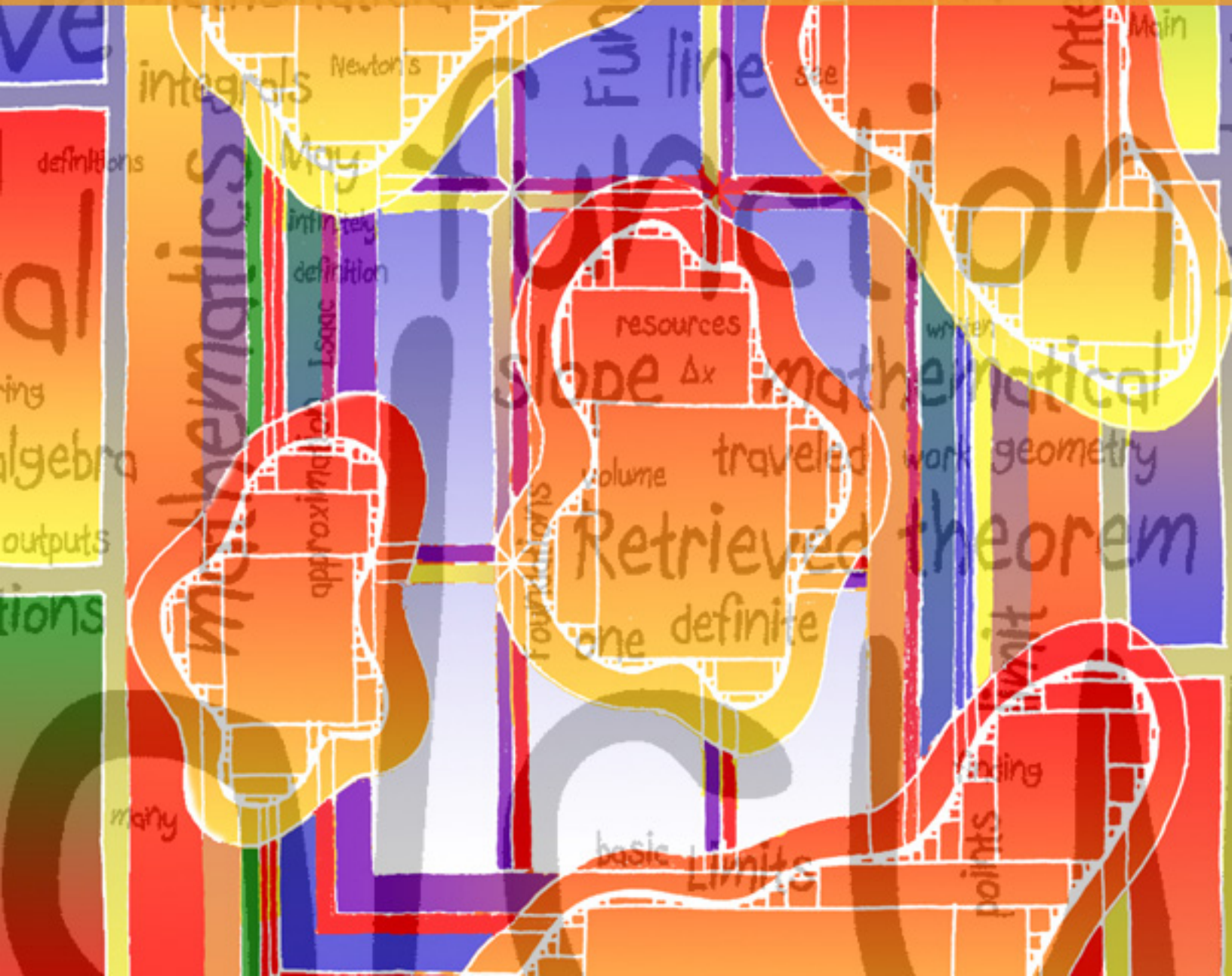
Calculus

Sarah Channon



First Edition, 2009

ISBN 978 93 80168 11 1

Published by:

Global Media
1819, Bhagirath Palace,
Chandni Chowk, Delhi-110 006
Email: globalmedia@dkpd.com

Table of Contents

1. Introduction
2. Functions
3. Limits
4. Differentiation
5. Integration
6. Infinite Series
7. Multivariable & Differential Calculus
8. Extensions

Introduction

Calculus is a tool used almost everywhere in the modern world to describe change and motion. Its use is widespread in science, engineering, medicine, business, industry, and many other fields. Just as algebra introduces students to new ways of thinking about arithmetic problems (by way of variables, equations, functions, and graphs), calculus introduces new ways of thinking about *algebra* problems (considering, for example, how the height of a point moving along a graph changes as its horizontal position changes).

What is calculus?

Calculus is the branch of mathematics dealing with instantaneous rates of change of continuously varying quantities. For example, consider a moving car. It is possible to create a function describing the *displacement* of the car (where it is located in relation to a reference point) at any point in time as well as a function describing the *velocity* (speed and direction of movement) of the car at any point in time. If the car were traveling at a constant velocity, then algebra would be sufficient to determine the position of the car at any time; if the velocity is unknown but still constant, the position of the car could be used (along with the time) to find the velocity.

However, the velocity of a car cannot jump from zero to 35 miles per hour at the beginning of a trip, stay constant throughout, and then jump back to zero at the end. As the accelerator is pressed down, the velocity rises gradually, and usually not at a constant *rate* (i.e., the driver may push on the gas pedal harder at the beginning, in order to speed up). Describing such motion and finding velocities and distances at particular times cannot be done using methods taught in pre-calculus, but it is not only possible but straightforward with calculus.

Calculus has two basic applications: *differential calculus* and *integral calculus*. The simplest introduction to differential calculus involves an explicit series of numbers. Given the series (42, 43, 3, 18, 34), the differential of this series would be (1, -40, 15, 16). The new series is derived from the difference of successive numbers which gives rise to its name "differential". Rarely, if ever, are differentials used on an explicit series of numbers as done here. Instead, they are derived from a series of numbers defined by a continuous function which are described later.

Integral calculus, like differential calculus, can also be introduced via series of numbers. Notice that in the previous example, the original series can almost be derived solely from its differential. Instead of taking the difference, however, integration involves taking the sum. Given the first number of the original series, 42 in this case, the rest of the original series can be derived by adding each successive number in its differential (42, 42+1, 43+(-40), 3+15, 18+16). Note that knowledge of the first number in the original series is crucial in deriving the integral. As with differentials, integration is performed on continuous functions rather than explicit series of numbers, but the concept is still the same. Integral calculus allows us to calculate the area under a curve of almost any shape; in the car example, this enables you to find the displacement of the car based on the

velocity curve. This is because the area under the curve is the total distance moved, as we will soon see.

Why learn calculus?

Calculus is essential for many areas of science and engineering. Both make heavy use of mathematical functions to describe and predict physical phenomena that are subject to continual change, and this requires the use of calculus. Take our car example: if you want to design cars, you need to know how to calculate forces, velocities, accelerations, and positions. All require calculus. Calculus is also necessary to study the motion of gases and particles, the interaction of forces, and the transfer of energy. It is also useful in business whenever rates are involved. For example, equations involving interest or supply and demand curves are grounded in the language of calculus.

Calculus also provided important tools in understanding functions and has led to the development of new areas of mathematics including real and complex analysis, topology, and non-euclidean geometry.

What is involved in learning calculus?

Learning calculus, like much of mathematics, involves two parts:

- Understanding the concepts: You must be able to explain what it means when you take a derivative rather than merely apply the formulas for finding a derivative. Otherwise, you will have no idea whether or not your solution is correct. Drawing diagrams, for example, can help clarify abstract concepts.

- Symbolic manipulation: Like other branches of mathematics, calculus is written in symbols that represent concepts. You will learn what these symbols mean and how to use them. A good working knowledge of trigonometry and algebra is a must, especially in integral calculus. Sometimes you will need to manipulate expressions into a usable form before it is possible to perform operations in calculus.

What you should know before using this text

There are some basic skills that you need before you can use this text. Continuing with our example of a moving car:

- You will need to describe the motion of the car in symbols. This involves understanding functions.

- You need to manipulate these functions. This involves algebra.

- You need to translate symbols into graphs and vice verse. This involves understanding the graphing of functions.

- It also helps (although it isn't necessarily essential) if you understand the functions used in trigonometry since these functions appear frequently in science.

Functions

Classical understanding of functions

To provide the classical understanding of functions, a *function* can be thought of as a machine. Machines take in raw materials, change them in a predictable way, and give out a finished product. The kinds of functions we consider here, for the most part, take in a real number, change it in a formulaic way, and give out a real number (which in special cases could be the same as the one we put in). You can think of this as an *input-output machine*. You give a function an input and it gives you an output. For example, the squaring function gives the output value 16 when the input is 4 and the output value 1 when the input is -1.

A function is usually symbolized *f* or *g* or something similar, though it doesn't have to be. A function is always defined as "of a variable" which tells the reader what to replace in the formula for the function.

For instance, $f(x) = 3x + 2$ tells the reader:

- The function *f* is a function of *x*.
- To evaluate the function at a certain number, replace the *x* with that number.
- Replacing *x* with that number in the right side of the function will produce the function's output for that certain input.
- In English, the definition of f is interpreted, "Given a number, *f* will return *two more than the triple of that number*."

Thus, if we want to know the value (or output) of the function at 3:

$f(x) = 3x + 2$
$f(3) = 3(3) + 2$ We evaluate the function at *x* = 3.
$f(3) = 9 + 2 = 11$ The value of f at 3 is 11.

See? It's easy!

Note that $f(3)$ means the value of the dependent variable when x takes on the value of 3. So we see that the number *11* is the output of the function when we give the number *3* as the input. We refer to the input as the **argument** of the function (or the **independent variable**), and to the output as the **value** of the function at the given argument (or the **dependent variable**). A good way to think of it is the dependent variable $f(x)$ 'depends' on the value of the independent variable x. This is read as "the value of *f* at three is eleven", or simply "*f* of three equals eleven".

Notation

Functions are used so much that there is a special notation for them. The notation is somewhat ambiguous, so familiarity with it is important in order to understand the intention of an equation or formula.

Though there are no strict rules for naming a function, it is standard practice to use the letters *f*, *g*, and *h* to denote functions, and the variable *x* to denote an independent variable. *y* is used for both dependent and independent variables.

When discussing or working with a function *f*, it's important to know not only the function, but also its independent variable *x*. Thus, when referring to a function *f*, you usually do not write *f*, but instead *f*(*x*). The function is now referred to as "*f* of *x*". The name of the function is adjacent to the independent variable (in parentheses). This is useful for indicating the value of the function at a particular value of the independent variable. For instance, if

$$f(x) = 7x + 1,$$

and if we want to use the value of *f* for *x* equal to 2, then we would substitute 2 for *x* on both sides of the definition above and write

$$f(2) = 7(2) + 1 = 14 + 1 = 15$$

This notation is more informative than leaving off the independent variable and writing simply '*f*', but can be ambiguous since the parentheses can be misinterpreted as multiplication.

Modern understanding of functions

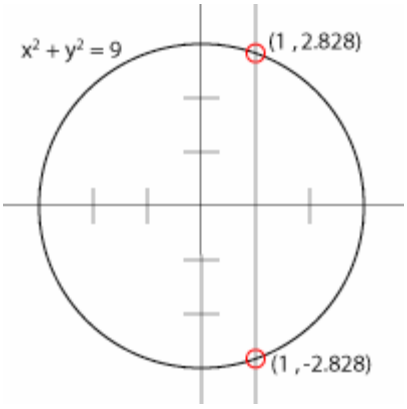
The formal definition of a function states that a function is actually a *rule* that associates elements of one set called the *domain* of the function, with the elements of another set called the *range* of the function. For each value we select from the domain of the function, there exists exactly one corresponding element in the range of the function. The definition of the function tells us which element in the range corresponds to the element we picked from the domain. Classically, the element picked from the domain is pictured as something that is fed into the function and the corresponding element in the range is pictured as the output. Since we "pick" the element in the domain whose corresponding element in the range we want to find, we have control over what element we pick and hence this element is also known as the "independent variable". The element mapped in the range is beyond our control and is "mapped to" by the function. This element is hence also known as the "dependent variable", for it depends on which independent variable we pick. Since the elementary idea of functions is better understood from the classical viewpoint, we shall use it hereafter. However, it is still important to remember the correct definition of functions at all times.

To make it simple, for the function $f(x)$, all of the possible x values constitute the domain, and all of the values $f(x)$ (y on the x-y plane) constitute the range.

Remarks

The following arise as a direct consequence of the definition of functions:

1. By definition, for each "input" a function returns only one "output", corresponding to that input. While the same output may correspond to more than one input, one input cannot correspond to more than one output. This is expressed graphically as the *vertical line test*: a line drawn parallel to the axis of the dependent variable (normally vertical) will intersect the graph of a function only once. However, a line drawn parallel to the axis of the independent variable (normally horizontal) may intersect the graph of a function as many times as it likes. Equivalently, this has an algebraic (or formula-based) interpretation. We can always say if $a = b$, then $f(a) = f(b)$, but if we only know that $f(a) = f(b)$ then we can't be sure that $a = b$.
2. Each function has a set of values, the function's *domain*, which it can accept as input. Perhaps this set is all positive real numbers; perhaps it is the set {pork, mutton, beef}. This set must be implicitly/explicitly defined in the definition of the function. You cannot feed the function an element that isn't in the domain, as the function is not defined for that input element.
3. Each function has a set of values, the function's *range*, which it can output. This may be the set of real numbers. It may be the set of positive integers or even the set {0,1}. This set, too, must be implicitly/explicitly defined in the definition of the function.



This is an example of an expression which fails the vertical line test.

The vertical line test

The vertical line test, mentioned in the preceding paragraph, is a systematic test to find out if an equation involving x and y can serve as a function (with x the independent variable and y the dependent variable). Simply graph the equation and draw a vertical line

through each point of the *x*-axis. If any vertical line ever touches the graph at more than one point, then the equation is not a function; if the line always touches at most one point of the graph, then the equation is a function.

(There are a lot of useful curves, like circles, that aren't functions (see picture). Some people call these graphs with multiple intercepts, like our circle, "multi-valued functions"; they would refer to our "functions" as "single-valued functions".)

Important functions

In order of degree (or complexity, informally said)

Constant function

$$f(x) = c$$

It disregards the input and always outputs the constant *c*, and is a polynomial of the *zeroth* degree where $f(x) = cx^0 = c(1) = c$. Its graph is a horizontal line.

Identity function

$$f(x) = x$$

The output is always the input. A polynomial of the *first* degree, $f(x) = x^1 = x$. Special case of a linear function.

Linear function

$$f(x) = mx + c$$

Takes an input, multiplies by *m* and adds *c*. It is a polynomial of the *first* degree. Its graph is a line (slanted, except $m = 0$).

Quadratic function

$$f(x) = ax^2 + bx + c$$

A polynomial of the *second* degree. Its graph is a parabola, unless $a = 0$. (Don't worry if you don't know what this is.)

Polynomial function

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

The number *n* is called the *degree*.

Signum function

$$\operatorname{sgn}(x) = \begin{cases} -1 & : \quad x < 0 \\ 0 & : \quad x = 0 \\ 1 & : \quad x > 0. \end{cases}$$

Determines the sign of the argument *x*.

Example functions

Some more simple examples of functions have been listed below.

$h(x) = \begin{cases} 1, & \text{if } x > 0 \\ -1, & \text{if } x < 0 \end{cases}$ Gives 1 if input is positive, -1 if input is negative. Note that the function only accepts negative and positive numbers, not 0. Mathematics describes this condition by saying 0 is not in the domain of the function.
$g(y) = y^2$ Takes an input and squares it. $g(z) = z^2$ Exactly the same function, rewritten with a different independent variable. This is perfectly legal and sometimes done to prevent confusion (e.g. when there are already too many uses of *x* or *y* in the same paragraph.)
$f(x) = \begin{cases} 5^{x^2}, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases}$ Note that we can define a function by a totally arbitrary rule.

It is possible to replace the independent variable with any mathematical expression, not just a number. For instance, if the independent variable is itself a function of another variable, then it could be replaced with that function. This is called composition, and is discussed later.

Manipulating functions

Functions can be manipulated in the same manner as any other variable; they can be added, multiplied, raised to powers, etc. For instance, let

$$f(x) = 3x + 2 \text{ and}$$
$$g(x) = x^2.$$

Then

$$\begin{aligned} f+g &= (f+g)(x) \\ &= f(x)+g(x) \\ &= (3x+2)+(x^2) \\ &= x^2+3x+2 \end{aligned},$$

$$\begin{aligned} f-g &= (f-g)(x) \\ &= f(x)-g(x) \\ &= (3x+2)-(x^2) \\ &= -x^2+3x+2 \end{aligned},$$

$$\begin{aligned} f\times g &= (f\times g)(x) \\ &= f(x)\times g(x) \\ &= (3x+2)\times(x^2) \\ &= 3x^3+2x^2 \end{aligned},$$

$$\begin{aligned} \frac{f}{g} &= \left(\frac{f}{g}\right)(x) \\ &= \frac{f(x)}{g(x)} \\ &= \frac{3x+2}{x^2} \\ &= \frac{3}{x}+\frac{2}{x^2} \end{aligned}.$$

Composition of functions

However, there is one particular way to combine functions which cannot be done with other variables. The value of a function *f* depends upon the value of another variable *x*; however, that variable could be equal to another function *g*, so its value depends on the value of a third variable. If this is the case, then the first variable is a function *h* of the third variable; this function (*h*) is called the **composition** of the other two functions (*f* and *g*). Composition is denoted by

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

This can be read as either "f composed with g" or "f of g of x."

For instance, let

$$f(x) = 3x + 2 \text{ and}$$
$$g(x) = x^2.$$

Then

$$\begin{aligned} h(x) &= f(g(x)) \\ &= f(x^2) \\ &= 3(x^2) + 2 \\ &= 3x^2 + 2 \end{aligned}.$$

Here, *h* is the composition of *f* and *g* and we write $h = f \circ g$. Note that composition is not commutative:

$$f(g(x)) = 3x^2 + 2, \text{ and}$$
$$\begin{aligned} g(f(x)) &= g(3x + 2) \\ &= (3x + 2)^2 \\ &= 9x^2 + 12x + 4 \end{aligned}$$
so $f(g(x)) \neq g(f(x))$.

Composition of functions is very common, mainly because functions themselves are common. For instance, squaring and sine are both functions:

$$\text{square}(x) = x^2,$$
$$\text{sine}(x) = \sin x$$

Thus, the expression $\sin^2 x$ is a composition of functions:

$$\begin{aligned} \sin^2 x &= \text{square}(\sin x) \\ &= \text{square}(\text{sine}(x)). \end{aligned}$$

(Note that this is *not* the same as $\text{sine}(\text{square}(x)) = \sin x^2$.) Since the function sine equals 1 / 2 if $x = \pi / 6$,

$$\text{square}(\text{sine}(\pi/6)) = \text{square}(1/2).$$

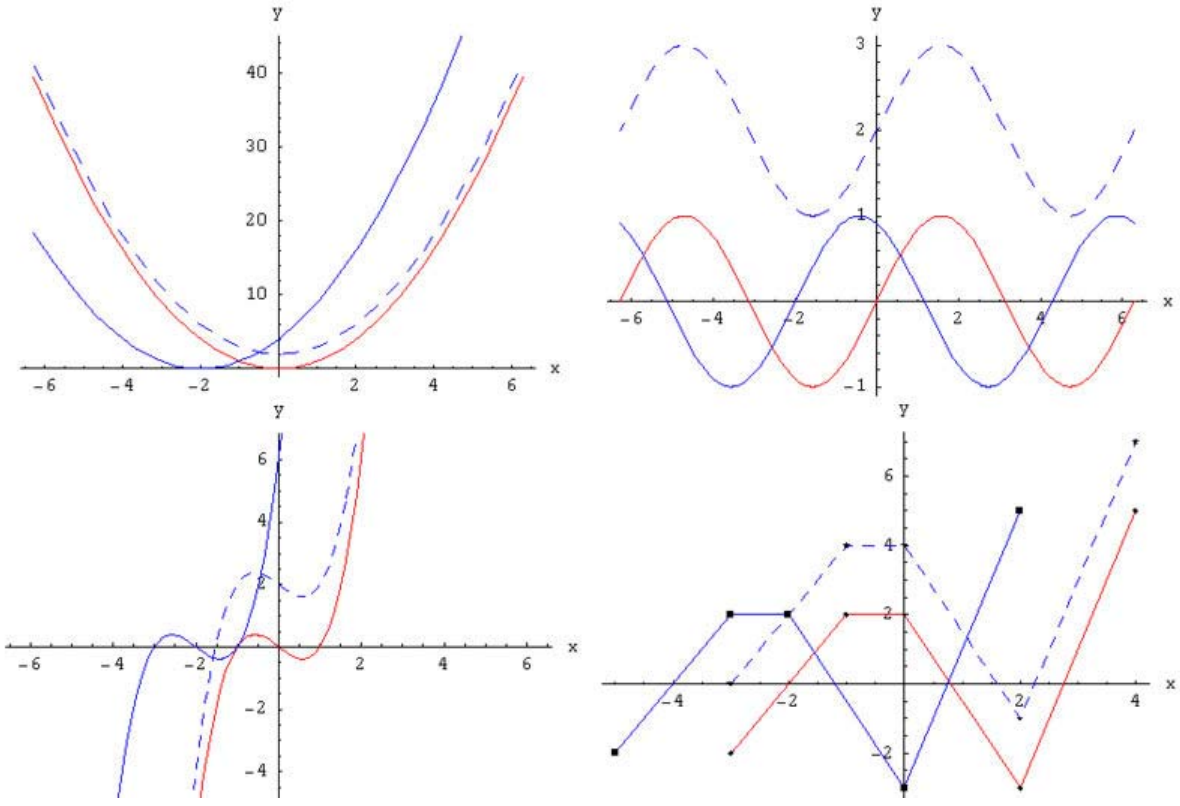
Since the function square equals 1 / 4 if $x = 1 / 2$,

$$\sin^2 \pi/6 = \text{square}(\text{sine}(\pi/6)) = \text{square}(1/2) = 1/4.$$

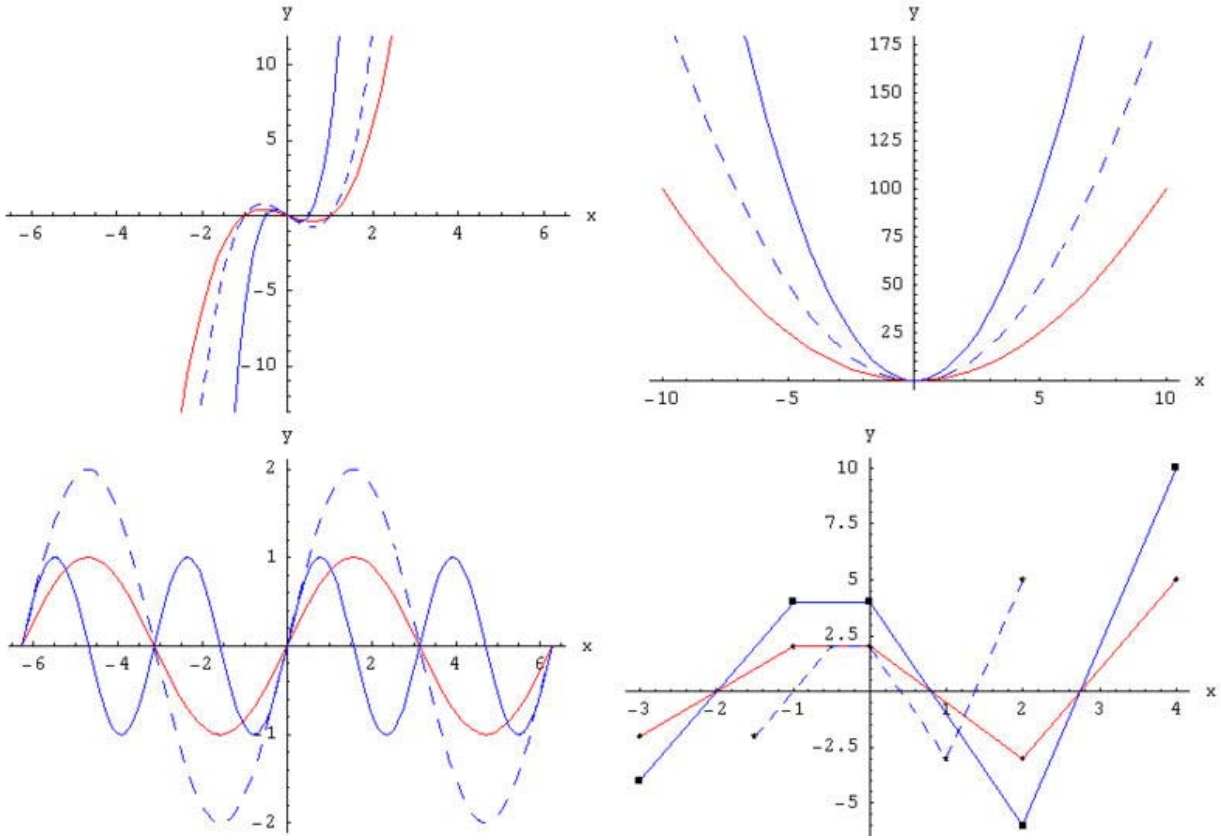
Transformations

Transformations are a type of function manipulation that are very common. They consist of multiplying, dividing, adding or subtracting constants to either the input or the output. Multiplying by a constant is called **dilation** and adding a constant is called **translation**. Here are a few examples:

$f(2 \times x)$ Dilation
$f(x + 2)$ Translation
$2 \times f(x)$ Dilation
$2 + f(x)$ Translation



Examples of horizontal and vertical translations



Examples of horizontal and vertical dilations

Translations and dilations can be either horizontal or vertical. Examples of both vertical and horizontal translations can be seen at right. The red graphs represent functions in their 'original' state, the solid blue graphs have been translated (shifted) horizontally, and the dashed graphs have been translated vertically.

Dilations are demonstrated in a similar fashion. The function

$$f(2 \times x)$$

has had its input doubled. One way to think about this is that now any change in the input will be doubled. If I add one to *x*, I add two to the input of *f*, so it will now change twice as quickly. Thus, this is a horizontal dilation by $\frac{1}{2}$because the distance to the *y*-axis has been **halved**. A vertical dilation, such as

$$2 \times f(x)$$

is slightly more straightforward. In this case, you double the output of the function. The output represents the distance from the *x*-axis, so in effect, you have made the graph of the function 'taller'. Here are a few basic examples where *a* is any positive constant:

Original graph	$f(x)$	Reflection about origin	$-f(-x)$
Horizontal translation by *a* units **left**	$f(x-a)$	Horizontal translation by *a* units **right**	$f(x+a)$
Horizontal dilation by a factor of *a*	$f(x \times \frac{1}{a})$	Vertical dilation by a factor of *a*	$a \times f(x)$
Vertical translation by *a* units **down**	$f(x)-a$	Vertical translation by *a* units **up**	$f(x)+a$
Reflection about *x*-axis	$-f(x)$	Reflection about *y*-axis	$f(-x)$

Domain and Range

Notation

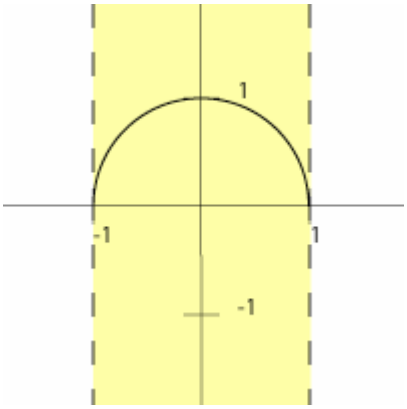
The domain and range of functions are commonly expressed using interval notation. This notation is very simple, but sometimes ambiguous because of the similarity to ordered pair notation:

Meaning	Interval Notation	Set Notation
All values greater than or equal to *a* and less than or equal to *b*	$[a,b]$	$\{x : a \leq x \leq b\}$
All values greater than *a* and less than *b*	(a,b)	$\{x : a < x < b\}$
All values greater than or equal to *a* and less than *b*	$[a,b)$	$\{x : a \leq x < b\}$
All values greater than *a* and less than or equal to *b*	$(a,b]$	$\{x : a < x \leq b\}$
All values greater than or equal to *a*.	$[a,\infty)$	$\{x : x \geq a\}$
All values greater than *a*.	(a,∞)	$\{x : x > a\}$

All values less than or equal to *a*.	$(-\infty, a]$	$\{x : x \leq a\}$
All values less than *a*.	$(-\infty, a)$	$\{x : x < a\}$
All values.	$(-\infty, \infty)$	$\{x : x \in \mathbb{R}\}$

Note that $\infty$and $-\infty$must always have an exclusive parenthesis rather than an inclusive bracket. This is because $\infty$is not a number, and therefore cannot be in our set. $\infty$is really just a symbol that makes things easier to write, like the intervals above.
Note: *(* is also denoted by *]*, and *)* by *[*, i.e., (a,b) is the same as]a,b[, and [a,b) is [a,b[. This is a source of funny misunderstandings.

Domain



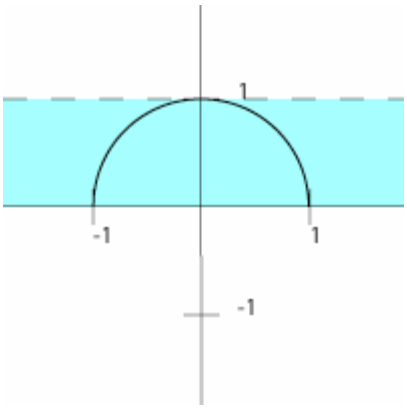
The domain of the function is the interval from -1 to 1

The **domain** of a function is the set of all points over which it is defined. More simply, it represents the set of x-values which the function can accept as input. For instance, if

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

then *f*(*x*) is only defined for values of *x* between - 1 and 1, because the square root function is not defined (in real numbers) for negative values. Thus, the domain, in interval notation, is $[-1, 1]$. In other words,

$$f(x) \text{is defined for } x \in [-1, 1], \text{or} \{x : -1 \leq x \leq 1\}.$$



The range of the function is the interval from 0 to 1

Range

The **range** of a function is the set of all values which it attains (i.e. the y-values). For instance, if:

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

Then, $f(x)$ can only equal values in the interval from 0 to 1. Thus, the range of f is $[0, 1]$.

One-to-one Functions

A function $f(x)$ is **one-to-one** (or less commonly **injective**) if, for every value of f, there is only one value of x that corresponds to that value of f. For instance, the function $f(x) = \sqrt{1 - x^2}$ is not one-to-one, because both $x = 1$ and $x = -1$ result in $f(x) = 0$. However, the function $f(x) = x + 2$ is one-to-one, because, for every possible value of $f(x)$, there is exactly one corresponding value of x. Other examples of one-to-one functions are $f(x) = x^3 + ax$, where $a \in [0, \infty)$. Note that if you have a one-to-one function and translate or dilate it, it remains one-to-one. (Of course you can't multiply x or f by a zero factor).

If you know what the graph of a function looks like, it is easy to determine whether or not the function is one-to-one. If every horizontal line intersects the graph in at most one point, then the function is one-to-one. This is known as the Horizontal Line Test.

Inverse functions

We call $g(x)$ the inverse function of $f(x)$ if, for all x:

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

A function $f(x)$ has an inverse function if and only if $f(x)$ is one-to-one. For example, the inverse of $f(x) = x + 2$ is $g(x) = x - 2$. The function $f(x) = \sqrt{1 - x^2}$ has no inverse.

Notation

The inverse function of f is denoted as $f^{-1}(x)$. Thus, $f^{-1}(x)$ is defined as the function that follows this rule

$f(f^{-1}(x)) = f^{-1}(f(x)) = x$:

To determine $f^{-1}(x)$ when given a function f, substitute $f^{-1}(x)$ for x and substitute x for $f(x)$. Then solve for $f^{-1}(x)$, provided that it is also a function.

Example: Given $f(x) = 2x - 7$, find $f^{-1}(x)$.

Substitute $f^{-1}(x)$ for x and substitute x for $f(x)$. Then solve for $f^{-1}(x)$:

$$\begin{aligned} f(x) &= 2x - 7 \\ x &= 2[f^{-1}(x)] - 7 \\ x + 7 &= 2[f^{-1}(x)] \\ \frac{x + 7}{2} &= f^{-1}(x) \end{aligned}$$

To check your work, confirm that $f^{-1}(f(x)) = x$:

$f^{-1}(f(x)) =$

$f^{-1}(2x - 7) =$

$$\frac{2x - 7 + 7}{2} = \frac{2x}{2} = x$$

If f isn't one-to-one, then, as we said before, it doesn't have an inverse. Then this method will fail.

Example: Given $f(x) = x^2$, find $f^{-1}(x)$.

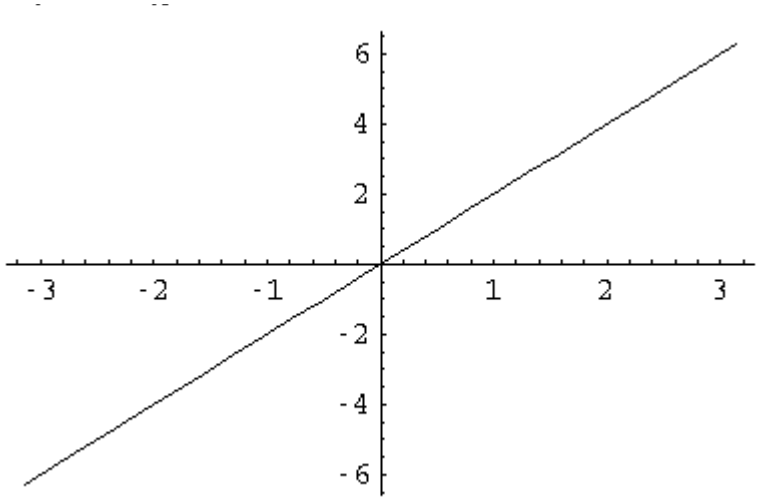
Substitute $f^{-1}(x)$ for x and substitute x for $f(x)$. Then solve for $f^{-1}(x)$:

$$f(x) = x^2$$

$$x = (f^{-1}(x))^2$$
$$f^{-1}(x) = \pm\sqrt{x}$$

Since there are two possibilities for $f^{-1}(x)$, it's not a function. Thus $f(x) = x^2$ doesn't have an inverse. Of course, we could also have found this out from the graph by applying the Horizontal Line Test. It's useful, though, to have lots of ways to solve a problem, since in a specific case some of them might be very difficult while others might be easy. For example, we might only know an algebraic expression for $f(x)$ but not a graph.

Graphing Functions



Graph of y=2x

It is sometimes difficult to understand the behavior of a function given only its definition; a visual representation or graph can be very helpful. A **graph** is a set of points in the Cartesian plane, where each point (x,y) indicates that $f(x) = y$. In other words, a graph uses the position of a point in one direction (the *vertical-axis* or *y-axis*) to indicate the value of f for a position of the point in the other direction (the *horizontal-axis* or *x-axis*).

Functions may be graphed by finding the value of f for various x and plotting the points $(x, f(x))$ in a Cartesian plane. For the functions that you will deal with, the parts of the function between the points can generally be approximated by drawing a line or curve between the points. Extending the function beyond the set of points is also possible, but becomes increasingly inaccurate.

Plotting points like this is laborious. Fortunately, many functions' graphs fall into general patterns. For a simple case, consider functions of the form

$$f(x) = ax$$

The graph of *f* is a single line, passing through (0,0) and (1,*a*). Thus, after plotting the two points, a straightedge may be used to draw the graph as far as is needed. After having learned calculus, you will know many more techniques for drawing good graphs of functions.

Algebraic manipulation

Purpose of review

This section is intended to review algebraic manipulation. It is important to understand algebra in order to do calculus. If you have a good knowledge of algebra, you should probably just skim this section to be sure you are familiar with the ideas.

Rules of arithmetic and algebra

The following rules are always true.

- Addition
  - Commutative Law: $a + b = b + a$.
  - Associative Law: $(a + b) + c = a + (b + c)$.
  - Additive Identity: $a + 0 = a$.
  - Additive Inverse: $a + (-a) = 0$.
- Subtraction
  - Definition: $a - b = a + (-b)$.
- Multiplication
  - Commutative law: $a \times b = b \times a$.
  - Associative law: $(a \times b) \times c = a \times (b \times c)$.
  - Multiplicative Identity: $a \times 1 = a$.
  - Multiplicative Inverse: $a \times \frac{1}{a} = 1$, whenever $a \neq 0$
  - Distributive law: $a \times (b + c) = a \times b + a \times c$.
- Division
  - Definition: $\frac{a}{b} = a \times \frac{1}{b}$, whenever $b \neq 0$.

The above laws are true for all *a*, *b*, and *c*, whether *a*, *b*, and *c* are numbers, variables, functions, or other expressions. For instance,

$$\begin{aligned}\frac{(x+2)(x+3)}{x+3} &= (((x+2)\times(x+3))\times(\frac{1}{x+3})) \\ &= ((x+2)\times((x+3)\times(\frac{1}{x+3}))) \\ &= ((x+2)\times(1)), \qquad x\neq -3 \\ &= x+2, \qquad x\neq -3.\end{aligned}$$

Of course, the above is much longer than simply cancelling $x + 3$ out in both the numerator and denominator. But, when you are cancelling, you are really just doing the above steps, so it is important to know what the rules are so as to know when you are allowed to cancel. Occasionally people do the following, for instance, which is incorrect:

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

The correct simplification is

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

where the number 2 cancels out in both the numerator and the denominator.

Exercises

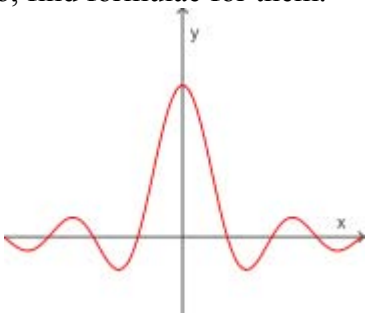
Functions

1. Let $f(x) = x^2$.
    1. Compute $f(0)$ and $f(2)$.
    2. What are the domain and range of f?
    3. Does f have an inverse? If so, find a formula for it.
2. Let $f(x) = x + 2$, $g(x) = 1 / x$.
    1. Give formulae for
        1. $f + g$,
        2. $f - g$,
        3. $g - f$,

4. $f \times g$,
5. f / g,
6. g / f,
7. $f \circ g$ and
8. $g \circ f$.

2. Compute $f(g(2))$ and $g(f(2))$.
3. Do f and g have inverses? If so, find formulae for them.

3. Does this graph represent a function?



Solutions-

1.
    1. $f(0) = 0$, $f(2) = 4$
    2. The domain is $(-\infty, \infty)$; the range is $[0, \infty)$.
    3. No, since f isn't one-to-one; for example, $f(-1) = f(1) = 1$.
2.
    1.
        1. $(f + g)(x) = x + 2 + 1 / x = (x^2 + 2x + 1) / x$.
        2. $(f - g)(x) = x + 2 - 1 / x = (x^2 + 2x - 1) / x$.
        3. $(g - f)(x) = 1 / x - x - 2 = (1 - x^2 - 2x) / x$.
        4. $(f \times g)(x) = (x + 2)/x$.
        5. $(f / g)(x) = x(x + 2)$ provided $x \neq 0$. Note that 0 is not in the domain of f / g, since it's not in the domain of g, and you can't divide by something that doesn't exist!
        6. $(g / f)(x) = 1 / [x(x + 2)]$. Although 0 is still not in the domain, we don't need to state it now, since 0 isn't in the domain of the expression $1 / [x(x + 2)]$ either.
        7. $(f \circ g)(x) = 1/x + 2 = (2x + 1)/x$.
        8. $(g \circ f)(x) = 1/(x + 2)$.
    2. $f(g(2)) = 5 / 2$; $g(f(2)) = 1 / 4$.
    3. Yes; $f^{-1}(x) = x - 2$ and $g^{-1}(x) = 1 / x$. Note that g and its inverse are the same.
3. As pictured, by the Vertical Line test, this graph represents a function.

Limits

Intuitive Look

A limit looks at what happens to a function when the input approaches a certain value. The general notation for a limit is as follows:

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

This is read as "The limit of $f(x)$ as x approaches a". We'll take up later the question of how we can determine whether a limit exists for $f(x)$ at a and, if so, what it is. For now, we'll look at it from an intuitive standpoint.

Let's say that the function that we're interested in is $f(x) = x^2$, and that we're interested in its limit as x approaches 2. Using the above notation, we can write the limit that we're interested in as follows:

$$\lim_{x \to 2} x^2$$

One way to try to evaluate what this limit is would be to choose values near 2, compute $f(x)$ for each, and see what happens as they get closer to 2. This is implemented as follows:

x	1.7	1.8	1.9	1.95	1.99	1.999
$f(x) = x^2$	2.89	3.24	3.61	3.8025	3.9601	3.996001

Here we chose numbers smaller than 2, and approached 2 from below. We can also choose numbers larger than 2, and approach 2 from above:

x	2.3	2.2	2.1	2.05	2.01	2.001
$f(x) = x^2$	5.29	4.84	4.41	4.2025	4.0401	4.004001

We can see from the tables that as x grows closer and closer to 2, $f(x)$ seems to get closer and closer to 4, regardless of whether x approaches 2 from above or from below. For this reason, we feel reasonably confident that the limit of x^2 as x approaches 2 is 4, or, written in limit notation,

$$\lim_{x \to 2} x^2 = 4.$$

Now let's look at another example. Suppose we're interested in the behavior of the function $f(x) = \frac{1}{x-2}$ as x approaches 2. Here's the limit in limit notation:

$$\lim_{x \to 2} \frac{1}{x-2}$$

Just as before, we can compute function values as x approaches 2 from below and from above. Here's a table, approaching from below:

x	1.7	1.8	1.9	1.95	1.99	1.999
$f(x) = \frac{1}{x-2}$	-3.333	-5	-10	-20	-100	-1000

And here from above:

x	2.3	2.2	2.1	2.05	2.01	2.001
$f(x) = \frac{1}{x-2}$	3.333	5	10	20	100	1000

In this case, the function doesn't seem to be approaching any value as x approaches 2. In this case we would say that the limit doesn't exist.

Both of these examples may seem trivial, but consider the following function:

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

This function is the same as

$$f(x) = \begin{cases} x^2 & \text{if } x \neq 2 \\ \text{undefined} & \text{if } x = 2 \end{cases}$$

Note that these functions are really completely identical; not just "almost the same," but actually, in terms of the definition of a function, completely the same; they give exactly the same output for every input.

In algebra, we would simply say that we can cancel the term $(x - 2)$, and then we have the function $f(x) = x^2$. This, however, would be a bit dishonest; the function that we have now is not really the same as the one we started with, because it is defined at $x = 2$, and our original function was, specifically, not defined at $x = 2$. In algebra we were willing to

ignore this difficulty because we had no better way of dealing with this type of function. Now, however, in calculus, we can introduce a better, more correct way of looking at this type of function. What we want is to be able to say that, even though at $x = 2$ the function doesn't exist, it works almost as though it does, and it's 4. It may not get there, but it gets really, really close. The only question that we have is: what do we mean by "close"?

Informal definition of a limit

As the precise definition of a limit is a bit technical, it is easier to start with an informal definition; we'll explain the formal definition later.

We suppose that a function f is defined for x near c (but we do not require that it be defined when $x = c$).

Definition: (Informal definition of a limit)
We call L the **limit of $f(x)$ as x approaches c** if $f(x)$ becomes close to L when x is close (but not equal) to c.

When this holds we write

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

or

$$f(x) \to L \quad \text{as} \quad x \to c.$$

Notice that the definition of a limit is not concerned with the value of $f(x)$ when $x = c$ (which may exist or may not). All we care about are the values of $f(x)$ when x is close to c, on either the left or the right (i.e. less or greater).

Limit rules

Now that we have defined, informally, what a limit is, we will list some rules that are useful for working with and computing limits. These will all be proven, or left as exercises, once we formally define the fundamental concept of the limit of a function.

First, the **constant rule** states that if $f(x) = b$ (that is, f is constant for all x) then the limit as x approaches c must be equal to b. In other words

$$\lim_{x \to c} b = b$$

.

Second, the **identity rule** states that if $f(x) = x$ (that is, f just gives back whatever number you put in) then the limit of f as x approaches c is equal to c. That is,

$$\lim_{x \to c} x = c$$.

The next few rules tell us how, given the values of some limits, to compute others. Suppose that $\lim_{x \to c} f(x) = L$ and $\lim_{x \to c} g(x) = M$ and that k is constant. Then

$$\begin{aligned}
\lim_{x \to c} kf(x) &= k \cdot \lim_{x \to c} f(x) = kL \\
\lim_{x \to c}[f(x) + g(x)] &= \lim_{x \to c} f(x) + \lim_{x \to c} g(x) = L + M \\
\lim_{x \to c}[f(x) - g(x)] &= \lim_{x \to c} f(x) - \lim_{x \to c} g(x) = L - M \\
\lim_{x \to c}[f(x) \cdot g(x)] &= \lim_{x \to c} f(x) \cdot \lim_{x \to c} g(x) = L \cdot M \\
\lim_{x \to c} \frac{f(x)}{g(x)} &= \frac{\lim_{x \to c} f(x)}{\lim_{x \to c} g(x)} = \frac{L}{M} \text{ as long as } M \neq 0
\end{aligned}$$

Notice that in the last rule we need to require that M is not equal to zero (otherwise we would be dividing by zero which is an undefined operation).

These rules are known as **identities**; they are the scalar product, sum, difference, product, and quotient rules for limits. (A scalar is a constant, and, when you multiply a function by a constant, we say that you are performing **scalar multiplication**.)

Using these rules we can deduce another. In particular, using the rule for products many times we get that

$$\lim_{x \to c} f(x)^n = \left(\lim_{x \to c} f(x) \right)^n$$ for a positive integer n.

This is called the **power rule**.

Examples

Example 1 Find the limit $\lim_{x \to 2} 4x^3$.

We need to simplify the problem, since we have no rules about this expression by itself. We know from the identity rule above that $\lim_{x \to 2} x = 2$. By the power rule,

$\lim_{x\to 2} x^3 = \left(\lim_{x\to 2} x\right)^3 = 2^3 = 8$. Lastly, by the scalar multiplication rule, we get $\lim_{x\to 2} 4x^3 = 4\lim_{x\to 2} x^3 = 4 \cdot 8 = 32$.

Example 2

Find the limit $\lim_{x\to 2}[4x^3 + 5x + 7]$.

To do this informally, we split up the expression, once again, into its components. As above, $\lim_{x\to 2} 4x^3 = 32$.

Also $\lim_{x\to 2} 5x = 5 \cdot \lim_{x\to 2} x = 5 \cdot 2 = 10$ and $\lim_{x\to 2} 7 = 7$. Adding these together gives

$$\lim_{x\to 2} 4x^3 + 5x + 7 = \lim_{x\to 2} 4x^3 + \lim_{x\to 2} 5x + \lim_{x\to 2} 7 = 32 + 10 + 7 = 49$$

.

Example 3

Find the limit, $\lim_{x\to 2} \frac{4x^3 + 5x + 7}{(x-4)(x+10)}$.

From the previous example the limit of the numerator is $\lim_{x\to 2} 4x^3 + 5x + 7 = 49$.
The limit of the denominator is

$$\lim_{x\to 2}(x-4)(x+10) = \lim_{x\to 2}(x-4) \cdot \lim_{x\to 2}(x+10) = (2-4)\cdot(2+10) = -24.$$

As the limit of the denominator is not equal to zero we can divide which gives

$$\lim_{x\to 2} \frac{4x^3 + 5x + 7}{(x-4)(x+10)} = -\frac{49}{24}.$$

Example 4

Find the limit, $\lim_{x\to 4} \frac{x^4 - 16x + 7}{4x - 5}$.

We apply the same process here as we did in the previous set of examples;

$$\lim_{x\to 4}\frac{x^4-16x+7}{4x-5}=\frac{\lim_{x\to 4}(x^4-16x+7)}{\lim_{x\to 4}(4x-5)}=\frac{\lim_{x\to 4}(x^4)-\lim_{x\to 4}(16x)+\lim_{x\to 4}(7)}{\lim_{x\to 4}(4x)-\lim_{x\to 4}5}.$$

We can evaluate each of these;
$\lim_{x\to 4}(x^4)=256, \lim_{x\to 4}(16x)=64, \lim_{x\to 4}(7)=7, \lim_{x\to 4}(4x)=16$ and
$\lim_{x\to 4}(5)=5$. Thus, the answer is $\frac{199}{11}$.

Example 5

Find the limit $\lim_{x\to 0}\frac{1-\cos x}{x}$.

To evaluate this seemingly complex limit, we will need to recall some sine and cosine identities. We will also have to use two new facts. First, if *f*(*x*) is a trigonometric function (that is, one of sine, cosine, tangent, cotangent, secant or cosecant) and is defined at *a*, then, $\lim_{x\to a} f(x)=f(a)$. Second, $\lim_{x\to 0}\frac{\sin x}{x}=1$.

To evaluate the limit, recognize that 1 − cos*x* can be multiplied by 1 + cos*x* to obtain (1 − $\cos^2 x$) which, by our trig identities, is $\sin^2 x$. So, multiply the top and bottom by 1 + cos*x*. (This is allowed because it is identical to multiplying by one.) This is a standard trick for evaluating limits of fractions; multiply the numerator and the denominator by a carefully chosen expression which will make the expression simplify somehow. In this case, we should end up with:

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

Our next step should be to break this up into $\lim_{x\to 0}\frac{\sin x}{x}\cdot\lim_{x\to 0}\frac{\sin x}{1+\cos x}$ by the product rule. As mentioned above, $\lim_{x\to 0}\frac{\sin x}{x}=1$.

Next, $\lim_{x\to 0}\frac{\sin x}{1+\cos x}=\frac{\lim_{x\to 0}\sin x}{\lim_{x\to 0}(1+\cos x)}=\frac{0}{1+\cos 0}=0$.

Thus, by multiplying these two results, we obtain 0.

We will now present an amazingly useful result, even though we cannot prove it yet. We can find the limit at *c* of any polynomial or rational function, as long as that rational function is defined at *c* (so we are not dividing by zero). That is, *c* must be in the domain of the function.

Limits of Polynomials and Rational functions

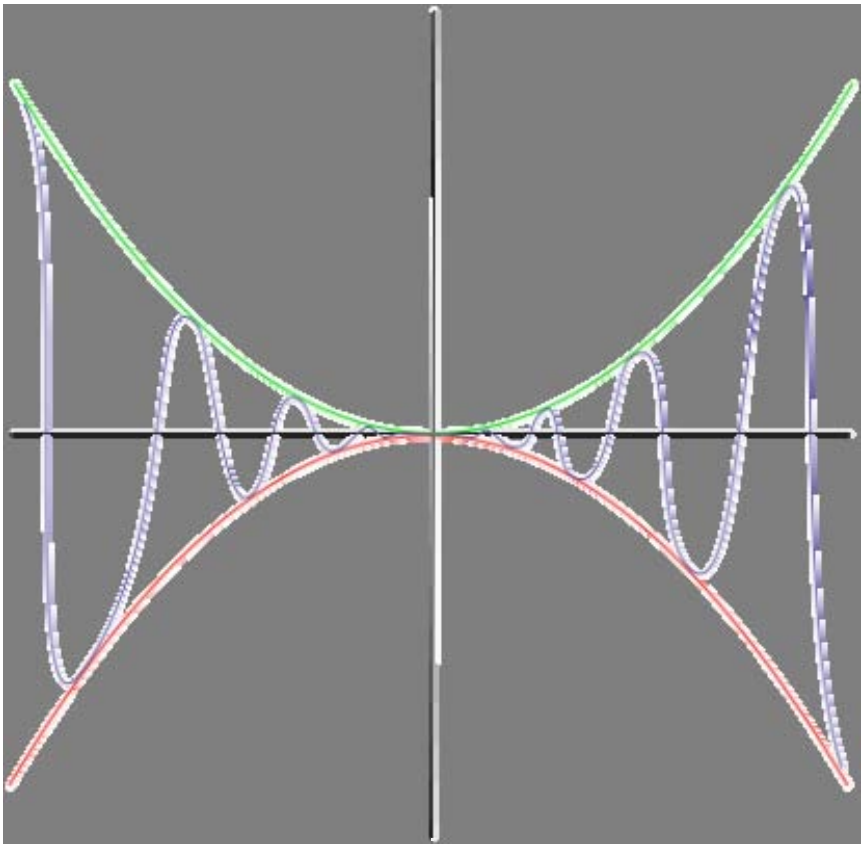
If *f* is a polynomial or rational function that is defined at *c* then

$$\lim_{x \to c} f(x) = f(c)$$

We already learned this for trigonometric functions, so we see that it is easy to find limits of polynomial, rational or trigonometric functions wherever they are defined. In fact, this is true even for combinations of these functions; thus, for example,

$$\lim_{x \to 1} (\sin x^2 + 4\cos^3(3x - 1)) = \sin 1^2 + 4\cos^3(3(1) - 1)$$

The Squeeze Theorem



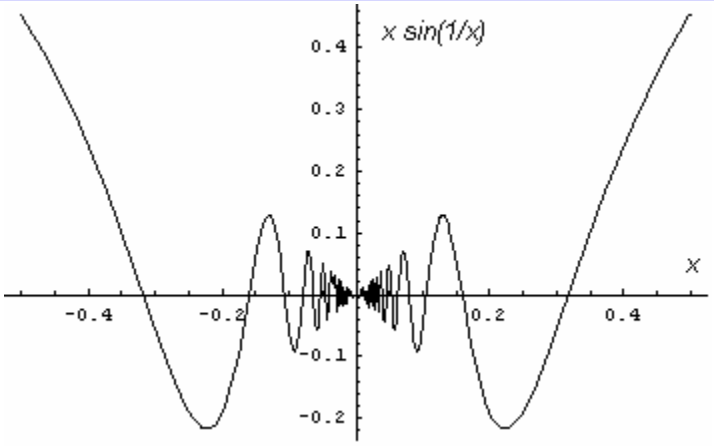
Graph showing *f* being squeezed between *g* and *h*

The Squeeze Theorem is very important in calculus, where it is typically used to find the limit of a function by comparison with two other functions whose limits are known.

It is called the Squeeze Theorem because it refers to a function *f* whose values are squeezed between the values of two other functions *g* and *h*, both of which have the same limit *L*. If the value of *f* is trapped between the values of the two functions *f* and *g*, the values of *f* must also approach *L*.

Expressed more precisely:

Theorem: (Squeeze Theorem)
Suppose that $g(x) \le f(x) \le h(x)$ holds for all *x* in some open interval containing *a*, except possibly at *x* = *a* itself. Suppose also that $\lim_{x \to a} g(x) = \lim_{x \to a} h(x) = L$. Then $\lim_{x \to a} f(x) = L$ also.



Plot of x*sin(1/x) for -0.5 < x <0.5

Example: Compute $\lim_{x \to 0} x \sin(1/x)$. Note that the sine of anything is in the interval [− 1,1]. That is, $-1 \le \sin x \le 1$ for all *x*. If *x* is positive, we can multiply these inequalities by *x* and get $-x \le x \sin(1/x) \le x$. If *x* is negative, we can similarly multiply the inequalities by the positive number - *x* and get $x \le x \sin(1/x) \le -x$. Putting these together, we can see that, for all nonzero *x*, $-|x| \le x \sin(1/x) \le |x|$. But it's easy to see that $\lim_{x \to 0} -|x| = \lim_{x \to 0} |x| = 0$. So, by the Squeeze Theorem, $\lim_{x \to 0} x \sin(1/x) = 0$.

Finding limits

Now, we will discuss how, in practice, to find limits. First, if the function can be built out of rational, trigonometric, logarithmic and exponential functions, then if a number *c* is in the domain of the function, then the limit at *c* is simply the value of the function at *c*.

If *c* is not in the domain of the function, then in many cases (as with rational functions) the domain of the function includes all the points near *c*, but not *c* itself. An example would be if we wanted to find $\lim_{x \to 0} \frac{x}{x}$, where the domain includes all numbers besides 0.

In that case, in order to find $\lim_{x \to c} f(x)$ we want to find a function *g*(*x*) similar to *f*(*x*), except with the hole at *c* filled in. The limits of *f* and *g* will be the same, as can be seen from the definition of a limit. By definition, the limit depends on *f*(*x*) only at the points where *x* is close to *c* but not equal to it, so the limit at *c* does not depend on the value of the function at *c*. Therefore, if $\lim_{x \to c} g(x) = L$, $\lim_{x \to c} f(x) = L$ also. And since the domain of our new function *g* includes *c*, we can now (assuming *g* is still built out of rational, trigonometric, logarithmic and exponential functions) just evaluate it at *c* as before. Thus we have $\lim_{x \to c} f(x) = g(c)$.

In our example, this is easy; canceling the *x*'s gives *g*(*x*) = 1, which equals *f*(*x*) = *x* / *x* at all points except 0. Thus, we have $\lim_{x \to 0} \frac{x}{x} = \lim_{x \to 0} 1 = 1$. In general, when computing limits of rational functions, it's a good idea to look for common factors in the numerator and denominator.

Lastly, note that the limit might not exist at all. There are a number of ways in which this can occur:

1. "Gap"
   There is a gap (not just a single point) where the function is not defined. As an example, in
   $f(x) = \sqrt{x^2 - 16}$

   $\lim_{x \to c} f(x)$ does not exist when $-4 \leq c \leq 4$. There is no way to "approach" the middle of the graph. Note also that the function also has no limit at the endpoints of the two curves generated (at *c* = − 4 and *c* = 4). For the limit to exist, the point must be approachable from *both* the left and the right. Note also that there is no limit at a totally isolated point on the graph.

