligned}
f(x) &= f(x_0) + \int_{x_0}^{x}\left\{f'(x_0) + \int_{x_0}^{s_1}\left[f''(x_0) + \int_{x_0}^{s_2}\left(f'''(x_0) + \int_{x_0}^{s_3}\cdots\right.\right.\right.\\
&\quad \left.\left.\left.\int_{x_0}^{s_{n-1}}\left\{f^{(n)}(x_0) + \int_{x_0}^{s_n} f^{(n+1)}(s_{n+1})\,ds_{n+1}\right\} ds_n\cdots ds_4\right) ds_3\right] ds_2\right\} ds_1\\
&= f(x_0) + f'(x_0)\int_{x_0}^{x} ds_1 + f''(x_0)\int_{x_0}^{x}\int_{x_0}^{s_1} ds_2\,ds_1 +\\
&\quad f'''(x_0)\int_{x_0}^{x}\int_{x_0}^{s_1}\int_{x_0}^{s_2} ds_3\,ds_2\,ds_1 + \cdots + f^{(n)}(x_0)\int_{x_0}^{x}\int_{x_0}^{s_1}\cdots\int_{x_0}^{s_{n-1}} ds_n\cdots ds_2\,ds_1 +\\
&\quad \int_{x_0}^{x}\int_{x_0}^{s_1}\cdots\int_{x_0}^{s_n} f^{(n+1)}(s_{n+1})\,ds_{n+1}\cdots ds_2\,ds_1\\
&= f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2}f''(x_0)(x-x_0)^2 + \frac{1}{6}f'''(x_0)(x-x_0)^3 +\\
&\quad \cdots + \frac{f^{(n)}}{n!}(x-x_0)^n + \int_{x_0}^{x}\int_{x_0}^{s_1}\cdots\int_{x_0}^{s_n} f^{(n+1)}(s_{n+1})\,ds_{n+1}\cdots ds_2\,ds_1\\
&= \sum_{j=0}^{n}\frac{f^{(j)}(x_0)}{j!}(x-x_0)^j + \int_{x_0}^{x}\int_{s_{n+1}}^{x}\cdots\int_{s_2}^{x} f^{(n+1)}(s_{n+1})\,ds_1\cdots ds_n\,ds_{n+1}\\
&= \sum_{j=0}^{n}\frac{f^{(j)}(x_0)}{j!}(x-x_0)^j + \int_{x_0}^{x} f^{(n+1)}(s_{n+1})\int_{s_{n+1}}^{x}\cdots\int_{s_2}^{x} ds_1\cdots ds_n\,ds_{n+1}\\
&= \sum_{j=0}^{n}\frac{f^{(j)}(x_0)}{j!}(x-x_0)^j + \int_{x_0}^{x}\frac{f^{(n+1)}(s_{n+1})}{n!}(x-s_{n+1})^n\,ds_{n+1}\\
&= T_n(x) + R_n(x).
\end{aligned}$$

The transformation of the $(n+1)$-fold integral between the third and fourth steps follows from the relations

$$x_0 \le s_{n+1} \le s_n \le \cdots \le s_2 \le s_1 \le x,$$

when $x_0 \le x$. Similar relations hold when $x_0 \ge x$.

In summary, the Taylor polynomial and Taylor remainder are given by

$$T_n(x) = f(x_0)+f'(x_0)(x-x_0)+\frac{f''(x_0)}{2}(x-x_0)^2+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n = \sum_{j=0}^{n}\frac{f^{(j)}(x_0)}{j!}(x-x_0)^j,$$

and

$$R_n(x) = \int_{x_0}^{x}\frac{f^{(n+1)}(s)}{n!}(x-s)^n\,ds.$$

Using the Weighted Mean Value Theorem for Integrals, as before, we obtain the alternative form of the remainder,

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{n!}\int_{x_0}^{x}(x-s)^n\,ds = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1},$$

where ξ lies between x_0 and x. By computing an upper bound on $|f^{(n+1)}(x)|$ for x near x_0, we can estimate the accuracy of an approximation of $f(x)$ near x_0 by $T_n(x)$.

Example The 3rd-degree Taylor polynomial for $f(x) = e^x$, centered at $x_0 = 0$, is given by

$$T_3(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6},$$

with remainder

$$R_3(x) = \frac{e^\xi}{24}x^4,$$

where ξ lies between 0 and x. If we let $x = 2.5$, it follows that the error in an approximation of e^x by $T_3(x)$ is at most

$$|R_3(2.5)| = \left|\frac{e^\xi}{24}(2.5)^4\right| \leq \frac{e^{2.5}}{24}(39.0625) \approx 19.828,$$

which is unacceptably large.

On the other hand, if we use the 3rd-degree Taylor polynomial centered at $x_0 = 2$,

$$T_3(x) = e^2 + e^2(x-2) + \frac{e^2}{2}(x-2)^2 + \frac{e^2}{6}(x-2)^3,$$

with remainder

$$R_3(x) = \frac{e^\xi}{24}(x-2)^4,$$

where ξ lies between 2 and x, we obtain the error bound

$$|R_3(2.5)| = \left|\frac{e^\xi}{24}(0.5)^4\right| \leq \frac{e^{2.5}}{24}(0.0625) \approx 0.0317,$$

which implies that using the Taylor polynomial centered at $x_0 = 2$ is far more accurate. □

Taylor polynomials are quite useful for approximating functions in such a way that they can easily be evaluated by a computer or calculator, because once the coefficients are known, the computation can be performed using only basic arithmetic operations. However, as the preceding example shows, a Taylor polynomial for a given function $f(x)$ must be used judiciously, because it can accurately approximate $f(x)$ only when x is near the center x_0, especially when the derivatives of f are large.

Previously, we learned how to approximate a function $f(x)$ by its Taylor polynomial

$$T_n(x) = \sum_{j=0}^{n} \frac{f^{(j)}(x_0)}{j!}(x - x_0)^j,$$

for a given center x_0. Because the Taylor polynomial of degree n consists of the first $n + 1$ terms of its Taylor series

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n,$$

it is helpful to be able to compute all of the coefficients of the Taylor series of $f(x)$, so that the coefficients of a Taylor polynomial of any degree are available.

Computing a Taylor series in its entirety requires that the coefficients fit a pattern that can easily be discerned from the first few terms of the series. In particular, the derivatives of $f(x)$ at x_0 must fit an identifiable pattern, as in the case of the Maclaurin series for the exponential function,

$$e^x = \sum_{n=0}^{\infty} \frac{1}{n!}x^n.$$

In a previous example, we used the Ratio Test to conclude that this series converges for all x.

In addition to the task of computing the coefficients of a Taylor series, it is also essential to determine the interval of convergence of the series. A Taylor polynomial cannot be used to approximate a function at a point at

which the Taylor series does not converge, because the Taylor remainder does not converge to zero as the degree of the polynomial increases. We now illustrate the process of computing Taylor series of functions, and their intervals of convergence, through some examples.

Example Consider the function

$$f(x) = \frac{1}{1-x}.$$

Its derivatives are

$$f'(x) = \frac{1}{(1-x)^2}, \quad f''(x) = \frac{2}{(1-x)^3}, \quad f^{(n)}(x) = \frac{n!}{(1-x)^{n+1}}.$$

It follows that $f^{(n)}(0) = n!$, and therefore the Taylor series for f centered at 0, also known as its Maclaurin series, is

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n = \sum_{n=0}^{\infty} x^n,$$

which is simply a geometric series with $a = 1$ and $r = x$. We conclude that this series has a radius of convergence of $R = 1$, and an interval of convergence of $(-1, 1)$. □

Example The function $f(x) = \sin x$ has derivatives

$$f'(x) = \cos x, \quad f''(x) = -\sin x, \quad f'''(x) = -\cos x, \quad f^{(4)}(x) = \sin x,$$

or, in general,

$$f^{(4n)}(x) = \sin x, \quad f^{(4n+1)}(x) = \cos x, \quad f^{(4n+2)}(x) = -\sin x, \quad f^{(4n+3)}(x) = -\cos x,$$

for each nonnegative integer n. Substituting $x = 0$ yields

$$f^{(4n)}(0) = 0, \quad f^{(4n+1)}(0) = 1, \quad f^{(4n+2)}(0) = 0, \quad f^{(4n+3)}(0) = -1,$$

or, more concisely,

$$f^{(2n)}(0) = 0, \quad f^{(2n+1)}(0) = (-1)^n, \quad n \geq 0.$$

It follows that the Maclaurin series for $\sin x$ includes only odd powers of x. We have

$$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}.$$

By applying the Ratio Test, we find that this series converges for all x. That is, $R = \infty$ and the interval of convergence $(-\infty, \infty)$. □

Example To compute the Maclaurin series for $f(x) = \cos x$, we can use the information about derivatives from the previous example to obtain

$$f^{(4n)}(x) = \cos x, \quad f^{(4n+1)}(x) = -\sin x, \quad f^{(4n+2)}(x) = -\cos x, \quad f^{(4n+3)}(x) = \sin x,$$

for each nonnegative integer n. Substituting $x = 0$ yields

$$f^{(4n)}(0) = 1, \quad f^{(4n+1)}(0) = 0, \quad f^{(4n+2)}(0) = -1, \quad f^{(4n+3)}(0) = 0,$$

or, more concisely,

$$f^{(2n+1)}(0) = 0, \quad f^{(2n)}(0) = (-1)^n, \quad n \geq 0.$$

It follows that the Maclaurin series for $\cos x$ includes only even powers of x. We have

$$\cos x = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n}.$$

By applying the Ratio Test, we find that this series converges for all x. That is, $R = \infty$ and the interval of convergence $(-\infty, \infty)$. □

Example Consider the function $f(x) = (1 + x)^k$, where k is a real number (not necessarily an integer). Its derivatives are

$$f'(x) = k(1+x)^{k-1}, \quad f''(x) = k(k-1)(1+x)^{k-2}, \quad f'''(x) = k(k-1)(k-2)(1+x)^{k-3},$$

and, in general,

$$f^{(n)}(x) = k(k-1)(k-2)\cdots(k-n+1)(1+x)^{k-n},$$

and therefore

$$f^{(n)}(0) = k(k-1)(k-2)\cdots(k-n+1).$$

It follows that the Maclaurin series for $(1 + x)^k$ is the *binomial series*

$$(1+x)^k = \sum_{n=0}^{\infty} \frac{k(k-1)(k-2)\cdots(k-n+1)}{n!} x^k = \sum_{n=0}^{\infty} \binom{k}{n} x^k,$$

where

$$\binom{k}{n} = \frac{k(k-1)(k-2)\cdots(k-n+1)}{n!}$$

is called a *binomial coefficient*. It is a generalization of

$$\left(\begin{array}{c} k \\ n \end{array} \right) = \frac{k(k-1)(k-2)\cdots(k-n+1)}{n!} = \frac{k!}{n!(n-k)!},$$

where k and n are non-negative integers. In this case, this binomial coefficient represents the number of ways to choose n objects from a set of k objects.

To determine the interval of convergence, we again use the Ratio Test, which yields

$$\lim_{n\to\infty} \left| \frac{k(k-1)\cdots(k-n)}{k(k-1)\cdots(k-n+1)} \frac{n!}{(n+1)!} \frac{x^{n+1}}{x^n} \right| = \lim_{n\to\infty} \left| \frac{k-n}{n+1} \right| |x| = \lim_{n\to\infty} \left| \frac{\frac{k}{n}-1}{1+\frac{1}{n}} \right| |x| = |x|,$$

which implies that the series converges for $|x| < 1$. □

Previously, we have learned how to compute Taylor series for certain functions by computing the coefficients directly and recognizing the pattern that they fit. However, for many functions, this approach is not feasible, because any pattern in the coefficients is too complex to recognize. We now illustrate how other techniques for obtaining power series can be used, sometimes in tandem, to obtain Taylor series for certain well-known functions.

Example Recall the sum of a geometric series,

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

provided $|x| < 1$. That is, the radius of convergence is $R = 1$. Because the power series representation of a function at a given center is unique, it follows that the Maclaurin series for $1/(1-x)$ is the above geometric series.

Substituting $-x^2$ for x yields

$$\sum_{n=0}^{\infty} (-x^2)^n = \sum_{n=0}^{\infty} (-1)^n x^{2n} = \frac{1}{1-(-x^2)} = \frac{1}{1+x^2}.$$

Because $|x^2| < 1$ implies that $|x| < 1$, this series has the same radius of convergence, $R = 1$. Furthermore, the above series is the Maclaurin series for $1/(1+x^2)$.

By anti-differentiating this series term-by-term, we can obtain a Maclaurin series for $\tan^{-1} x$, for $|x| < 1$, which is an antiderivative of $1/(1+x^2)$:

$$\tan^{-1} x = \int_0^x \frac{1}{1+s^2}\, ds + C$$

$$\begin{aligned} &= \int_0^x \sum_{n=0}^{\infty} (-1)^n s^{2n}\, ds + C \\ &= \sum_{n=0}^{\infty} (-1)^n \int_0^x s^{2n}\, ds + C \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1} + C. \end{aligned}$$

The unknown constant C is included because we have computed a specific antiderivative of $1/(1+x^2)$, and all antiderivatives of a given function differ from one another by a constant. To compute C, we substitute $x = 0$, and because $\tan^{-1} 0 = 0$, we obtain the equation $0 = 0 + C$, so we conclude that $C = 0$ and the Maclaurin series for $\tan^{-1} x$ is

$$\tan^{-1} x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1}, \quad |x| < 1.$$

□

Two Taylor series with the same center can trivially be added or subtracted in order to obtain a Taylor series for the sum or difference of the corresponding functions. The radius of convergence of the sum or difference is the minimum of the radii of convergence of the two series that are being added or subtracted.

It is also possible to multiply or divide Taylor series to obtain new Taylor series, as the following examples show.

Example Recall the Maclaurin series for e^x and $\cos x$:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots,$$

$$\cos x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots.$$

The Taylor series for $e^x \cos x$ can be obtained by multiplying these series, in the same way that polynomials are multiplied. This is because for each non-negative integer n, the number of terms from each series that are needed to compute the coefficient c_n in the product is finite.

We have

$$e^x \cos x = \left(\sum_{n=0}^{\infty} \frac{x^n}{n!} \right) \left(\sum_{m=0}^{\infty} (-1)^m \frac{x^{2m}}{(2m)!} \right)$$

$$\begin{aligned} &= \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(-1)^m\frac{x^{2m+n}}{n!(2m)!} \\ &= \sum_{\ell=0}^{\infty}\sum_{m=0}^{\lfloor \ell/2\rfloor}(-1)^m\frac{x^{\ell}}{(\ell-2m)!(2m)!} \\ &= 1+x-\frac{1}{3}x^3-\frac{1}{6}x^4-\frac{1}{30}x^5+\cdots, \end{aligned}$$

where $\ell = 2m+n$, and, for any real number x, $\lfloor x \rfloor$ is the "floor" of x, which is the greatest integer that is less than or equal to x. This series converges for all x, like the series for e^x and $\cos x$. □

Example Recall the Maclaurin series for $\sin x$ and $\cos x$:

$$\sin x = \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{(2n+1)!} = x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\cdots,$$

$$\cos x = \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n}}{(2n)!} = 1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+\cdots.$$

To compute the Maclaurin series for $\tan x = \sin x/\cos x$, we can use polynomial division to divide the Maclaurin series for $\sin x$ by that of $\cos x$, even though these series are not polynomials themselves, but sums of infinitely many monomials.

First, we divide the leading terms of the two series. The first term of the series for $\sin x$ is x, while the first term of the series for $\cos x$ is 1. Therefore, the first term of the quotient of the series, and therefore the series for $\tan x$, is $x/1 = x$.

Next, we subtract x times the series for $\cos x$ from that of $\sin x$ to obtain the remainder:

$$\sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{(2n+1)!}-\sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{(2n)!} = \sum_{n=0}^{\infty}(-1)^{n+1}\frac{2n}{(2n+1)!}x^{2n+1}.$$

Then, we divide the leading term of *this* series, which is $x^3/3$, by the leading term of the series for $\cos x$, which is 1, to obtain the second term of the quotient, which is $x^3/3$.

Continuing this process yields the series

$$\tan x = x+\frac{1}{3}x^3+\frac{2}{15}x^5+\frac{17}{315}x^7+\frac{62}{2835}x^9+\cdots$$

The coefficients arise from the following pattern:

$$\begin{aligned}
\frac{1}{3} &= -\frac{1}{3!} + \frac{1}{2!}, \\
\frac{2}{15} &= \frac{1}{5!} - \frac{1}{4!} + \frac{1}{3}\frac{1}{2!}, \\
\frac{17}{315} &= -\frac{1}{7!} + \frac{1}{6!} - \frac{1}{3}\frac{1}{4!} + \frac{2}{15}\frac{1}{2!}, \\
\frac{62}{2835} &= \frac{1}{9!} - \frac{1}{8!} + \frac{1}{3}\frac{1}{6!} - \frac{2}{15}\frac{1}{4!} + \frac{17}{315}\frac{1}{2!}.
\end{aligned}$$

Note that each coefficient is expressed in terms of the previous coefficients. □

1.9.1 Summary

- If a function $f(x)$ can be represented by a power series centered at x_0, then the coefficients are given by $c_n = f^{(n)}(x_0)/n!$, where $f^{(n)}(x_0)$ is the nth derivative of f evaluated at x_0.

- The power series centered at x_0 that represents a function f is called the *Taylor series of f centered at x_0*. If $x_0 = 0$, the Taylor series is also called the *Maclaurin series of f*.

- If the Taylor series of f is truncated after the first $(n + 1)$ terms, the result is a polynomial of degree n, called the nth-degree *Taylor polynomial of f centered at x_0*.

- The difference between $f(x)$ and its nth-degree Taylor polynomial $T_n(x)$, centered at x_0, is called the *Taylor remainder of f*. It closely resembles the term of degree $n+1$ in the Taylor series, except that the $(n + 1)$-st derivative of f is evaluated at an unknown point between x_0 and x. By bounding this derivative between x_0 and x, one can estimate the error in approximating $f(x)$ by $T_n(x)$.

- One approach to computing the Taylor series of a function $f(x)$, centered at x_0, is to compute the first few coefficients, given by $c_n = f^{(n)}(x_0)/n!$, and trying to recognize a pattern that applies for all nonnegative integers n.

- This approach leads to the following known Maclaurin series:

–
$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

–
$$\sin x = x - \frac{x^3}{6} + \frac{x^5}{120} - \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}$$

–
$$\cos x = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!}$$

–
$$(1+x)^k = 1 + kx + k(k-1)x^2 + \cdots = \sum_{n=0}^{\infty} \left(\begin{array}{c} k \\ n \end{array} \right) x^k$$

All of these series converge for all x, except for the binomial series for $(1+x)^k$, which converges for $|x| < 1$.

- Given a Maclaurin series for a function $f(x)$, a Maclaurin series for $f(x^p)$, where p is a positive integer, can be obtained by substitution of x^p for x in the series for $f(x)$.

- A Taylor series for $f(x)$ can be differentiated term-by-term to obtain a Taylor series for $f'(x)$. The center and radius of convergence of the series for $f'(x)$ are the same as that of $f(x)$.

- A Taylor series for $f(x)$ can be anti-differentiated term-by-term to obtain a Taylor series for $\int^x f(s)\,ds$. The center and radius of convergence of the series for $\int^x f(s)\,ds$ are the same as that of $f(x)$.

- A Taylor series for $f(x)$ can be integrated term-by-term from a to b to compute the integral from a to b of $f(x)$, provided that (a, b) lies with in the interval of convergence of the series.

- Two Taylor series, for functions $f(x)$ and $g(x)$, with the same center x_0 and radius of convergence R, can be multiplied to obtain a Taylor series for $f(x)g(x)$ whose radius of convergence is also R.

- Two Taylor series, for functions $f(x)$ and $g(x)$, with the same center x_0, can be divided using polynomial long division to obtain a Taylor series for $f(x)/g(x)$, provided that the leading coefficient c_0 of the series for $g(x)$ is nonzero.

1.10 Review

You should now be able to complete the following types of problems:

- Computing a formula for the terms of a sequence, given the first few terms of the sequence. This involves detecting patterns such as a^n for some constant a, or n^p for some integer p, or $(-1)^n$ for terms that oscillate.

- Determining whether a sequence converges or diverges, and, if convergent, computing its limit. This may require techniques such as dividing the numerator and denominator by the highest power of n, multiplying and dividing by the conjugate of an expression involving square roots, or using l'Hospital's Rule. An understanding of what kinds of functions grow faster than others can be helpful.

- Computing the limit of a recursively defined sequence $\{a_n\}$, in which a_{n+1} is defined in terms of a_n. You should also be able to show that such a sequence is convergent using the Monotonic Sequence Theorem, which applies if the sequence is monotonic and bounded.

- Determining whether a geometric series $\sum ar^n$ converges or diverges, by computing r and checking whether $|r| < 1$, and if it converges, computing its limit $a/(1-r)$.

- Determining whether a series diverges using the Divergence Test. A series diverges if its terms do not converge to zero.

- Determining whether a telescoping series converges, and if it does, computing its limit. This may require partial fraction decomposition.

- Determining whether a series converges or diverges using the Integral Test, Comparison Test, or the Limit Comparison Test. Note that these tests should only be applied to series whose terms are positive and decreasing.

- Determining whether an alternating series converges, using the Alternating Series Test, and if it converges, computing an error estimate using the Alternating Series Estimation Theorem.

- Determining whether a series converges or diverges using the Ratio Test or the Root Test.

- Computing a Taylor series of a given function $f(x)$ around a given center x_0, using the general formula for a Taylor series,

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n.$$

- Using a Taylor series of a given function $f(x)$, at a given center x_0, to compute the value of a function to a given accuracy. To save time, you can use the Alternating Series Estimation Theorem to determine when sufficient accuracy has been achieved. This theorem states that if s is the sum of an alternating series

$$s = \sum_{n=0}^{\infty}(-1)^n b_n,$$

  and s_n is the nth partial sum, then

$$|s - s_n| \leq |b_{n+1}|.$$

  That is, the error in a partial sum is no larger than the first excluded term. If the series is not alternating, then the Taylor remainder can be used instead.

- Computing the Taylor series of a function by differentiating or integrating a known (or more easily obtained) Taylor series term-by-term.

- Computing a definite or indefinite integral by representing the integrand as a Taylor series, and integrating term-by-term. For a definite integral, you will be asked to compute the integral to a given accuracy.

- Using Taylor series to compute the sum of a given infinite series, by recognizing the series as the evaluation of a known Taylor series at a specific value of x.

When trying to determine which test should be applied to a series, it is advisable to use a thought process such as the following:

- Do the terms even converge to zero? If not, you can simply apply the Divergence Test to conclude that the series diverges.

- Does the series fit the form of a geometric series $\sum ar^n$?

- Is it a telescoping series? Such a series may have terms that are expressed as a single fraction, which can then be rewritten as a difference of two fractions using partial fraction decomposition.
- Do the terms of the series alternate in sign, due to a factor such as $(-1)^n$? In that case, the Alternating Series Test may apply.
- If none of the above tests apply, consider the Ratio Test or the Root Test. The Root test is particularly useful if each term a_n of the series includes an expression that is raised to the nth power. If either the Ratio Test or Root Test is inconclusive, you should not attempt the other test, for it will be inconclusive as well.
- Next, consider the Integral Test, if the terms of a series describe a function that can readily be integrated.
- Next, consider the Comparison Test, unless the terms are not clearly decreasing, in which case the Limit Comparison Test may be more helpful.

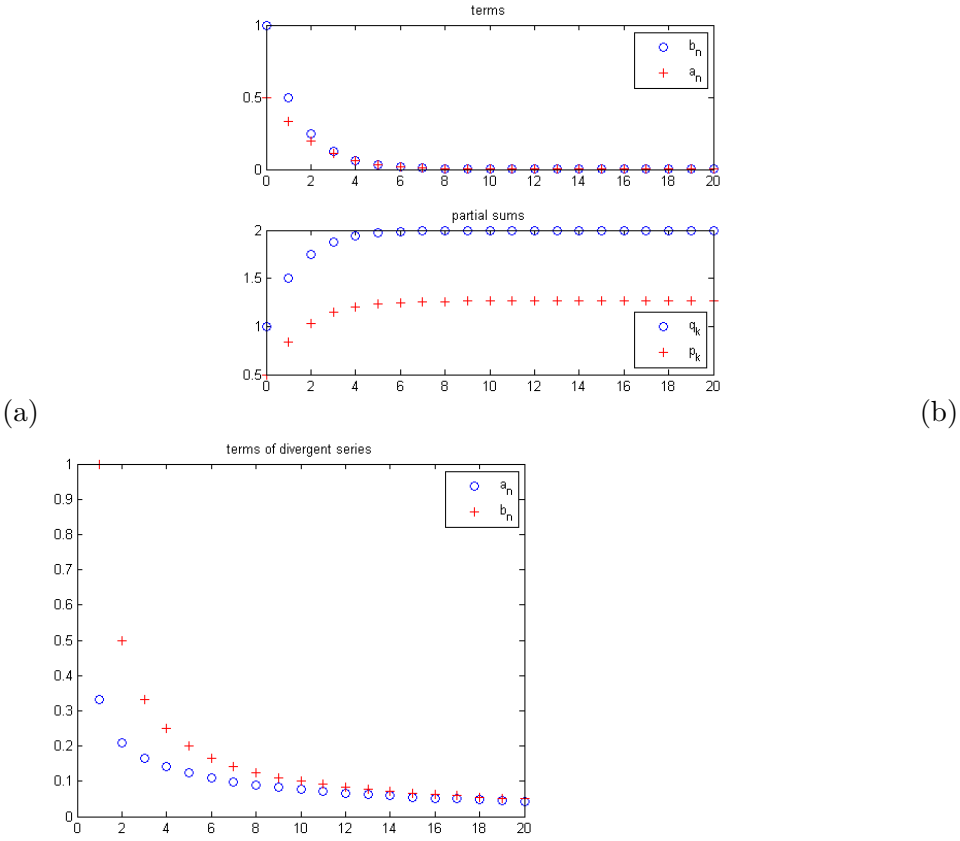

Figure 1.6: (a) Top plot: terms of convergent series $\sum a_n$ (crosses) and $\sum b_n$ (circles), for which $b_n \geq a_n \geq 0$. Bottom plot: partial sums of $\sum a_n$ (crosses) and $\sum b_n$ (circles). (b) Terms of two divergent series $\sum a_n$ (circles) and $\sum b_n$ (crosses) that have a ratio of 1 as $n \to \infty$.

Chapter 2

Vectors and the Geometry of Space

In this chapter, we will learn how to work with locations and directions in three-dimensional space, in order to easily describe objects such as lines, planes and curves. This will set the stage for the study of functions of two variables, the graphs of which are surfaces in space.

2.1 Three-Dimensional Coordinate Systems

2.1.1 Points in Three-Dimensional Space

Previously, we have identified a point in the xy-plane by an ordered pair that consists of two real numbers, an x-coordinate and y-coordinate, which denote signed distances along the x-axis and y-axis, respectively, from the origin, which is the point $(0, 0)$. These axes, which are collectively referred to as the coordinate axes, divided the plane into four quadrants.

We now generalize these concepts to three-dimensional space, or xyz-space. In this space, a point is represented by an *ordered triple* (x, y, z) that consists of three numbers, an x-coordiante, a y-coordinate, and a z-coordinate. As in the two-dimensional xy-plane, these coordinates indicate the signed distance along the *coordinate axes*, the x-axis, y-axis and z-axis, respectively, from the origin, denoted by O, which has coordinates $(0, 0, 0)$. There is a one-to-one correspondence between a point in xyz-space and a triple in $\mathbb{R}^3$, which is the set of all ordered triples of real numbers. This correspondence is known as a *three-dimensional rectangular coordinate system.*

Example Figure 2.1 displays the point $(2, 3, 1)$ in xyz-space, denoted by

the letter P, along with its projections onto the coordinate planes (described below). The origin is denoted by the letter O. □

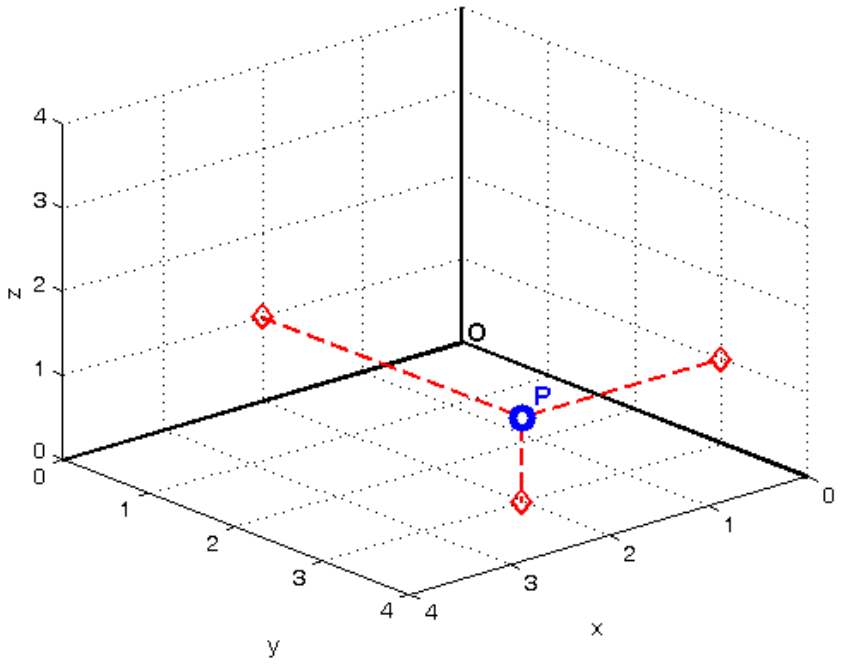

Figure 2.1: The point $(2, 3, 1)$ in xyz-space, denoted by the letter P. The origin is denoted by the letter O. The projections of P onto the coordinate planes are indicated by the diamonds. The dashed lines are line segments perpendicular to the coordinate planes that connect P to its projections.