2. "Jump"
   If the graph suddenly jumps to a different level, there is no limit. For example, let *f*(*x*) be the greatest integer $\leq x$. Then, if *c* is an integer, when *x* approaches *c*

from the right $f(x) = c$, while when $f(x)$ approaches from the left $f(x) = c - 1$. Thus $\lim_{x \to c} f(x)$ will not exist.

3. Vertical asymptote
   In
   $$f(x) = \frac{1}{x^2}$$
   the graph gets arbitrarily high as it approaches 0, so there is no limit. (In this case we sometimes say the limit is infinite; see the next section.)

4. Infinite oscillation
   These next two can be tricky to visualize. In this one, we mean that a graph continually rises above and falls below a horizontal line. In fact, it does this infinitely often as you approach a certain x-value. This often means that there is no limit, as the graph never approaches a particular value. However, if the height (and depth) of each oscillation diminishes as the graph approaches the x-value, so that the oscillations get arbitrarily smaller, then there might actually be a limit. The use of oscillation naturally calls to mind the trigonometric functions. An example of a trigonometric function that does not have a limit as x approaches 0 is
   $$f(x) = \sin \frac{1}{x}.$$
   As x gets closer to 0 the function keeps oscillating between - 1 and 1. In fact, $\sin(1 / x)$ oscillates an infinite number of times on the interval between 0 and any positive value of x. The sine function is equal to zero whenever $x = k\pi$, where k is a positive integer. Between every two integers k, sin$x$ goes back and forth between 0 and - 1 or 0 and 1. Hence, $\sin(1 / x) = 0$ for every $x = 1 / (k\pi)$. In between consecutive pairs of these values, $1 / (k\pi)$ and $1 / [(k + 1)\pi]$, $\sin(1 / x)$ goes back and forth from 0, to either - 1 or 1 and back to 0. We may also observe that there are an infinite number of such pairs, and they are all between 0 and $1 / \pi$. There are a finite number of such pairs between any positive value of x and $1 / \pi$, so there must be infinitely many between any positive value of x and 0. From our reasoning we may conclude that, as x approaches 0 from the right, the function $\sin(1 / x)$ does not approach any specific value. Thus,
   $\lim_{x \to 0} \sin(1/x)$ does not exist.

Using limit notation to describe asymptotes

Now consider the function

$$g(x) = \frac{1}{x^2}.$$

What is the limit as x approaches zero? The value of $g(0)$ does not exist; it is not defined.

$$g(0) = \frac{1}{0^2}$$

Notice, also, that we can make $g(x)$ as large as we like, by choosing a small x, as long as $x \neq 0$. For example, to make $g(x)$ equal to one trillion, we choose x to be 10^{-6}. Thus, $\lim_{x \to 0} 1/x^2$ does not exist.

However, we *do* know something about what happens to $g(x)$ when x gets close to 0 without reaching it. We want to say we can make $g(x)$ arbitrarily large (as large as we like) by taking x to be sufficiently close to zero, but not equal to zero. We express this symbolically as follows:

$$\lim_{x \to 0} g(x) = \lim_{x \to 0} \frac{1}{x^2} = \infty$$

Note that the limit does not exist at 0; for a limit, being ∞ is a special kind of not existing. In general, we make the following definition.

Definition: Informal definition of a limit being $\pm\infty$
We say the **limit of *f*(*x*) as *x* approaches *c* is infinity** if $f(x)$ becomes very big (as big as we like) when x is close (but not equal) to c.

In this case we write

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

or

$$f(x) \to \infty \quad \text{as} \quad x \to c.$$

Similarly, we say the **limit of *f*(*x*) as *x* approaches *c* is negative infinity** if $f(x)$ becomes very negative when x is close (but not equal) to c.

In this case we write

$$\lim_{x \to c} f(x) = -\infty$$

or

$$f(x) \to -\infty \quad \text{as} \quad x \to c.$$

An example of the second half of the definition would be that $\lim_{x \to 0} -1/x^2 = -\infty$.

Key application of limits

To see the power of the concept of the limit, let's consider a moving car. Suppose we have a car whose position is linear with respect to time (that is, a graph plotting the position with respect to time will show a straight line). We want to find the velocity. This is easy to do from algebra; we just take the slope, and that's our velocity.

But unfortunately, things in the real world don't always travel in nice straight lines. Cars speed up, slow down, and generally behave in ways that make it difficult to calculate their velocities.

Now what we really want to do is to find the velocity at a given moment (the instantaneous velocity). The trouble is that in order to find the velocity we need two points, while at any given time, we only have one point. We can, of course, always find the average speed of the car, given two points in time, but we want to find the speed of the car at one precise moment.

This is the basic trick of differential calculus, the first of the two main subjects of this book. We take the average speed at two moments in time, and then make those two moments in time closer and closer together. We then see what the limit of the slope is as these two moments in time are closer and closer, and say that this limit is the slope at a single instant.

We will study this process in much greater depth later in the book. First, however, we will need to study limits more carefully.

Continuity

We are now ready to define the concept of a function being **continuous**. The idea is that we want to say that a function is continuous if you can draw its graph without taking your pencil off the page. But sometimes this will be true for some parts of a graph but not for others. Therefore, we want to start by defining what it means for a function to be continuous at *one point*. The definition is simple, now that we have the concept of limits:

Definition: (continuity at a point)

If $f(x)$ is defined on an open interval containing c, then $f(x)$ is said to be **continuous at c** if and only if $\lim_{x \to c} f(x) = f(c)$.

Note that for *f* to be continuous at *c*, the definition in effect requires three conditions:

1. that *f* is defined at *c*, so *f*(*c*) exists,
2. the limit as *x* approaches *c* exists, and
3. the limit and *f*(*c*) are equal.

If any of these do not hold then *f* is not continuous at *c*.

The idea of the definition is that the point of the graph corresponding to *c* will be close to the points of the graph corresponding to nearby *x*-values. Now we can define what it means for a function to be continuous in general, not just at one point.

A function is said to be **continuous** if it is continuous at every point in its domain.

Discontinuities

A **discontinuity** is a point where a function is not continuous. There are lots of possible ways this could happen, of course. Here we'll just discuss two simple ways.

Removable Discontinuities

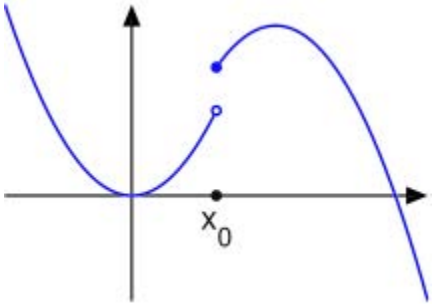
The function $f(x) = \frac{x^2 - 9}{x - 3}$ is not continuous at $x = 3$. It is discontinuous at that point because the fraction then becomes $\frac{0}{0}$, which is indeterminate. Therefore the function fails the first of our three conditions for continuity at the point 3; 3 is just not in its domain.

However, we say that this discontinuity is **removable**. This is because, if we modify the function at that point, we can eliminate the discontinuity and make the function continuous. To see how to make the function *f*(*x*) continuous, we have to simplify *f*(*x*), getting $f(x) = \frac{x^2 - 9}{x - 3} = \frac{(x+3)(x-3)}{(x-3)} = \frac{x+3}{1} \cdot \frac{x-3}{x-3}$. We can define a new function *g*(*x*) where *g*(*x*) = *x* + 3. Note that the function *g*(*x*) is not the same as the original function *f*(*x*), because *g*(*x*) is defined at *x* = 3, while *f*(*x*) is not. Thus, *g*(*x*) is continuous at *x* = 3, since $\lim_{x \to 3}(x+3) = 6 = g(3)$. However, whenever $x \neq 3$, *f*(*x*) = *g*(*x*); all we did to *f* to get *g* was to make it defined at *x* = 3.

In fact, this kind of simplification is always possible with a discontinuity in a rational function. We can divide the numerator and the denominator by a common factor (in our example *x* − 3) to get a function which is the same except where that common factor was

0 (in our example at $x = 3$). This new function will be identical to the old except for being defined at new points where previously we had division by 0.

Jump Discontinuities



Unfortunately, not all discontinuities can be removed from a function. Consider this function:

$$k(x) = \begin{cases} 1, & \text{if } x > 0 \\ -1, & \text{if } x \leq 0 \end{cases}$$

Since $\lim_{x \to 0} k(x)$ does not exist, there is no way to redefine *k* at one point so that it will be continuous at 0. These sorts of discontinuities are called *nonremovable* discontinuities.

Note, however, that both one-sided limits exist; $\lim_{x \to 0^-} k(x) = -1$ and $\lim_{x \to 0^+} k(x) = 1$. The problem is that they are not equal, so the graph "jumps" from one side of 0 to the other. In such a case, we say the function has a *jump* discontinuity. (Note that a jump discontinuity is a kind of nonremovable discontinuity.)

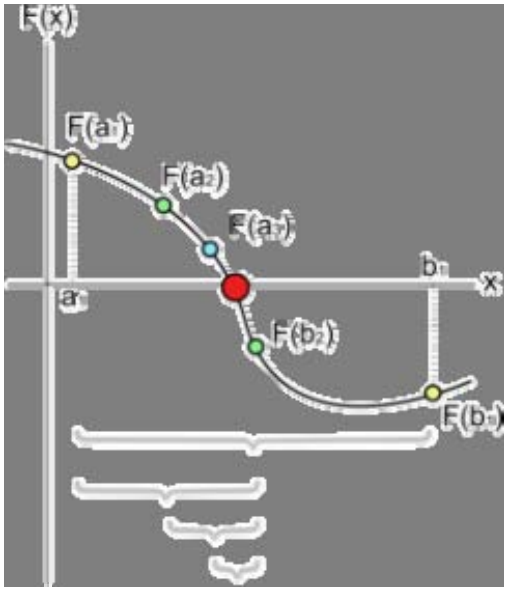
One-Sided Continuity

Just as a function can have a one-sided limit, a function can be continuous from a particular side.

Intermediate value theorem (IVT)

The intermediate value theorem is a very important theorem in calculus and analysis. It says:

If a function is continuous on a closed interval [a,b], then for every value k between f(a) and f(b) there is a value c on [a,b] such that f(c)=k.



A few steps of the bisection method applied over the starting range $[a_1;b_1]$. The bigger red dot is the root of the function.

The bisection method is the simplest and most reliable algorithm to find roots to an equation.

Suppose we want to solve the equation f(x) = 0. Given two points a and b such that f(a) and f(b) have opposite signs, we know by the intermediate value theorem that f must have at least one root in the interval [a, b] as long as f is continuous on this interval. The bisection method divides the interval in two by computing c = (a+b) / 2. There are now two possibilities: either f(a) and f(c) have opposite signs, or f(c) and f(b) have opposite signs. The bisection algorithm is then applied recursively to the sub-interval where the sign change occurs. In this way we home in to a small sub-interval containing the root. The mid point of that small sub-interval is usually taken as the root.

What is differentiation?

Differentiation is a method that allows us to find a function that relates the **rate of change** of one variable with respect to another variable.

Informally, we may suppose that we're tracking the position of a car on a two-lane road with no passing lanes. Assuming the car never pulls off the road, we can abstractly study the car's position by assigning it a variable, x. Since the car's position changes as the time changes, we say that x is dependent on time, or x = x(t). But this is not enough to study how the car's position changes as the time changes; so far, we only have a way to tell where the car is at a specific time. Differentiation allows us to study *dx* / *dt*, which is the mathematical expression for how the car's position changes with respect to time.

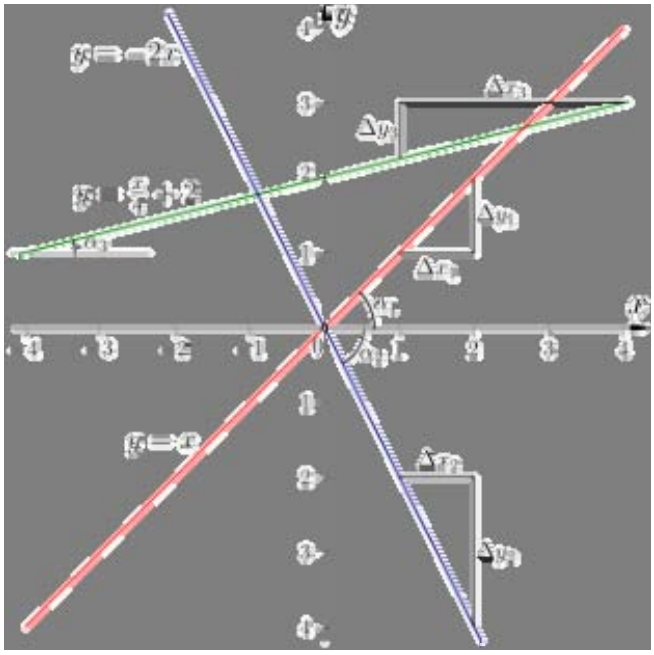
Formally, we are able to find the slope at any point of a non-linear function (compare with determining the slope of a linear function: quite easy). Suppose we have determined the position of a particle at any time t modeled by the function:

$$x = 3t^2$$

Differentiating this would give us the amount of **change in distance with respect to time** (a rate, or speed).

The Definition of Slope

On a Line



Three lines with different gradients

The **slope** of a line, also called the **gradient** of the line, is a measure of its inclination. A line that is horizontal has slope 0, a line from the bottom left to the top right has a positive slope, a line from the top left to the bottom right has a negative slope.

Gradient can be defined in two (equivalent) ways. The first way is to express it as how much the line climbs for a given "step" horizontally. We denote a step in a quantity using a delta (Δ) symbol. Thus, a step in *x* is written as Δ*x*. We can therefore write this definition of gradient as:

$$\text{Gradient} = \frac{\Delta y}{\Delta x}$$

Alternatively, we can define gradient as the "tangent function" of the line:

$$\text{Gradient} = \tan(\alpha),$$

where α is the angle between the line to the horizontal (measured clockwise). Those who know how the tangent function is generated (opposite side over adjacent side) will be able to spot the equivalence here.

If we have two points on a line, $P(x_1, y_1)$ and $Q(x_2, y_2)$, the step in *x* from *P* to *Q* is given by:

$$\Delta x = x_2 - x_1$$

Likewise, the step in *y* from *P* to *Q* is given by:

$$\Delta y = y_2 - y_1$$

This leads to the very important result below.

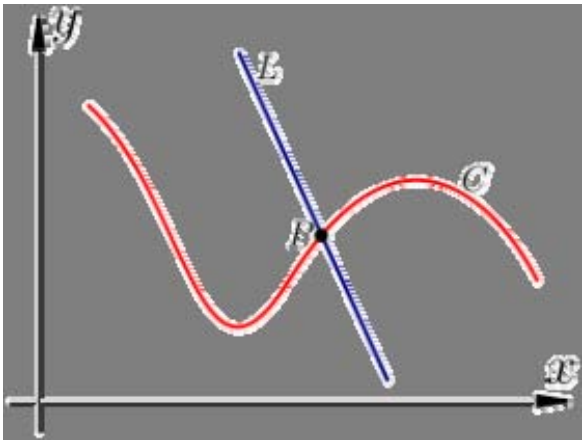
The definition of slope, *m*, between two points (x_1,y_1) and (x_2,y_2) on a line is

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

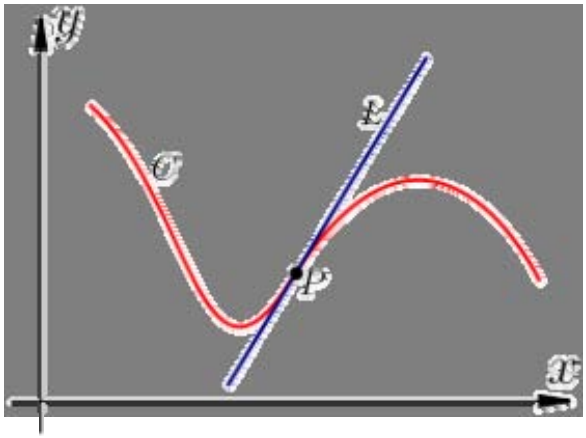
On a Function

Most functions we are interested in are not straight lines (although they can be). We cannot define a gradient of a curved function in the same way as we can for a line. In order for us to understand how to find the gradient of a function at a point, we will first have to cover the idea of **tangency**. A **tangent** is a line which *just* touches a curve at a point, such that the angle between them at that point is zero. Consider the following four curves and lines:

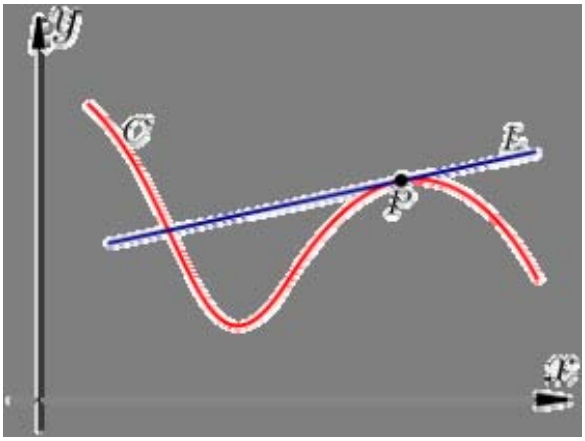
(i)



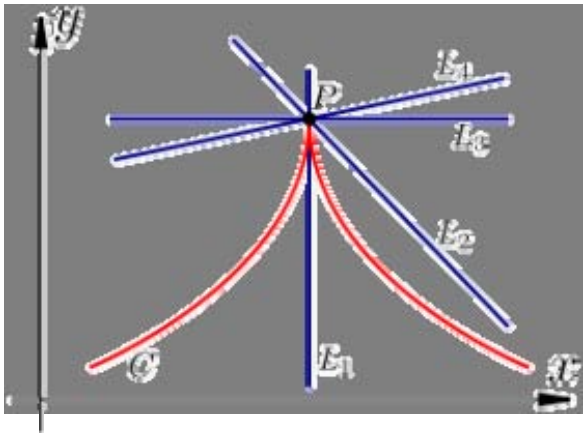
(ii)



(iii)

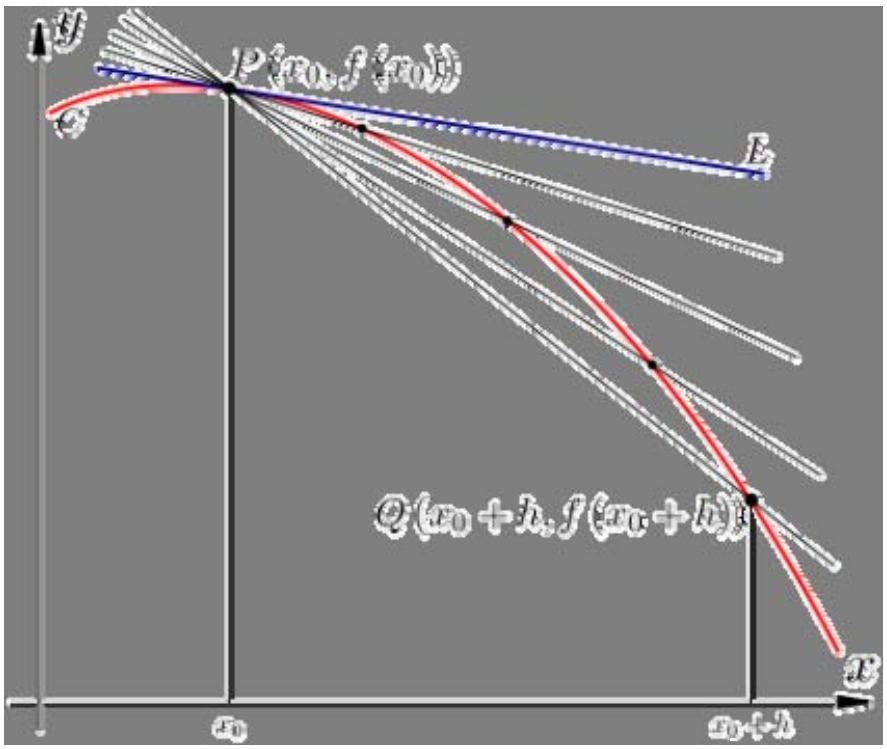


(iv)



i. The line *L* crosses, but is not tangent to *C* at *P*.
ii. The line *L* crosses, and is tangent to *C* at *P*.
iii. The line *L* crosses *C* at more than one point, but is tangent to *C* at *P*.
iv. There are many lines that cross *C* at *P*, but none are tangent. In fact, this curve has an *undefined* tangent at 'P.

A **secant** is a line drawn through two points on a curve. We can construct a definition of the tangent as the limit of a secant of the curve drawn as the separation between the points tends to zero. Consider the diagram below.



As the distance *h* tends to zero, the secant line becomes the tangent at the point x_0. The two points we draw our line through are:

$$P\left(x_0, f\left(x_0\right)\right)$$

and

$$Q\left(x_0 + h, f\left(x_0 + h\right)\right)$$

As a secant line is simply a line and we know two points on it, we can find its slope, *m*, from before:

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

Substituting in the points on the line,

$$m = \frac{f\left(x_0 + h\right) - f\left(x_0\right)}{\left(x_0 + h\right) - x_0}.$$

This simplifies to

$$m = \frac{f\left(x_0 + h\right) - f\left(x_0\right)}{h}.$$

This expression is called the **difference quotient**. Note that *h* can be positive or negative — it is perfectly valid to take a secant with a secondary point to the left.

Now, to find the slope of the tangent, m_0 we let *h* be zero. We cannot simply set it to zero as this would imply division of zero by zero which would yield an undefined result. Instead we must find the limit of the above expression as *h* **tends** to zero:

$$m_0 = \lim_{h \to 0} \left[\frac{f(x_0 + h) - f(x_0)}{h} \right]$$

The Slope at a Point

Consider the formula for average velocity in the x-direction, $\frac{\Delta x}{\Delta t}$. This formula can be used to find approximate results for speed, but is rarely exact. To correct this we look at the **change in position as the change in time approaches 0**. Mathematically this is written as: $\lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} = \frac{dx}{dt}$, where d denotes change, x denotes distance, and t denotes time. Compare the operator d with Δ. The delta should be familiar to you from studying slope. They both indicate a difference between two numbers, however d denotes an infinitesimal difference.

(Note that the letter s is often used to denote distance, which would yield $\frac{ds}{dt}$. The letter d is often avoided in denoting distance due to the ambiguity in the expression $\frac{dd}{dt}$.)

If a function f(x) is plotted on an (x,y) Cartesian coordinate system, differentiation will yield a function which describes the rate of change of y with respect to x.

The rate of change at a specific point is called the instantaneous rate of change. The line with a slope equal to the instantaneous rate of change and that only touches the graph at that specific point is known as a tangent.

Historically, the primary motivation for the study of **differentiation** was the tangent line problem: for a given curve, find the slope of the straight line that is tangent to the curve at a given point. The word tangent comes from the Latin word *tangens*, which means touching. Thus, to solve the tangent line problem, we need to find the slope of a line that is "touching" a given curve at a given point. But what exactly do we mean by "touching"?

The solution is obvious in some cases: for example, a line $y = mx + c$ is its own tangent; the slope at any point is m. For the parabola $y = x^2$, the slope at the point (0,0) is 0 (the tangent line is flat). In fact, at any vertex of any smooth function the slope is zero, because the slope of the tangent lines on either side of the point are opposite signs.

But how can you find the slope of, say, $y = \sin x + x^2$ at $x = 1.5$?

The easiest way to find slopes for any function is by differentiation. Differentiation results in another function whose value for any value x is the slope of the original function at x. This function is known as the derivative of the original function, and is denoted by either a prime sign, as in $f'(x)$(read "f prime of x"), the quotient notation, $\frac{df}{dx}$ or $\frac{d}{dx}[f]$, which is more useful in some cases, or the differential operator notation, $D_x[f(x)]$, which is generally just written as $Df(x)$.

Most of the time the brackets are not needed, but are useful for clarity if we are dealing with something like $D(fg)$ for a product.

Example: If $f(x) = 3x + 5$, $f'(x) = 3$, no matter what x.

Example: If $f(x) = |x|$ (the absolute value function) then

$$f'(x) = \begin{cases} -1, & x < 0 \\ \text{undefined}, & x = 0 \\ 1, & x > 0 \end{cases}.$$

Here, $f(x)$ is not smooth (though it is continuous) at $x = 0$ and so the limits $\lim_{x \to 0+} f'(x)$ and $\lim_{x \to 0-} f'(x)$(the limits as 0 is approached from the right and left respectively) are not equal. f'(0) is said to be undefined, and so f'(x) has a discontinuity at 0. This sort of point of non-differentiability is called a cusp. Functions may also not be differentiable because they go to infinity at a point, or oscillate infinitely frequently.

The Definition of the Derivative

The derivative is the formula m_{tan} (slope of tangent line) on a curve at a specific point.

An example

Draw a curve defined as $y = 3x^2$ and select any point on it. We select the point at which $x = 4$; *what is the slope at this point?* We can do it "the hard (and imprecise) way", *without* using differentiation, as follows, using a calculator and using small differences below and above the given point:

When $x = 3.999$, $y = 47.976003$.

When $x = 4.001$, $y = 48.024003$.

Then the difference between the two values of x is $\Delta x = 0.002$.

Then the difference between the two values of y is $\Delta y = 0.048$.

Thus, the slope $= \frac{\Delta y}{\Delta x} = 24$ at the point of the graph at which $x = 4$.

Now using differentiation rules (see below) to solve this problem again, when $y = 3x^2$ then the slope at any point of that curve is found by evaluating $y' = 6x$.

Our x is 4, so that $y' = \frac{\Delta y}{\Delta x} = 6 * 4 = 24$ again. No need for a calculator!

$$f'(x) = \lim_{\Delta x \to 0} \left[\frac{f(x + \Delta x) - f(x)}{\Delta x} \right]$$

This is the definition of the derivative. If the limit exists we say that f is differentiable at x and its derivative at x is $f'(x)$. A visual explanation of this formula is that the slope of the tangent line is the limit of the slope of a secant line when the difference of the points (Δx) tends to zero.

Example

Let us try this for a simple function:

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

$$= \frac{1}{2}$$

This is consistent with the definition of the derivative as the slope of a function.

Sometimes, the slope of a function varies with *x*. This is demonstrated by the function $f(x) = x^2$,

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

Understanding the Derivative Notation

The derivative notation is special and unique in mathematics. The most common use of derivatives you'll run into when first starting out with differentiating is the Leibniz notation, expressed as $\frac{dy}{dx}$. You may think of this as "rate of change in y with respect to x". You may also think of it as "infinitesimal value of y divided by infinitesimal value of x". Either way is a good way of thinking. Often, in an equation, you will see just $\frac{d}{dx}$, which literally means "derivative with respect to x". You may safely assume that it is the equivalent of $\frac{dy}{dx}$ for now.

As you advance through your studies, you will see that *dy* and *dx* can act as separate entities that can be multiplied and divided (to a certain degree). Eventually you will see derivatives such as $\frac{dx}{dy}$, which sometimes will be written $\frac{d}{dy}$. Or, you may see a derivative in polar coordinates marked as $\frac{d\theta}{dr}$.

All of the following are equivalent for expressing the derivative of $y = x^2$

- $\frac{dy}{dx} = 2x$
- $\frac{d}{dx}x^2 = 2x$
- $dy = 2x dx$
- $f'(x) = 2x$
- $D(f(x)) = 2x_{\text{S}}$

Exercises

Using the definition of the derivative find the derivative of the function $f(x) = 2x + 3$

Using the definition of the derivative find the derivative of the function $f(x) = x^3$. Now try $f(x) = x^4$. Can you see a pattern? In the next section we will find the derivative of $f(x) = x^n$ for all n.

The text states that the derivative of $|x|$ is not defined at $x = 0$. Use the definition of the derivative to show this.

Graph the derivative to $y = 4x^2$ on a piece of graph paper without solving for dy / dx. Then, solve for dy / dx and graph that; compare the two graphs.

Use the definition of the derivative to show that the derivative of sin$x$ is cos$x$. Hint: Use a suitable sum to product formula and the fact $\lim_{t \to 0} \frac{\sin t}{t} = 1$

Differentiation rules

The process of differentiation is tedious for large functions. Therefore, rules for differentiating general functions have been developed, and can be proved with a little effort. Once sufficient rules have been proved, it will be possible to differentiate a wide variety of functions. Some of the simplest rules involve the derivative of linear functions.

Derivative of a Constant Function

For any fixed real number c,

$$\frac{d}{dx}[c] = 0$$

Intuition

The function f(x) = c is a horizontal line, which has a constant slope of zero. Therefore, it should be expected that the derivative of this function is zero, regardless of the value of x.

Proof

From the definition of a derivative:

$$\lim_{\Delta x \to 0} \left[\frac{f(x + \Delta x) - f(x)}{\Delta x} \right]$$

Let $f(x) = c$. Then $f(x + \Delta x) = c$ because there are no x's in the function with which to plug in $x + \Delta x$. Therefore:

$$\frac{d}{dx}[c] = \lim_{\Delta x \to 0} \left[\frac{(c) - c}{\Delta x} \right] = 0$$

Example

$$\frac{d}{dx}[3] = 0$$

Example

$$\frac{d}{dx}[z] = 0$$

Derivative of a Linear Function

For any fixed real numbers *m* and *c*,

$$\frac{d}{dx}[mx + c] = m$$

The special case $\frac{dx}{dx} = 1$ shows the advantage of the $\frac{d}{dx}$ notation -- rules are intuitive by basic algebra, though this does not constitute a proof, and can lead to misconceptions to what exactly *dx* and *dy* actually are.

Constant multiple and addition rules

Since we already know the rules for some very basic functions, we would like to be able to take the derivative of more complex functions and break them up into simpler

functions. Two tools that let us do this are the constant multiple rules and the addition rule.

The Constant Rule

For any fixed real number *c*,

$$\frac{d}{dx}\left[cf(x)\right] = c\frac{d}{dx}\left[f(x)\right]$$

The reason, of course, is that one can factor *c* out of the numerator, and then of the entire limit, in the definition.

Example

We already know that

$$\frac{d}{dx}\left[x^2\right] = 2x$$

.

Suppose we want to find the derivative of $3x^2$

$$\begin{aligned}\frac{d}{dx}\left[3x^2\right] &= 3\frac{d}{dx}\left[x^2\right] \\ &= 3 \times 2x \\ &= 6x\end{aligned}$$

Another simple rule for breaking up functions is the addition rule.

The Addition and Subtraction Rules

$$\frac{d}{dx}\left[f(x) \pm g(x)\right] = \frac{d}{dx}\left[f(x)\right] \pm \frac{d}{dx}\left[g(x)\right]$$

Proof

From the definition:

$$\lim_{\Delta x \to 0}\left[\frac{\left[f(x+\Delta x) \pm g(x+\Delta x)\right] - \left[f(x) \pm g(x)\right]}{\Delta x}\right]$$

$$= \lim_{\Delta x \to 0} \left[\frac{[f(x+\Delta x) - f(x)] \pm [g(x+\Delta x) - g(x)]}{\Delta x} \right]$$

$$= \lim_{\Delta x \to 0} \left[\frac{[f(x+\Delta x) - f(x)]}{\Delta x} \right] \pm \lim_{\Delta x \to 0} \left[\frac{[g(x+\Delta x) - g(x)]}{\Delta x} \right]$$

By definition then, this last term is $\frac{d}{dx}[f(x)] \pm \frac{d}{dx}[g(x)]$

Example:

$$\begin{aligned} \frac{d}{dx}\left[3x^2 + 5x\right] &= \frac{d}{dx}\left[3x^2 + 5x\right] \\ &= \frac{d}{dx}\left[3x^2\right] + \frac{d}{dx}[5x] \\ &= 6x + \frac{d}{dx}[5x] \\ &= 6x + 5 \end{aligned}$$

The fact that both of these rules work is extremely significant mathematically because it means that differentiation is **linear**. You can take an equation, break it up into terms, figure out the derivative individually and build the answer back up, and nothing odd will happen.

We now need only one more piece of information before we can take the derivatives of any polynomial.

The Power Rule

$$\frac{d}{dx}[x^n] = nx^{n-1}, x \neq 0$$

For example, in the case of x^2 the derivative is $2x^1 = 2x$ as was established earlier. This rule is actually in effect in linear equations too, since $x^{n-1} = x^0$ when n=1, and of course, any real number or variable to the zero power is one.

This rule also applies to fractional and negative powers. Therefore

$$\frac{d}{dx}\left[\sqrt{x}\right] = \frac{d}{dx}\left[x^{1/2}\right]$$

$$= \frac{1}{2} x^{-1/2}$$
$$= \frac{1}{2\sqrt{x}}$$

Since polynomials are sums of monomials, using this rule and the addition rule lets you differentiate any polynomial. A relatively simple proof for this can be derived from the binomial expansion theorem.

Derivatives of polynomials

With these rules in hand, you can now find the derivative of any polynomial you come across. Rather than write the general formula, let's go step by step through the process.

$$\frac{d}{dx}\left[6x^5 + 3x^2 + 3x + 1\right]$$

The first thing we can do is to use the addition rule to split the equation up into terms:

$$\frac{d}{dx}\left[6x^5\right] + \frac{d}{dx}\left[3x^2\right] + \frac{d}{dx}\left[3x\right] + \frac{d}{dx}\left[1\right].$$

We can immediately use the linear and constant rules to get rid of some terms:

$$\frac{d}{dx}\left[6x^5\right] + \frac{d}{dx}\left[3x^2\right] + 3 + 0.$$

Now you may use the constant multiplier rule to move the constants outside the derivatives:

$$6\frac{d}{dx}\left[x^5\right] + 3\frac{d}{dx}\left[x^2\right] + 3.$$

Then use the power rule to work with the individual monomials:

$$6\left(5x^4\right) + 3\left(2x\right) + 3.$$

And then do some algebra to get the final answer:

$$30x^4 + 6x + 3.$$

These are not the only differentiation rules. There are other, more advanced, differentiation rules, which will be described in a later chapter.

Exercises

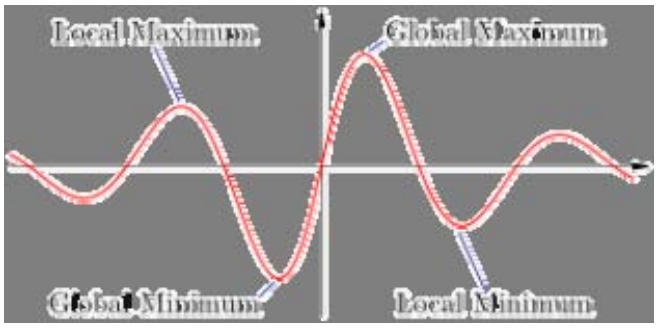
- Find the derivatives of the following equations:

  $f(x) = 42$
  $f(x) = 6x + 10$
  $f(x) = 2x^2 + 12x + 3$

- Use the definition of a derivative to prove the derivative of a constant function, of a linear function, and the constant rule and addition or subtraction rules.
- Answers:

  $f'(x) = 0$
  $f'(x) = 6$
  $f'(x) = 4x + 12$

Extrema and Points of Inflexion



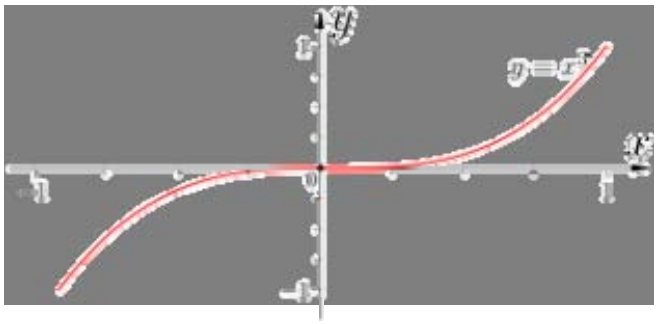
The four types of extrema.

Maxima and **minima** are points where a function reaches a highest or lowest value, respectively. There are two kinds of **extrema** (a word meaning maximum *or* minimum): **global** and **local**, sometimes referred to as "absolute" and "relative", respectively. A global maximum is a point that takes the largest value on the entire range of the function, while a global minimum is the point that takes the smallest value on the range of the function. On the other hand, local extrema are the largest or smallest values of the function in the immediate vicinity.

All extrema look like the crest of a hill or the bottom of a bowl on a graph of the function. A global extremum is always a local extremum too, because it is the largest or smallest value on the entire range of the function, and therefore also its vicinity. It is also possible to have a function with no extrema, global or local: $y=x$ is a simple example.

At any extremum, the slope of the graph is necessarily zero, as the graph must stop rising or falling at an extremum, and begin to fall or rise. Because of this, extrema are also commonly called **stationary points** or **turning points**. Therefore, the first derivative of a function is equal to zero at extrema. If the graph has one or more of these stationary

points, these may be found by setting the first derivative equal to zero and finding the roots of the resulting equation.



The function $f(x)=x^3$, which contains a point of inflexion at the point (0,0).

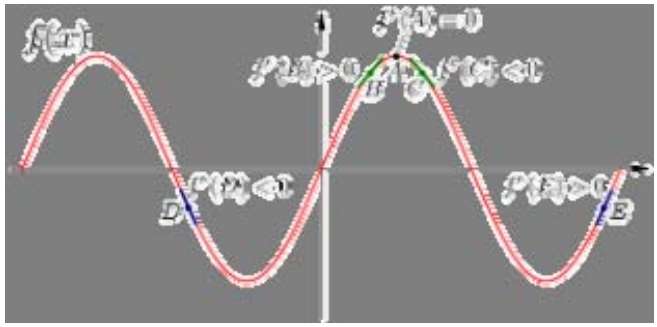
However, a slope of zero does not guarantee a maximum or minimum: there is a third class of stationary point called a point of inflexion. Consider the function

$$f(x) = x^3.$$

The derivative is

$$f'(x) = 3x^2$$

The slope at x=0 is 0. We have a slope of zero, but while this makes it a stationary point, this doesn't mean that it is a maximum or minimum. Looking at the graph of the function you will see that x=0 is neither, it's just a spot at which the function flattens out. True extrema require the a sign change in the first derivative. This makes sense - you have to rise (positive slope) to and fall (negative slope) from a maximum. In between rising and falling, on a smooth curve, there will be a point of zero slope - the maximum. A minimum would exhibit similar properties, just in reverse.



Good (*B* and *C*, green) and bad (*D* and *E*, blue) points to check in order to classify the extremum (*A*, black). The bad points lead to an incorrect classification of *A* as a minimum.

This leads to a simple method to classify a stationary point - plug x values slightly left and right into the derivative of the function. If the results have opposite signs then it is a

true maximum/minimum. You can also use these slopes to figure out if it is a maximum or a minimum: the left side slope will be positive for a maximum and negative for a minimum. However, you must exercise caution with this method, as, if you pick a point too far from the extremum, you could take it on the far side of *another* extremum and incorrectly classify the point.

The Extremum Test

A more rigorous method to classify a stationary point is called the **extremum test**. As we mentioned before, the sign of the first derivative must change for a stationary point to be a true extremum. Now, the *second* derivative of the function tells us the rate of change of the first derivative. It therefore follows that if the second derivative is positive at the stationary point, then the gradient is increasing. The fact that it is a stationary point in the first place means that this can only be a minimum. Conversely, if the second derivative is negative at that point, then it is a maximum.

Now, if the second derivative is zero, we have a problem. It could be a point of inflexion, or it could still be an extremum. Examples of each of these cases are below - all have a second derivative equal to zero at the stationary point in question:

- $y = x^3$ has a point of inflexion at $x = 0$
- $y = x^4$ has a minimum at $x = 0$
- $y = -x^4$ has a maximum at $x = 0$

However, this is not an insoluble problem. What we must do is continue to differentiate until we get, at the *(n+1)*th derivative, a non-zero result at the stationary point:

$$f'(x) = 0,\ f''(x) = 0,\ \ldots,\ f^{(n)}(x) = 0,\ f^{(n+1)}(x) \neq 0$$

If *n* is odd, then the stationary point is a true extremum. If the *(n+1)*th derivative is positive, it is a minimum; if the *(n+1)*th derivative is negative, it is a maximum. If *n* is even, then the stationary point is a point of inflexion.

As an example, let us consider the function

$$f(x) = -x^4$$

We now differentiate until we get a non-zero result at the stationary point at x=0 (assume we have already found this point as usual):

$$\begin{aligned} f'(x) &= -4x^3 \\ f''(x) &= -12x^2 \\ f'''(x) &= -24x \\ f^{(4)}(x) &= -24 \end{aligned}$$

Therefore, *(n+1)* is 4, so *n* is 3. This is odd, and the fourth derivative is negative, so we have a maximum. Note that none of the methods given can tell you if this is a global extremum or just a local one. To do this, you would have to set the function equal to the height of the extremum and look for other roots.

More Differentiation Rules

Chain Rule

We know how to differentiate regular polynomial functions. For example:

$$\frac{d}{dx}(3x^3 - 6x^2 + x) = 9x^2 - 12x + 1$$

$$f(x) = (x^2 + 5)^2$$
$$f(x) = x^4 + 10x^2 + 25$$
$$f'(x) = 4x^3 + 20x$$

However, there is a useful rule known as the **chain method rule**. The function above ($f(x) = (x^2 + 5)^2$) can be consolidated into two nested parts $f(x) = u^2$, where $u = m(x) = (x^2 + 5)$. Therefore:

if
$g(u) = u^2$ and
$u = m(x) = x^2 + 5$

Then:

$$f(x) = g(m(x))$$

Then

$$f'(x) = g'(m(x))m'(x)$$

The **chain rule** states that if we have a function of the form $y(u(x))$ (i.e. y can be written as a function of u and u can be written as a function of x) then:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

Chain Rule

If a function F(x) is composed to two differentiable functions g(x) and m(x), so that F(x)=g(m(x)), then F(x) is differentiable and,

$$F'(x) = g'(m(x))m'(x)$$

We can now investigate the original function:

$$\frac{dy}{du} = 2u$$
$$\frac{du}{dx} = 2x$$

Therefore

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} = 2u \cdot 2x = 2(x^2 + 5)(2x) = 4x^3 + 20x$$

This can be performed for more complicated equations. If we consider:

$$\frac{d}{dx}\sqrt{1 + x^2}$$

and let $y=\sqrt{u}$ and $u=1+x^2$, so that $dy/du=1/2\sqrt{u}$ and $du/dx=2x$, then, by applying the chain rule, we find that

$$\frac{d}{dx}\sqrt{1 + x^2} = \frac{1}{2\sqrt{1 + x^2}} \cdot 2x = \frac{x}{\sqrt{1 + x^2}}$$

So, in just plain words, for the chain rule you take the normal derivative of the **whole thing** (make the exponent the coefficient, then multiply by original function but decrease the exponent by 1) then multiply by the derivative of the inside.

Product and Quotient Rules

When we wish to differentiate a more complicated expression such as:

$$h(x) = (x^2 + 5)^5 \cdot (x^3 + 2)^3$$

our only way (up to this point) to differentiate the expression is to expand it and get a polynomial, and then differentiate that polynomial. This method becomes very complicated and is particularly error prone when doing calculations by hand. It is advantageous to find the derivative of $h(x)$ using just the functions $f(x) = (x^2+5)^5$ and $g(x) = (x^3 + 2)^3$ and their derivatives.

Derivatives of products (Product rule)

$$\frac{d}{dx}\left[f(x)\cdot g(x)\right] = f'(x)\cdot g(x) + f(x)\cdot g'(x)$$

What this rule basically means is that if one has a function that is the product of two functions, then all one has to do is differentiate the first function, multiply it by the other undifferentiated function, add that to the first function undifferentiated multiplied by the differentiated second function. For example, if one were to take the function

$$H(x) = (x+2)^3\cdot 2(x+1)^2$$

its derivative would ***not*** be

$$3(x+2)^2\cdot 4(x+1)$$

Instead it would be

$$3(x+2)^2\cdot 2(x+1)^2 + (x+2)^3\cdot 4(x+1)$$

Another way of approaching this is if one were to have a function that was a product of the two functions A and B

$$h(x) = A\cdot B$$

Its derivative would be

$$h'(x) = A'\cdot B + A\cdot B'$$

Proof

Proving this rule is relatively straightforward, first let us state the equation for the derivative:

$$\frac{d}{dx}\left[f(x)\cdot g(x)\right] = \lim_{h\to 0}\frac{f(x+h)\cdot g(x+h) - f(x)\cdot g(x)}{h}$$

We will then apply one of the oldest tricks in the book—adding a term that cancels itself out to the middle:

$$\frac{d}{dx}\left[f(x)\cdot g(x)\right] = \lim_{h\to 0}\frac{f(x+h)\cdot g(x+h) - \mathbf{f(x)\cdot g(x+h) + f(x)\cdot g(x+h)} - f(x)\cdot g(x)}{h}$$

Notice that those terms sum to zero, and so all we have done is add 0 to the equation.

Now we can split the equation up into forms that we already know how to solve:

$$\frac{d}{dx}\left[f(x)\cdot g(x)\right] = \lim_{h\to 0}\left[\frac{f(x+h)\cdot g(x+h) - f(x)\cdot g(x+h)}{h} + \frac{f(x)\cdot g(x+h) - f(x)\cdot g(x)}{h}\right]$$

Looking at this, we see that we can separate the common terms out of the numerators to get:

$$\frac{d}{dx}\left[f(x)\cdot g(x)\right] = \lim_{h\to 0}\left[g(x+h)\frac{f(x+h) - f(x)}{h} + f(x)\frac{g(x+h) - g(x)}{h}\right]$$

Which, when we take the limit, becomes:

$$\frac{d}{dx}\left[f(x)\cdot g(x)\right] = f(x)\cdot g'(x) + g(x)\cdot f'(x)$$, or the mnemonic "one D-two plus two D-one"

This can be extended to 3 functions:

$$\frac{d}{dx}[fgh] = f(x)g(x)h'(x) + f(x)g'(x)h(x) + f'(x)g(x)h(x)$$

For any number of functions, the derivative of their product is the sum, for each function, of its derivative times each other function.

Application, proof of the power rule

The product rule can be used to give a proof of the power rule for whole numbers. The proof proceeds by mathematical induction. We begin with the base case $n = 1$. If $f_1(x) = x$ then from the definition is easy to see that

$$f_1'(x) = \lim_{h \to 0} \frac{x + h - x}{h} = 1$$

Next we suppose that for fixed value of N, we know that for $f_N(x) = x^N$, $f_N'(x) = Nx^{N-1}$. Consider the derivative of $f_{N+1}(x) = x^{N+1}$,

$$f'_{N+1}(x) = (x \cdot x^N)' = (x)'x^N + x \cdot (x^N)' = x^N + x \cdot N \cdot x^{N-1} = (N+1)x^N.$$

We have shown that the statement $f_n'(x) = n \cdot x^{n-1}$ is true for $n = 1$ and that if this statement holds for $n = N$, then it also holds for $n = N + 1$. Thus by the principle of mathematical induction, the statement must hold for $n = 1, 2, \ldots$.

Quotient rule

For quotients, where one function is divided by another function, the equation is more complicated but it is simply a special case of the product rule.

$$\frac{f(x)}{g(x)} = f(x) \cdot g(x)^{-1}$$

Then we can just use the product rule and the chain rule:

$$\frac{d}{dx}\frac{f(x)}{g(x)} = f'(x) \cdot g(x)^{-1} - f(x) \cdot g'(x) \cdot g(x)^{-2}$$

We can then multiply through by 1, or more precisely: $g(x)^2 / g(x)^2$, which cancels out into 1, to get:

$$\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{f'(x) \cdot g(x)}{g(x)^2} - \frac{f(x) \cdot g'(x)}{g(x)^2}$$

This leads us to the so-called "quotient rule":

Derivatives of quotients (Quotient Rule)

$$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{f'(x)\cdot g(x) - f(x)\cdot g'(x)}{g(x)^2}$$

Which some people remember with the mnemonic "low D-high minus high D-low over the square of what's below."

Examples

The derivative of $(4x - 2) / (x^2 + 1)$ is:

$$\begin{aligned}\frac{d}{dx}\left[\frac{(4x-2)}{x^2+1}\right] &= \frac{(x^2+1)(4) - (4x-2)(2x)}{(x^2+1)^2} \\ &= \frac{(4x^2+4) - (8x^2-4x)}{(x^2+1)^2} \\ &= \frac{-4x^2+4x+4}{(x^2+1)^2}\end{aligned}$$

Remember: the derivative of a product/quotient **is not** the product/quotient of the derivatives. (That is, differentiation does not distribute over multiplication or division.) However one can distribute before taking the derivative. That is

$$\frac{d}{dx}((a+b)\times(c+d)) = \frac{d}{dx}(ac+ad+bc+bd)$$

Implicit Differentiation

Generally, one will encounter functions expressed in explicit form, that is, $y = f(x)$ form. You might encounter a function that contains a mixture of different variables. Many times it is inconvenient or even impossible to solve for y. A good example is the function $y^2 + 2yx + 3 = 5x$. It is too cumbersome to isolate y in this function. One can utilize implicit differentiation to find the derivative. To do so, consider y to be a nested function that is defined implicitly by x. You need to employ the chain rule whenever you take the derivative of a variable with respect to a different variable: i.e., $\frac{d}{dx}$(the derivative with respect to x) of x is 1; $\frac{d}{dx}$of y is $\frac{dy}{dx}$.

Remember:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

Therefore:

$$\frac{d}{dx}(y^3) = \frac{d}{dy}(y^3)\frac{dy}{dx} = 3y^2 \cdot \frac{dy}{dx}$$

Examples

$$xy = 1$$

can be solved as:

$$y = \frac{1}{x}$$

then differentiated:

$$\frac{dy}{dx} = -\frac{1}{x^2}$$

However, it can also be differentiated like this:

$$\frac{d}{dx}[xy] = \frac{d}{dx}[1]$$
$$x\frac{dy}{dx} + y = 0 \quad \text{(use the product rule)}$$
$$\frac{dy}{dx} = -\frac{y}{x} \quad \text{(solve for } \frac{dy}{dx}\text{)}$$

Note that, if we substitute $y = \frac{1}{x}$ into $\frac{dy}{dx} = -\frac{y}{x}$, we end up with $\frac{dy}{dx} = -\frac{1}{x^2}$ again.

- Find the derivative of $y^2 + x^2 = 25$ with respect to x.

You are seeking $\frac{dy}{dx}$.

Take the derivative of each side of the equation with respect to x.

$$\frac{d(y^2+x^2)}{dx} = \frac{d(25)}{dx}$$
$$2y \cdot \frac{dy}{dx} + 2x = 0$$
$$2y \cdot \frac{dy}{dx} = -2x$$
$$\frac{dy}{dx} = -\frac{x}{y}$$

Exponential, logarithmic, and trigonometric functions

Exponential

To determine the derivative of an exponent requires use of the *symmetric difference* equation for determining the derivative:

$$\frac{d}{dx}f(x) = \lim_{h \to 0} \frac{f(x+h) - f(x-h)}{2h}$$

First we will solve this for the specific case of an exponent with a base of *e* and then extend it to the general case with a base of *a* where *a* is a positive real number.

First we set up our problem using $f(x) = e^x$:

$$\frac{d}{dx}e^x = \lim_{h \to 0} \frac{e^{x+h} - e^{x-h}}{2h}$$

Then we apply some basic algebra with powers (specifically that $a^{b+c} = a^b\, a^c$):

$$\frac{d}{dx}e^x = \lim_{h \to 0} \frac{e^x e^h - e^x e^{-h}}{2h}$$

Treating e^x as a constant with respect to what we are taking the limit of, we can use the limit rules to move it to the outside, leaving us with:

$$\frac{d}{dx}e^x = e^x \cdot \lim_{h \to 0} \frac{e^h - e^{-h}}{2h}$$

A careful examination of the limit reveals a hyperbolic sine:

$$\frac{d}{dx}e^x = e^x \cdot \lim_{h \to 0} \frac{\sinh(h)}{h}$$

The limit of $\frac{\sinh(h)}{h}$ as h approaches 0 is equal to *1*, leaving us with:

Derivative of the exponential function

$$\frac{d}{dx}e^x = e^x$$

in which $f'(x) = f(x)$.

Now that we have derived a specific case, let us extend things to the general case. Assuming that a is a positive real constant, we wish to calculate:

$$\frac{d}{dx}a^x$$

One of the oldest tricks in mathematics is to break a problem down into a form that we already know we can handle. Since we have already determined the derivative of e^x, we will attempt to rewrite a^x in that form.

Using that $e^{\ln(c)} = c$ and that $\ln(a^b) = b \cdot \ln(a)$, we find that:

$$a^x = e^{x \cdot \ln(a)}$$

Thus, we simply apply the chain rule:

$$\frac{d}{dx}e^{x \cdot \ln(a)} = \left[\frac{d}{dx}x \cdot \ln(a)\right] e^{x \cdot \ln(a)}$$

In which we can solve for the derivative and substitute back with $e^{x \cdot \ln(a)} = a^x$ to get:

Derivative of the exponential function

$$\frac{d}{dx}a^x = \ln(a)\, a^x$$

Logarithms

Closely related to the exponentiation is the logarithm. Just as with exponents, we will derive the equation for a specific case first (the natural log, where the base is *e*), and then work to generalize it for any logarithm.

First let us create a variable *y* such that:

$$y = \ln(x)$$

It should be noted that what we want to find is the derivative of *y* or $\frac{dy}{dx}$.

Next we will put both sides to the power of *e* in an attempt to remove the logarithm from the right hand side:

$$e^y = x$$

Now, applying the chain rule and the property of exponents we derived earlier, we take the derivative of both sides:

$$\frac{dy}{dx} \cdot e^y = 1$$

This leaves us with the derivative:

$$\frac{dy}{dx} = \frac{1}{e^y}$$

Substituting back our original equation of $x = e^y$, we find that:

Derivative of the Natural Logarithm

$$\frac{d}{dx}\ln(x) = \frac{1}{x}$$

If we wanted, we could go through that same process again for a generalized base, but it is easier just to use properties of logs and realize that:

$$\log_b(x) = \frac{\ln(x)}{\ln(b)}$$

Since 1 / ln(*b*) is a constant, we can just take it outside of the derivative:

$$\frac{d}{dx}\log_b(x) = \frac{1}{\ln(b)} \cdot \frac{d}{dx}\ln(x)$$

which leaves us with the generalized form of:

Derivative of the Logarithm

$$\frac{d}{dx}\log_b(x) = \frac{1}{x\ln(b)}$$

Trigonometric Functions

Sine, Cosine, Tangent, Cosecant, Secant, Cotangent: These are functions that crop up continuously in mathematics and engineering and have a lot of practical applications. They also appear in more advanced mathematics, particularly when dealing with things such as line integrals with complex numbers and alternate representations of space like spherical and cylindrical coordinate systems.

We use the definition of the derivative, i.e.,

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

to work these first two out.

Let us find the derivative of sin *x*, using the above definition.

$f(x) = \sin x$

$f'(x) = \lim_{h\to 0}\frac{\sin(x+h) - \sin x}{h}$ Definition of derivative

$= \lim_{h\to 0}\frac{\cos(x)\sin(h) + \cos(h)\sin(x) - sin(x)}{h}$ trigonometric identity

$$= \lim_{h \to 0} \frac{\cos(x)\sin(h) + (\cos(h) - 1)\sin(x)}{h}$$ factoring

$$= \lim_{h \to 0} \frac{\cos(x)\sin(h)}{h} + \lim_{h \to 0} \frac{(\cos(h) - 1)\sin(x)}{h}$$
separation of terms

$$= \cos x \times 1 + \sin x \times 0$$ application of limit

$$= \cos x$$ solution

Now for the case of *cos x*

$$f(x) = \cos x$$

$$f'(x) = \lim_{h \to 0} \frac{\cos(x + h) - \cos x}{h}$$ Definition of derivative

$$= \lim_{h \to 0} \frac{\cos(x)\cos(h) - \sin(h)\sin(x) - cos(x)}{h}$$
trigonometric identity

$$= \lim_{h \to 0} \frac{\cos(x)(\cos(h) - 1) - \sin(x)\sin(h)}{h}$$ factoring

$$= \lim_{h \to 0} \frac{\cos(x)(\cos(h) - 1)}{h} - \lim_{h \to 0} \frac{\sin(x)\sin(h)}{h}$$
separation of terms

$= \cos x \times 0 - \sin x \times 1$application of limit

$= -\sin x$solution

Therefore we have established

Derivative of Sine and Cosine

$$\frac{d}{dx}\sin(x) = \cos(x)$$
$$\frac{d}{dx}\cos(x) = -\sin(x)$$

To find the derivative of the tangent, we just remember that:

$$\tan(x) = \frac{\sin(x)}{\cos(x)}$$

which is a quotient. Applying the quotient rule, we get:

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

Then, remembering that $\cos^2(x) + \sin^2(x) = 1$, we simplify:

$$\begin{aligned}\frac{\cos^2(x) + \sin^2(x)}{\cos^2(x)} &= \frac{1}{\cos^2(x)} \\ &= \sec^2(x)\end{aligned}$$

Derivative of the Tangent

$$\frac{d}{dx}\tan(x) = \sec^2(x)$$

For secants, we just need to apply the chain rule to the derivations we have already determined.

$$\sec(x) = \frac{1}{\cos(x)}$$

So for the secant, we state the equation as:

$$\sec(x) = \frac{1}{u}$$

$u(x) = \cos(x)$

Take the derivative of both equations, we find:

$$\frac{d}{dx}\sec(x) = \frac{-1}{u^2} \cdot \frac{du}{dx}$$

$$\frac{du}{dx} = -\sin(x)$$

Leaving us with:

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

Simplifying, we get:

Derivative of the Secant

$$\frac{d}{dx}\sec(x) = \sec(x)\tan(x)$$

Using the same procedure on cosecants:

$$\csc(x) = \frac{1}{\sin(x)}$$

We get:

Derivative of the Cosecant

$$\frac{d}{dx}\csc(x) = -\csc(x)\cot(x)$$

Using the same procedure for the cotangent that we used for the tangent, we get:

Derivative of the Cotangent

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

Inverse Trigonometric Functions

Arcsine, arccosine, arctangent: These are the functions that allow you to determine the angle given the sine, cosine, or tangent of that angle.