2.1.2 Planes in Three-Dimensional Space

Unlike two-dimensional space, which consists of a single plane, the xy-plane, three-dimensional space contains infinitely many planes, just as two-dimensional space consists of infinitely many lines. Three planes are of particular importance: the xy-plane, which contains the x- and y-axes; the yz-plane, which contains the y- and z-axes; and the xz-plane, which contains the x- and z-axes.

Alternatively, the xy-plane can be described as the set of all points (x, y, z) for which $z = 0$. Similarly, the yz-plane is the set of all points

of the form $(0, y, z)$, while the xz-plane is the set of all points of the form $(x, 0, z)$.

Just as the x-axis and y-axis divide the xy-plane into four quadrants, these three planes divide xyz-space into eight *octants*. Within each octant, all x-coordiantes have the same sign, as do all y-coordinates, and all z-coordinates. In particular, the *first octant* is the octant in which all three coordinates are positive.

2.1.3 Plotting Points in xyz-space

Graphing in xyz-space can be difficult because, unlike graphing in the xy-plane, depth perception is required. To simplify plotting of points, one can make use of *projections* onto the coordinate planes. The projection of a point (x, y, z) onto the xy-plane is obtained by connecting the point to the xy-plane by a line segment that is perpendicular to the plane, and computing the intersection of the line segment with the plane.

Later, we will learn more about how to compute projections of points onto planes, but in this relatively simple case, it follows from our working definition that the projection of the point (x, y, z) onto the xy-plane is the point $(x, y, 0)$. Similarly, the projection of this point onto the yz-plane is the point $(0, y, z)$, and the projection of this point onto the xz-plane is the point $(x, 0, z)$. Figure 2.1 illustrates these projections, and how they can be used to plot a point in xyz-space. One can first plot the point's projections, which is similar to the task of plotting points in the xy-plane, and then use line segments originating from these projections and perpendicular to the coordinate planes to "locate" the point in xyz-space.

2.1.4 The Distance Formula

The distance between two points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ in the xy-plane is given by the distance formula,

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Similarly, the distance between two points $P_1 = (x_1, y_1, z_1)$ and $P_2 = (x_2, y_2, z_2)$ in xyz-space is given by the following generalization of the distance formula,

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}.$$

This can be proved by repeated application of the Pythagorean Theorem.

Example The distance between $P_1 = (2, 3, 1)$ and $P_2 = (8, -5, 0)$ is

$$d(P_1, P_2) = \sqrt{(8-2)^2 + (-5-3)^2 + (0-1)^2} = \sqrt{36 + 64 + 1} = \sqrt{101} \approx 10.05.$$

□

2.1.5 Equations of Surfaces

In two dimensions, the solution set of a single equation involving the coordinates x and/or y is a curve. In three dimensions, the solution set of an equation involving x, y and/or z is a surface.

Example The equation $z = 3$ describes a plane that is parallel to the xy-plane, and is 3 units "above" it; that is, it lies 3 units along the positive z-axis from the xy-plane. On the other hand, the equation $x = y$ describes a plane consisting of all points whose x- and y-coordinates are equal. It is not parallel to any coordinate plane, but it contains the z-axis, which consists of all points whose x- and y-coordinates are both zero, and it intersects the xy-plane at the line $y = x$. □

The equation of a sphere with center $C = (h, k, \ell)$ and radius r is

$$(x-h)^2 + (y-k)^2 + (z-\ell)^2 = r^2.$$

The *unit sphere* has center $O = (0, 0, 0)$ and radius 1:

$$x^2 + y^2 + z^2 = 1.$$

We now illustrate how to work with equations of spheres.

Example The equation of a sphere with center $C = (-3, -1, 1)$ and radius $r = 10$ is

$$(x-(-3))^2 + (y-(-1))^2 + (z-1)^2 = 10^2,$$

or

$$(x+3)^2 + (y+1)^2 + (z-1)^2 = 100.$$

Expanding, we obtain

$$x^2 + y^2 + z^2 + 6x + 2y - 2z = 89,$$

which obscures the center and radius, but it is still possible to detect that the equation represents a sphere, due to the fact that the x^2, y^2 and z^2 terms have equal coefficients. □

Example The equation

$$4x^2 + 4y^2 + 4z^2 - 8x - 16y - 16 = 0$$

describes a sphere, as can be seen by the equal coefficients in front of the x^2, y^2 and z^2. To determine the radius and center of the sphere, we complete the square in x and y:

$$\begin{aligned} 0 &= 4x^2 + 4y^2 + 4z^2 - 8x - 16y - 16 \\ &= 4(x^2 - 2x) + 4(y^2 - 4y) + 4z^2 - 16 \\ &= 4(x^2 - 2x + 1 - 1) + 4(y^2 - 4y + 4 - 4) + 4z^2 - 16 \\ &= 4[(x-1)^2 - 1] + 4[(y-2)^2 - 4] + 4z^2 - 16 \\ &= 4(x-1)^2 + 4(y-2)^2 + 4z^2 - 36. \end{aligned}$$

Rearranging, we obtain the standard form of the equation of the sphere:

$$(x-1)^2 + (y-2)^2 + z^2 = 9,$$

which reveals that the center is at the point $C = (1, 2, 0)$, and the radius is $r = 3$. □

Example The region consisting of all points that lie between the spheres centered at $(-1, 1, 2)$, with radii 3 and 5, can be described by the inequalities

$$9 < (x+1)^2 + (y-1)^2 + (z-2)^2 < 25.$$

The points that lie on these spheres are excluded by these inequalities. To include them, $\leq$ should be used instead of $<$. These inequalities use the fact that the equation of the sphere with center $(-1, 1, 2)$ and radius r is

$$(x+1)^2 + (y-1)^2 + (z-2)^2 = r^2.$$

□

Example The inequality

$$x^2 + y^2 + z^2 > 4x$$

describes the set of all points *outside* the sphere with center $(2, 0, 0)$ and radius 2. To see this, we rewrite the inequality as

$$x^2 - 4x + y^2 + z^2 > 0$$

and then complete the square to obtain

$$x^2 - 4x + 4 - 4 + y^2 + z^2 > 0.$$

Factoring the perfect square that has been completed, and then rearranging, yields

$$(x-2)^2 + y^2 + z^2 > 4.$$

This inequality, when changed to an equality, describes the sphere with center $(2, 0, 0)$ and radius 2. The inequality prescribes that the distance between (x, y, z) and $(2, 0, 0)$ be *greater* than 2, so (x, y, z) must lie outside this sphere. □

2.1.6 Summary

- The three-dimensional rectangular coordinate system is the one-to-one correspondence between each point P in three-dimensional space, or xyz-space, and an ordered triple (x, y, z) in $\mathbb{R}^3$. The numbers x, y and z are the x-, y- and z-coordinates of P. The origin O is the point with coordinates $(0, 0, 0)$.

- The coordinate planes are: the xy-plane, the set of all points whose z-coordinate is zero; the yz-plane, the set of all points whose x-coordinate is zero; and the xz-plane, the set of all points whose y-coordinate is zero.

- The projection of a point $P = (x, y, z)$ onto the xy-plane is the point $(x, y, 0)$. The projection of P onto the yz-plane is the point $(0, y, z)$. The projection of P onto the xz-plane is the point $(x, 0, z)$.

- The distance formula states that the distance between two points in xyz-space is the square root of the sum of the squares of the differences between corresponding coordinates. That is, given $P_1 = (x_1, y_1, z_1)$ and $P_2 = (x_2, y_2, z_2)$, the distance between P_1 and P_2 is given by $d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$.

- The equation of a sphere with center $C = (x_0, y_0, z_0)$ and radius r is $(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = r^2$.

- An equation in which x^2, y^2 and z^2 have the same coefficients describes a sphere; the center and radius can be determined by completing the square in x, y and z.

2.2 Vectors

When an object is displaced, it is moved a certain distance, and also in some direction. Both the distance and direction are required in order to fully describe the object's motion. Similarly, an object's velocity incorporates the rate at which it is traveling, and its direction of travel. A *vector* refers to a quantity, such as displacement or velocity, that has, as properties, both magnitude and direction.

Visually, a vector is represented by an arrow. The length of the arrow indicates the magnitude of the vector, and the direction of the arrow is the direction of the vector. The point at the tail of the arrow is called the *initial point* of the vector, and the tip of the arrow is called the *terminal point*. A typical vector is shown in Figure 2.2.

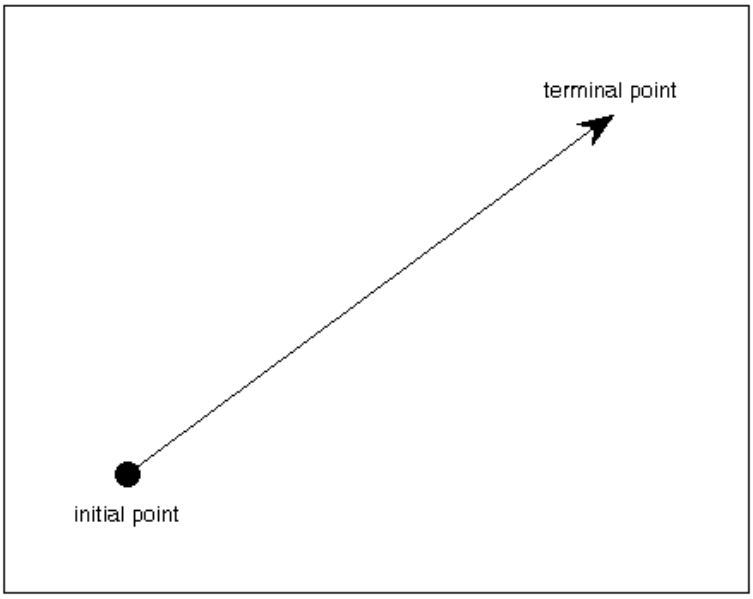

Figure 2.2: A representation of a vector by an arrow, which indicates the length and direction of the vector.

Two arrows that are pointing in the same direction, and the same length,

but are located at different positions, represent vectors that are indistinguishable from one another, because they have the same magnitude and direction. We say that these vectors are *equivalent*, or *equal*.

We will normally denote vectors by bold lowercase letters, such as $\mathbf{u}$ or $\mathbf{v}$. A vector that is of particular importance is the *zero vector*, denoted by $\mathbf{0}$. It is a vector whose magnitude, or length, is zero. As such, it does not have any particular direction.

2.2.1 Combining Vectors

Let $\mathbf{u}$ and $\mathbf{v}$ be two vectors that are positioned in such a way that the initial point of $\mathbf{v}$ coincides with the terminal point of $\mathbf{u}$. The *sum* of $\mathbf{u}$ and $\mathbf{v}$ is the vector $\mathbf{u}+\mathbf{v}$ whose initial point is the initial point of $\mathbf{u}$, and whose terminal point is the terminal point of $\mathbf{v}$.

The vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{u}+\mathbf{v}$ form a triangle, which is why this definition of the sum of two vectors is called the *Triangle Law*. This law is illustrated in Figure 2.3(a). However, if the two vectors are added in reverse order, the sum $\mathbf{v}+\mathbf{u}$ turns out to be equal to $\mathbf{u}+\mathbf{v}$. That is, addition of vectors is *commutative*. To see this, we note that if the initial point of $\mathbf{u}$ in the sum $\mathbf{u}+\mathbf{v}$ coincides with the initial point of $\mathbf{v}$ in the sum $\mathbf{v}+\mathbf{u}$, the two copies of $\mathbf{u}$ and $\mathbf{v}$ form a parallelogram, with the two sums coinciding with one of its diagonals. Therefore, commutativity of vector addition is called the *Parallelogram Law*, which is illustrated in Figure 2.3(b).

Another important operation that can be performed on vectors is called *scalar multiplication*. This operation entails scaling the magnitude of a vector $\mathbf{u}$ by a number s, which, in this context, is called a *scalar*. The direction of the resulting vector, denoted by $s\mathbf{u}$ and called a *scalar multiple* of $\mathbf{u}$, is the same as that of $\mathbf{u}$ if $s > 0$, and the opposite of that of $\mathbf{u}$ if $s < 0$. If $s = 0$, $s\mathbf{u} = \mathbf{0}$, the zero vector.

Two vectors that have the same direction are said to be parallel. It follows from the definition of scalar multiplication that two vectors are parallel if and only if one is a scalar multiple of the other. One scalar multiple of particular importance is the *negative* of a vector. Given a vector $\mathbf{u}$, its negative is the vector $-\mathbf{u}$ obtained by scaling $\mathbf{u}$ by -1. Its direction is the opposite of that of $\mathbf{u}$, but its magnitude is the same.

Example Figure 2.4 displays a vector $\mathbf{u}$ and some examples of scalar multiples of $\mathbf{u}$, including $2\mathbf{u}$, $-\mathbf{u}$, and $-3\mathbf{u}$. □

The negative is used to define the operation of *vector subtraction*. Given two vectors $\mathbf{u}$ and $\mathbf{v}$, we define $\mathbf{u}-\mathbf{v}$ to be the sum $\mathbf{u}+(-\mathbf{v})$. This operation

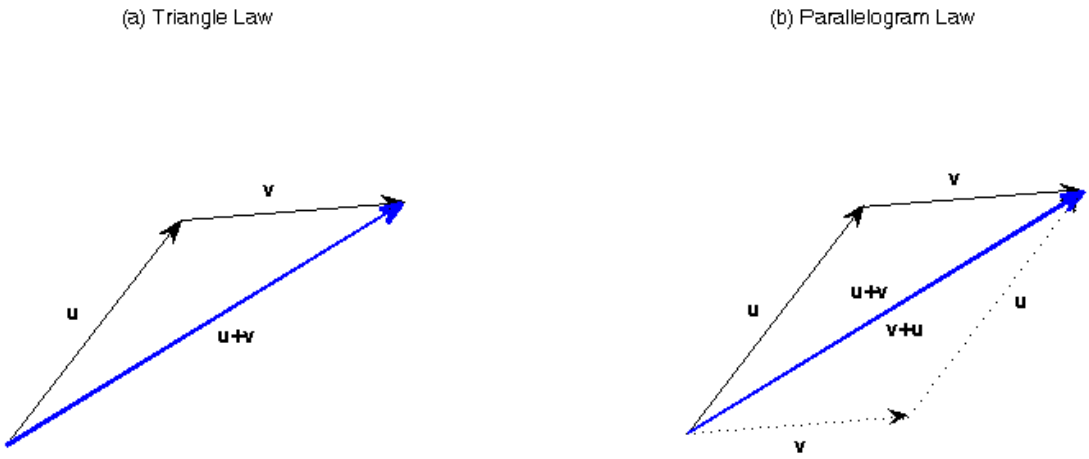

Figure 2.3: (a) The Triangle Law. (b) The Parallelogram Law.

gives us another way of characterizing equivalent vectors: two vectors are equivalent, or equal, if and only if their difference is the zero vector: $\mathbf{u}-\mathbf{v}=\mathbf{0}$ implies $\mathbf{u}=\mathbf{v}$.

2.2.2 Components

So far, we have defined vectors, and operations on them, geometrically, in terms of vectors' magnitude and direction. However, in many situations, it is easier to treat vectors algebraically. This requires the introduction of a coordinate system. In the three-dimensional rectangular coordinate system, we identify a vector $\mathbf{u}$ with a point in $\mathbb{R}^3$ by positioning the initial point of $\mathbf{u}$ at the origin, and defining the *components* of $\mathbf{u}$ to be the coordinates (u_1, u_2, u_3) of its terminal point. To avoid confusing a vector $\mathbf{u}$ with the point (u_1, u_2, u_3), we denote its components by $\langle u_1, u_2, u_3\rangle$.

If two vectors are equivalent, then they have the same components. Although they may have different initial points and terminal points, the difference between the coordinates of the terminal point and those of the initial

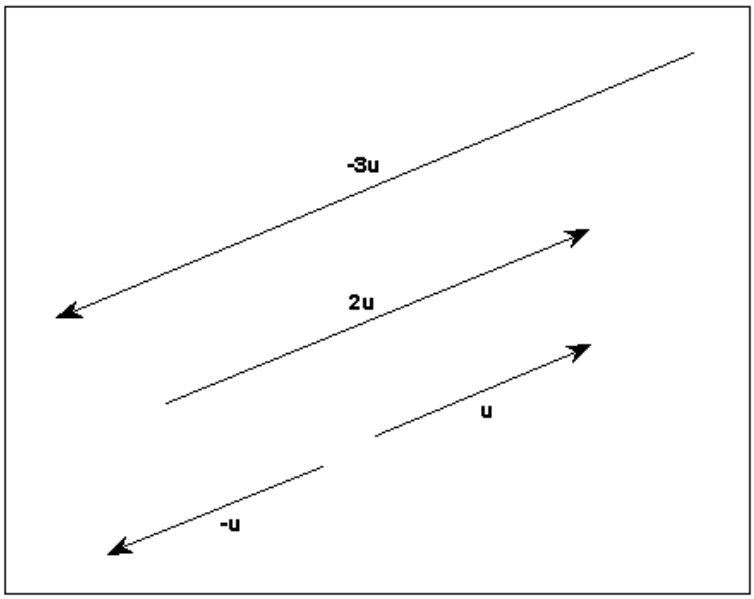

Figure 2.4: A vector **u** and various scalar multiples.

point are always equal to these components. That is, if a vector **v** has initial point $P_1 = (x_1, y_1, z_1)$ and terminal point $P_2 = (x_2, y_2, z_2)$, then the components of the vector **v** are

$$\mathbf{v} = \langle x_2 - x_1, y_2 - y_1, z_2 - z_1 \rangle.$$

If the initial point is the origin $O = (0, 0, 0)$, we refer to the vector as the *position vector* of the terminal point, since the components of the vector are the same as the coordinates of the terminal point.

Example Let **u** be a vector with initial point $U_1 = (0, 0, 0)$ and terminal point $U_2 = (2, 3, 4)$. Let **v** be a vector with initial point $V_1 = (1, 2, -1)$ and terminal point $V_2 = (3, 6, 3)$. These vectors are equivalent, and both have components $\langle 2, 3, 4 \rangle$. □

The *magnitude*, or *length*, of a vector $\mathbf{v} = \langle v_1, v_2, v_3 \rangle$, denoted by $|\mathbf{v}|$, is the distance between its initial and terminal points. It follows from the

definiton of the vector's components that

$$|\mathbf{v}| = \sqrt{v_1^2 + v_2^2 + v_3^2}.$$

Example Let $\mathbf{v}$ be a vector with initial point $P_1 = (-3, 4, -1)$ and terminal point $P_2 = (1, 5, 0)$. Then

$$|\mathbf{v}| = \sqrt{(1-(-3))^2 + (5-4)^2 + (0-(-1))^2} = \sqrt{18} \approx 4.24.$$

□

We have learned what it means to add, scale and subtract vectors, but only in a geometric sense. It is essential to be able to understand these operations in terms of the components of vectors.

To that end, let $\mathbf{u} = \langle u_1, u_2, u_3 \rangle$ and $\mathbf{v} = \langle v_1, v_2, v_3 \rangle$ be vectors. Then the components of the sum $\mathbf{u} + \mathbf{v}$ are given by

$$\mathbf{u} + \mathbf{v} = \langle u_1 + v_1, u_2 + v_2, u_3 + v_3 \rangle.$$

We see that operating on the vectors involves applying the same operation to corresponding components.

Similarly, given a scalar s, the vector $s\mathbf{u}$ is given by

$$s\mathbf{u} = \langle su_1, su_2, su_3 \rangle.$$

It follows that

$$\mathbf{u} - \mathbf{v} = \langle u_1 - v_1, u_2 - v_2, u_3 - v_3 \rangle.$$

Example Let $\mathbf{u} = \langle 7, -8, 9 \rangle$ and $\mathbf{v} = \langle -3, 4, -6 \rangle$. Then

$$\mathbf{u} + \mathbf{v} = \langle 7 + (-3), -8 + 4, 9 + (-6) \rangle = \langle 4, -4, 3 \rangle,$$

$$\mathbf{u} - \mathbf{v} = \langle 7 - (-3), -8 - 4, 9 - (-6) \rangle = \langle 10, -12, 15 \rangle,$$

$$2\mathbf{u} = \langle 2(7), 2(-8), 2(9) \rangle = \langle 14, -16, 18 \rangle,$$

and

$$3\mathbf{u} - 2\mathbf{v} = \langle 3(7) - 2(-3), 3(-8) - 2(4), 3(9) - 2(-6) \rangle = \langle 27, -30, 39 \rangle.$$

□

We have defined these operations for vectors in three dimensions, but they generalize in a natural to way to vectors of any length. We define V_3

to be the set of all vectors in three dimensions, and V_2 to be the set of all vectors in two dimensions.

More generally, we define V_n to be the set of all vectors in n-dimensional space. Each vector $\mathbf{v}$ in V_n is identified with an ordered n-tuple $(v_1, v_2, \ldots, v_n)$ in $\mathbb{R}^n$, where $v_1, v_2, \ldots, v_n$ are the components of $\mathbf{v}$. As in the three-dimensional case, addition of vectors in $\mathbf{v}_n$ is performed by adding corresponding components, and scalar multiplication of vectors is performed by scaling components.

A set of vectors, together with the operations of addition and scalar multiplication defines a *vector space*. A vector space has the following properties. For any vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$, and for any scalars c and d,

1. Commutativity: $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$
2. Associativity: $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$
3. Additive identity: $\mathbf{u} + \mathbf{0} = \mathbf{u}$
4. Additive inverse: $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$
5. Distributive property: $c(\mathbf{u} + \mathbf{v}) = c\mathbf{u} + c\mathbf{v}$
6. Distributive property: $(c + d)\mathbf{u} = c\mathbf{u} + d\mathbf{u}$
7. Associativity: $(cd)\mathbf{u} = c(d\mathbf{u})$
8. Multiplicative identity: $1\mathbf{u} = \mathbf{u}$

Another essential property of vector spaces is that they are *closed* under the operations of vector addition and scalar multiplication. That is, if $\mathbf{u}$ and $\mathbf{v}$ are vectors in V_n, and c is a scalar, then $\mathbf{u} + \mathbf{v}$ and $c\mathbf{u}$ are also vectors in V_n.

In V_3, three vectors that are of particular importance are the *standard basis vectors*

$$\mathbf{i} = \langle 1, 0, 0 \rangle, \quad \mathbf{j} = \langle 0, 1, 0 \rangle, \quad \mathbf{k} = \langle 0, 0, 1 \rangle.$$

Each of these vectors has length 1, and point in the direction of one of the coordinate axes. They are illustrated in Figure 2.5. In two dimensions, there are two standard basis vectors:

$$\mathbf{i} = \langle 1, 0 \rangle, \quad \mathbf{j} = \langle 0, 1 \rangle.$$

In general, V_n has n standard basis vectors, usually denoted by $e_{\mathbf{i}}$, for $i = 1, 2, \ldots, n$, where the ith component of $\mathbf{e}_i$ is equal to 1, and all other components are zero.

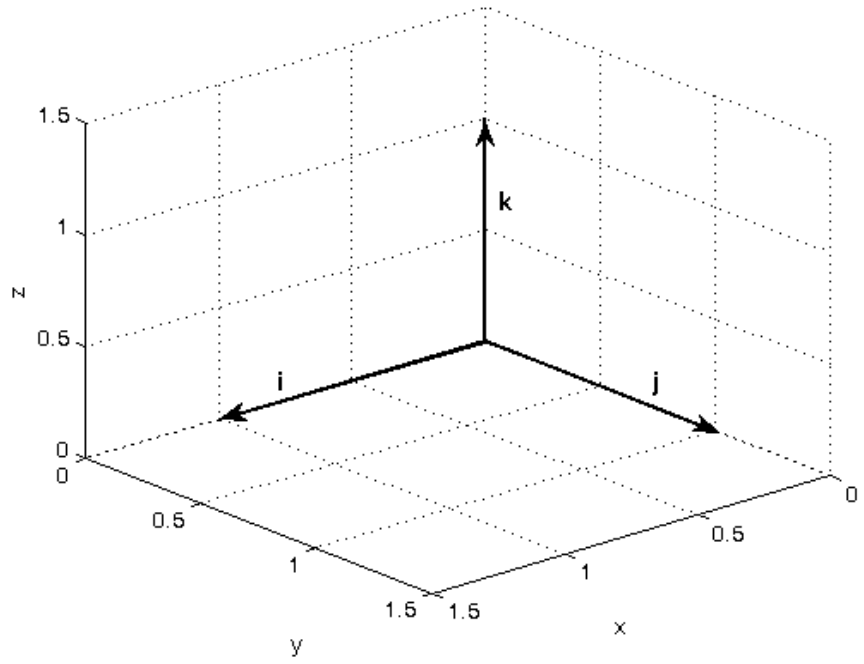

Figure 2.5: The standard basis vectors **i**, **j** and **k**

.

In general, a *basis* of a vector space V is a set of vectors that has as few elements as possible, but has the property that *every* vector in V can be expressed as a linear combination of members of the basis. That is, every vector can be obtained from the vectors in the basis using addition and scalar multiplication. In V_3, we have, for any vector $\mathbf{v} = \langle v_1, v_2, v_3 \rangle$,

$$\begin{aligned} \mathbf{v} &= \langle v_1, v_2, v_3 \rangle \\ &= \langle v_1, 0, 0 \rangle + \langle 0, v_2, 0 \rangle + v_3 \langle 0, 0, v_3 \rangle \\ &= v_1 \langle 1, 0, 0 \rangle + v_2 \langle 0, 1, 0 \rangle + v_3 \langle 0, 0, 1 \rangle \\ &= v_1 \mathbf{i} + v_2 \mathbf{j} + v_3 \mathbf{k}. \end{aligned}$$

Similarly, for any vector $\mathbf{v} = \langle v_1, v_2 \rangle$ in V_2, we can write

$$\mathbf{v} = v_1 \mathbf{i} + v_2 \mathbf{j}.$$

Example Let $\mathbf{u} = \langle 2, -1, 3\rangle$ and $\mathbf{v} = \langle -4, 5, 0\rangle$. Then we can write

$$\mathbf{u} = 2\mathbf{i} - \mathbf{j} + 3\mathbf{k}, \quad \mathbf{v} = -4\mathbf{i} + 5\mathbf{j},$$

and then

$$\mathbf{u} + \mathbf{v} = 2\mathbf{i} - \mathbf{j} + 3\mathbf{k} - 4\mathbf{i} + 5\mathbf{j} = (2-4)\mathbf{i} + (-1+5)\mathbf{j} + 3\mathbf{k} = -2\mathbf{i} + 4\mathbf{j} + 3\mathbf{k}.$$

□

A vector has both magnitude and direction, but often we are only interested in the direction of a nonzero vector $\mathbf{v}$. It is helpful to "filter out" the magnitude by *normalizing* it so that its magnitude is 1. This is accomplished by dividing by its magnitude, which yields a **unit vector u** that has the same direction of $\mathbf{v}$. Specifically,

$$\mathbf{u} = \frac{1}{|\mathbf{v}|}\mathbf{v}.$$

Then

$$|\mathbf{u}| = \left|\frac{1}{|\mathbf{v}|}\mathbf{v}\right| = \frac{|\mathbf{v}|}{|\mathbf{v}|} = 1.$$

Example Let $\mathbf{v} = \langle 3, -4, 5\rangle$. Then

$$|\mathbf{v}| = \sqrt{3^2 + (-4)^2 + 5^2} = \sqrt{50} = 5\sqrt{2},$$

and the unit vector $\mathbf{u}$ with the same direction as $\mathbf{v}$ is

$$\mathbf{u} = \frac{\sqrt{2}}{10}\langle 3, -4, 5\rangle = \left\langle \frac{3\sqrt{2}}{10}, -\frac{2\sqrt{2}}{5}, \frac{\sqrt{2}}{2} \right\rangle.$$

□

Example Consider the vector $\mathbf{w} = \langle 2, 4, -3\rangle$. We wish to compute a vector $\mathbf{z}$ that is in the opposite direction of $\mathbf{w}$, and has magnitude 5. First, we need to identify the magnitude and direction of $\mathbf{w}$. The magnitude is given by

$$|\mathbf{w}| = \sqrt{2^2 + 4^2 + (-3)^2} = \sqrt{29}.$$

Then, a unit vector in the same direction of $\mathbf{w}$ is

$$\mathbf{u} = \left\langle \frac{2}{\sqrt{29}}, \frac{4}{\sqrt{29}}, -\frac{3}{\sqrt{29}} \right\rangle.$$

It follows that a unit vector in the opposite direction of $\mathbf{w}$ is

$$-\mathbf{u} = \left\langle -\frac{2}{\sqrt{29}}, -\frac{4}{\sqrt{29}}, \frac{3}{\sqrt{29}} \right\rangle.$$

We conclude that a vector in the opposite direction of $\mathbf{w}$, with magnitude 5, is

$$\mathbf{z} = -5\mathbf{u} = -\frac{5}{|\mathbf{w}|}\mathbf{w} = \left\langle -\frac{10}{\sqrt{29}}, -\frac{20}{\sqrt{29}}, \frac{15}{\sqrt{29}} \right\rangle.$$

□

Example Let $\mathbf{v}$ be a vector in the xy-plane with initial point $(0, 0)$, and magnitude $\sqrt{2}$, that makes an angle of $3\pi/4$ with the positive x-axis. To find the components of $\mathbf{v}$, we note that these components are equal to the coordinates of its terminal point, because its initial point is at the origin. These coordinates are obtained from the magnitude and angle with the positive x-axis as follows:

$$\mathbf{v} = \sqrt{2}\left\langle \cos\frac{3\pi}{4}, \sin\frac{3\pi}{4} \right\rangle = \sqrt{2}\left\langle \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2} \right\rangle = \langle 1, 1\rangle.$$

□

Example Suppose that a plane is steered northeast, at an *air speed* of 400 mph, while the wind is blowing westward at 100 mph. To compute the *ground speed* of the plane, which accounts for both the steering and the wind, we need to add the velocity vectors corresponding to the plane and the wind. This yields a *resultant* velocity vector, whose magnitude is the ground speed.