First, let us start with the arcsine such that:

$y = \arcsin(x)$

To find dy/dx we first need to break this down into a form we can work with:

$x = \sin(y)$

Then we can take the derivative of that:

$$1 = \cos(y) \cdot \frac{dy}{dx}$$

...and solve for dy / dx:

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

At this point we need to go back to the unit triangle. Since *y* is the angle and the opposite side is sin(*y*) (which is equal to *x*), the adjacent side is cos(*y*) (which is equal to the square root of 1 minus x^2, based on the Pythagorean theorem), and the hypotenuse is 1. Since we have determined the value of cos(*y*) based on the unit triangle, we can substitute it back in to the above equation and get:

Derivative of the Arcsine

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

We can use an identical procedure for the arccosine and arctangent:

Derivative of the Arccosine

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

Derivative of the Arctangent

$$\frac{d}{dx}\arctan(x) = \frac{1}{1+x^2}$$

Exercises

By using the above rules, practice differentiation on the following.

1. $\frac{d}{dx}[(x^3+5)^{10}]$
2. $\frac{d}{dx}[x^3+3x]$
3. $\frac{d}{dx}[(x+4)\cdot(x+2)\cdot(x-3)]$
4. $\frac{d}{dx}[\frac{x+1}{3x^2}]$
5. $\frac{d}{dx}[3\cdot x^3]$

6. $\frac{d}{dx}[sinx \cdot x^4]$
7. $\frac{d}{dx}[2^x]$
8. $\frac{d}{dx}[e^{x^2}]$
9. $\frac{d}{dx}[e^{2^x}]$
10. $\frac{d}{dx}[x^x]$

Applications of Derivatives

Newton's Method

Newton's Method (also called the Newton-Raphson method) is a recursive algorithm for approximating the root of a differentiable function. We know simple formulas for finding the roots of linear and quadratic equations, and there are also more complicated formulae for cubic and quartic equations. At one time it was hoped that there would be formulas found for equations of quintic and higher-degree, though it was later shown by Neils Henrik Abel that no such equations exist. The Newton-Raphson method is a method for approximating the roots of polynomial equations of any order. In fact the method works for any equation, polynomial or not, as long as the function is differentiable in a desired interval.

Newton's Method

Let $f(x)$ be a differentiable function. Select a point x_1 based on a first approximation to the root, arbitrarily close to the function's root. To approximate the root you then recursively calculate using:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

As you recursively calculate, the x_n's become increasingly better approximations of the function's root.

For n number of approximations,

$$x_n = x_0 - \sum_{i=0}^{n} \frac{f(x_i)}{f'(x_i)}$$

Examples

Find the root of the function $f(x) = x^2$.

$$x_1 = f(2) = 4$$
$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 2$$
$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} = 1$$
$$x_4 = x_3 - \frac{f(x_3)}{f'(x_3)} = \frac{1}{2}$$
$$x_5 = x_4 - \frac{f(x_4)}{f'(x_4)} = \frac{1}{4}$$
$$x_6 = x_5 - \frac{f(x_5)}{f'(x_5)} = \frac{1}{8}$$
$$x_7 = x_6 - \frac{f(x_6)}{f'(x_6)} = \frac{1}{16}$$
$$x_8 = x_7 - \frac{f(x_7)}{f'(x_7)} = \frac{1}{32}$$

As you can see x_n is gradually approaching zero (which we know is the root of *f(x)*). One can approach the function's root with arbitrary accuracy.

Answer: $f(x) = x^2$ has a root at $x = 0$.

Notes

This method fails when $f'(x) = 0$. In that case, one should choose a new starting place. Occasionally it may happen that $f(x) = 0$ and $f'(x) = 0$ have a common root. To detect whether this is true, we should first find the solutions of $f'(x) = 0$, and then check the value of $f(x)$ at these places.

Newton's method also may not converge for every function, take as an example:

$$f(x) = \begin{cases} \sqrt{x-r}, & \text{for} x \geq r \\ -\sqrt{r-x}, & \text{for} x \leq r \end{cases}$$

For this function choosing any $x_1 = r - h$ then $x_2 = r + h$ would cause successive approximations to alternate back and forth, so no amount of iteration would get us any closer to the root than our first guess.

Related Rates

Process for solving related rates problems:

- Write out any relevant formulas and information.
- Take the derivative of the primary equation with respect to time.
- Solve for the desired variable.
- Plug-in known information and simplify.

As stated, when doing related rates, you generate a function which compares the rate of change of one value with respect to change in time. For example, velocity is the rate of change of distance over time. Likewise, acceleration is the rate of change of velocity over time. Therefore, for the variables for distance, velocity, and acceleration, respectively x, v, and a, and time, t:

$$v = \frac{dx}{dt}$$
$$a = \frac{dv}{dt}$$

Using derivatives, you can find the functions for velocity and acceleration from the distance function. This is the basic idea behind related rates: the rate of change of a function is the derivative of that function with respect to time.

Common Applications

Filling Tank

This is the easiest variant of the most common textbook related rates problem: the filling water tank.

- The tank is a cube, with volume 1000L.
- You have to fill the tank in ten minutes or you die.
- You want to escape with your life and as much money as possible, so you want to find the smallest pump that can finish the task.

We need a pump that will fill the tank 1000L in ten minutes. So, for pump rate p, volume of water pumped v, and minutes t:

$$p = \frac{dv}{dt}$$

Examples

Related rates can get complicated very easily.

Example 1:

A cone with a circular base is being filled with water. Find a formula which will find the rate with which water is pumped.

- Write out any relevant formulas or pieces of information.

$$V = \frac{1}{3}\pi r^2 h$$

- Take the derivative of the equation above with respect to time. Remember to use the Chain Rule and the Product Rule.

$$V = \frac{1}{3}\pi r^2 h$$
$$\frac{dV}{dt} = \frac{\pi}{3}\left(r^2 \cdot \frac{dh}{dt} + 2rh \cdot \frac{dr}{dt}\right)$$

Answer: $$\frac{dV}{dt} = \frac{\pi}{3}\left(r^2 \cdot \frac{dh}{dt} + 2rh \cdot \frac{dr}{dt}\right)$$

Example 2:

A spherical hot air balloon is being filled with air. The volume is changing at a rate of 2 cubic feet per minute.
How is the radius changing with respect to time when the radius is equal to 2 feet?

- Write out any relevant formulas and pieces of information.

$$V_{sphere} = \frac{4}{3}\pi r^3$$

$$\frac{dV}{dt} = 2$$
$$r = 2$$

- Take the derivative of both sides of the volume equation with respect to time.

$$V = \frac{4}{3}\pi r^3$$
$$\frac{dV}{dt} = \frac{4}{3} \cdot 3 \cdot \pi r^2 \cdot \frac{dr}{dt}$$
$$= 4\pi r^2 \cdot \frac{dr}{dt}$$

- Solve for $\frac{dr}{dt}$.

$$\frac{dr}{dt} = \frac{1}{4\pi r^2} \cdot \frac{dV}{dt}$$

- Plug-in known information.

$$\frac{dr}{dt} = \frac{1}{16\pi} \cdot 2$$
$$\frac{dr}{dt} = \frac{1}{8\pi}$$

Answer: $\frac{dr}{dt}$ ft/min.

Example 3:

An airplane is attempting to drop a box onto a house. The house is 300 feet away in horizontal distance and 400 feet in vertical distance. The rate of change of the horizontal distance with respect to time is the same as the rate of change of the vertical distance with respect to time. How is the distance between the box and the house changing with respect to time at the moment? The rate of change in the horizontal direction with respect to time is -50 feet per second.

Note: Because the vertical distance is downward in nature, the rate of change of y is negative. Similarly, the horizontal distance is decreasing, therefore it is negative (it is getting closer and closer).

The easiest way to describe the horizontal and vertical relationships of the plane's motion is the Pythagorean Theorem.

- Write out any relevant formulas and pieces of information.

$x^2 + y^2 = s^2$ (where s is the distance between the plane and the house)
$x = 300$
$y = 400$
$s = \sqrt{x^2 + y^2} = \sqrt{300^2 + 400^2} = 500$
$\frac{dx}{dt} = \frac{dy}{dt} = -50$

- Take the derivative of both sides of the distance formula with respect to time.

$$x^2 + y^2 = s^2$$
$$2x \cdot \frac{dx}{dt} + 2y \cdot \frac{dy}{dt} = 2s \cdot \frac{ds}{dt}$$

- Solve for $\frac{ds}{dt}$.

$$\frac{ds}{dt} = \frac{1}{2s}\left(2x \cdot \frac{dx}{dt} + 2y \cdot \frac{dy}{dt}\right)$$

- Plug-in known information

$$\frac{ds}{dt} = \frac{1}{2(500)}\left[2(300) \cdot (-50) + 2(400) \cdot (-50)\right]$$
$$= \frac{1}{1000}(-70000)$$
$$= -70 \text{ft/s}$$

Answer: $\frac{ds}{dt} = -70$ ft/sec.

Example 4:

Sand falls onto a cone shaped pile at a rate of 10 cubic feet per minute. The radius of the pile's base is always 1/2 of its altitude. When the pile is 5 ft deep, how fast is the altitude of the pile increasing?

- Write down any relevant formulas and information.

$$V = \frac{1}{3}\pi r^2 h$$
$$\frac{dV}{dt} = 10$$
$$r = \frac{1}{2}h$$
$$h = 5$$

Substitute $r = \frac{1}{2}h$ into the volume equation.

$$V = \frac{1}{3}\pi r^2 h$$
$$= \frac{1}{3}\pi h \cdot \left(\frac{h^2}{4}\right)$$
$$= \frac{1}{12}\pi h^3$$

- Take the derivative of the volume equation with respect to time.

$$V = \frac{1}{12}\pi h^3$$
$$\frac{dV}{dt} = \frac{1}{4}\pi h^2 \cdot \frac{dh}{dt}$$

- Solve for $\frac{dh}{dt}$.

$$\frac{dh}{dt} = \frac{4}{\pi h^2} \cdot \frac{dV}{dt}$$

- Plug-in known information and simplify.

$$\frac{dh}{dt} = \frac{4}{\pi(5)^2} \cdot 10$$
$$= \frac{8}{5\pi} \text{ ft/min}$$

Answer: $\frac{dh}{dt} = \frac{8}{5\pi}$ ft/min.

Example 5:

A 10 ft long ladder is leaning against a vertical wall. The foot of the ladder is being pulled away from the wall at a constant rate of 2 ft/sec. When the ladder is exactly 8 ft from the wall, how fast is the top of the ladder sliding down the wall?

- Write out any relevant formulas and information.

Use the Pythagorean Theorem to describe the motion of the ladder.

$x^2 + y^2 = l^2$ (where l is the length of the ladder)
$l = 10$
$\frac{dx}{dt} = 2$
$x = 8$
$y = \sqrt{l^2 - x^2} = \sqrt{100 - 64} = \sqrt{36} = 6$

- Take the derivative of the equation with respect to time.

$2x \cdot \frac{dx}{dt} + 2y \cdot \frac{dy}{dt} = 0$ (l^2 so $\frac{dl}{dt} = 0$.)

- Solve for $\frac{dy}{dt}$.

$2x \cdot \frac{dx}{dt} + 2y \cdot \frac{dy}{dt} = 0$
$2y \cdot \frac{dy}{dt} = -2x \cdot \frac{dx}{dt}$
$\frac{dy}{dt} = -\frac{x}{y} \cdot \frac{dx}{dt}$

- Plug-in known information and simplify.

$\frac{dy}{dt} = \left(-\frac{8}{6}\right)(-2)$

$= \frac{8}{3}$ ft/sec

Answer: $\frac{dy}{dt} = \frac{8}{3}$ ft/sec.

Exercises

Problem Set

Here's a few problems for you to try:

1. A spherical balloon is inflated at a rate of **100 ft^3/min**. Assuming the rate of inflation remains constant, how fast is the radius of the balloon increasing at the instant the radius is **4 ft**?
2. Water is pumped from a cone shaped reservoir (the vertex is pointed down) **10 ft** in diameter and **10 ft** deep at a constant rate of **3 ft^3/min**. How fast is the water level falling when the depth of the water is **6 ft**?
3. A boat is pulled into a dock via a rope with one end attached to the bow of a boat and the other end held by a man standing **6 ft** above the bow of the boat. If the man pulls the rope at a constant rate of **2 ft/sec**, how fast is the boat moving toward the dock when **10 ft** of rope is out?

Solution Set

1. $\frac{25}{16\pi} \frac{ft}{min}$
2. $\frac{1}{3\pi} \frac{ft}{min}$
3. $\frac{5}{2} \frac{ft}{sec}$

Kinematics

Kinematics or the study of motion is a very relevant topic in calculus.

This section uses the following conventions:

- $x(t)$ represents the position equation
- $v(t)$ represents the velocity equation
- $a(t)$ represents the acceleration equation

Differentiation

Average Velocity and Acceleration

Average velocity and acceleration problems use the algebraic definitions of velocity and acceleration.

- $v_{avg} = \frac{\Delta x}{\Delta t}$
- $a_{avg} = \frac{\Delta v}{\Delta t}$

Examples

Example 1:

A particle's position is defined by the equation $x(t) = t^3 - 2t^2 + t$. Find the
average velocity over the interval *[2,7]*.

- Find the average velocity over the interval *[2,7]*:

$$\begin{aligned} v_{avg} &= \frac{x(7) - x(2)}{7 - 2} \\ &= \frac{252 - 2}{5} \\ &= 50 \end{aligned}$$

Answer: $v_{avg} = 50$.

Instantaneous Velocity and Acceleration

Instantaneous velocity and acceleration problems use the derivative definitions of velocity and acceleration.

- $v(t) = \frac{dx}{dt}$
- $a(t) = \frac{dv}{dt}$

Examples

Example 2:

A particle moves along a path with a position that can be determined by the function $x(t) = 4t^3 + e^t$.
Determine the acceleration when $t = 3$.

- Find $v(t) = \frac{ds}{dt}$.

  $$\frac{ds}{dt} = 12t^2 + e^t$$

- Find $a(t) = \frac{dv}{dt} = \frac{d^2s}{dt^2}$.

  $$\frac{d^2s}{dt^2} = 24t + e^t$$

- Find $a(3) = \frac{d^2s}{dt^2}|_{t=3}$

  $$\begin{aligned}\frac{d^2s}{dt^2}|_{t=3} &= 24(3) + e^3 \\ &= 72 + e^3 \\ &= 92.08553692...\end{aligned}$$

Answer: $a(3) = 92.08553692...$

Integration

- $x_2 - x_1 = \int_{t_1}^{t_2} v(t)dt$
- $v_2 - v_1 = \int_{t_1}^{t_2} a(t)dt$

Optimization

Optimization is the use of Calculus in the real world. Let us assume we are a pizza parlor and wish to maximize profit. Perhaps we have a flat piece of cardboard and we need to make a box with the greatest volume. How does one go about this process?

Obviously, this requires the use of maximums and minimums. We know that we find maximums and minimums via derivatives. Therefore, one can conclude that Calculus will be a useful tool for maximizing or minimizing (also known as "Optimizing") a situation.

Examples

Example 1:

A box manufacturer desires to create a box with a surface area of 100 inches squared.
What is the maximum size volume that can be formed by bending this material into a box?
The box is to be closed. The box is to have a square base, square top, and rectangular sides.

- Write out known formulas and information

$$A_{base} = x^2$$
$$A_{side} = x \cdot h$$
$$A_{total} = 2x^2 + 4x \cdot h = 100$$
$$V = l \cdot w \cdot h = x^2 \cdot h$$

- Eliminate the variable *h* in the volume equation

$$2x^2 + 4xh = 100$$
$$x^2 + 2xh = 50$$
$$2xh = 50 - x^2$$
$$h = \frac{50 - x^2}{2x}$$
$$V = (x^2)\left(\frac{50 - x^2}{2x}\right)$$
$$= \frac{1}{2}(50x - x^3)$$

- Find the derivative of the volume equation in order to maximize the volume

$$\frac{dV}{dx} = \frac{1}{2}(50 - 3x^2)$$

- Set $\frac{dV}{dx} = 0$ and solve for x

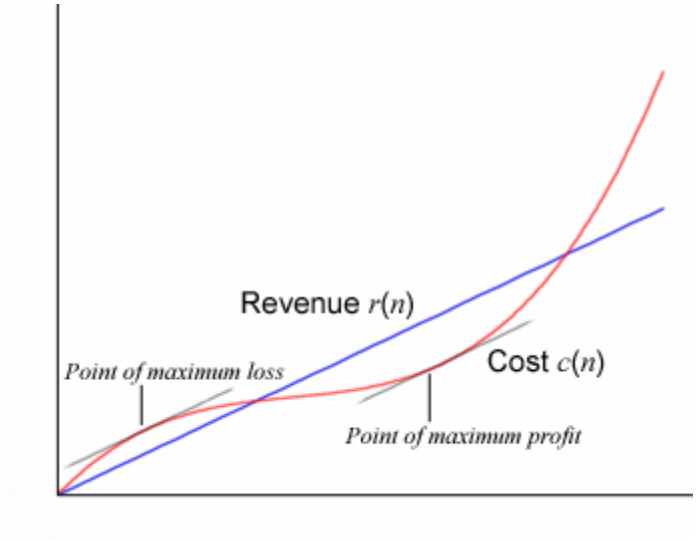
$$\frac{1}{2}(50 - 3x^2) = 0$$
$$50 - 3x^2 = 0$$
$$3x^2 = 50$$
$$x = \pm\frac{\sqrt{50}}{\sqrt{3}}$$

- Plug-in the *x* value into the volume equation and simplify

$$V = \frac{1}{2}\left[50 \cdot \sqrt{\frac{50}{3}} - \left(\sqrt{\frac{50}{3}}\right)^3\right]$$
$$= 68.04138174..$$

Answer: $V_{max} = 68.04138174..$

Sales Example



A small retailer can sell n units of a product for a revenue of r(n)=8.1n and at a cost of $c(n)=n^3-7n^2+18n$, with all amounts in thousands. How many units does it sell to maximize its profit?

The retailer's profit is defined by the equation p(n)=r(n) - c(n), which is the revenue generated less the cost. The question asks for the maximum amount of profit which is the maximum of the above equation. As previously discussed, the maxima and minima of a

graph are found when the slope of said graph is equal to zero. To find the slope one finds the derivative of p(n). By using the subtraction rule p'(n)=r'(n) - c'(n):

$$\begin{aligned} p(n) &= r(n) - c(n) \\ p'(n) &= \frac{d}{dn}\left[8.1n\right] - \frac{d}{dn}\left[n^3 - 7n^2 + 18n\right] \\ &= -3n^2 + 14n - 9.9 \end{aligned}$$

Therefore, when $-3n^2 + 14n - 9.9 = 0$ the profit will be maximized or minimized. Use the quadratic formula to find the roots, giving {3.798,0.869}. To find which of these is the maximum and minimum the function can be tested:

p(0.869) = - 3.97321, p(3.798) = 8.58802

Because we only consider the functions for all n >= 0 (i.e. you can't have n=-5 units), the only points that can be minima or maxima are those two listed above. To show that 3.798 is in fact a maximum (and that the function doesn't remain constant past this point) check if the sign of p'(n) changes at this point. It does, and for n greater than 3.798 P'(n) the value will remain decreasing. Finally, this shows that for this retailer selling 3,798 units would return a profit of $8,588.02.

Integration

Definition of the Integral

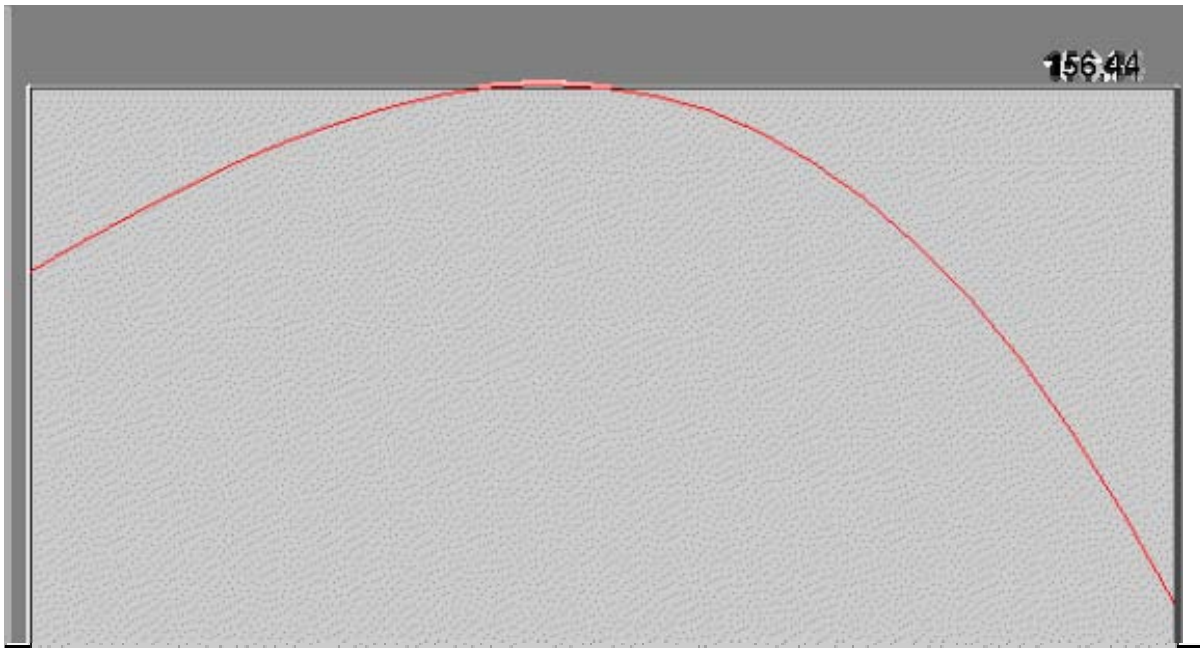


Figure 1

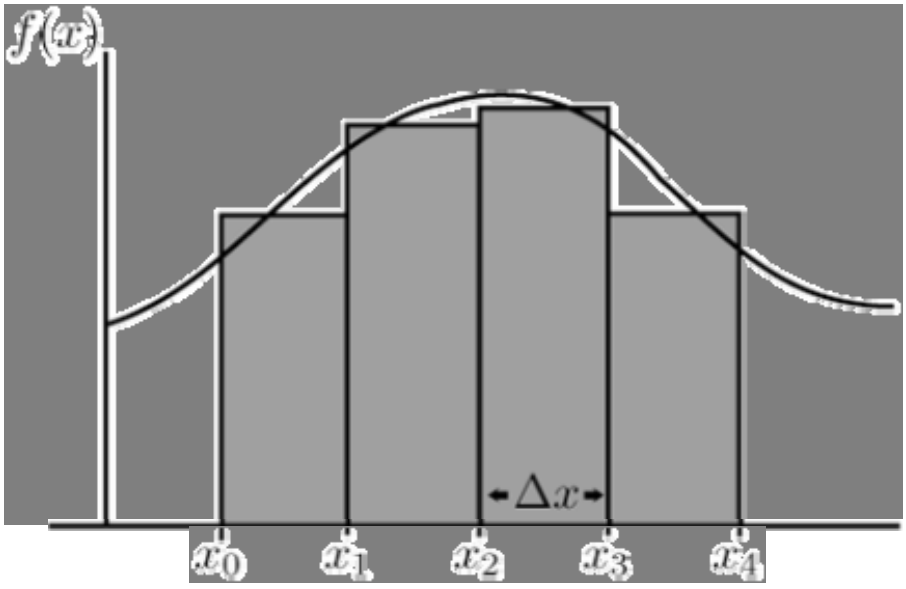


Figure 2

The rough idea of defining the area under the graph of *f* is to approximate this area with a finite number of rectangles. Since we can easily work out the area of the rectangles we get an estimate of the area under the graph. If we use a larger number of rectangles we expect a better approximation, and the limit as we approach an infinite number of rectangles will give the exact area.

Suppose first that *f* is positive and *a*<*b*. We pick an integer *n* and divide the interval *[a,b]* into n subintervals of equal width (see Figure 2). As the interval *[a,b]* has width *b-a* each subinterval has width $\Delta x = \dfrac{b-a}{n}$. We denote the endpoints of the subintervals by $x_0, x_1, \ldots, x_n$ so

$$x_i = a + i\Delta x \text{ for } i = 0, 1, \ldots, n.$$

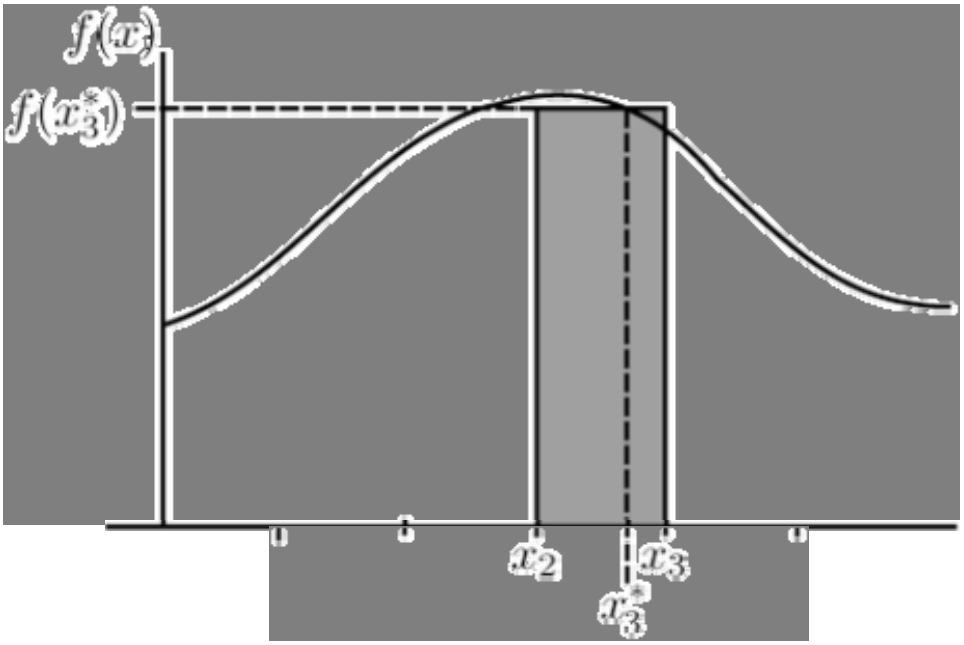


Figure 3

Now for each $i = 1, \ldots, n$ pick a *sample point* x_i^* in the interval $[x_{i-1}, x_i]$ and consider the rectangle of height $f(x_i^*)$ and width Δ*x* (see Figure 3). The area of this rectangle is $f(x_i^*)\Delta x$. By adding up the area of all the rectangles for $i = 1, \ldots, n$ we get that the area *S* is approximated by

$$A_n = f(x_1^*)\Delta x + f(x_2^*)\Delta x + \cdots + f(x_n^*)\Delta x.$$

A more convenient way to write this is with the summation notation as

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

For each number *n* we get a different approximation. As *n* gets larger the width of the rectangles gets smaller which yields a better approximation (see Figures 4 and 5). In the limit as A_n as *n* tends to infinity we get the area of *S*.

Definition of the Definite Integral Suppose *f* is a continuous function on *[a,b]* and $\Delta x = \frac{b-a}{n}$. Then the *definite integral* of *f* between *a* and *b* is

$$\int_a^b f(x)\ dx = \lim_{n\to\infty} A_n = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\Delta x.$$

where x_i^* are any sample points in the interval $[x_{i-1},x_i]$.

It is a fact that if *f* is continuous on *[a,b]* then this limit always exists and does not depend on the choice of the points $x_i^* \in [x_{i-1}, x_i]$. For instance they may be evenly spaced, or distributed ambiguously throughout the interval. The proof of this is technical and is beyond the scope of this section.

Notation When considering the expression $\int_a^b f(x)\ dx$ the function *f* is called the *integrand* and the interval *[a,b]* is the interval of integration. Also *a* is called the *lower limit* and *b* the *upper limit* of integration.

One important feature of this definition is that we also allow functions which take negative values. If *f(x)<0* for all *x* then $f(x_i^*) < 0$ so $f(x_i^*)\Delta x < 0$. So the definite integral of *f* will be strictly negative. More generally if *f* takes on both positive an negative values then $\int_a^b f(x)dx$ will be the area under the positive part of the graph of *f* **minus** the area under the graph of the negative part of the graph (see Figure 6). For this reason we say that $\int_a^b f(x)dx$ *is the* ***signed area*** *under the graph.*

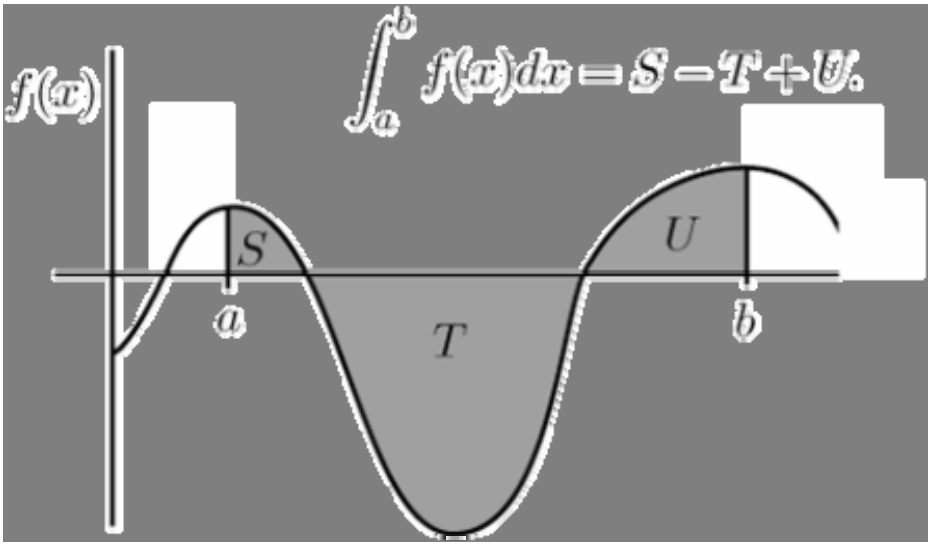
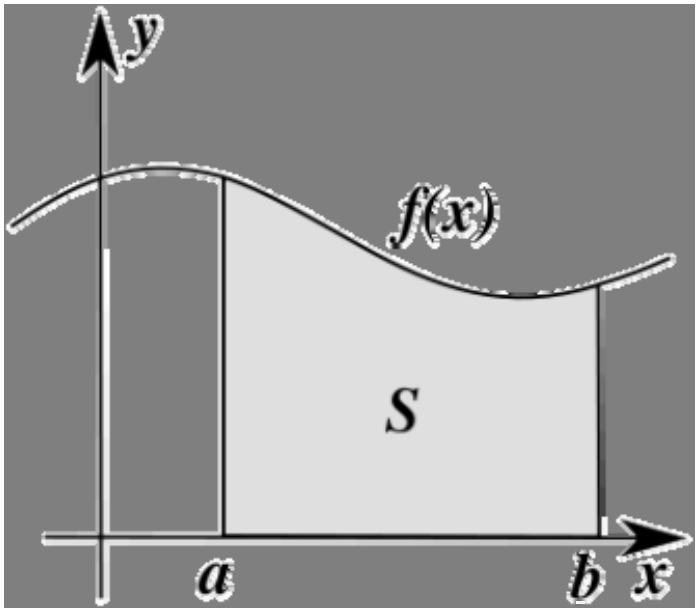
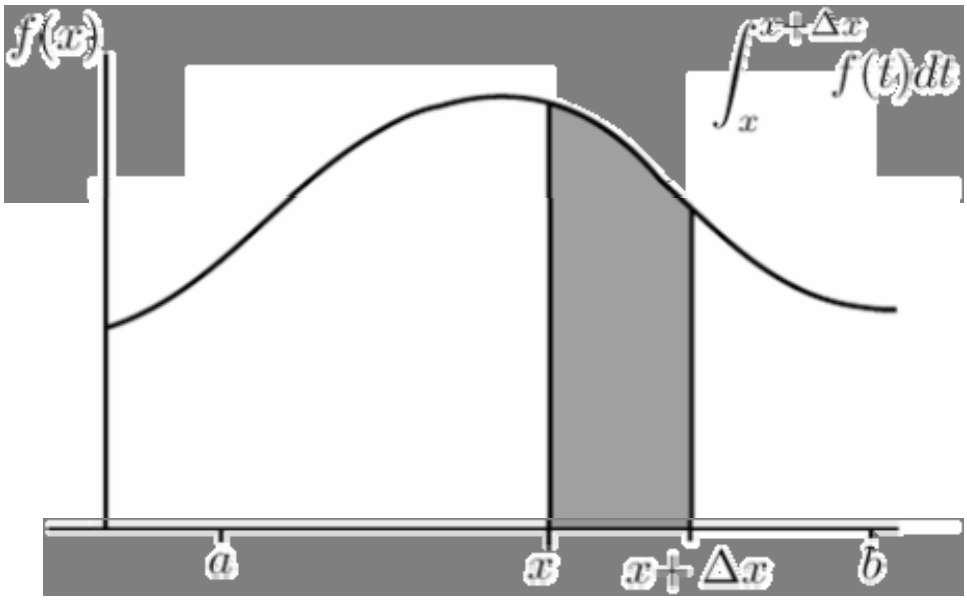


Figure 6



Figure



The area spanned by $f(x)$

A geometrical proof that anti-derivative gives the area

Suppose we have a function F(x) which returns the area between x and some unknown point u. (Actually, u is the first number before x which satisfies F(u) = 0, but our solution is independent from u, so we won't bother ourselves with it.) We don't even know if something like F exists or not, but we're going to investigate what clue do we have if it *does* exist.

We can use F to calculate the area between a and b, for instance, which is obviously F(b)-F(a); F is something general. Now, consider a rather peculiar situation, the area bounded at x and $x + \Delta x$, in the limit of $\Delta x \to 0$. Of course it can be calculated by using F, but we're looking for another solution this time. As the right border approaches the left one, the shape seems to be an infinitesimal rectangle, with the height of f(x) and width of Δx. So, the area reads:

$$\text{Infinitesimal area} = \lim_{\Delta x \to 0} f(x)\Delta x$$

Of course, we could use F to calculate this area as well:

$$\text{Infinitesimal area} = \lim_{\Delta x \to 0} F(x + \Delta x) - F(x)$$

By combining these equations, we have

$$\lim_{\Delta x \to 0} F(x + \Delta x) - F(x) = \lim_{\Delta x \to 0} f(x)\Delta x$$

If we divide both sides by Δx, we get

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

which is an interesting result, because the left-hand side is the derivative of F with respect to x. This remarkable result doesn't tell us what F itself is, however it tells us what the *derivative of F* is, and it is *f*.

Independence of Variable

It is important to notice that the variable *x* did not play an important role in the definition of the integral. In fact we can replace it with any other letter, so the following are all equal: $\int_a^b f(x)dx = \int_a^b f(t)dt = \int_a^b f(u)du = \int_a^b f(w)dw.$ Each of these is the signed area under the graph of *f* between *a* and *b*.

Left and Right Handed Riemann Sums

These methods are sometimes referred to as L-RAM and R-RAM, RAM standing for "Rectangular Approximation Method."

We could have decided to choose all our sample points x_i^* to be on the right hand side of the interval $[x_{i-1}, x_i]$ (see Figure 7). Then $x_i^* = x_i$ for all *i* and the approximation that we called A_n for the area becomes

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

This is called the *right-handed Riemann sum*, and the integral is the limit

$$\int_a^b f(x)\, dx = \lim_{n \to \infty} A_n = \lim_{n \to \infty} \sum_{i=1}^{n} f(x_i)\Delta x.$$

Alternatively we could have taken each sample point on the left hand side of the interval. In this case $x_i^* = x_{i-1}$ (see Figure 8) and the approximation becomes

$$A_n = \sum_{i=1}^{n} f(x_{i-1})\Delta x.$$

Then the integral of *f* is

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} A_n = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_{i-1})\Delta x.$$

The key point is that, as long as *f* is continuous, these two definitions give the same answer for the integral.

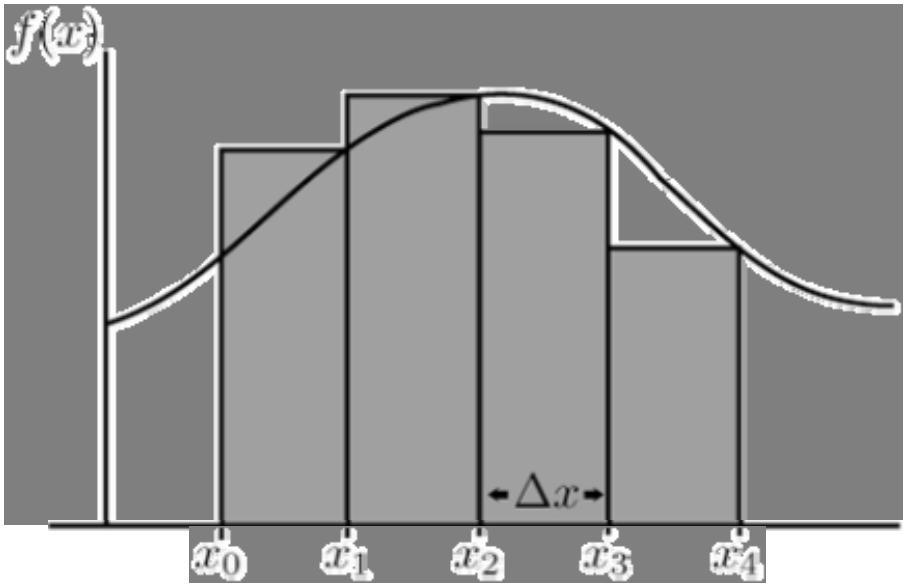


Figure 7

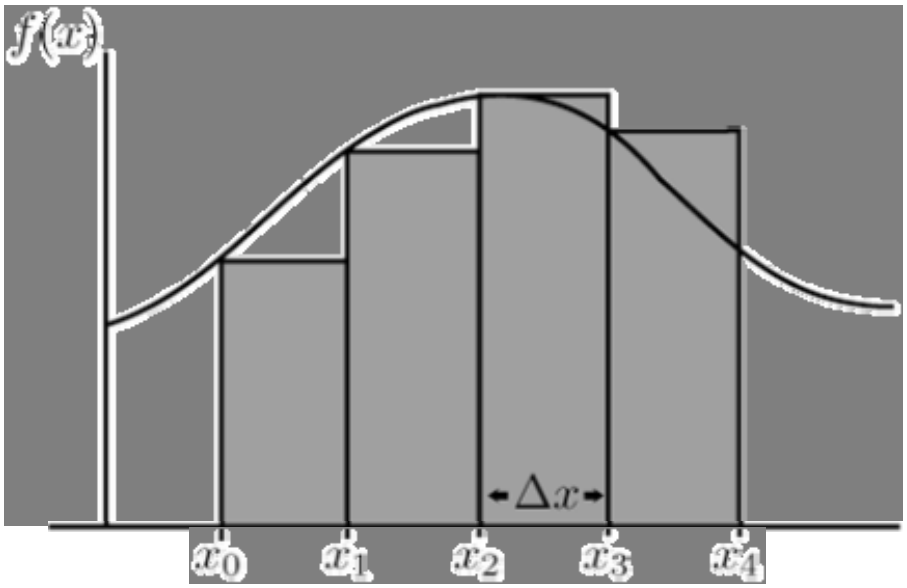


Figure 8

Example 1

In this example we will calculate the area under the curve given by the graph of $f(x) = x$ for x between 0 and 1. First we fix an integer n and divide the interval [0,1] into n subintervals of equal width. So each subinterval has width

$$\Delta x = \frac{1}{n}.$$

To calculate the integral we will use the right-handed Riemann Sum. (We could have used the left-handed sum instead, and this would give the same answer in the end). For the right-handed sum the sample points are

$$x_i^* = 0 + i\Delta x = \frac{i}{n} \quad i = 1, \ldots, n$$

Notice that $f(x_i^*) = x_i^* = \frac{i}{n}$. Putting this into the formula for the approximation,

$$A_n = \sum_{i=1}^{n} f(x_i^*)\Delta x = \sum_{i=1}^{n} f(i/n)\Delta x = \sum_{i=1}^{n} \frac{i}{n} \cdot \frac{1}{n} = \frac{1}{n^2} \sum_{i=1}^{n} i.$$

Now we use the formula

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

to get

$$A_n = \frac{1}{n^2}\frac{n(n+1)}{2} = \frac{n(n+1)}{2n^2}.$$

To calculate the integral of *f* between *0* and *1* we take the limit as *n* tends to infinity,

$$\int_0^1 f(x)dx = \lim_{n\to\infty} \frac{n(n+1)}{2n^2} = \frac{1}{2}.$$

Example 2

Next we show how to find the integral of the function $f(x) = x^2$ between $x=a$ and $x=b$. This time the interval *[a,b]* has width *b-a* so

$$\Delta x = \frac{b-a}{n}.$$

Once again we will use the right-handed Riemann Sum. So the sample points we choose are

$$x_i^* = a + i\Delta x = a + \frac{i(b-a)}{n}.$$

Thus

$$\begin{aligned} A_n &= \sum_{i=1}^{n} f(x_i^*)\Delta x \\ &= \sum_{i=1}^{n} f\left(a + \frac{(b-a)i}{n}\right)\Delta x \\ &= \frac{b-a}{n}\sum_{i=1}^{n}\left(a + \frac{(b-a)i}{n}\right)^2 \\ &= \frac{b-a}{n}\sum_{i=1}^{n}\left(a^2 + \frac{2a(b-a)i}{n} + \frac{(b-a)^2 i^2}{n^2}\right) \end{aligned}$$

We have to calculate each piece on the right hand side of this equation. For the first two,

$$\sum_{i=1}^{n} a^2 = a^2 \sum_{i=1}^{n} 1 = na^2$$
$$\sum_{i=1}^{n} \frac{2a(b-a)i}{n} = \frac{2a(b-a)}{n} \sum_{i=1}^{n} i = \frac{2a(b-a)}{n} \cdot \frac{n(n+1)}{2}.$$

For the third sum we have to use a formula

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

to get

$$\sum_{i=1}^{n} \frac{(b-a)^2 i^2}{n^2} = \frac{(b-a)^2}{n^2} \frac{n(n+1)(2n+1)}{6}.$$

Putting this together

$$A_n = \frac{b-a}{n} \left(na^2 + \frac{2a(b-a)}{n} \cdot \frac{n(n+1)}{2} + \frac{(b-a)^2}{n^2} \frac{n(n+1)(2n+1)}{6} \right).$$

Taking the limit as *n* tend to infinity gives

$$\begin{aligned} \int_a^b x^2 dx &= (b-a)\left(a^2 + a(b-a) + \frac{1}{3}(b-a)^2\right) \\ &= (b-a)\left(a^2 + ab - a^2 + \frac{1}{3}(b^2 - 2ab + a^2)\right) \\ &= \frac{1}{3}(b-a)(b^2 + ab + a^2) \\ &= \frac{1}{3}(b^3 - a^3). \end{aligned}$$

Basic Properties of the Integral

The Constant Rule

From the definition of the integral we can deduce some basic properties. We suppose that *f* and *g* are continuous on *[a,b]*.

Integrating Constants

If *c* is constant then $\int_a^b c\,dx = c\,(b-a).$

When $c > 0$ and $a < b$ this integral is the area of a rectangle of height *c* and width *b-a* which equals *c(b-a)*.

Example

$$\int_1^3 9dx = 9(3-1) = 9*2 = 18.$$
$$\int_{-2}^6 11dx = 11(6-(-2)) = 11*8 = 88.$$
$$\int_2^{17} 0dx = 0*(17-2) = 0.$$

Constant Rule

$$\int_a^b cf(x)dx = c\int_a^b f(x)dx.$$

When *f* is positive, the height of the function *cf* at a point *x* is *c* times the height of the function *f*. So the area under *cf* between *a* and *b* is *c* times the area under *f*. We can also give a proof using the definition of the integral, using the constant rule for limits,

$$\int_a^b c\,f(x)dx = \lim_{n\to\infty}\sum_{i=1}^n cf(x_i^*) = c\lim_{n\to\infty}\sum_{i=1}^n f(x_i^*) = c\int_a^b f(x)dx.$$

Example We saw in the previous section that

$$\int_0^1 xdx = \frac{1}{2}.$$

Using the constant rule we can use this to calculate that

$$\int_0^1 3xdx = 3\int_0^1 xdx = 3.\frac{1}{2} = \frac{3}{2},$$
$$\int_0^1 -7xdx = -7\int_0^1 xdx = (-7).\frac{1}{2} = -\frac{7}{2}.$$

Example We saw in the previous section that

$\int_a^b x^2 dx = \frac{1}{3}(b^3 - a^3)$. We can use this and the constant rule to calculate that

$$\int_1^3 2x^2 dx = 2\int_1^3 x^2 dx = 2.\frac{1}{3}.(3^3 - 1^3) = \frac{2}{3}(27-1) = \frac{52}{3}.$$

The addition and subtraction rule

Addition and Subtraction Rules of Integration

$$\int_a^b (f(x) + g(x))dx = \int_a^b f(x)dx + \int_a^b g(x)dx.$$
$$\int_a^b (f(x) - g(x))dx = \int_a^b f(x)dx - \int_a^b g(x)dx.$$

As with the constant rule, the addition rule follows from the addition rule for limits:

$$\begin{aligned}\int_a^b (f(x) + g(x))dx &= \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*) + g(x_i^*) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*) + \lim_{n\to\infty} \sum_{i=1}^{n} g(x_i^*) \\ &= \int_a^b f(x)dx + \int_a^b g(x)dx.\end{aligned}$$

The subtraction rule can be proved in a similar way.

Example From above $\int_1^3 9dx = 18$ and $\int_1^3 2x^2 dx = \frac{52}{3}$ so

$$\int_1^3 (2x^2 + 9)dx = \int_1^3 2x^2 dx + \int_1^3 9dx = \frac{52}{3} + 18 = \frac{106}{3},$$
$$\int_1^3 (2x^2 - 9)dx = \int_1^3 2x^2 dx - \int_1^3 9dx = \frac{52}{3} - 18 = -\frac{2}{3}.$$

Example

$$\int_0^2 4x^2 + 14dx = 4\int_0^2 x^2 dx + \int_0^2 14dx = 4\cdot\frac{1}{3}(2^3 - 0^3) + 2\cdot 14 = \frac{32}{3} + 28 = \frac{116}{3}.$$

The Comparison Rule

Comparison Rule

- Suppose $f(x) \geq 0$ for all *x* in *[a,b]*. Then

$$\int_a^b f(x)dx \geq 0.$$

- Suppose $f(x) \geq g(x)$ for all *x* in *[a,b]*. Then

$$\int_a^b f(x)dx \geq \int_a^b g(x)dx.$$

- Suppose $M \geq f(x) \geq m$ for all *x* in *[a,b]*. Then

$$M(b-a) \geq \int_a^b f(x)dx \geq m(b-a).$$

If $f(x) \geq 0$ then each of the rectangles in the Riemann sum to calculate the integral of *f* will be above the *y* axis, so the area will be non-negative. If $f(x) \geq g(x)$ then $f(x) - g(x) \geq 0$ and by linearity of the integral we get the second property. Finally if $M \geq f(x) \geq m$ then the area under the graph of *f* will be greater than the area of rectangle with height *m* and less than the area of the rectangle with height *M* (see Figure 9). So

$$M(b-a) = \int_a^b M \geq \int_a^b f(x)dx \geq \int_a^b m = m(b-a).$$

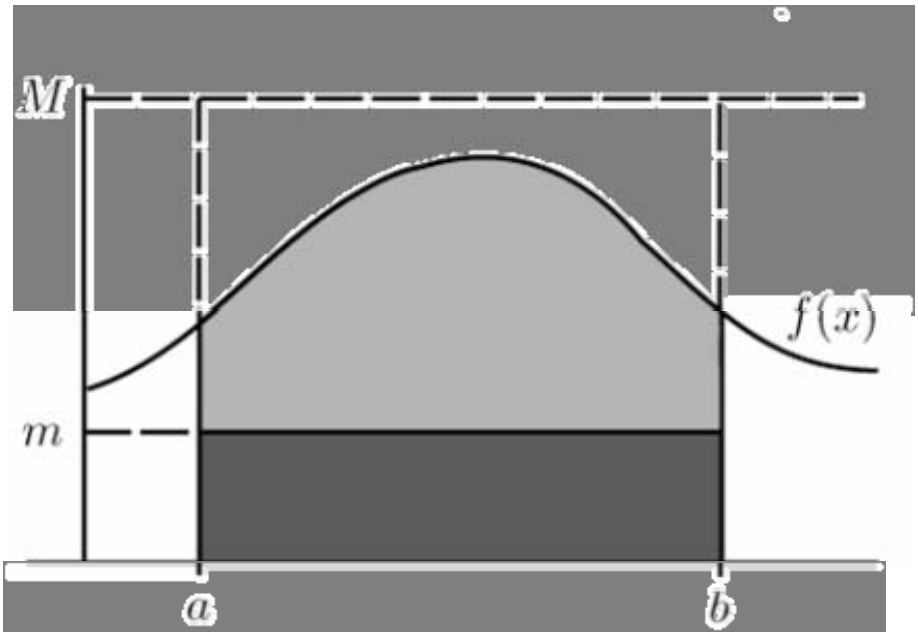


Figure 9

Linearity with respect to endpoints

Additivity with respect to endpoints Suppose $a < c < b$. Then

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx.$$

Again suppose that *f* is positive. Then this property should be interpreted as saying that the area under the graph of *f* between *a* and *b* is the area between *a* and *c* plus the area between *c* and *b* (see Figure 8)

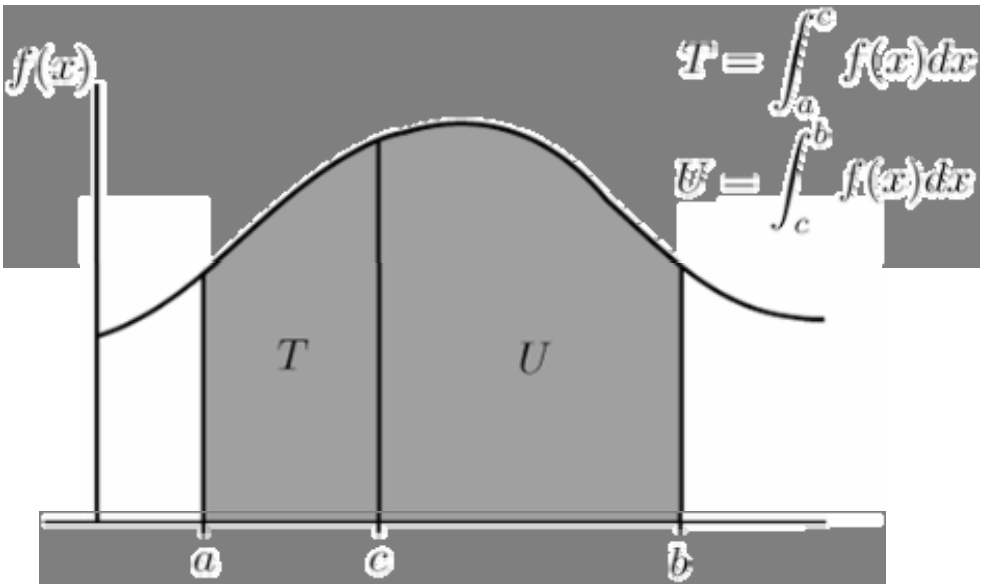


Figure 8

Extension of Additivity with respect to limits of integration

When $a = b$ we have that $\Delta x = \frac{b-a}{n} = 0$ so

$$\int_a^a f(x)dx = 0.$$

Also in defining the integral we assumed that $a<b$. But the definition makes sense even when $b<a$ in which case $\Delta x = \frac{1}{n}(b-a)$ so has changed sign. This gives

$$\int_b^a f(x)dx = -\int_a^b f(x)dx.$$

With these definitions,

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$$

whatever the order of a,b,c.

Fundamental Theorem of Calculus

Statement of the Fundamental Theorem

Suppose that f is continuous on $[a,b]$. We can define a function F by

$$F(x) = \int_a^x f(t)\ dt \text{ for } x \text{ in } [a, b].$$

Fundamental Theorem of Calculus Part I Suppose *f* is continuous on *[a,b]* and *F* is defined by

$$F(x) = \int_a^x f(t)\ dt.$$

Then *F* is differentiable on *(a,b)* and for all $x \in (a, b)$,

$$F'(x) = f(x).$$

Now recall that *F* is said to be an antiderivative of *f* if $F'(x) = f(x)$.

Fundamental Theorem of Calculus Part II Suppose that *f* is continuous on *[a,b]* and that *F* is any antiderivative of *f*. Then

$$\int_a^b f(x)\ dx = F(b) - F(a).$$

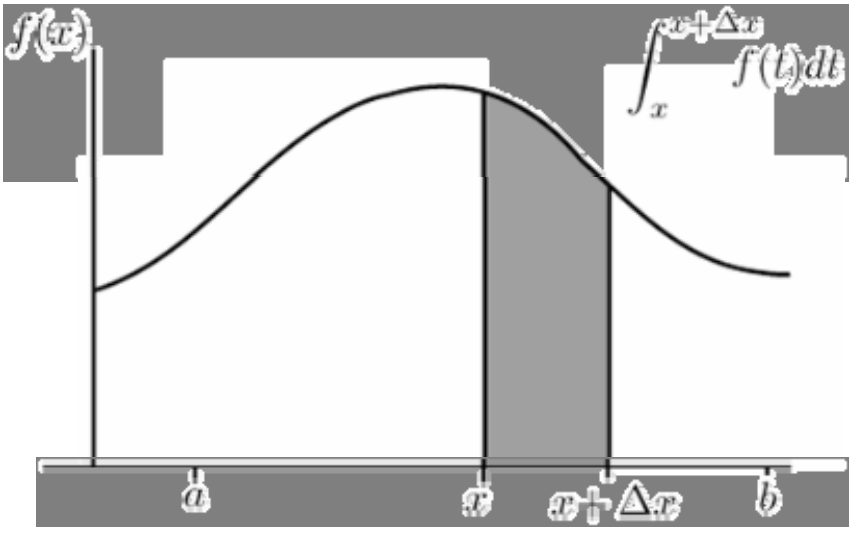


Figure 1

Note: a minority of mathematicians refer to part one as two and part two as one. All mathematicians refer to what is stated here as part 2 as The Fundamental Theorem of Calculus.

Proofs

Proof of Fundamental Theorem of Calculus Part I

Suppose x is in (a,b). Pick Δx so that $x + \Delta x$ is also in (a, b). Then

$$F(x) = \int_a^x f(t)dt$$

and

$$F(x + \Delta x) = \int_a^{x+\Delta x} f(t)dt.$$

Subtracting the two equations gives

$$F(x + \Delta x) - F(x) = \int_a^{x+\Delta x} f(t)dt - \int_a^x f(t)dt.$$

Now

$$\int_a^{x+\Delta x} f(t)dt = \int_a^x f(t)dt + \int_x^{x+\Delta x} f(t)dt$$

so rearranging this we have

$$F(x + \Delta x) - F(x) = \int_x^{x+\Delta x} f(t)dt.$$

According to the mean value theorem for integration, there exists a c in $[x, x + \Delta x]$ such that

$$\int_x^{x+\Delta x} f(t)dt = f(c)\Delta x.$$

Notice that c depends on Δx. Anyway what we have shown is that

$$F(x + \Delta x) - F(x) = f(c)\Delta x,$$

and dividing both sides by Δx gives

$$\frac{F(x + \Delta x) - F(x)}{\Delta x} = f(c).$$

Take the limit as $\Delta x \to 0$ we get the definition of the derivative of F at x so we have

$$F'(x) = \lim_{\Delta x \to 0} f(c).$$

To find the other limit, we will use the squeeze theorem. The number c is in the interval $[x, x + \Delta x]$, so $x \leq c \leq x + \Delta x$. Also, $\lim_{\Delta x \to 0} x = x$ and $\lim_{\Delta x \to 0} x + \Delta x = x$. Therefore, according to the squeeze theorem,

$$\lim_{\Delta x \to 0} c = x$$

As f is continuous we have

$$F'(x) = \lim_{\Delta x \to 0} f(c) = f\left(\lim_{\Delta x \to 0} c\right) = f(x)$$

which completes the proof.

Proof of Fundamental Theorem of Calculus Part II

Define $P(x) = \int_a^x f(t)dt.$ Then by the Fundamental Theorem of Calculus part I we know that P is differentiable on (a,b) and for all $x \in (a, b)$

$$P'(x) = f(x).$$

So P is an antiderivative of f. Now we were assuming that F was also an antiderivative so for all $x \in (a, b)$,

$$P'(x) = F'(x) = f(x).$$

A consequence of the Mean Value Theorem is that this implies there is a constant C such that for all $x \in (a, b)$,

$$P(x) = F(x) + C,$$

and as P and F are continuous we see this holds when $x=a$ and when $x=b$ as well. Since we know that $P(a)=0$ we can put $x=a$ into the equation to get $0=F(a) +C$ so $C=-F(a)$. And putting $x=b$ gives

$$\int_a^b f(t)dx = P(b) = F(b) + C = F(b) - F(a).$$

Integration of Polynomials

Using the power rule for differentiation we can find a formula for the integral of a power using the Fundamental Theorem of Calculus. Let $f(x) = x^n$. We want to find an antiderivative for f. Since the differentiation rule for powers lowers the power by 1, we have that

$$\frac{d}{dx}x^{n+1} = (n+1)x^n.$$

As long as $n+1 \neq 0$ we can divide by $n+1$ to get

$$\frac{d}{dx}\left(\frac{x^{n+1}}{n+1}\right) = x^n = f(x).$$

So the function $F(x) = \frac{x^{n+1}}{n+1}$ is an antiderivative of f. If $a,b>0$ then F is continuous on $[a,b]$ and we can apply the Fundamental Theorem of Calculus we can calculate the integral of f to get the following rule.