To perform the addition of the velocity vectors, we need their components. For the plane's vector, the magnitude is 400, and its direction is identified by an angle of $\pi/4$ radians with the positive x-axis, corresponding to northeast. It follows that the components of this vector are

$$\mathbf{v} = 400\left\langle \cos\frac{\pi}{4}, \sin\frac{\pi}{4} \right\rangle = 400\left\langle \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2} \right\rangle = \langle 200\sqrt{2}, 200\sqrt{2}\rangle.$$

Similarly, the wind's velocity vector, which makes an angle of π radians with the positive x-axis, corresponding to west, is

$$\mathbf{w} = 100\langle \cos\pi, \sin\pi\rangle = 100\langle -1, 0\rangle = \langle -100, 0\rangle.$$

We now perform the addition to obtain the plane's true course:

$$\mathbf{c} = \mathbf{v} + \mathbf{w} = \langle 200\sqrt{2}, 200\sqrt{2}\rangle + \langle -100, 0\rangle = \langle 200\sqrt{2} - 100, 200\sqrt{2}\rangle.$$

To obtain the ground speed, we compute its magnitude, and obtain

$$|\mathbf{c}| = \sqrt{(200\sqrt{2} - 100)^2 + (200\sqrt{2})^2} = 100\sqrt{(2\sqrt{2} - 1)^2 + (2\sqrt{2})^2} = 100\sqrt{17 - 4\sqrt{2}} \approx 336.796.$$

We conclude that the ground speed of the plane is approximately 336.796 mph, which is slower than the plane's air speed, because the wind is largely blowing against the direction in which the plane is steered. On the other hand, if the wind were to blow to the north, the ground speed of the plane would be faster than its air speed. □

Example Suppose a box is being pulled along the ground with a force of 50 newtons, applied at an angle of 60° from the horizontal. Then, the force vector **F** has a magnitude of 50, and a direction

$$\langle \cos 60^\circ, \sin 60^\circ \rangle = \left\langle \frac{1}{2}, \frac{\sqrt{3}}{2} \right\rangle.$$

It follows that **F** can be written as a sum of horizontal and vertical components,

$$\mathbf{F} = 50\left\langle \frac{1}{2}, \frac{\sqrt{3}}{2} \right\rangle = 50\left\langle \frac{1}{2}, 0 \right\rangle + 50\left\langle 0, \frac{\sqrt{3}}{2} \right\rangle = \langle 25, 0\rangle + \langle 0, 25\sqrt{3}\rangle = 25\mathbf{i} + 25\sqrt{3}\mathbf{j}.$$

□

2.2.3 Summary

- A vector is a quantity that has both a magnitude and a direction. It can be represented, geometrically, by an arrow that points in the vector's direction, and has a length equal to the vector's magnitude. The tip of the arrow is called the vector's terminal point, and the tail of the arrow is its initial point.
- The zero vector is a vector whose magnitude is zero. It is the only vector that has no specific direction.
- Geometrically, the sum of two vectors **u** and **v** is the vector $\mathbf{u} + \mathbf{v}$ obtained by placing the terminal point of **u** at the initial point of **v**, and defining the sum to be the vector with **u**'s initial point and **v**'s terminal point. This is called the Triangle Law.

- Vector addition is commutative: $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$. This is called the Parallelogram Law.
- Given a number s and vector $\mathbf{u}$, the vector $s\mathbf{u}$ is a scalar multiple of $\mathbf{u}$ and s is called a scalar. It has the same direction as $\mathbf{u}$, and its magnitude is s times the magnitude of $\mathbf{u}$.
- The negative of a vector $\mathbf{u}$, denoted by $-\mathbf{u}$, is defined by $(-1)\mathbf{u}$. It has the opposite direction of $\mathbf{u}$, and the same magnitude.
- The difference of two vectors $\mathbf{u}$ and $\mathbf{v}$, denoted by $\mathbf{u} - \mathbf{v}$, is defined to be $\mathbf{u} + (-\mathbf{v})$.
- The components u_1, u_2 and u_3 of a vector $\mathbf{u} = \langle u_1, u_2, u_3 \rangle$ are defined to be the rectangular coordinates of its terminal point (u_1, u_2, u_3), when its initial point is the origin.
- The magnitude of a vector $\mathbf{v} = \langle v_1, v_2, v_3 \rangle$ is equal to the distance between its initial point and terminal point, which is $\sqrt{v_1^2 + v_2^2 + v_3^2}$.
- The components of the sum of two vectors are the sums of corresponding components of the two vectors.
- The components of a vector multiplied by a scalar are the components multiplied by the same scalar.
- The standard basis vectors $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are vectors of magnitude 1 that point in the directions of the positive x-, y- and z-axes, respectively. Any vector in three-dimensional space can be expressed as a linear combination of these vectors.
- A unit vector is a vector of magnitude 1. A nonzero vector can be normalized by dividing it by its magnitude, to obtain a unit vector in the same direction.

2.3 The Dot Product

One of the most fundamental problems concerning vectors is that of computing the angle between two given vectors. It has numerous applications in mathematics and other sciences. In physics, it plays a role in the decomposition of forces into component forces that act in various directions. In computer science, it is useful for creating two-dimensional visualizations of three-dimensional objects. In computational mathematics, it is a vital

ingredient in algorithms for data fitting, approximation of functions, and other essential problems.

Let $\mathbf{u} = \langle u_1, u_2, u_3 \rangle$ and $\mathbf{v} = \langle v_1, v_2, v_3 \rangle$ be two vectors with a common initial point. Then $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{u} - \mathbf{v}$ form a triangle, as shown in Figure 2.6. By the Law of Cosines,

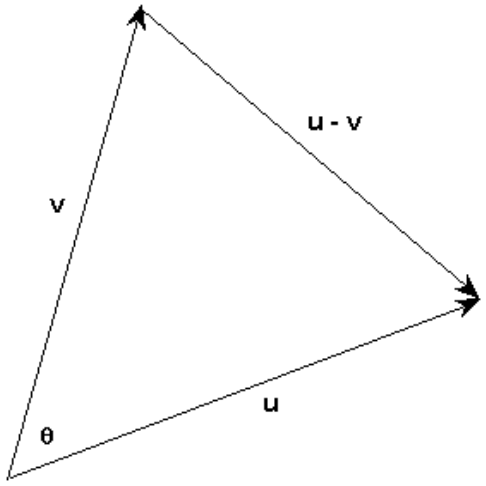

Figure 2.6: By the Triangle Law, the vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{u}-\mathbf{v}$ form a triangle. The angle between $\mathbf{u}$ and $\mathbf{v}$ is θ.

$$|\mathbf{u} - \mathbf{v}|^2 = |\mathbf{u}|^2 + |\mathbf{v}|^2 - 2|\mathbf{u}||\mathbf{v}| \cos\theta,$$

where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$. Using the formula for the magnitude of a vector, we obtain

$$(u_1-v_1)^2+(u_2-v_2)^2+(u_3-v_3)^2 = (u_1^2+u_2^2+u_3^2)+(v_1^2+v_2^2+v_3^2)-2|\mathbf{u}||\mathbf{v}| \cos\theta.$$

Simplifying yields

$$u_1v_1 + u_2v_2 + u_3v_3 = |\mathbf{u}||\mathbf{v}| \cos\theta.$$

We therefore define the *dot product*, also known as the *inner product*, of $\mathbf{u}$ and $\mathbf{v}$ to be the number $\mathbf{u} \cdot \mathbf{v}$ given by

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + u_3 v_3.$$

An equivalent definition, typically used in physics, is

$$\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos\theta,$$

where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$.

Example Let $\mathbf{u} = \langle 1, -1, 2 \rangle$ and $\mathbf{v} = \langle -2, 1, 3 \rangle$. Then

$$\mathbf{u} \cdot \mathbf{v} = 1(-2) + (-1)(1) + 2(3) = -2 - 1 + 6 = 3.$$

To obtain the angle θ between $\mathbf{u}$ and $\mathbf{v}$, we compute

$$\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|} = \frac{3}{\sqrt{6}\sqrt{14}} = \frac{3}{2\sqrt{21}},$$

which yields $\theta \approx 1.237$ radians, or 70.893 degrees. □

Example Let $\mathbf{u}$ and $\mathbf{v}$ be vectors such that $|\mathbf{u}| = 3$, $|\mathbf{v}| = 4$, and the angle between them is $\pi/3$ radians, or 60 degrees. Then

$$\mathbf{u} \cdot \mathbf{v} = 3(4)\cos\frac{\pi}{3} = 12\frac{1}{2} = 6.$$

□

2.3.1 Properties

The dot product has the following properties, which can be proved from the definition.

1. $\mathbf{u} \cdot \mathbf{u} = |\mathbf{u}|^2$
2. Commutativity: $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$
3. Distributive property: $\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$
4. $(c\mathbf{u}) \cdot \mathbf{v} = c(\mathbf{u} \cdot \mathbf{v}) = \mathbf{u} \cdot (c\mathbf{v})$, for any scalar c
5. $\mathbf{0} \cdot \mathbf{u} = 0$

Example By the first, second and third properties, the length of a sum of vectors $\mathbf{u} + \mathbf{v}$ can be expressed in terms of inner products as follows:

$$\begin{aligned} |\mathbf{u}+\mathbf{v}|^2 &= (\mathbf{u}+\mathbf{v})\cdot(\mathbf{u}+\mathbf{v}) \\ &= \mathbf{u}\cdot\mathbf{u}+\mathbf{v}\cdot\mathbf{u}+\mathbf{u}\cdot\mathbf{v}+\mathbf{v}\cdot\mathbf{v} \\ &= |\mathbf{u}|^2+2\mathbf{u}\cdot\mathbf{v}+|\mathbf{v}|^2. \end{aligned}$$

□

2.3.2 Orthogonality

Suppose that two *nonzero* vectors $\mathbf{u}$ and $\mathbf{v}$ have an angle between them that is $\theta = \pi/2$. That is, $\mathbf{u}$ and $\mathbf{v}$ are *perpendicular*, or *orthogonal*. Then, we have

$$\mathbf{u}\cdot\mathbf{v} = |\mathbf{u}||\mathbf{v}|\cos\frac{\pi}{2} = 0.$$

On the other hand, if $\mathbf{u}\cdot\mathbf{v} = 0$, then we must have $\cos\theta = 0$, where θ is the angle between them, which implies that $\theta = \pi/2$, and therefore $\mathbf{u}$ and $\mathbf{v}$ are orthogonal. In summary, $\mathbf{u}\cdot\mathbf{v} = 0$ if and only if $\mathbf{u}$ and $\mathbf{v}$ are orthogonal.

Example Let $\mathbf{u} = \langle\alpha, \beta\rangle$ be any nonzero vector in V_2. Then a vector that has the same length as $\mathbf{u}$, and is orthogonal to $\mathbf{u}$ is $\mathbf{v} = \langle\beta, -\alpha\rangle$. To verify this, we compute

$$\mathbf{u}\cdot\mathbf{v} = \langle\alpha,\beta\rangle\cdot\langle\beta,-\alpha\rangle = \alpha\beta+\beta(-\alpha) = 0.$$

By the fourth property of the dot product, $\mathbf{w} = \langle-\beta, \alpha\rangle$ also satisfies $|\mathbf{w}| = |\mathbf{u}|$, and is orthogonal to $\mathbf{u}$. □

2.3.3 Projections

When a three-dimensional object is to be displayed on a two-dimensional surface, it is necessary to determine where each point in the object appears within that surface. When a force is applied to an object from a particular direction, for example to push the object along a horizontal surface, it is helpful to decompose the force into component forces that each act from orthogonal directions, such as vertical, corresponding to gravity, and horizontal, corresponding to the direction of displacement.

These are examples of applications in which it is necessary to *project* a given vector $\mathbf{u} = \langle u_1, u_2, u_3\rangle$ onto another vector $\mathbf{v} = \langle v_1, v_2, v_3\rangle$. Projection requires solving this problem: what vector, in the direction of $\mathbf{v}$, is the *best approximation* of $\mathbf{u}$?

In other words, we wish to find the vector $\mathbf{w} = c\mathbf{v}$, where c is an unknown scalar, that minimizes $|\mathbf{u} - \mathbf{w}|$, the distance between $\mathbf{u}$ and $\mathbf{w}$. Using the propeties of the dot product, we obtain

$$\begin{aligned} |\mathbf{u} - \mathbf{w}|^2 &= (\mathbf{u} - c\mathbf{v}) \cdot (\mathbf{u} - c\mathbf{v}) \\ &= |\mathbf{u}|^2 - 2c\mathbf{v} \cdot \mathbf{u} + c^2|\mathbf{v}|^2. \end{aligned}$$

Differentiating with respect to c, we obtain the equation

$$-2\mathbf{v} \cdot \mathbf{u} + 2c|\mathbf{v}|^2 = 0.$$

It follows that $\mathbf{w} = c\mathbf{v}$ is the best approximation to $\mathbf{u}$ if

$$c = \frac{\mathbf{v} \cdot \mathbf{u}}{|\mathbf{v}|^2}.$$

We therefore define the *vector projection* of $\mathbf{u}$ onto $\mathbf{v}$ by

$$\mathbf{w} = \text{proj}_{\mathbf{v}}\mathbf{u} = c\mathbf{v} = \frac{\mathbf{v} \cdot \mathbf{u}}{|\mathbf{v}|^2}\mathbf{v}.$$

This projection can also be written as

$$\text{proj}_{\mathbf{v}}\mathbf{u} = \frac{\mathbf{v} \cdot \mathbf{u}}{|\mathbf{v}|}\frac{\mathbf{v}}{|\mathbf{v}|} = (\mathbf{q} \cdot \mathbf{u})\mathbf{q} = (\text{comp}_{\mathbf{v}}\mathbf{u})\mathbf{q}$$

where

$$\mathbf{q} = \frac{\mathbf{v}}{|\mathbf{v}|}$$

is the unit vector in the direction of $\mathbf{v}$, and

$$\text{comp}_{\mathbf{v}}\mathbf{u} = \mathbf{q} \cdot \mathbf{u} = \frac{\mathbf{v} \cdot \mathbf{u}}{|\mathbf{v}|}$$

is called the *scalar projection* of $\mathbf{u}$ onto $\mathbf{v}$. Note that the absolute value of the scalar projection is also the magnitude of the vector projection.

The scalar projection can also be written as

$$\text{comp}_{\mathbf{v}}\mathbf{u} = \mathbf{q} \cdot \mathbf{u} = |\mathbf{q}||\mathbf{u}| \cos\theta = |\mathbf{u}| \cos\theta,$$

where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$, which is also the angle between $\mathbf{u}$ and $\mathbf{q}$. Furthermore, since the vector projection of $\mathbf{u}$ onto $\mathbf{v}$ is the best approximation of $\mathbf{u}$ by a vector in the direction of $\mathbf{v}$, the "error" in this approximation is orthogonal to $\mathbf{v}$. That is,

$$\mathbf{v} \cdot (\mathbf{u} - \text{proj}_{\mathbf{v}}\mathbf{u}) = \mathbf{v} \cdot \mathbf{u} - \frac{\mathbf{v} \cdot \mathbf{u}}{|\mathbf{v}|^2}\mathbf{v} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u} - \mathbf{v} \cdot \mathbf{u} = 0,$$

by the properties of the dot product.

Example Let $\mathbf{u} = \langle 2, 4\rangle$ and $\mathbf{v} = \langle 1, 1\rangle$. Then the unit vector in the direction of $\mathbf{v}$ is

$$\mathbf{q} = \frac{\langle 1, 1\rangle}{\sqrt{2}} = \left\langle \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right\rangle.$$

It follows that

$$\text{comp}_{\mathbf{v}}\mathbf{u} = \mathbf{q} \cdot \mathbf{u} = \left\langle \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right\rangle \cdot \langle 2, 4\rangle = \frac{6}{\sqrt{2}} = 3\sqrt{2},$$

and

$$\text{proj}_{\mathbf{v}}\mathbf{u} = (\text{comp}_{\mathbf{v}}\mathbf{u})\mathbf{q} = 3\sqrt{2}\left\langle \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right\rangle = \langle 3, 3\rangle.$$

This projection is shown in Figure 2.7. □

Example Let $\mathbf{u} = \langle 3, 4, 5\rangle$ and $\mathbf{v} = \langle 1, 1, 1\rangle$. Then the unit vector in the direction of $\mathbf{v}$ is

$$\mathbf{q} = \frac{\langle 1, 1, 1\rangle}{\sqrt{3}} = \left\langle \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right\rangle.$$

It follows that

$$\text{comp}_{\mathbf{v}}\mathbf{u} = \mathbf{q} \cdot \mathbf{u} = \left\langle \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right\rangle \cdot \langle 3, 4, 5\rangle = \frac{12}{\sqrt{3}} = 4\sqrt{3},$$

and

$$\text{proj}_{\mathbf{v}}\mathbf{u} = (\text{comp}_{\mathbf{v}}\mathbf{u})\mathbf{q} = 4\sqrt{3}\left\langle \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right\rangle = \langle 4, 4, 4\rangle.$$

This projection is shown in Figure 2.8. □

Example Suppose that a box is pulled 10 m along the ground by a constant force of 50 N that is applied at an angle of 30° above the horizontal. Then, the horizontal force can be obtained by computing the vector projection of the force vector $\mathbf{F} = 50\langle \cos 30°, \sin 30°\rangle$ onto the *displacement vector* $\mathbf{D} = 10\langle 1, 0\rangle$:

$$\text{comp}_{\mathbf{D}}\mathbf{F} = \frac{50\left\langle \frac{\sqrt{3}}{2}, \frac{1}{2} \right\rangle \cdot 10\langle 1, 0\rangle}{|10\langle 1, 0\rangle|} = 25\sqrt{3},$$

$$\text{proj}_{\mathbf{D}}\mathbf{F} = \text{comp}_{\mathbf{D}}\mathbf{F}\frac{\mathbf{D}}{|\mathbf{D}|} = 25\sqrt{3}\frac{10\langle 1, 0\rangle}{|10\langle 1, 0\rangle} = 25\sqrt{3}\langle 1, 0\rangle.$$

Since the displacement vector and the horizontal force vector have the same direction, it follows that the total *work* done, which is defined to be

the product of force and distance when the force is applied in the direction of displacement, is equal to the product of the *magnitudes* of the horizontal force and displacement vectors, which is

$$W = |\text{proj}_{\mathbf{D}}\mathbf{F}||\mathbf{D}| = (25\sqrt{3})(10) = 250\sqrt{3}.$$

More generally, the amount of work done by a force $\mathbf{F}$ to displace an object along the displacement vector $\mathbf{D}$ is given by

$$W = (\text{proj}_{\mathbf{D}}\mathbf{F}) \cdot \mathbf{D} = \frac{(\mathbf{F} \cdot \mathbf{D})\mathbf{D}}{|\mathbf{D}|^2} \cdot \mathbf{D} = \frac{(\mathbf{F} \cdot \mathbf{D})(\mathbf{D} \cdot \mathbf{D})}{\mathbf{D} \cdot \mathbf{D}} = \mathbf{F} \cdot \mathbf{D}.$$

This is a natural generalization of the basic formula for work, $W = FD$, where F is the amount of force applied to the object along the direction of displacement, and D is the distance that the object is moved. □

2.3.4 Summary

- The dot product, or inner product, of two vectors, is the sum of the products of corresponding components. Equivalently, it is the product of their magnitudes, times the cosine of the angle between them.
- The dot product of a vector with itself is the square of its magnitude.
- The dot product of two vectors is commutative; that is, the order of the vectors in the product does not matter.
- Multiplying a vector by a constant multiplies its dot product with any other vector by the same constant.
- The dot product of a vector with the zero vector is zero.
- Two nonzero vectors are perpendicular, or orthogonal, if and only if their dot product is equal to zero.
- The scalar projection of a vector $\mathbf{u}$ onto a vector $\mathbf{v}$ is $\mathbf{q} \cdot \mathbf{u}$, where $\mathbf{q}$ is the unit vector in the direction of $\mathbf{v}$.
- The vector projection of $\mathbf{u}$ onto $\mathbf{v}$ is the scalar projection of $\mathbf{u}$ onto $\mathbf{v}$ times $\mathbf{q}$, where $\mathbf{q}$ is the unit vector in the direction of $\mathbf{v}$.
- The vector projection of $\mathbf{u}$ onto $\mathbf{v}$ is the best approximation of $\mathbf{u}$ in the direction of $\mathbf{v}$, in the sense that the difference between $\mathbf{u}$ and its vector projection onto $\mathbf{v}$ is orthogonal to $\mathbf{v}$.

- The work done by a force that is applied at an angle to the displacement vector can be computed by projecting the force vector onto the displacement vector, and then multiplying the magnitudes of the force and displacement vectors.

2.4 The Cross Product

In three-dimensional space, or xyz-space, the set of solutions of an equation in x, y and z defines a surface. When the equation is linear, such as

$$y = 3, \quad x = y, \quad \text{or} \quad 2x + 3y - z = 0,$$

the solutions define a plane. On the other hand, a plane can also be defined in terms of any two vectors contained with in it. For example, the xy-plane is the set of all *displacement vectors* of the form $\mathbf{v} = x\mathbf{i} + y\mathbf{j}$, where x and y are real numbers. That is, any vector in the xy-plane is a *linear combination* of $\mathbf{i}$ and $\mathbf{j}$.

This definition of a plane, however, does not readily lend itself to an equation that characterizes it. To obtain such an equation, it is helpful to note that a plane can be viewed as the set of all vectors that are orthogonal to a particular direction. Therefore, given two vectors $\mathbf{u} = \langle u_1, u_2, u_3 \rangle$ and $\mathbf{v} = \langle v_1, v_2, v_3 \rangle$ that define a plane, we need to be able to compute a vector $\mathbf{w} = \langle w_1, w_2, w_3 \rangle$ that is orthogonal to both of them. Then, by the properties of the dot product, any vector $\mathbf{z} = c\mathbf{u} + d\mathbf{v}$, which is contained in this plane, satisfies

$$\mathbf{w} \cdot \mathbf{z} = \mathbf{w} \cdot (c\mathbf{u} + d\mathbf{v}) = c\mathbf{w} \cdot \mathbf{u} + d\mathbf{w} \cdot \mathbf{v} = 0,$$

so $\mathbf{w}$ is orthogonal to the entire plane.

To compute $\mathbf{w}$, we note that because $\mathbf{u} \cdot \mathbf{w}$ and $\mathbf{v} \cdot \mathbf{w} = 0$, its components satisfy the equations

$$\begin{aligned} u_1 w_1 + u_2 w_2 + u_3 w_3 &= 0, \\ v_1 w_1 + v_2 w_2 + v_3 w_3 &= 0. \end{aligned}$$

For simplicity, we assume $u_1 \neq 0$. To eliminate w_1 from the second equation, we subtract v_1/u_1 times the first equation from the second to obtain

$$\begin{aligned} u_1 w_1 + u_2 w_2 + u_3 w_3 &= 0, \\ \left(v_2 - \frac{v_1}{u_1} u_2 \right) w_2 + \left(v_3 - \frac{v_1}{u_1} u_3 \right) w_3 &= 0. \end{aligned}$$

Multiplying the second equation by u_1 yields

$$\begin{aligned} u_1w_1 + u_2w_2 + u_3w_3 &= 0, \\ (u_1v_2 - u_2v_1)\,w_2 + (u_1v_3 - u_3v_1)\,w_3 &= 0. \end{aligned}$$

By viewing the second equation as a dot product being equal to zero, and Using the fact that the vector $\langle -\beta, \alpha\rangle$ is orthogonal to the vector $\langle \alpha, \beta\rangle$, we obtain

$$w_2 = u_3v_1 - u_1v_3, \quad w_3 = u_1v_2 - u_2v_1.$$

Substituting these values into the first equation yields

$$w_1 = -\frac{u_2}{u_1}(u_3v_1 - u_1v_3) - \frac{u_3}{u_1}(u_1v_2 - u_2v_1) = u_2v_3 - u_3v_2.$$

We conclude that the vector

$$\mathbf{w} = \langle u_2v_3 - u_3v_2, u_3v_1 - u_1v_3, u_1v_2 - u_2v_1\rangle$$

is orthogonal to both $\mathbf{u}$ and $\mathbf{v}$. We define $\mathbf{w}$ to be the *cross product* of $\mathbf{u}$ and $\mathbf{v}$, and write

$$\mathbf{w} = \mathbf{u} \times \mathbf{v}.$$

We note that the above system of equations has infinitely many solutions. We have chosen the particular solution denoted by $\mathbf{w}$ because its components can be computed without having to divide by any of the components of $\mathbf{u}$ and $\mathbf{v}$, thus avoiding any requirement that any of these components be nonzero. This condition uniquely determines the *magnitude* of $\mathbf{w}$, but not its direction; this ambiguity will be resolved shortly.

Example Let $\mathbf{u} = \langle 2, -1, 1\rangle$ and $\mathbf{v} = \langle 1, 2, 1\rangle$. Then

$$\mathbf{u} \times \mathbf{v} = \langle (-1)1 - 2(1), 1(1) - 2(1), 2(2) - (-1)(1)\rangle = \langle -3, -1, 5\rangle.$$

The vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{u} \times \mathbf{v}$ are shown in Figure 2.9. □

Example Consider the plane that contains the three points $P_1 = (1, 1, 0)$, $P_2 = (-1, 3, 1)$, and $P_3 = (-1, -1, -1)$. We wish to compute the vector $\mathbf{w}$ that is orthogonal to all vectors contained in this plane. This can be accomplished by computing vectors $\mathbf{u}$ and $\mathbf{v}$ with initial and terminal points taken from the three points P_1, P_2, and P_3.

Specifically, we define

$$\mathbf{u} = \vec{P_1P_2} = \langle -1 - 1, 3 - 1, 1 - 0\rangle = \langle -2, 2, 1\rangle,$$

$$\mathbf{v} = \vec{P_1P_3} = \langle -1-1, -1-1, -1-0\rangle = \langle -2, -2, -1\rangle.$$

Then, we have

$$\begin{aligned}
\mathbf{w} &= \mathbf{u} \times \mathbf{v} \\
&= \langle -2, 2, 1\rangle \times \langle -2, -2, -1\rangle \\
&= \langle 2(-1) - 1(-2), 1(-2) - (-2)(-1), -2(-2) - 2(-2)\rangle \\
&= \langle 0, -4, 8\rangle.
\end{aligned}$$

These vectors are shown in Figure 2.10. □

Another way to derive the cross product is in terms of the determinant of the matrix

$$A = \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{bmatrix},$$

where, as before, **i**, **j** and **k** are the standard basis vectors. We have

$$\mathbf{w} = \mathbf{u} \times \mathbf{v} = \det(A).$$

This definition follows from Cramer's rule, which expresses the solution to a system of linear equations in terms of determinants.

We now compute the magnitude of the cross product. We have

$$\begin{aligned}
|\mathbf{w}|^2 &= w_1^2 + w_2^2 + w_3^2 \\
&= (u_2v_3 - u_3v_2)^2 + (u_3v_1 - u_1v_3)^2 + (u_1v_2 - u_2v_1)^2 \\
&= u_2^2v_3^2 - 2u_2u_3v_2v_3 + u_3^2v_2^2 + u_3^2v_1^2 - 2u_1u_3v_1v_3 + u_1^2v_3^2 + u_1^2v_2^2 - 2u_1u_2v_1v_2 + u_2^2v_1^2 \\
&= u_2^2v_3^2 - 2u_2u_3v_2v_3 + u_3^2v_2^2 + u_3^2v_1^2 - 2u_1u_3v_1v_3 + u_1^2v_3^2 + u_1^2v_2^2 - 2u_1u_2v_1v_2 + u_2^2v_1^2 + \\
&\quad (u_1^2v_1^2 + u_2^2v_2^2 + u_3^2v_3^2) - (u_1^2v_1^2 + u_2^2v_2^2 + u_3^2v_3^2) \\
&= (u_1^2 + u_2^2 + u_3^2)(v_1^2 + v_2^2 + v_3^2) - \\
&\quad (2u_2v_2u_3v_3 + 2u_1v_1u_3v_3 + 2u_1v_1u_2v_2 + u_1^2v_1^2 + u_2^2v_2^2 + u_3^2v_3^2) \\
&= |\mathbf{u}|^2|\mathbf{v}|^2 - (\mathbf{u} \cdot \mathbf{v})^2 \\
&= |\mathbf{u}|^2|\mathbf{v}|^2 - |\mathbf{u}|^2|\mathbf{v}|^2 \cos^2\theta \\
&= |\mathbf{u}|^2|\mathbf{v}|^2(1 - \cos^2\theta) \\
&= |\mathbf{u}|^2|\mathbf{v}|^2 \sin^2\theta.
\end{aligned}$$

where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$, with $0 \leq \theta \leq \pi$. We conclude

$$|\mathbf{u} \times \mathbf{v}| = |\mathbf{u}||\mathbf{v}| \sin\theta.$$

The direction of $\mathbf{u} \times \mathbf{v}$ is not uniquely determined by the fact that it is orthogonal to both $\mathbf{u}$ and $\mathbf{v}$, for its negative also has this property. The proper direction is determined by the *right-hand rule*. This rule states that if the fingers of the right hand are curled in the direction of rotation from $\mathbf{u}$ to $\mathbf{v}$, through an angle less than 180°, then the thumb is pointing in the direction of $\mathbf{u} \times \mathbf{v}$.

2.4.1 Parallel Vectors

We have learned that the Cross Product can be used to compute a vector $\mathbf{w}$ in V_3 that is orthogonal to two given vectors $\mathbf{u}$ and $\mathbf{v}$ that, together, define a plane in xyz-space. However, two given vectors do not necessarily define a plane; it is required that they not be parallel. If $\mathbf{u}$ and $\mathbf{v}$ are parallel, then, by the definition of the cross product, we obtain

$$\mathbf{w} = \mathbf{u} \times \mathbf{v} = \mathbf{0}.$$

This is consistent with the fact that $|\mathbf{u} \times \mathbf{v}| = |\mathbf{u}||\mathbf{v}| \sin\theta$, where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$. If $\mathbf{u}$ and $\mathbf{v}$ are parallel, then this angle is either 0 or π radians, and in either case, the sine of that angle is zero.

Example The vectors $\mathbf{u} = \langle 2, 1, 2 \rangle$ and $\mathbf{v} = \langle -4, -2, -4 \rangle$ are parallel, as $\mathbf{v} = -2\mathbf{u}$. That is, $\mathbf{v}$ has twice the magnitude of $\mathbf{u}$, and the opposite direction. It follows that $\mathbf{u} \times \mathbf{v} = \mathbf{0}$. □

2.4.2 Properties

The following properties are useful for performing complex vector operations that involve the cross product.

1. Anticommutativity: $\mathbf{v} \times \mathbf{u} = -(\mathbf{u} \times \mathbf{v})$
2. Scalar multiplication: $(c\mathbf{u}) \times \mathbf{v} = \mathbf{u} \times (c\mathbf{v})$
3. Distribution over addition: $\mathbf{w} \times (\mathbf{u} + \mathbf{v}) = \mathbf{w} \times \mathbf{u} + \mathbf{w} \times \mathbf{v}$, $(\mathbf{u} + \mathbf{v}) \times \mathbf{w} = \mathbf{u} \times \mathbf{w} + \mathbf{v} \times \mathbf{w}$.
4. $\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v}) = (\mathbf{w} \times \mathbf{u}) \cdot \mathbf{v}$
5. $\mathbf{w} \times (\mathbf{u} \times \mathbf{v}) = (\mathbf{w} \cdot \mathbf{v})\mathbf{u} - (\mathbf{w} \cdot \mathbf{u})\mathbf{v}$

2.4.3 Areas

The magnitude of the cross product has useful geometric applications. Consider a parallelogram with vectors $\mathbf{u}$ and $\mathbf{v}$ as adjacent sides, such that the angle θ between them, that is *internal* to the parallelogram, is not more than

90 degrees. The area of this parallelogram is the product of the lengths of its base and its height. If we choose $\mathbf{u}$ to be its base, then, by right triangle trigonometry, its height is given by $|\mathbf{v}|\sin\theta$, where $\sin\theta > 0$ due to the fact that $0 < \theta < 90°$.

It follows that the area A of the parallelogram is

$$A = |\mathbf{u}||\mathbf{v}|\sin\theta = |\mathbf{u} \times \mathbf{v}|.$$

That is, the magnitude of the cross product of $\mathbf{u}$ and $\mathbf{v}$ yields the area of a parallelogram with adjacent sides $\mathbf{u}$ and $\mathbf{v}$. We have assumed that the angle between them is no more than 90 degrees, but if this is not the case, we can obtain the same result by instead defining the parallelogram in terms of $-\mathbf{u}$ and $\mathbf{v}$, in view of the fact that $|-\mathbf{u}| = |\mathbf{u}|$.

This same formula can be used to compute the area of a triangle with adjacent sides defined by $\mathbf{u}$ and $\mathbf{v}$. The area A of such a triangle is given by

$$A = \frac{1}{2}|\mathbf{u} \times \mathbf{v}|,$$

since a parallelogram consists of two triangles of equal area that share a side.

Example We will compute the area of the triangle whose vertices are the points $P_1 = (1, 1, 0)$, $P_2 = (3, 0, 4)$ and $P_3 = (0, 5, 2)$. This triangle is half of a parallelogram with adjacent sides

$$\mathbf{u} = \vec{P_1P_2} = \langle 2, -1, 4\rangle, \quad \mathbf{v} = \vec{P_1P_3} = \langle -1, 4, 2\rangle.$$

The area of this parallelogram is

$$|\mathbf{u}\times\mathbf{v}| = |\langle -1(2)-4(4), 4(-1)-2(2), 2(4)-(-1)(-1)\rangle| = |\langle -18, -8, 7\rangle| = \sqrt{437} \approx 23.345.$$

Therefore, the area of the triangle is $\sqrt{545}/2 \approx 10.452$.

Alternatively, we can compute the angle θ between $\mathbf{u}$ and $\mathbf{v}$, using the dot product:

$$\cos\theta = \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{u}||\mathbf{v}|} = \frac{1}{\sqrt{26}\sqrt{21}} = \frac{1}{\sqrt{546}}.$$

This yields $\theta = \cos^{-1}(1/\sqrt{546}) \approx 1.528$ radians, or $87.547°$. Then, we compute

$$|\mathbf{u}\times\mathbf{v}| = |\mathbf{u}||\mathbf{v}|\sin\theta = \sqrt{21}\sqrt{26}\sqrt{1-\cos^2\theta} = \sqrt{546}\sqrt{1-\frac{1}{546}} = \sqrt{546}\sqrt{\frac{545}{546}} = \sqrt{545},$$

where we have used the identity $\cos^2\theta + \sin^2\theta = 1$. □

2.4.4 Volumes

The cross product and dot product can be used together to compute the volume of a parallelepiped. Let $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ be three vectors that do not lie within the same plane. Then, they define three sides of a parallelepiped, in which any two of them define a parallelogram that serves as one of its six faces.

The volume of the parallelepiped is the area of the base parallelogram times the height. Let the parallelogram defined by $\mathbf{u}$ and $\mathbf{v}$ be the base. We know that its area is $|\mathbf{u} \times \mathbf{v}|$. To obtain the height, we first note that the height must be measured along a direction that is perpendicular to the base; that is, it should be in the direction of $\mathbf{z}$ or $-\mathbf{z}$.

Then, by right-triangle trigonometry, the height is $|\mathbf{w}||\cos\theta|$, where θ is the angle between $\mathbf{w}$ and $\mathbf{z}$. It follows that the volume V of the parallelepiped is

$$V = |\mathbf{z}||\mathbf{w}||\cos\theta| = |\mathbf{w} \cdot \mathbf{z}| = |\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})|.$$

We refer to this combination of the dot and cross products as the *triple product* of $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$.

Example Consider the parallelepiped, shown in Figure 2.11, defined by the vectors

$$\mathbf{u} = \langle 3, 1, 1\rangle, \quad \mathbf{v} = \langle 1, 5, 1\rangle, \quad \mathbf{w} = \langle 0, 1, 2\rangle.$$

To compute its volume, we define the base to be the parallelogram with edges $\mathbf{u}$ and $\mathbf{v}$. Then, we compute

$$\mathbf{z} = \mathbf{u} \times \mathbf{v} = \langle 1(1) - 1(5), 1(1) - 3(1), 3(5) - 1(1)\rangle = \langle -4, -2, 14\rangle.$$

Finally, we obtain the volume by the triple product

$$|\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})| = |\mathbf{w} \cdot \mathbf{z}| = |\langle 0, 1, 2\rangle \cdot \langle -4, -2, 14\rangle| = 26.$$

□

2.4.5 Summary

- The cross product of two nonzero vectors $\mathbf{u}$ and $\mathbf{v}$ is a vector $\mathbf{w} = \mathbf{u} \times \mathbf{v}$ that is orthogonal to both $\mathbf{u}$ and $\mathbf{v}$.

- The magnitude of $\mathbf{u} \times \mathbf{v}$ is $|\mathbf{u}||\mathbf{v}|\sin\theta$, where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$, with $0 \leq \theta \leq \pi$.

- The direction of $\mathbf{u} \times \mathbf{v}$ is determined by the direction of the thumb of the right hand, when the fingers are rotated from $\mathbf{u}$ to $\mathbf{v}$ through the angle θ between $\mathbf{u}$ and $\mathbf{v}$, with $0 \leq \theta \leq \pi$.
- A vector that is orthogonal to all vectors in a given plane can be computed from any three points in the plane by first subtracting the coordinates of the points to obtain two vectors in the plane, and then computing their cross product.
- The cross product of two parallel vectors is the zero vector.
- The magnitude of the cross product of two vectors is the area of a parallelogram with the two vectors defining adjacent sides.
- The triple product of three vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$, given by $|\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})|$, is the volume of the parallelepiped defined by $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$.

2.5 Equations of Lines and Planes

2.5.1 Equations of Lines

A line can be viewed, conceptually, as the set of all points in space that satisfy two criteria:

1. They contain a particular point, which we identify by a position vector $\mathbf{r}_0$.
2. The vector between $\mathbf{r}_0$ and any position vector $\mathbf{r}$ on the line is parallel to a given vector $\mathbf{v}$.

The vector with initial point $\mathbf{r}_0$ and terminal point $\mathbf{r}$ is given by $\mathbf{s} = \mathbf{r} - \mathbf{r}_0$. This vector must be parallel to $\mathbf{v}$, which implies that $\mathbf{s} = t\mathbf{v}$ for some scalar t. It follows that *any* position vector $\mathbf{r}$, corresponding to a point P on the line with coordinates equal to the components of r, has the form

$$\mathbf{r} = \mathbf{r}_0 + t\mathbf{v},$$

where t is a scalar that is called a *parameter*. The parameter t is allowed to assume *any* value, so that any point P on the line can be obtained from this equation, which is called the *vector equation* of the line. Such an equation is illustrated in Figure 2.12.

Let $\mathbf{r} = \langle x, y, z \rangle$ be the position vector of any point $P = (x, y, z)$ on a line, and let $\mathbf{r}_0 = \langle x_0, y_0, z_0 \rangle$ be the position vector of a *particular* point

$P_0 = (x_0, y_0, z_0)$ on the line. Furthermore, let $\mathbf{v} = \langle a, b, c \rangle$ be the vector that indicates the direction of the line. By writing the vector equation of the line in terms of components, we obtain the *parametric equations* of the line,

$$x = x_0 + at, \quad y = y_0 + bt, \quad z = z_0 + ct.$$

The components a, b and c of $\mathbf{v}$ are called the *direction numbers* of the line.

Example Let $\mathbf{r}_0 = \langle 1, 2, 0 \rangle$ and $\mathbf{v} = \langle 1, -3, 2 \rangle$. Then the vector equation of the line containing $\mathbf{r}_0$ and parallel to $\mathbf{v}$ is

$$\mathbf{r} = \langle 1, 2, 0 \rangle + t\langle 1, -3, 2 \rangle = \langle 1 + t, 2 - 3t, 2t \rangle.$$

The corresponding parametric equations are

$$x = 1 + t, \quad y = 2 - 3t, \quad z = 2t.$$

The direction numbers of the line are 1, -3 and 2. □

The direction numbers can be used to describe the line using equations that do not require the parameter t. If all of the direction numbers are nonzero, we can solve each of the parametric equations for t. This yields three expressions for t that can be equated, resulting in the *symmetric equations*

$$\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c}.$$

If any of the direction numbers are equal to zero, then t can still be eliminated using those parametric equations with nonzero direction numbers.

Example To obtain the symmetric equations of the line from the previous example, we solve each of the parametric equations and obtain

$$t = x - 1, \quad t = \frac{y - 2}{-3}, \quad t = \frac{z}{2}.$$

Setting the three expressions for t equal to each other yields the symmetric equations

$$\frac{x - 1}{1} = \frac{y - 2}{-3} = \frac{z}{2}.$$

On the other hand, the symmetric equations for the line whose parametric equations are

$$x = 2t + 3, \quad y = -1, \quad z = -4t + 5,$$

and whose direction numbers are 2, 0 and -4, are

$$\frac{x - 3}{2} = \frac{z - 5}{-4}, \quad y = -1.$$

□

Example Consider the line with parametric equations

$$x = 7t - 6, \quad y = -3t + 4, \quad z = 2t + 9.$$

To determine the point at which this line intersects the xy-plane, we note that this plane consists of all points (x, y, z) in space such that $z = 0$. We therefore set $z = 0$ in the third parametric equation to obtain the equation $0 = 2t+9$, which we then solve for t to obtain $t = -9/2$. We then substitute $t = -9/2$ into the parametric equations for x and y to obtain the intersection point

$$x = 7\left(-\frac{9}{2}\right) - 6 = -\frac{75}{2}, \quad y = -3\left(-\frac{9}{2}\right) + 4 = \frac{35}{2}, \quad z = 0.$$

□

The vector equation of a line can also be used to describe a (finite) line segment, as opposed to an entire line, which extends infinitely in both directions. Let $\mathbf{r}_0$ be the initial point of the line segment, and $\mathbf{r}_1$ be the terminal point. Then, let $\mathbf{v} = \mathbf{r}_1 - \mathbf{r}_0$. It follows that

$$\mathbf{r} = \mathbf{r}_0 + t\mathbf{v}, \quad 0 \leq t \leq 1,$$

describes a point on the line segment. When $t = 0$, we have $\mathbf{r} = \mathbf{r}_0$, whereas when $t = 1$, $\mathbf{r} = \mathbf{r}_0 + (\mathbf{r}_1 - \mathbf{r}_0) = \mathbf{r}_1$. When $0 < t < 1$, $\mathbf{r}$ is the position vector of a point P in the interior of the segment. Using the definition of $\mathbf{v}$, the equation of the line segment becomes

$$\mathbf{r} = (1 - t)\mathbf{r}_0 + t\mathbf{r}_1, \quad 0 \leq t \leq 1.$$

Given any two lines in space, there are three ways in which they can interact:

- They can intersect at a point P. To determine if this is the case, we can equate the formulas for corresponding components (x, y, and z) and try to find a set of parameter values, one for each line, that satisfies *all* of the resulting equations. If this is the case, then the lines intersect at a point whose coordinates can be obtained by substituting one of these parameter values into the appropriate parametric equations.

- They can be parallel. This is the case if and only if the vectors that describe the directions of the lines are parallel. That is, these vectors must be (nonzero) scalar multiples of one another.

- They are skew. This is the case if the lines do not intersect, and are not parallel. Intuitively, one example of skew lines are lines that are contained in parallel planes, but have different direction numbers.

Example Consider the lines with parametric equations

$$x = 3t - 4, \quad y = 2t + 6, \quad z = -t + 10,$$

$$x = 5s + 2, \quad y = 3s - 1, \quad z = -2s + 7.$$

We use different letters for the parameters of the two lines, because they are independent of one another. To compute the intersection point, if it exists, we set the two expressions for x equal to each other to obtain the equation

$$3t - 4 = 5s + 2.$$

Similarly, we set the expressions for y and z equal to one another to obtain the additional equations

$$2t + 6 = 3s - 1, \quad -t + 10 = -2s - 3.$$

We must find values of s and t that satisfy *all three* equations. To do this, we consider the first two equations. Solving the second equation, $2t + 6 = 3s - 1$, for t yields

$$t = \frac{3s}{2} - \frac{7}{2}.$$

Substituting this expression for t into the first equation, and solving for s, yields $s = -33$. Substituting this value of s in the above expression for t yields $t = -53$.

We then substitute both of these values into the third equation, and find that it is not satisfied, as we obtain the falsehood $-13 = -3$. Therefore, the lines are skew. If the parameter values $s = -33$ and $t = -53$, obtained from the first two equations, *had* satisfied the third equation, then the lines would have intersected. In that case, the value of t could be substituted into the equations for the first line to obtain the coordinates of the intersection point. □

Example The lines whose parametric equations are

$$x = 3t - 4, \quad y = 2t + 6, \quad z = -t + 10,$$

$$x = 6t + 2, \quad y = 4s - 1, \quad z = -2s + 7,$$

are parallel. To see this, we note that the first line has direction numbers 3, 2 and -1, while the second line has direction numbers 6, 4 and -2. The direction numbers of the second line are twice those of the first. Because all of the ratios of corresponding direction numbers are the same, the lines are parallel. □

2.5.2 Systems of Linear Equations

The process of determining whether two non-parallel lines are intersect, or whether they are skew, entails solving a system of three *linear equations* for two *unknowns*, which are the parameters from the equations of the two lines. Therefore, it is helpful to examine such systems of equations and learn about techniques for solving them.

First, we consider a system of only two linear equations, with two unknowns s and t:

$$\begin{aligned} as + bt &= u, \\ cs + dt &= v. \end{aligned}$$

The numbers a, b, c and d are called the *coefficients* of the system, and the numbers u and v are the *right-hand side* values, since, by convention, the coefficients and unknowns are normally written on the left-hand side of the equal sign.

We now attempt to solve this system for s and t. We assume that $a \neq 0$, for otherwise, we can interchange the equations, provided that $c \neq 0$. The case of $a = c = 0$ will be discussed later. We then multiply the first equation by c/a, and subtract it from the second equation to obtain the modified system

$$\begin{aligned} as + bt &= u, \\ \left(c - \frac{c}{a}a\right)s + \left(d - \frac{c}{a}b\right)t &= v - \frac{c}{a}u. \end{aligned}$$

Multiplying the second equation by a yields

$$\begin{aligned} as + bt &= u, \\ (ad - bc)t &= av - cu. \end{aligned}$$

If $ad - bc \neq 0$, then the solution of the system can be obtained by solving the second equation for t, and then substituting this value into the first equation to solve for s. This yields

$$t = \frac{av - cu}{ad - bc}, \quad s = \frac{u - bt}{a} = \frac{u(ad - bc) - b(av - cu)}{a(ad - bc)} = \frac{du - bv}{ad - bc}.$$

We see that if $ad - bc \neq 0$, then the system of equations has a *unique* solution. However, what if $ad - bc = 0$? In that case, the second equation becomes $0 = av - cu$. This leads to two possibilities:

- If $av - cu \neq 0$, then the second equation is a contradiction, because the left side is zero while the right side is nonzero. It follows that there is *no solution* to the system of equations, and we say that the equations are *inconsistent*.

- If $av - cu = 0$, then the second equation is the *identity* $0 = 0$. That is, it is satisfied regardless of the values of s and t. These unknowns are only constrained by the first equation $as + bt = u$, which actually has *infinitely many* solutions, because for *any* value of t, we can solve for s and obtain $s = (u - bt)/a$.

To this point, we have assumed that either a or c is nonzero. If both are zero, then we have similar outcomes, for then the system of equations reduces to

$$\begin{aligned} bt &= u, \\ dt &= v. \end{aligned}$$

Each equation can easily be solved for t to obtain $t = u/b$ or $t = v/d$, assuming that b and d are nonzero. If the two values of t are equal, then the system has *infinitely many* solutions, because s can assume any value. Otherwise, the system is inconsistent, and there is *no solution*.

Now, suppose that $b = 0$ in the above system. If $u = 0$, then the first equation reduces to the identity $0 = 0$, and the second equation can be used to determine t, assuming that $d \neq 0$. However, if $u \neq 0$, then we obtain the contradiction $0 = u$ for some nonzero u, and therefore the system has no solution. The only scenario that remains is if a, b, c and d are all zero. If both u and v are zero, then any value of s and t satisfies this trivial system; otherwise, one or both equations is a contradiction.

To summarize, there are three possible outcomes of a system of two equations and two unknowns:

- There is a unique solution for s and t, which is the case if and only if $ad - bc \neq 0$.

- There are infinitely many solutions.

- There is no solution.

It is interesting to note that if we write the system in matrix-vector form

$$A\mathbf{s} = \mathbf{u},$$

where

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad \mathbf{s} = \begin{bmatrix} s \\ t \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u \\ v \end{bmatrix},$$

then the *determinant* of the matrix A is given by

$$\det(A) = ad - bc.$$

That is, the system of equations has a unique solution if and only if the determinant of its coefficient matrix is nonzero, which is also true for systems of n equations and n unknowns, for any positive integer n. Furthermore, we can express the previously stated formulas for the solutions s and t as

$$\mathbf{s} = \frac{1}{ad-bc}\begin{bmatrix} du - bv \\ av - cu \end{bmatrix} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = A^{-1}\mathbf{u},$$

where A^{-1} is the *inverse* of A. This is the generalization of the reciprocal of a nonzero number to a matrix with nonzero determinant, which is said to be *invertible*.

When computing the intersection point of two lines in three-dimensional space, if one exists, we need to solve three equations with only two unknowns. Generally, such systems are inconsistent, which is interpreted as the lines being skew. However, to be sure, we can use the following procedure:

1. Attempt to compute the solution of the system of the first two equations, corresponding to the x- and y-coordinates.

2. If the first two equations have a unique solution (that is, $ad - bc \neq 0$), then substitute the computed values of s and t into the third equation. If the equation is satisfied, then the lines intersect; otherwise, they are skew.

3. If the first two equations are inconsistent (that is, $av - cu \neq 0$), then the lines are skew.

4. If the first two equations have infinitely many solutions, then the lines are guaranteed to intersect. The simplest course of action is to return to step 1 and repeat the process with a different set of two equations chosen from the original system of three. Because the lines are not parallel, the new system is guaranteed to have a unique solution, and there is no need to verify the solution by substituting the computed values of s and t into the remaining equation.

We now illustrate this procedure with some examples.

Example Consider the lines with parametric equations

$$x = 2s + 3, \quad y = 4s + 3, \quad z = s + 1,$$

$$x = 4t + 5, \quad y = 5t + 6, \quad z = 2t + 9.$$

These lines are not parallel, because the direction numbers are not all proportional. To determine whether they intersect, we set the expressions for the x- and y-coordinates equal to one another:

$$2s + 3 = 4t + 5, \quad 4s + 3 = 5t + 6,$$

to obtain the system of equations

$$\begin{aligned} 2s - 4t &= 2, \\ 4s - 5t &= 3. \end{aligned}$$

In the notation of the general system of two equations discussed previously, $a = 2$, $b = -4$, $c = 4$, $d = -5$, $u = 2$, and $v = 3$. We have

$$ad - bc = 2(-5) - (-4)(4) = -10 - (-16) = 6 \neq 0,$$

so this system has a unique solution. Using the formulas for s and t above, we obtain

$$s = \frac{du - bv}{ad - bc} = \frac{(-5)(2) - (-4)(3)}{6} = \frac{1}{3}, \quad t = \frac{av - cu}{ad - bc} = \frac{2(3) - 4(2)}{6} = -\frac{1}{3}.$$

Substituting these values into the equation for the z-coordinate, which is $s - 2t = 8$, we obtain

$$\frac{1}{3} - 2\left(-\frac{1}{3}\right) = \frac{1}{3} + \frac{2}{3} = 1 \neq 8,$$

so the equation is not satisfied. We conclude that the lines are skew. □

Example Consider the lines with parametric equations

$$x = 2s + 3, \quad y = 3s + 4, \quad z = 4s + 3,$$

$$x = 4t + 5, \quad y = 6t + 7, \quad z = 5t + 9.$$

These lines are not parallel, because the direction numbers are not all proportional. To determine whether they intersect, we set the expressions for the x- and y-coordinates equal to one another:

$$2s + 3 = 4t + 5, \quad 3s + 4 = 6t + 7,$$

to obtain the system of equations

$$\begin{aligned} 2s - 4t &= 2, \\ 3s - 6t &= 3. \end{aligned}$$

In the notation of the general system of two equations discussed previously, $a = 2$, $b = -4$, $c = 3$, $d = -6$, $u = 2$, and $v = 3$. We have

$$ad - bc = 2(-6) - (-4)(3) = -12 - (-12) = 0,$$

so this system does not have a unique solution. However, we also have

$$av - cu = 2(3) - 3(2) = 0,$$

so the system actually has infinitely many solutions. Therefore, we instead work with the equations for the x- and z-coordinates, which yields the new system

$$\begin{aligned} 2s - 4t &= 2, \\ 4s - 5t &= 6. \end{aligned}$$

In this system, $a = 2$, $b = -4$, $c = 4$, $d = -5$, $u = 2$ and $v = 6$. We have

$$ad - bc = 2(-5) - (-4)(4) = -10 - (-16) = 6 \neq 0,$$

so this system has a unique solution. Using the formulas for s and t above, we obtain

$$s = \frac{du - bv}{ad - bc} = \frac{(-5)(2) - (-4)(6)}{6} = \frac{7}{3}, \quad t = \frac{av - cu}{ad - bc} = \frac{2(6) - 4(2)}{6} = \frac{2}{3}.$$

We conclude that the lines intersect when $s = 7/3$ and $t = 2/3$. Substituting $s = 7/3$ into the equation for the first line, we obtain the coordinates of the intersection point,

$$x = 2\frac{7}{3} + 3 = \frac{23}{3}, \quad y = 3\frac{7}{3} + 4 = 11, \quad z = 4\frac{7}{3} + 3 = \frac{37}{3}.$$

□

Example Consider the lines with parametric equations

$$x = 2s + 3, \quad y = s + 1, \quad z = 4s + 3,$$

$$x = 4t + 5, \quad y = 2t + 9, \quad z = 5t + 6.$$

These lines are not parallel, because the direction numbers are not all proportional.

To determine whether they intersect, we set the expressions for the x- and y-coordinates equal to one another:

$$2s + 3 = 4t + 5, \quad s + 1 = 2t + 9,$$

to obtain the system of equations

$$\begin{aligned} 2s - 4t &= 2, \\ s - 2t &= 8. \end{aligned}$$

We see that the coefficients in the equations are proprtional. That is, the coefficients in the first equation are double of the corresponding coefficients of the second.

However, the quantities on the right-hand side of the equations do not have the same ratio. Therefore, these equations are inconsistent, meaning that they do not have a solution. We immediately conclude that the lines are skew. □

Example Consider the lines with parametric equations

$$x = 2s + 3, \quad y = 3s + 4, \quad z = 4s + 3,$$

$$x = 4t + 5, \quad y = 6t + 7, \quad z = 5t + 9.$$

These lines are not parallel, because the direction numbers are not all proportional.

To determine whether they intersect, we set the expressions for the x- and y-coordinates equal to one another:

$$2s + 3 = 4t + 5, \quad 3s + 4 = 6t + 7,$$

to obtain the system of equations

$$\begin{aligned} 2s - 4t &= 2, \\ 3s - 6t &= 3. \end{aligned}$$

The coefficients, and right-hand side quantities, of these equations are all proportional. That is, we can obtain the second equation by multiplying the first by 3/2. In other words, these equations are *dependent* on one another, and therefore do not have a unique solution. In fact, they have infinitely many solutions.

We instead work with the equations for the x- and z-coordinates, which yields the new system

$$\begin{aligned} 2s - 4t &= 2, \\ 4s - 5t &= 6. \end{aligned}$$

These equations are guaranteed to be independent, and thus have a unique solution, since the lines are already known to not be parallel.

By doubling the first equation, which yields $4s - 8t = 4$, and subtracting it from the second, we obtain $3t = 2$, and therefore $t = 2/3$. Substituting this value into the first equation yields the equation $2s - 4(2/3) = 2$, which has the solution $s = 7/3$.

We conclude that the lines intersect when $s = 7/3$ and $t = 2/3$. Substituting $s = 7/3$ into the equation for the first line, we obtain the coordinates of the intersection point,

$$x = 2\frac{7}{3} + 3 = \frac{23}{3}, \quad y = 3\frac{7}{3} + 4 = 11, \quad z = 4\frac{7}{3} + 3 = \frac{37}{3}.$$

□

2.5.3 Equations of Planes

Previously, we learned how to describe lines using various types of equations. Now, we will do the same with planes. Suppose that we are given three points $\mathbf{r}_0$, $\mathbf{r}_1$ and $\mathbf{r}_2$ that are not co-linear. Then, these points define a plane, and the vectors $\mathbf{v}_1 = \mathbf{r}_1 - \mathbf{r}_0$ and $\mathbf{v}_2 = \mathbf{r}_2 - \mathbf{r}_0$ are vectors contained within the plane, that are also not parallel to one another.

A plane consists of all vectors that are orthogonal to a given direction $\mathbf{n}$, which is said to be *normal* to the plane, and passes through a given point $\mathbf{r}_0$. The *normal vector* $\mathbf{n}$ can be obtained by computing

$$\mathbf{n} = \mathbf{v}_1 \times \mathbf{v}_2.$$

Let $\mathbf{r}$ be *any* point in the plane. Then the vector $\mathbf{u} = \mathbf{r} - \mathbf{r}_0$ is orthogonal to $\mathbf{n}$. That is,

$$\mathbf{n} \cdot \mathbf{u} = \mathbf{n} \cdot (\mathbf{r} - \mathbf{r}_0) = 0.$$

This equation is called the *vector equation* of the plane.

If we write

$$\mathbf{n} = \langle a, b, c\rangle, \quad \mathbf{r} = \langle x, y, z\rangle, \quad \mathbf{r}_0 = \langle x_0, y_0, z_0\rangle,$$

then the vector equation can be rewritten as

$$ax + by + cz + d = 0,$$

where $d = -\mathbf{n} \cdot \mathbf{r}_0 = -(ax_0 + by_0 + cz_0)$. This is a *linear equation* in the *unknowns* x, y, and z.

Example Consider the plane containing the points $P_0 = (1, 4, 1)$, $P_1 = (5, 1, -1)$ and $P_2 = (4, 4, 4)$, which we identify with the position vectors

$$\mathbf{r}_0 = \langle 1, 4, 1\rangle, \quad \mathbf{r}_1 = \langle 5, 1, -1\rangle, \quad \mathbf{r}_2 = \langle 4, 4, 4\rangle.$$

We wish to find a linear equation that describes this plane. First, we need to compute a vector $\mathbf{n}$ that is normal to the plane, which can be obtained by computing the cross product of two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ that are contained within the plane. We have

$$\begin{aligned}
\mathbf{n} &= \mathbf{v}_1 \times \mathbf{v}_2 \\
&= (\mathbf{r}_1 - \mathbf{r}_0) \times (\mathbf{r}_2 - \mathbf{r}_0) \\
&= \langle 4, -3, -2\rangle \times \langle 3, 0, 3\rangle \\
&= \langle -3(3) - (-2)0, -2(3) - 4(3), 4(0) - (-3)(3)\rangle \\
&= \langle -9, -18, 9\rangle.
\end{aligned}$$

It follows that the vector equation of the plane is

$$\mathbf{n} \cdot (\mathbf{r} - \mathbf{r}_0) = \langle -9, -18, 9\rangle \cdot (\langle x, y, z\rangle - \langle 1, 4, 1\rangle = 0,$$

which can also be written as

$$-9x - 18y + 9z + 72 = 0,$$

since

$$\mathbf{n} \cdot \mathbf{r}_0 = -72.$$

□

2.5.4 Intersecting Planes

The *angle between planes* is defined to be the angle between their normal vectors. If this angle is either 0 or π, then the normal vectors are parallel, and we say that the planes are parallel. Otherwise, the planes intersect, and this intersection is a line.

To determine the line formed by this intersection, we need to solve the system of equations consisting of the equations of the two planes. Because this system of equations has three unknowns, but there are only two equations, there will be infinitely many points that satisfy the system, and the set of all such solutions constitutes a line.