Power Rule of Integration I

$$\int_a^b x^n dx = \left[\frac{x^{n+1}}{n+1}\right]_a^b = \frac{b^{n+1} - a^{n+1}}{n+1}.$$ as long as $n \neq -1$ and $a,b > 0$.

Notice that we allow all values of n, even negative or fractional. If $n>0$ then this works even if a or b are negative.

Power Rule of Integration II

$$\int_a^b x^n dx = \left[\frac{x^{n+1}}{n+1}\right]_a^b = \frac{b^{n+1} - a^{n+1}}{n+1}.$$ as long as $n > 0$.

Example

To find $\int_1^2 x^3 dx$ we raise the power by 1 and have to divide by 4. So

$$\int_1^2 x^3 dx = \left[\frac{x^4}{4}\right]_1^2 = \frac{2^4}{4} - \frac{1^4}{4} = \frac{15}{4}.$$

Example

The power rule also works for negative powers. For instance

$$\int_1^3 \frac{1}{x^3}dx = \int_1^3 x^{-3}dx = \left[\frac{x^{-2}}{-2}\right]_1^3 = \frac{1}{-2}\left(3^{-2}-1^{-2}\right) = -\frac{1}{2}\left(\frac{1}{3^2}-1\right) = -\frac{1}{2}\left(\frac{1}{9}-1\right) = \frac{1}{2}\cdot\frac{8}{9} = \frac{4}{9}.$$

Example

We can also use the power rule for fractional powers. For instance

$$\int_0^5 \sqrt{x}dx = \int_0^5 x^{\frac{1}{2}}dx = \left[\frac{x^{\frac{3}{2}}}{\frac{3}{2}}dx\right]_0^5 = \frac{2}{3}\left(5^{\frac{3}{2}}-0^{\frac{3}{2}}\right) = \frac{2}{3}\left(5^{\frac{3}{2}}\right)$$

Example

Using the linearity rule we can now integrate any polynomial. For example

$$\int_0^3 (3x^2+4x+2)dx = \left[x^3+2x^2+2x\right]_0^3 = 3^3+2\cdot 3^2+2\cdot 3-0 = 27+18+6 = 51.$$

Indefinite Integrals

The Fundamental Theorem of Calculus tells us that if *f* is continuous then the function $F(x) = \int_a^x f(t)dt$ is an antiderivative of *f* (i.e. *F*'(*x*) = *f*(*x*). However it is not the only antiderivative. We can add any constant to *F* without changing the derivative.

We write $F(x) = \int f(x)dx + C$ if the derivative of *F* is $F'(x) = f(x)$.

Example

Since the derivative of x^4 is $4x^3$ the general antiderivative of $4x^3$ is x^4 plus a constant. Thus

$$\int 4x^3dx = x^4 + C.$$

Example: Finding antiderivatives

Let us return to the previous example, that of $6x^2$. How would we go about finding the integral of this function? Recall the rule from differentiation that

$$\frac{d}{dx}x^n = nx^{n-1}$$

In our circumstance, we have:

$$\frac{d}{dx}x^3 = 3x^2$$

This is a start! We now know that the function we seek will have a power of 3 in it. How would we get the constant of 6? Well,

$$2\frac{d}{dx}x^3 = 2 \times 3x^2 = 6x^2.$$

Thus, we say that $2x^3$ is an antiderivative of $6x^2$.

Basic Properties of Indefinite Integrals

Constant Rule for indefinite integrals

If *c* is constant then $\int cf(x)dx = c\int f(x)dx.$

Sum/Difference Rule for indefinite integrals

$$\int f(x) + g(x)dx = \int f(x)dx + \int g(x)dx.$$
$$\int f(x) - g(x)dx = \int f(x)dx - \int g(x)dx.$$

Indefinite integrals of Polynomials

Since

$$\frac{d}{dx}\frac{1}{n+1}x^{n+1} = x^n$$

we have the following rule for indefinite integrals.

Power rule for indefinite integrals If $n \neq -1$, then

$$\int x^n dx = \frac{1}{n+1}x^{n+1} + C.$$

Integral of the Inverse function

Since

$$\frac{d}{dx}\ln x = \frac{1}{x}$$

We know that

$$\int \frac{dx}{x} = \ln|x| + C$$

Note that the polynomial integration rule does not apply when the exponent is -1. This technique of integration must be used instead. Since the argument of the natural logarithm function must be positive (on the real line), the absolute value signs are added around its argument to ensure that the argument is positive.

Integral of Sine and Cosine

In this section we will concern ourselves with determining the integrals of the *sin* and *cosine* function.

Recall that

$$\frac{d}{dx}\sin x = \cos x$$
$$\frac{d}{dx}\cos x = -\sin x.$$

So *sin x* is an antiderivative of *cos x* and -*cos x* is an antiderivative of *sin x*. Hence we get the following rules for integrating *sin x* and *cos x*.

$$\int \cos x \; dx = \sin x + C$$
$$\int \sin x \; dx = -\cos x + C$$

We will find how to integrate more complicated trigonometric functions in the chapter on Further integration techniques.

Integral of the Exponential function

Since

$$\frac{d}{dx}e^x = e^x$$

we see that e^x is its own antiderivative. Perhaps a more useful definition of this rule can be given as:

$$\frac{d}{dx}e^{f(x)} = f'(x) \cdot e^{f(x)}$$

hence:

$$\frac{d}{dx}e^{x} = (1)e^{x}$$

Where the exponent (x) is differentiated to give a value of 1

Simplified: $\frac{d}{dx}e^{x} = (1)e^{x}$

Becomes: $\frac{d}{dx}e^{x} = e^{x}$

So the integral of an exponential function can be found thusly:

$$\int e^{x}dx = e^{x} + C$$

Integration Rules

$$\int ndx = nx + C$$
$$\int x^{n}dx = \frac{x^{n+1}}{n+1} + C, n \neq -1$$
$$\int e^{x}dx = e^{x} + C$$
$$\int a^{x}dx = \frac{a^{x}}{\ln(a)} + C$$
$$\int \frac{1}{x}dx = \ln|x| + C, x > 0$$
$$\int \sin x dx = -\cos x + C$$
$$\int \cos x dx = \sin x + C$$
$$\int \tan x dx = -\ln|\cos x| + C$$
$$\int \csc x dx = -\ln|\csc x + \cot x| + C$$

$$\int \sec x dx = \ln|\sec x + \tan x| + C$$
$$\int \cot x dx = \ln|\sin x| + C$$
$$\int \sec^2 x dx = \tan x + C$$
$$\int \csc^2 x dx = -\cot x + C$$
$$\int \sec x \tan x dx = \sec x + C$$
$$\int \csc x \cot x dx = -\csc x + C$$
$$\int \frac{1}{\sqrt{a^2 - b^2x^2}} dx = \frac{1}{b} \arcsin \frac{bx}{a} + C$$
$$\int \frac{1}{a^2 + b^2x^2} dx = \frac{1}{ab} \arctan \frac{bx}{a} + C$$
$$\int \frac{1}{x\sqrt{x^2 - a^2}} dx = \frac{1}{a} \operatorname{arcsec} \frac{|x|}{a} + C$$

The Substitution Rule

Suppose that we want to find $\int (\cos(x^2) \cdot x) dx.$

The Fundamental theorem of calculus tells us that we want to find an antiderivative of the function:

$$f(x) = \cos(x^2) \cdot x$$

Since sin(x) differentiates to cos(x) as a first guess we might try the function sin(x^2). But by the Chain Rule

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

which is almost what we wanted apart from the fact that there is an extra factor of 2 in front. But this is easily dealt with because we can divide by any constant so

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

So using the Fundamental Theorem of Calculus, $\int x \cos(x^2) dx = \frac{\sin(x^2)}{2} + C.$

In fact this technique will work for more general integrands. Suppose *u* is a differentiable function. Then to evaluate $\int u'(x)cos(u(x))dx$ we just have to notice that by the Chain Rule

$$\frac{d}{dx}\sin(u(x)) = \cos(u(x))\frac{du}{dx} = u'(x)\cos(u(x)).$$

As long as *u*' is continuous the Fundamental Theorem applies and tells us that

$$\int \cos(u(x))u'(x)dx = \sin(u(x)) + C.$$

Now the right hand side of this equation is just the integral of cos(*u*) but with respect to *u*. If we write *u* instead of *u(x)* this becomes

$$\int \cos(u(x))u'(x)dx = \sin(u)du + C = \int \cos(u)du.$$

So for instance if $u(x) = x^3$ we have worked out that

$$\int (\cos(x^3) \cdot 3x^2)dx = \sin(x^3) + C.$$

Now there was nothing special about using the cosine function in the discussion above, and it could be replaced by any other function. Doing this gives us the substitution rule for indefinite integrals.

Substitution rule for indefinite integrals Assume *u* is differentiable with continuous derivative and that *f* is continuous on the range of *u*. Then

$$\int f(u(x))\frac{du}{dx}dx = \int f(u)du.$$

Notice that it looks like you can *cancel* in the expression $\frac{du}{dx}dx$ to leave just a *du*. This does not really make any sense as $\frac{du}{dx}$ is **not a fraction**, but is a good way to remember the substitution rule.

There is a similar rule for definite integrals, but we have to change the endpoints.

Substitution rule for definite integrals Assume *u* is differentiable with continuous derivative and that *f* is continuous on the range of *u*. Suppose $c = u(a)$ and $d = u(b)$. Then

$$\int_a^b f(u(x))\frac{du}{dx}dx = \int_c^d f(u)du.$$

Examples

Consider the integral

$$\int_0^2 x\cos(x^2+1)\,dx$$

By using the substitution $u = x^2 + 1$, we obtain $du = 2x\,dx$ and

$$\begin{aligned}\int_0^2 x\cos(x^2+1)\,dx &= \frac{1}{2}\int_0^2 \cos(x^2+1)2x\,dx \\ &= \frac{1}{2}\int_1^5 \cos(u)\,du \\ &= \frac{1}{2}(\sin(5)-\sin(1)).\end{aligned}$$

Note how the lower limit $x = 0$ was transformed into $u = 0^2 + 1 = 1$ and the upper limit $x = 2$ into $u = 2^2 + 1 = 5$.

Proof of the substitution rule

We will now prove the substitution rule for definite integrals. Let F be an anti derivative of f so

$F'(x) = f(x)$. By the Fundamental Theorem of Calculus

$$\int_c^d f(u)du = F(d) - F(c).$$

Next we define a function G by the rule

$$G(x) = F(u(x))\,.$$

Then by the Chain rule G is differentiable with derivative

$$G'(x) = F'(u(x))u'(x) = f(u(x))u'(x)\,.$$

Integrating both sides with respect to x and using the Fundamental Theorem of Calculus we get

$$\int_a^b f(u(x))u'(x)dx = \int_a^b G'(x)dx = G(b) - G(a).$$

But by the definition of F this equals

$$G(b) - G(a) = F(u(b)) - F(u(a)) = F(d) - F(c) = \int_c^d f(u)du.$$

Hence

$$\int_a^b f(u(x))u'(x)dx = \int_c^d f(u)du.$$

which is the substitution rule for definite integrals.

Integration of even and odd functions

Recall that a function f is called odd if it satisfies $f(-x) = -f(x)$ and is called even if $f(-x) = f(x)$.

Suppose f is a continuous odd function then for any a,

$$\int_{-a}^a f(x)dx = 0.$$

If f is a continuous even function then for any a,

$$\int_{-a}^a f(x)dx = 2\int_0^a f(x)dx.$$

Caution: For improper integrals (e.g. if a *is infinity, or if the function approaches infinity at 0 or* a*, etc.), the first equation above is only true if* $\int_0^a f(x)dx$ *exists. Otherwise the integral is undefined, and only the Cauchy principal value is 0.*

Suppose f is an odd function and consider first just the integral from $-a$ to 0. We make the substitution $u=-x$ so $du=-dx$. Notice that if $x=-a$ then $u=a$ and if $x=0$ then $u=0$. Hence

$$\int_{-a}^0 f(x)dx = -\int_a^0 f(-u)du = \int_0^a f(-u)du.$$

Now as f is odd, $f(-u) = -f(u)$ so the integral becomes

$$\int_{-a}^0 f(x)dx = -\int_0^a f(u)du.$$

Now we can replace the dummy variable u with any other variable. So we can replace it with the letter x to give

$$\int_{-a}^0 f(x)dx = -\int_0^a f(u)du = -\int_0^a f(x)dx.$$

Now we split the integral into two pieces

$$\int_{-a}^a f(x)dx = \int_{-a}^0 f(x)dx + \int_0^a f(x)dx = -\int_0^a f(x)dx + \int_0^a f(x)dx = 0.$$

The proof of the formula for even functions is similar, and is left as an exercise.

Integration by Parts

Integration by parts for indefinite integrals Suppose *f* and *g* are differentiable and their derivatives are continuous. Then

$$\int f(x)g'(x)dx = f(x)g(x) - \int f'(x)g(x)dx.$$

If we write *u=f(x)* and *v=g(x)* then using the Leibnitz notation *du=f'(x) dx* and *dv=g'(x) dx* and the integration by parts rule becomes

$$\int udv = uv - \int vdu.$$

For definite integrals the rule is essentially the same, as long as we keep the endpoints.

Integration by parts for definite integrals Suppose *f* and *g* are differentiable and their derivatives are continuous. Then

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

This can also be expressed in Leibniz notation.

$$\int_a^b udv = \Big[uv\Big]_a^b - \int_a^b vdu.$$

Example Find

$$\int x\cos(x)\,dx$$

Here we let:

$u = x$, so that $du = dx$,
$dv = \cos(x)dx$, so that $v = \sin(x)$.

Then:

$$\begin{aligned}\int x\cos(x)\,dx &= \int u\,dv \\ &= uv - \int v\,du \\ \int x\cos(x)\,dx &= x\sin(x) - \int \sin(x)\,dx \\ \int x\cos(x)\,dx &= x\sin(x) + \cos(x) + C\end{aligned}$$

where C is an arbitrary constant of integration.

Example

$$\int x^2 e^x\,dx$$

In this example we will have to use integration by parts twice.

Here we let

$u = x^2$, so that $du = 2xdx$,
$dv = e^x dx$, so that $v = e^x$.

Then:

$$\begin{aligned}\int x^2 e^x\,dx &= \int u\,dv \\ &= uv - \int v\,du \\ \int x^2 e^x\,dx &= x^2 e^x - \int 2xe^x\,dx = x^2 e^x - 2\int xe^x\,dx.\end{aligned}$$

Now to calculate the last integral we use integration by parts again. Let

$u =$ "x, so that $du = dx$,
$dv = e^x dx$, so that $v = e^x$

and integrating by parts gives

$$\int xe^x\,dx = xe^x - \int e^x\,dx = xe^x - e^x.$$

So in the end

$$\int x^2 e^x \, dx = x^2 e^x - 2(xe^x - e^x) = x^2 e^x - 2xe^x + 2e^x = e^x(x^2 - 2x + 2).$$

Example Find

$$\int \ln(x) \, dx.$$

The trick here is to write this integral as

$$\int \ln(x) \cdot 1 \, dx.$$

Now let

$u = ln(x)$ so $du = 1 / x dx$,
$v = x$ so $dv = 1 dx$.

Then using integration by parts,

$$\begin{aligned} \int \ln(x) \, dx &= x \ln(x) - \int \frac{x}{x} \, dx \\ &= x \ln(x) - \int 1 \, dx \\ \int \ln(x) \, dx &= x \ln(x) - x + C \\ \int \ln(x) \, dx &= x(\ln(x) - 1) + C \end{aligned}$$

where, again, C is an arbitrary constant.

Example Find $\int \arctan(x) dx.$

Again the trick here is to write the integrand as $\arctan(x) = \arctan(x) \cdot 1$. Then let

$u = \arctan(x)$; $du = 1/(1+x^2)\, dx$
$v = x$; $dv = 1 \cdot dx$

so using integration by parts,

$$\int \arctan(x) \, dx = x \arctan(x) - \int \frac{x}{1 + x^2} \, dx$$

$$= x \arctan(x) - \frac{1}{2} \ln\left(1 + x^2\right) + C.$$

Example Find : $\int e^x \cos(x)\, dx$ This example uses integration by parts twice. First let,

$u = e^x$; thus $du = e^x dx$
$dv = \cos(x)dx$; thus $v = \sin(x)$

so

$$\int e^x \cos(x)\, dx = e^x \sin(x) - \int e^x \sin(x)\, dx$$

Now, to evaluate the remaining integral, we use integration by parts again, with

$u = e^x$; $du = e^x dx$
$v = -\cos(x)$; $dv = \sin(x)dx$

Then

$$\begin{aligned} \int e^x \sin(x)\, dx &= -e^x \cos(x) - \int -e^x \cos(x)\, dx \\ &= -e^x \cos(x) + \int e^x \cos(x)\, dx \end{aligned}$$

Putting these together, we get

$$\int e^x \cos(x)\, dx = e^x \sin(x) + e^x \cos(x) - \int e^x \cos(x)\, dx$$

Notice that the same integral shows up on both sides of this equation. So we can simply add the integral to both sides to get:

$$\begin{aligned} 2\int e^x \cos(x)\, dx &= e^x(\sin(x) + \cos(x)) \\ \int e^x \cos(x)\, dx &= \frac{e^x(\sin(x) + \cos(x))}{2} \end{aligned}$$

Integration techniques-

Infinite Sums

The most basic, and arguably the most difficult, type of evaluation is to use the formal definition of a Riemann integral.

Exact Integrals as Limits of Sums

Using the definition of an integral, we can evaluate the limit as n goes to infinity. This technique requires a fairly high degree of familiarity with summation identities. This technique is often referred to as evaluation "by definition," and can be used to find definite integrals, as long as the integrands are fairly simple. We start with definition of the integral:

$$\int_a^b f(x)\ dx = \lim_{n\to\infty} \frac{b-a}{n} \sum_{i=1}^{n} f(x_i^*)$$

Then picking x_i^* to be $x_i = a + i\frac{b-a}{n}$ we get,

$$= \lim_{n\to\infty} \frac{b-a}{n} \sum_{i=1}^{n} f(a + i\frac{b-a}{n}).$$

In some simple cases, this expression can be reduced to a real number, which can be interpreted as the area under the curve if f(x) is positive on [a,b].

Example 1

Find $\int_0^2 x^2\ dx$ by writing the integral as a limit of Riemann sums.

$$\begin{aligned}
\int_0^2 x^2\ dx &= \lim_{n\to\infty} \frac{b-a}{n} \sum_{i=1}^{n} f(x_i^*) \\
&= \lim_{n\to\infty} \frac{2}{n} \sum_{i=1}^{n} f(\frac{2i}{n}) \\
&= \lim_{n\to\infty} \frac{2}{n} \sum_{i=1}^{n} \left(\frac{2i}{n}\right)^2 \\
&= \lim_{n\to\infty} \frac{2}{n} \sum_{i=1}^{n} \frac{4i^2}{n^2}
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \frac{8}{n^3} \sum_{i=1}^{n} i^2 \\
&= \lim_{n\to\infty} \frac{8}{n^3} \frac{n(n+1)(2n+1)}{6} \\
&= \lim_{n\to\infty} \frac{4}{3} \frac{2n^2+3n+1}{n^2} \\
&= \lim_{n\to\infty} \frac{8}{3} + \frac{4}{n} + \frac{4}{3n^2} \\
&= \frac{8}{3}
\end{aligned}
$$

In other cases, it is even possible to evaluate indefinite using the formal definition. We can define the indefinite integral as follows:

$$
\begin{aligned}
\int f(x)\, dx = \int_0^x f(t)\, dt &= \lim_{n\to\infty} \frac{x-0}{n} \sum_{i=1}^{n} f(t_i^*) \\
&= \lim_{n\to\infty} \frac{x}{n} \sum_{i=1}^{n} f\left(0 + \frac{(x-0)\cdot i}{n}\right) \\
&= \lim_{n\to\infty} \frac{x}{n} \sum_{i=1}^{n} f\left(\frac{x\cdot i}{n}\right)
\end{aligned}
$$

Example 2

Suppose $f(x) = x^2$, then we can evaluate the indefinite integral as follows.

$$
\begin{aligned}
\int_0^x f(t)\, dt &= \lim_{n\to\infty} \frac{x}{n} \sum_{i=1}^{n} f\left(\frac{x\cdot i}{n}\right) \\
&\lim_{n\to\infty} \frac{x}{n} \sum_{i=1}^{n} \left(\frac{x\cdot i}{n}\right)^2 \\
&\lim_{n\to\infty} \frac{x}{n} \sum_{i=1}^{n} \frac{x^2\cdot i^2}{n^2} \\
&\lim_{n\to\infty} \frac{x^3}{n^3} \sum_{i=1}^{n} i^2 \\
&\lim_{n\to\infty} \frac{x^3}{n^3} \sum_{i=1}^{n} i^2
\end{aligned}
$$

$$\lim_{n\to\infty} \frac{x^3}{n^3}\frac{n(n+1)(2n+1)}{6}$$
$$\lim_{n\to\infty} \frac{x^3}{n^3}\frac{n(n+1)(2n+1)}{6}$$
$$\lim_{n\to\infty} \frac{x^3}{n^3}\frac{2n^3+3n^2+n}{6}$$
$$x^3 \lim_{n\to\infty} \left(\frac{2n^3}{6n^3}+\frac{3n^2}{6n^3}+\frac{n}{6n^3}\right)$$
$$x^3 \lim_{n\to\infty} \left(\frac{1}{3}+\frac{1}{2n}+\frac{1}{6n^2}\right)$$
$$x^3 \cdot \left(\frac{1}{3}\right)$$
$$\frac{x^3}{3}$$

If we are to write this formally, we remember our arbitrary constant, and we get $\frac{x^3}{3}+C$.

Recognizing Derivatives and the Substitution Rule

After learning a simple list of antiderivatives, it is time to move on to more complex integrands, which are not at first readily integrable. In these first steps, we notice certain special case integrands which can be easily integrated in a few steps.

Recognizing Derivatives and Reversing Derivative Rules

If we recognize a function $g(x)$ as being the derivative of a function $f(x)$, then we can easily express the antiderivative of $g(x)$:

$$\int g(x)\,dx = f(x) + C.$$

For example, since

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

we can conclude that

$$\int \cos x \, dx = \sin x + C.$$

Similarly, since we know e^x is its own derivative,

$$\int e^x \, dx = e^x + C.$$

The power rule for derivatives can be reversed to give us a way to handle integrals of powers of x. Since

$$\frac{d}{dx}x^n = nx^{n-1}$$

,

we can conclude that

$$\int nx^{n-1} \, dx = x^n + C,$$

or, a little more usefully,

$$\int x^n \, dx = \frac{x^{n+1}}{n+1} + C$$

.

Integration by Substitution

For many integrals, a substitution can be used to transform the integrand and make possible the finding of an antiderivative. There are a variety of such substitutions, each depending on the form of the integrand.

Integrating with the derivative present

If a component of the integrand can be viewed as the derivative of another component of the integrand, a substitution can be made to simplify the integrand.

For example, in the integral

$$\int 3x^2(x^3+1)^5\,dx$$

we see that $3x^2$ is the derivative of $x^3 + 1$. Letting

$u = x^3 + 1$

we have

$$\frac{du}{dx} = 3x^2$$

or, in order to apply it to the integral,

$du = 3x^2dx$.

With this we may write

$$\int 3x^2(x^3+1)^5\,dx = \int u^5\,du = \frac{1}{6}u^6 + C = \frac{1}{6}(x^3+1)^6 + C.$$

Note that it was not necessary that we had *exactly* the derivative of u in our integrand. It would have been sufficient to have any constant multiple of the derivative.

For instance, to treat the integral

$$\int x^4\sin(x^5)\,dx$$

we may let $u = x^5$. Then

$$du = 5x^4\,dx$$

and so

$$\frac{1}{5}du = x^4\,dx$$

the right-hand side of which is a factor of our integrand. Thus,

$$\int x^4\sin(x^5)\,dx = \int \frac{1}{5}\sin u\,du = -\frac{1}{5}\cos u + C = -\frac{1}{5}\cos x^5 + C.$$

In general, the integral of a power of a function times that function's derivative may be integrated in this way. Since $\frac{d[g(x)]}{dx} = g'(x)$,

we have $dx = \frac{d[g(x)]}{g'(x)}.$

Therefore, $$\begin{aligned}\int g'(x)[g(x)]^n &= \int g'(x)[g(x)]^n \frac{d[g(x)]}{g'(x)} \\ &= \int [g(x)]^n d[g(x)] \\ &= \frac{[g(x)]^{n+1}}{n+1}\end{aligned}$$

Integration by Parts

If y = u v where u and v are functions of x,

Then *y*' = (*uv*)' = *v*'*u* + *u*'*v*

Rearranging, *uv*' = (*uv*)' − *vu*'

Therefore, $\int uv'\,dx = \int (uv)'\,dx - \int vu'\,dx$

Therefore, $\int uv'\,dx = uv - \int vu'\,dx$, or

$$\int u\,dv = uv - \int v\,du.$$

This is the integration by parts formula. It is very useful in many integrals involving products of functions, as well as others.

For instance, to treat

$$\int x \sin x\,dx$$

we choose *u* = *x* and $dv = \sin x\,dx$. With these choices, we have *du* = *dx* and *v* = − cos*x*, and we have

$$\int x \sin x\,dx = -x\cos x - \int (-\cos x)\,dx = -x\cos x + \int \cos x\,dx = -x\cos x + \sin x + C.$$

Note that the choice of u and dv was critical. Had we chosen the reverse, so that $u = \sin x$ and $dv = x\,dx$, the result would have been

$$\frac{1}{2}x^2 \sin x - \int \frac{1}{2}x^2 \cos x \, dx.$$

The resulting integral is no easier to work with than the original; we might say that this application of integration by parts took us in the wrong direction.

So the choice is important. One general guideline to help us make that choice is, if possible, to choose u to be the factor of the integrand which *becomes simpler* when we differentiate it. In the last example, we see that $\sin x$ does not become simpler when we differentiate it: $\cos x$ is no simpler than $\sin x$.

An important feature of the integration by parts method is that we often need to apply it more than once. For instance, to integrate

$$\int x^2 e^x \, dx,$$

we start by choosing $u = x^2$ and $dv = e^x$ to get

$$\int x^2 e^x \, dx = x^2 e^x - 2\int x e^x \, dx.$$

Note that we still have an integral to take care of, and we do this by applying integration by parts again, with $u = x$ and $dv = e^x\,dx$, which gives us

$$\int x^2 e^x \, dx = x^2 e^x - 2\int x e^x \, dx = x^2 e^x - 2(x e^x - e^x) + C = x^2 e^x - 2x e^x + 2e^x + C.$$

So, two applications of integration by parts were necessary, owing to the power of x^2 in the integrand.

Note that *any power of x* does become simpler when we differentiate it, so when we see an integral of the form

$$\int x^n f(x) \, dx$$

one of our first thoughts ought to be to consider using integration by parts with $u = x^n$. Of course, in order for it to work, we need to be able to write down an antiderivative for dv.

Example

Use integration by parts to evaluate the integral

$$\int \sin(x)e^x\,dx$$

Solution: If we let $u = sin(x)$ and $v' = e^x$, then we have $u' = \cos(x)$ and $v = e^x$. Using our rule for integration by parts gives

$$\int \sin(x)e^x\,dx = e^x\sin(x) - \int \cos(x)e^x\,dx$$

We do not seem to have made much progress. But if we integrate by parts again with $u = cos(x)$ and $v' = e^x$ and hence $u' = -\sin(x)$ and $v = e^x$, we obtain

$$\int \sin(x)e^x\,dx = e^x\sin(x) - e^x\cos(x) - \int e^x\sin(x)\,dx$$

We may solve this identity to find the anti-derivative of $e^x\sin(x)$ and obtain

$$\int \sin(x)e^x\,dx = \frac{1}{2}e^x(\sin(x) - \cos(x)) + C$$

Integration by Complexifying

This technique requires an understanding and recognition of complex numbers. Specifically Euler's formula:

$$\cos\theta + i\cdot\sin\theta = e^{i\cdot\theta}$$

Recognize, for example, that the real portion:

$$\mathrm{Re}\{e^{i\cdot\theta}\} = \cos\theta$$

Given an integral of the general form:

$$\int e^x\cos 2x\,dx$$

We can complexify it:

$$\int \mathrm{Re}\{e^x(\cos 2x + i\cdot\sin 2x)\}\,dx$$

$$\int \mathrm{Re}\{e^{x}(e^{i2x})\}\,dx$$

With basic rules of exponents:

$$\int \mathrm{Re}\{e^{x+i2x}\}\,dx$$

It can be proven that the "real portion" operator can be moved outside the integral:

$$\mathrm{Re}\{\int e^{x(1+2i)}\,dx\}$$

The integral easily evaluates:

$$\mathrm{Re}\{\frac{e^{x(1+i2)}}{1+2i}\}$$

Multiplying and dividing by (1-2i):

$$\mathrm{Re}\{\frac{1-2i}{5}e^{x(1+i2)}\}$$

Which can be rewritten as:

$$\mathrm{Re}\{\frac{1-2i}{5}e^{x}e^{i2x}\}$$

Applying Euler's forumula:

$$\mathrm{Re}\{\frac{1-2i}{5}e^{x}(\cos 2x+i\cdot\sin 2x)\}$$

Expanding:

$$\mathrm{Re}\{\frac{e^{x}}{5}(\cos 2x+2\sin 2x)+i\cdot\frac{e^{x}}{5}(\sin 2x-2\cos 2x)\}$$

Taking the Real part of this expression:

$$\frac{e^{x}}{5}(\cos 2x+2\sin 2x)$$

So:

$$\int e^x \cos 2x \, dx = \frac{e^x}{5}(\cos 2x + 2\sin 2x)$$

Partial Fraction Decomposition

Suppose we want to find $\int \frac{3x+1}{x^2+x}dx$. One way to do this is to simplify the integrand by finding constants *A* and *B* so that

$$\frac{3x+1}{x^2+x} = \frac{3x+1}{x(x+1)} = \frac{A}{x} + \frac{B}{x+1}.$$

This can be done by cross multiplying the fraction which gives $\frac{3x+1}{x(x+1)} = \frac{A(x+1)+Bx}{x(x+1)}$. As both sides have the same denominator we must have $3x + 1 = A(x + 1) + Bx$. This is an equation for x so must hold whatever value x is. If we put in $x = 0$ we get $1 = A$ and putting $x = -1$ gives $-2 = -B$ so $B = 2$. So we see that

$$\frac{3x+1}{x^2+x} = \frac{1}{x} + \frac{2}{x+1}$$

Returning to the original integral

$$\begin{aligned}\int \frac{3x+1}{x^2+x}dx &= \int \frac{dx}{x} + \int \frac{2}{x+1}dx \\ &= \ln|x| + 2\ln|x+1| + C\end{aligned}$$

Rewriting the integrand as a sum of simpler fractions has allowed us to reduce the initial integral to a sum of simpler integrals. In fact this method works to integrate any rational function.

Method of Partial Fractions:

- **Step 1** Use long division to ensure that the degree of $P(x)$ less than the degree of $Q(x)$.
- **Step 2** Factor Q(x) as far as possible.
- **Step 3** Write down the correct form for the partial fraction decomposition (see below) and solve for the constants.

To factor Q(x) we have to write it as a product of linear factors (of the form $ax + b$) and irreducible quadratic factors (of the form $ax^2 + bx + c$ with $b^2 - 4ac < 0$).

Some of the factors could be repeated. For instance if $Q(x) = x^3 - 6x^2 + 9x$ we factor $Q(x)$ as

$$Q(x) = x(x^2 - 6x + 9) = x(x - 3)(x - 3) = x(x - 3)^2.$$

It is important that in each quadratic factor we have $b^2 - 4ac < 0$, otherwise it is possible to factor that quadratic piece further. For example if $Q(x) = x^3 - 3x^2 - 2x$ then we can write

$$Q(x) = x(x^2 - 3x + 2) = x(x - 1)(x + 2)$$

We will now show how to write $P(x) / Q(x)$ as a sum of terms of the form

$$\frac{A}{(ax+b)^k} \text{ and } \frac{Ax+B}{(ax^2+bx+c)^k}.$$

Exactly how to do this depends on the factorization of $Q(x)$ and we now give four cases that can occur.

Case (a) Q(x) is a product of linear factors with no repeats.

This means that $Q(x) = (a_1x + b_1)(a_2x + b_2)...(a_nx + b_n)$ where no factor is repeated and no factor is a multiple of another.

For each linear term we write down something of the form $\frac{A}{(ax+b)}$, so in total we write

$$\frac{P(x)}{Q(x)} = \frac{A_1}{(a_1x+b_1)} + \frac{A_2}{(a_2x+b_2)} + \cdots + \frac{A_n}{(a_nx+b_n)}$$

Example 1

Find $\int \frac{1+x^2}{(x+3)(x+5)(x+7)}dx$

Here we have $P(x) = 1 + x^2$,$Q(x) = (x + 3)(x + 5)(x + 7)$ and *Q(x)* is a product of linear factors. So we write

$$\frac{1+x^2}{(x+3)(x+5)(x+7)} = \frac{A}{x+3} + \frac{B}{x+5} + \frac{C}{x+7}$$

Multiply both sides by the denominator

$1 + x^2 = A(x + 5)(x + 7) + B(x + 3)(x + 7) + C(x + 3)(x + 5)$

Substitute in three values of x to get three equations for the unknown constants,

$$\begin{array}{ll} x = -3 & 1 + 3^2 = 2 \cdot 4A \\ x = -5 & 1 + 5^2 = -2 \cdot 2B \\ x = -7 & 1 + 7^2 = (-4) \cdot (-2)C \end{array}$$

so $A = 5 / 4, B = -13 / 2, C = 25 / 4$, and

$$\frac{1+x^2}{(x+3)(x+5)(x+7)} = \frac{5}{4x+12} - \frac{13}{2x+10} + \frac{25}{4x+28}$$

We can now integrate the left hand side.

$$\int \frac{1+x^2\,dx}{(x+3)(x+5)(x+7)} = \frac{5}{4}\ln|x+3| - \frac{13}{2}\ln|x+5| + \frac{25}{4}\ln|x+7| + C$$

Case (b) Q(x) is a product of linear factors some of which are repeated.

If $(ax + b)$ appears in the factorisation of $Q(x)$ k-times. Then instead of writing the piece $\frac{A}{(ax+b)}$ we use the more complicated expression

$$\frac{A_1}{ax+b} + \frac{A_2}{(ax+b)^2} + \frac{A_3}{(ax+b)^3} + \cdots + \frac{A_k}{(ax+b)^k}$$

Example 2

Find $\int \frac{1}{(x+1)(x+2)^2} dx$

Here *P(x)=1" and "Q(x)=(x+1)(x+2)*^{2} We write

$$\frac{1}{(x+1)(x+2)^2} = \frac{A}{x+1} + \frac{B}{x+2} + \frac{C}{(x+2)^2}$$

Multiply both sides by the denominator $1 = A(x+2)^2 + B(x+1)(x+2) + C(x+1)$

Substitute in three values of x to get 3 equations for the unknown constants,

$$\begin{aligned} x = 0 \quad & 1 = 2^2 A + 2B + C \\ x = -1 \quad & 1 = A \\ x = -2 \quad & 1 = -C \end{aligned}$$

so A=1, B=-1, C=-1, and

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

We can now integrate the left hand side.

$$\int \frac{1}{(x+1)(x+2)^2} dx = \ln \frac{x+1}{x+2} + \frac{1}{x+2} + C$$

Case (c) Q(x) contains some quadratic pieces which are not repeated.

If $(ax^2 + bx + c)$ appears we use $\dfrac{Ax+B}{(ax^2+bx+c)}$.

Case (d) Q(x) contains some repeated quadratic factors.

If $(ax^2 + bx + c)$ appears k-times then use

$$\frac{A_1 x + B_1}{(ax^2+bx+c)} + \frac{A_2 x + B_2}{(ax^2+bx+c)^2} + \frac{A_3 x + B_3}{(ax^2+bx+c)^3} + \cdots + \frac{A_k x + B_k}{(ax^2+bx+c)^k}$$

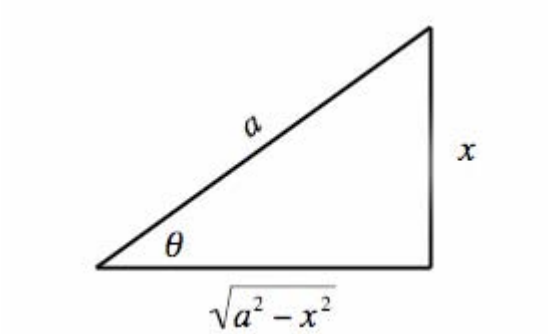
Trigonometric Substitution

If the integrand contains a single factor of one of the forms $\sqrt{a^2 - x^2}$ or $\sqrt{a^2 + x^2}$ or $\sqrt{x^2 - a^2}$ we can try a trigonometric substitution.

- If the integrand contains $\sqrt{a^2 - x^2}$ let $x = a\sin\theta$ and use the identity $1 - \sin^2\theta = \cos^2\theta$.

- If the integrand contains $\sqrt{a^2 + x^2}$ let $x = a\tan\theta$ and use the identity $1 + \tan^2\theta = \sec^2\theta$.
- If the integrand contains $\sqrt{x^2 - a^2}$ let $x = a\sec\theta$ and use the identity $\sec^2\theta - 1 = \tan^2\theta$.

Sine substitution



This substitution is easily derived from a triangle, using the Pythagorean Theorem.

If the integrand contains a piece of the form $\sqrt{a^2 - x^2}$ we use the substitution

$$x = a\sin\theta \quad dx = a\cos\theta d\theta$$

This will transform the integrand to a trigonometic function. If the new integrand can't be integrated on sight then the tan-half-angle substitution described below will generally transform it into a more tractable algebraic integrand.

Eg, if the integrand is $\sqrt{(1-x^2)}$,

$$\begin{aligned}\int_0^1 \sqrt{1-x^2}dx &= \int_0^{\pi/2} \sqrt{1-\sin^2\theta}\cos\theta\, d\theta \\ &= \int_0^{\pi/2} \cos^2\theta\, d\theta \\ &= \tfrac{1}{2}\int_0^{\pi/2} 1+\cos 2\theta\, d\theta \\ &= \tfrac{\pi}{4}\end{aligned}$$

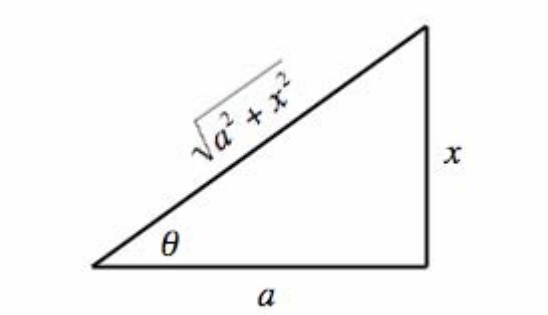
If the integrand is $\sqrt{(1+x)}/\sqrt{(1-x)}$, we can rewrite it as

$$\sqrt{\frac{1+x}{1-x}} = \sqrt{\frac{1+x}{1+x}\frac{1+x}{1-x}} = \frac{1+x}{\sqrt{1-x^2}}$$

Then we can make the substitution

$$\begin{aligned}
\int_0^a \frac{1+x}{\sqrt{1-x^2}}dx &= \int_0^\alpha \frac{1+\sin\theta}{\cos\theta}\cos\theta\, d\theta \qquad 0<a<1 \\
&= \int_0^\alpha 1+\sin\theta\, d\theta \qquad \alpha = \sin^{-1} a \\
&= \alpha + [-\cos\theta]_0^\alpha \\
&= \alpha + 1 - \cos\alpha \\
&= 1 + \sin^{-1} a - \sqrt{1-a^2}
\end{aligned}$$

Tangent substitution



This substitution is easily derived from a triangle, using the Pythagorean Theorem.

When the integrand contains a piece of the form $\sqrt{a^2+x^2}$ we use the substitution

$$x = a\tan\theta \quad \sqrt{x^2+a^2} = a\sec\theta \quad dx = a\sec^2\theta d\theta$$

E.g, if the integrand is $(x^2+a^2)^{-3/2}$ then on making this substitution we find

$$\begin{aligned}
\int_0^z (x^2+a^2)^{-\frac{3}{2}}\, dx &= a^{-2}\int_0^\alpha \cos\theta\, d\theta \qquad z>0 \\
&= a^{-2}[\sin\theta]_0^\alpha \qquad \alpha = \tan^{-1}(z/a) \\
&= a^{-2}\sin\alpha \\
&= a^{-2}\frac{z/a}{\sqrt{1+z^2/a^2}} \qquad = \frac{1}{a^2}\frac{z}{\sqrt{a^2+z^2}}
\end{aligned}$$

If the integral is

$$I = \int_0^z \sqrt{x^2+a^2} \quad z>0$$

then on making this substitution we find

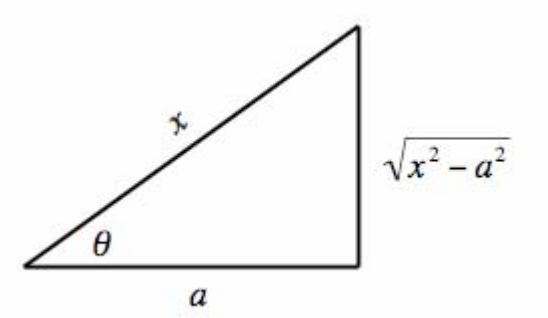
$$\begin{aligned} I &= a^2 \int_0^\alpha \sec^3\theta\, d\theta && \alpha = \tan^{-1}(z/a) \\ &= a^2 \int_0^\alpha \sec\theta\, d\tan\theta \\ &= a^2[\sec\theta\tan\theta]_0^\alpha - a^2 \int_0^\alpha \sec\theta\tan^2\theta\, d\theta \\ &= a^2\sec\alpha\tan\alpha - a^2 \int_0^\alpha \sec^3\theta\, d\theta + a^2 \int_0^\alpha \sec\theta\, d\theta \\ &= a^2\sec\alpha\tan\alpha - I + a^2 \int_0^\alpha \sec\theta\, d\theta \end{aligned}$$

After integrating by parts, and using trigonometric identities, we've ended up with an expression involving the original integral. In cases like this we must now rearrange the equation so that the original integral is on one side only

$$\begin{aligned} I &= \tfrac{1}{2}a^2\sec\alpha\tan\alpha + \tfrac{1}{2}a^2 \int_0^\alpha \sec\theta\, d\theta \\ &= \tfrac{1}{2}a^2\sec\alpha\tan\alpha + \tfrac{1}{2}a^2 \left[\ln(\sec\theta+\tan\theta)\right]_0^\alpha \\ &= \tfrac{1}{2}a^2\sec\alpha\tan\alpha + \tfrac{1}{2}a^2 \ln(\sec\alpha+\tan\alpha) \\ &= \tfrac{1}{2}a^2\left(\sqrt{1+\tfrac{z^2}{a^2}}\right)\tfrac{z}{a} + \tfrac{1}{2}a^2\ln\left(\sqrt{1+\tfrac{z^2}{a^2}}+\tfrac{z}{a}\right) \\ &= \tfrac{1}{2}z\sqrt{z^2+a^2} + \tfrac{1}{2}a^2\ln\left(\tfrac{z}{a}+\sqrt{1+\tfrac{z^2}{a^2}}\right) \end{aligned}$$

As we would expect from the integrand, this is approximately $z^2/2$ for large z.

Secant substitution



This substitution is easily derived from a triangle, using the Pythagorean Theorem.

If the integrand contains a factor of the form $\sqrt{x^2-a^2}$ we use the substitution

$$x = a\sec\theta \quad dx = a\sec\theta\tan\theta d\theta \quad \sqrt{x^2-a^2} = a\tan\theta.$$

Example 1

Find $\displaystyle\int_1^z \frac{\sqrt{x^2-1}}{x}dx.$

$$\begin{aligned}
\int_1^z \frac{\sqrt{x^2-1}}{x}dx &= \int_1^\alpha \frac{\tan\theta}{\sec\theta}\sec\theta\tan\theta\,d\theta & z>1\\
&= \int_0^\alpha \tan^2\theta\,d\theta & \alpha=\sec^{-1}z\\
&= \left[\tan\theta-\theta\right]_0^\alpha & \tan\alpha=\sqrt{\sec^2\alpha-1}\\
&= \tan\alpha-\alpha & \tan\alpha=\sqrt{z^2-1}\\
&= \sqrt{z^2-1}-\sec^{-1}z
\end{aligned}$$

Example 2

Find $\displaystyle\int_1^z \frac{\sqrt{x^2-1}}{x^2}dx.$

$$\begin{aligned}
\int_1^z \frac{\sqrt{x^2-1}}{x^2}dx &= \int_1^\alpha \frac{\tan\theta}{\sec^2\theta}\sec\theta\tan\theta\,d\theta & z>1\\
&= \int_0^\alpha \frac{\sin^2\theta}{\cos\theta}d\theta & \alpha=\sec^{-1}z
\end{aligned}$$

We can now integrate by parts

$$\begin{aligned}
\int_1^z \frac{\sqrt{x^2-1}}{x^2}dx &= -\left[\tan\theta\cos\theta\right]_0^\alpha+\int_0^\alpha\sec\theta\,d\theta\\
&= -\sin\alpha+\left[\ln(\sec\theta+\tan\theta)\right]_0^\alpha\\
&= \ln(\sec\alpha+\tan\alpha)-\sin\alpha\\
&= \ln(z+\sqrt{z^2-1})-\frac{\sqrt{z^2-1}}{z}
\end{aligned}$$

Trigonometric Integrals

When the integrand is primarily or exclusively based on trigonometric functions, the following techniques are useful.

Powers of Sine and Cosine

We will give a general method to solve generally integrands of the form $\cos^m(x)\sin^n(x)$. First let us work through an example.

$$\int(\cos^3 x)(\sin^2 x)\,dx$$

Notice that the integrand contains an odd power of cos. So rewrite it as

$$\int(\cos^2 x)(\sin^2 x)\cos x\,dx$$

We can solve this by making the substitution $u = \sin(x)$ so $du = \cos(x)\,dx$. Then we can write the whole integrand in terms of u by using the identity

$$\cos(x)^2 = 1 - \sin^2(x) = 1 - u^2.$$

So

$$\begin{aligned}
\int (\cos^3 x)(\sin^2 x)\,dx &= \int (\cos^2 x)(\sin^2 x)\cos x\,dx \\
&= \int (1-u^2)u^2\,du \\
&= \int u^2\,du - \int u^4\,du \\
&= \tfrac{1}{3}u^3 + \tfrac{1}{5}u^5 + C \\
&= \tfrac{1}{3}\sin^3 x - \tfrac{1}{5}\sin^5 x + C
\end{aligned}.$$

This method works whenever there is an odd power of sine or cosine.

To evaluate $\int (\cos^m x)(\sin^n x)\,dx$ when **either** m or n is **odd**.

- If m is odd substitute $u = \sin x$ and use the identity $\cos^2 x = 1 - \sin^2 x = 1 - u^2$.
- If n is odd substitute $u = \cos x$ and use the identity $\sin^2 x = 1 - \cos^2 x = 1 - u^2$.

Example

Find $\int_0^{\pi/2} \cos^{40}(x)\sin^3(x)dx$.

As there is an odd power of sin we let $u = \cos x$ so $du = -\sin(x)dx$. Notice that when $x=0$ we have $u=\cos(0)=1$ and when $x = \pi / 2$ we have $u = \cos(\pi / 2) = 0$.

$$\begin{aligned}
\int_0^{\pi/2} \cos^{40}(x)\sin^3(x)dx &= \int_0^{\pi/2} \cos^{40}(x)\sin^2(x)\sin(x)dx \\
&= -\int_1^0 u^{40}(1-u^2)du \\
&= \int_0^1 u^{40}(1-u^2)du \\
&= \int_0^1 u^{40} - u^{42}du \\
&= [\tfrac{1}{41}u^{41} - \tfrac{1}{43}u^{43}]_0^1 \\
&= \tfrac{1}{41} - \tfrac{1}{43}.
\end{aligned}$$

When both m and n are even things get a little more complicated.

To evaluate $\int (\cos^m x)(\sin^n x)\,dx$ when both *m* and *n* are **even**.
Use the identities $\sin^2 x = 1/2\ (1- \cos 2x)$ and $\cos^2 x = 1/2\ (1+ \cos 2x)$.

Example

Find $\int \sin^2 x \cos^4 x\,dx.$

As $\sin^2 x = 1/2\ (1- \cos 2x)$ and $\cos^2 x = 1/2\ (1+ \cos 2x)$ we have

$$\int \sin^2 x \cos^4 x\,dx = \int \left(\frac{1}{2}(1-\cos 2x)\right)\left(\frac{1}{2}(1+\cos 2x)\right)^2\,dx,$$

and expanding, the integrand becomes

$$\frac{1}{8}\left(\int 1-\cos^2 2x+\cos 2x-\cos^3 2x\,dx\right).$$

Using the multiple angle identities

$$\begin{aligned} I &= \tfrac{1}{8}\left(\int 1\,dx-\int \cos^2 2x\,dx+\int \cos 2x\,dx-\int \cos^3 2x\,dx\right) \\ &= \tfrac{1}{8}\left(x-\tfrac{1}{2}\int (1+\cos 4x)\,dx+\tfrac{1}{2}\sin 2x-\int \cos^2 2x\cos 2x\,dx\right) \\ &= \tfrac{1}{16}\left(x+\sin 2x+\int \cos 4x\,dx-2\int (1-\sin^2 2x)\cos 2x\,dx\right) \end{aligned}$$

then we obtain on evaluating

$$I = \frac{x}{16}-\frac{\sin 4x}{64}+\frac{\sin^3 2x}{48}+C$$

Powers of Tan and Secant

To evaluate $\int (\tan^m x)(\sec^n x)\,dx$.

1. If *n* is even and $n \geq 2$ then substitute $u=\tan x$ and use the identity $\sec^2 x = 1 + \tan^2 x$.
2. If *n* and *m* are both odd then substitute $u=\sec x$ and use the identity $\tan^2 x = \sec^2 x-1$.
3. If *n* is odd and *m* is even then use the identity $\tan^2 x = \sec^2 x-1$ and apply a reduction formula to integrate $\sec^j x dx$.

Example 1

Find $\int \sec^2 x dx$.

There is an even power of sec$x$. Substituting $u = \tan x$ gives $du = \sec^2 x dx$ so

$$\int \sec^2 x dx = \int du = u + C = \tan x + C.$$

Example 2

Find $\int \tan x dx$.

Let $u = \cos x$ so $du = -\sin x dx$. Then

$$\begin{aligned} \int \tan x dx &= \int \frac{\sin x}{\cos x} dx \\ &= \int \frac{-1}{u} du \\ &= -\ln|u| + C \\ &= -\ln|\cos x| + C \\ &= \ln|\sec x| + C. \end{aligned}$$

Example 3

Find $\int \sec x dx$.

The trick to do this is to multiply and divide by the same thing like this:

$$\begin{aligned} \int \sec x dx &= \int \sec x \frac{\sec x + \tan x}{\sec x + \tan x} dx \\ &= \int \frac{\sec^2 x + \sec x \tan x}{\sec x + \tan x} \end{aligned}.$$

Making the substitution $u = \sec x + \tan x$ so $du = \sec x \tan x + \sec^2 x dx$,

$$\begin{aligned} \int \sec x dx &= \int \frac{1}{u} du \\ &= \ln|u| + C \\ &= \ln|\sec x + \tan x| + C \end{aligned}.$$

More trigonometric combinations

For the integrals $\int \sin nx \cos mx \, dx$ or $\int \sin nx \sin mx \, dx$ or $\int \cos nx \cos mx \, dx$ use the identities

- $\sin a \cos b = \frac{1}{2}(\sin(a+b) + \sin(a-b))$
- $\sin a \sin b = \frac{1}{2}(\cos(a-b) - \cos(a+b))$
- $\cos a \cos b = \frac{1}{2}(\cos(a-b) + \cos(a+b))$

Example 1

Find $\int \sin 3x \cos 5x \, dx.$

We can use the fact that sin *a* cos *b*=(1/2)(sin(*a*+*b*)+sin(*a*-*b*)), so

$$\sin 3x \cos 5x = (\sin 8x + \sin(-2x))/2$$

Now use the oddness property of sin(*x*) to simplify

$$\sin 3x \cos 5x = (\sin 8x - \sin 2x)/2$$

And now we can integrate

$$\begin{aligned} \int \sin 3x \cos 5x \, dx &= \tfrac{1}{2}\int \sin 8x - \sin 2x dx \\ &= \tfrac{1}{2}(-\tfrac{1}{8}\cos 8x + \tfrac{1}{2}\cos 2x) + C \end{aligned}$$

Example 2

Find: $\int \sin x \sin 2x \, dx$.

Using the identities

$$\sin x \sin 2x = \frac{1}{2}(\cos(-x) - \cos(3x)) = \frac{1}{2}(\cos x - \cos 3x).$$

Then

$$\begin{aligned}\int \sin x \sin 2x \, dx &= \tfrac{1}{2}\int (\cos x - \cos 3x)\, dx \\ &= \tfrac{1}{2}(\sin x - \tfrac{1}{3}\sin 3x) + C\end{aligned}$$

Reduction Formula

A *reduction formula* is one that enables us to solve an integral problem by *reducing* it to a problem of solving an easier integral problem, and then reducing that to the problem of solving an easier problem, and so on.

For example, if we let

$$I_n = \int x^n e^x \, dx$$

Integration by parts allows us to simplify this to

$$\begin{aligned}I_n &= x^n e^x - n\int x^{n-1} e^x \, dx = \\ I_n &= x^n e^x - n I_{n-1}\end{aligned}$$

which is our desired reduction formula. Note that we stop at

$$I_0 = e^x.$$

Similarly, if we let

$$I_n = \int_0^{\alpha} \sec^n \theta \, d\theta$$

then integration by parts lets us simplify this to

$$I_n = \sec^{n-2}\alpha \tan\alpha - (n-2)\int_0^{\alpha} \sec^{n-2}\theta \tan^2\theta \, d\theta$$

Using the trigonometric identity, $\tan^2=\sec^2-1$, we can now write

$$\begin{aligned}I_n &= \sec^{n-2}\alpha\tan\alpha + (n-2)\left(\int_0^{\alpha}\sec^{n-2}\theta\, d\theta - \int_0^{\alpha}\sec^n\theta\, d\theta\right) \\ &= \sec^{n-2}\alpha\tan\alpha + (n-2)\left(I_{n-2} - I_n\right)\end{aligned}$$

Rearranging, we get

$$I_n = \frac{1}{n-1}\sec^{n-2}\alpha\tan\alpha + \frac{n-2}{n-1}I_{n-2}$$

Note that we stop at n=1 or 2 if n is odd or even respectively.

As in these two examples, integrating by parts when the integrand contains a power often results in a reduction formula.

Rational Functions Using Trig

Here we look at using trigonometry to simplify various rational integrands.

The "tan half angle" substitution

Another useful change of variables is

$$t = \tan(x/2)$$

With this transformation, using the double-angle trig identities,

$$\sin x = \frac{2t}{1+t^2} \quad \cos x = \frac{1-t^2}{1+t^2} \quad \tan x = \frac{2t}{1-t^2} \quad dx = \frac{2dt}{1+t^2}$$

This transforms a trigonometric integral into a algebraic integral, which may be easier to integrate.

For example, if the integrand is 1/(1 + sin x) then

$$\begin{aligned} \int_0^{\pi/2} \frac{dx}{1+\sin x} &= \int_0^1 \frac{2dt}{(1+t)^2} \\ &= \left[-\frac{2}{1+t}\right]_0^1 \\ &= 1 \end{aligned}$$

This method can be used to further simplify trigonometric integrals produced by the changes of variables described earlier.

For example, if we are considering the integral

$$I = \int_{-1}^{1} \frac{\sqrt{1-x^2}}{1+x^2} dx$$

we can first use the substitution x= sin θ, which gives

$$I = \int_{-\pi/2}^{\pi/2} \frac{\cos^2\theta}{1+\sin^2\theta} d\theta$$

then use the tan-half-angle substition to obtain

$$I = \int_{-1}^{1} \frac{(1-t^2)^2}{1+6t^2+t^4} \frac{2dt}{1+t^2}$$

In effect, we've removed the square root from the original integrand. We could do this with a single change of variables, but doing it in two steps gives us the opportunity of doing the trigonometric integral another way.

Having done this, we can split the new integrand into partial fractions, and integrate.

$$\begin{aligned} I &= \int_{-1}^{1} \frac{2-\sqrt{2}}{t^2+3-\sqrt{8}} dt + \int_{-1}^{1} \frac{2+\sqrt{2}}{t^2+3+\sqrt{8}} dt - \int_{-1}^{1} \frac{2}{1+t^2} dt \\ &= \frac{4-\sqrt{8}}{\sqrt{3-\sqrt{8}}} \tan^{-1}(\sqrt{3+\sqrt{8}}) + \frac{4+\sqrt{8}}{\sqrt{3+\sqrt{8}}} \tan^{-1}(\sqrt{3-\sqrt{8}}) - \pi \end{aligned}$$

This result can be further simplified by use of the identities

$$3 \pm \sqrt{8} = \left(\sqrt{2} \pm 1\right)^2 \quad \tan\left(\sqrt{2} \pm 1\right) = \left(\frac{1}{4} \pm \frac{1}{8}\right)\pi$$

ultimately leading to

$$I = (\sqrt{2}-1)\pi$$

In principle, this approach will work with any integrand which is the square root of a quadratic multiplied by the ratio of two polynomials. However, it should not be applied automatically.