Let the equations of two planes be given by

$$a_1x + b_1y + c_1z + d_1 = 0, \quad a_2x + b_2y + c_2z + d_2 = 0,$$

and let the corresponding normal vectors be

$$\mathbf{n}_1 = \langle a_1, b_1, c_1 \rangle, \quad \mathbf{n}_2 = \langle a_2, b_2, c_2 \rangle.$$

To solve this system of equations, we first check whether the equations are independent if we were to set $z = 0$. That is, we must check whether

$$a_1b_2 - b_1a_2 = 0,$$

or, equivalently, whether a_1 and b_1 are proportional to a_2 and b_2. If they are *not* proportional, then we can set $z = 0$ to obtain the system of equations

$$\begin{aligned} a_1x + b_1y &= -d_1, \\ a_2x + b_2y &= -d_2, \end{aligned}$$

which is now guaranteed to have a unique solution. This gives us a point on the line that is common to both planes. If $a_1b_2 - b_1a_2 = 0$, then we cannot necessarily substitute $z = 0$, for the resulting system of equations might be inconsistent. Instead, we can set $y = 0$, in which case the resulting system of equations, for the unknowns x and z, will have a unique solution.

To determine the *direction* of the line of intersection, we note that any vector in a plane is orthogonal to its normal vector. Because this line belongs to *both* planes, a vector in the direction of the line is orthogonal to both normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$. It follows that a vector $\mathbf{v}$ in the direction of the line of intersection can be found by computing

$$\mathbf{v} = \mathbf{n}_1 \times \mathbf{n}_2.$$

This direction, and the previously computed point on the line, can be used to obtain the parametric or symmetric equations of the line.

Example Consider two planes defined by the equations

$$x + 3y - 2z + 10 = 0, \quad 2x - 4y + 3z - 5 = 0.$$

These planes are not parallel, because their normal vectors $\mathbf{n}_1 = \langle 1, 3, -2 \rangle$ and $\mathbf{n}_2 = \langle 2, -4, 3 \rangle$ are not parallel. Their intersection is a line that is parallel to the vector

$$\mathbf{v} = \mathbf{n}_1 \times \mathbf{n}_2 = \langle 1, 3, -2 \rangle \times \langle 2, -4, 3 \rangle = \langle 3(3) - (-2)(-4), (-2)(2) - 1(3), 1(-4) - 3(2) \rangle = \langle 1, -7, -10 \rangle.$$

To write down the equation of the line of intersection, we need to compute the coordinates of a point on the line. Substituting $z = 0$ into the equations of the plane yields the system

$$\begin{aligned} x + 3y &= -10, \\ 2x - 4y &= 5. \end{aligned}$$

This system has a unique solution, because the coefficients of the equations are not proportional. Subtracting twice the first equation from the second yields the simpler equation $-10y = 25$, so $y = -5/2$. Substituting this value into the first equation yields $x = -10 - 3(-5/2) = -5/2$. We conclude that the line can be described using the parametric equations

$$x = -5/2 + t, \quad y = -5/2 - 7t, \quad z = -10t.$$

We can also describe the line using the symmetric equations

$$\frac{x + 5/2}{1} = \frac{y + 5/2}{-7} = \frac{z}{-10}.$$

□

2.5.5 Distance from a Point to a Plane

Let $\mathbf{p}_1 = \langle x_1, y_1, z_1 \rangle$ be a position vector corresponding to a point $P = (x_1, y_1, z_1)$. Let $(\mathbf{r} - \mathbf{r}_0) \cdot \mathbf{n} = 0$ be the equation of a plane, where $\mathbf{r}_0 = \langle x_0, y_0, z_0 \rangle$ is the position vector for a point $R_0 = \langle x_0, y_0, z_0 \rangle$ in the plane, and $\mathbf{n}$ is the plane's normal vector. We consider the problem of computing the distance D between the point P_1 and this plane.

Intuitively, it makes sense to define this distance as the distance from P_1 to some point P_2 contained within the plane. However, we need to

determine what a suitable point P_2 would be. We choose P_2 to be the *best approximation* of P_1 by a point in the plane, just as the vector projection of a vector $\mathbf{v}$ onto a vector $\mathbf{u}$ was previously defined to be the best approximation of $\mathbf{v}$ by a vector that is parallel to $\mathbf{u}$.

The key characteristic of the best approximation of the point P_1 by a point P_2 in the plane is that the error in this approximation, that is, the vector between P_1 and P_2, should be orthogonal to the plane. That is, this vector should be parallel to $\mathbf{n}$, the normal to the plane. To determine the length of this vector, we form a triangle with the points $\mathbf{p}_1$, $\mathbf{p}_2$ and $\mathbf{r}_0$, where $\mathbf{p}_2$ is the position vector for the point P_2.

Because $\mathbf{p}_1 - \mathbf{p}_2$ is parallel to $\mathbf{n}$, which is orthogonal to $\mathbf{p}_2 - \mathbf{r}_0$, this triangle is a right triangle, with the hypotenuse defined by $\mathbf{p}_1 - \mathbf{r}_0$. Therefore, we can use right triangle trigonometry to determine that the distance D is given by

$$D = |\mathbf{v}_1| \cos\theta$$

where $\mathbf{v}_1 = \mathbf{p}_1 - \mathbf{r}_0$ and θ is the angle between $\mathbf{v}_1$ and $\mathbf{n}$, with $\mathbf{n}$ chosen so that $0 \leq \theta < \pi/2$. It follows that

$$D = \frac{|\mathbf{v}_1 \cdot \mathbf{n}|}{|\mathbf{n}|}.$$

It is interesting to note that from this formula, we can see that D is also the absolute value of the scalar projection of $\mathbf{v}_1$ onto $\mathbf{n}$, or, equivalently, the magnitude of the *vector* projection of $\mathbf{v}_1$ onto $\mathbf{n}$.

If $\mathbf{n} = \langle a, b, c \rangle$, and we write the equation of the plane in the form $ax + by + cz + d = 0$, then we can express this distance as

$$D = \frac{|a(x_1 - x_0) + b(y_1 - y_0) + c(z_1 - z_0)|}{\sqrt{a^2 + b^2 + c^2}} = \frac{|ax_1 + by_1 + cz_1 + d|}{\sqrt{a^2 + b^2 + c^2}},$$

because the equations of the plane, $(\mathbf{r} - \mathbf{r}_0) \cdot \mathbf{n} = 0$ and $ax + by + cz + d = 0$, are related by $d = -\mathbf{r}_0 \cdot \mathbf{n} = -(ax_0 + by_0 + cz_0)$.

Example We wish to compute the distance D between the point $P_1 = (4, 5, 6)$ and the plane described by the linear equation

$$2x + 3y - 4z + 15 = 0.$$

The normal vector for this plane is $\mathbf{n} = \langle a, b, c \rangle = \langle 2, 3, -4 \rangle$. It follows that the distance D is given by

$$D = \frac{|2(4) + 3(5) - 4(6) + 15|}{\sqrt{2^2 + 3^2 + (-4)^2}} = \frac{14}{\sqrt{29}} \approx 2.6.$$

□

The formula for the distance between a point and a plane can be used to compute the distance between two parallel planes. The idea is to identify one point in the first plane, and then compute the distance between this point and the second plane. Because the planes are parallel, this distance will be the same, regardless of which point from the first plane is chosen.

2.5.6 Summary

- The vector equation of a line is of the form $\mathbf{r} = \mathbf{r}_0 + t\mathbf{v}$, where $\mathbf{r}_0$ is the position vector of a particular point on the line, t is a scalar parameter, $\mathbf{v}$ is a vector that describes the direction of the line, and $\mathbf{r}$ is the position vector of the point on the line corresponding to the value of t.
- The parametric equations of the line are the components of the vector equation, and have the form $x = x_0 + at$, $y = y_0 + bt$, and $z = z_0 + ct$. The components a, b and c of $\mathbf{v}$ are called the direction numbers of the line.
- The symmetric equations of a line are obtained by eliminating the parameter t from the parametric equations. The expressions $(x - x_0)/a$, $(y - y_0)/b$, and $(z - z_0)/b$, for which the direction numbers are nonzero, are equated.
- The line segment between $\mathbf{r}_0$ and $\mathbf{r}_1$ can be described by the equation $\mathbf{r} = (1 - t)\mathbf{r}_0 + t\mathbf{r}_1$, where $0 \leq t \leq 1$.
- If the expressions for x, y and z in the parametric equations of two lines are equal for the same values of the parameters, then the lines intersect at the point obtained by substituting the common parameters into the appropriate parametric equations.
- If two lines have direction vectors that are parallel, then the lines are also parallel.
- If two lines do not intersect, and are not parallel, then they are skew.
- A system of two linear equations with two unknowns can have a unique solution, no solution, or infinitely many solutions. It has a unique solution if the determinant of the system is nonzero.
- To determine whether two non-parallel lines intersect, one can try to find a solution to a system of three equations in two unknowns,

obtained by equating the parametric equations of corresponding components. The unknowns correspond to the parameters of the two lines.

- If the system of any two of the three equations is inconsistent, then the lines are skew.
- If the system of any two of the three equations has a unique solution, then the lines intersect if that solution satisfies the remaining equation, and the lines are skew if it does not.
- If the system of any two of the three equations has infinitely many solutions, then the lines intersect, and the parameters of the intersection point can be obtained by computing the solution of any other pair of equations, which is guaranteed to be unique.
- The vector equation of a plane is $\mathbf{n} \cdot (\mathbf{r} - \mathbf{r}_0) = 0$, where $\mathbf{n}$ is a vector that is normal to the plane, $\mathbf{r}$ is any position vector in the plane, and $\mathbf{r}_0$ is a given position vector in the plane. The normal vector $\mathbf{n}$ can be obtained by computing the cross product of any two non-parallel vectors in the plane.
- Two planes are parallel if and only if their normal vectors are parallel.
- If two planes are not parallel, their intersection is a line. The direction of the line is a vector that is orthogonal to the planes' normal vectors. A point on the line can be found by finding a solution of the system of equations consisting of the equations of the planes, which can be accomplished by setting one of the coordinates equal to zero.
- The distance between a point $\mathbf{p}$ and a plane $\mathbf{n} \cdot (\mathbf{r} - \mathbf{r}_0) = 0$ is the absolute value of the dot product of the unit (normalized) normal vector $\mathbf{n}/|\mathbf{n}|$ and the vector between $\mathbf{p}$ and $\mathbf{r}_0$.

2.6 Review

You should now be able to complete the following types of problems:

- Determining the center and radius of a sphere, given the equation of the sphere, by completing the square.
- Using inequalities to describe a region in three-dimensional space, given a verbal description of the region.

- Normalizing and scaling vectors in order to obtain a vector of a specified length and direction.

- Computing the resultant of given velocity vectors by vector addition, and computing the actual speed of an object, which is the magnitude of the resultant, as well as the actual direction. This involves computing the components of a vector $\mathbf{v}$ in the xy-plane from its magnitude and direction, using the formula

$$\mathbf{v} = \langle r\cos\theta, r\sin\theta\rangle,$$

  where $r = |\mathbf{v}|$ and θ is the angle that $\mathbf{v}$ makes with the positive x-axis. To compute the angle of a vector $\mathbf{v} = \langle x, y\rangle$ from its components, use the relationship

$$\tan\theta = \frac{y}{x}, \quad x \neq 0,$$

  keeping in mind that if $x < 0$, it is necessary to add π to the angle obtained by computing $\tan^{-1}(y/x)$ on your calculator. If $x = 0$ and $y > 0$, the angle is $\pi/2$, whereas if $x = 0$ and $y < 0$, the angle is $-\pi/2$.

- Computing the angle θ between two vectors $\mathbf{u}$ and $\mathbf{v}$, using the relation $\mathbf{u}\cdot\mathbf{v} = |\mathbf{u}||\mathbf{v}|\cos\theta$.

- Computing the scalar projection of $\mathbf{u}$ onto $\mathbf{v}$, defined by

$$\text{comp}_{\mathbf{v}}\mathbf{u} = \frac{\mathbf{v}\cdot\mathbf{u}}{|\mathbf{v}|},$$

  and the vector projection of $\mathbf{u}$ onto $\mathbf{v}$, defined by

$$\text{proj}_{\mathbf{v}}\mathbf{u} = \frac{\mathbf{v}\cdot\mathbf{u}}{|\mathbf{v}|^2}\mathbf{v}.$$

  Note that if we first compute the unit vector $\mathbf{w}$ in the direction of $\mathbf{v}$,

$$\mathbf{w} = \frac{\mathbf{v}}{|\mathbf{v}|},$$

  then these projections can be defined using the simpler formulas

$$\text{comp}_{\mathbf{v}}\mathbf{u} = \mathbf{w}\cdot\mathbf{u}, \quad \text{proj}_{\mathbf{v}}\mathbf{u} = (\mathbf{w}\cdot\mathbf{u})\mathbf{w}.$$

- Computing the work done by a force $\mathbf{F}$ applied to an object to move it along a displacement vector $\mathbf{D}$, which can be defined in terms of its initial and terminal points. The work is given by

$$W = \mathbf{F}\cdot\mathbf{D} = |\mathbf{F}||\mathbf{D}|\cos\theta,$$

where θ is the angle between $\mathbf{F}$ and $\mathbf{D}$, because $|\mathbf{F}|\cos\theta$ is the magnitude of the force applied to the object along its direction of motion, $|\mathbf{D}|$ is the distance that the object is moved, and work is the product of force and distance.

- Computing a unit vector $\mathbf{u}$ that is orthogonal to given vectors $\mathbf{v}$ and $\mathbf{w}$. This vector can be obtained by computing the cross product $\mathbf{v}\times\mathbf{w}$, and then normalizing this vector (that is, dividing by its magnitude) to obtain a unit vector.

- Computing the area of a parallelogram, given two vectors that define adjacent edges. This is accomplished by computing the magnitude of the cross product of the vectors.

- Computing the area of a triangle, given its vertices. This is accomplished by computing vectors between pairs of vertices, and then computing the magnitude of their cross product. This yields the area of the parallelogram defined by these vectors, which is twice the area of the triangle.

- Computing the volume of a parallelepiped defined by three given vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$. This is accomplished by computing the absolute value of the triple product, $|\mathbf{u}\cdot(\mathbf{v}\times\mathbf{w})|$.

- Determining whether two given lines are parallel, intersecting or skew. The lines are parallel if their direction numbers are proportional. Otherwise, equate corresponding parametric equations and determine if the resulting system of three equations has a solution. This is accomplished by trying to solve any two of the equations, written in the form

$$\begin{aligned} as + bt &= u, \\ cs + dt &= v, \end{aligned}$$

for the parameters of the lines, s and t. If there is no solution (which is true if the coefficients a, b and c, d are proportional but the right-hand side values u and v are not), the lines are skew. If there is a unique solution, which is true if these coefficients are not proportional, then the lines intersect if this solution satisfies the remaining equation; otherwise they are skew. If the two chosen equations have infinitely many solutions (that is, if the coefficients and right-hand side values are *all* proportional), then choose two other equations and repeat the process.

- Computing an equation of the plane containing three given points. This is accomplished by computing vectors between pairs of points, and taking their cross product to obtain a normal vector $\mathbf{n}$. Then, the vector equation of the plane has the form $\mathbf{n} \cdot (\mathbf{r} - \mathbf{r}_0) = 0$, where $\mathbf{r}_0$ is the position vector for any of the three points, and $\mathbf{r} = \langle x, y, z \rangle$ is the position vector for any point in the plane.

- Computing an equation of the line of intersection of two non-parallel planes. This is accomplished by first computing the direction numbers of the line, which can be obtained by computing the cross product of the normal vectors of the planes. Then, it is necessary to compute a point that is in both planes by finding a solution of their equations. This can be accomplished by setting one of the coordinates x, y or z equal to zero in both equations, and solving for the other two coordinates. If a solution cannot be found, then choose a different coordinate to set to zero.

- Computing the distance between a point and a plane. If the equation of the plane is $ax + by + cz + d = 0$, and the point is $P = (x_1, y_1, z_1)$, then the distance D is given by

$$D = \frac{|ax_1 + by_1 + cz_1 + d|}{\sqrt{a^2 + b^2 + c^2}}.$$

  Equivalently, if the plane is described by a vector equation $\mathbf{n} \cdot (\mathbf{r} - \mathbf{r}_0)$, and the point P is identified by a position vector $\mathbf{r}_1 = \langle x_1, y_1, z_1 \rangle$, then

$$D = \frac{|\mathbf{n} \cdot (\mathbf{r}_1 - \mathbf{r}_0)|}{|\mathbf{n}|}.$$

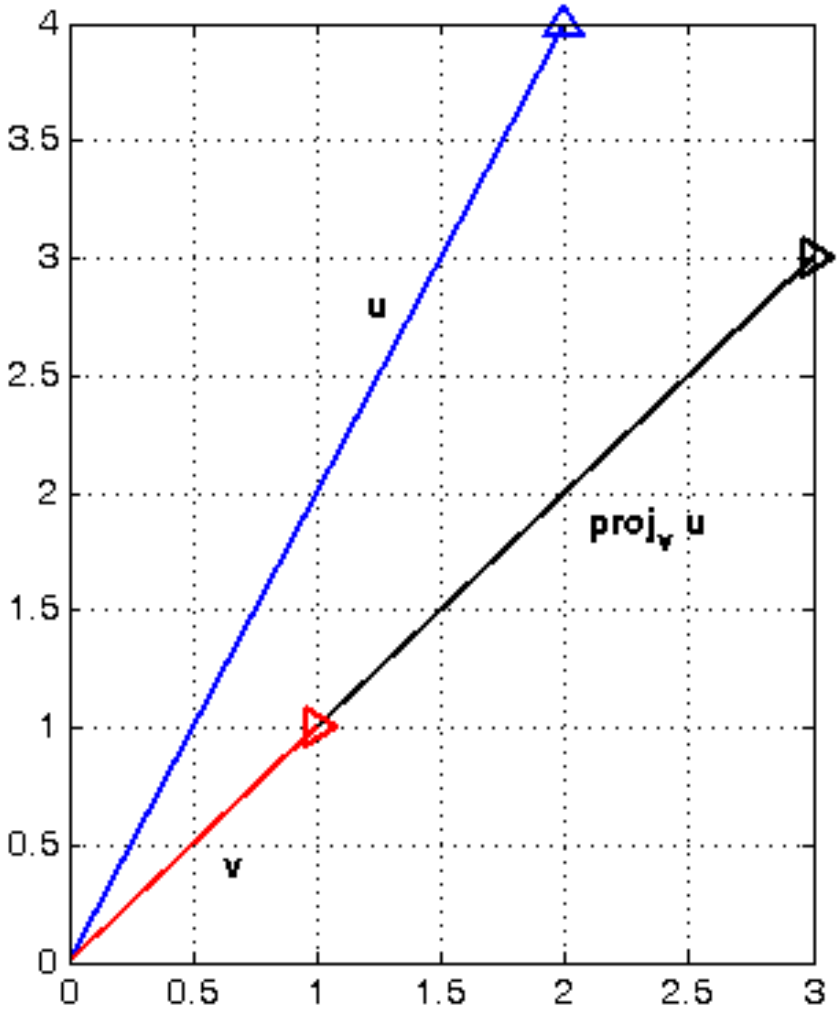

Figure 2.7: The vector projection (black vector) of $\mathbf{u} = \langle 2, 4 \rangle$ (blue vector) onto $\mathbf{v} = \langle 1, 1 \rangle$ (red vector).

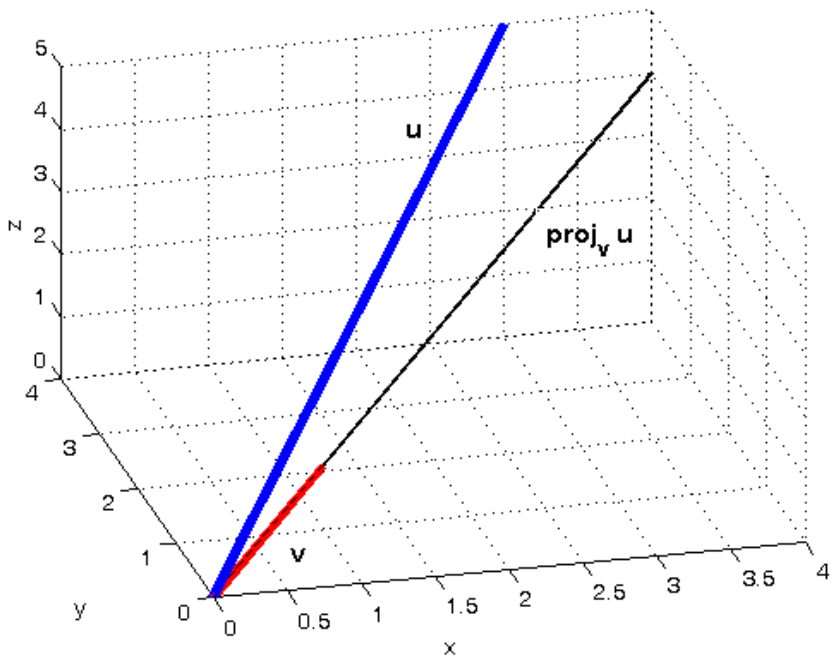

Figure 2.8: The vector projection of $\mathbf{u} = \langle 3, 4, 5 \rangle$ onto $\mathbf{v} = \langle 1, 1, 1 \rangle$.

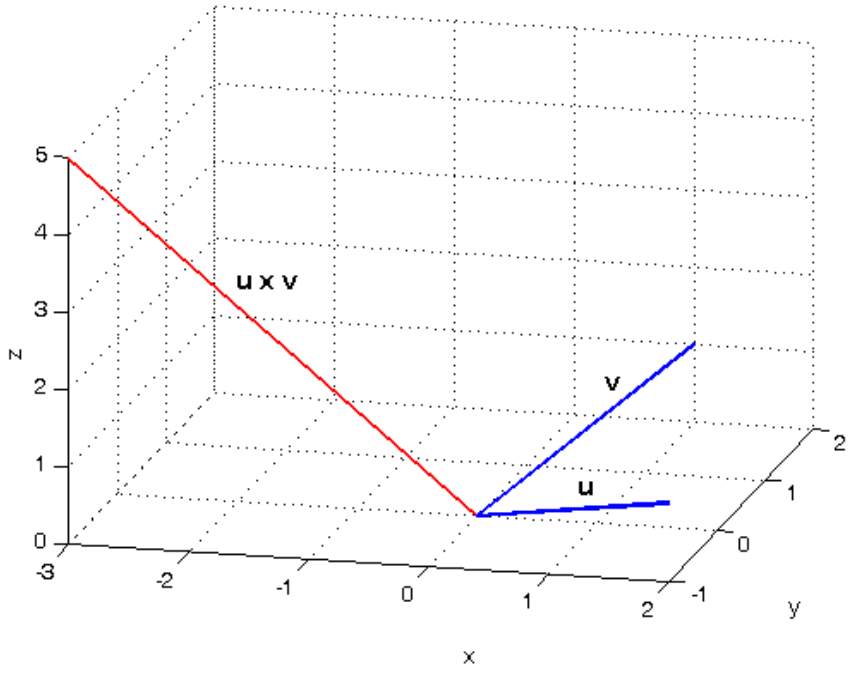

Figure 2.9: The cross product of $\mathbf{u} = \langle 2, -1, 1 \rangle$ and $\mathbf{v} = \langle 1, 2, 2 \rangle$ is the vector $\mathbf{u} \times \mathbf{v} = \langle -3, -1, 5 \rangle$.

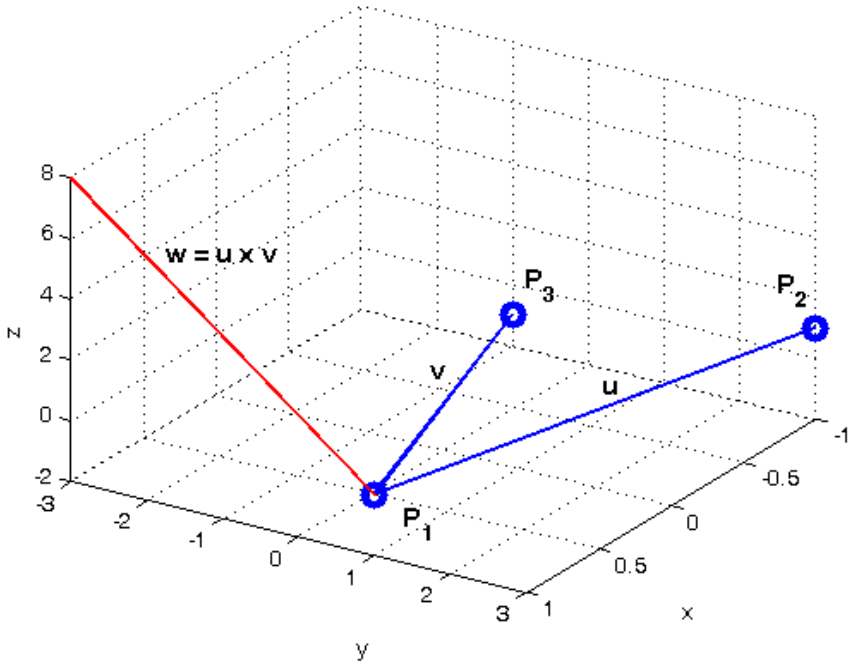

Figure 2.10: The vector $\mathbf{w} = \mathbf{u} \times \mathbf{v}$ is orthogonal to the plane determined by the points P_1, P_2, and P_3.

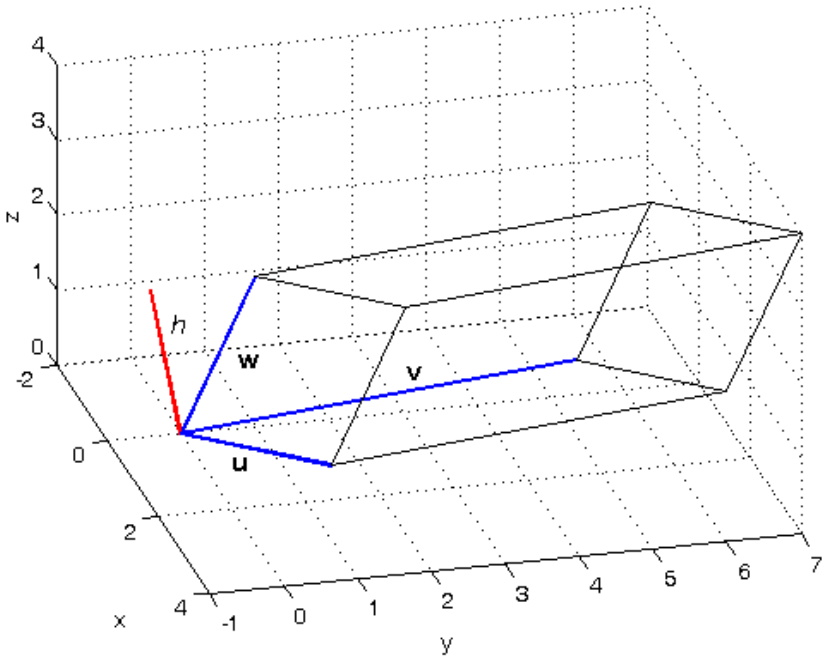

Figure 2.11: A parallelepiped defined by the vectors $\mathbf{u} = \langle 3, 1, 1\rangle$, $\mathbf{v} = \langle 1, 5, 1\rangle$, and $\mathbf{w} = \langle 0, 1, 2\rangle$. The height, which is measured along the direction of $\mathbf{z} = \mathbf{u} \times \mathbf{v}$, is indicated by the scalar h.

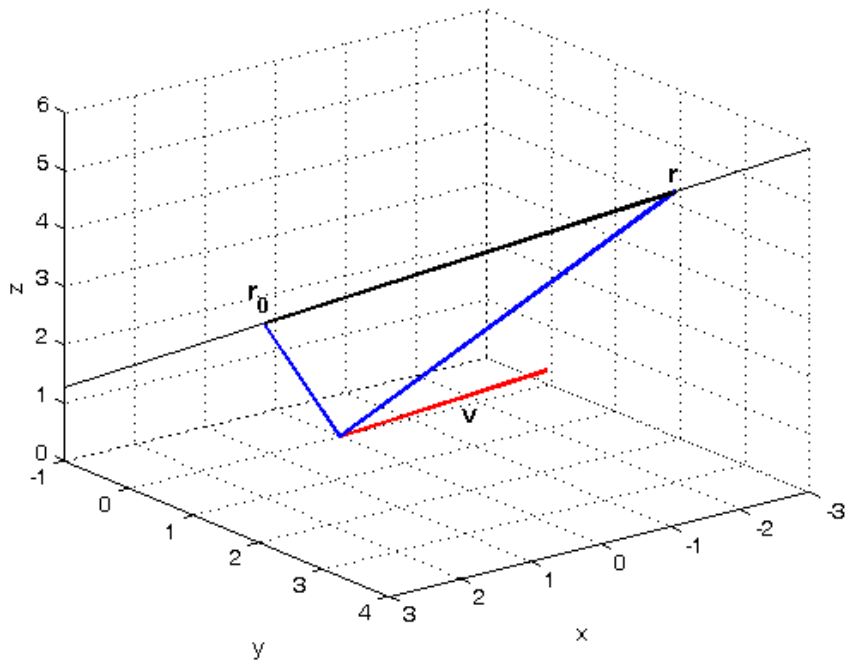

Figure 2.12: The line described by the vector equation $\mathbf{r} = \mathbf{r}_0 + t\mathbf{v}$, where $\mathbf{r}_0 = \langle 2, 1, 3\rangle$ and $\mathbf{v} = \langle -2, 1, 1\rangle$. The position vector $\mathbf{r}$ shown is $\langle -2, 3, 5\rangle$.

Chapter 3

Parametric Curves and Polar Coordinates

3.1 Parametric Curves

There are many useful curves that cannot be described by an equation of the form $y = f(x)$, because f is a function and therefore requires that only one y-value be associated with every x-value. For example, a complete circle cannot be described by such an equation. In such cases, we can instead describe the curve by *parametric equations*

$$x = f(t), \quad y = g(t),$$

where the variable t is called a *parameter*, and the curve defined by these equations is called a *parametric curve*. For example, a circle of radius r can be defined by the parametric equations

$$x = r\cos t, \quad y = r\sin t.$$

The parameter t is typically restricted to some interval $[a, b]$. The point $x = f(a)$, $y = g(a)$ is then called the *initial point* of the curve, and the point $x = f(b)$, $y = g(b)$ is called the *terminal point* of the curve.