E.g, in this last example, once we deduced

$$I = \int_{-\pi/2}^{\pi/2} \frac{\cos^2\theta}{1+\sin^2\theta} d\theta$$

we could have used the double angle formula, since this contains only even powers of cos and sin. Doing that gives

$$I = \int_{-\pi/2}^{\pi/2} \frac{1+\cos 2\theta}{3-\cos 2\theta} d\theta = \frac{1}{2} \int_{-\pi}^{\pi} \frac{1+\cos\phi}{3-\cos\phi} d\phi$$

Using tan-half-angle on this new, simpler, integrand gives

$$\begin{aligned} I &= \int_{-\infty}^{\infty} \frac{1}{1+2t^2} \frac{dt}{1+t^2} \\ &= \int_{-\infty}^{\infty} \frac{2dt}{1+2t^2} - \int_{-\infty}^{\infty} \frac{dt}{1+t^2} \end{aligned}$$

This can be integrated on sight to give

$$I = \frac{4}{\sqrt{2}}\frac{\pi}{2} - 2\frac{\pi}{2} = (\sqrt{2} - 1)\pi$$

This is the same result as before, but obtained with less algebra, which shows why it is best to look for the most straightforward methods at every stage.

A more direct way of evaluating the integral I is to substitute $t = \tan\theta$ right from the start, which will directly bring us to the line

$$I = \int_{-\infty}^{\infty} \frac{1}{1+2t^2}\frac{dt}{1+t^2}$$

above. More generally, the substitution $t = \tan x$ gives us

$$\sin x = \frac{t}{\sqrt{1+t^2}} \quad \cos x = \frac{1}{\sqrt{1+t^2}} \quad dx = \frac{dt}{1+t^2}$$

so this substitution is the preferable one to use if the integrand is such that all the square roots would disappear after substitution, as is the case in the above integral.

Alternate Method

In general, to evaluate integrals of the form

$$\int \frac{A + B\cos x + C\sin x}{a + b\cos x + c\sin x}\,dx$$

,

it is extremely tedious to use the aforementioned "tan half angle" substitution directly, as one easily ends up with a rational function with a 4th degree denominator. Instead, we may first write the numerator as

$$A + B\cos x + C\sin x \equiv p(a + b\cos x + c\sin x) + q\frac{d}{dx}(a + b\cos x + c\sin x) + r$$

.

Then the integral can be written as

$$\int \left(p + \frac{q\frac{d}{dx}(a + b\cos x + c\sin x)}{a + b\cos x + c\sin x} + \frac{r}{a + b\cos x + c\sin x} \right) dx$$

which can be evaluated much more easily.

Example

Evaluate $\int \frac{\cos x + 2}{\cos x + \sin x}\, dx$.

Let

$$\cos x + 2 \equiv p(\cos x + \sin x) + q\frac{d}{dx}(\cos x + \sin x) + r.$$

Then

$$\cos x + 2 \equiv p(\cos x + \sin x) + q(-\sin x + \cos x) + r$$
$$\cos x + 2 \equiv (p+q)\cos x + (p-q)\sin x + r.$$

Comparing coefficients of cos *x*, sin *x* and the constants on both sides, we obtain

$$\begin{cases} p+q &= 1 \\ p-q &= 0 \\ r &= 2 \end{cases}$$

yielding $p = q = 1/2$, r = 2. Substituting back into the integrand,

$$\int \frac{\cos x + 2}{\cos x + \sin x}\, dx = \int \frac{1}{2}\, dx + \frac{1}{2}\int \frac{d(\cos x + \sin x)}{\cos x + \sin x} + \int \frac{2}{\cos x + \sin x}\, dx$$

.

The last integral can now be evaluated using the "tan half angle" substitution described above, and we obtain

$$\int \frac{2}{\cos x + \sin x}\, dx = \sqrt{2}\ln\left|\frac{\tan\frac{x}{2} - 1 + \sqrt{2}}{\tan\frac{x}{2} - 1 - \sqrt{2}}\right| + C.$$

The original integral is thus

$$\int \frac{\cos x + 2}{\cos x + \sin x}\, dx = \frac{x}{2} + \frac{1}{2}\ln|\cos x + \sin x| + \sqrt{2}\ln\left|\frac{\tan\frac{x}{2} - 1 + \sqrt{2}}{\tan\frac{x}{2} - 1 - \sqrt{2}}\right| + C$$

.

Irrational Functions

Integration of irrational functions is more difficult than rational functions, and many cannot be done. However, there are some particular types that can be reduced to rational forms by suitable substitutions.

Type 1

Integrand contains $\sqrt[n]{\frac{ax+b}{cx+d}}$

Use the substitution $u = \sqrt[n]{\frac{ax+b}{cx+d}}$.

Example

Find $\int \frac{1}{x}\sqrt{\frac{1-x}{x}}\,dx$.

$\int \frac{x}{\sqrt[3]{ax+b}}\,dx$

Type 2

Integral is of the form $\int \frac{Px+Q}{\sqrt{ax^2+bx+c}}\,dx$

Write $Px + Q$ as $Px + Q = p\frac{d}{dx}(ax^2+bx+c) + q$.

Example

Find $\int \frac{4x-1}{\sqrt{5-4x-x^2}}\,dx$.

Type 3

Integrand contains $\sqrt{a^2-x^2}$, $\sqrt{a^2+x^2}$ or $\sqrt{x^2-a^2}$

This was discussed in "trigonometric substitutions above". Here is a summary:

1. For $\sqrt{a^2-x^2}$, use $x = a\sin\theta$.

2. For $\sqrt{a^2 + x^2}$, use $x = a\tan\theta$.
3. For $\sqrt{x^2 - a^2}$, use $x = a\sec\theta$.

Type 4

Integral is of the form $\int \frac{1}{(px+q)\sqrt{ax^2+bx+c}}\,dx$

Use the substitution $u = \frac{1}{px+q}$.

Example

Find $\int \frac{1}{(1+x)\sqrt{3+6x+x^2}}\,dx$.

Type 5

Other rational expressions with the irrational function $\sqrt{ax^2+bx+c}$

1. If $a > 0$, we can use $u = \sqrt{ax^2+bx+c} \pm \sqrt{a}x$.
2. If $c > 0$, we can use $u = \frac{\sqrt{ax^2+bx+c} \pm \sqrt{c}}{x}$.
3. If $ax^2 + bx + c$ can be factored as $a(x - \alpha)(x - \beta)$, we can use $u = \sqrt{\frac{a(x-\alpha)}{x-\beta}}$.
4. If $a < 0$ and $ax^2 + bx + c$ can be factored as $-a(\alpha - x)(x - \beta)$, we can use $x = \alpha\cos^2\theta + \beta\sin^2\theta$, / *theta* + β.

Numerical Approximations

It is often the case, when evaluating definite integrals, that an antiderivative for the integrand cannot be found, or is extremely difficult to find. In some instances, a numerical approximation to the value of the definite value will suffice. The following techniques can be used, and are listed in rough order of ascending complexity.

Riemann Sum

This comes from the definition of an integral. If we pick n to be finite, then we have:

$$\int_a^b f(x)\ dx \approx \sum_{i=1}^{n} f(x_i^*)\Delta x$$

where x_i^* is any point in the i-th sub-interval $[x_{i-1},x_i]$ on [a,b].

Right Rectangle

A special case of the Riemann sum, where we let $x_i^* = x_i$, in other words the point on the far right-side of each sub-interval on, [a,b]. Again if we pick n to be finite, then we have:

$$\int_a^b f(x)\ dx \approx \sum_{i=1}^{n} f(x_i)\Delta x$$

Left Rectangle

Another special case of the Riemann sum, this time we let $x_i^* = x_{i-1}$, which is the the point on the far left side of each sub-interval on [a,b]. As always, this is an approximation when n is finite. Thus, we have:

$$\int_a^b f(x)\ dx \approx \sum_{i=1}^{n} f(x_{i-1})\Delta x$$

Trapezoidal Rule

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2n}\left[f(x_0) + 2\sum_{i=1}^{n-1}(f(x_i)) + f(x_n)\right] = \frac{b-a}{2n}(f(x_0)+2f(x_1)+2f(x_2)+\cdots+2f(x_{n-1})+f(x_n))$$

Simpson's Rule

Remember, n must be even,

$$\begin{aligned}\int_a^b f(x)\ dx &\approx \frac{b-a}{3n}\left[f(x_0) + \sum_{i=1}^{n-1}\left((3-(-1)^i)f(x_i)\right) + f(x_n)\right] \\ &= \frac{b-a}{3n}\left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \cdots + 4f(x_{n-1}) + f(x_n)\right]\end{aligned}$$

Improper Integrals

In a definite integral ($\int_a^b f(x)\,dx$) the function has defined intervals and the function itself is continuous. In this section, we deal with integrals of functions where the interval is infinite (type I) or the function has infinite discontinuity in the intervals [a,b] (type II).

Type I: Infinite Integrals

An integral with infinite region includes ±∞ included in the interval such as $\int_{-\infty}^{\infty} f(x)\,dx$. We cannot simply find the antiderivative and plug in ∞. We can however rewrite the integral using a limit. Let $\int_1^{\infty} \frac{1}{x^2}\,dx = \lim_{b\to\infty} \int_1^b \frac{1}{x^2}\,dx$

Now this represents a definite integral so we can find the antiderivative and see if the integral converges. $\lim_{b\to\infty} \int_1^b \frac{1}{x^2}\,dx = \lim_{b\to\infty} \left[-\frac{1}{x}\right]_1^b = \lim_{b\to\infty} -\frac{1}{b} + 1 = 1$

We can now define the the type 1 integral:

(a) If there is some value *b* where $b \geq a$and $\int_a^b f(x)\,dx$ exists, then

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

(b) If there is some value *a* where $a \leq b$and $\int_a^b f(x)\,dx$ exists, then

$$\int_{-\infty}^b f(x)\,dx = \lim_{a\to-\infty} \int_a^b f(x)\,dx$$

(c) We can also define $\int_{-\infty}^{\infty} f(x)\,dx$ as

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{a} f(x)\,dx + \int_a^{\infty} f(x)\,dx$$ assuming that both integrals converge.

**note that if the limits fail to exist, we say that the integral diverges and if the improper integrals yield a finite solution, the integral converges.

Lets look at an example: Evaluate the integral if it converges. $\int_{-\infty}^{\infty} xe^{x^{-2}}\,dx$

$\int_{-\infty}^{\infty} xe^{x^{-2}}\,dx = \lim_{a\to-\infty} \int_a^0 xe^{x^{-2}}\,dx + \lim_{b\to\infty} \int_0^b xe^{x^{-2}}\,dx$ Use the chain rule to

find the antiderivative with $u = -x^2, dx = -\frac{du}{2x}$

$$\lim_{a\to\infty} \int_a^0 xe^{x^{-2}}\,dx = \frac{-e^{(0)^{-2}}}{2} - \lim_{a\to-\infty} \frac{-e^{(a)^{-2}}}{2} = (-\frac{1}{2}) - (0)$$

$$\lim_{b\to\infty} \int_0^b xe^{x^{-2}}\,dx = \lim_{b\to\infty} \frac{-e^{(b)^{-2}}}{2} - \frac{-e^{(0)^{-2}}}{2} = (0) + \frac{1}{2}$$

$\frac{1}{2} - \frac{1}{2} = 0$ This shows that the integral converges to 0.

Type II: Infinite Discontinuity
Integrating a function that contains a vertical asymptote.

Applications of Integration

Area

Finding the area between two curves, usually given by two explicit functions, is often useful in calculus.

In general the rule for finding the area between two curves is

$A = A_{top} - A_{bottom}$ or

If f(x) is the upper function and g(x) is the lower function

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

This is true whether the functions are in the first quadrant or not.

Area between two curves

Suppose we are given two functions $y_1=f(x)$ and $y_2=g(x)$ and we want to find the area between them on the interval $[a,b]$. Also assume that $f(x)\geq g(x)$ for all x on the interval $[a,b]$. Begin by partitioning the interval $[a,b]$ into n equal subintervals each having a length of $\Delta x=(b-a)/n$. Next choose any point in each subinterval, x_i^*. Now we can 'create' rectangles on each interval. At the point x_i^*, the height of each rectangle is $f(x_i^*)-g(x_i^*)$

and the width is Δx. Thus the area of each rectangle is $[f(x_i^*)-g(x_i^*)]\Delta x$. An *approximation* of the area, A, between the two curves is

$$A := \sum_{i=1}^{n} [f(x_i^*) - g(x_i^*)]\Delta x$$

.

Now we take the limit as n approaches infinity and get

$$A = \lim_{n\to\infty} \sum_{i=1}^{n} [f(x_i^*) - g(x_i^*)]\Delta x$$

which gives the exact area. Recalling the definition of the definite integral we notice that

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

.

This formula of finding the area between two curves is sometimes known as applying integration with respect to the x-axis since the rectangles used to approximate the area have their bases lying parallel to the x-axis. It will be most useful when the two functions are of the form $y_1=f(x)$ and $y_2=g(x)$. Sometimes however, one may find it simpler to integrate with respect to the y-axis. This occurs when integrating with respect to the x-axis would result in more than one integral to be evaluated. These functions take the form $x_1=f(y)$ and $x_2=g(y)$ on the interval $[c,d]$. Note that $[c,d]$ are values of y. The derivation of this case is completely identical. Similar to before, we will assume that $f(y)\geq g(y)$ for all y on $[c,d]$. Now, as before we can divide the interval into n subintervals and create rectangles to approximate the area between $f(y)$ and $g(y)$. It may be useful to picture each rectangle having their 'width', Δy, parallel to the y-axis and 'height', $f(y_i^*)-g(y_i^*)$ at the point y_i^*, parallel to the x-axis. Following from the work above we may reason that an *approximation* of the area, A, between the two curves is

$$A := \sum_{i=1}^{n} [f(y_i^*) - g(y_i^*)]\Delta y$$

.

As before, we take the limit as n approaches infinity to arrive at

$$A = \lim_{n\to\infty} \sum_{i=1}^{n} [f(y_i^*) - g(y_i^*)]\Delta y$$

,

which is nothing more than a definite integral, so

$$A = \int_c^d [f(y) - g(y)]\, dy$$

.

Regardless of the form of the functions, we basically use the same formula.

Volume

In this section we will learn how to find the volume of a shape. The procedure is very similar to calculating the Area. The basic procedure is:

- Partition the shape in
- Calculate basal area of every partition
- Multiply by height of the partition
- Sum up all the volumes

So, given a function *f*(*x*) that gives us the basal area at a given height *x*, we can write it up as follows:

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

Now limit it to infinity:

$$\lim_{n\to\infty} \sum_{i=1}^{n} f(x_i)\Delta x$$

This is a Riemann's Sum, so we can rewrite it as:

$$\int_a^b f(x)\,dx$$

Examples

Calculate the volume of a square pyramid of base *b* and height *h*.

The basal shape is a square, and depends on the height *x* at which it is taken. For simplicity, we will consider an inverted pyramid, so that $f(x) = \left(\frac{b}{h}x\right)^2$. We integrate in the proper range (0 to h):

$$\int_0^h \left(\frac{b}{h}x\right)^2 dx = \left(\frac{b}{h}\right)^2 \int_0^h x^2\,dx = \left(\frac{b}{h}\right)^2 \frac{h^3}{3} = \frac{b^2 \cdot h}{3}$$

Volume of solids of revolution

Revolution solids

A solid is said to be of revolution when it is formed by rotating a given curve against an axis. For example, rotating a circle positioned at *(0,0)* against the *y-axis* would create a revolution solid, namely, a sphere.

Calculating the volume

Calculating the volume of this kind of solid is very similar to calculating any volume: we calculate the basal area, and then we integrate through the height of the volume.

Say we want to calculate the volume of the shape formed rotating over the *x-axis* the area contained between the curves $f(x)$ and $g(x)$ in the range [a,b]. First calculate the basal area:

$$| \pi f(x)^2 - \pi g(x)^2 |$$

And then integrate in the range [a,b]:

$$\int_a^b |\pi f(x)^2 - \pi g(x)^2| \, dx = \pi \int_a^b |f(x)^2 - g(x)^2| \, dx$$

Alternatively, if we want to rotate in the *y-axis* instead, f and g must be invertible in the range [a,b], and, following the same logic as before:

$$\pi \int_a^b |f^{-1}(x)^2 - g^{-1}(x)^2| \, dx$$

Arc length

Suppose that we are given a function f and we want to calculate the length of the curve drawn out by the graph of f. If the graph were a straight line this would be easy — the formula for the length of the line is given by Pythagoras' theorem. And if the graph were a polygon we can calculate the length by adding up the length of each piece.

The problem is that most graphs are not polygons. Nevertheless we can estimate the length of the curve by approximating it with straight lines. Suppose the curve C is given by the formula $y=f(x)$ for $a \le x \le b$. We divide the interval $[a,b]$ into n subintervals with equal width Δx and endpoints $x_0, x_1, \ldots, x_n$. Now let $y_i = f(x_i)$ so $P_i = (x_i,y_i)$ is the point on the curve above x_i. The length of the straight line between P_i and P_{i+1} is

$$|P_iP_{i+1}| = \sqrt{(y_{i+1} - y_i)^2 + (x_{i+1} - x_i)^2}.$$

So an estimate of the length of the curve C is the sum

$$\sum_{i=0}^{n-1} |P_iP_{i+1}|$$

As we divide the interval *[a,b]* into more pieces this gives a better estimate for the length of C. In fact we make that a definition.

Definition (Length of a Curve)

The length of the curve *y=f(x)* for $a \leq x \leq b$is defined to be

$$L = \lim_{n\to\infty} \sum_{i=0}^{n-1} |P_{i+1}P_i|.$$

The Arclength Formula

Suppose that *f'* is continuous on *[a,b]*. Then the length of the curve given by $y = f(x)$ between *a* and *b* is given by

$$L = \int_a^b \sqrt{1 + (f'(x))^2}dx$$

And in Leibniz notation

$$L = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2}dx$$

Proof: Consider $y_{i+1} - y_i = f(x_{i+1}) - f(x_i)$. By the Mean Value Theorem there is a point z_i in (x_{i+1},x_i) such that

$$y_{i+1} - y_i = f(x_{i+1}) - f(x_i) = f'(z_i)(x_{i+1} - x_i).$$

So

$$\begin{aligned} |P_iP_{i+1}| &= \sqrt{(y_{i+1} - y_i)^2 + (x_{i+1} - x_i)^2} \\ &= \sqrt{(f'(z_i))^2(x_{i+1} - x_i)^2 + (x_{i+1} - x_i)^2} \\ &= \sqrt{(1 + (f'(z_i))^2)(x_{i+1} - x_i)^2} \\ &= \sqrt{(1 + (f'(z_i))^2)}\Delta x. \end{aligned}$$

Putting this into the definiton of the length of *C* gives

$$L = \lim_{n\to\infty} \sum_{i=0}^{n-1} \sqrt{(1 + (f'(z_i))^2)}\Delta x.$$

Now this is the definition of the integral of the function $g(x) = \sqrt{1 + (f'(x))^2}$ between a and b (notice that g is continuous because we are assuming that f' is continuous). Hence

$$L = \int_a^b \sqrt{1 + (f'(x))^2}dx$$

as claimed.

Arclength of a parametric curve

For a parametric curve, that is, a curve defined by $x = f(t)$ and $y = g(t)$, the formula is slightly different:

$$L = \int_a^b \sqrt{(f'(t))^2 + (g'(t))^2}\,dt$$

Proof: The proof is analogous to the previous one: Consider $y_{i+1} - y_i = g(t_{i+1}) - g(t_i)$ and $x_{i+1} - x_i = f(t_{i+1}) - f(t_i)$. By the Mean Value Theorem there are points c_i and d_i in (t_{i+1},t_i) such that

$$y_{i+1} - y_i = g(t_{i+1}) - g(t_i) = g'(c_i)(t_{i+1} - t_i) \text{ and}$$
$$x_{i+1} - x_i = f(t_{i+1}) - f(t_i) = f'(d_i)(t_{i+1} - t_i).$$

So

$$\begin{aligned}|P_iP_{i+1}| &= \sqrt{(y_{i+1} - y_i)^2 + (x_{i+1} - x_i)^2} \\ &= \sqrt{(g'(c_i))^2(t_{i+1} - t_i)^2 + (f'(d_i))^2(t_{i+1} - t_i)^2} \\ &= \sqrt{(f'(d_i))^2) + (g'(c_i))^2)(t_{i+1} - t_i)^2} \\ &= \sqrt{(f'(d_i))^2 + (g'(c_i))^2}\Delta t.\end{aligned}$$

Putting this into the definiton of the length of the curve gives

$$L = \lim_{n\to\infty} \sum_{i=0}^{n-1} \sqrt{(f'(d_i))^2 + (g'(c_i))^2}\Delta t.$$

This is equivalent to:

$$L = \int_a^b \sqrt{(f'(t))^2 + (g'(t))^2}\, dt$$

Surface area

Suppose we are given a function *f* and we want to calculate the surface area of the function *f* rotated around a given line. The calculation of surface area of revolution is related to the arc length calculation.

If the function *f* is a straight line, other methods such as surface area formulas for cylinders and conical frustra can be used. However, if *f* is not linear, an integration technique must be used.

Recall the formula for the lateral surface area of a conical frustrum:

$$A = 2\pi r l$$

where *r* is the average radius and *l* is the slant height of the frustrum.

For *y=f(x)* and $a \leq x \leq b$, we divide *[a,b]* into subintervals with equal width Δx and endpoints $x_0, x_1, \ldots, x_n$. We map each point $y_i = f(x_i)$ to a conical frustrum of width Δx and lateral surface area A_i.

We can estimate the surface area of revolution with the sum

$$A = \sum_{i=0}^{n} A_i$$

As we divide *[a,b]* into smaller and smaller pieces, the estimate gives a better value for the surface area.

Definition (Surface of Revolution)

The surface area of revolution of the curve *y=f(x)* about a line for $a \leq x \leq b$ is defined to be

$$A = \lim_{n\to\infty} \sum_{i=0}^{n} A_i$$

The Surface Area Formula

Suppose *f* is a continuous function on the interval *[a,b]* and *r(x)* represents the distance from *f(x)* to the axis of rotation. Then the lateral surface area of revolution about a line is given by

$$A = 2\pi \int_a^b r(x)\sqrt{1 + (f'(x))^2}dx$$

And in Leibniz notation

$$A = 2\pi \int_a^b r(x)\sqrt{1 + \left(\frac{dy}{dx}\right)^2}dx$$

Proof:

$$\begin{aligned} A &= \lim_{n\to\infty} \sum_{i=1}^{n} A_i \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} 2\pi r_i l_i \\ &= 2\pi \lim_{n\to\infty} \sum_{i=1}^{n} r_i l_i \end{aligned}$$

As $n \to \infty$ and $\Delta x \to 0$, we know two things:

1. the average radius of each conical frustrum r_i approaches a single value

2. the slant height of each conical frustrum l_i equals an infitesmal segment of arc length

From the arc length formula discussed in the previous section, we know that

$$l_i = \sqrt{1 + (f'(x_i))^2}$$

Therefore

$$\begin{aligned} A &= 2\pi \lim_{n\to\infty} \sum_{i=1}^{n} r_i l_i \\ &= 2\pi \lim_{n\to\infty} \sum_{i=1}^{n} r_i \sqrt{1 + (f'(x_i))^2}\Delta x \end{aligned}$$

Because of the definition of an integral $\int_a^b f(x)dx = \lim_{n\to\infty}\sum_{i=1}^{n} f(c_i)\Delta x_i$, we can simplify the sigma operation to an integral.

$$A = 2\pi \int_a^b r(x)\sqrt{1+(f'(x))^2}dx$$

Or if *f* is in terms of *y* on the interval *[c,d]*

$$A = 2\pi \int_c^d r(y)\sqrt{1+(f'(y))^2}dy$$

Work

$$W = \int F\,dr = \int ma\,dr = \int m\frac{dv}{dt}dr = m\int \frac{dr}{dt}dv = m\int vdv = \frac{1}{2}mv^2 = \Delta E_k$$

Infinite Series

Series

An arithmetic series is the sum of a sequence of terms with a common difference. A geometric series is the sum of terms with a common ratio. For example, an interesting series which appears in many practical problems in science, engineering, and mathematics is the geometric series $r + r^2 + r^3 + r^4 + ...$ where the ... indicates that the series continues indefinitely. A common way to study a particular series (following Cauchy) is to define a sequence consisting of the sum of the first n terms. For example, to study the geometric series we can consider the sequence which adds together the first n terms:

$$S_n(r) = \sum_{i=1}^{n} r^i.$$

Generally by studying the sequence of partial sums we can understand the behavior of the entire infinite series.

Two of the most important questions about a series are

- Does it converge?
- If so, what does it converge to?

For example, it is fairly easy to see that for $r > 1$, the geometric series $S_n(r)$ will not converge to a finite number (i.e., it will diverge to infinity). To see this, note that each time we increase the number of terms in the series, $S_n(r)$ increases by r^{n+1}, since $r^{n+1} > 1$ for all $r > 1$ (as we defined), $S_n(r)$ must increase by a number greater than one every term. When increasing the sum by more than one for every term, it will diverge.

Perhaps a more surprising and interesting fact is that for $| r | < 1$, $S_n(r)$ will converge to a finite value. Specifically, it is possible to show that

$$\lim_{n\to\infty} S_n(r) = \frac{r}{1-r}.$$

Indeed, consider the quantity

$$(1-r)S_n(r) = (1-r)\sum_{i=1}^{n} r^n = \sum_{i=1}^{n} r^n - \sum_{i=2}^{n+1} r^n = r - r^{n+1}$$

Since $r^{n+1} \to 0$as $n \to \infty$for | *r* | < 1, this shows that $(1-r)S_n(r) \to r$as $n \to \infty$. The quantity 1 - *r* is non-zero and doesn't depend on *n* so we can divide by it and arrive at the formula we want.

We'd like to be able to draw similar conclusions about any series.

Unfortunately, there is no simple way to sum a series. The most we will be able to do in most cases is determine if it converges. The geometric and the telescoping series are the only types of series in which we can easily find the sum of.

Convergence

It is obvious that for a series to converge, the a_n must tend to zero (because sum of any infinite terms is infinity, except when the sequence approaches 0), but even if the limit of the sequence is 0, is not sufficient to say it converges.

Consider the harmonic series, the sum of 1/*n*, and group terms

$$\begin{array}{rcccccccccc} \sum_1^{2^m} \frac{1}{n} & = 1 + \frac{1}{2} + & \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + & \dots & + \sum_{1+2^{n-1}}^{2^n} \frac{1}{p} & & \\ & > \frac{3}{2} + & \frac{1}{4}2 + & \frac{1}{8}4 + & \dots & & + \frac{1}{2^n}2^{n-1} \\ & = \frac{3}{2} + & \frac{1}{2} + & \frac{1}{2} + & \dots & + \frac{1}{2} & (m \text{ terms}) \end{array}$$

As *m* tends to infinity, so does this final sum, hence the series diverges.

We can also deduce something about how quickly it diverges. Using the same grouping of terms, we can get an upper limit on the sum of the first so many terms, the *partial sums*.

$$1 + \frac{m}{2} \leq \sum_1^{2^m} \frac{1}{n} \leq 1 + m$$

or

$$1 + \frac{\ln_2 m}{m} \leq \sum_1^{m} \frac{1}{n} \leq 1 + \ln_2 m$$

and the partial sums increase like log *m*, very slowly.

Notice that to discover this, we compared the terms of the harmonic series with a series we knew diverged.

This is a *convergence test* (also known as the direct comparison test) we can apply to any pair of series.

- If b_n converges and $|a_n| \leq |b_n|$ then a_n converges.
- If b_n diverges and $|a_n| \geq |b_n|$ then a_n diverges.

There are many such tests, the most important of which we'll describe in this chapter.

Absolute convergence

Theorem: If the series of **absolute** values, $\sum_{n=1}^{\infty} |a_n|$, converges, then so does the series $\sum_{n=1}^{\infty} a_n$

We say such a series *converges absolutely*.

Proof:

Let $\varepsilon > 0$

According to the Cauchy criterion for series convergence, exists N so that for all $N < m,n$:

$$\sum_{k=n}^{m} |a_k| < \epsilon$$

We know that:

$$|\sum_{k=n}^{m} a_k| \leq \sum_{k=n}^{m} |a_k|$$

And then we get:

$$|\sum_{k=n}^{m} a_k| \leq \sum_{k=n}^{m} |a_k| < \epsilon$$

Now we get:

$$|\sum_{k=n}^{m} a_k| < \epsilon$$

Which is exactly the Cauchy criterion for series convergence.

Q.E.D

The converse does not hold. The series 1-1/2+1/3-1/4 ... converges, even though the series of its absolute values diverges.

A series like this that converges, but not absolutely, is said to *converge conditionally.*

If a series converges absolutely, we can add terms in any order we like. The limit will still be the same.

If a series converges conditionally, rearranging the terms changes the limit. In fact, we can make the series converge to any limit we like by choosing a suitable rearrangement.

E.g, in the series 1-1/2+1/3-1/4 ..., we can add only positive terms until the partial sum exceeds 100, subtract 1/2, add only positive terms until the partial sum exceeds 100, subtract 1/4, and so on, getting a sequence with the same terms that converges to 100.

This makes absolutely convergent series easier to work with. Thus, all but one of convergence tests in this chapter will be for series all of whose terms are positive, which must be absolutely convergent or divergent series. Other series will be studied by considering the corresponding series of absolute values.

Ratio test

For a series with terms a_n, all positive, if

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = r$$

then

- the series converges if $r<1$
- the series diverges if $r>1$
- the series could do either if $r=1$, the test is not conclusive in this case.

E.g, suppose

$$a_n = \frac{n!n!}{(2n)!}$$

then

$$\frac{a_{n+1}}{a_n} = \frac{(n+1)^2}{(2n+1)(2n+2)} = \frac{n+1}{4n+2} \to \frac{1}{4}$$

so this series converges.

Integral test

If $f(x)$ is a monotonically decreasing, always positive function, then the series

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

converges if *and only if* the integral

$$\int_1^{\infty} f(x)dx$$

converges.

E.g, consider $f(x)=1/x^p$, for a fixed p.

- If $p=1$ this is the harmonic series, which diverges.
- If $p<1$ each term is larger than the harmonic series, so it diverges.
- If $p>1$ then

$$\begin{aligned}\int_1^{\infty} x^{-p}dx &= \lim_{s\to\infty}\int_1^{s} x^{-p}dx \\ &= \lim_{s\to\infty}\frac{-1}{(p-1)x^{p-1}}\Big|_1^{s} \\ &= \lim_{s\to\infty}\left(\frac{1}{p-1}-\frac{1}{(p-1)s^{p-1}}\right) = \frac{1}{p-1}\end{aligned}$$

The integral converges, for $p>1$, so the series converges.

We can prove this test works by writing the integral as

$$\int_1^{\infty} f(x)dx = \sum_{n=1}^{\infty}\int_n^{n+1} f(x)dx$$

and comparing each of the integrals with rectangles, giving the inequalities

$$f(n) \geq \int_n^{n+1} f(x)dx \geq f(n+1)$$

Applying these to the sum then shows convergence.

Limit Comparison

- If b_n converges, and the limit

$$\lim_{n\to\infty} \frac{a_n}{b_n}$$

exists and is not zero, then a_n converges

- If c_n diverges, and

$$\lim \frac{|a_n|}{c_n} > 0$$

then a_n diverges

Example:

$$a_n = n^{-\frac{n+1}{n}}$$

For large *n*, the terms of this series are similar to, but smaller than, those of the harmonic series. We compare the limits.

$$\lim \frac{|a_n|}{c_n} = \lim \frac{n}{n^{\frac{n+1}{n}}} = \lim \frac{1}{n^{\frac{1}{n}}} = 1 > 0$$

so this series diverges.

Alternating series

If the signs of the a_n alternate,

$$a_n = (-1)^n |a_n|$$

then we call this an ***alternating*** *series.* The series sum converges provided that

$$\lim_{n\to\infty} a_n = 0 \text{ and } |a_{n+1}| < |a_n|.$$

The error in a partial sum of an alternating series is smaller than the first omitted term.

$$\left| \sum_{n=1}^{\infty} a_n - \sum_{n=1}^{m} a_n \right| < |a_{m+1}|$$

Geometric series

The geometric series can take either of the following forms

$$\sum_{n=0}^{\infty} ar^n \quad \text{or} \quad \sum_{n=1}^{\infty} ar^{n-1}$$

As you have seen at the start, the sum of the geometric series is

$$S_n = \frac{a}{1-r} \quad \text{for } |r| < 1$$

Telescoping series

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

Expanding (or "telescoping") this type of series is informative. If we expand this series, we get:

$$\sum_{n=0}^{k}(b_n - b_{n+1}) = (b_0 - b_1) + (b_1 - b_2) + ... + (b_{k-1} - b_k)$$

Additive cancellation leaves:

$$\sum_{n=0}^{k}(b_n - b_{n+1}) = b_0 - b_k$$

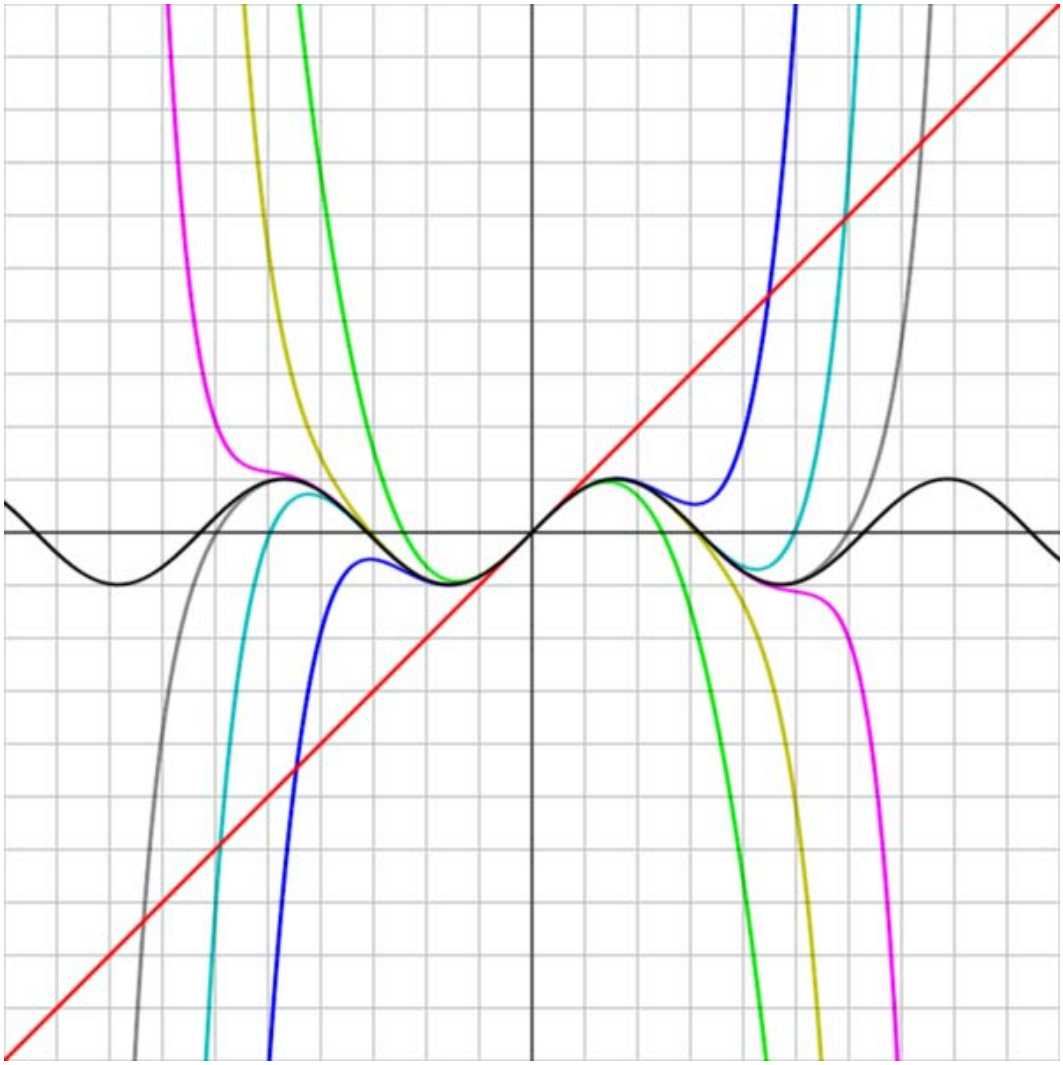
Thus,

$$\sum_{n=0}^{\infty}(b_n - b_{n+1}) = \lim_{k\to\infty}\sum_{n=0}^{k}(b_n - b_{n+1}) = \lim_{k\to\infty}(b_0 - b_k) = b_0 - \lim_{k\to\infty} b_k$$

and all that remains is to evaluate the limit.

There are other tests that can be used, but these tests are sufficient for all commonly encountered series.

Taylor Series



sin(x) and Taylor approximations, polynomials of degree 1, 3, 5, 7, 9, 11 and 13.

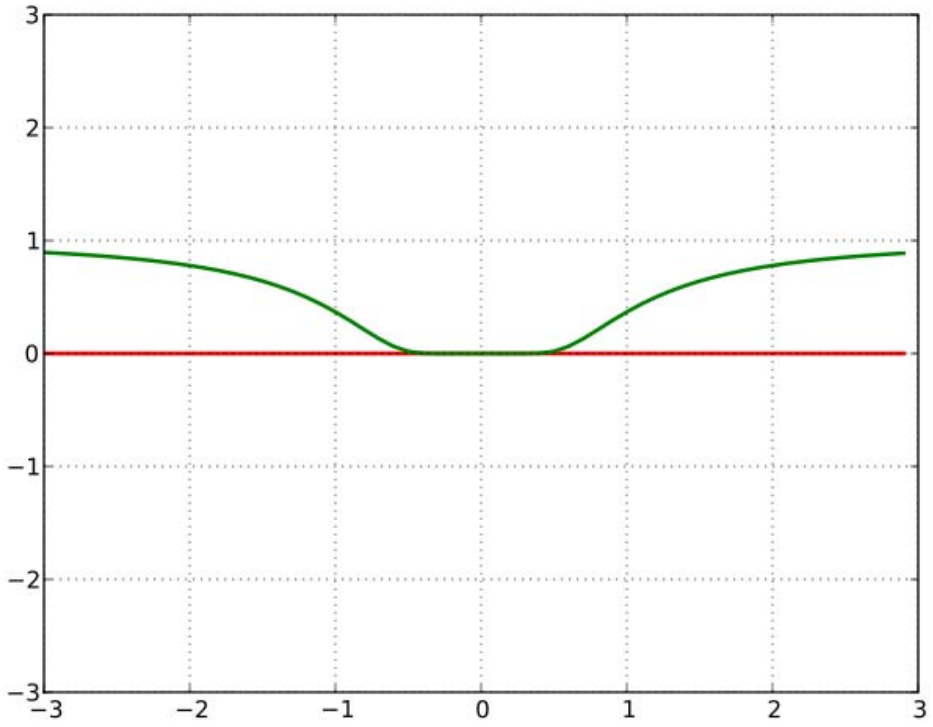
The **Taylor series** of an infinitely often differentiable real (or complex) function f defined on an open interval $(a-r, a+r)$ is the power series

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

Here, $n!$ is the factorial of n and $f^{(n)}(a)$ denotes the nth derivative of f at the point a. If this series converges for every x in the interval $(a-r, a+r)$ and the sum is equal to $f(x)$, then the function $f(x)$ is called **analytic**. To check whether the series converges towards $f(x)$, one normally uses estimates for the remainder term of Taylor's theorem. A function is analytic if and only if a power series converges to the function; the coefficients in that power series are then necessarily the ones given in the above Taylor series formula.

If $a = 0$, the series is also called a **Maclaurin series**.

The importance of such a power series representation is threefold. First, differentiation and integration of power series can be performed term by term and is hence particularly easy. Second, an analytic function can be uniquely extended to a holomorphic function defined on an open disk in the complex plane, which makes the whole machinery of complex analysis available. Third, the (truncated) series can be used to approximate values of the function near the point of expansion.



The function e^{-1/x^2} is not analytic: the Taylor series is 0, although the function is not.

Note that there are examples of infinitely often differentiable functions $f(x)$ whose Taylor series converge, but are *not* equal to $f(x)$. For instance, for the function defined piecewise by saying that $f(x) = \exp(-1/x^2)$ if $x \neq 0$ and $f(0) = 0$, all the derivatives are zero at $x = 0$, so the Taylor series of $f(x)$ is zero, and its radius of convergence is infinite, even though the function most definitely is not zero. This particular pathology does not afflict complex-valued functions of a complex variable. Notice that $\exp(-1/z^2)$ does not approach 0 as z approaches 0 along the imaginary axis.

Some functions cannot be written as Taylor series because they have a singularity; in these cases, one can often still achieve a series expansion if one allows also negative powers of the variable x; see Laurent series. For example, $f(x) = \exp(-1/x^2)$ can be written as a Laurent series.

The Parker-Sockacki theorem is a recent advance in finding Taylor series which are solutions to differential equations. This theorem is an expansion on the Picard iteration.

List of Taylor series

Several important Taylor series expansions follow. All these expansions are also valid for complex arguments *x*.

Exponential function and natural logarithm:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \quad \text{for all } x$$

$$\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n \quad \text{for } |x| < 1$$

Geometric series:

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \quad \text{for } |x| < 1$$

Binomial series:

$$(1+x)^\alpha = \sum_{n=0}^{\infty} C(\alpha, n) x^n \quad \text{for all } |x| < 1 \quad \text{and all complex } \alpha$$

Trigonometric functions:

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

$$\cos x = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n} \quad \text{for all } x$$

$$\tan x = \sum_{n=1}^{\infty} \frac{B_{2n}(-4)^n(1-4^n)}{(2n)!} x^{2n-1} \quad \text{for } |x| < \frac{\pi}{2}$$

$$\sec x = \sum_{n=0}^{\infty} \frac{(-1)^n E_{2n}}{(2n)!} x^{2n} \quad \text{for } |x| < \frac{\pi}{2}$$

$$\arcsin x = \sum_{n=0}^{\infty} \frac{(2n)!}{4^n (n!)^2 (2n+1)} x^{2n+1} \quad \text{for } |x| < 1$$

$$\arctan x = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} x^{2n+1} \quad \text{for } |x| < 1$$

Hyperbolic functions:

$$\sinh x = \sum_{n=0}^{\infty} \frac{1}{(2n+1)!} x^{2n+1} \quad \text{for all } x$$
$$\cosh x = \sum_{n=0}^{\infty} \frac{1}{(2n)!} x^{2n} \quad \text{for all } x$$
$$\tanh x = \sum_{n=1}^{\infty} \frac{B_{2n} 4^n (4^n - 1)}{(2n)!} x^{2n-1} \quad \text{for } |x| < \frac{\pi}{2}$$
$$\sinh^{-1} x = \sum_{n=0}^{\infty} \frac{(-1)^n (2n)!}{4^n (n!)^2 (2n+1)} x^{2n+1} \quad \text{for } |x| < 1$$
$$\tanh^{-1} x = \sum_{n=0}^{\infty} \frac{1}{2n+1} x^{2n+1} \quad \text{for } |x| < 1$$

Lambert's W function:

$$W_0(x) = \sum_{n=1}^{\infty} \frac{(-n)^{n-1}}{n!} x^n \quad \text{for } |x| < \frac{1}{e}$$

The numbers B_k appearing in the expansions of tan(*x*) and tanh(*x*) are the Bernoulli numbers. The C(α,*n*) in the binomial expansion are the binomial coefficients. The E_k in the expansion of sec(*x*) are Euler numbers.

Multiple dimensions

The Taylor series may be generalized to functions of more than one variable with

$$\sum_{n_1=0}^{\infty} \cdots \sum_{n_d=0}^{\infty} \frac{\partial^{n_1}}{\partial x^{n_1}} \cdots \frac{\partial^{n_d}}{\partial x^{n_d}} \frac{f(a_1, \cdots, a_d)}{n_1! \cdots n_d!} (x_1 - a_1)^{n_1} \cdots (x_d - a_d)^{n_d}$$

History

The Taylor series is named for mathematician Brook Taylor, who first published the power series formula in 1715.

Constructing a Taylor Series

Several methods exist for the calculation of Taylor series of a large number of functions. One can attempt to use the Taylor series as-is and generalize the form of the coefficients, or one can use manipulations such as substitution, multiplication or division, addition or subtraction of standard Taylor series (such as those above) to construct the Taylor series of a function, by virtue of Taylor series being power series. In some cases, one can also derive the Taylor series by repeatedly applying integration by parts. The use of computer

algebra systems to calculate Taylor series is common, since it eliminates tedious substitution and manipulation.

Example 1

Consider the function

$$f(x) = \ln(1 + \cos x),$$

for which we want a Taylor series at 0.

We have for the natural logarithm

$$\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \quad \text{for } |x| < 1$$

and for the cosine function

$$\cos x = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots \quad \text{for all } x \in \mathbb{C}.$$

We can simply substitute the second series into the first. Doing so gives

$$\left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots\right) - \frac{1}{2}\left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots\right)^2 + \frac{1}{3}\left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots\right)^3 - \cdots$$

Expanding by using multinomial coefficients gives the required Taylor series. Note that cosine and therefore *f* are even functions, meaning that $f(x) = f(-x)$, hence the coefficients of the odd powers x, x^3, x^5, x^7 and so on have to be zero and don't need to be calculated. The first few terms of the series are

$$\ln(1 + \cos x) = \ln 2 - \frac{x^2}{4} - \frac{x^4}{96} - \frac{x^6}{1440} - \frac{17x^8}{322560} - \frac{31x^{10}}{7257600} - \cdots$$

The general coefficient can be represented using Faà di Bruno's formula. However, this representation does not seem to be particularly illuminating and is therefore omitted here.

Example 2

Suppose we want the Taylor series at 0 of the function

$$g(x) = \frac{e^x}{\cos x}.$$

We have for the exponential function

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

and, as in the first example,

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots$$

Assume the power series is

$$\frac{e^x}{\cos x} = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \cdots$$

Then multiplication with the denominator and substitution of the series of the cosine yields

$$\begin{aligned} e^x &= (c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \cdots) \cos x \\ &= \left(c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 + \cdots\right) \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots\right) \\ &= c_0 - \frac{c_0}{2} x^2 + \frac{c_0}{4!} x^4 + c_1 x - \frac{c_1}{2} x^3 + \frac{c_1}{4!} x^5 + c_2 x^2 - \frac{c_2}{2} x^4 + \frac{c_2}{4!} x^6 + c_3 x^3 - \frac{c_3}{2} x^5 + \frac{c_3}{4!} x^7 + \cdots \end{aligned}$$

Collecting the terms up to fourth order yields

$$= c_0 + c_1 x + \left(c_2 - \frac{c_0}{2}\right) x^2 + \left(c_3 - \frac{c_1}{2}\right) x^3 + \left(c_4 + \frac{c_0}{4!} - \frac{c_2}{2}\right) x^4 + \cdots$$

Comparing coefficients with the above series of the exponential function yields the desired Taylor series

$$\frac{e^x}{\cos x} = 1 + x + x^2 + \frac{2x^3}{3} + \frac{x^4}{2} + \cdots$$

Power Series

The study of **power series** is aimed at investigating series which can approximate some function over a certain interval.

Motivations

Elementary calculus (differentiation) is used to obtain information on a line which touches a curve at one point (i.e. a tangent). This is done by calculating the gradient, or slope of the curve, at a single point. However, this does not provide us with reliable information on the curve's actual *value* at given points in a wider interval. This is where the concept of power series becomes useful.

An example

Consider the curve of $y = \cos(x)$, about the point $x = 0$. A naïve approximation would be the line $y = 1$. However, for a more accurate approximation, observe that $\cos(x)$ looks like an inverted parabola around $x = 0$ - therefore, we might think about which parabola could approximate the shape of $\cos(x)$ near this point. This curve might well come to mind:

$$y = \frac{1 - x^2}{2}$$

In fact, this is the best estimate for $\cos(x)$ which uses polynomials of degree 2 (i.e. a highest term of x^2) - but how do we know this is true? This is the study of power series: finding optimal approximations to functions using polynomials.

Definition

A *power series* is a series of the form

$$a_0x^0 + a_1x^1 + ... + a_nx^n$$

or, equivalently,

$$\sum_{j=0}^{n} a_j x^j$$

Radius of convergence

When using a power series as an alternative method of calculating a function's value, the equation

$$f(x) = \sum_{j=0}^{n} a_j x^j$$

can only be used to study $f(x)$ where the power series converges - this may happen for a finite range, or for all real numbers.

The size of the interval (around its center) in which the power series converges to the function is known as the *radius of convergence*.

An example

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n$$ (a geometric series)

this converges when $|x| < 1$, the range $-1 < x < +1$, so the radius of convergence - centered at 0 - is **1**. It should also be observed that at the *extremities* of the radius, that is where $x = 1$ and $x = -1$, the power series does not converge.

Another example

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

Using the ratio test, this series converges when the ratio of successive terms is less than one:

$$\lim_{n\to\infty} \left| \frac{x^{(n+1)}}{(n+1)!} \frac{n!}{x^n} \right| < 1$$

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

or $$\lim_{n\to\infty} \left| \frac{x}{n+1} \right| < 1$$

which is always true - therefore, this power series has an infinite radius of convergence. In effect, this means that the power series can *always* be used as a valid alternative to the original function, e^x.

Abstraction

If we use the ratio test on an arbitrary power series, we find it converges when

$$\lim \frac{|a_{n+1}x|}{|a_n|} < 1$$

and diverges when

$$\lim \frac{|a_{n+1}x|}{|a_n|} > 1$$

The radius of convergence is therefore

$$r = \lim \frac{|a_n|}{|a_{n+1}|}$$

If this limit diverges to infinity, the series has an infinite radius of convergence.

Differentiation and Integration

Within its radius of convergence, a power series can be differentiated and integrated term by term.

$$\frac{d}{dx}\sum_{j=0}^{\infty} a_j x^j = \sum_{j=0}^{\infty}(j+1)a_{j+1}x^j$$
$$\int \sum_{j=0}^{\infty} a_j z^j dz = \sum_{j=1}^{\infty} \frac{a_{j-1}}{j}x^j$$

Both the differential and the integral have the same radius of convergence as the original series.

This allows us to sum exactly suitable power series. For example,

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \dots$$

This is a geometric series, which converges for | *x* | < 1. Integrating both sides, we get

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} \dots$$

which will also converge for | *x* | < 1. When *x* = -1 this is the harmonic series, which *diverges'; when* x *= 1 this is an alternating series with diminishing terms, which* converges *to ln 2 - this is testing the extremities.*

It also lets us write power series for integrals we cannot do exactly such as the error function:

$$e^{-x^2} = \sum (-1)^n \frac{x^{2n}}{n!}$$

The left hand side can not be integrated exactly, but the right hand side can be.

$$\int_0^z e^{-x^2} dx = \sum \frac{(-1)^n z^{2n+1}}{(2n+1)n!}$$

This gives us a power series for the sum, which has an infinite radius of convergence, letting us approximate the integral as closely as we like.

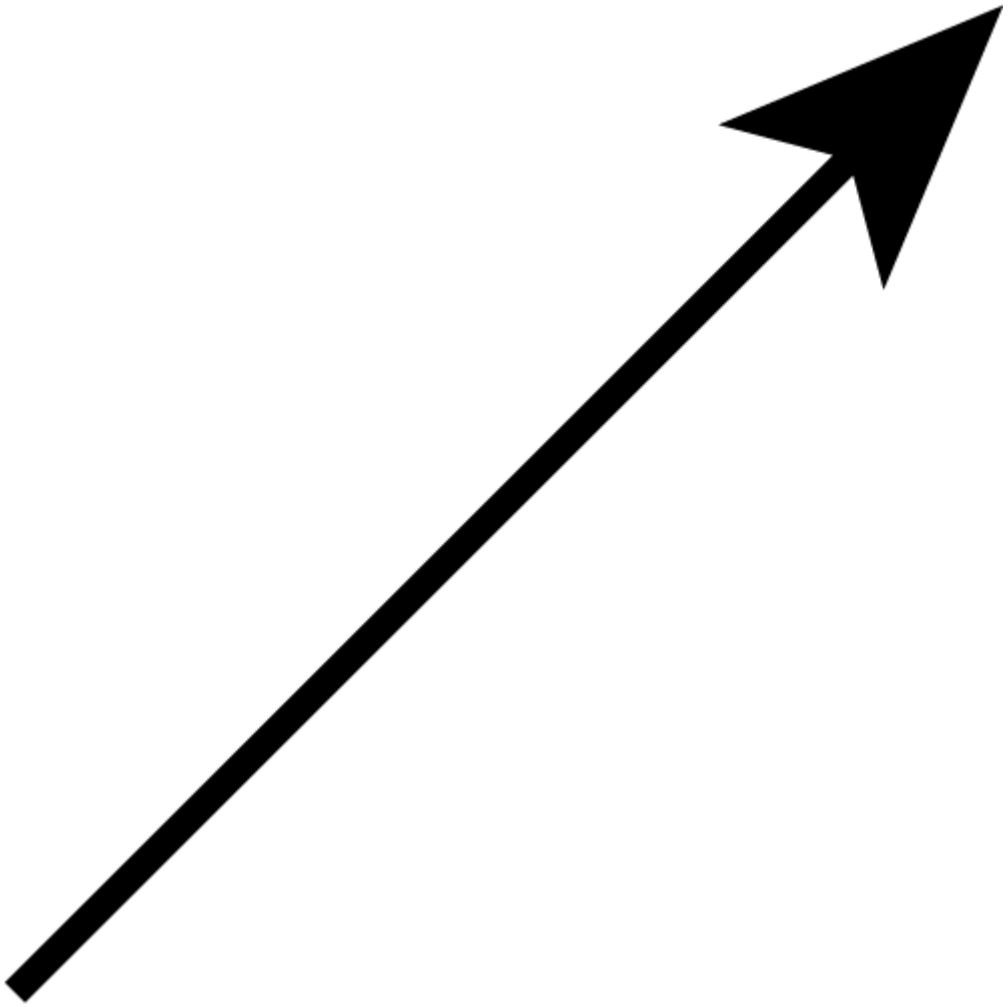
Multivariable & Differential Calculus

Two-Dimensional Vectors

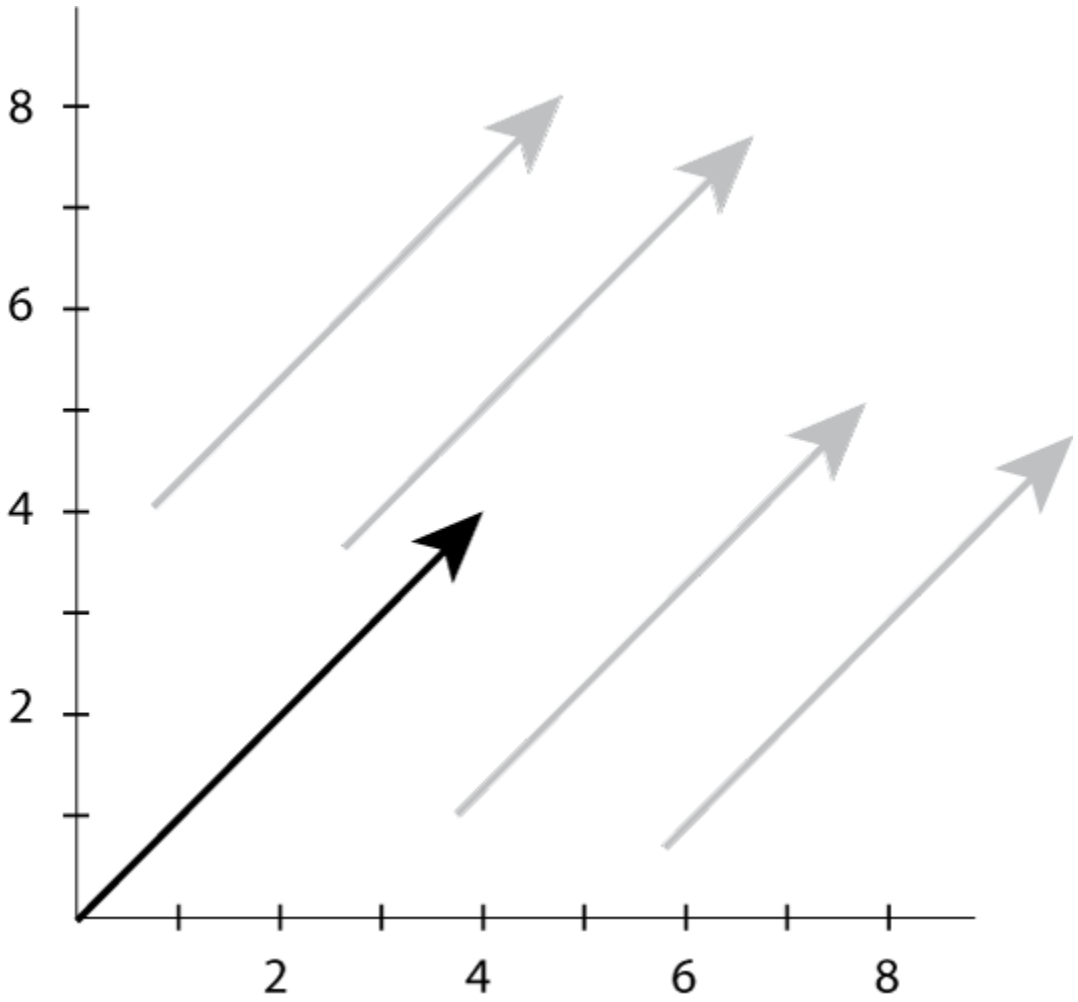
Introduction

In most mathematics courses up until this point, we deal with **scalars**. These are quantities which only need one number to express. For instance, the amount of gasoline used to drive to the grocery store is a scalar quantity because it only needs one number: 2 gallons.