Because any functions $f(t)$ and $g(t)$ can be chosen to define the x-coordinates and y-coordinates, respectively, of points on the curve, there is no requirement that each x-value is associated with only y-value, as with a curve defined by the equation $y = f(x)$. It follows that *any* curve in the plane can be defined using parametric equations.

Example Construct parametric equations of the form

$$x = f(t), \quad y = g(t)$$

that describe the unit circle.

Solution The unit circle is a circle of radius 1 with center at the origin $(0, 0)$. It is described by the equation

$$x^2 + y^2 = 1.$$

Choosing $f(t) = \cos t$ and $g(t) = \sin t$, where $0 \leq t \leq 2\pi$, we find that x and y satisfy this equation and describe the entire circle. If we let t denote time, and let $(x, y) = (f(t), g(t))$ denote the position of a particle at time t, then the particle begins at the point $(1, 0)$ (corresponding to $t = 0$) and moves once around the circle in the counterclockwise direction, at constant speed.

An alternative description of this circle is given by the parametric equations

$$x = \sin e^t, \quad y = \cos e^t, \quad \ln \pi \leq t \leq \ln 3\pi.$$

In this case, a particle whose motion is described by these equations starts at the point $(0, -1)$ and travels once around the circle in the clockwise direction, at steadily increasing speed. □

Example Describe the differences between the following sets of parametric equations that represent the curve $y = x^3$, where $-\infty < t < \infty$:

1. $x = t$, $y = t^3$
2. $x = t^2$, $y = t^6$
3. $x = \sin t$, $y = \sin^3 t$.

Solution

1. These equations describe the entire curve $y = x^3$. A particle whose motion is described by these equations traces the curve from left to right, at constant speed in the x-direction.
2. These equations describe the portion of the curve in the right-half plane $x \geq 0$. A particle whose motion is described by these equations traces the curve from right to left as t increases from $-\infty$, until $t = 0$, at which point the particle turns around and retraces the curve from left to right, at constant speed in the x-direction.
3. These equations describe the portion of the curve for which $-1 \leq x \leq 1$ and $-1 \leq y \leq 1$. A particle whose motion is described by these equations traces the curve from left to right until reaching the point $(1, 1)$, at which point it turns around and retraces the curve from right to left until reaching the point $(-1, -1)$. This process continues forever as t increases.

□

Example Find parametric equations for the astroid $x^{2/3} + y^{2/3} = 1$.

Solution Writing the equation for the astroid as

$$(x^{1/3})^2 + (y^{1/3})^2 = 1,$$

we see that $x^{1/3}$ and $y^{1/3}$ can only assume values between -1 and 1. Therefore, we can use the identity $\sin^2\theta + \cos^2\theta = 1$ and let $x^{1/3} = \cos t$ and $y^{1/3} = \sin t$, which yields the equations

$$x = \cos^3 t, \quad y = \sin^3 t,$$

where $0 \leq t \leq 2\pi$. □

Example Find parametric equations for the ellipse

$$4x^2 + 9y^2 = 36.$$

Solution Rewriting the equation as

$$(2x)^2 + (3y)^2 = 6^2,$$

we see that $2x$ and $3y$ can only assume values between -6 and 6. Equating $2x = 6\cos t$ and $3y = 6\sin t$ yields the equations

$$x = 3\cos t, \quad y = 2\sin t,$$

where $0 \leq t < 2\pi$. □

Example Sketch the curve described by the parametric equations

$$x = \sin t, \quad y = \sin 2t,$$

where $0 \leq t \leq 2\pi$.

Solution The curve can be sketched by choosing several values of t in the interval $[0, 2\pi]$ and computing the corresponding values of x and y for each value of t. In Figure 3.1, the curve is plotted by using Matlab$^{\text{TM}}$ to compute x and y for $t = 0, 0.01, 0.02, \ldots$ all the way up to 2π, plotting the resulting points, and then connecting the points to obtain a smooth curve. □

Example Given a curve defined by the parametric equations

$$x = 3t + 2, \quad y = t - 1,$$

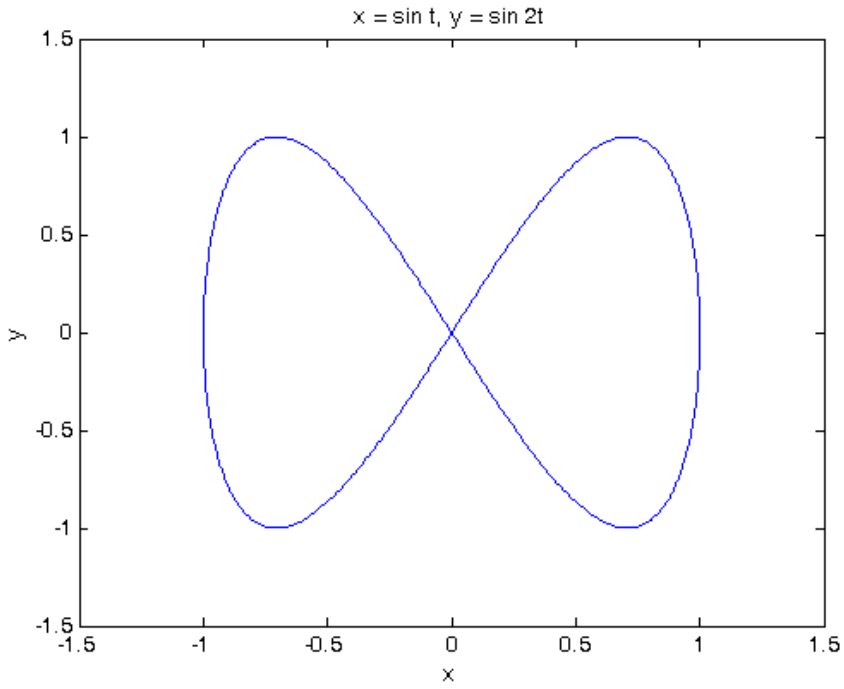

Figure 3.1: Curve defined by the parametric equations $x = \sin t$, $y = \sin 2t$.

eliminate the parameter t and obtain a Cartesian equation for the curve.

Solution By a Cartesian equation, we mean an equation of the form $y = f(x)$ or $x = f(y)$. In this case, we can obtain either type of equation since both x and y are one-to-one functions of t. We choose to obtain an equation of the form $y = f(x)$. Solving the equation $x = 3t + 2$ for t, we obtain

$$t = \frac{x-2}{3}.$$

Substituting this expression for t into the equation $y = t - 1$, we obtain the equation

$$y = f(x) = \frac{x-2}{3} - 1 = \frac{x-5}{3}.$$

□

Example Given a curve defined by the parametric equations

$$x = \sqrt{t+1}, \quad y = e^t,$$

where $t \geq 0$, eliminate the parameter t and obtain a Cartesian equation for the curve.

Solution Since x is a one-to-one function of t, we can solve the equation $x = \sqrt{t+1}$ for t and obtain

$$t = x^2 - 1,$$

where $x \geq 1$. Substituting this relation into the equation $y = e^t$, we obtain the Cartesian equation

$$y = e^{x^2-1}.$$

Since y is also a one-to-one function of t, we have the relation

$$t = \ln y,$$

where $y \geq 1$. We can substitute this relation into the equation $x = \sqrt{t+1}$ to obtain the alternative representation of the curve,

$$x = \sqrt{\ln y + 1}.$$

□

3.1.1 Summary

- A parametric curve in the xy-plane is a curve that is described by parametric equations $x = f(t)$ and $y = g(t)$, which define the x- and y-coordinates of each point on the curve as functions of a parameter t, where t belongs to an interval $[a, b]$.
- The initial point of the curve is $(f(a), g(a))$, and the terminal point is $(f(b), g(b))$.
- Any curve can be described by parametric equations, because parametric equations do not require that each x-value is associated with only one y-value, unlike an equation of the form $y = f(x)$.
- A curve defined by an equation of the form $[f(x)]^2 + [g(y)]^2 = r^2$ can be converted to parametric equations by equating $f(x) = r\cos t$ and $g(y) = r\sin t$, and solving for x and y.
- Parametric equations $x = f(t)$, $y = g(t)$ can be converted to an equation of the form $y = f(x)$ by solving $x = f(t)$ for t, if possible, and substituting the resulting expression for t into the equation $y = g(t)$.

3.2 Calculus with Parametric Curves

3.2.1 Arc Length

In this section, we will learn how to use calculus to compute the length of a curve that is described by an equation of the form $y = f(x)$, for some given function $f(x)$. Just as we learned how to compute the area under such a curve as the limit of a sum of areas of simpler regions (namely, rectangles), we can compute the length of the curve by interpreting the length as a limit of a sum of lengths of the simplest curves known, which are line segments.

Suppose that we wish to compute the length of the curve $y = f(x)$ between $x = a$ and $x = b$. We can approximate this length by dividing the interval $[a, b]$ into subintervals of length $\Delta x = (b - a)/n$, just as we did when we were trying to compute the approximate area under $y = f(x)$. Consider any subinterval $[x_{i-1}, x_i]$. Then, if Δx is chosen to be sufficiently small, the length of the curve $y = f(x)$ between $x = x_{i-1}$ and $x = x_i$ can be well approximated by the length of the line segment between the points $(x_{i-1}, f(x_{i-1}))$ and $(x_i, f(x_i))$. This line segment is the hypotenuse of a right triangle with legs of length Δx and $|f(x_i) - f(x_{i-1})|$, and therefore the length L_i of the curve $y = f(x)$ between $x = x_{i-1}$ and $x = x_i$ is approximately

$$L_i \approx \sqrt{\Delta x^2 + (f(x_i) - f(x_{i-1}))^2} = \Delta x \sqrt{1 + \left(\frac{f(x_i) - f(x_{i-1})}{\Delta x}\right)^2}. \quad (3.1)$$

It follows that the length L of the curve between $x = a$ and $x = b$ is approximated by

$$L \approx \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} \sqrt{1 + \left(\frac{f(x_i) - f(x_{i-1})}{\Delta x}\right)^2} \Delta x = \sum_{i=1}^{n} \sqrt{1 + \left(\frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}\right)^2} \Delta x. \quad (3.2)$$

As n, the number of line segments, approaches ∞, Δx approaches zero, so the length of each subinterval $[x_{i-1}, x_i]$ tends to zero. It follows from the definition of the derivative that

$$\lim_{\Delta x \to 0} \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} = \lim_{\Delta x \to 0} \frac{f(x_{i-1} + \Delta x) - f(x_{i-1})}{\Delta x} = f'(x_{i-1}) \quad (3.3)$$

and therefore the sum converges to the definite integral

$$L = \lim_{n \to \infty} \sum_{i=1}^{n} L_i = \int_a^b \sqrt{1 + [f'(x)]^2}\, dx. \quad (3.4)$$

The value of this integral is called the *arc length* of the curve $y = f(x)$ from $x = a$ to $x = b$. Similarly, if a curve is defined by the equation $x = f(y)$ from $y = c$ to $y = d$, the arc length of the curve is given by the definite integral

$$\int_c^d \sqrt{1 + [f'(y)]^2}\, dy. \tag{3.5}$$

Example 1 Compute the arc length of the curve $y = 2x + 3$, where $0 \leq x \leq 2$.

Solution Since the curve is just a line segment, we can simply use the distance formula to compute the arc length, since the arc length is the distance between the endpoints of the segment. The endpoints are $(0, 3)$ and $(2, 7)$, and therefore the arc length is

$$\sqrt{(2-0)^2 + (7-3)^2} = \sqrt{2^2 + 4^2} = \sqrt{20} = 2\sqrt{5}. \tag{3.6}$$

Using the arc length formula, we have $y' = 2$, and therefore the arc length is given by the integral

$$\int_0^2 \sqrt{1 + 2^2}\, dx = \int_0^2 \sqrt{5}\, dx = \sqrt{5}x\Big|_0^2 = 2\sqrt{5}. \tag{3.7}$$

□

Example 2 Compute the arc length of the curve $y = \sin x$ from $x = 0$ to $x = \pi$.

Solution Since $y' = \cos x$, the arc length is given by the integral

$$\int_0^\pi \sqrt{1 + \cos^2 x}\, dx. \tag{3.8}$$

Unfortunately, this integral cannot be evaluated using the Fundamental Theorem of Calculus. Using an approximation method such as the Composite Simpson's Rule, the value of the integral is seen to be approximately 3.8202. □

Example 3 Compute the arc length of the *astroid* described by the equation $x^{2/3} + y^{2/3} = 1$.

Solution We consider only the portion of the astroid in the upper quadrant $x \geq 0$, $y \geq 0$, which has endpoints $(0, 1)$ and $(1, 0)$. In this quadrant, the astroid can be described by the equation

$$y = (1 - x^{2/3})^{3/2}. \tag{3.9}$$

It follows that the arc length L of this segment of the astroid is given by the integral

$$\begin{aligned}
L &= \int_0^1 \sqrt{1+(y')^2}\,dx \\
&= \int_0^1 \sqrt{1+((3/2)(1-x^{2/3})^{1/2}(-(2/3)x^{-1/3})^2}\,dx \\
&= \int_0^1 \sqrt{1+(-(1-x^{2/3})^{1/2}x^{-1/3})^2}\,dx \\
&= \int_0^1 \sqrt{1+(1-x^{2/3})x^{-2/3}}\,dx \\
&= \int_0^1 \sqrt{1+(x^{-2/3}-1)}\,dx \\
&= \int_0^1 \sqrt{x^{-2/3}}\,dx \\
&= \int_0^1 x^{-1/3}\,dx \\
&= \frac{3}{2}x^{2/3}\Big|_0^1 \\
&= \frac{3}{2}. \qquad (3.10)
\end{aligned}$$

□

Example 4 Prove that the shortest distance between two given points is a straight line.

Solution For simplicity, we assume that the two points lie on the same horizontal line; specifically, the points are (a,k) and (b,k). Let $y=f(x)$ describe a curve connecting the two points. Then, the arc length of the curve is given by

$$\int_a^b \sqrt{1+[f'(x)]^2}\,dx. \qquad (3.11)$$

Since the integrand $\sqrt{1+[f'(x)]^2}$ is always positive, we can minimize the arc length by choosing $f(x)$ so that the integrand itself is minimized. This is the case when $f'(x)=0$; i.e., $f(x)$ is constant. Therefore the arc length is minimized when $f(x)=k$ and the corresponding curve is a straight line connecting the two points. □

In some cases, it is desirable to compute the arc length of a curve $y = f(x)$ as a function of its endpoints. For example, if the left endpoint of the curve is fixed at the point $(a, f(a))$ and we wish to know the arc length along this curve from the left endpoint to any other point $(x, f(x))$, then we can obtain this length as a function of x from the integral

$$s(x) = \int_a^x \sqrt{1 + [f'(t)]^2}\, dt. \tag{3.12}$$

The function $s(x)$ is known as the *arc length function*.

3.2.2 Arc Length of Parametrically Defined Curves

We have learned how to compute the arc length of a curve described by an equation of the form $y = f(x)$, where $a \le x \le b$. The arc length L of such a curve is given by the definite integral

$$L = \int_a^b \sqrt{1 + [f'(x)]^2}\, dx.$$

Now, suppose that this curve can also be defined by parametric equations

$$x = g(t), \quad y = h(t), \tag{3.13}$$

where $c \le t \le d$. It follows that

$$y = h(t) = f(g(t)),$$

and therefore, by the Chain Rule,

$$\frac{dy}{dt} = h'(t) = f'(g(t))g'(t) = f'(g(t))\frac{dx}{dt} = f'(x)\frac{dx}{dt}.$$

In the integral defining the arc length of the curve, we make the substitution $x = g(t)$ and obtain

$$\begin{aligned} L &= \int_a^b \sqrt{1 + [f'(x)]^2}\, dx \\ &= \int_c^d \sqrt{1 + [f'(g(t))]^2}\, g'(t)\, dt \\ &= \int_c^d \sqrt{[g'(t)]^2 + [f'(g(t))g'(t)]^2}\, dt \\ &= \int_c^d \sqrt{[g'(t)]^2 + [h'(t)]^2}\, dt \\ &= \int_c^d \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt. \end{aligned}$$

It turns out that this formula for the arc length applies to *any* curve that is defined by parametric equations of the form (3.13), as long as x and y are differentiable functions of the parameter t. To derive the formula in the general case, one can proceed as in the case of a curve defined by an equation of the form $y = f(x)$, and define the arc length as the limit as $n \to \infty$ of the sum of the lengths of n line segments whose endpoints lie on the curve.

Example Compute the length of the curve

$$x = 2\cos^2\theta, \quad y = 2\cos\theta\sin\theta,$$

where $0 \le \theta \le \pi$.

Solution This curve is plotted in Figure 3.2; it is a circle of radius 1 centered at the point $(1, 0)$. It follows that its length, which we will denote by L, is the circumference of the circle, which is 2π. Using the arc length formula, we can obtain the same result as follows:

$$\begin{aligned}
L &= \int_0^\pi \sqrt{\left(\frac{dx}{d\theta}\right)^2 + \left(\frac{dy}{d\theta}\right)^2}\, d\theta \\
&= \int_0^\pi \sqrt{(-4\cos\theta\sin\theta)^2 + \left(2\cos^2\theta - 2\sin^2\theta\right)^2}\, d\theta \\
&= 2\int_0^\pi \sqrt{(2\cos\theta\sin\theta)^2 + \left(\cos^2\theta - \sin^2\theta\right)^2}\, d\theta \\
&= 2\int_0^\pi \sqrt{4\cos^2\theta\sin^2\theta + \cos^4\theta - 2\cos^2\theta\sin^2\theta + \sin^4\theta}\, d\theta \\
&= 2\int_0^\pi \sqrt{\cos^4\theta + 2\cos^2\theta\sin^2\theta + \sin^4\theta}\, d\theta \\
&= 2\int_0^\pi \sqrt{(\cos^2\theta + \sin^2\theta)^2}\, d\theta \\
&= 2\int_0^\pi \sqrt{1^2}\, d\theta \\
&= 2\int_0^\pi d\theta \\
&= 2\pi.
\end{aligned}$$

Note: Double-angle and half-angle formulas could have been used in this example, but little would have been gained except during the differentiation stage, so I chose not to use them. □

Example Compute the length of the curve

$$x = t\sin t, \quad y = t\cos t$$

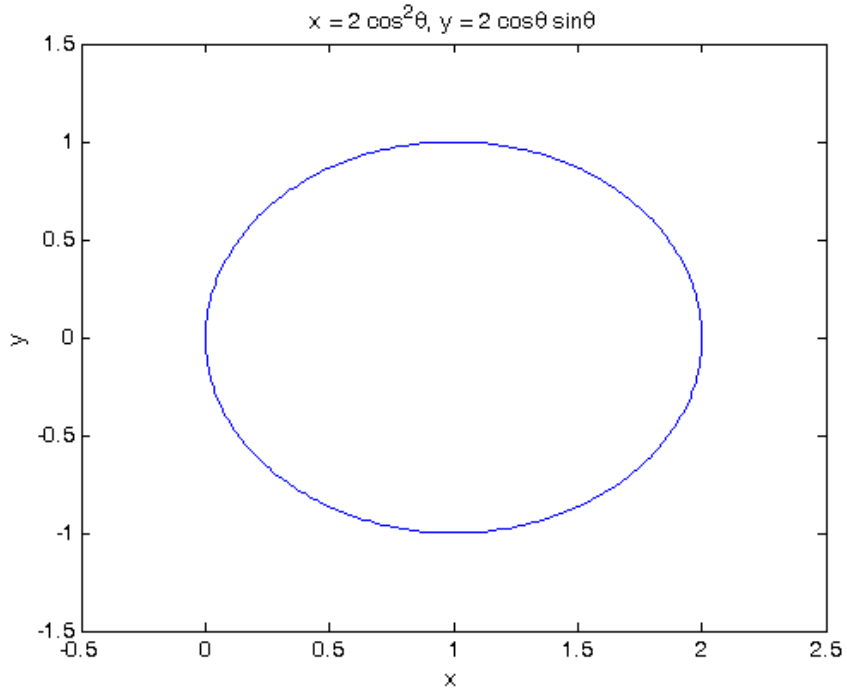

Figure 3.2: Curve defined by $x = \cos^2\theta$, $y = 2\cos\theta\sin\theta$

from $t = 0$ to $t = 2\pi$.

Solution The length L is given by the integral

$$\begin{aligned}
L &= \int_0^{2\pi} \sqrt{(\sin t + t\cos t)^2 + (\cos t - t\sin t)^2}\, dt \\
&= \int_0^{2\pi} \sqrt{\sin^2 t + 2\sin t\cos t + t^2\cos^2 t + \cos^2 t - 2\sin t\cos t + t^2\sin^2 t}\, dt \\
&= \int_0^{2\pi} \sqrt{(\sin^2 t + \cos^2 t) + t^2(\cos^2 t + \sin^2 t)}\, dt \\
&= \int_0^{2\pi} \sqrt{1 + t^2}\, dt \\
&= \int_0^{\tan^{-1}(2\pi)} \sqrt{1 + \tan^2\theta}\,\sec^2\theta\, d\theta
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^{\tan^{-1}(2\pi)} \sec^3\theta\, d\theta \\
&= \frac{1}{2}\left[\sec\theta\tan\theta + \ln|\sec\theta+\tan\theta|\right]\Big|_0^{\tan^{-1}(2\pi)} \\
&= \frac{1}{2}\left[\sqrt{1+\tan^2\theta}\tan\theta + \ln|\sqrt{1+\tan^2\theta}+\tan\theta|\right]\Big|_0^{\tan^{-1}(2\pi)} \\
&= \frac{1}{2}\left[\sqrt{1+(2\pi)^2}(2\pi) + \ln|\sqrt{1+(2\pi)^2}+2\pi|\right] \\
&\approx 21.2563.
\end{aligned}
$$

The integral of $\sec^3\theta$ is obtained using integration by parts, with $u = \sec\theta$ and $dv = \sec^2\theta\, d\theta$. The curve is displayed in Figure 3.3. □

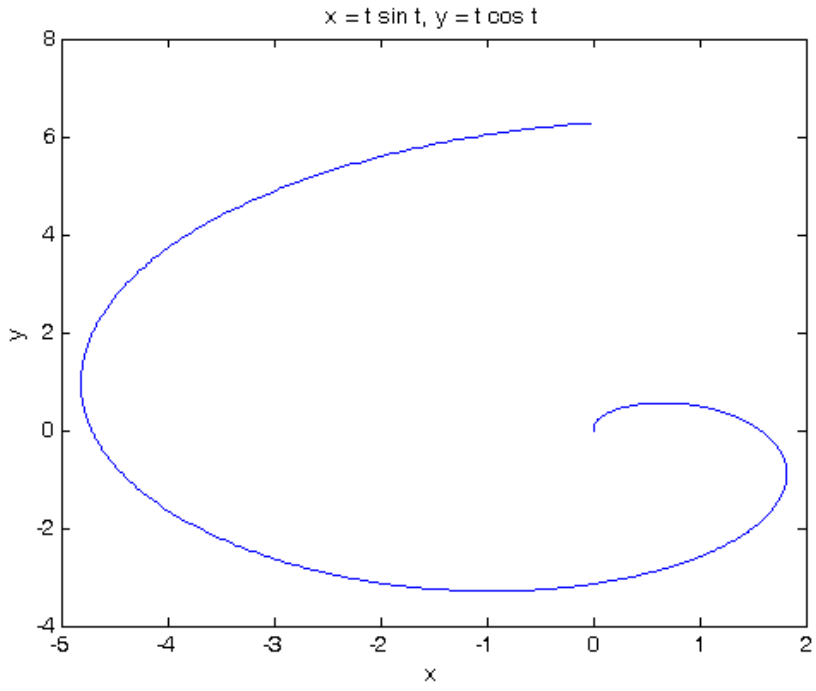

Figure 3.3: Curve defined by $x = t\sin t$, $y = t\cos t$

Example Find the distance traveled by a particle with position $x = \sin^2 t$, $y = \cos^2 t$, as t varies within the interval $[0, 3\pi]$. Compare with the length

of the curve.

Solution The distance traveled, D, can be computed using the arc length formula:

$$\begin{aligned}
D &= \int_0^{3\pi} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt \\
&= \int_0^{3\pi} \sqrt{(2\sin t\cos t)^2 + (-2\cos t\sin t)^2}\, dt \\
&= \int_0^{3\pi} \sqrt{8\sin^2 t\cos^2 t}\, dt \\
&= 2\sqrt{2}\int_0^{3\pi} |\sin t\cos t|\, dt \\
&= \sqrt{2}\int_0^{3\pi} |\sin 2t|\, dt \\
&= 6\sqrt{2}\int_0^{\pi/2} \sin 2t\, dt \\
&= 6\sqrt{2}\,\frac{1}{2}(-\cos 2t)\Big|\,0^{\pi/2} \\
&= 6\sqrt{2}.
\end{aligned}$$

In the sixth step, we used the fact that the *net* area under $\sin 2t$ from $t = \pi/2 + k\pi$ to $t = (k+1)\pi$, where k is an integer, is the negative of the area under $\sin 2t$ from $t = 0$ to $t = \pi/2$, as shown in Figure 3.4. Because we need to compute the area under $|\sin 2t|$, these areas must be negated to obtain the correct distance.

The length of the curve is $\sqrt{2}$, because the curve is traced once as t increases from 0 to $\pi/2$, and then retraced repeatedly every $\pi/2$ units in t. Therefore, from $t = 0$ to $t = 3\pi$, the curve is traced six times. □

3.2.3 Tangents of Parametric Curves

When a curve is described by an equation of the form $y = f(x)$, we know that the slope of the tangent line of the curve at the point $(x_0, y_0) = (x_0, f(x_0))$ is given by

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

However, if the curve is defined by parametric equations

$$x = f(t), \quad y = g(t),$$

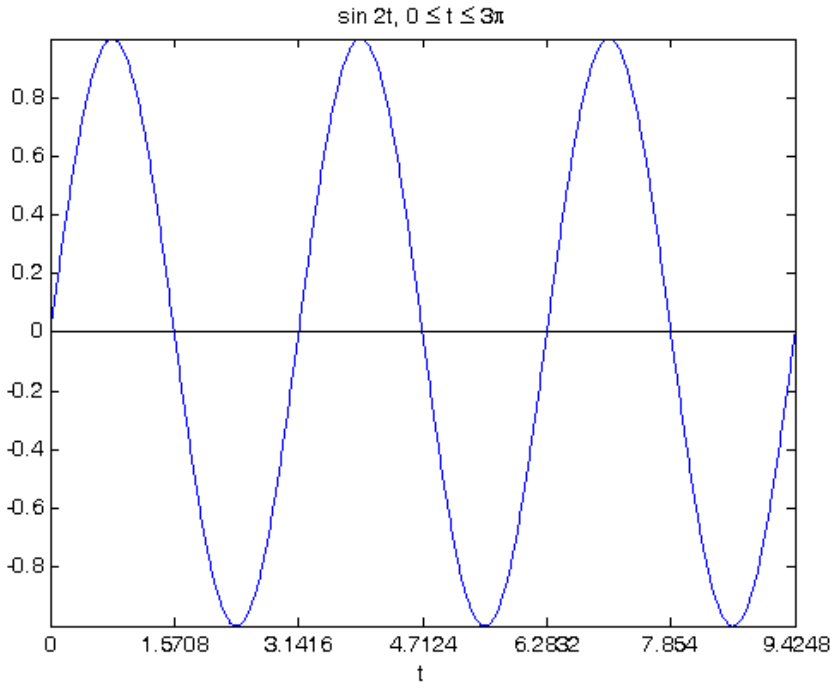

Figure 3.4: Graph of $\sin 2t$ for $0 \leq t \leq 3\pi$.

then we may not have a description of the curve as a function of x in order to compute the slope of the tangent line in this way. Instead, we apply the Chain Rule to obtain

$$\frac{dy}{dt} = \frac{dy}{dx}\frac{dx}{dt}.$$

Solving for dy/dx yields

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}}.$$

This allows us to express dy/dx as a function of the parameter t.

Example The slope of the tangent to the spiraling curve defined by

$$x = t\sin t, \quad y = t\cos t$$

which is shown in Figure 3.3, is given by

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{\cos t - t\sin t}{\sin t + t\cos t}.$$

At the point $(\pi/2, 0)$, which corresponds to $t = \pi/2$, the slope of the tangent is

$$m = \frac{\cos\frac{\pi}{2} - \frac{\pi}{2}\sin\frac{\pi}{2}}{\sin\frac{\pi}{2} + \frac{\pi}{2}\cos\frac{\pi}{2}} = \frac{0 - \frac{\pi}{2}\cdot 1}{1 + \frac{\pi}{2}\cdot 0} = -\frac{\pi}{2}.$$

From the point-slope form of the equation of a line, we see the equation of the tangent line of the curve at this point is given by

$$y - 0 = -\frac{\pi}{2}\left(x - \frac{\pi}{2}\right).$$

□

We know that a curve defined by the equation $y = f(x)$ has a horizontal tangent if $dy/dx = 0$, and a vertical tangent if $f'(x)$ has a vertical asymptote. For parametric curves, we also can identify a horizontal tangent by determining where $dy/dx = 0$. This is the case whenever $dy/dt = 0$, provided that $dx/dt = 0$, thus excluding the case where dy/dx is the indeterminate form $0/0$. Similarly, the tangent line is vertical whenever $dx/dt = 0$, but $dy/dt \neq 0$.

Example Consider the unit circle, which can be parametrized by the equations

$$x = \cos t, \quad y = \sin t, \quad 0 \leq t < 2\pi.$$

The slope of the tangent at any point on the circle is given by

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{\cos t}{-\sin t} = -\cot t.$$

A horizontal tangent occurs whenever $\cos t = 0$, and $\sin t \neq 0$. This is the case whenever $t = \pi/2$ or $t = 3\pi/2$. Substituting these parameter values into the parametric equations, we see that the circle has two horizontal tangents, at the points $(0, 1)$ and $(0, -1)$.

A vertical tangent occurs whenever $\sin t = 0$, and $\cos t \neq 0$. This is the case whenever $t = 0$ or $t = \pi$. Substituting these parameter values into the parametric equations, we see that the circle has two vertical tangents, at the points $(1, 0)$ and $(-1, 0)$. □

It is important to note that unlike a curve defined by $y = f(x)$, a point on the curve may have more than one tangent line, because a parametric curve is allowed to intersect itself.

Example Consider the curve defined by the parametric equations

$$x = t^2, \quad y = (t^2 - 4)\sin t.$$

This curve has two tangents at the point $(\pi^2, 0)$. To see this, we first note that $x = t^2 = \pi$ when $t = \pm\sqrt{\pi}$. Substituting these values into the equation for y, we obtain $y = 0$, since $\sin t = 0$ when $t = \pm\pi$. Therefore, there are two distinct parameter values corresponding to this point on the curve.

Next, we must compute dy/dx for both values of t. We have

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{(t^2 - 4)\cos t + 2t\sin t}{2t} = \sin t + \frac{t^2 - 4}{t}\cos t.$$

Substituting $t = -\pi$ yields

$$\frac{dy}{dx} = \sin(-\pi) + \frac{(-\pi)^2 - 4}{-\pi}\cos(-\pi) = \frac{\pi^2 - 4}{\pi} \approx 1.8684.$$

On the other hand, substituting $t = \pi$ yields

$$\frac{dy}{dx} = \sin\pi + \frac{\pi^2 - 4}{\pi}\cos\pi = -\frac{\pi^2 - 4}{\pi} \approx -1.8684.$$

The curve is illustrated in Figure 3.5. □

In order to graph curves, it is helpful to know where the curve is concave up or concave down. For a curve defined by $y = f(x)$, this is determined by computing its second derivative $d^2y/dx^2 = f''(x)$ and checking its sign. For a parametric curve, we can compute d^2y/dx^2 in the same way as dy/dx, by using the Chain Rule. First, we note that

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right).$$

Then, from the Chain Rule,

$$\frac{d}{dt}\left(\frac{dy}{dx}\right) = \frac{d}{dx}\left(\frac{dy}{dx}\right)\frac{dx}{dt} = \frac{d^2y}{dx^2}\frac{dx}{dt}.$$

Solving for d^2y/dx^2 yields

$$\frac{d^2y}{dx^2} = \frac{\frac{d}{dx}\left(\frac{dy}{dx}\right)}{\frac{dx}{dt}}.$$

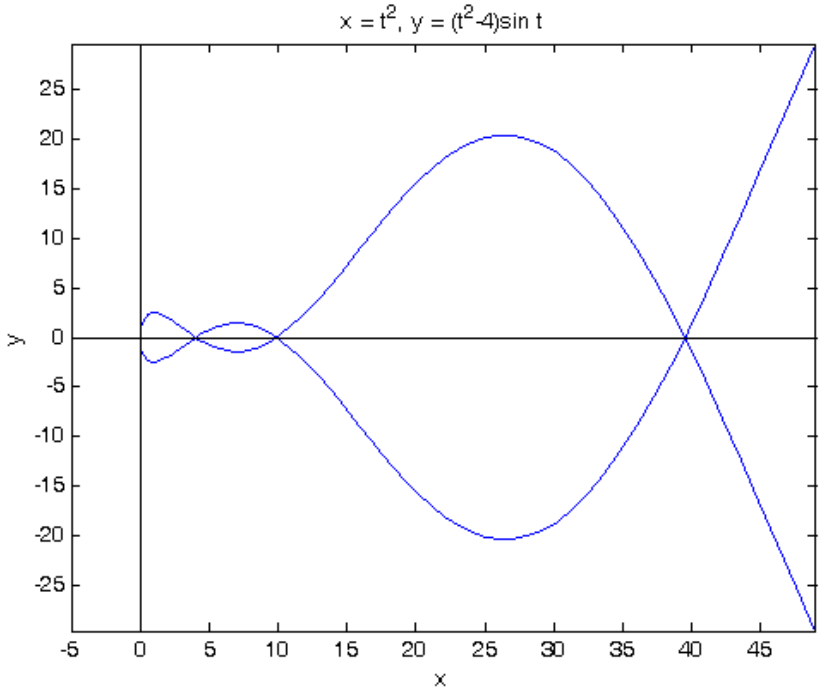

Figure 3.5: Graph of the parametric curve $x = t^2$, $y = (t^2 - 4)\sin t$.

To use this formula, one first computes dy/dx in terms of dy/dt and dx/dt, as described above. Then, dy/dx is a function of t, which can be differentiated with respect to t in the usual way, before being divided by dx/dt to obtain d^2y/dx^2.

It is possible to obtain a formula for d^2y/dx^2 that uses only derivatives of x and y with respect to t. By applying the Quotient Rule to differentiate dy/dx with respect to t, we obtain

$$\frac{d^2y}{dx^2} = \frac{\frac{dx}{dt}\frac{d^2y}{dt^2} - \frac{dy}{dt}\frac{d^2x}{dt^2}}{\left(\frac{dx}{dt}\right)^3},$$

although the first formula may be easier to remember.

Example Consider the astroid, defined by the parametric equations

$$x = \cos^3 t, \quad y = \sin^3 t, \quad 0 \leq t < 2\pi.$$

This curve is illustrated in Figure 3.6. To determine where the curve is concave up or concave down, we first compute dy/dx as a function of t:

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{3\sin^2 t\cos t}{-3\cos^2 t\sin t} = -\tan t.$$

Next, we use this to compute d^2y/dx^2:

$$\frac{d^2y}{dx^2} = \frac{\frac{d}{dt}\left(\frac{dy}{dx}\right)}{\frac{dx}{dt}} = \frac{-\sec^2 t}{-3\cos^2 t\sin t} = \frac{1}{3\cos^4 t\sin t}.$$

We conclude that the astroid is concave up whenever $\sin t > 0$, which is the case when $y > 0$. It is concave down whenever $\sin t < 0$, which is the case whenever $y < 0$. □

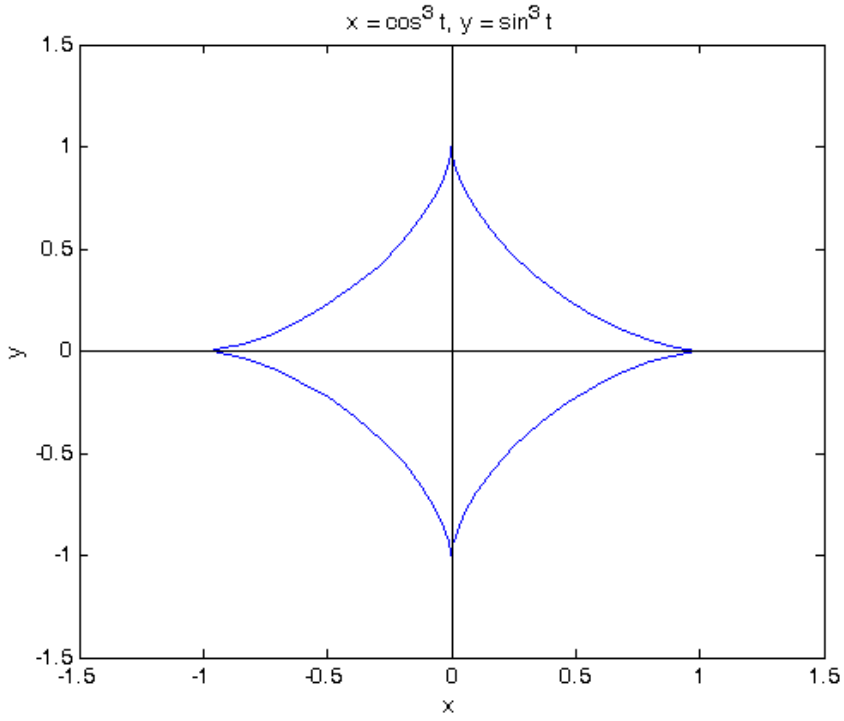

Figure 3.6: Graph of the astroid $x = \cos^3 t$, $y = \sin^3 t$, for $0 \le t < 2\pi$.

Example A string is wound around a circle and then unwound while being held taut. The curve traced by the point P at the end of the string is called

the *involute* of the circle. If the circle has radius r and center O and the initial position of P is $(r, 0)$, and if the parameter θ is chosen as in the figure (see page 495 of the text), show that parametric equations of the involute are

$$x = r(\cos\theta + \theta\sin\theta), \quad y = r(\sin\theta - \theta\cos\theta).$$

Solution For any angle θ, the position vector for the point P is the sum of two vectors $\mathbf{u}$ and $\mathbf{v}$, where $\mathbf{u}$ is the position vector $\mathbf{u}$ for the point T on the circle corresponding to θ, and $\mathbf{v}$ is the vector from T to P. The magnitude of $\mathbf{u}$ is r, the radius of the circle, so its components are $\mathbf{u} = \langle r\cos\theta, r\sin\theta\rangle$.

The magnitude of $\mathbf{v}$ is the length of the portion of the string that has been unwound, which is the length of the arc of the circle that begins at the point $(r, 0)$ and ends after sweeping counterclockwise (for increasing θ) through θ radians; this length is $r\theta$.

Because the string is held taut, it is tangent to the circle at T, so $\mathbf{v}$ and $\mathbf{u}$ are perpendicular; in fact, the angle made by $\mathbf{v}$ with the positive x-axis, when $\mathbf{v}$ is translated so that its initial point is at the origin, is always 90 degrees, or $\pi/2$ radians, less than that of $\mathbf{u}$. That is,

$$\mathbf{v} = \left\langle r\theta\cos\left(\theta - \frac{\pi}{2}\right), r\theta\sin\left(\theta - \frac{\pi}{2}\right)\right\rangle.$$

Using the identities

$$\cos(A-B) = \cos A\cos B + \sin A\sin B, \quad \sin(A-B) = \sin A\cos B - \cos A\sin B,$$

with $A = \theta$ and $B = \pi/2$, we obtain $\cos(\theta - \pi/2) = \sin\theta$ and $\sin(\theta - \pi/2) = -\cos\theta$. By adding the components of $\mathbf{u}$ and $\mathbf{v}$, we obtain the desired parametric equations for the involute. □

3.2.4 Areas Under Parametric Curves

Recall that the area A of the region bounded by the curve $y = F(x)$, the vertical lines $x = a$ and $x = b$, and the x-axis is given by the integral

$$A = \int_a^b F(x)\,dx.$$

Now, suppose that the curve $y = F(x)$ is also defined by the parametric equations $x = f(t)$, $y = g(t)$, for $\alpha \le t \le \beta$. Furthermore, suppose that $f(\alpha) = a$ and $f(\alpha) = b$. If the curve is traversed only once as t increases

from α to β, then the area can also be computed by integrating with respect to t as follows:

$$A = \int_a^b F(x)\,dx = \int_\alpha^\beta g(t)f'(t)\,dt.$$

On the other hand, if $t = \alpha$ corresponds to the *right* endpoint of the curve, and $t = \beta$ corresponds to the left endpoint, then limits of integration must be reversed:

$$A = \int_\beta^\alpha g(t)f'(t)\,dt = -\int_\alpha^\beta g(t)f'(t)\,dt.$$

Example The upper half-circle with center $(0, 0)$ and radius 1 can be defined by the parametric equations $x = \cos t$, $y = \sin t$, for $0 \le t \le \pi$. Because $t = 0$ corresponds to the right endpoint of this curve, and $t = \pi$ corresponds to the left endpoint, the area bounded by the upper half-circle and the x-axis is given by

$$A = \int_\pi^0 \sin t(-\sin t)\,dt = \int_\pi^0 -\sin^2 t\,dt = \int_0^\pi \sin^2 t\,dt = \int_0^\pi \frac{1-\cos 2t}{2}\,dt = \frac{t}{2} - \frac{\sin 2t}{4}\bigg|_0^\pi = \frac{\pi}{2},$$

which, as expected, is half of the area of the circle. □

Example Use the parametric equations of an ellipse, $x = a\cos\theta$, $y = b\sin\theta$, $0 \le \theta \le 2\pi$, to find the area that it encloses.

Solution Because of the symmetry of the ellipse, we can compute the area by first computing the are of the region bounded by the given curve, for $0 \le \theta \le \pi/2$, and the lines $x = 0$ and $y = 0$, and then multiply the result by 4. The area A of this region is given by the integral

$$A = \int_0^a y\,dx.$$

Substituting $x = a\cos\theta$, which yields $dx = -a\sin\theta$, and substituting $y = b\sin\theta$, we obtain an integral for the area in terms of θ,

$$A = \int_{\pi/2}^0 (b\sin\theta)(-a\sin\theta)\,d\theta = ab\int_0^{\pi/2} \sin^2\theta\,d\theta.$$

The limits $\theta = 0$ and $\theta = \pi/2$ arise from the fact that when $\theta = 0$, $x = a\cos 0 = a$, and when $\theta = \pi/2$, $x = a\cos\frac{\pi}{2} = 0$. In the last step, we interchanged the limits of integration, which changed the sign of the integral. Using the identity $\sin^2\theta = (1 - \cos 2\theta)/2$, we obtain

$$A = ab\int_0^{\pi/2} \left[\frac{1}{2} - \frac{1}{2}\cos 2\theta\right]\,d\theta = ab\left[\frac{\theta}{2} - \frac{1}{4}\sin 2\theta\right]\bigg|_0^{\pi/2} = \frac{\pi}{4}ab.$$

Multiplying by 4 yields the area of the entire ellipse, πab. □

Example Find the area of the region bounded by the curve $x = \cos t$, $y = e^t$, $0 \leq t \leq \pi/2$, and the lines $y = 1$ and $x = 0$. The region is shown in Figure 3.7.

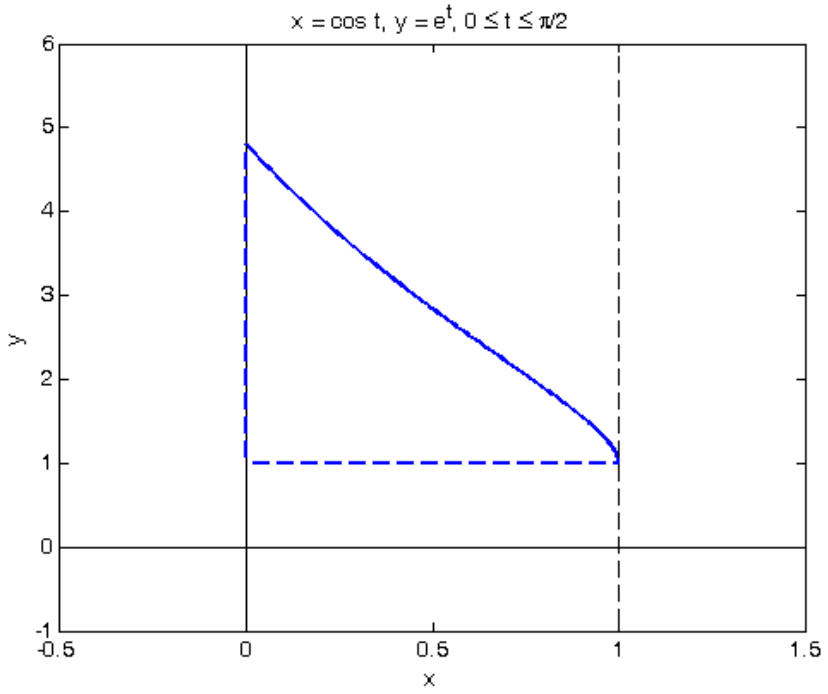

Figure 3.7: The region bounded by the lines $x = 0$, $y = 1$, and the curve $x = \cos t$, $y = e^t$, for $0 \leq t \leq \pi/2$.

Solution First, we note that as t increases from 0 to $\pi/2$, x decreases from 1 to 0, while y increases from 1 to $e^{\pi/2} \approx 4.81$. It follows that if we wish to compute the area by integrating y with respect to x, the limits of integration should be $x = 0$ and $x = 1$. Furthermore, because the region is bounded below by $y = 1$, we can compute the area of the prescribed region by first computing the area of the region bounded by $x = \cos t$, $y = e^t$, $y = 0$ and $x = 0$. Then, we can subtract the area of the rectangle bounded by the lines $x = 0$, $x = 1$, $y = 0$ and $y = 1$, which is 1.

We then have

$$A = \int_0^1 y\,dx - 1 = \int_{\pi/2}^0 e^t(-\sin t)\,dt - 1 = \int_0^{\pi/2} e^t \sin t\,dt.$$

Using integration by parts with $u = e^t$ and $dv = \sin t\,dt$, and a second time with $u = e^t$, $dv = \cos t\,dt$, yields

$$\begin{aligned}\int_0^{\pi/2} e^t \sin t\,dt &= -e^t \cos t\,dt\Big|_0^{\pi/2} + \int_0^{\pi/2} e^t \cos t\,dt \\ &= 1 + e^t \sin t\Big|_0^{\pi/2} - \int_0^{\pi/2} e^t \sin t\,dt \\ &= \frac{1}{2} + \frac{1}{2}e^{\pi/2}.\end{aligned}$$

In the last step, we moved the integral on the right side to the left side, because they are the same except for their sign, and then we divided by 2 to obtain the result. Subtracting 1 from this integral yields $A = \frac{1}{2}e^{\pi/2} - \frac{1}{2}$ for the area of the region.

Another way to compute the area is to integrate x with respect to y. In this case, the limits of integration are $y = 1$ and $y = e^{\pi/2}$. We then have

$$A = \int_1^{\exp(\pi/2)} x\,dy = \int_0^{\pi/2} e^t \cos t\,dt.$$

Using integration by parts as before yields the area. □

3.2.5 Summary

- The arc length of a curve defined by the equation $y = f(x)$, for $a \leq x \leq b$, is the integral of $\sqrt{1 + [f'(x)]^2}$ from a to b.
- Often, this integral cannot be computed analytically using known integration rules, so the arc length must instead be approximated numerically using a technique such as the Composite Simpson's Rule.
- If a curve is defined by parametric equations $x = g(t)$, $y = h(t)$ for $c \leq t \leq d$, the arc length of the curve is the integral of $\sqrt{(dx/dt)^2 + (dy/dt)^2} = \sqrt{[g'(t)]^2 + [h'(t)]^2}$ from c to d.
- The slope of the tangent line of a parametric curve defined by parametric equations $x = f(t)$, $y = g(t)$ is given by $dy/dx = (dy/dt)/(dx/dt)$.

- A parametric curve has a horizontal tangent wherever $dy/dt = 0$ and $dx/dt \neq 0$. It has a vertical tangent wherever $dx/dt = 0$ and $dy/dt \neq 0$.
- The concavity of a parametric curve at a point can be determined by computing $d^2y/dx^2 = d(dy/dx)/dt/(dx/dt)$, where dy/dt is best represented as a function of t, not x. The curve is concave up when d^2y/dx^2 is positive, and concave down if it is negative.
- A parametric curve $x = f(t)$, $y = g(t)$ can have two tangents at a point (x_0, y_0) on its graph, if there are two distinct values of the parameter t, t_1 and t_2, such that $f(t_1) = f(t_2) = x_0$ and $g(t_1) = g(t_2) = y_0$.
- The area of the region bounded by the parametric curve $x = f(t)$, $y = g(t)$, the x-axis, the line $x = a$, and the line $x = b$, where $f(\alpha) = a$ and $g(\beta) = b$, is the integral from α to β of $g(t)f'(t)\,dt$, provided that the curve is only traversed once as t increases from α to β.

3.3 Polar Coordinates

Throughout this course, we have denoted a point in the plane by an ordered pair (x, y), where the numbers x and y denote the directed (i.e., signed positive or negative) distance between the point and each of two perpendicular lines, the x-axis and the y-axis. The elements of this ordered pair are called *coordinates*, and the coordinates used in this particular method of identifying points in the plane are called *Cartesian coordinates*.

In this section, we introduce an alternative coordinate system known as the *polar coordinate system*. In this system, a point in the plane is identified by an ordered pair (r, θ), where:

- r is the directed distance from a point designated as the *pole*, and
- θ is the angle, in radians, that a ray between the pole and the point makes with a ray designated as the *polar axis*.

The coordinates r and θ are called *polar coodinates*.

The pole is the point $(0, 0)$ in Cartesian coordinates, and has polar coordinates $(0, \theta)$ for *any* value of θ. The polar axis corresponds to the positive x-axis. An angle θ is considered positive if measured in the counterclockwise direction from the polar axis, and negative if measured in the clockwise direction.

Example Sketch the region in the plane consisting of points whose polar coordinates satisfy $2 < r < 3$, $5\pi/3 \leq \theta \leq 7\pi/3$.

Solution A sketch of the region described by the given inequalities is obtained by first sketching two concentric circles, of radii 2 and 3, and the rays extending from the origin at angles $5\pi/3$ and $7\pi/3$. The region is illustrated in Figure 3.8. □

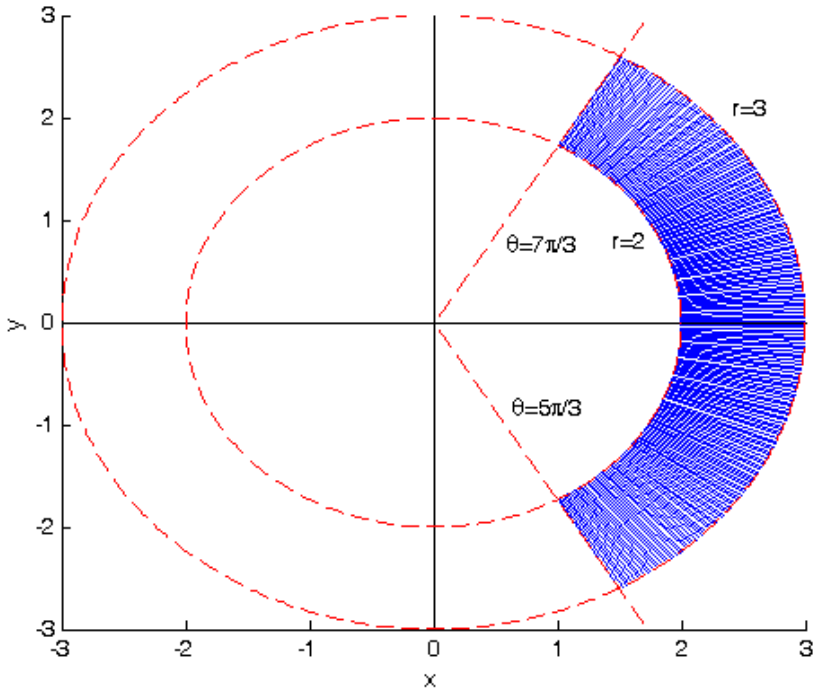

Figure 3.8: Region satisfying $2 < r < 3$ and $5\pi/3 \leq \theta \leq 7\pi/3$.

3.3.1 Conversion Between Cartesian and Polar Coordinates

Using these conventions, the Cartesian coordinates of a point can easily be obtained from the polar coordinates using the relations

$$x = r\cos\theta, \quad y = r\sin\theta.$$

Since $\sin\theta$ and $\cos\theta$ are not one-to-one, and since r is allowed to assume negative values, it follows that each point in the plane has infinitely many representations in polar coordinates.

Example Compute the Cartesian coordinates of the following points whose polar coordinates are given.

1. $(1, \pi/4)$
2. $(-1, 5\pi/4)$
3. $(1, 9\pi/4)$

Solution Using the relations

$$x = r\cos\theta, \quad y = r\sin\theta,$$

we have:

1. $x = 1 \cdot \cos(\pi/4) = \sqrt{2}/2$, $y = 1 \cdot \sin(\pi/4) = \sqrt{2}2$
2. $x = -\cos(5\pi/4) = -(-\sqrt{2}/2) = \sqrt{2}/2$, $y = -\sin(\pi/4) = -(-\sqrt{2}2) = \sqrt{2}/2$
3. $x = 1 \cdot \cos(9\pi/4) = \cos(\pi/4) = \sqrt{2}/2$, $y = 1 \cdot \sin(9\pi/4) = \sin(\pi/4) = \sqrt{2}/2$

□

The polar coordinates of a point can be obtained from the Cartesian coordinates as follows:

$$r = x^2 + y^2, \quad \tan\theta = \frac{y}{x}.$$

It should be noted that because $\tan\theta$ is not one-to-one on the interval $0 \leq \theta < 2\pi$, it is necessary to consider the signs of x and y in order to make sure that the proper value of θ is used to represent the point (x, y). Otherwise, the point (r, θ) may lie in the wrong quadrant of the plane.

Example Compute the polar coordinates of the following points whose Cartesian coordinates are given.

1. $(-\sqrt{3}/2, 1/2)$
2. $(-1, -1)$

Solution Using the relations

$$r^2 = x^2 + y^2, \quad \tan\theta = \frac{y}{x},$$

we have:

1. $$r^2 = (-\sqrt{3}/2)^2 + (1/2)^2 = 3/4 + 1/4 = 1, \quad \tan\theta = -\frac{1}{\sqrt{3}}.$$

   It follows that $r = 1$. Because the x-coordinate of the point is negative, we should seek a value of θ that lies in the interval $(\pi/2, 3\pi/2)$. However, the range of the inverse tangent function lies in the interval $(-\pi/2, \pi/2)$, and therefore $\theta = \tan^{-1}(-1/\sqrt{3}) = -\pi/6$. Since tangent has a *period* of π, it follows that

   $$\tan(\theta + \pi) = \tan\theta$$

   for any θ; in other words, its values repeat after every π units. Since

   $$\tan(5\pi/6) = \tan(-\pi/6 + \pi) = \tan(-\pi/6) = -\frac{1}{\sqrt{3}},$$

   it follows that $\theta = 5\pi/6$ satisfies the relation $\tan\theta = y/x$, and since $5\pi/6$ lies in the interval $(\pi/2, 3\pi/2)$, it is a correct value of θ for this point.

2. $$r^2 = (-1)^2 + (-1)^2 = 2, \quad \tan\theta = \frac{-1}{-1} = 1.$$

   It follows that $r = \sqrt{2}$. Because the x-coordinate of the point is negative, we should seek a value of θ that lies in the interval $(\pi/2, 3\pi/2)$. However, we have $\tan^{-1}(1) = \pi/4$, which is not in that interval. Since

   $$\tan(5\pi/4) = \tan(\pi/4 + \pi) = \tan(\pi/4) = 1,$$

   it follows that $\theta = 5\pi/4$ satisfies the relation $\tan\theta = y/x$, and since $5\pi/4$ lies in the interval $(\pi/2, 3\pi/2)$, it is a correct value of θ for this point.

□

3.3.2 Polar Equations

A *polar equation* is an equation of the form $r = f(\theta)$. Such an equation defines a curve in the plane by assigning a distance from the pole to each angle θ via the function $f(\theta)$. For example, the simple polar equation $r = k$, where k is a constant, describes a circle of radius k. The *graph* of a polar equation is the set of all points in the plane that can be described using polar coordinates that satisfy the equation. This definition is worded as such in

order to take into account that each point in the plane can have infinitely many representations in polar coordinates.

Example Sketch the curve described by the polar equation

$$r = \cos 2\theta, \quad 0 \leq \theta \leq 2\pi.$$

Solution This curve can be sketched by evaluating $r = \cos 2\theta$ at several values of θ. For each such value, the point $(r, \theta) = (\cos 2\theta, \theta)$ can be plotted by traversing r units along the ray that makes the angle θ with the polar axis (which is the x-axis), if r is positive; otherwise, use the ray that makes the angle $\theta + \pi$ with the polar axis. The curve $r = \cos 2\theta$ is illustrated in Figure 3.9. The circles indicate the points corresponding to $\theta = 0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4$, and 2π. □

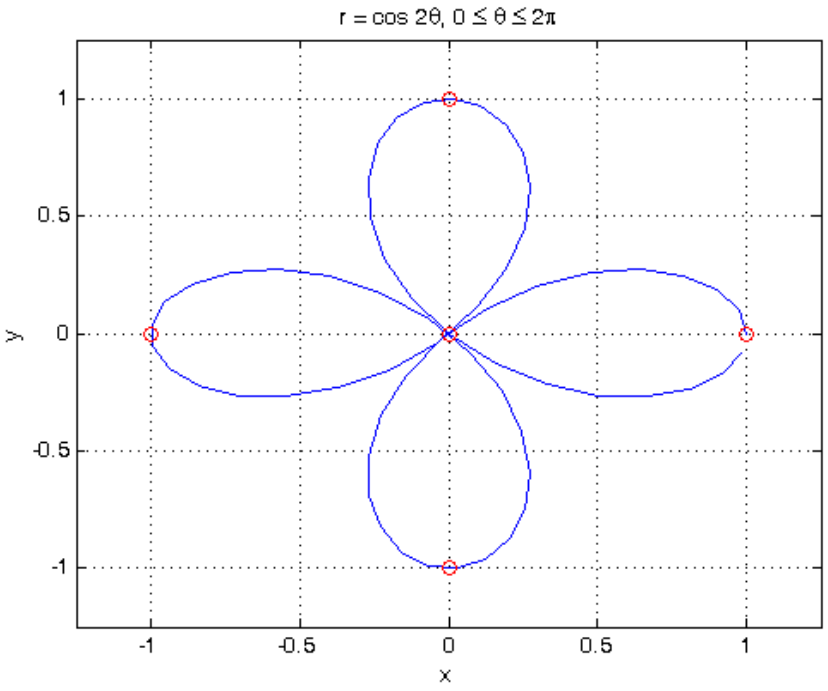

Figure 3.9: Curve described by the polar equation $r = \cos 2\theta$, where $0 \leq \theta \leq 2\pi$.