In this unit, we deal with **vectors**. A vector is a **directed line segment** -- that is, a line segment that points one direction or the other. As such, it has an **initial point** and a **terminal point**. The vector starts at the initial point and ends at the terminal point, and the vector points towards the terminal point. A vector is drawn as a line segment with an arrow at the terminal point:



The same vector can be placed anywhere on the coordinate plane and still be the same vector -- the only two bits of information a vector represents are the **magnitude** and the **direction**. The magnitude is simply the length of the vector, and the direction is the angle at which it points. Since neither of these specify a starting or ending *location*, the same vector can be placed anywhere. To illustrate, all of the line segments below can be defined as the vector with magnitude $4\sqrt{2}$and angle 45 degrees:

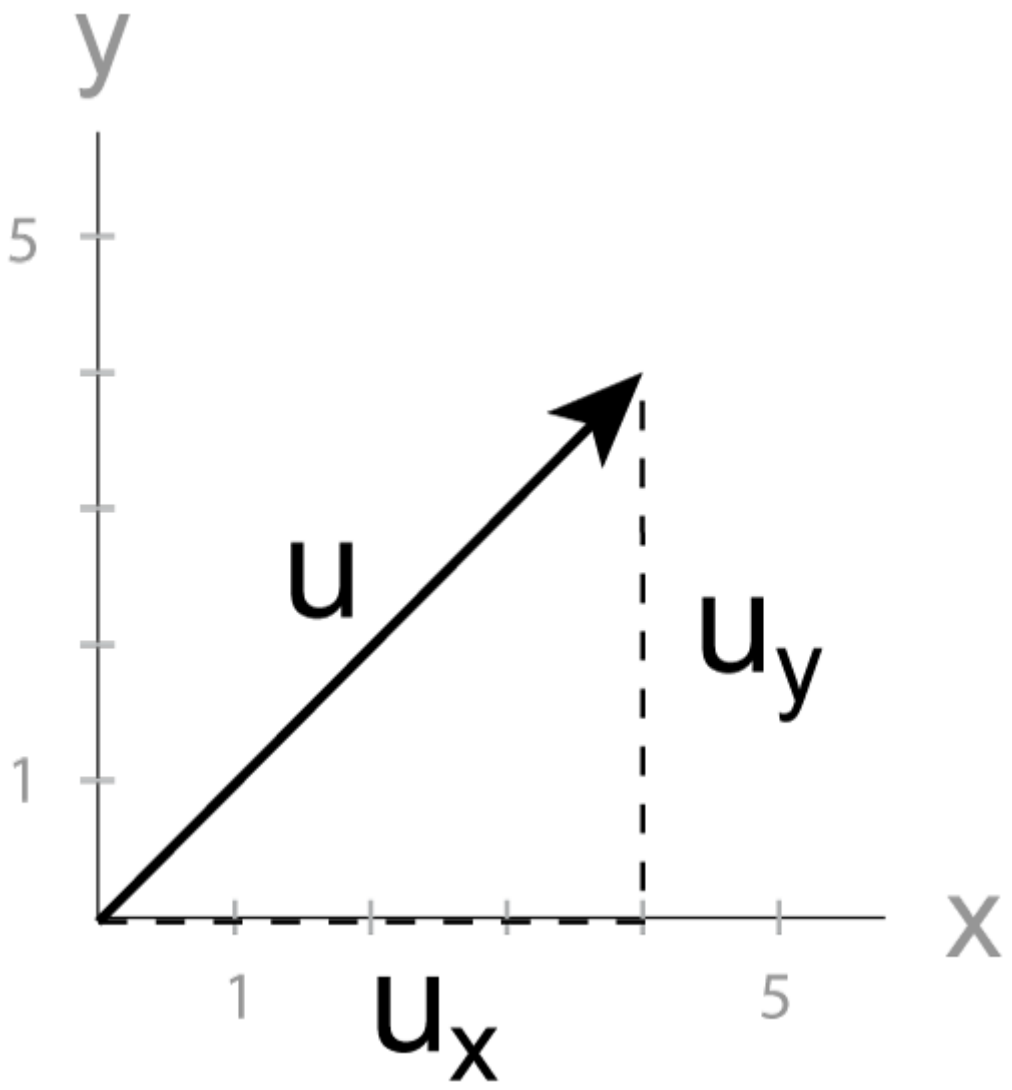


It is customary, however, to place the vector with the initial point at the origin as indicated by the black vector. This is called the **standard position**.

Component Form

In standard practice, we don't express vectors by listing the length and the direction. We instead use **component form**, which lists the height (rise) and width (run) of the vectors. It is written as follows:

$$\begin{pmatrix} \text{run} \\ \text{rise} \end{pmatrix}$$



From the diagram we can now see the benefits of the standard position: the two numbers for the terminal point's coordinates are the same numbers for the vector's rise and run. Note that we named this vector *u*. Just as you can assign numbers to variables in algebra (usually x, y, and z), you can assign vectors to variables in calculus. The letters u, v, and w are usually used, and either boldface or an arrow over the letter is used to identify it as a vector.

When expressing a vector in component form, it is no longer obvious what the magnitude and direction are. Therefore, we have to perform some calculations to find the magnitude and direction.

Magnitude

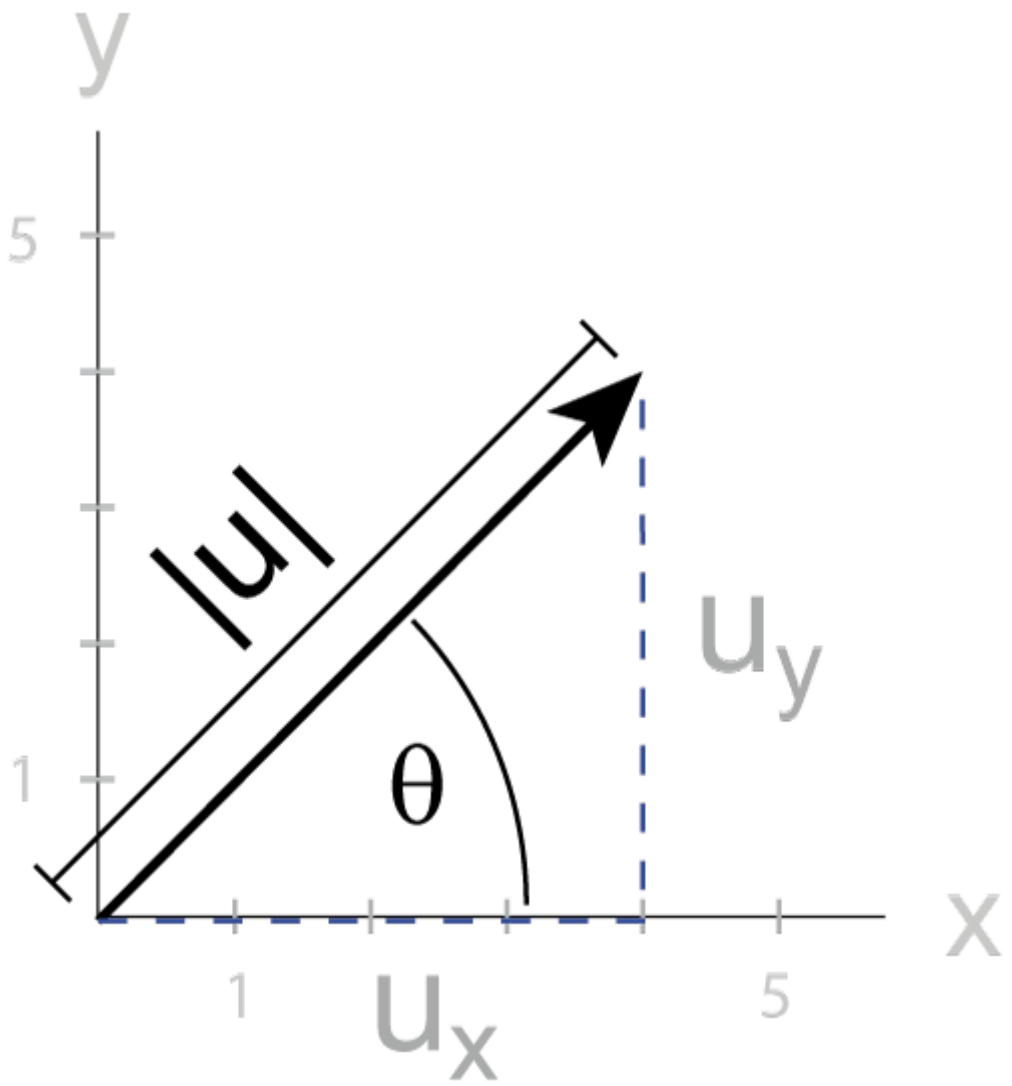
$$|\mathbf{u}| = \sqrt{u_x^2 + u_y^2}$$

where u_x is the width, or run, of the vector; u_y is the height, or rise, of the vector. You should recognize this formula as the Pythagorean theorem. It is -- the magnitude is the distance between the initial point and the terminal point.

The magnitude of a vector can also be called the norm.

Direction

$$\tan\theta = \frac{u_y}{u_x}$$



where θ is the direction of the vector. This formula is simply the tangent formula for right triangles.

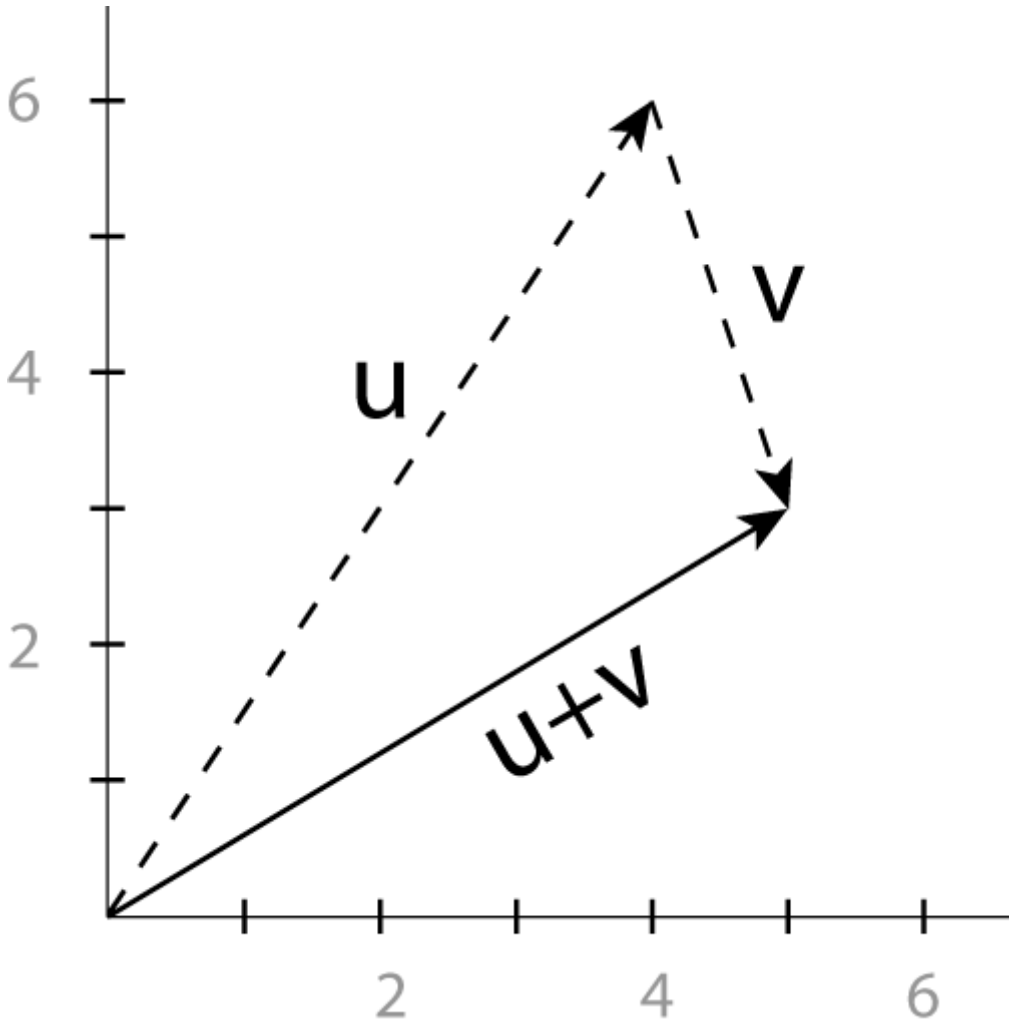
Vector Operations

For these definitions, assume:

$$\mathbf{u} = \begin{pmatrix} u_x \\ u_y \end{pmatrix} \mathbf{v} = \begin{pmatrix} v_x \\ v_y \end{pmatrix}$$

Vector Addition

Graphically, adding two vectors together places one vector at the end of the other. This is called *tip-to-tail* addition: The **resultant vector**, or solution, is the vector drawn from the initial point of the first vector to the terminal point of the second vector when they are drawn tip-to-tail:



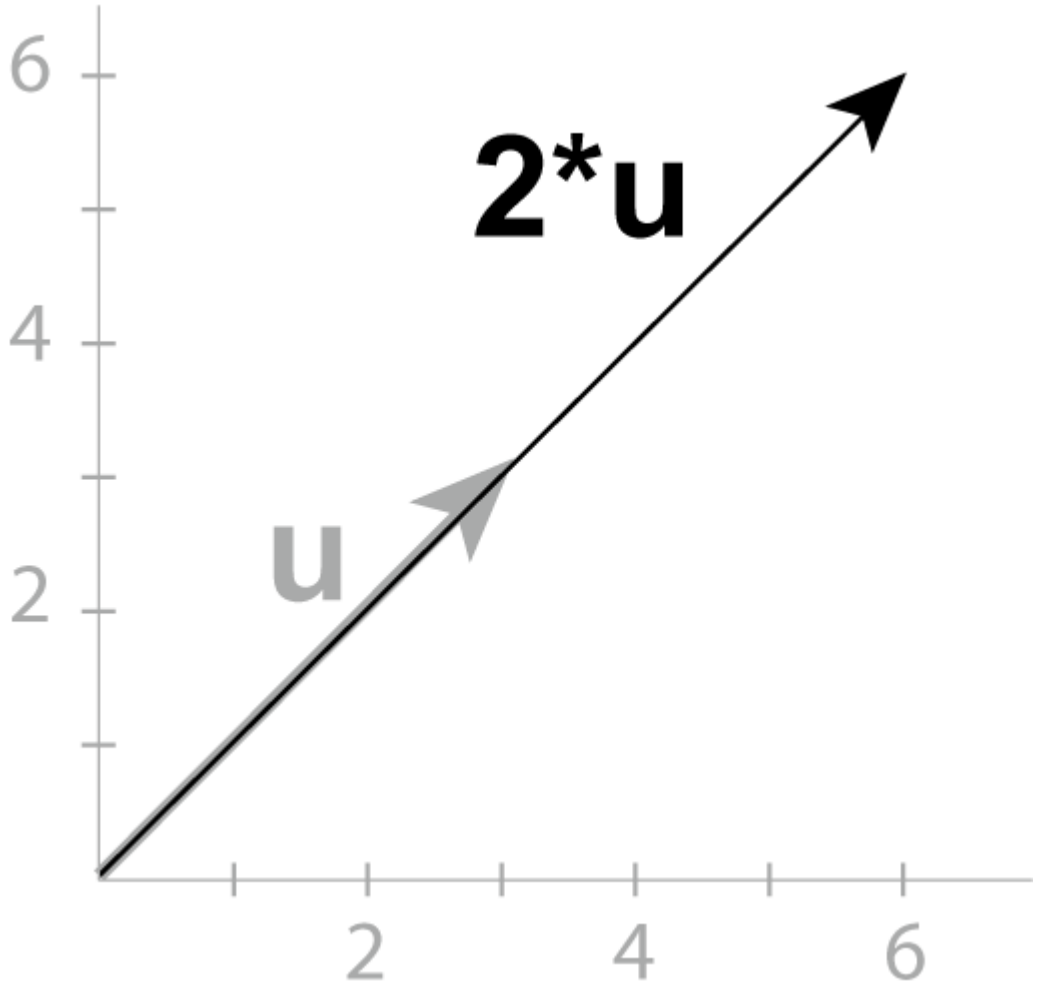
$$\begin{pmatrix} 4 \\ 6 \end{pmatrix} + \begin{pmatrix} 1 \\ -3 \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \end{pmatrix}$$

Numerically:

$$\mathbf{u} + \mathbf{v} = \begin{pmatrix} u_x + v_x \\ u_y + v_y \end{pmatrix}$$

Scalar Multiplication

Graphically, multiplying a vector by a scalar changes only the magnitude of the vector by that same scalar. That is, multiplying a vector by 2 will "stretch" the vector to twice its original magnitude, keeping the direction the same.



$$2 \cdot \begin{pmatrix} 3 \\ 3 \end{pmatrix} = \begin{pmatrix} 6 \\ 6 \end{pmatrix}$$

Numerically, you calculate the resultant vector with this formula:

$$c\mathbf{u} = \begin{pmatrix} cu_x \\ cu_y \end{pmatrix}$$, where c is a constant scalar.

As previously stated, the magnitude is changed by the same constant:

$$|c\mathbf{u}| = c|\mathbf{u}|$$

Since multiplying a vector by a constant results in a vector in the same direction, we can reason that two vectors are parallel if one is a constant multiple of the other -- that is, that $\mathbf{u} || \mathbf{v}$ if $\mathbf{u} = c\mathbf{v}$ for some constant c.

We can also divide by a non-zero scalar by instead multiplying by the reciprocal, as with dividing regular numbers:

$$\frac{\mathbf{u}}{c} = \frac{1}{c}\mathbf{u}, c \neq 0$$

Dot Product

The dot product, sometimes confusingly called the scalar *product*, of two vectors is given by:

$$\mathbf{u} \cdot \mathbf{v} = u_x v_x + u_y v_y$$

or which is equivalent to:

$$\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos\theta$$

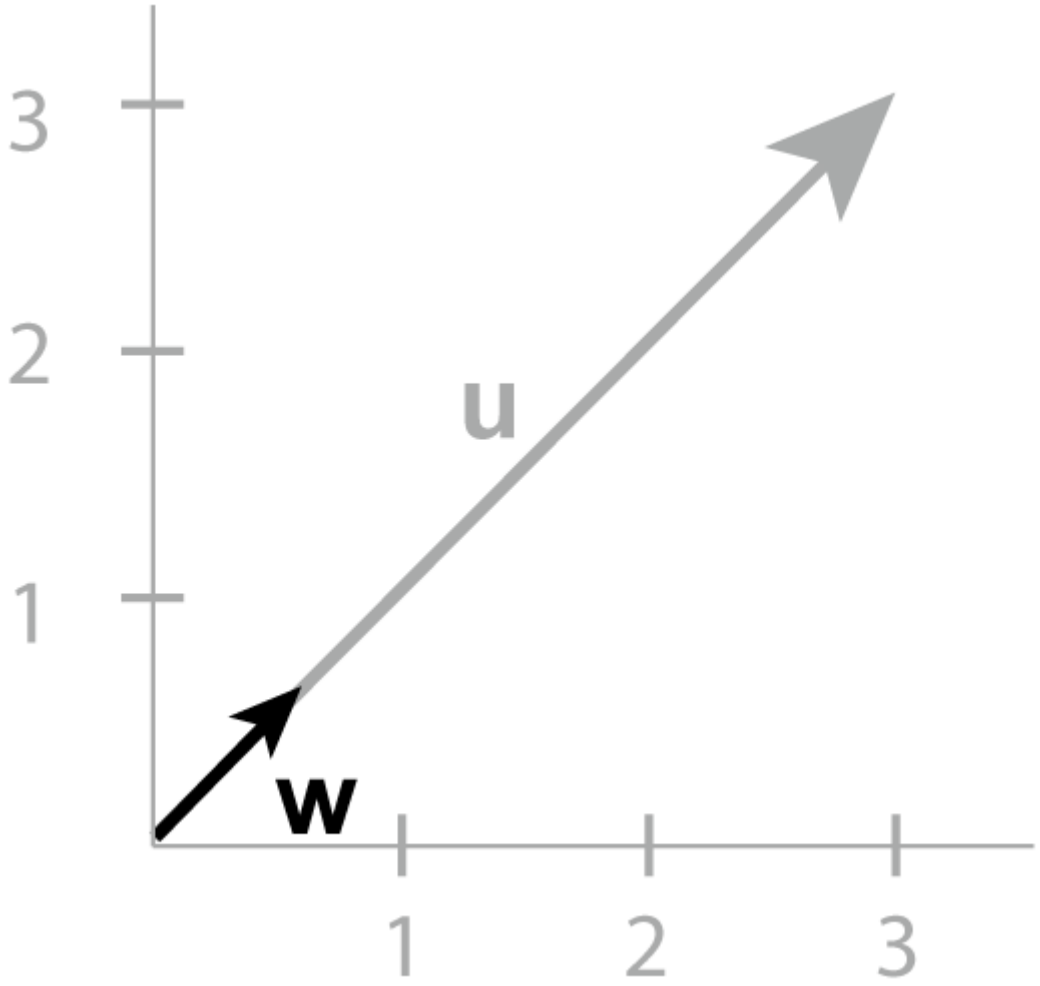
where θ is the angle difference between the two vectors. This provides a convenient way of finding the angle between two vectors:

$$\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|}$$

Applications of Scalar Multiplication and Dot Product

Unit Vectors

A **unit vector** is a vector with a magnitude of 1. The **unit vector of u** is a vector in the same direction as $\mathbf{u}$, but with a magnitude of 1:



The process of finding the unit vector of u is called **normalization**. As mentioned in **scalar multiplication**, multiplying a vector by constant C will result in the magnitude being multiplied by C. We know how to calculate the magnitude of $\mathbf{u}$. We know that dividing a vector by a constant will divide the magnitude by that constant. Therefore, if that constant is the magnitude, dividing the vector by the magnitude will result in a unit vector in the same direction as $\mathbf{u}$:

$$\mathbf{w} = \frac{\mathbf{u}}{|\mathbf{u}|}$$, where $\mathbf{w}$is the unit vector of $\mathbf{u}$

Standard Unit Vectors

A special case of *Unit Vectors* are the *Standard Unit Vectors* **i** and **j**: **i** points one unit directly right in the x direction, and **j** points one unit directly up in the y direction:

$$\mathbf{i} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\mathbf{j} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

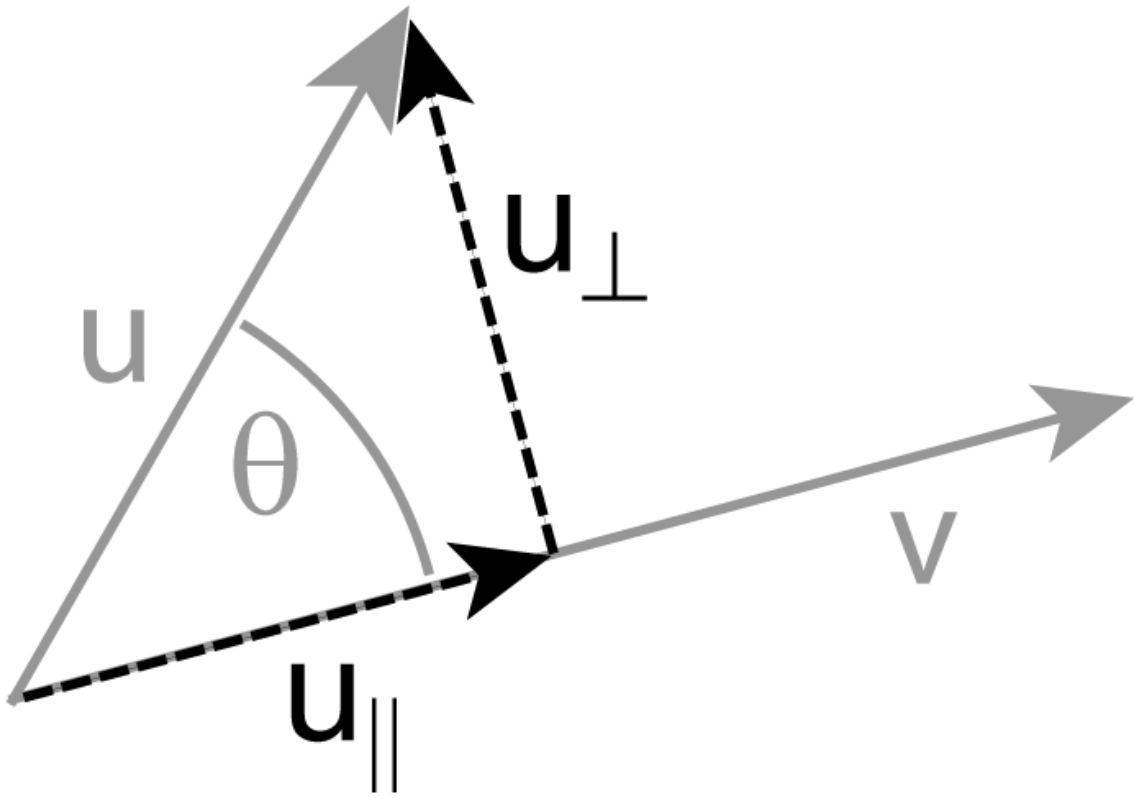
Using the scalar multiplication and vector addition rules, we can then express vectors in a different way:

$$\begin{pmatrix} x \\ y \end{pmatrix} = x\mathbf{i} + y\mathbf{j}$$

If we work that equation out, it makes sense. Multiplying x by **i** will result in the vector $\begin{pmatrix} x \\ 0 \end{pmatrix}$. Multiplying y by **j** will result in the vector $\begin{pmatrix} 0 \\ y \end{pmatrix}$. Adding these two together will give us our original vector, $\begin{pmatrix} x \\ y \end{pmatrix}$. Expressing vectors using i and j is called **standard form**.

Projection and Decomposition of Vectors

Sometimes it is necessary to decompose a vector $\mathbf{u}$ into two components: one component parallel to a vector $\mathbf{v}$, which we will call $\mathbf{u}_{\parallel}$; and one component perpendicular to it, $\mathbf{u}_{\perp}$.



Since the length of $\mathbf{u}_{\parallel}$is ($|\mathbf{u}| \cdot \cos\theta$), it is straightforward to write down the formulas for $\mathbf{u}_{\perp}$and $\mathbf{u}_{\parallel}$:

$$\mathbf{u}_{\parallel} = |\mathbf{u}| * \frac{(\mathbf{u} \cdot \mathbf{v})}{(|\mathbf{u}||\mathbf{v}|)} * \frac{\mathbf{v}}{|\mathbf{v}|} = (\mathbf{u} \cdot \mathbf{v})/(|\mathbf{v}|^2)\mathbf{v}$$

and

$$\mathbf{u}_{\perp} = \mathbf{u} - \mathbf{u}_{\parallel}$$

Length of a vector

The length of a vector is given by the dot product of a vector with itself, and θ = 0*deg*:

$$\mathbf{u} \cdot \mathbf{u} = |\mathbf{u}||\mathbf{u}| \cos\theta = |\mathbf{u}|^2$$

Perpendicular vectors

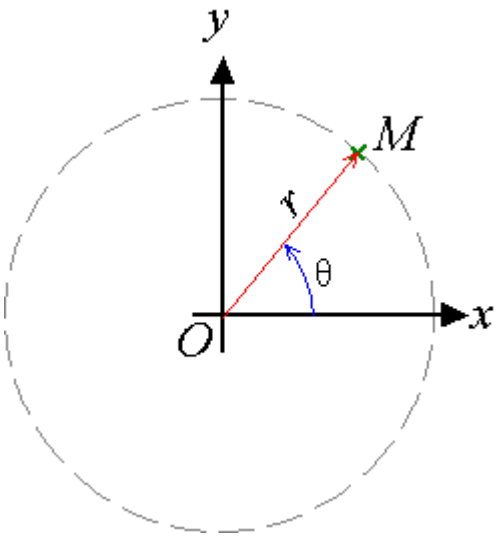
If the angle θ between two vectors is 90 degrees or $\frac{\pi}{2}$(if the two vectors are orthogonal to each other), that is the vectors are perpendicular, then the dot product is 0. This provides

us with an easy way to find a perpendicular vector: if you have a vector $\mathbf{u} = \begin{pmatrix} u_x \\ u_y \end{pmatrix}$, a perpendicular vector can easily be found by either

$$\mathbf{v} = \begin{pmatrix} -u_y \\ u_x \end{pmatrix} = -\begin{pmatrix} u_y \\ -u_x \end{pmatrix}$$

Polar coordinates

Polar coordinates are an alternative two-dimensional coordinate system, which is often useful when rotations are important. Instead of specifying the position along the *x* and *y* axes, we specify the distance from the origin, *r*, and the direction, an angle θ.



Looking at this diagram, we can see that the values of *x* and *y* are related to those of *r* and θ by the equations

$$\begin{aligned} x &= r\cos\theta \quad & r &= \sqrt{x^2 + y^2} \\ y &= r\sin\theta & \tan\theta &= \tfrac{y}{x} \end{aligned}$$

Because $\tan^{-1}$ is multivalued, care must be taken to select the right value.

Just as for Cartesian coordinates the unit vectors that point in the *x* and *y* directions are special, so in polar coordinates the unit vectors that point in the *r* and θ directions are also special.

We will call these vectors $\hat{\mathbf{r}}$ and $\hat{\boldsymbol{\theta}}$, pronounced r-hat and theta-hat. Putting a circumflex over a vector this way is often used to mean the unit vector in that direction.

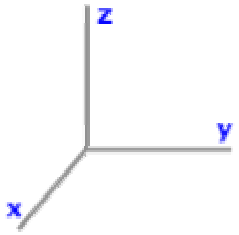
Again, on looking at the diagram we see,

$$\mathbf{i} = \hat{\mathbf{r}}\cos\theta - \hat{\boldsymbol{\theta}}\sin\theta \qquad \hat{\mathbf{r}} = \frac{x}{r}\mathbf{i} + \frac{y}{r}\mathbf{j}$$
$$\mathbf{j} = \hat{\mathbf{r}}\sin\theta + \hat{\boldsymbol{\theta}}\cos\theta \qquad \hat{\boldsymbol{\theta}} = -\frac{y}{r}\mathbf{i} + \frac{x}{r}\mathbf{j}$$

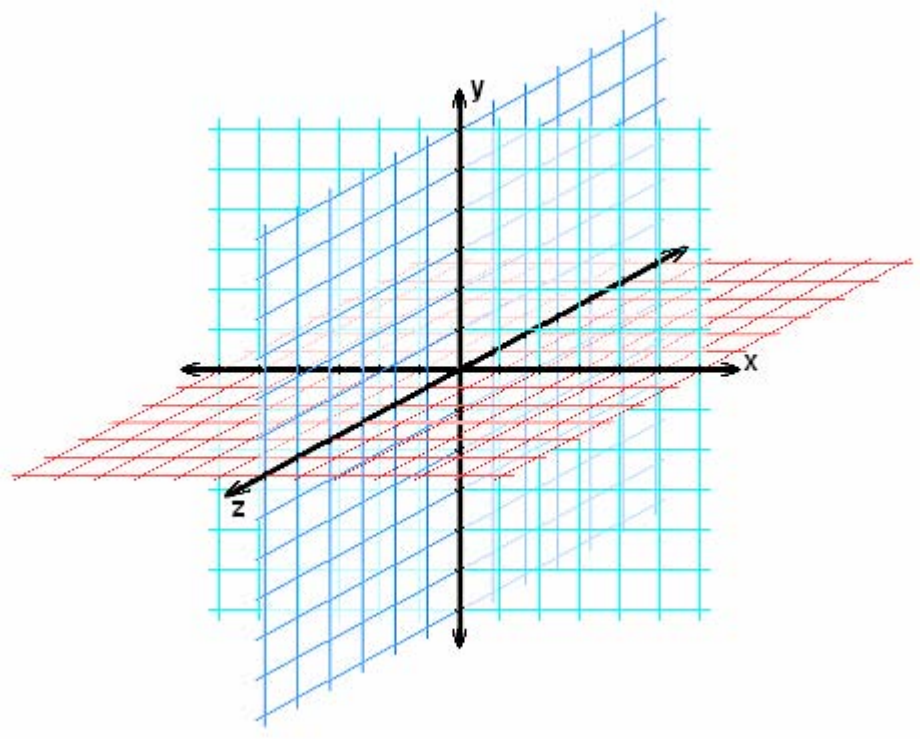
Three-Dimensional Coordinates and Vectors

Basic definition

Two-dimensional Cartesian coordinates as we've discussed so far can be easily extended to three-dimensions by adding one more value: 'z'. If the standard (x,y) coordinate axes are drawn on a sheet of paper, the 'z' axis would extend upwards off of the paper.



Similar to the two coordinate axes in two-dimensional coordinates, there are three **coordinate planes** in space. These are the **xy-plane**, the **yz-plane**, and the **xz-plane**. Each plane is the "sheet of paper" that contains both axes the name mentions. For instance, the yz-plane contains both the y and z axes and is perpendicular to the x axis.



Therefore, vectors can be extended to three dimensions by simply adding the 'z' value.

$$\mathbf{u} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

To facilitate standard form notation, we add another standard unit vector:

$$\mathbf{k} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Again, both forms (component and standard) are equivalent.

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = 1\mathbf{i} + 2\mathbf{j} + 3\mathbf{k}$$

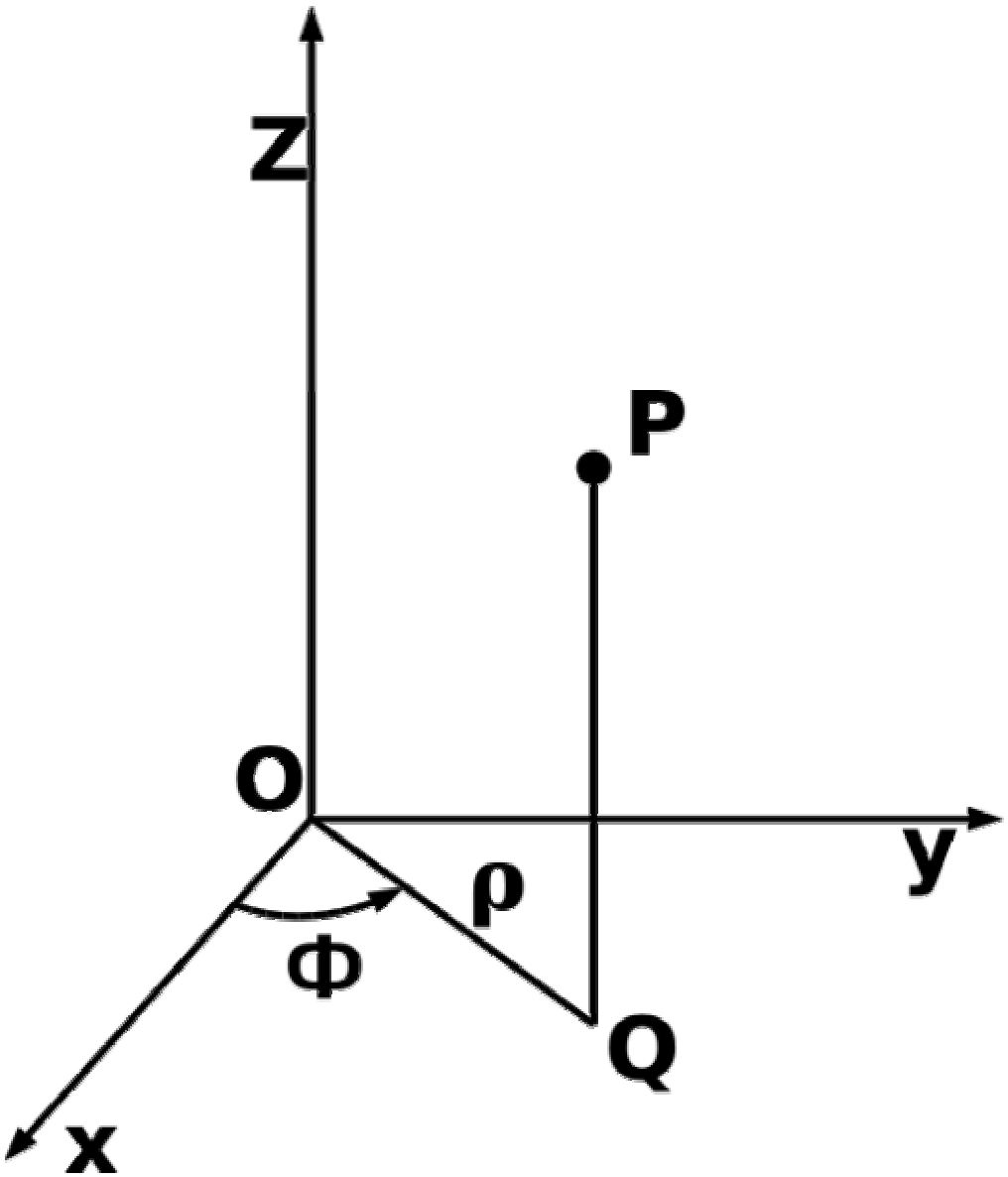
Magnitude: Magnitude in three dimensions is the same as in two dimensions, with the addition of a 'z' term in the radicand.

$$|\mathbf{u}| = \sqrt{u_x^2 + u_y^2 + u_z^2}$$

Three dimensions

The polar coordinate system is extended into three dimensions with two different coordinate systems, the cylindrical and spherical coordinate systems, both of which include two-dimensional or planar polar coordinates as a subset. In essence, the cylindrical coordinate system extends polar coordinates by adding an additional distance coordinate, while the spherical system instead adds an additional angular coordinate.

Cylindrical coordinates



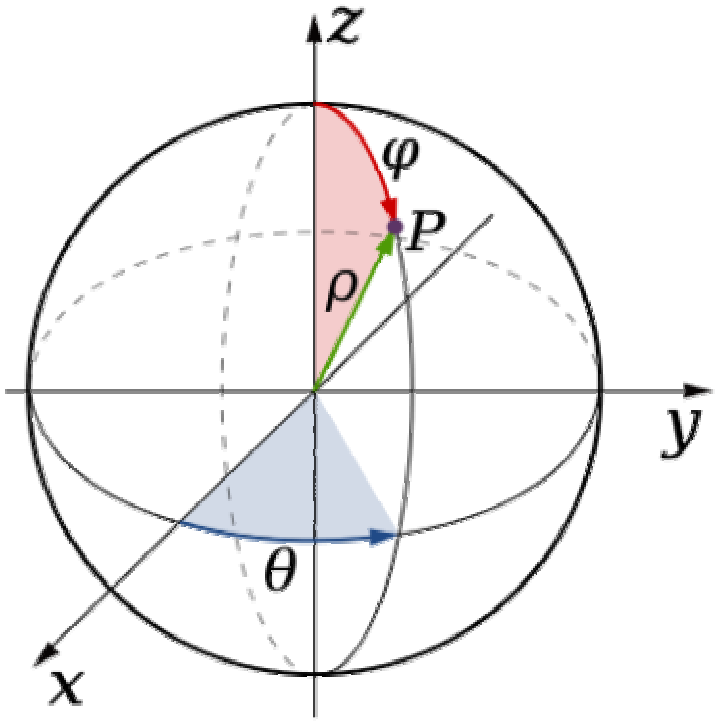
a point plotted with cylindrical coordinates

The *cylindrical coordinate system* is a coordinate system that essentially extends the two-dimensional polar coordinate system by adding a third coordinate measuring the height of a point above the plane, similar to the way in which the Cartesian coordinate system is extended into three dimensions. The third coordinate is usually denoted *h*, making the three cylindrical coordinates (*r*, θ, *h*).

The three cylindrical coordinates can be converted to Cartesian coordinates by

$$x = r\ \cos\theta$$
$$y = r\ \sin\theta$$
$$z = h.$$

Spherical coordinates



A point plotted using spherical coordinates

Polar coordinates can also be extended into three dimensions using the coordinates (ρ, φ, θ), where ρ is the distance from the origin, φ is the angle from the z-axis (called the colatitude or zenith and measured from 0 to 180°) and θ is the angle from the x-axis (as in the polar coordinates). This coordinate system, called the *spherical coordinate system*, is similar to the latitude and longitude system used for Earth, with the origin in the centre of Earth, the latitude δ being the complement of φ, determined by δ = 90° − φ, and the longitude *l* being measured by *l* = θ − 180°.

The three spherical coordinates are converted to Cartesian coordinates by

$$\begin{aligned} x &= \rho \sin\phi \cos\theta \\ y &= \rho \sin\phi \sin\theta \\ z &= \rho \cos\phi. \end{aligned}$$

$$\begin{aligned} r &= \sqrt{x^2 + y^2 + z^2}, \\ \theta &= \arctan\frac{y}{x}, \\ \phi &= \arccos\frac{z}{r}, \end{aligned}$$

Cross Product

The cross product of two vectors is a determinant:

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_x & u_y & u_z \\ v_x & v_y & v_z \end{vmatrix}$$

and is also a pseudovector.

The cross product of two vectors is orthogonal to both vectors. The magnitude of the cross product is the product of the magnitude of the vectors and the sin of the angle between them.

$$|\mathbf{u} \times \mathbf{v}| = |\mathbf{u}||\mathbf{v}| \sin\theta$$

This magnitude is the area of the parallelogram defined by the two vectors.

The cross product is *linear* and *anticommutative*. For any numbers *a* and *b*,

$$\mathbf{u} \times (a\mathbf{v} + b\mathbf{w}) = a\mathbf{u} \times \mathbf{v} + b\mathbf{u} \times \mathbf{w} \quad \mathbf{u} \times \mathbf{v} = -\mathbf{v} \times \mathbf{u}$$

If both vectors point in the same direction, their cross product is zero.

Triple Products

If we have three vectors we can combine them in two ways, a triple scalar product,

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})$$

and a triple vector product

$$\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$$

The triple scalar product is a determinant

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = \begin{vmatrix} u_x & u_y & u_z \\ v_x & v_y & v_z \\ w_x & w_y & w_z \end{vmatrix}$$

If the three vectors are listed clockwise, looking from the origin, the sign of this product is positive. If they are listed anticlockwise the sign is negative.

The order of the cross and dot products doesn't matter.

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = (\mathbf{u} \times \mathbf{v}) \cdot \mathbf{w}$$

Either way, the absolute value of this product is the volume of the parallelepiped defined by the three vectors, **u**, **v**, and **w**

The triple vector product can be simplified

$$\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = (\mathbf{u} \cdot \mathbf{w})\mathbf{v} - (\mathbf{u} \cdot \mathbf{v})\mathbf{w}$$

This form is easier to do calculations with.

The triple vector product is **not** associative.

$$\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) \neq (\mathbf{u} \times \mathbf{v}) \times \mathbf{w}$$

There are special cases where the two sides are equal, but in general the brackets matter. They must not be omitted.

Three-Dimensional Lines and Planes

We will use **r** to denote the position of a point.

The multiples of a vector, **a** all lie on a line through the origin. Adding a constant vector **b** will shift the line, but leave it straight, so the equation of a line is,

$$\mathbf{r} = \mathbf{a}s + \mathbf{b}$$

This is a *parametric equation*. The position is specified in terms of the parameter *s*.

Any linear combination of two vectors, **a** and **b** lies on a single plane through the origin, provided the two vectors are not colinear. We can shift this plane by a constant vector again and write

$$\mathbf{r} = \mathbf{a}s + \mathbf{b}t + \mathbf{c}$$

If we choose **a** and **b** to be *orthonormal* vectors in the plane (i.e unit vectors at right angles) then *s* and *t* are Cartesian coordinates for points in the plane.

These parametric equations can be extended to higher dimensions.

Instead of giving parametric equations for the line and plane, we could use constraints. E.g, for any point in the *xy* plane *z=0*

For a plane through the origin, the single vector normal to the plane, **n**, is at right angle with every vector in the plane, by definition, so

$$\mathbf{r} \cdot \mathbf{n} = 0$$

is a plane through the origin, normal to **n**.

For planes not through the origin we get

$$(\mathbf{r} - \mathbf{a}) \cdot \mathbf{n} = 0 \quad \mathbf{r} \cdot \mathbf{n} = a$$

A line lies on the intersection of two planes, so it must obey the constraint for both planes, i.e

$$\mathbf{r} \cdot \mathbf{n} = a \quad \mathbf{r} \cdot \mathbf{m} = b$$

These constraint equations can also be extended to higher dimensions.

Vector-Valued Functions

Vector-Valued Functions are functions that instead of giving a resultant scalar value, give a resultant vector value. These aid in the create of direction and vector fields, and are therefore used in physics to aid with visualizations of electric, magnetic, and many other fields. They are of the following form:

$$\mathbf{F(t)} = \begin{pmatrix} \mathbf{a_1(t)} \\ \mathbf{a_2(t)} \\ \mathbf{a_3(t)} \\ . \\ . \\ \mathbf{a_n(t)} \end{pmatrix}$$

Limits, Derivatives, and Integrals

Put simply, the limit of a vector-valued function is the limit of its parts.

Proof:

Suppose $$\lim_{t \to c} \mathbf{F}(t) = \mathbf{L} = \begin{pmatrix} \mathbf{a_1} \\ \mathbf{a_2} \\ \mathbf{a_3} \\ . \\ . \\ \mathbf{a_n} \end{pmatrix}$$

Therefore for any ε > 0 there is a φ > 0 such that

$$0 < |t - c| < \phi \implies |\mathbf{F}(t) - \mathbf{L}| < \epsilon$$

But by the triangle inequality

$$|a_1| \leq |\mathbf{F}| \leq |a_1| + |a_2| + |a_3| + ... + |a_n \| a_1(t) - a_1| \leq |\mathbf{F}(t) - \mathbf{L}|$$

So

$$0 < |t - c| < \phi \implies |a_1(t) - a_1| < \epsilon$$

Therefore $\lim_{t \to c} a_1(t) = a_1$ A similar argument can be used through parts a_n(t)

$$\lim_{t\to c}\mathbf{F}(t)=\mathbf{L}=\begin{pmatrix}\mathbf{a_1}\\ \mathbf{a_2}\\ \mathbf{a_3}\\ .\\ .\\ \mathbf{a_n}\end{pmatrix}$$

Now let again, and that for any ε>0 there is a corresponding φ>0 such 0<|t-c|<φ implies

$$|a_n(t)-a_n|<\frac{\epsilon}{n}$$

Then

$$0<|t-c|<\phi \implies |\mathbf{F}(t)-\mathbf{L}|\leq \frac{\epsilon_1}{n}+...+\frac{\epsilon_n}{n}=\epsilon$$

therefore:

$$\lim_{t\to c}\mathbf{F}(t)=\mathbf{L}=\begin{pmatrix}\mathbf{a_1}\\ \mathbf{a_2}\\ \mathbf{a_3}\\ .\\ .\\ \mathbf{a_n}\end{pmatrix}=\begin{pmatrix}\lim_{t\to c}\mathbf{a_1(t)}\\ \lim_{t\to c}\mathbf{a_2(t)}\\ \lim_{t\to c}\mathbf{a_3(t)}\\ .\\ .\\ \lim_{t\to c}\mathbf{a_n(t)}\end{pmatrix}$$

From this we can then create an accurate definition of a derivative of a vector-valued function:

$$\mathbf{F'}(t)=\lim_{h\to 0}\frac{\mathbf{F}(t+h)-\mathbf{F}(t)}{h}=\begin{pmatrix}\mathbf{a_1(t)}\\ \mathbf{a_2(t)}\\ \mathbf{a_3(t)}\\ .\\ .\\ \mathbf{a_n(t)}\end{pmatrix}$$

$$= \lim_{h \to 0} \frac{\begin{pmatrix} \mathbf{a_1(t+h)} \\ \mathbf{a_2(t+h)} \\ \mathbf{a_3(t+h)} \\ \cdot \\ \cdot \\ \mathbf{a_n(t+h)} \end{pmatrix} - \begin{pmatrix} \mathbf{a_1(t)} \\ \mathbf{a_2(t)} \\ \mathbf{a_3(t)} \\ \cdot \\ \cdot \\ \mathbf{a_n(t)} \end{pmatrix}}{h}$$

$$= \begin{pmatrix} \lim_{h \to 0} \frac{a_1(t+h) - a_1(t)}{h} \\ \lim_{h \to 0} \frac{a_2(t+h) - a_2(t)}{h} \\ \lim_{h \to 0} \frac{a_3(t+h) - a_3(t)}{h} \\ \cdot \\ \cdot \\ \lim_{h \to 0} \frac{a_n(t+h) - a_n(t)}{h} \end{pmatrix}$$

The final step was accomplished by taking what we just did with limits.

By the Fundamental Theorem of Calculus integrals can be applied to the vector's components.

In other words: the limit of a vector function is the limit of its parts, the derivative of a vector function is the derivative of its parts, and the integration of a vector function is the integration of it parts.

Velocity, Acceleration, Curvature, and a brief mention of the Binormal

Assume we have a vector-valued function which starts at the origin and as its independent variables changes the points that the vectors point at trace a path.

We will call this vector $\mathbf{r}(t)$, which is commonly known as the **position vector**.

If $\mathbf{r}$then represents a position and t represents time, then in model with Physics we know the following:

$\mathbf{r}(t+h) - \mathbf{r}(t)$is displacement. $\mathbf{r}'(t) = \mathbf{v}(t)$where $\mathbf{v}(t)$is the velocity vector. $|\mathbf{v}(t)|$is the speed. $\mathbf{r}''(t) = \mathbf{v}'(t) = \mathbf{a}(t)$where $\mathbf{a}(t)$is the acceleration vector.

The only other vector that comes in use at times is known as the curvature vector.

The vector $\mathbf{T}(t)$ used to find it is known as the unit tangent vector, which is defined as $\frac{\mathbf{v}(t)}{|\mathbf{v}(t)|}$ or shorthand $\hat{\mathbf{v}}$.

The vector normal $\mathbf{N}$ to this then is $\frac{\mathbf{T}'(t)}{|\mathbf{v}(t)|}$.

We can verify this by taking the dot product

$$\mathbf{T} \cdot \mathbf{N} = 0$$

Also note that $|\mathbf{v}(t)| = \frac{ds}{dt}$

and

$$\mathbf{T}(t) = \frac{v}{|v|} = \frac{\frac{dr}{dt}}{\frac{ds}{dt}} = \frac{dr}{ds}$$

and

$$\mathbf{N} = \frac{\mathbf{T}'(t)}{|\mathbf{v}(t)|} = \frac{\frac{dT}{dt}}{\frac{ds}{dt}} = \frac{dT}{ds}$$

Then we can actually verify:

$$\frac{d}{ds}(\mathbf{T} \cdot \mathbf{T}) = \frac{d}{ds}(1)$$

$$\frac{dT}{ds} \cdot \mathbf{T} + \mathbf{T} \cdot \frac{dT}{ds} = 0$$

$$2 * \mathbf{T} \cdot \frac{dT}{ds} = 0$$

$$\mathbf{T} \cdot \frac{dT}{ds} = 0$$

$$\mathbf{T} \cdot \mathbf{N} = 0$$

Therefore $\mathbf{N}$ is perpendicular to $\mathbf{T}$

What this gives rise to is the **Unit Normal Vector** $\frac{\frac{dT}{ds}}{|\frac{dT}{ds}|}$ of which the top-most vector is the Normal vector, but the bottom half $(|\frac{dT}{ds}|)^{-1}$ is known as the curvature. Since the Normal vector points toward the inside of a curve, the sharper a turn, the Normal vector has a large magnitude, therefore the curvature has a small value, and is used as an index in civil engineering to reflect the sharpness of a curve (clover-leaf highways, for instance).

The only other thing not mentioned is the Binormal that occurs in 3-d curves $\mathbf{T} \times \mathbf{N} = \mathbf{B}$, which is useful in creating planes parallel to the curve. < Calculus/Outline

In your previous study of calculus, we have looked at functions and their behavior. Most of these functions we have examined have been all in the form

$$f(x) : \mathbf{R} \to \mathbf{R},$$

and only occasional examination of functions of two variables. However, the study of functions of *several* variables is quite rich in itself, and has applications in several fields.

We write functions of vectors - many variables - as follows:

$$\boldsymbol{f} : \mathbf{R}^m \to \mathbf{R}^n$$

and $\boldsymbol{f}(\boldsymbol{x})$ for the function that maps a vector in $\mathbf{R}^m$ to a vector in $\mathbf{R}^n$.

Before we can do calculus in $\mathbf{R}^n$, we must familiarize ourselves with the structure of $\mathbf{R}^n$. We need to know which properties of $\mathbf{R}$ can be extended to $\mathbf{R}^n$

Topology in $\mathbf{R}^n$

We are already familiar with the nature of the regular real number line, which is the set $\mathbf{R}$, and the two-dimensional plane, $\mathbf{R}^2$. This examination of *topology* in $\mathbf{R}^n$ attempts to look at a generalization of the nature of *n*-dimensional spaces - $\mathbf{R}$, or $\mathbf{R}^{23}$, or $\mathbf{R}^n$.

Lengths and distances

If we have a vector in $\mathbf{R}^2$, we can calculate its length using the Pythagorean theorem. For instance, the length of the vector (2, 3) is

$$\sqrt{3^2 + 2^2} = \sqrt{13}$$

We can generalize this to $\mathbf{R}^n$. We define a vector's length, written $|\boldsymbol{x}|$, as the square root of the sum of the squares of each of its components. That is, if we have a vector $\boldsymbol{x}=(x_1,...,x_n)$,

$$|\mathbf{x}| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

Now that we have established some concept of length, we can establish the distance between two vectors. We define this distance to be the length of the two vectors' difference. We write this distance $d(\boldsymbol{x}, \boldsymbol{y})$, and it is

$$d(\mathbf{x},\mathbf{y}) = |\mathbf{x} - \mathbf{y}| = \sqrt{\sum (x_i - y_i)^2}$$

This distance function is sometimes referred to as a *metric*. Other metrics arise in different circumstances. The metric we have just defined is known as the *Euclidean* metric.

Open and closed balls

In **R**, we have the concept of an *interval*, in that we choose a certain number of other points about some central point. For example, the interval [-1, 1] is centered about the point 0, and includes points to the left and right of zero.

In $\mathbf{R}^2$ and up, the idea is a little more difficult to carry on. For $\mathbf{R}^2$, we need to consider points to the left, right, above, and below a certain point. This may be fine, but for $\mathbf{R}^3$ we need to include points in more directions.

We generalize the idea of the interval by considering all the points that are a given, fixed distance from a certain point - now we know how to calculate distances in $\mathbf{R}^n$, we can make our generalization as follows, by introducing the concept of an *open ball* and a *closed ball* respectively, which are analogous to the open and closed interval respectively.

an *open ball*
$B(\mathbf{a}, r)$
is a set in the form $\{ \boldsymbol{x} \in \mathbf{R}^n | d(\boldsymbol{x}, \boldsymbol{a}) < r\}$
an *closed ball*
$\overline{B}(\mathbf{a}, r)$
is a set in the form $\{ \boldsymbol{x} \in \mathbf{R}^n | d(\boldsymbol{x}, \boldsymbol{a}) \leq r\}$

In **R**, we have seen that the open ball is simply an open interval centered about the point $x=a$. In $\mathbf{R}^2$ this is a circle with no boundary, and in $\mathbf{R}^3$ it is a sphere with no outer surface. (*What would the closed ball be?*)

Boundary points

If we have some area, say a field, then the common sense notion of the *boundary* is the points 'next to' both the inside and outside of the field. For a set, S, we can define this rigorously by saying the boundary of the set contains all those points such that we can find points both inside and outside the set. We call the set of such points ∂S

Typically, when it exists the dimension of ∂S is one lower than the dimension of S. e.g, the boundary of a volume is a surface and the boundary of a surface is a curve.

This isn't always true; but it is true of all the sets we will be using.

A set *S* is *bounded* if there is some positive number such that we can encompass this set by a closed ball about **0**. --> if every point in it is within a finite distance of the origin, i.e there exists some $r>0$ such that **x** is in S implies $|\mathbf{x}|<r$.

Curves and parameterizations

If we have a function $\boldsymbol{f} : \mathbf{R} \to \mathbf{R}^n$, we say that $\boldsymbol{f}$'s image (the set $\{\boldsymbol{f}(t) \mid t \in \mathbf{R}\}$ - or some subset of **R**) is a *curve* in $\mathbf{R}^n$ and $\boldsymbol{f}$ is its parametrization.

Parameterizations are not necessarily unique - for example, $\boldsymbol{f}(t) = (\cos t, \sin t)$ such that $t \in [0, 2\pi)$ is one parametrization of the unit circle, and $\boldsymbol{g}(t) = (\cos at, \sin at)$ such that $t \in [0, 2\pi/a)$ is a whole family of parameterizations of that circle.

Collision and intersection points

Say we have two different curves. It may be important to consider

- when the two curves cross each other - where they *intersect*
- when the two curves hit each other at the same time - where they *collide*.

Intersection points

Firstly, we have two parameterizations $\boldsymbol{f}(t)$ and $\boldsymbol{g}(t)$, and we want to find out when they intersect, this means that we want to know when the function values of each parametrization are the same. This means that we need to solve

$$\boldsymbol{f}(t) = \boldsymbol{g}(s)$$

because we're seeking the function values independent of the times they intersect.

For example, if we have $\boldsymbol{f}(t) = (t, 3t)$ and $\boldsymbol{g}(t) = (t, t^2)$, and we want to find intersection points:

$$\boldsymbol{f}(t) = \boldsymbol{g}(s)$$
$$(t, 3t) = (s, s^2)$$

$t = s$ and $3t = s^2$

with solutions $(t, s) = (0, 0)$ and $(3, 3)$

So, the two curves intersect at the points (0, 0) and (3, 3).

Collision points

However, if we want to know when the points "collide", with $\boldsymbol{f}(t)$ and $\boldsymbol{g}(t)$, we need to know when both the function values *and* the times are the same, so we need to solve instead

$$\boldsymbol{f}(t) = \boldsymbol{g}(t)$$

For example, using the same functions as before, $\boldsymbol{f}(t) = (t, 3t)$ and $\boldsymbol{g}(t) = (t, t^2)$, and we want to find collision points:

$\boldsymbol{f}(t) = \boldsymbol{g}(t)$
$(t, 3t) = (t, t^2)$
$t = t$ and $3t = t^2$

which gives solutions $t = 0, 3$ So the collision points are (0, 0) and (3, 9).

We may want to do this to actually model physical problems, such as in ballistics.

Continuity and differentiability

If we have a parametrization $\boldsymbol{f} : \mathbf{R} \to \mathbf{R}^n$, which is built up out of *component functions* in the form $\boldsymbol{f}(t) = (f_1(t),...,f_n(t))$, $\boldsymbol{f}$ is continuous if and only if each component function is also.

In this case the derivative of $\boldsymbol{f}(t)$ is

$a_i = (f_1'(t),...,f_n'(t))$. This is actually a specific consequence of a more general fact we will see later.

Tangent vectors

Recall in single-variable calculus that on a curve, at a certain point, we can draw a line that is tangent to that curve at exactly at that point. This line is called a *tangent*. In the several variable case, we can do something similar.

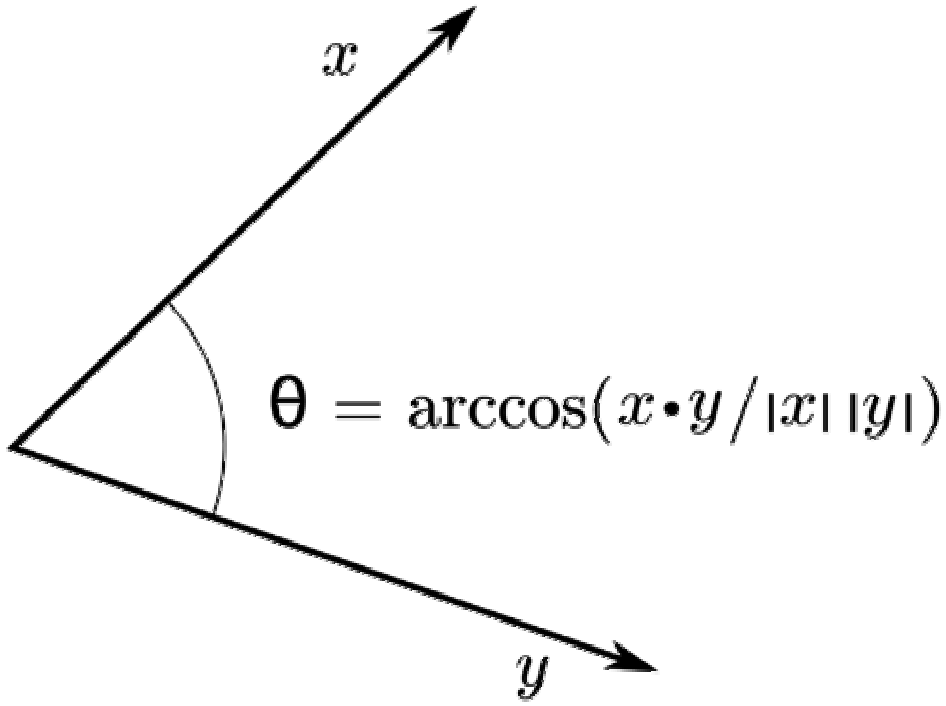
We can expect the *tangent vector* to depend on $\boldsymbol{f}'(t)$ and we know that a line is its own tangent, so looking at a parametrised line will show us precisely how to define the tangent vector for a curve.

An arbitrary line is $\mathbf{f}(t)=\mathbf{a}t+\mathbf{b}$, with $:f_i(t)=a_it+b_i$, so

$f_i'(t)=a_i$ and
$f'(t)=\mathbf{a}$, which is the direction of the line, its tangent vector.

Similarly, for any curve, the tangent vector is $f'(t)$.

Angle between curves



We can then formulate the concept of the *angle* between two curves by considering the angle between the two tangent vectors. If two curves, parametrized by $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ intersect at some point, which means that

$\boldsymbol{f}_1(s)=\boldsymbol{f}_2(t)=c$,

the angle between these two curves at c is the angle between the tangent vectors $\boldsymbol{f}_1'(s)$ and $\boldsymbol{f}_2'(t)$ is given by

$$\arccos \frac{\mathbf{f}_1'(s) \cdot \mathbf{f}_2'(t)}{|\mathbf{f}_1'(s)||\mathbf{f}_2'(t)|}$$

Tangent lines

With the concept of the tangent vector as being analogous to being the gradient of the line in the one variable case, we can form the idea of the *tangent line*. Recall that we need a point on the line and its direction.

If we want to form the tangent line to a point on the curve, say $\boldsymbol{p}$, we have the direction of the line $\boldsymbol{f}'(\boldsymbol{p})$, so we can form the tangent line

$$\boldsymbol{x}(t)=\boldsymbol{p}+t\,\boldsymbol{f}'(\boldsymbol{p})$$

Different parameterizations

One such parametrization of a curve is not necessarily unique. Curves can have several different parametrizations. For example, we already saw that the unit circle can be parametrized by $\boldsymbol{g}(t) = (\cos at, \sin at)$ such that $t \in [0, 2\pi/a)$.

Generally, if $\boldsymbol{f}$ is one parametrization of a curve, and $\boldsymbol{g}$ is another, with

$$\boldsymbol{f}(t_0) = \boldsymbol{g}(s_0)$$

there is a function $u(t)$ such that $u(t_0)=s_0$, and $\boldsymbol{g}'(u(t)) = \mathbf{f}(t)$ *near* t_0.

This means, in a sense, the function $u(t)$ "speeds up" the curve, but keeps the curve's shape.

Surfaces

A surface in space can be described by the image of a function $\boldsymbol{f} : \boldsymbol{R}^2 \rightarrow \boldsymbol{R}^n$. **f** *is said to be the parametrization of that surface.*

For example, consider the function

$$\boldsymbol{f}(\alpha, \beta) = \alpha(2,1,3)+\beta(-1,2,0)$$

This describes an infinite plane in $\mathbf{R}^3$. If we restrict α and β to some domain, we get a parallelogram-shaped surface in $\mathbf{R}^3$.

Surfaces can also be described explicitly, as the graph of a function $z = f(x, y)$ which has a standard parametrization as $\boldsymbol{f}(x,y)=(x, y, f(x,y))$, or implictly, in the form $f(x, y, z)=c$.

Level sets

The concept of the *level set* (or *contour*) is an important one. If you have a function $f(x, y, z)$, a level set in $\mathbf{R}^3$ is a set of the form $\{(x,y,z)|f(x,y,z)=c\}$. Each of these level sets is a surface.

Level sets can be similarly defined in any $\mathbf{R}^n$

Level sets in two dimensions may be familiar from maps, or weather charts. Each line represents a level set. For example, on a map, each contour represents all the points where the height is the same. On a weather chart, the contours represent all the points where the air pressure is the same.

Limits and continuity

Before we can look at derivatives of multivariate functions, we need to look at how limits work with functions of several variables first, just like in the single variable case.

If we have a function $\boldsymbol{f} : \mathbf{R}^m \to \mathbf{R}^n$, we say that $\boldsymbol{f}(\boldsymbol{x})$ approaches $\boldsymbol{b}$ (in $\mathbf{R}^n$) as $\boldsymbol{x}$ approaches $\boldsymbol{a}$ (in $\mathbf{R}^m$) if, for all positive ε, there is a corresponding positive number δ, $|\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{b}| < \varepsilon$ whenever $|\boldsymbol{x}-\boldsymbol{a}| < \delta$, with $\boldsymbol{x} \neq \boldsymbol{a}$.

This means that by making the difference between $\boldsymbol{x}$ and $\boldsymbol{a}$ smaller, we can make the difference between $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{b}$ as small as we want.

If the above is true, we say

- $\boldsymbol{f}(\boldsymbol{x})$ has *limit* $\boldsymbol{b}$ at $\boldsymbol{a}$
- $\lim_{\mathbf{x}\to\mathbf{a}} \mathbf{f}(\mathbf{x}) = \mathbf{b}$
- $\boldsymbol{f}(\boldsymbol{x})$ approaches $\boldsymbol{b}$ as $\boldsymbol{x}$ approaches $\boldsymbol{a}$
- $\boldsymbol{f}(\boldsymbol{x}) \to \boldsymbol{b}$ as $\boldsymbol{x} \to \boldsymbol{a}$

These four statements are all equivalent.

Rules

Since this is an almost identical formulation of limits in the single variable case, many of the limit rules in the one variable case are the same as in the multivariate case.

For $\boldsymbol{f}$ and $\boldsymbol{g}$, mapping $\mathbf{R}^m$ to $\mathbf{R}^n$, and $h(\boldsymbol{x})$ a scalar function mapping $\mathbf{R}^m$ to $\mathbf{R}$, with

- $\boldsymbol{f}(\boldsymbol{x}) \to \boldsymbol{b}$ as $\boldsymbol{x} \to \boldsymbol{a}$
- $\boldsymbol{g}(\boldsymbol{x}) \to \boldsymbol{c}$ as $\boldsymbol{x} \to \boldsymbol{a}$

- $h(\boldsymbol{x}) \to H$ as $\boldsymbol{x} \to \boldsymbol{a}$

then:

- $\lim_{\mathbf{x}\to\mathbf{a}} (\mathbf{f} + \mathbf{g}) = \mathbf{b} + \mathbf{c}$
- $\lim_{\mathbf{x}\to\mathbf{a}} (h\mathbf{f}) = H\mathbf{b}$

and consequently

- $\lim_{\mathbf{x}\to\mathbf{a}} (\mathbf{f} \cdot \mathbf{g}) = \mathbf{b} \cdot \mathbf{c}$
- $\lim_{\mathbf{x}\to\mathbf{a}} (\mathbf{f} \times \mathbf{g}) = \mathbf{b} \times \mathbf{c}$

when H≠0

- $\lim_{\mathbf{x}\to\mathbf{a}} (\frac{\mathbf{f}}{h}) = \frac{\mathbf{b}}{H}$

Continuity

Again, we can use a similar definition to the one variable case to formulate a definition of continuity for multiple variables.

If $\boldsymbol{f} : \mathbf{R}^m \to \mathbf{R}^n$, f is continuous at a point $\boldsymbol{a}$ in $\mathbf{R}^m$ if $\boldsymbol{f}(\boldsymbol{a})$ is defined and

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

Just as for functions of one dimension, if $\boldsymbol{f}$, $\boldsymbol{g}$ are both continuous at $\boldsymbol{p}$, $\boldsymbol{f}+\boldsymbol{g}$, $\lambda\boldsymbol{f}$ (for a scalar λ), $\boldsymbol{f}\cdot\boldsymbol{g}$, and $\boldsymbol{f}\times\boldsymbol{g}$ are continuous also. If $\varphi : \mathbf{R}^m \to \mathbf{R}$ is continus at $\boldsymbol{p}$, $\varphi\boldsymbol{f}$, $\boldsymbol{f}/\varphi$ are too if φ is never zero.

From these facts we also have that if A is some matrix which is $n\times m$ in size, with $\boldsymbol{x}$ in $\mathbf{R}^m$, a function $\boldsymbol{f}(\boldsymbol{x})=A\,\boldsymbol{x}$ is continuous in that the function can be expanded in the form $x_1\boldsymbol{a}_1+...+x_m\boldsymbol{a}_m$, which can be easily verified from the points above.

If $\mathbf{f} : \mathbf{R}^m \to \mathbf{R}^n$ which is in the form $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}),...,f_n(\boldsymbol{x})$ is continuous if and only if each of its component functions are a polynomial or rational function, whenever they are defined.

Finally, if $\boldsymbol{f}$ is continuous at $\boldsymbol{p}$, $\boldsymbol{g}$ is continuous at $\boldsymbol{f}(\boldsymbol{p})$, $\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))$ is continuous at $\boldsymbol{p}$.

Special note about limits

It is important to note that we can approach a point *in more than one direction*, and thus, the direction that we approach that point counts in our evaluation of the limit. It may be the case that a limit may exist moving in one direction, but not in another.

Differentiable functions

We will start from the one-variable definition of the derivative at a point *p*, namely

$$\lim_{x \to p} \frac{f(x) - f(p)}{x - p} = f'(p)$$

Let's change above to equivalent form of

$$\lim_{x \to p} \frac{f(x) - f(p) - f'(p)(x - p)}{x - p} = 0$$

which achieved after pulling f'(p) inside and putting it over a common denominator.

We can't divide by vectors, so this definition can't be immediately extended to the multiple variable case. Nonetheless, we don't have to: the thing we took interest in was the quotient of two small distances (magnitudes), not their other properties (like sign). It's worth noting that 'other' property of vector neglected is its direction. Now we can divide by the absolute value of a vector, so lets rewrite this definition in terms of absolute values

$$\lim_{x \to p} \frac{|f(x) - f(p) - f'(p)(x - p)|}{|x - p|} = 0$$

Another form of formula above is obtained by letting $h = x - p$ we have $x = p + h$ and if $x \to p$, the $h = x - p \to 0$, so

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

,

where *h* can be thought of as a 'small change'.

So, how can we use this for the several-variable case?

If we switch all the variables over to vectors and replace the constant (which performs a linear map in one dimension) with a matrix (which denotes also a linear map), we have

$$\lim_{\mathbf{x} \to \mathbf{p}} \frac{|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{p}) - \mathbf{A}(\mathbf{x} - \mathbf{p})|}{|\mathbf{x} - \mathbf{p}|} = 0$$

or

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{|\mathbf{f}(\mathbf{p}+\mathbf{h}) - \mathbf{f}(\mathbf{p}) - \mathbf{A}\mathbf{h}|}{|\mathbf{h}|} = 0$$

If this limit exists for some $\boldsymbol{f} : \mathbf{R}^m \to \mathbf{R}^n$, and there is a linear map $\boldsymbol{A} : \mathbf{R}^m \to \mathbf{R}^n$ (denoted by matrix A which is $m \times n$), we refer to this map as being the derivative and we write it as $\mathrm{D}_{\boldsymbol{p}} \boldsymbol{f}$.

A point on terminology - in referring to the action of taking the derivative (giving the linear map $\boldsymbol{A}$), we write $\mathrm{D}_{\boldsymbol{p}} \boldsymbol{f}$, but in referring to the matrix A itself, it is known as the *Jacobian* matrix and is also written $J_{\boldsymbol{p}} \boldsymbol{f}$.

Properties

There are a number of important properties of this formulation of the derivative.

Affine approximations

If $\boldsymbol{f}$ is differentiable at $\boldsymbol{p}$ for $\boldsymbol{x}$ close to $\boldsymbol{p}$, $|\boldsymbol{f}(\boldsymbol{x})-(\boldsymbol{f}(\boldsymbol{p})+A(\boldsymbol{x}-\boldsymbol{p}))|$ is small compared to $|\boldsymbol{x}-\boldsymbol{p}|$, which means that $\boldsymbol{f}(\boldsymbol{x})$ is approximately equal to $\boldsymbol{f}(\boldsymbol{p})+A(\boldsymbol{x}-\boldsymbol{p})$.

We call an expression of the form $g(x)+c$ affine, when $g(x)$ is linear and c is a constant. $\boldsymbol{f}(\boldsymbol{p})+A(\boldsymbol{x}-\boldsymbol{p})$ is an affine approximation to $\boldsymbol{f}(\boldsymbol{x})$.

Jacobian matrix and partial derivatives

The Jacobian matrix of a function is in the form

$$(J_{\mathbf{p}}\mathbf{f})_{ij} = \left.\frac{\partial f_i}{\partial x_j}\right|_{\mathbf{p}}$$

for a $\boldsymbol{f} : \mathbf{R}^m \to \mathbf{R}^n$, $J_{\boldsymbol{p}} \boldsymbol{f}'$ *is a* m×n *matrix.*

The consequence of this is that if $\boldsymbol{f}$ is differentiable at $\boldsymbol{p}$, *all* the partial derivatives of $\boldsymbol{f}$ exist at $\boldsymbol{p}$.

However, it is possible that all the partial derivatives of a function exist at some point yet that function is not differentiable there, so it's very important not to mix derivative (linear map) with the Jacobian (matrix) especially when cited situation arised.

Continuity and differentiability

Furthermore, if all the partial derivatives exist, and are continuous in some neighbourhood of a point $\boldsymbol{p}$, then $\boldsymbol{f}$ is differentiable at $\boldsymbol{p}$. This has the consequence that for

a function ***f*** which has its component functions built from continuous functions (such as rational functions, differentiable functions or otherwise), ***f*** is differentiable everywhere ***f*** is defined.

We use the terminology *continuously differentiable* for a function differentiable at ***p*** which has all its partial derivatives existing and are continuous in some neighbourhood at ***p***.

Rules of taking Jacobians

If ***f*** : $\mathbf{R}^m \to \mathbf{R}^n$, and $h(x) : \mathbf{R}^m \to \mathbf{R}$ are differentiable at '**p**':

- $J_{\mathbf{p}}(\mathbf{f} + \mathbf{g}) = J_{\mathbf{p}}\mathbf{f} + J_{\mathbf{p}}\mathbf{g}$
- $J_{\mathbf{p}}(h\mathbf{f}) = hJ_{\mathbf{p}}\mathbf{f} + \mathbf{f}(\mathbf{p})J_{\mathbf{p}}h$
- $J_{\mathbf{p}}(\mathbf{f} \cdot \mathbf{g}) = \mathbf{g}^T J_{\mathbf{p}}\mathbf{f} + \mathbf{f}^T J_{\mathbf{p}}\mathbf{g}$

Important: make sure the order is right - matrix multiplication is not commutative!

Chain rule

The chain rule for functions of several variables is as follows. For ***f*** : $\mathbf{R}^m \to \mathbf{R}^n$ and ***g*** : $\mathbf{R}^n \to \mathbf{R}^p$, and ***g*** o ***f*** differentiable at ***p***, then the Jacobian is given by

$$\left(J_{\mathbf{f}(\mathbf{p})}\mathbf{g}\right)\left(J_{\mathbf{p}}\mathbf{f}\right)$$

Again, we have matrix multiplication, so one must preserve this exact order. Compositions in one order may be defined, but not necessarily in the other way.

Alternate notations

For simplicity, we will often use various standard abbreviations, so we can write most of the formulae on one line. This can make it easier to see the important details.

We can abbreviate partial differentials with a subscript, e.g,

$$\partial_x h(x,y) = \frac{\partial h}{\partial x} \quad \partial_x \partial_y h = \partial_y \partial_x h$$

When we are using a subscript this way we will generally use the Heaviside *D* rather than ∂,

$$D_x h(x,y) = \frac{\partial h}{\partial x} \quad D_x D_y h = D_y D_x h$$

Mostly, to make the formulae even more compact, we will put the subscript on the function itself.

$$D_x h = h_x \quad h_{xy} = h_{yx}$$

If we are using subscripts to label the axes, x_1, x_2 …, then, rather than having two layers of subscripts, we will use the number as the subscript.

$$h_1 = D_1 h = \partial_1 h = \partial_{x_1} h = \frac{\partial h}{\partial x_1}$$

We can also use subscripts for the components of a vector function, **u**=(u_x, u_y, u_y) or **u**=(u_1,u_2…u_n)

If we are using subscripts for both the components of a vector and for partial derivatives we will separate them with a comma.

$$u_{x,y} = \frac{\partial u_x}{\partial y}$$

The most widely used notation is h_x. Both h_1 and $\partial_1 h$ are also quite widely used whenever the axes are numbered. The notation $\partial_x h$ is used least frequently.

We will use whichever notation best suits the equation we are working with.

Directional derivatives

Normally, a partial derivative of a function with respect to one of its variables, say, x_j, takes the derivative of that "slice" of that function parallel to the x_j'th axis.

More precisely, we can think of cutting a function $\boldsymbol{f}(x_1,...,x_n)$ in space along the x_j'th axis, with keeping everything but the x_j variable constant.

From the definition, we have the partial derivative at a point $\boldsymbol{p}$ of the function along this slice as

$$\frac{\partial \mathbf{f}}{\partial x_j} = \lim_{t \to 0} \frac{\mathbf{f}(\mathbf{p} + t\mathbf{e}_j) - \mathbf{f}(\mathbf{p})}{t}$$

provided this limit exists.

Instead of the basis vector, which corresponds to taking the derivative along that axis, we can pick a vector in any direction (which we usually take as being a unit vector), and we take the *directional derivative* of a function as

$$\frac{\partial \mathbf{f}}{\partial \mathbf{d}} = \lim_{t \to 0} \frac{\mathbf{f}(\mathbf{p} + t\mathbf{d}) - \mathbf{f}(\mathbf{p})}{t}$$

where ***d*** is the direction vector.

If we want to calculate directional derivatives, calculating them from the limit definition is rather painful, but, we have the following: if $\boldsymbol{f} : \mathbf{R}^n \to \mathbf{R}$ is differentiable at a point ***p***, $|\boldsymbol{p}|=1$,

$$\frac{\partial \mathbf{f}}{\partial \mathbf{d}} = D_{\mathbf{p}}\mathbf{f}(\mathbf{d})$$

There is a closely related formulation which we'll look at in the next section.

Gradient vectors

The partial derivatives of a scalar tell us how much it changes if we move along one of the axes. What if we move in a different direction?

We'll call the scalar *f*, and consider what happens if we move an infintesimal direction ***dr***=(*dx,dy,dz*), using the chain rule.

$$\mathbf{df} = dx\frac{\partial f}{\partial x} + dy\frac{\partial f}{\partial y} + dz\frac{\partial f}{\partial z}$$

This is the dot product of **dr** with a vector whose components are the partial derivatives of ***f***, called the gradient of ***f***

$$\mathrm{grad}\mathbf{f} = \nabla \mathbf{f} = \left(\frac{\partial \mathbf{f}(\mathbf{p})}{\partial x_1}, \cdots, \frac{\partial \mathbf{f}(\mathbf{p})}{\partial x_n} \right)$$

We can form directional derivatives at a point ***p***, in the direction ***d*** then by taking the dot product of the gradient with ***d***

$$\frac{\partial \mathbf{f}(\mathbf{p})}{\partial \mathbf{d}} = \mathbf{d} \cdot \nabla \mathbf{f}(\mathbf{p})$$

.

Notice that grad *f* looks like a vector multiplied by a scalar. This particular combination of partial derivatives is commonplace, so we abbreviate it to

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right)$$

We can write the action of taking the gradient vector by writing this as an *operator*. Recall that in the one-variable case we can write *d*/*dx* for the action of taking the derivative with respect to *x*. This case is similar, but acts like a vector.

We can also write the action of taking the gradient vector as:

$$\nabla = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \cdots \frac{\partial}{\partial x_n} \right)$$

Properties of the gradient vector

Geometry

- Grad ***f***(***p***) is a vector pointing in the direction of steepest slope of ***f***. |grad ***f***(***p***)| is the rate of change of that slope at that point.

For example, if we consider h(*x*, *y*)=x^2+y^2. The level sets of *h* are concentric circles, centred on the origin, and

$$\nabla h = (h_x, h_y) = 2(x, y) = 2\mathbf{r}$$

grad *h* points directly away from the origin, at right angles to the contours.

- Along a level set, (□***f***)(***p***) is perpendicular to the level set {***x***|***f***(***x***)=***f***(***p***) at ***x***=***p***}.

If **dr** points along the contours of *f*, where the function is constant, then *df* will be zero. Since *df* is a dot product, that means that the two vectors, **df** and grad *f*, must be at right angles, i.e the gradient is at right angles to the contours.

Algebraic properties

Like *d/dx*, □ is linear. For any pair of constants, *a* and *b*, and any pair of scalar functions, *f* and *g*

$$\frac{d}{dx}(af + bg) = a\frac{d}{dx}f + b\frac{d}{dx}g \quad \nabla(af + bg) = a\nabla f + b\nabla g$$

Since it's a vector, we can try taking its dot and cross product with other vectors, and with itself.

Divergence

If the vector function **u** maps $\mathbf{R}^n$ to itself, then we can take the dot product of **u** and □. This dot product is called the divergence.

$$\operatorname{div}\mathbf{u} = \nabla \cdot \mathbf{u} = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \cdots \frac{\partial u_n}{\partial x_n}$$

If we look at a vector function like **v**=(1+x^2,xy) we can see that to the left of the origin all the **v** vectors are converging towards the origin, but on the right they are diverging away from it.

Div **u** tells us how much **u** is converging or diverging. It is positive when the vector is diverging from some point, and negative when the vector is converging on that point.

> *Example:*
> For *v*=(1+x^2, xy), div **v**=3*x*, which is positive to the right of the origin, where **v** is diverging, and negative to the left of the origin, where **v** is converging.

Like grad, div is linear.

$$\nabla \cdot (a\mathbf{u} + b\mathbf{v}) = a\nabla \cdot \mathbf{u} + b\nabla \cdot \mathbf{v}$$

Later in this chapter we will see how the divergence of a vector function can be integrated to tell us more about the behaviour of that function.

To find the divergence we took the dot product of and a vector with on the left. If we reverse the order we get

$$\mathbf{u} \cdot \nabla = u_x D_x + u_y D_y + u_z D_z$$

To see what this means consider **i**· This is D_x, the partial differential in the **i** direction. Similarly, **u**· is the the partial differential in the **u** direction, multiplied by |**u**|

Curl

If **u** is a three-dimensional vector function on $\mathbf{R}^3$ then we can take its cross product with □. This cross product is called the *curl*.

$$\operatorname{curl}\mathbf{u} = \nabla \times \mathbf{u} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ D_x & D_y & D_z \\ u_x & u_y & u_z \end{vmatrix}$$

Curl **u** tells us if the vector **u** is rotating round a point. The direction of curl **u** is the axis of rotation.

We can treat vectors in two dimensions as a special case of three dimensions, with u_z=0 and $D_z\mathbf{u}$=0. We can then extend the definition of curl **u** to two-dimensional vectors

$$\operatorname{curl}\mathbf{u} = D_y u_x - D_x u_y$$

This two dimensional curl is a scalar. In four, or more, dimensions there is no vector equivalent to the curl.

Example:
Consider **u**=(-*y, x*). These vectors are tangent to circles centred on the origin, so appear to be rotating around it anticlockwise.

$$\operatorname{curl}\mathbf{u} = D_y(-y) - D_x x = -2$$

Example
Consider **u**=(-*y, x-z, y*), which is similar to the previous example.

$$\operatorname{curl}\mathbf{u} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ D_x & D_y & D_z \\ -y & x-z & y \end{vmatrix} = 2\mathbf{i} + 2\mathbf{k}$$

This **u** is rotating round the axis **i**+**k**

Later in this chapter we will see how the curl of a vector function can be integrated to tell us more about the behaviour of that function.

Product and chain rules

Just as with ordinary differentiation, there are product rules for grad, div and curl.

- If *g* is a scalar and **v** is a vector, then

  the divergence of *g***v** is
  $\nabla \cdot (g\mathbf{v}) = g\nabla \cdot \mathbf{v} + (\mathbf{v} \cdot \nabla)g$
  the curl of *g***v** is
  $\nabla \times (g\mathbf{v}) = g(\nabla \times \mathbf{v}) + (\nabla g) \times \mathbf{v}$

- If *u* and *v* are both vectors then

  the gradient of their dot product is
  $\nabla(\mathbf{u} \cdot \mathbf{v}) = \mathbf{u} \times (\nabla \times \mathbf{v}) + \mathbf{v} \times (\nabla \times \mathbf{u}) + (\mathbf{u} \cdot \nabla)\mathbf{v} + (\mathbf{v} \cdot \nabla)\mathbf{u}$
  the divergence of their cross product is
  $\nabla \cdot (\mathbf{u} \times \mathbf{v}) = \mathbf{v} \cdot (\nabla \times \mathbf{u}) - \mathbf{u} \cdot (\nabla \times \mathbf{v})$
  the curl of their cross product is
  $\nabla \times (\mathbf{u} \times \mathbf{v}) = (\mathbf{v} \cdot \nabla)\mathbf{u} - (\mathbf{u} \cdot \nabla)\mathbf{v} + \mathbf{u}(\nabla \cdot \mathbf{v}) - \mathbf{v}(\nabla \cdot \mathbf{u})$

We can also write chain rules. In the general case, when both functions are vectors and the composition is defined, we can use the Jacobian defined earlier.

$$\nabla \mathbf{u}(\mathbf{v})|_{\mathbf{r}} = \mathbf{J}_{\mathbf{v}} \, \nabla \mathbf{v}|_{\mathbf{r}}$$

where $\mathbf{J_u}$ is the Jacobian of **u** at the point **v**.

Normally **J** is a matrix but if either the range or the domain of **u** is $\mathbf{R}^1$ then it becomes a vector. In these special cases we can compactly write the chain rule using only vector notation.

- If *g* is a scalar function of a vector and *h* is a scalar function of *g* then

$$\nabla h(g) = \frac{dh}{dg} \nabla g$$

- If *g* is a scalar function of a vector then

$$\nabla = (\nabla g) \frac{d}{dg}$$

This substitution can be made in any of the equations containing

Second order differentials

We can also consider dot and cross products of with itself, whenever they can be defined. Once we know how to simplify products of two 's we'll know out to simplify products with three or more.

The divergence of the gradient of a scalar *f* is

$$\nabla^2 f(x_1, x_2, \ldots x_n) = \frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2} + \ldots + \frac{\partial^2 f}{\partial x_n^2}$$

This combination of derivatives is the **Laplacian** of *f*. It is commmonplace in physics and multidimensional calculus because of its simplicity and symmetry.

We can also take the Laplacian of a vector,

$$\nabla^2 \mathbf{u}(x_1, x_2, \ldots x_n) = \frac{\partial^2 \mathbf{u}}{\partial x_1^2} + \frac{\partial^2 \mathbf{u}}{\partial x_2^2} + \ldots + \frac{\partial^2 \mathbf{u}}{\partial x_n^2}$$

The Laplacian of a vector is not the same as the divergence of its gradient

$$\nabla(\nabla \cdot \mathbf{u}) - \nabla^2\mathbf{u} = \nabla \times (\nabla \times \mathbf{u})$$

Both the curl of the gradient and the divergence of the curl are always zero.

$$\nabla \times \nabla f = 0 \quad \nabla \cdot (\nabla \times \mathbf{u}) = 0$$

This pair of rules will prove useful.

Integration

We have already considered differentiation of functions of more than one variable, which leads us to consider how we can meaningfully look at integration.

In the single variable case, we interpret the definite integral of a function to mean the area under the function. There is a similar interpretation in the multiple variable case: for example, if we have a paraboloid in $\mathbf{R}^3$, we may want to look at the integral of that paraboloid over some region of the *xy* plane, which will be the *volume* under that curve and inside that region.

Riemann sums

When looking at these forms of integrals, we look at the Riemann sum. Recall in the one-variable case we divide the interval we are integrating over into rectangles and summing the areas of these rectangles as their widths get smaller and smaller. For the multiple-variable case, we need to do something similar, but the problem arises how to split up $\mathbf{R}^2$, or $\mathbf{R}^3$, for instance.

To do this, we extend the concept of the interval, and consider what we call a *n*-interval. An *n*-interval is a set of points in some rectangular region with sides of some fixed width in each dimension, that is, a set in the form $\{\boldsymbol{x} \in \mathbf{R}^n | a_i \leq x_i \leq b_i \text{ with } i = 0,...,n\}$, and its area/size/volume (which we simply call its *measure* to avoid confusion) is the product of the lengths of all its sides.

So, an *n*-interval in $\mathbf{R}^2$ could be some rectangular partition of the plane, such as $\{(x,y) \mid x \in [0,1] \text{ and } y \in [0, 2]|\}$. Its measure is 2.

If we are to consider the Riemann sum now in terms of sub-*n*-intervals of a region Ω, it is

$$\sum_{i;S_i \subset \Omega} f(x_i^*)m(S_i)$$

where $m(S_i)$ is the measure of the division of Ω into *k* sub-*n*-intervals S_i, and x^*_i is a point in S_i. The index is important - we only perform the sum where S_i falls completely within Ω - any S_i that is not completely contained in Ω we ignore.

As we take the limit as k goes to infinity, that is, we divide up Ω into finer and finer sub-n-intervals, and this sum is the same no matter how we divide up Ω, we get the *integral* of f over Ω which we write

$$\int_{\Omega} f$$

For two dimensions, we may write

$$\int \int_{\Omega} f$$

and likewise for n dimensions.

Iterated integrals

Thankfully, we need not always work with Riemann sums every time we want to calculate an integral in more than one variable. There are some results that make life a bit easier for us.

For $\mathbf{R}^2$, if we have some region bounded between two functions of the other variable (so two functions in the form $f(x) = y$, or $f(y) = x$), between a constant boundary (so, between $x = a$ and $x = b$ or $y = a$ and $y = b$), we have

$$\int_a^b \int_{f(x)}^{g(x)} h(x, y)\, dydx$$

An important theorem (called *Fubini's theorem*) assures us that this integral is the same as

$$\int \int_{\Omega} f .$$

Order of integration

In some cases the first integral of the entire iterated integral is difficult or impossible to solve, therefore, it can be to our advantage to change the order of integration.

$$\int_a^b \int_{f(x)}^{g(x)} h(x, y)\, dxdy$$

$$\int_c^d \int_{e(y)}^{f(y)} h(x, y)\, dydx$$

As of the writing of this, there is no set method to change an order of integration from dxdy to dydx or some other variable. Although, it is possible to change the order of integration in an x and y simple integration by simply switching the limits of integration around also, in non-simple x and y integrations the best method as of yet is to recreate the limits of the integration from the graph of the limits of integration.

In higher order integration that can't be graphed, the process can be very tedious. For example, dxdydz can be written into dzdydx, but first dxdydz must be switched to dydxdz and then to dydzdx and then to dzdydx (but since 3-dimensional cases can be graphed, doing this would be seemingly idiotic).

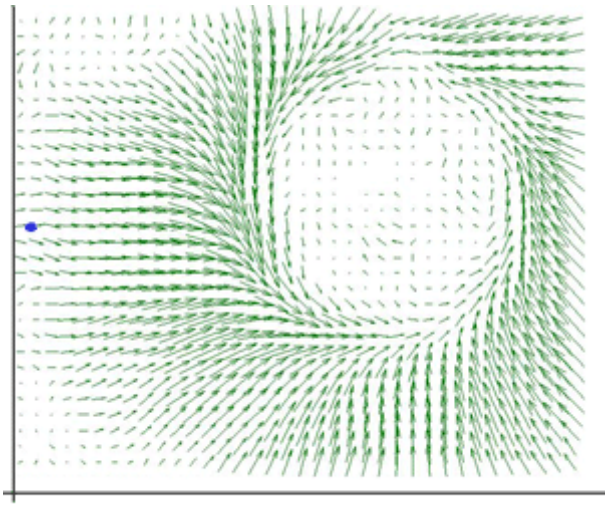
Parametric integrals

If we have a vector function, **u**, of a scalar parameter, *s*, we can integrate with respect to *s* simply by integrating each component of **u** seperately.

$$\mathbf{v}(s) = \int \mathbf{u}(s)\,ds \Rightarrow v_i(s) = \int u_i(s)\,ds$$

Similarly, if **u** is given a function of vector of parameters, **s**, lying in $\mathbf{R}^n$, integration with respect to the parameters reduces to a multiple integral of each component.

Line integrals



In one dimension, saying we are integrating from *a* to *b* uniquely specifies the integral.

In higher dimensions, saying we are integrating from **a** to **b** is not sufficient. In general, we must also specify the path taken between **a** and **b**.

We can then write the integrand as a function of the arclength along the curve, and integrate by components.

E.g, given a scalar function $h(\mathbf{r})$ we write

$$\int_C h(\mathbf{r})\,d\mathbf{r} = \int_C h(\mathbf{r})\frac{d\mathbf{r}}{ds}\,ds = \int_C h(\mathbf{r}(s))\mathbf{t}(s)\,ds$$

where C is the curve being integrated along, and $\mathbf{t}$ is the unit vector tangent to the curve.

There are some particularly natural ways to integrate a vector function, $\mathbf{u}$, along a curve,

$$\int_C \mathbf{u}\,ds \quad \int_C \mathbf{u}\cdot d\mathbf{r} \quad \int_C \mathbf{u}\times d\mathbf{r} \quad \int_C \mathbf{u}\cdot \mathbf{n}ds$$

where the third possibility only applies in 3 dimensions.

Again, these integrals can all be written as integrals with respect to the arclength, s.

$$\int_C \mathbf{u}\cdot d\mathbf{r} = \int_C \mathbf{u}\cdot\mathbf{t}\,ds \quad \int_C \mathbf{u}\times d\mathbf{r} = \int_C \mathbf{u}\times\mathbf{t}\,ds$$

If the curve is planar and u a vector lying in the same plane, the second integral can be usefully rewritten. Say,

$$\mathbf{u} = u_t\mathbf{t} + u_n\mathbf{n} + u_b\mathbf{b}$$

where $\mathbf{t}$, $\mathbf{n}$, and $\mathbf{b}$ are the tangent, normal, and binormal vectors uniquely defined by the curve.

Then

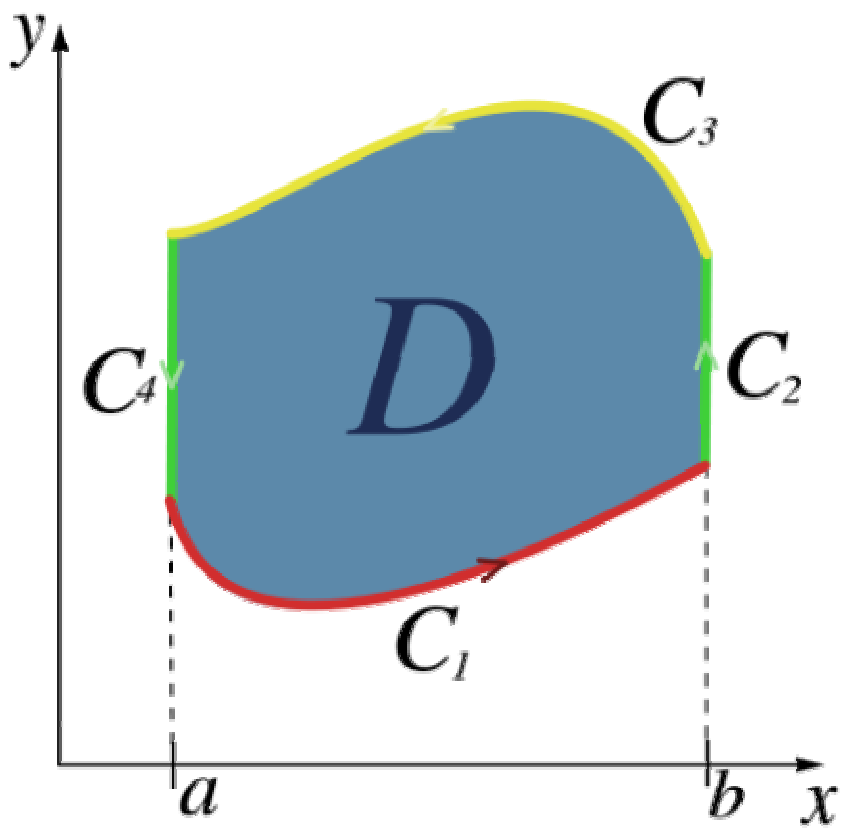
$$\mathbf{u}\times\mathbf{t} = -\mathbf{b}u_n + \mathbf{n}u_b$$

For the 2-d curves specified $\mathbf{b}$ is the constant unit vector normal to their plane, and u_b is always zero.

Therefore, for such curves,

$$\int_C \mathbf{u}\times d\mathbf{r} = \int_C \mathbf{u}\cdot\mathbf{n}ds$$

Green's Theorem



Let C be a piecewise smooth, simple closed curve that bounds a region S on the Cartesian plane. If two function M(x,y) and N(x,y) are continuous and their partial derivatives are continuous, then

$$\int\int_S (\frac{\partial N}{\partial x} - \frac{\partial M}{\partial y})dA = \oint_C M dx + N dy = \oint_C \mathbf{F} \cdot d\mathbf{r}$$

In order for Green's theorem to work there must be no singularities in the vector field within the boundaries of the curve.

Green's theorem works by summing the circulation in each infinitesimal segment of area enclosed within the curve.

Inverting differentials

We can use line integrals to calculate functions with specified divergence, gradient, or curl.

- If grad $V = \mathbf{u}$

$$V(\mathbf{p}) = \int_{\mathbf{p}_0}^{\mathbf{p}} \mathbf{u} \cdot d\mathbf{r} + h(\mathbf{p})$$

where *h* is any function of zero gradient and curl **u** must be zero.

- If div **u** = *V*

$$\mathbf{u}(\mathbf{p}) = \int_{\mathbf{p}_0}^{\mathbf{p}} V \, d\mathbf{r} + \mathbf{w}(\mathbf{p})$$

where **w** is any function of zero divergence.

- If curl **u** = **v**

$$\mathbf{u}(\mathbf{p}) = \frac{1}{2} \int_{\mathbf{p}_0}^{\mathbf{p}} \mathbf{v} \times d\mathbf{r} + \mathbf{w}(\mathbf{p})$$

where **w** is any function of zero curl.

For example, if $V=r^2$ then

$$\mathrm{grad} V = 2(x, y, z) = 2\mathbf{r}$$

and

$$\begin{aligned} \int_{\mathbf{0}}^{\mathbf{r}} 2\mathbf{u} \cdot d\mathbf{u} &= \int_{\mathbf{0}}^{\mathbf{r}} 2\left(u du + v dv + w dw\right) \\ &= \left[u^2\right]_{\mathbf{0}}^{\mathbf{r}} + \left[v^2\right]_{\mathbf{0}}^{\mathbf{r}} + \left[w^2\right]_{\mathbf{0}}^{\mathbf{r}} \\ &= x^2 + y^2 + z^2 = r^2 \end{aligned}$$

so this line integral of the gradient gives the original function.

Similarly, if **v**=**k** then

$$\mathbf{u}(\mathbf{p}) = \int_{\mathbf{p}_0}^{\mathbf{p}} \mathbf{k} \times d\mathbf{r}$$

Consider any curve from **0** to **p**=(*x,* y', *z*), given by **r**=**r**(*s*) with **r**(0)=**0** and **r**(*S*)=**p** for some *S*, and do the above integral along that curve.

$$\begin{aligned}
\mathbf{u}(\mathbf{p}) &= \int_0^S \mathbf{k} \times \tfrac{d\mathbf{r}}{ds}\, ds \\
&= \int_0^S \left(\tfrac{dr_x}{ds}\mathbf{j} - \tfrac{dr_y}{ds}\mathbf{i} \right) ds \\
&= \mathbf{j} \int_0^S \tfrac{dr_x}{ds}\, ds - \mathbf{i} \int_0^S \tfrac{dr_y}{ds}\, ds \\
&= \mathbf{j}[r_x(s)]_0^S - \mathbf{i}[r_y(s)]_0^S \\
&= p_x\mathbf{j} - p_y\mathbf{i} = x\mathbf{j} - y\mathbf{i}
\end{aligned}$$

and curl **u** is

$$\frac{1}{2}\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ D_x & D_y & D_z \\ -y & x & 0 \end{vmatrix} = \mathbf{k} = \mathbf{v}$$

as expected.

We will soon see that these three integrals do not depend on the path, apart from a constant.

Surface and Volume Integrals

Just as with curves, it is possible to parameterise surfaces then integrate over those parameters without regard to geometry of the surface.

That is, to integrate a scalar function *V* over a surface *A* parameterised by *r* and *s* we calculate

$$\int_A V(x,y,z)\, dS = \int\int_A V(r,s) \det J\, dr ds$$

where *J* is the Jacobian of the tranformation to the parameters.

To integrate a vector this way, we integrate each component seperately.

However, in three dimensions, every surface has an associated normal vector **n**, which can be used in integration. We write ***dS***=**n***dS*.

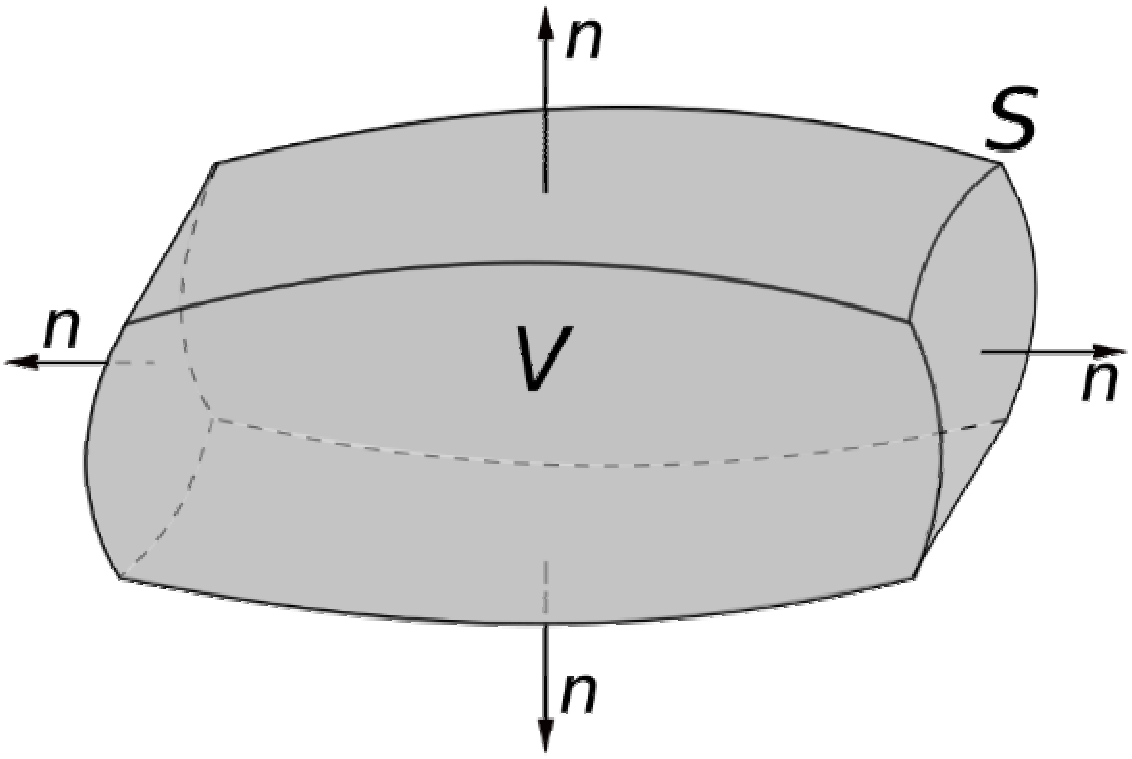
For a scalar function, *V*, and a vector function, **v**, this gives us the integrals

$$\int_A V\, \mathbf{dS} \qquad \int_A \mathbf{v} \cdot \mathbf{dS} \qquad \int_A \mathbf{v} \times \mathbf{dS}$$

These integrals can be reduced to parametric integrals but, written this way, it is clear that they reflect more of the geometry of the surface.

When working in three dimensions, dV is a scalar, so there is only one option for integrals over volumes.

Gauss's divergence theorem



We know that, in one dimension,

$$\int_a^b Df\,dx = f|_a^b$$

Integration is the inverse of differentiation, so integrating the differential of a function returns the original function.

This can be extended to two or more dimensions in a natural way, drawing on the analogies between single variable and multivariable calculus.

The analog of D is □, so we should consider cases where the integrand is a divergence.

Instead of integrating over a one-dimensional interval, we need to integrate over a n-dimensional volume.

In one dimension, the integral depends on the values at the edges of the interval, so we expect the result to be connected with values on the boundary.

This suggests a theorem of the form,

$$\int_V \nabla \cdot \mathbf{u}\, dV = \int_{\partial V} \mathbf{n} \cdot \mathbf{u} dS$$

This is indeed true, for vector fields in any number of dimensions.

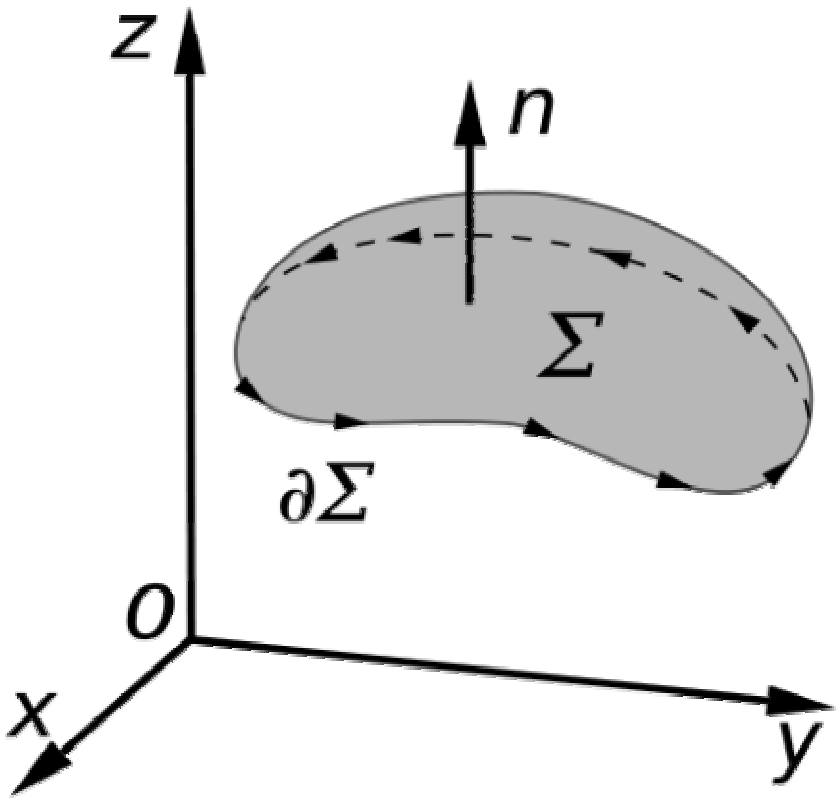
This is called *Gauss's theorem*.

There are two other, closely related, theorems for grad and curl:

- $\int_V \nabla u\, dV = \int_{\partial V} u\mathbf{n} dS$,
- $\int_V \nabla \times \mathbf{u}\, dV = \int_{\partial V} \mathbf{n} \times \mathbf{u} dS$,

with the last theorem only being valid where curl is defined.

Stokes' curl theorem



These theorems also hold in two dimensions, where they relate surface and line integrals. Gauss's divergence theorem becomes

$$\int_S \nabla \cdot \mathbf{u}\, dS = \oint_{\partial S} \mathbf{n} \cdot \mathbf{u} ds$$

where *s* is arclength along the boundary curve and the vector **n** is the unit normal to the curve that lies in the surface *S*, i.e in the tangent plane of the surface at its boundary, which is not necessarily the same as the unit normal associated with the boundary curve itself.

Similarly, we get

$$\int_S \nabla \times \mathbf{u}\, dS = \int_C \mathbf{n} \times \mathbf{u} ds \quad (1)$$

,

where *C* is the boundary of *S*

In this case the integral does not depend on the surface *S*.

To see this, suppose we have different surfaces, S_1 and S_2, spanning the same curve C, then by switching the direction of the normal on one of the surfaces we can write

$$\int_{S_1+S_2} \nabla \times \mathbf{u}\, dS = \int_S \nabla \times \mathbf{u}\, dS - \int_S \nabla \times \mathbf{u}\, dS \quad (2)$$

The left hand side is an integral over a closed surface bounding some volume V so we can use Gauss's divergence theorem.

$$\int_{S_1+S_2} \nabla \times \mathbf{u}\, dS = \int_V \nabla \cdot \nabla \times \mathbf{u}\, dV$$

but we know this integrand is always zero so the right hand side of (2) must always be zero, i.e the integral is independent of the surface.

This means we can choose the surface so that the normal to the curve lying in the surface is the same as the curves intrinsic normal.

Then, if **u** itself lies in the surface, we can write

$$\mathbf{u} = (\mathbf{u} \cdot \mathbf{n})\, \mathbf{n} + (\mathbf{u} \cdot \mathbf{t})\, \mathbf{t}$$

just as we did for line integrals in the plane earlier, and substitute this into (1) to get

$$\int_S \nabla \times \mathbf{u}\, dS = \int_C \mathbf{u} \cdot d\mathbf{r}$$

Ordinary differential equations involve equations containing:

- variables
- functions
- their derivatives

and their solutions.

In studying integration, you *already* have considered solutions to very simple differential equations. For example, when you look to solving

$$\int f(x)\, dx = g(x)$$

for g(x), you are really solving the differential equation

$$g'(x) = f(x)$$

Notations and terminology

The notations we use for solving differential equations will be crucial in the ease of solubility for these equations.

This document will be using **three** notations primarily:

- f' to denote the derivative of f
- Df to denote the derivative of f
- $\frac{df}{dx}$ to denote the derivative of f (for separable equations).

Terminology

Consider the differential equation

$$3f''(x) + 5xf(x) = 11$$

Since the equation's highest derivative is 2, we say that the differential equation is of *order* 2.

Some simple differential equations

A key idea in solving differential equations will be that of integration.

Let us consider the second order differential equation (remember that a function acts on a value).

$$f''(x) = 2$$

How would we go about solving this? It tells us that on differentiating twice, we obtain the constant 2 so, if we integrate twice, we should obtain our result.

Integrating once first of all:

$$\int f''(x)\,dx = \int 2\,dx$$
$$f'(x) = 2x + C_1$$

We have transformed the apparently difficult second order differential equation into a rather simpler one, viz.

$$f'(x) = 2x + C_1$$

This equation tells us that if we differentiate a function once, we get $2x + C_1$. If we integrate once more, we should find the solution.

$$\int f'(x)\,dx = \int 2x + C_1\,dx$$
$$f(x) = x^2 + C_1 x + C_2$$

This is the *solution* to the differential equation. We will get $f'' = 2$ for *all* values of C_1 and C_2.

The values C_1 and C_2 are related to quantities known as *initial conditions*.

Why are initial conditions useful? ODEs (ordinary differential equations) are useful in modeling physical conditions. We may wish to model a certain physical system which is initially at rest (so one initial condition may be zero), or wound up to some point (so an initial condition may be nonzero and be say 5 for instance) and we may wish to see how the system reacts under such an initial condition.

When we solve a system with given initial conditions, we substitute them after our process of integration.

Example

When we solved $f'' = 2$ say we had the initial conditions $f'(0) = 3$ and $f(0) = 2$. (Note, initial conditions need not occur at f(0)).

After we integrate we make substitutions:

$$f'(0) = 2(0) + C_1$$
$$3 = C_1$$
$$\int f'(x)\,dx = \int 2x + 3\,dx$$
$$f(x) = x^2 + 3x + C_2$$
$$f(0) = 0^2 + 3(0) + C_2$$
$$2 = C_2$$
$$f(x) = x^2 + 3x + 2$$

Without initial conditions, the answer we obtain is known as the *general solution* or the solution to the *family of equations*. With them, our solution is known as a *specific solution*.

Basic first order Differential Equations

In this section we will consider *four* main types of differential equations:

- separable
- homogeneous
- linear
- exact

There are many other forms of differential equation, however, and these will be dealt with in the next section

Separable equations

A *separable* equation is in the form (using dy/dx notation which will serve us greatly here)

$$\frac{dy}{dx} = f(x)/g(y)$$

Previously we have only dealt with simple differential equations with g(*y*)=1. How do we solve such a separable equation as above?

We group *x* and *dx* terms together, and *y* and *dy* terms together as well.

$$g(y)\ dy = f(x)\ dx$$

Integrating both sides with respect to y on the left hand side and x on the right hand side:

$$\int g(y)\,dy = \int f(x)\,dx + C$$

we will obtain the solution.

Worked example

Here is a worked example illustrating the process.

We are asked to solve

$$\frac{dy}{dx} = 3x^2y$$

Separating

$$\frac{dy}{y} = (3x^2)\,dx$$

Integrating

$$\int \frac{dy}{y} = \int 3x^2\,dx$$
$$\ln y = x^3 + C$$
$$y = e^{x^3+C}$$

Letting $k = e^C$ where k is a constant we obtain

$$y = ke^{x^3}$$

which is the general solution.

Verification

This step does not need be part of your working, but if you have time, you can verify your answer by differentiation. We obtained

$$y = ke^{x^3}$$

as the solution to

$$\frac{dy}{dx} = 3x^2y$$

Differentiating the solution,

$$\frac{dy}{dx} = 3kx^2e^{x^3}$$

Since $y = ke^{x^3}$, we can write

$$\frac{dy}{dx} = 3x^2y$$

We see that we obtain our original differential equation, so we can confirm our working as being correct.

Homogeneous equations

A *homogeneous* equation is in the form

$$\frac{dy}{dx} = f(y/x)$$

This looks difficult as it stands, however we can utilize the substitution

$$v = \frac{y}{x}$$

so that we are now dealing with F(v) rather than F(y/x).

Now we can express y in terms of v, as $y=xv$ and use the product rule.

The equation above then becomes, using the product rule

$$\frac{dy}{dx} = v + x\frac{dv}{dx}$$

Then

$$\begin{aligned} v + x\frac{dv}{dx} &= f(v) \\ x\frac{dv}{dx} &= f(v) - v \\ \frac{dv}{dx} &= \frac{f(v) - v}{x} \end{aligned}$$

which is a separable equation and can be solved as above.

However let's look at a worked equation to see how homogeneous equations are solved.