Example Sketch the curve described by the polar equation

$$r = \sin\theta, \quad 0 \le \theta \le 2\pi.$$

Solution Figure 3.10 displays the curve, which can be plotted using the same approach as in the previous example. The circles indicate the points corresponding to $\theta = 0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4$, and 2π. The circle is traced twice, once for $0 \le \theta \le \pi$, and again for $\pi \le \theta \le 2\pi$. □

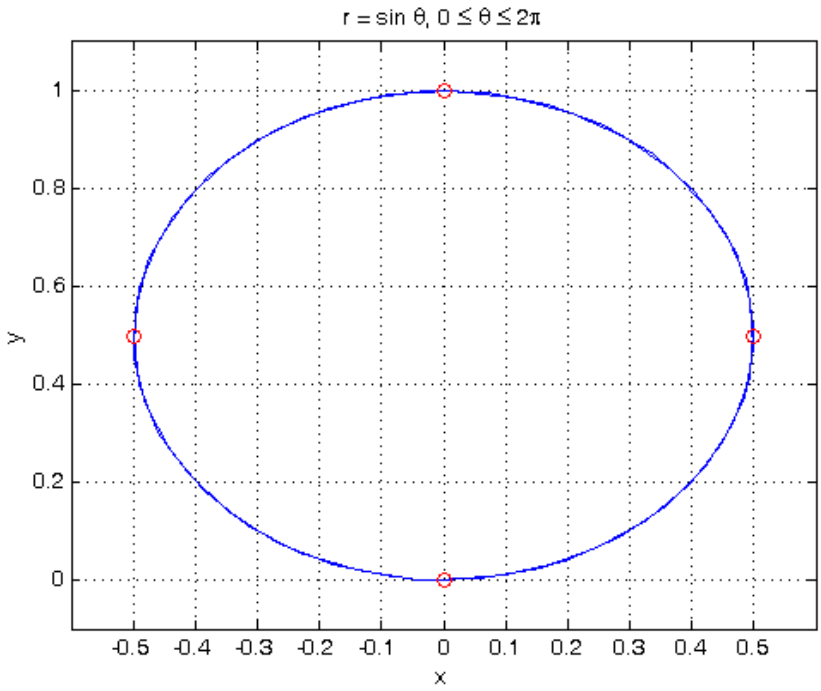

Figure 3.10: Curve described by the polar equation $r = \sin\theta$, where $0 \le \theta \le 2\pi$.

Example Identify the curve $r = 3\sin\theta$ by finding a Cartesian equation for the curve.

Solution Multiplying both sides by r yields $r^2 = 3r\sin\theta$. Using the relations $y = r\sin\theta$ and $x^2 + y^2 = r^2$, we obtain the Cartesian equation

$$x^2 + y^2 = 3y.$$

Moving the $3y$ to the left side and completing the square yields

$$x^2 + y^3 - 3y + \frac{9}{4} = \frac{9}{4},$$

which, upon factoring, becomes

$$x^2 + \left(y - \frac{3}{2}\right)^2 = \left(\frac{3}{2}\right)^2.$$

This is the equation of a circle with center $(0, 3/2)$ and radius $3/2$. □

Example Identify the curve $r = \csc\theta$ by finding a Cartesian equation for the curve.

Solution Applying $\csc\theta = 1/\sin\theta$ and multiplying both sides by $\sin\theta$ yields $r\sin\theta = 1$, which, in Cartesian coordinates, becomes $y = 1$. This curve is a horizontal line that lies one unit above the x-axis. □

Example Find a polar equation for the curve represented by the Cartesian equation $x^2 + y^2 = 2cx$.

Solution Using the relations $x^2 + y^2 = r^2$ and $x = r\cos\theta$, we obtain the polar equation $r^2 = 2cr\cos\theta$, which can then be divided through by r to obtain a polar equation in standard form, $r = 2c\cos\theta$. This curve is a circle with center $(c, 0)$ with radius c. □

3.3.3 Tangents to Polar Curves

We now determine the slope of a tangent line of a polar curve. If the curve can be described by an equation of the form $y = F(x)$ for some differentiable function F, then, by the Chain Rule,

$$\frac{dy}{d\theta} = F'(x)\frac{dx}{d\theta},$$

but since $F'(x) = dy/dx$, it follows that

$$\frac{dy}{dx} = \frac{dy/d\theta}{dx/d\theta}.$$

Expressing x and y in polar coordinates and applying the Product Rule yields

$$\frac{dy}{dx} = \frac{\frac{dy}{d\theta}}{\frac{dx}{d\theta}} = \frac{\frac{dr}{d\theta}\sin\theta + r\cos\theta}{\frac{dr}{d\theta}\cos\theta - r\sin\theta}.$$

It can be shown that this result also holds for curves that cannot be described by an equation of the form $y = F(x)$.

We make the following observations about tangents to polar curves, based on the above expression for their slope:

- Horizontal tangents occur when $dy/d\theta = 0$, provided that $dx/d\theta \neq 0$.
- Vertical tangents occur when $dx/d\theta = 0$, provided that $dy/d\theta \neq 0$.
- At the pole, when $r = 0$, the slope of the tangent is given by

$$\frac{dy}{dx} = \frac{\frac{dr}{d\theta}\sin\theta}{\frac{dr}{d\theta}\cos\theta} = \tan\theta$$

  provided $dr/d\theta \neq 0$.

Example Given the curve defined by the polar equation $r = \sin\theta$, where $0 \leq \theta \leq \pi$, determine the values of θ at which the tangent to the curve is either horizontal or vertical.

Solution As we learned in the previous example, this curve is a circle with center $(0, 1/2)$ and radius $1/2$. The curve is displayed in Figure 3.10. Using the formula for the slope of the tangent, we have, by double-angle formulas,

$$\begin{aligned}
\frac{dy}{dx} &= \frac{\frac{dy}{d\theta}}{\frac{dx}{d\theta}} \\
&= \frac{\frac{dr}{d\theta}\sin\theta + r\cos\theta}{\frac{dr}{d\theta}\cos\theta - r\sin\theta} \\
&= \frac{(\cos\theta)\sin\theta + (\sin\theta)\cos\theta}{(\cos\theta)\cos\theta - (\sin\theta)\sin\theta} \\
&= \frac{2\sin\theta\cos\theta}{\cos^2\theta - \sin^2\theta} \\
&= \frac{\sin 2\theta}{\cos 2\theta} \\
&= \tan 2\theta.
\end{aligned}$$

Alternatively, we can use the relations

$$x = r\cos\theta, \quad y = r\sin\theta$$

to compute $dy/d\theta$ and $dx/d\theta$ directly. In this case, we have

$$x = \sin\theta\cos\theta, \quad y = \sin^2\theta$$

or, from double-angle and half-angle formulas,

$$x = \frac{\sin 2\theta}{2}, \quad y = \frac{1 - \cos 2\theta}{2}$$

and therefore

$$\frac{dy}{d\theta} = \sin 2\theta, \quad \frac{dx}{d\theta} = \cos 2\theta,$$

which yields $dy/dx = \tan 2\theta$ as before.

The tangent is horizontal when $dy/d\theta = 0$ and $dx/d\theta \neq 0$. This occurs when $\theta = 0, \pi/2$, or π. The tangent is vertical when $dx/d\theta = 0$ and $dy/d\theta \neq 0$. This occurs when $\theta = \pi/4$ and $\theta = 3\pi/4$. □

Example Let P be any point (except the origin) on the curve $r = f(\theta)$. If ψ is the angle between the tangent line at P and the radial line OP, show that

$$\tan\psi = \frac{r}{dr/d\theta}$$

[*Hint:* Observe that $\psi = \phi - \theta$ in the figure on page 504 in the text.]

Solution We have

$$\tan\psi = \tan(\phi - \theta) = \frac{\tan\phi - \tan\theta}{1 + \tan\phi\tan\theta}.$$

Because ϕ is the angle that the tangent line makes with the positive x-axis,

$$\tan\phi = \frac{dy}{dx} = \frac{\frac{dr}{d\theta}\sin\theta + r\cos\theta}{\frac{dr}{d\theta}\cos\theta - r\sin\theta}.$$

It follows that if we write $r' = dr/d\theta$, then

$$\tan\psi = \frac{\frac{r'\sin\theta + r\cos\theta}{r'\cos\theta - r\sin\theta} - \frac{\sin\theta}{\cos\theta}}{1 + \frac{r'\sin\theta + r\cos\theta}{r'\cos\theta - r\sin\theta}\frac{\sin\theta}{\cos\theta}},$$

we can simplify by putting all fractions over a common denominator. Because the common denominators are both equal to $\cos\theta(r'\cos\theta - r\sin\theta)$, they cancel, and we obtain

$$\tan\psi = \frac{(r'\sin\theta + r\cos\theta)\cos\theta - (r'\cos\theta - r\sin\theta)\sin\theta}{(r'\cos\theta - r\sin\theta)\cos\theta + (r'\sin\theta + r\cos\theta)\sin\theta}.$$

Expanding, and using $\sin^2\theta + \cos^2\theta = 1$, we obtain $\tan\psi = r/r'$. □

3.3.4 Summary

- A point can be represented by polar coordinates (r, θ), where r is the distance between the point and the origin, or pole, and θ is the angle that a line segment from the pole to the point makes with the positive x-axis.
- To convert from polar coordinates to Cartesian coordinates (x, y), one can use the formulas $x = r\cos\theta$ and $y = r\sin\theta$.
- To convert from Cartesian coordinates to polar coordinates, one can use $r = \sqrt{x^2 + y^2}$, and $\theta = \tan^{-1}(y/x)$ if $x < 0$. If $x < 0$, then $\theta = \tan^{-1}(y/x) + \pi$. If $x = 0$, $\theta = \pi/2$ if $y > 0$, and $-\pi/2$ if $y < 0$.
- To graph a curve defined by a polar equation of the form $r = f(\theta)$, one can compute r for various values of θ, and use polar coordinates to plot the corresponding points on the curve.
- To compute the slope of the tangent to a polar curve $r = f(\theta)$, one can differentiate $x = f(\theta)\cos\theta$ and $y = f(\theta)\sin\theta$ with respect to θ, and then use the relation $dy/dx = (dy/d\theta)/(dx/d\theta)$.

3.4 Areas and Lengths in Polar Coordinates

In this section, we learn how to compute areas of regions and lengths of curves, for regions and curves that are most easily described using polar equations instead of Cartesian equations.

3.4.1 Area

Consider a region bounded by a curve with polar equation $r = f(\theta)$ and the rays $\theta = a$ and $\theta = b$. The area of such a region would be difficult to compute if working in Cartesian coordinates, but can be obtained in polar coordinates using the formula for the area of a sector of a circle of radius r and central angle θ,

$$A = \frac{1}{2}r^2\theta.$$

To compute the area of such a region, we can divide the interval $[a, b]$ into subintervals of uniform width $\Delta\theta = (b - a)/n$. Then, we can approximate the region with n circular sectors. The subinterval has endpoints $[\theta_{i-1}, \theta_i]$, where $\theta_i = a + i\Delta\theta$, and the corresponding sector has central angle $\Delta\theta$ and radius $f(\theta_i^*)$, where $\theta_{i-1} \leq \theta_i^a st \leq \theta_i$.

It follows that the area of the region can be approximated by the sum

$$A \approx \sum_{i=1}^{n} \frac{1}{2}[f(\theta_i^*)]^2 \Delta\theta.$$

This sum is a Riemann sum, and therefore, as $n \to \infty$, the sum converges to the definite integral

$$A = \int_a^b \frac{1}{2}[f(\theta)]^2 d\theta.$$

Similarly, the area of a region bounded by the rays $\theta = a$ and $\theta = b$, as well as the curves $r = f(\theta)$ and $r = g(\theta)$, where $f(\theta) \geq g(\theta) \geq 0$ for $a \leq \theta \leq b$, is given by the integral

$$A = \frac{1}{2} \int_a^b [f(\theta)]^2 - [g(\theta)]^2 \, d\theta.$$

Example Compute the area A of the region bounded by the curve $r = \theta$ and the rays $\theta = 0$ and $\theta = 3\pi/2$.

Solution We have

$$\begin{aligned} A &= \frac{1}{2} \int_0^{3\pi/2} \theta^2 \, d\theta \\ &= \frac{1}{2} \frac{\theta^3}{3} \Big|_0^{3\pi/2} \\ &= \frac{27\pi^3}{48} \\ &= \frac{9\pi^3}{16}. \end{aligned}$$

The curve is illustrated in Figure 3.11. □

Example Compute the area A of the region bounded by the curve $r = \sqrt{\sin\theta \cos\theta}$ and the rays $\theta = 0$ and $\theta = \pi/4$.

Solution We have

$$\begin{aligned} A &= \frac{1}{2} \int_0^{\pi/4} (\sqrt{\sin\theta \cos\theta})^2 \, d\theta \\ &= \frac{1}{2} \int_0^{\pi/4} \sin\theta \cos\theta \, d\theta \end{aligned}$$

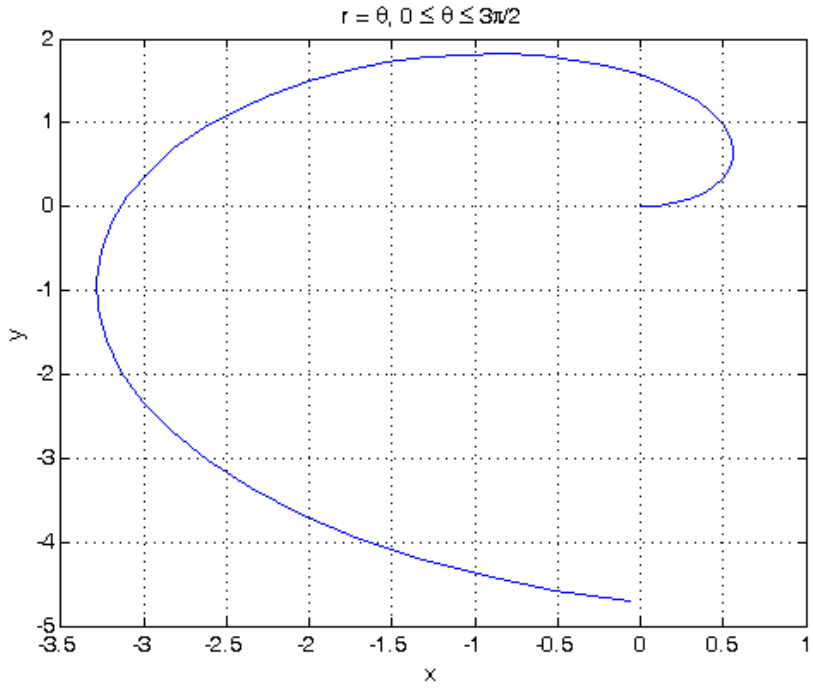

Figure 3.11: Curve $r = \theta$, $0 \le \theta \le 3\pi/2$

$$\begin{aligned} &= \frac{1}{2}\int_0^{\pi/4} \frac{\sin 2\theta}{2}\, d\theta \\ &= \frac{-\cos 2\theta}{8}\Big|_0^{\pi/4} \\ &= \frac{1}{8}[-\cos 2(\pi/4) - (-\cos 2(0))] \\ &= \frac{1}{8}[-\cos \pi/2 + \cos 0] \\ &= \frac{1}{8}. \end{aligned}$$

The curve is illustrated in Figure 3.12. □

Example Find the area of the region (see page 508 in the text) bounded by the curve $r = 4 + 3\sin\theta$ and the line $x = 0$.

Solution Points on the line $x = 0$ correspond to $\theta = -\pi/2$ (for $y < 0$) and

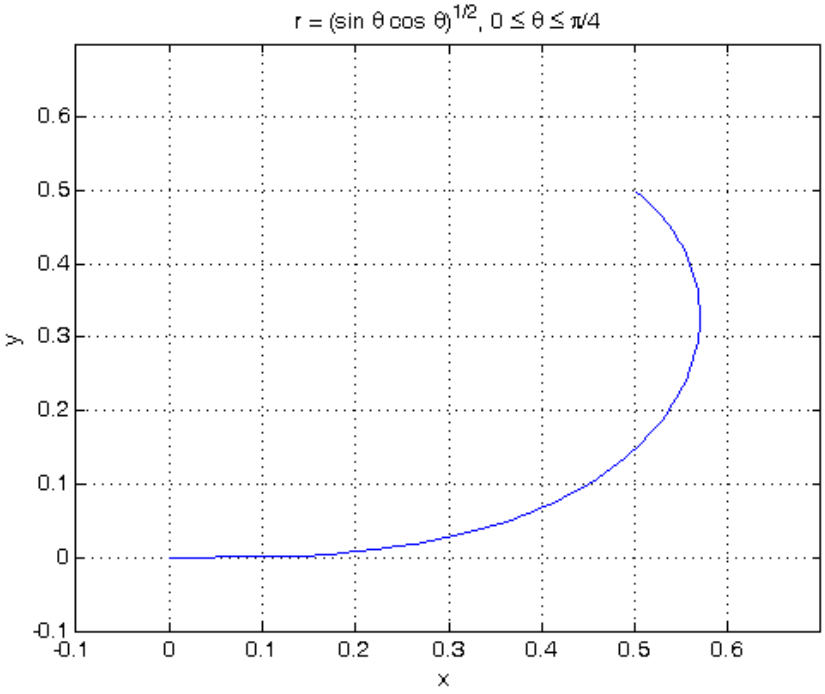

Figure 3.12: Curve $r = \sqrt{\sin\theta\cos\theta}$, $0 \le \theta \le \pi/4$

$\theta = \pi/2$ (for $y > 0$). It follows that the area A of the region is given by the integral

$$\begin{aligned} A &= \frac{1}{2}\int_{-\pi/2}^{\pi/2} (4 + 3\sin\theta)^2 \, d\theta \\ &= \frac{1}{2}\int_{-\pi/2}^{\pi/2} 16 + 24\sin\theta + 9\sin^2\theta \, d\theta \\ &= \frac{1}{2}\int_{-\pi/2}^{\pi/2} 16 + 24\sin\theta + 9\frac{1-\cos 2\theta}{2} \, d\theta \\ &= \frac{1}{2}\left[16\theta - 24\cos\theta + \frac{9}{2}\left(\theta - \frac{\sin 2\theta}{2}\right)\right]\Bigg|_{-\pi/2}^{\pi/2} \\ &= \frac{1}{2}\left[16\pi + \frac{9}{2}\pi\right] \end{aligned}$$

$$= \frac{41\pi}{4}.$$

□

Example Find the area of the region enclosed by one loop of the curve $r = \sin 2\theta$.

Solution We have $r = 0$ when $\theta = 0$ and $\theta = \pi/2$, so the interval $0 \leq \theta \leq \pi/2$ corresponds to one loop of the curve. Therefore, the area of the region enclosed by this loop is

$$\frac{1}{2}\int_0^{\pi/2} \sin^2 2\theta \, d\theta = \frac{1}{2}\int_0^{\pi/2} \frac{1-\cos 4\theta}{2} \, d\theta = \frac{1}{4}\left[\theta - \frac{\sin 4\theta}{2}\right]\Bigg|_0^{\pi/2} = \frac{\pi}{8},$$

where a double-angle identity was used to rewrite $\sin^2 2\theta$ in such a way that it can be integrated. □

Example Find the area of the region
that lies inside the curve $r = 3\cos\theta$ and outside the curve $r = 1 + \cos\theta$.

Solution These curves intersect when $3\cos\theta = 1 + \cos\theta$, or $\cos\theta = 1/2$, which is the case when $\theta = \pm\pi/3$. Therefore, the area A of the region between them is given by

$$\begin{aligned}
A &= \frac{1}{2}\int_{-\pi/3}^{\pi/3} [3\cos\theta]^2 - [1+\cos\theta]^2 \, d\theta \\
&= \frac{1}{2}\int_{-\pi/3}^{\pi/3} 9\cos^2\theta - (1 + 2\cos\theta + \cos^2\theta) \, d\theta \\
&= \frac{1}{2}\int_{-\pi/3}^{\pi/3} 8\cos^2\theta - 1 - 2\cos\theta \, d\theta \\
&= \frac{1}{2}\int_{-\pi/3}^{\pi/3} 8\frac{1+\cos 2\theta}{2} - 1 - 2\cos\theta \, d\theta \\
&= \frac{1}{2}\int_{-\pi/3}^{\pi/3} 3 + 4\cos 2\theta - 2\cos\theta \, d\theta \\
&= \frac{1}{2}\left[3\theta + 4\frac{\sin 2\theta}{2} - 2\sin\theta\right]\Bigg|_{-\pi/3}^{\pi/3} \\
&= \pi.
\end{aligned}$$

□

Example Find the area of the region that lies inside both of the curves $r = \sin\theta$ and $r = \cos\theta$.

Solution These curves intersect when $\theta = \pi/4$. Therefore, if we divide the region that lies inside both curves with the ray $\theta = \pi/4$, we can obtain its area A by computing

$$\begin{aligned} A &= \frac{1}{2}\int_0^{\pi/4} \sin^2\theta\, d\theta + \frac{1}{2}\int_{\pi/4}^{\pi/2} cos^2\theta\, d\theta \\ &= \frac{1}{2}\int_0^{\pi/4} \frac{1-\cos 2\theta}{2}\, d\theta + \frac{1}{2}\int_{\pi/4}^{\pi/2} \frac{1+\cos 2\theta}{2}\, d\theta \\ &= \left[\frac{\theta}{2} - \frac{\sin 2\theta}{4}\right]\Bigg|_0^{\pi/4} + \left[\frac{\theta}{2} + \frac{\sin 2\theta}{4}\right]\Bigg|_{\pi/4}^{\pi/2} \\ &= \frac{\pi}{8} - \frac{1}{4} + \frac{\pi}{8} - \frac{1}{4} \\ &= \frac{1}{8}(\pi - 2). \end{aligned}$$

□

3.4.2 Arc Length

Recall that the arc length of a curve represented by parametric equations

$$x = f(\theta), \quad y = g(\theta)$$

is given by

$$L = \int_a^b \sqrt{\left(\frac{dx}{d\theta}\right)^2 + \left(\frac{dy}{d\theta}\right)^2}\, d\theta,$$

where θ is the parameter, and the curve is defined for $a \le \theta \le b$. Ideally we would like to be able to compute the arc length of a curve directly from its polar equation, instead of always having to convert to Cartesian coordinates.

Since Cartesian coordinates and polar coordinates are related by the equations $x = r\cos\theta$, $y = r\sin\theta$, it follows from the Product Rule that

$$\frac{dx}{d\theta} = \frac{dr}{d\theta}\cos\theta - r\sin\theta, \quad \frac{dy}{d\theta} = \frac{dr}{d\theta}\sin\theta + r\cos\theta.$$

We then have

$$\left(\frac{dx}{d\theta}\right)^2 + \left(\frac{dy}{d\theta}\right)^2 = \left(\frac{dr}{d\theta}\cos\theta - r\sin\theta\right)^2 + \left(\frac{dr}{d\theta}\sin\theta + r\cos\theta\right)^2$$

$$\begin{aligned}
&= \left(\frac{dr}{d\theta}\right)^2 \cos^2\theta - 2r\frac{dr}{d\theta}\sin\theta\cos\theta + r^2\sin^2\theta + \\
&\quad \left(\frac{dr}{d\theta}\right)^2 \sin^2\theta + 2r\frac{dr}{d\theta}\sin\theta\cos\theta + r^2\cos^2\theta \\
&= \left(\frac{dr}{d\theta}\right)^2 \cos^2\theta + \left(\frac{dr}{d\theta}\right)^2 \sin^2\theta + r^2\sin^2\theta + r^2\cos^2\theta \\
&= \left(\frac{dr}{d\theta}\right)^2 + r^2.
\end{aligned}$$

It follows that the arc length of a curve can be obtained directly from its polar equation, using the formula

$$L = \int_a^b \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2}\, d\theta.$$

Example Compute the arc length L of the curve $r = \cos\theta$, where $0 \le \theta \le \pi/2$.

Solution We have $dr/d\theta = -\sin\theta$, and therefore

$$\begin{aligned}
L &= \int_0^{\pi/2} \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2}\, d\theta \\
&= \int_0^{\pi/2} \sqrt{\cos^2\theta + (-\sin\theta)^2}\, d\theta \\
&= \int_0^{\pi/2} \sqrt{\cos^2\theta + \sin^2\theta}\, d\theta \\
&= \int_0^{\pi/2} 1\, d\theta \\
&= \frac{\pi}{2}.
\end{aligned}$$

The curve is illustrated in Figure 3.13. □

3.4.3 Summary

- The area of the region bounded by the polar curve $r = f(\theta)$, the ray $\theta = a$, and the ray $\theta = b$ is half of the integral of r^2 from a to b.
- The arc length of a polar curve $r = f(\theta)$, where $a \le \theta \le b$, is the integral of $\sqrt{r^2 + (dr/d\theta)^2}$ from a to b.

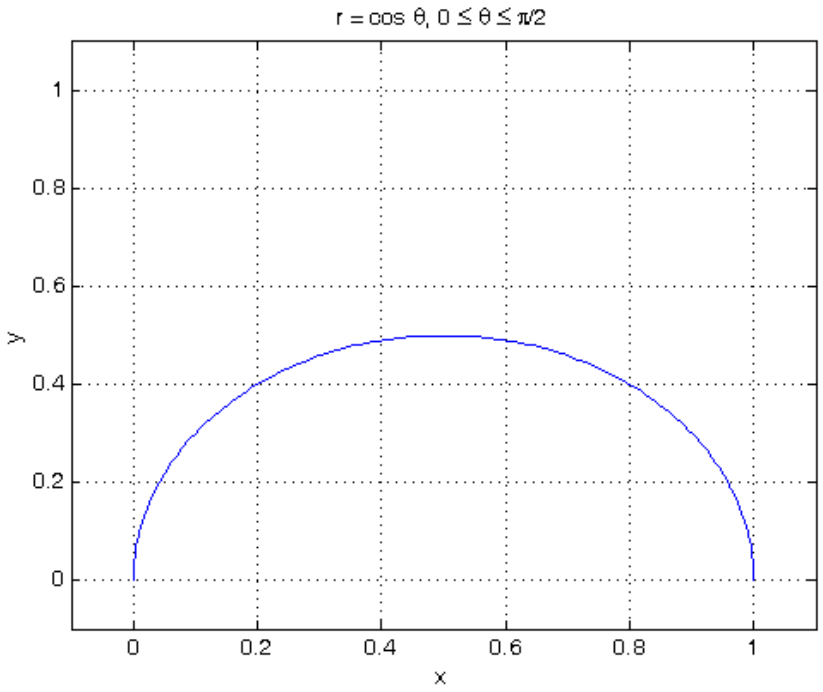

Figure 3.13: Curve $r = \cos\theta$, $0 \leq \theta \leq \pi/2$

3.5 Review

You should now be able to complete the following types of problems:

- Eliminating the parameter from the parametric equations $x = g(t)$, $y = h(t)$ of a given curve, and describing the curve using a Cartesian equation of the form $y = f(x)$. This requires solving $x = g(t)$ for t, and substituting this solution into $y = h(t)$ to obtain $y = h(g^{-1}(x))$.

- Computing the arc length of a curve $y = f(x)$, for $a \leq x \leq b$. This requires evaluting the integral

$$\int_a^b \sqrt{1 + [f'(x)]^2}\, dx.$$

  Evaluating this integral may require the use of a *trigonometric substitution* of the form $x = a\tan\theta$, for the purpose of integrating a function

involving $\sqrt{a^2+x^2}$. This substitution uses the identity $\tan^2\theta + 1 = \sec^2\theta$. Alternatively, evaluating the integral may require recognizing perfect squares of the form $(a-b)^2 = a^2 - 2ab + b^2$ or $(a+b)^2 = a^2 + 2ab + b^2$.

- Computing the arc length of a curve $x = f(t)$, $y = g(t)$, for $a \le t \le b$. This requires evaluating the integral

$$\int_a^b \sqrt{[f'(t)]^2 + [g'(t)]^2}\, dt.$$

  See the comments above for tips on how to evaluate such an integral.

- Computing the equation of the tangent line at a point on a parametric curve. This requires computing the slope $dy/dx = (dy/dt)/(dx/dt)$, and then using the point-slope form $y - y_0 = m(x - x_0)$, where m is the slope and (x_0, y_0) is a point on the line. It will also be necessary to determine the value of the parameter t that corresponds to the given point.

- Determining whether a given parametric curve is concave up or concave down at a given point. This involves computing

$$\frac{d^2y}{dx^2} = \frac{\frac{d}{dt}\left(\frac{dy}{dx}\right)}{\frac{dx}{dt}},$$

  where $dy/dx = (dy/dt)/(dx/dt)$ is computed as a function of t, so that it can readily be differentiated with respect to t.

- Computing the area of a region bounded by a parametric curve, as well as horizontal and vertical lines. This requires using one of the formulas

$$\int_a^b y\, dx = \int_\alpha^\beta g(t)f'(t)\, dt, \quad \int_c^d x\, dy = \int_\alpha^\beta f(t)g'(t)\, dt,$$

  where $x = f(t)$, $y = g(t)$, $\alpha \le t \le \beta$ describes the curve, and $f(\alpha) = a$, $f(\beta) = b$, $g(\alpha) = c$, and $g(\beta) = d$, with $a \le b$ and $c \le d$. The choice of formula depends on whether the region is bounded by two vertical lines, in which case the first formula should be used, or two horizontal lines, in which case the second formula should be used.

- Converting points from Cartesian to polar coordinates, using the relations $r^2 = x^2 + y^2$ and $\tan\theta = y/x$. Note that if $x < 0$, then the angle obtained from $\tan^{-1}(y/x)$ must be corrected by adding π radians to it. Also, if $x = 0$ and $y > 0$, then $\theta = \pi/2$, and if $x = 0$ and $y < 0$, then $\theta = -\pi/2$.

- Determining where a curve described by a given polar equation $r = f(\theta)$ has horizontal or vertical tangents. This requires computing

$$\frac{dy}{dx} = \frac{\frac{dr}{d\theta}\sin\theta + r\cos\theta}{\frac{dr}{d\theta}\cos\theta - r\sin\theta}.$$

  If the numerator is zero but the denominator is nonzero, then the tangent is horizontal. If the denominator is zero but the numerator is nonzero, then the tangent is vertical.

- Computing the arc length of a curve described by a polar equation $r = f(\theta)$, $a \le \theta \le b$. This requires evaluating the integral

$$\int_a^b \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2}\, d\theta,$$

  which may involve the techniques described for other problems involving arc length.

- Computing the area of a region bounded by a polar curve $r = f(\theta)$ and the rays $\theta = a$ and $\theta = b$. This requires evaluating the integral

$$\int_a^b \frac{1}{2}[f(\theta)]^2\, d\theta.$$

Index