Worked example

We have the equation

$$\frac{dy}{dx} = \frac{y^2 + x^2}{yx}$$

This does not appear to be immediately separable, but let us expand to get

$$\begin{aligned} \frac{dy}{dx} &= \frac{y^2}{yx} + \frac{x^2}{yx} \\ \frac{dy}{dx} &= \frac{x}{y} + \frac{y}{x} \end{aligned}$$

Substituting $y=xv$ which is the same as substituting $v=y/x$:

$$\frac{dy}{dx} = 1/v + v$$

Now

$$v + x\frac{dv}{dx} = 1/v + v$$

Canceling *v* from both sides

$$x\frac{dv}{dx} = 1/v$$

Separating

$$v\,dv = dx/x$$

Integrating both sides

$$\begin{aligned}
&\frac{1}{2}v^2 + C = \ln(x) \\
&\frac{1}{2}\left(\frac{y}{x}\right)^2 = \ln(x) - C \\
&y^2 = 2x^2\ln(x) - 2Cx^2 \\
&y = x\sqrt{2\ln(x) - 2C}
\end{aligned}$$

which is our desired solution.

Linear equations

A linear first order differential equation is a differential equation in the form

$$a(x)\frac{dy}{dx} + b(x)y = c(x)$$

Multiplying or dividing this equation by any non-zero function of *x* makes no difference to its solutions so we could always divide by *a*(*x*) to make the coefficient of the differential 1, but writing the equation in this more general form may offer insights.

At first glance, it is not possible to integrate the left hand side, but there is one special case. If *b* happens to be the differential of *a* then we can write

$$a(x)\frac{dy}{dx} + b(x)y = a(x)\frac{dy}{dx} + y\frac{da}{dx} = \frac{d}{dx}a(x)y$$

and integration is now straightforward.

Since we can freely multiply by any function, lets see if we can use this freedom to write the left hand side in this special form.

We multiply the entire equation by an arbitrary, *I*(*x*), getting

$$aI\frac{dy}{dx} + bIy = cI$$

then impose the condition

$$\frac{d}{dx}aI = bI$$

If this is satisfied the new left hand side will have the special form. Note that multiplying *I* by any constant will leave this condition still satisfied.

Rearranging this condition gives

$$\frac{1}{I}\frac{dI}{dx} = \frac{b - \frac{da}{dx}}{a}$$

We can integrate this to get

$$\ln I(x) = \int \frac{b(z)}{a(z)}dz - \ln a(x) + c \quad I(x) = \frac{k}{a(x)}e^{\int \frac{b(z)}{a(z)}dz}$$

We can set the constant *k* to be 1, since this makes no difference.

Next we use *I* on the original differential equation, getting

$$e^{\int \frac{b(z)}{a(z)}dz}\frac{dy}{dx} + e^{\int \frac{b(z)}{a(z)}dz}\frac{b(x)}{a(x)}y = e^{\int \frac{b(z)}{a(z)}dz}\frac{c(x)}{a(x)}$$

Because we've chosen *I* to put the left hand side in the special form we can rewrite this as

$$\frac{d}{dx}(ye^{\int \frac{b(z)}{a(z)}dz}) = e^{\int \frac{b(z)}{a(z)}dz}\frac{c(x)}{a(x)}$$

Integrating both sides and dividing by *I* we obtain the final result

$$y = e^{-\int \frac{b(z)}{a(z)}dz} \left(\int e^{\int \frac{b(z)}{a(z)}dz} \frac{c(x)}{a(x)} dx + C \right)$$

We call *I* an *integrating factor*. Similar techniques can be used on some other calclulus problems.

Example

Consider

$$\frac{dy}{dx} + y \tan x = 1 \quad y(0) = 0$$

First we calculate the integrating factor.

$$I = e^{\int \tan z dz} = e^{\ln \sec x} = \sec x$$

Multiplying the equation by this gives

$$\sec x \frac{dy}{dx} + y \sec x \tan x = \sec x$$

or

$$\frac{d}{dx} y \sec x = \sec x$$

We can now integrate

$$y = \cos x \int_0^x \sec z \, dz = \cos x \ln(\sec x + \tan x)$$

Exact equations

An exact equation is in the form

$$f(x, y)\, dx + g(x, y)\, dy = 0$$

and, has the property that

$$D_x f = D_y g$$

(If the differential equation does not have this property then we can't proceed any further).

As a result of this, if we have an exact equation then there exists a function h(x, y) such that

$$D_y h = f \text{ and } D_x h = g$$

So then the solutions are in the form

$$h(x, y) = c$$

by using the fact of the total differential. We can find then h(x, y) by integration

Basic second and higher order ODE's

The generic solution of a n^{th} order ODE will contain n constants of integration. To calculate them we need n more equations. Most often, we have either

> boundary conditions, the values of y and its derivatives take for two different values of x

or

> initial conditions, the values of y and its first $n-1$ derivatives take for one particular value of x.

Reducible ODE's

1. If the independent variable, x, does not occur in the differential equation then its order can be lowered by one. This will reduce a second order ODE to first order.

Consider the equation:

$$F\left(y, \frac{dy}{dx}, \frac{d^2y}{dx^2}\right) = 0$$

Define

$$u = \frac{dy}{dx}$$

Then

$$\frac{d^2y}{dx^2} = \frac{du}{dx} = \frac{du}{dy} \cdot \frac{dy}{dx} = \frac{du}{dy} \cdot u$$

Substitute these two expression into the equation and we get

$$F\left(y, u, \frac{du}{dy} \cdot u\right)_{=0}$$

which is a first order ODE

Example

Solve

$$1 + 2y^2 \mathrm{D}^2 y = 0$$

if at x=0, y=Dy=1

First, we make the substitution, getting

$$1 + 2y^2 u \frac{du}{dy} = 0$$

This is a first order ODE. By rearranging terms we can separate the variables

$$udu = -\frac{dy}{2y^2}$$

Integrating this gives

$u^2 / 2 = c + 1 / 2y$

We know the values of y and u when x=0 so we can find c

$$c = u^2/2 - 1/2y = 1^2/2 - 1/(2 \cdot 1) = 1/2 - 1/2 = 0$$

Next, we reverse the substitution

$$\frac{dy}{dx}^2 = u^2 = \frac{1}{y}$$

and take the square root

$$\frac{dy}{dx} = \pm\frac{1}{\sqrt{y}}$$

To find out which sign of the square root to keep, we use the initial condition, D*y*=1 at *x*=0, again, and rule out the negative square root. We now have another separable first order ODE,

$$\frac{dy}{dx} = \frac{1}{\sqrt{y}}$$

Its solution is

$$\frac{2}{3}y^{\frac{3}{2}} = x + d$$

Since *y*=1 when *x*=0, *d*=2/3, and

$$y = \left(1 + \frac{3x}{2}\right)^{\frac{2}{3}}$$

2. If the dependent variable, *y*, does not occur in the differential equation then it may also be reduced to a first order equation.

Consider the equation:

$$F\left(x, \frac{dy}{dx}, \frac{d^2y}{dx^2}\right) = 0$$

Define

$$u = \frac{dy}{dx}$$

Then

$$\frac{d^2y}{dx^2} = \frac{du}{dx}$$

Substitute these two expressions into the first equation and we get

$$F\left(x, u, \frac{du}{dx}\right)_{=0}$$

which is a first order ODE

Linear ODEs

An ODE of the form

$$\frac{d^n y}{dx^n} + a_1(x)\frac{d^{n-1} y}{dx^{n-1}} + ... + a_n y = F(x)$$

is called **linear**. Such equations are much simpler to solve than typical non-linear ODEs. Though only a few special cases can be solved exactly in terms of elementary functions, there is much that can be said about the solution of a generic linear ODE.

If *F(x)=0* for all *x* the ODE is called **homogeneous**

Two useful properties of generic linear equations are

1. Any linear combination of solutions of an homogeneous linear equation is also a solution.
2. If we have a solution of a nonhomogeneous linear equation and we add any solution of the corresponding homogenous linear equation we get another solution of the nonhomogeneous linear equation

Variation of constants

Suppose we have a linear ODE,

$$\frac{d^n y}{dx^n} + a_1(x)\frac{d^{n-1} y}{dx^{n-1}} + ... + a_n y = 0$$

and we know one solution, *y=w(x)*

The other solutions can always be written as *y=wz*. This substitution in the ODE will give us terms involving every differential of *z* upto the n^{th}, no higher, so we'll end up with an n^{th} order linear ODE for *z*.

We know that *z* is constant is one solution, so the ODE for *z* must not contain a *z* term, which means it will effectively be an $n\text{-}1^{th}$ order linear ODE. We will have reduced the order by one.

Lets see how this works in practice.

Example

Consider

$$\frac{d^2y}{dx^2} + \frac{2}{x}\frac{dy}{dx} - \frac{6}{x^2}y = 0$$

One solution of this is $y=x^2$, so substitute $y=zx^2$ into this equation.

$$\left(x^2\frac{d^2z}{dx^2} + 2x\frac{dz}{dx} + 2z\right) + \frac{2}{x}\left(x^2\frac{dz}{dx} + 2xz\right) - \frac{6}{x^2}x^2z = 0$$

Rearrange and simplify.

$$x^2D^2z + 6xDz = 0$$

This is first order for Dz. We can solve it to get

$$z = Ax^{-5} \quad y = Ax^{-3}$$

Since the equation is linear we can add this to any multiple of the other solution to get the general solution,

$$y = Ax^{-3} + Bx^2$$

Linear homogeneous ODE's with constant coefficients

Suppose we have a ODE

$$(D^n + a_1D^{n-1} + \ldots + a_{n-1}D + a_0)y = 0$$

we can take an inspired guess at a solution (motivate this)

$$y = e^{px}$$

For this function $D^ny=p^ny$ so the ODE becomes

$$(p^n + a_1p^{n-1} + \ldots + a_{n-1}p + a_0)y = 0$$

$y=0$ is a trivial solution of the ODE so we can discard it. We are then left with the equation

$$p^n + a_1 p^{n-1} + \ldots + a_{n-1} p + a_0) = 0$$

This is called the *characteristic* equation of the ODE.

It can have up to *n* roots, p_1, p_2 … p_n, each root giving us a different solution of the ODE.

Because the ODE is linear, we can add all those solution together in any linear combination to get a general solution

$$y = A_1 e^{p_1 x} + A_2 e^{p_2 x} + \ldots + A_n e^{p_n x}$$

To see how this works in practice we will look at the second order case. Solving equations like this of higher order uses the exact same principles; only the algebra is more complex.

Second order

If the ODE is second order,

$$D^2 y + bDy + cy = 0$$

then the characteristic equation is a quadratic,

$$p^2 + bp + c = 0$$

with roots

$$p_\pm = \frac{-b \pm \sqrt{b^2 - 4c}}{2}$$

What these roots are like depends on the sign of b^2-4c, so we have three cases to consider.

1) $b^2 > 4c$

In this case we have two different real roots, so we can write down the solution straight away.

$$y = A_+ e^{p+} + A_- e^{p-}$$

2) $b^2 < 4c$

In this case, both roots are imaginary. We could just put them directly in the formula, but if we are interested in real solutions it is more useful to write them another way.

Defining $k^2=4c-b^2$, then the solution is

$$y = A_+ e^{ikx-\frac{bx}{2}} + A_- e^{-ikx-\frac{bx}{2}}$$

For this to be real, the *A*'s must be complex conjugates

$$A_\pm = Ae^{\pm ia}$$

Make this substitution and we can write,

$$y = Ae^{-bx/2}\cos(kx + a)$$

If *b* is positive, this is a damped oscillation.

3) $b^2 = 4c$

In this case the characteristic equation only gives us one root, *p=-b/2*. We must use another method to find the other solution.

We'll use the method of variation of constants. The ODE we need to solve is,

$$D^2y - 2pDy + p^2y = 0$$

rewriting *b* and *c* in terms of the root. From the characteristic equation we know one solution is $y = e^{px}$ so we make the substitution $y = ze^{px}$, giving

$$(e^{px}D^2z + 2pe^{px}Dz + p^2e^{px}z) - 2p(e^{px}Dz + pe^{px}z) + p^2e^{px}z = 0$$

This simplifies to $D^2z=0$, which is easily solved. We get

$$z = Ax + B \quad y = (Ax + B)e^{px}$$

so the second solution is the first multiplied by *x*.

Higher order linear constant coefficient ODE's behave similarly: an exponential for every real root of the characteristic and a exponent multiplied by a trig factor for every complex conjugate pair, both being multiplied by a polynomial if the root is repeated.

E.g, if the characteristic equation factors to

$$(p - 1)^4(p - 3)(p^2 + 1)^2 = 0$$

the general solution of the ODE will be

$$y = (A + Bx + Cx^2 + Dx^3)e^x + Ee^{3x} + F\cos(x + a) + Gx\cos(x + b)$$

The most difficult part is finding the roots of the characteristic equation.

Linear nonhomogeneous ODEs with constant coefficients

First, let's consider the ODE

$$Dy - y = x$$

a nonhomogeneous first order ODE which we know how to solve.

Using the integrating factor e^{-x} we find

$$y = ce^{-x} + 1 - x$$

This is the sum of a solution of the corresponding homogeneous equation, and a polynomial.

Nonhomogeneous ODE's of higher order behave similarly.

If we have a single solution, y_p of the nonhomogeneous ODE, called a *particular* solution,

$$(D^n + a_1 D^{n-1} + \cdots + a_n)y = F(x)$$

then the general solution is $y=y_p+y_h$, where y_h is the general solution of the homogeneous ODE.

Find y_p for an arbitrary $F(x)$ requires methods beyond the scope of this chapter, but there are some special cases where finding y_p is straightforward.

Remember that in the first order problem y_p for a polynomial $F(x)$ was itself a polynomial of the same order. We can extend this to higher orders.

Example:

$$D^2y + y = x^3 - x + 1$$

Consider a particular solution

$$y_p = b_0 + b_1x + b_2x^2 + x^3$$

Substitute for y and collect coefficients

$$x^3 + b_2x^2 + (6 + b_1)x + (2b_2 + b_0) = x^3 - x + 1$$

So $b_2=0$, $b_1=-7$, $b_0=1$, and the general solution is

$$y = a\sin x + b\cos x + 1 - 7x + x^3$$

This works because all the derivatives of a polynomial are themselves polynomials.

Two other special cases are

$$F(x) = P_n e^{kx} \quad y_p(x) = Q_n e^{kx}$$
$$F(x) = A_n \sin kx + B_n \cos kx \quad y_p(x) = P_n \sin kx + Q_n \cos kx$$

where P_n,Q_n,A_n, and B_n are all polynomials of degree n.

Making these substitutions will give a set of simultaneous linear equations for the coefficients of the polynomials.

Non-Linear ODEs

If the ODE is not linear, first check if it is reducible. If it is neither linear nor reducible there is no generic method of solution. You may, with sufficient ingenuity and algebraic skill, be able to transform it into a linear ODE.

First order

Any partial differential equation of the form

$$h_1 \frac{\partial u}{\partial x_1} + h_2 \frac{\partial u}{\partial x_2} \cdots + h_n \frac{\partial u}{\partial x_n} = b$$

where h_1, h_2 … h_n, and b are all functions of both u and $\mathbf{R}^n$ can be reduced to a set of ordinary differential equations.

To see how to do this, we will first consider some simpler problems.

Special cases

We will start with the simple PDE

$$u_z(x, y, z) = u(x, y, z) \quad (1)$$

Because u is only differentiated with respect to z, for any fixed x and y we can treat this like the ODE, $du/dz=u$. The solution of that ODE is ce^z, where c is the value of u when $z=0$, *for the fixed x and y*

Therefore, the solution of the PDE is

$u(x,y,z) = u(x,y,0)e^z$

Instead of just having a constant of integration, we have an arbitary function. This will be true for any PDE.

Notice the shape of the solution, an arbitary function of points in the *xy*, plane, which is normal to the 'z' axis, and the solution of an ODE in the 'z' direction.

Now consider the slightly more complex PDE

$$a_x u_x + a_y u_y + a_z u_z = h(u) \quad (2)$$

where h can be any function, and each *a* is a real constant.

We recognize the left hand side as being $\mathbf{a}\cdot\quad$, so this equation says that the differential of *u* in the **a** direction is *h(u)*. Comparing this with the first equation suggests that the solution can be written as an arbitary function on the plane normal to **a** combined with the solution of an ODE.

Remembering from Calculus/Vectors that any vector **r** can be split up into components parallel and perpendicular to *a*,

$$\mathbf{r} = \mathbf{r}_\perp + \mathbf{r}_\parallel = \left(\mathbf{r} - \frac{(\mathbf{r}\cdot\mathbf{a})\mathbf{a}}{|\mathbf{a}|^2}\right) + \frac{(\mathbf{r}\cdot\mathbf{a})\mathbf{a}}{|\mathbf{a}|^2}$$

we will use this to split the components of **r** in a way suggested by the analogy with (1).

Let's write

$$\mathbf{r} = (x,y,z) = \mathbf{r}_\perp + s\mathbf{a} \quad s = \frac{\mathbf{r}\cdot\mathbf{a}}{\mathbf{a}\cdot\mathbf{a}}$$

and substitute this into (2), using the chain rule. Because we are only differentiating in the **a** direction, adding any function of the perpendicular vector to *s* will make no difference.

First we calculate grad *s*, for use in the chain rule,

$$\nabla s = \frac{\mathbf{a}}{a^2}$$

On making the substitution into (2), we get,

$$h(u) = \mathbf{a}\cdot\nabla s\frac{d}{ds}u(s) = \frac{\mathbf{a}\cdot\mathbf{a}}{\mathbf{a}\cdot\mathbf{a}}\frac{d}{ds}u(s) = \frac{du}{ds}$$

which is an ordinary differential equation with the solution

$$s = c(\mathbf{r}_\perp) + \int^u \frac{dt}{h(t)}$$

The constant *c* can depend on the perpendicular components, but not upon the parallel coordinate. Replacing *s* with a monotonic scalar function of *s* multiplies the ODE by a function of *s*, which doesn't affect the solution.

Example:

$u(x,t)_x = u(x,t)_t$

For this equation, **a** is (1, -1), $s=x-t$, and the perpendicular vector is (x+t)(1, 1). The reduced ODE is $du/ds=0$ so the solution is

u=f(x+t)

To find *f* we need initial conditions on *u*. Are there any constraints on what initial conditions are suitable?

Consider, if we are given

- $u(x,0)$, this is exactly $f(x)$,
- $u(3t,t)$, this is $f(4t)$ and $f(t)$ follows immediately
- $u(t^3+2t,t)$, this is $f(t^3+3t)$ and $f(t)$ follows, on solving the cubic.
- $u(-t,t)$, then this is $f(0)$, so if the given function isn't constant we have a inconsistency, and if it is the solution isn't specified off the initial line.

Similarly, if we are given *u* on any curve which the lines $x+t=c$ intersect only once, and to which they are not tangent, we can deduce *f*.

For any first order PDE with constant coefficients, the same will be true. We will have a set of lines, parallel to $r=\mathbf{a}t$, along which the solution is gained by integrating an ODE with initial conditions specified on some surface to which the lines aren't tangent.

If we look at how this works, we'll see we haven't actually used the constancy of **a**, so let's drop that assumption and look for a similar solution.

The important point was that the solution was of the form $u=f(x(s),y(s))$, where $(x(s),y(s))$ is the curve we integrated along -- a straight line in the previous case. We can add constant functions of integration to *s* without changing this form.

Consider a PDE,

$a(x,y)u_x + b(x,y)u_y = c(x,y,u)$

For the suggested solution, $u=f(x(s),y(s))$, the chain rule gives

$$\frac{du}{ds} = \frac{dx}{ds}u_x + \frac{dy}{ds}u_y$$

Comparing coefficients then gives

$$\frac{dx}{ds} = a(x,y) \quad \frac{dy}{ds} = b(x,y) \quad \frac{du}{ds} = c(x,y,u)$$

so we've reduced our original PDE to a set of simultaneous ODE's. This procedure can be reversed.

The curves $(x(s),y(s))$ are called *characteristics* of the equation.

Example: Solve $yu_x = xu_y$ given $u=f(x)$ for $x\geq 0$ The ODE's are

$$\frac{dx}{ds} = y \quad \frac{dy}{ds} = -x \quad \frac{du}{ds} = 0$$

subject to the initial conditions at $s=0$,

$$x(0) = r \quad y(0) = 0 \quad u(0) = f(r) \quad r \geq 0$$

This ODE is easily solved, giving

$$x(s) = r\cos s \quad y(s) = r\sin s \quad u(s) = f(r)$$

so the characteristics are concentric circles round the origin, and in polar coordinates $u(r,\theta)=f(r)$

Considering the logic of this method, we see that the independence of a and b from u has not been used either, so that assumption too can be dropped, giving the general method for equations of this *quasilinear* form.

Quasilinear

Summarising the conclusions of the last section, to solve a PDE

$$a_1(u,\mathbf{x})\frac{\partial u}{\partial x_1} + a_2(u,\mathbf{x})\frac{\partial u}{\partial x_2}\cdots + a_n(u,\mathbf{x})\frac{\partial u}{\partial x_n} = b(u,\mathbf{x})$$

subject to the initial condition that on the surface, $(x_1(r_1,\ldots,r_{n-1}, \ldots x_n(r_1,\ldots,r_{n-1})$, $u=f(r_1,\ldots,r_{n-1})$ --this being an arbitary paremetrisation of the initial surface--

- we transform the equation to the equivalent set of ODEs,

$$\frac{dx_1}{ds} = a_1 \quad \ldots \quad \frac{dx_n}{ds} = a_n \qquad \frac{du}{ds} = b$$

subject to the initial conditions

$$x_i(0) = f(r_1, \ldots, r_{n-1}) \qquad u = f(r_1, r_2, \ldots r_{n-1})$$

- Solve the ODE's, giving x_i as a function of s and the r_i.
- Invert this to get s and the r_i as functions of the x_i.
- Substitute these inverse functions into the expression for u as a function of s and the r_i obtained in the second step.

Both the second and third steps may be troublesome.

The set of ODEs is generally non-linear and without analytical solution. It may even be easier to work with the PDE than with the ODEs.

In the third step, the r_i together with s form a coordinate system adapted for the PDE. We can only make the inversion at all if the Jacobian of the transformation to Cartesian coordinates is not zero,

$$\begin{vmatrix} \frac{\partial x_1}{\partial r_1} & \cdots & \frac{\partial x_1}{\partial r_{n-1}} & a_1 \\ \vdots & \ddots & & \vdots \\ \frac{\partial x_n}{\partial r_1} & \cdots & \frac{\partial x_n}{\partial r_{n-1}} & a_n \end{vmatrix} \neq 0$$

This is equivalent to saying that the vector (a_1, &hellip:, a_n) is never in the tangent plane to a surface of constant s.

If this condition is not false when s=0 it may become so as the equations are integrated. We will soon consider ways of dealing with the problems this can cause.

Even when it is technically possible to invert the algebraic equations it is obviously inconvenient to do so.

Example

To see how this works in practice, we will
a/ consider the PDE,

$$uu_x + u_y + u_t = 0$$

with generic initial condition,

$$u = f(x,y) \text{ on } t = 0$$

Naming variables for future convenience, the corresponding ODE's are

$$\frac{dx}{d\tau} = u \quad \frac{dy}{d\tau} = 1 \quad \frac{dz}{d\tau} = 1 \quad \frac{du}{d\tau} = 0$$

subject to the initial conditions at τ=0

$$x = r \quad y = s \quad t = 0 \quad u = f(r, s)$$

These ODE's are easily solved to give

$$x = r + f(r, s)\tau \quad y = s + \tau \quad t = \tau \quad u = f(r, s)$$

These are the parametric equations of a set of straight lines, the characteristics.

The determinant of the Jacobian of this coordinate transformation is

$$\begin{vmatrix} 1 + \tau\frac{\partial f}{\partial r} & \tau\frac{\partial f}{\partial s} & f \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{vmatrix} = 1 + \tau\frac{\partial f}{\partial r}$$

This determinant is 1 when t=0, but if f_r is anywhere negative this determinant will eventually be zero, and this solution fails.

In this case, the failure is because the surface $sf_r = -1$ is an envelope of the characteristics.

For arbitary f we can invert the transformation and obtain an implicit expression for u

$u = f(x - tu, y - x)$

If f is given, this can be solved for u.

1/ $f(x,y) = ax$, The implicit solution is

$$u = a(x - tu) \Rightarrow u = \frac{ax}{1 + at}$$

This is a line in the u-x plane, rotating clockwise as t increases. If a is negative, this line eventually become vertical. If a is positive, this line tends towards u=0, and the solution is valid for all t.

2/ $f(x,y)=x^2$, The implicit solution is

$$u = (x - tu)^2 \Rightarrow u = \frac{1 + 2tx - \sqrt{1 + 4tx}}{2t^2}$$

This solution clearly fails when $1 + 4tx < 0$, which is just when $sf_r = -1$. For any $t>0$ this happens somewhere. As t increases, this point of failure moves toward the origin.

Notice that the point where $u=0$ stays fixed. This is true for any solution of this equation, whatever f is.

We will see later that we can find a solution after this time, if we consider discontinuous solutions. We can think of this as a shockwave.

3/ $f(x,y) = \sin(xy)$
The implicit solution is

$$u(x, y, t) = \sin\left((x - tu)(y - x)\right)$$

and we can not solve this explitly for u. The best we can manage is a numerical solution of this equation.

b/We can also consider the closely related PDE

$$uu_x + u_y + u_t = y$$

The corresponding ODE's are

$$\frac{dx}{d\tau} = u \quad \frac{dy}{d\tau} = 1 \quad \frac{dz}{d\tau} = 1 \quad \frac{du}{d\tau} = y$$

subject to the initial conditions at $\tau=0$

$$x = r \quad y = s \quad t = 0 \quad u = f(r, s)$$

These ODE's are easily solved to give

$$x = r + \tau f + \frac{1}{2}s\tau^2 + \frac{1}{6}\tau^3 \quad y = s + \tau \quad t = \tau \quad u = f + s\tau + \frac{1}{2}\tau^2$$

Writing f in terms of u, s, and τ, then substituting into the equation for x gives an implicit solution

$$u(x, y, t) = f(x - ut + \frac{1}{2}yt^2 - \frac{1}{6}t^3, y - t) + yt - \frac{1}{2}t^2$$

It is possible to solve this for u in some special cases, but in general we can only solve this equation numerically. However, we can learn much about the global properties of the solution from further analysis

Characteristic initial value problems

What if initial conditions are given on a characteristic, on an envelope of characteristics, on a surface with characteristic tangents at isolated points?

Discontinuous solutions

So far, we've only considered smooth solutions of the PDE, but this is too restrictive. We may encounter initial conditions which aren't smooth, e.g.

$$u_t = cu_x \quad u(x,0) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

If we were to simply use the general solution of this equation for smooth initial conditions,

$$u(x,t) = u(x+ct,0)$$

we would get

$$u(x,t) = \begin{cases} 1, & x+ct \geq 0 \\ 0, & x+ct < 0 \end{cases}$$

which appears to be a solution to the original equation. However, since the partial differentials are undefined on the characteristic $x+ct=0$, so it becomes unclear what it means to say that the equation is true at that point.

We need to investigate further, starting by considering the possible types of discontinuities.

If we look at the derivations above, we see we've never use any second or higher order derivatives so it doesn't matter if they aren't continuous, the results above will still apply.

The next simplest case is when the function is continuous, but the first derivative is not, e.g $|x|$. We'll initially restrict ourselves to the two-dimensional case, $u(x, t)$ for the generic equation.

$$a(x,t)u_x + b(x,t)u_t = c(u,x,t) \quad (1)$$

Typically, the discontinuity is not confined to a single point, but is shared by all points on some curve, $(x_0(s), t_0(s))$

Then we have

$$\begin{array}{ll} x > x_0 & \lim_{x \to x_0} = u_+ \\ x < x_0 & \lim_{x \to x_0} = u_- \end{array}$$

We can then compare *u* and its derivatives on both sides of this curve.

It will prove useful to name the *jumps* across the discontinuity. We say

$$[u] = u_+ - u_- \quad [u_x] = u_{x+} - u_{x-} \quad [u_t] = u_{t+} - u_{t-}$$

Now, since the equation (1) is true on both sides of the discontinuity, we can see that both u_+ and u_-, being the limits of solutions, must themselves satisfy the equation. That is,

$$\begin{array}{rcl} a(x,t)u_{+x} + b(x,t)u_{+t} & = & c(u_+, x, t) \\ a(x,t)u_{-x} + b(x,t)u_{-t} & = & c(u_-, x, t) \end{array} \text{ where } \begin{array}{c} x = x_0(s) \\ t = t_0(s) \end{array}$$

Subtracting then gives us an equation for the jumps in the differentials

$$a(x,t)[u_x] + b(x,t)[u_t] = 0$$

We are considering the case where *u* itself is continuous so we know that [*u*]=0. Differentiating this with respect to *s* will give us a second equation in the differential jumps.

$$\frac{dx_0}{ds}[u_x] + \frac{dx_0}{dt}[u_t] = 0$$

The last two equations can only be both true if one is a multiple of the other, but multiplying *s* by a constant also multiplies the second equation by that same constant while leaving the curve of discontinuity unchanged, hence we can without loss of generality define *s* to be such that

$$\frac{dx_0}{ds} = a \quad \frac{dt_0}{ds} = b$$

But these are the equations for a characteristic, i.e **discontinuities propagate along characteristics.** We could use this property as an alternative definition of characteristics.

We can deal similarly with discontinuous functions by first writing the equation in *conservation form*, so called because conservation laws can always be written this way.

$$(au)_x + (bu)_t = a_x u + b_t u + c \quad (1)$$

Notice that the left hand side can be regarded as the divergence of (*au*, *bu*). Writing the equation this way allows us to use the theorems of vector calculus.

Consider a narrow strip with sides parallel to the discontinuity and width *h*

We can integrate both sides of (1) over R, giving

$$\int_R (au)_x + (bu)_t \, dxdt = \int_R (a_x + b_t)u + c \, dxdt$$

Next we use Green's theorem to convert the left hand side into a line integral.

$$\oint_{\partial R} audt - budx = \int_R (a_x + b_t)u + c \, dxdt$$

Now we let the width of the strip fall to zero. The right hand side also tends to zero but the left hand side reduces to the difference between two integrals along the part of the boundary of *R* parallel to the curve.

$$\int au_+ dt - bu_+ dx - \int au_- dt - bu_- dx = 0$$

The integrals along the opposite sides of *R* have different signs because they are in opposite directions.

For the last equation to always be true, the integrand must always be zero, i.e

$$\left(a\frac{dt_0}{ds} - b\frac{dx_0}{ds} \right) [u] = 0$$

Since, by assumption [*u*] isn't zero, the other factor must be, which immediately implies the curve of discontinuity is a characteristic.

Once again, *discontinuities propagate along characteristics.*

Above, we only considered functions of two variables, but it is straightforward to extend this to functions of *n* variables.

The initial condition is given on an *n*-1 dimensional surface, which evolves along the characteristics. Typical discontinuities in the initial condition will lie on a *n*-2 dimensional surface embedded within the initial surface. This surface of discontinuity will propagate along the characteristics that pass through the initial discontinuity.

The jumps themselves obey ordinary differential equations, much as *u* itself does on a characteristic. In the two dimensional case, for *u* continuous but not smooth, a little algebra shows that

$$\frac{d[u_x]}{ds} = [u_x]\left(\frac{\partial c}{\partial u} + a\frac{b_x}{b} - a_x\right)$$

while *u* obeys the same equation as before,

$$\frac{du}{ds} = c$$

We can integrate these equations to see how the discontinuity evolves as we move along the characteristic.

We may find that, for some future *s*, $[u_x]$ passes through zero. At such points, the discontinuity has vanished, and we can treat the function as smooth at that characteristic from then on.

Conversely, we can expect that smooth functions may, under the righr circumstances, become discontinuous.

To see how all this works in practice we'll consider the solutions of the equation

$$u_t + uu_x = 0 \quad u(x,0) = f(x)$$

for three different initial conditions.

The general solution, using the techniques outlined earlier, is

$$u = f(x - tu)$$

u is constant on the characteristics, which are straight lines with slope dependent on *u*.

First consider *f* such that

$$f(x) = \begin{cases} 1 & x > a \\ \frac{x}{a} & a \geq x > 0 \quad a > 0 \\ 0 & x \leq 0 \end{cases}$$

While u is continuous its derivative is discontinuous at x=0, where u=0, and at x=a, where u=1. The characteristics through these points divide the solution into three regions.

All the characteristics to the right of the characteristic through x=a, t=0 intersect the x-axis to the right of x=1, where u=1 so u is 1 on all those characteristics, i.e whenever x-t>a.

Similarly the characteristic through the origin is the line x=0, to the left of which u remains zero.

We could find the value of u at a point in between those two characteristics either by finding which intermediate characteristic it lies on and tracing it back to the initial line, or via the general solution.

Either way, we get

$$f(x) = \begin{cases} 1 & x - t > a \\ \frac{x}{a+t} & a + t \geq x > 0 \\ 0 & x \leq 0 \end{cases}$$

At larger t the solution u is more spread out than at t=0 but still the same shape.

We can also consider what happens when a tends to 0, so that u itself is discontinuous at x=0.

If we write the PDE in conservation form then use Green's theorem, as we did above for the linear case, we get

$$[u]\frac{dx_0}{ds} = \frac{1}{2}[u^2]\frac{dt_0}{ds}$$

[u^2] is the difference of two squares, so if we take s=t we get

$$\frac{dx_0}{dt} = \frac{1}{2}(u_- + u_+)$$

In this case the discontinuity behaves as if the value of u on it were the average of the limiting values on either side.

However, there is a caveat.

Since the limiting value to the left is *u*- the discontinuity must lie on that characteristic, and similarly for *u*+; i.e *the jump discontinuity must be on an intersection of characteristics*, at a point where *u* would otherwise be multivalued.

For this PDE the characteristic can only intersect on the discontinuity if

$$u_- > u_+$$

If this is not true the discontinuity can not propagate. Something else must happen.

The limit *a*=0 is an example of a jump discontinuity for which this condition is false, so we can see what happens in such cases by studying it.

Taking the limit of the solution derived above gives

$$f(x) = \begin{cases} 1 & x > t \\ \frac{x}{t} & t \geq x > 0 \\ 0 & x \leq 0 \end{cases}$$

If we had taken the limit of any other sequence of initials conditions tending to the same limit we would have obtained a trivially equivalent result.

Looking at the characteristics of this solution, we see that at the jump discontinuity characteristics on which *u* takes every value betweeen 0 and 1 all intersect.

At later times, there are two slope discontinuities, at *x*=0 and *x*=*t*, but no jump discontinuity.

This behaviour is typical in such cases. The jump discontinuity becomes a pair of slope discontinuities between which the solution takes all appropriate values.

Now, lets consider the same equation with the initial condition

$$f(x) = \begin{cases} 1 & x \leq 0 \\ 1 - \frac{x}{a} & a \geq x > 0 \quad a > 0 \\ 0 & x > a \end{cases}$$

This has slope discontinuities at *x*=0 and *x*=a, dividing the solution into three regions.

The boundaries between these regions are given by the characteristics through these initial points, namely the two lines

$$x = t \quad x = a$$

These characteristics intersect at t=a, so the nature of the solution must change then.

In between these two discontinuities, the characteristic through $x=b$ at $t=0$ is clearly

$$x = \left(1 - \frac{b}{a}\right) t + b \quad 0 \leq b \leq a$$

All these characteristics intersect at the same point, $(x,t)=(a,a)$.

We can use these characteristics, or the general solution, to write u for $t<a$

$$u(x,t) = \begin{cases} 1 & x \leq t \\ \frac{a-x}{a-t} & a \geq x > t \\ 0 & x > a \end{cases} \quad a > t \geq 0$$

As t tends to a, this becomes a step function. Since u is greater to the left than the right of the discontinuity, it meets the condition for propagation deduced above, so for $t>a$ u is a step function moving at the average speed of the two sides.

$$u(x,t) = \begin{cases} 1 & x \leq \frac{a+t}{2} \\ 0 & x > \frac{a+t}{2} \end{cases} \quad t \geq a \geq 0$$

This is the reverse of what we saw for the initial condition previously considered, two slope discontinuities merging into a step discontinuity rather than vice versa. Which actually happens depends entirely on the initial conditions. Indeed, examples could be given for which both processes happen.

In the two examples above, we started with a discontinuity and investigated how it evolved. It is also possible for solutions which are initially smooth to become discontinuous.

For example, we saw earlier for this particular PDE that the solution with the initial condition $u=x^2$ breaks down when $2xt+1=0$. At these points the solution becomes discontinuous.

Typically, discontinuities in the solution of any partial differential equation, not merely ones of first order, arise when solutions break down in this way and propagate similarly, merging and splitting in the same fashion.

Fully non-linear PDEs

It is possible to extend the approach of the previous sections to reduce any equation of the form

$$F(x_1, x_2, \ldots, x_n, u, u_{x_1}, u_{x_2}, \ldots, u_{x_n}) = 0$$

to a set of ODE's, for *any* function, *F*.

We will not prove this here, but the corresponding ODE's are

$$\frac{dx_i}{d\tau} = \frac{\partial F}{\partial u_i} \quad \frac{du_i}{d\tau} = -\left(\frac{\partial F}{\partial x_i} + u_i \frac{\partial F}{\partial u}\right) \quad \frac{du}{d\tau} = \sum_{i=1}^{n} u_i \frac{\partial F}{\partial u_i}$$

If *u* is given on a surface parameterized by $r_1 \ldots r_n$ then we have, as before, *n* initial conditions on the *n*, x_i

$$\tau = 0 \quad x_i = f_i(r_1, r_2, \ldots, r_{n-1})$$

given by the parameterization and *one* initial condition on *u* itself,

$$\tau = 0 \quad u = f(r_1, r_2, \ldots, r_{n-1})$$

but, because we have an extra *n* ODEs for the u_i's, we need an extra *n* initial conditions.

These are, *n*-1 consistency conditions,

$$\tau = 0 \quad \frac{\partial f}{\partial r_i} = \sum_{j=1}^{n-1} u_i \frac{\partial f_i}{\partial r_j}$$

which state that the u_i's are the partial derivatives of *u* on the initial surface, and *one* initial condition

$$\tau = 0 \quad F(x_1, x_2, \ldots, x_n, u, u_1, u_2, \ldots, u_n) = 0$$

stating that the PDE itself holds on the initial surface.

These *n* initial conditions for the u_i will be a set of algebraic equations, which may have multiple solutions. Each solution will give a different solution of the PDE.

Example

Consider

$$u_t = u_x^2 + u_y^2, \quad u(x, y, 0) = x^2 + y^2$$

The initial conditions at τ=0 are

$$\begin{array}{llll} x = r & y = s & t = 0 & u = r^2 + s^2 \\ u_x = 2r & u_y = 2s & u_t = 4(r^2 + s^2) & \end{array}$$

and the ODE's are

$$\begin{array}{llll} \frac{dx}{d\tau} = -2u_x & \frac{dy}{d\tau} = -2u_y & \frac{dt}{d\tau} = 1 & \frac{du}{d\tau} = u_t - 2(u_x^2 + u_y^2) \\ \frac{du_x}{d\tau} = 0 & \frac{du_y}{d\tau} = 0 & \frac{du_t}{d\tau} = 0 & \end{array}$$

Note that the partial derivatives are constant on the characteristics. This always happen, when the PDE contains only partial derivatives, simplifying the procedure.

These equations are readily solved to give

$$x = r(1 - 4\tau) \quad y = s(1 - 4\tau) \quad t = \tau \quad u = (r^2 + s^2)(1 - 4\tau)$$

On eliminating the parameters we get the solution,

$$u = \frac{x^2 + y^2}{1 - 4t}$$

which can easily be checked.

Second order

Suppose we are given a second order linear PDE to solve

$$a(x,y)u_{xx} + b(x,y)u_{xy} + c(x,y)u_{yy} = d(x,y)u_x + e(x,y)u_y + p(x,y)u + q(x,y) \quad (1)$$

The natural approach, after our experience with ordinary differential equations and with simple algebraic equations, is attempt a factorisation. Let's see how for this takes us.

We would expect factoring the left hand of (1) to give us an equivalent equation of the form

$$a(x,y)(D_x + \alpha_+(x,y)D_y)(D_x + \alpha_-(x,y)D_y)u$$

and we can immediately divide through by *a*. This suggests that those particular combinations of first order derivatives will play a special role.

Now, when studying first order PDE's we saw that such combinations were equivalent to the derivatives along characteristic curves. Effectively, we changed to a coordinate system defined by the characteristic curve and the initial curve.

Here, we have two combinations of first order derivatives each of which may define a different characteristic curve. If so, the two sets of characteristics will define a natural coordinate system for the problem, much as in the first order case.

In the new coordinates we will have

$$D_x + \alpha_+(x,y)D_y = D_r \quad D_x + \alpha_-(x,y)D_y = D_s$$

with each of the factors having become a differentiation along its respective characteristic curve, and the left hand side will become simply u_{rs} giving us an equation of the form

$$u_{rs} = A(r,s)u_r + B(r,s)u_s + C(r,s)u + D(r,s)$$

If *A*, *B*, and *C* all happen to be zero, the solution is obvious. If not, we can hope that the simpler form of the left hand side will enable us to make progress.

However, before we can do all this, we must see if (1) can actually be factored.

Multiplying out the factors gives

$$u_{xx} + \frac{b(x,y)}{a(x,y)}u_{xy} + \frac{c(x,y)}{a(x,y)}u_{yy} = u_{xx} + (\alpha_+ + \alpha_-)u_{xy} + \alpha_+\alpha_- u_{yy}$$

On comparing coefficients, and solving for the α's we see that they are the roots of

$$a(x,y)\alpha^2 + b(x,y)\alpha + c(x,y) = 0$$

Since we are discussing real functions, we are only interested in real roots, so the existence of the desired factorization will depend on the discriminant of this quadratic equation.

- If $b(x,y)^2 > 4a(x,y)c(x,y)$

  then we have two factors, and can follow the procedure outlined above. Equations like this are called *hyperbolic*

- If $b(x,y)^2 = 4a(x,y)c(x,y)$

  then we have only factor, giving us a single characteristic curve. It will be natural to use distance along these curves as one coordinate, but the second must be determined by other considerations.
  The same line of argument as before shows that use the characteristic curve this way gives a second order term of the form u_{rr}, where we've only taken the second derivative with respect to one of the two coordinates. Equations like this are called *parabolic*

- If $b(x,y)^2 < 4a(x,y)c(x,y)$

  then we have no real factors. In this case the best we can do is reduce the second order terms to the simplest possible form satisfying this inequality, i.e $u_{rr}+u_{ss}$
  It can be shown that this reduction is always possible. Equations like this are called *elliptic*

It can be shown that, just as for first order PDEs, discontinuities propagate along characteristics. Since elliptic equations have no real characteristics, this implies that any discontinuities they may have will be restricted to isolated points; i.e, that the solution is almost everywhere smooth.

This is not true for hyperbolic equations. Their behavior is largely controlled by the shape of their characteristic curves.

These differences mean different methods are required to study the three types of second equation. Fortunately, changing variables as indicated by the factorisation above lets us reduce any second order PDE to one in which the coefficients of the second order terms are constant, which means it is sufficient to consider only three standard equations.

$$u_{xx} + u_{yy} = 0 \quad u_{xx} - u_{yy} = 0 \quad u_{xx} - u_y = 0$$

We could also consider the cases where the right hand side of these equations is a given function, or proportional to u or to one of its first order derivatives, but all the essential properties of hyperbolic, parabolic, and elliptic equations are demonstrated by these three standard forms.

While we've only demonstrated the reduction in two dimensions, a similar reduction applies in higher dimensions, leading to a similar classification. We get, as the reduced form of the second order terms,

$$a_1 \frac{\partial^2 u}{\partial x_1^2} + a_2 \frac{\partial^2 u}{\partial x_2^2} + \cdots + a_n \frac{\partial^2 u}{\partial x_n^2}$$

where each of the a_is is equal to either 0, +1, or -1.

If *all* the a_is have the *same sign* the equation is *elliptic*

If *any* of the a_is are *zero* the equation is *parabolic*

If *exactly one* of the a_is has the *opposite sign* to the rest the equation is *hyperbolic*

In 2 or 3 dimensions these are the only possibilities, but in 4 or more dimensions there is a fourth possibility: *at least two* of the a_is are *positive*, and *at least two* of the a_is are *negative*.

Such equations are called *ultrahyperbolic*. They are less commonly encountered than the other three types, so will not be studied here.

When the coefficients are not constant, an equation can be hyperbolic in some regions of the *xy* plane, and elliptic in others. If so, different methods must be used for the solutions in the two regions.

Elliptic

Standard form, Laplace's equation:

$$\nabla^2 h = 0$$

Quote equation in spherical and cylindrical coordinates, and give full solution for cartesian and cylindrical coordinates. Note averaging property Comment on physical significance, rotation invariance of laplacian.

Hyperbolic

Standard form, wave equation:

$$\nabla^2 h = c^2 h_{tt}$$

Solution, any sum of functions of the form

$$h = f(\mathbf{k} \cdot \mathbf{x} - \omega t) \quad \omega = |\mathbf{k}| c$$

These are waves. Compare with solution from separating variables. Domain of dependance, etc.

Parabolic

The canonical parabolic equation is the diffusion equation:

$$\nabla^2 h = h_t$$

Here, we will consider some simple solutions of the one-dimensional case.

The properties of this equation are in many respects intermediate between those of hyperbolic and elliptic equation.

As with hyperbolic equations but not elliptic, the solution is well behaved if the value is given on the initial surface t=0.

However, the characteristic surfaces of this equation are the surfaces of constant *t*, thus there is no way for discontinuities to propagate to positive t.

Therefore, as with elliptic equations but not hyberbolic, the solutions are typically smooth, even when the initial conditions aren't.

Furthermore, at a local maximum of *h*, its Laplacian is negative, so *h* is decreasing with *t*, while at local minima, where the Laplacian will be positive, *h* will increase with *t*. Thus, initial variations in *h* will be smoothed out as *t* increases.

In one dimension, we can learn more by integrating both sides,

$$\begin{array}{rcl} \int_{-a}^{b} h_t dt & = & \int_{-a}^{b} h_{xx} dx \\ \frac{d}{dt}\int_{-a}^{b} h\, dt & = & \left[h_x\right]_{-a}^{b} \end{array}$$

Provided that h_x tends to zero for large *x*, we can take the limit as *a* and *b* tend to infinity, deducing

$$\frac{d}{dt}\int_{-\infty}^{\infty} h\, dt$$

so the integral of *h* over all space is constant.

This means this PDE can be thought of as describing some conserved quantity, initially concentrated but spreading out, or diffusing, over time.

This last result can be extended to two or more dimensions, using the theorems of vector calculus.

We can also differentiate any solution with respect to any coordinate to obtain another solution. E.g if *h* is a solution then

$$\nabla^2 h_x = \partial_x \nabla^2 h = \partial_x \partial_t h = \partial_t h_x$$

so h_x also satisfies the diffusion equation.

Similarity solution

Looking at this equation, we might notice that if we make the change of variables

$$r = \alpha x \quad \tau = \alpha^2 t$$

then the equation retains the same form. This suggests that the combination of variables x^2/t, which is unaffected by this variable change, may be significant.

We therefore assume this equation to have a solution of the special form

$$h(x,t) = f(\eta) \text{ where } \eta = \frac{x}{t^{\frac{1}{2}}}$$

then

$$h_x = \eta_x f_\eta = t^{-\frac{1}{2}} f_\eta \quad h_t = \eta_t f_\eta = -\frac{\eta}{2t} f_\eta$$

and substituting into the diffusion equation eventually gives

$$f_{\eta\eta} + \frac{\eta}{2} f_\eta = 0$$

which is an ordinary differential equation.

Integrating once gives

$$f_\eta = Ae^{-\frac{\eta^2}{4}}$$

Reverting to *h*, we find

$$\begin{aligned} h_x &= \frac{A}{\sqrt{t}} e^{-\frac{\eta^2}{4}} \\ h &= \frac{A}{\sqrt{t}} \int_{-\infty}^{x} e^{-s^2/4t} ds + B \\ &= A \int_{-\infty}^{x/2\sqrt{t}} e^{-z^2} dz + B \end{aligned}$$

This last integral can not be written in terms of elementary functions, but its values are well known.

In particular the limiting values of *h* at infinity are

$$h(-\infty, t) = B \quad h(\infty, t) = B + A\sqrt{\pi},$$

taking the limit as *t* tends to zero gives

$$h = \begin{cases} B & x < 0 \\ B + A\sqrt{\pi} & x > 0 \end{cases}$$

We see that the initial discontinuity is immediately smoothed out. The solution at later times retains the same shape, but is more stretched out.

The derivative of this solution with respect to *x*

$$h_x = \frac{A}{\sqrt{t}} e^{-x^2/4t}$$

is itself a solution, with *h* spreading out from its initial peak, and plays a significant role in the further analysis of this equation.

The same similiarity method can also be applied to some non-linear equations.

Separating variables

We can also obtain some solutions of this equation by separating variables.

$$h(x,t) = X(x)T(t) \Rightarrow X''T = X\dot{T}$$

giving us the two ordinary differential equations

$$\frac{d^2X}{dx^2} + k^2X = 0 \quad \frac{dT}{dt} = -kT$$

and solutions of the general form

$$h(x,t) = Ae^{-kt}\sin(kx + \alpha)$$

Extensions

Systems of Ordinary Differential Equations

We have already examined cases where we have a single differential equation and found several methods to aid us in finding solutions to these equations. But what happens if we have two or more differential equations that depend on each other? For example, consider the case where

$$D_t x(t) = 3y(t)^2 + x(t)t$$

and

$$D_t y(t) = x(t) + y(t)$$

Such a set of differential equations is said to be *coupled*. Systems of ordinary differential equations such as these are what we will look into in this section.

First order systems

A general system of differential equations can be written in the form

$$D_t \mathbf{x} = \mathbf{F}(\mathbf{x}, t)$$

Instead of writing the set of equations in a vector, we can write out each equation explicitly, in the form:

$$\begin{aligned} D_t x_1 &= F_1(x_1, \ldots, x_n, t) \\ &\vdots \\ D_t x_i &= F_i(x_1, \ldots, x_n, t) \end{aligned}$$

If we have the system at the very beginning, we can write it as:

$$D_t \mathbf{x} = \mathbf{G}(\mathbf{x}, t)$$

where

$$\mathbf{x} = (x(t), y(t)) = (x, y)$$

and

$$\mathbf{G}(\mathbf{x},t) = (3y^2 + xt, x + y)$$

or write each equation out as shown above.

Why are these forms important? Often, this arises as a single, higher order differential equation that is changed into a simpler form in a system. For example, with the same example,

$D_t x(t) = 3y(t)^2 + x(t)t$
$D_t y(t) = x(t) + y(t)$

we can write this as a higher order differential equation by simple substitution.

$D_t y(t) - y(t) = x(t)$

then

$D_t x(t) = 3y(t)^2 + (D_t y(t) - y(t))t$
$D_t x(t) = 3y(t)^2 + tD_t y(t) - ty(t)$

Notice now that the vector form of the system is dependent on *t* since

$$\mathbf{G}(\mathbf{x},t) = (3y^2 + xt, x + y)$$

the first component is dependent on *t*. However, if instead we had

$$\mathbf{H}(\mathbf{x}) = (3y^2 + x, x + y)$$

notice the vector field is no longer dependent on *t*. We call such systems *autonomous*. They appear in the form

$$D_t\mathbf{x} = \mathbf{H}(\mathbf{x})$$

We can convert between an autonomous system and a non-autonomous one by simply making a substitution that involves *t*, such as ***y***=(***x***, t), to get a system:

$$D_t\mathbf{y} = (\mathbf{F}(\mathbf{y}), 1) = (\mathbf{F}(\mathbf{x},t), 1)$$

In vector form, we may be able to separate ***F*** in a linear fashion to get something that looks like:

$$\mathbf{F}(\mathbf{x},t) = A(t)\mathbf{x} + \mathbf{b}(t)$$

where $A(t)$ is a matrix and ***b*** is a vector. The matrix could contain functions or constants, clearly, depending on whether the matrix depends on *t* or not.

Formal limits

In preliminary calculus, the definition of a limit is probably the most difficult concept to grasp (if nothing else, it took some of the most brilliant mathematicians 150 years to arrive at it); it is also the most important and most useful.

The intuitive definition of a limit is adequate for manipulation most of the time, but is inadequate to understand the concept, or to prove anything with it. The issue here lies with our meaning of "arbitrarily close". We discussed earlier that the meaning of this term is that the closer x gets to the specified value, the closer the function must get to the limit, so that however close we want the function to the limit, we can find a corresponding x close to our value. We can express this concept as follows:

Definition: (Formal definition of a limit)
Let $f(x)$ be a function defined on an open interval that contains $x=c$, except possibly at $x=c$. Let L be an existing number. Then we say that,

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

if, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that for all $x \in D_f$ when

$$0 < |x - c| < \delta,$$

we have

$$|f(x) - L| < \varepsilon.$$

To further explain, earlier we said that "however close we want the function to the limit, we can find a corresponding x close to our value." Using our new notation of epsilon (ε) and delta (δ), we mean that if we want to find $f(x)$ within ε of L, the limit, then we know that there is an x within δ of c that puts it there.

Again, since this is tricky; let's resume our example from before: $f(x) = x^2$, at $x = 2$. To start, let's say we want $f(x)$ to be within .01 of the limit. We know here that the limit should be 4, so we say; for $\varepsilon = .01$, there is some delta so that as long as $0 < |x - c| < \delta$, then $|f(x) - L| < \varepsilon$.

To show this, we can pick *any* delta that is bigger than 0. To be sure, you might pick .00000000000001, because you are absolutely sure that if x is within .00000000000001 of 2, then $f(x)$ will be within .01 of 4. Of course, we can't just pick a specific value for epsilon, like .01, because we said in our definition "for **every** $\varepsilon > 0$." This means that

we need to be able to show an infinite number of deltas, one for each epsilon. We can't list an infinite number of deltas!

Of course, we know of a very good way to do this; we simply create a function, so that for every epsilon, it can give us a delta. In this case, it's a rather easy function; all we need is $\delta(\varepsilon) < \sqrt{(\varepsilon)}$.

So how do you show that $f(x)$ tends to L as x tends to c? Well imagine somebody gave you a small number ε(e.g., say $\varepsilon = 0.03$). Then you have to find a $\delta > 0$ and show that whenever $0 < |x - c| < \delta$ we have $|f(x) - L| < 0.03$. Now if that person gave you a smaller ε(say $\varepsilon = 0.002$) then you would have to find another δ, but this time with 0.03 replaced by 0.002. If you can do this for *any* choice of ε then you have shown that $f(x)$ tends to L as x tends to c.

Definition: (Limit of a function at infinity)
We call L the **limit** of $f(x)$ as x approaches ∞ if for every number $\varepsilon > 0$ there exists a δ such that whenever $x > \delta$ we have

$$|f(x) - L| < \varepsilon$$

When this holds we write

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

or

$$f(x) \to L \quad as \quad x \to \infty.$$

Similarly, we call L the **limit** of $f(x)$ as x approaches $-\infty$ if for every number $\varepsilon > 0$, there exists a number δ such that whenever $x < \delta$ we have

$$|f(x) - L| < \varepsilon$$

When this holds we write

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

or

$$f(x) \to L \quad as \quad x \to -\infty.$$

Notice the difference in these two definitions. For the limit of $f(x)$ as x approaches ∞ we are interested in those x such that $x > \delta$. For the limit of $f(x)$ as x approaches $-\infty$ we are interested in those x such that $x < \delta$.

Examples

Here are some examples on finding limits using the definition.

1) What is δ when ε = 0.01 for

$$\lim_{x \to 8} \frac{x}{4} = 2?$$

we start with the conclusion and substitute the given values for f(x) and ε

$$\left| \frac{x}{4} - 2 \right| < 0.01$$

and simplify

$$7.96 < x < 8.04$$

using the first part of the definition of a limit

$$-0.04 < x - 8 < 0.04$$

we normally choose the smaller of |-0.04| and 0.04 for δ but any smaller number will work so δ=0.04

2) What is the limit of $f(x) = x + 7$ as x approaches 4?

There are two steps to answering such a question; first we must determine the answer — this is where intuition and guessing is useful, as well as the informal definition of a limit. Then, we must prove that the answer is right. For this problem, the answer happens to be 11. Now, we must prove it using the definition of a limit:

Informal: 11 is the limit because when x is roughly equal to 4, $f(x) = x + 7$ approximately equals 4 + 7, which equals 11.

Formal: We need to prove that no matter what value of ε is given to us, we can find a value of δ such that

$$|f(x) - 11| < \varepsilon$$

whenever

$$|x-4| < \delta.$$

For this particular problem, letting δ equal ε works (see choosing delta for help in determining the value of delta to use). Now, we have to prove

$$|f(x)-11| < \varepsilon$$

given that

$$|x-4| < \delta = \varepsilon.$$

Since $|x-4| < \varepsilon$, we know

$$|f(x)-11| = |x+7-11| = |x-4| < \varepsilon$$

which is what we wished to prove.

3) What is the limit of $f(x) = x^2$ as x approaches 4?

Formal: Again, we pull two things out of thin air; the limit is 16 (use the informal definition to find the limit of *f(x)*), and δ equals √(ε+16) - 4. Note that δ is always positive for positive ε. Now, we have to prove

$$\left|x^2-16\right| < \varepsilon$$

given that

$$|x-4| < \delta = \sqrt{\varepsilon+16} - 4.$$

We know that |x + 4| = |(x - 4) + 8| ≤ |x - 4| + 8 < δ + 8 (because of the triangle inequality), thus

$$\begin{aligned} |x^2-16| &= |x-4| \cdot |x+4| \\ &< (\delta) \cdot (\delta+8) \\ &< (\sqrt{16+\varepsilon}-4) \cdot (\sqrt{16+\varepsilon}+4). \\ &< (\sqrt{16+\varepsilon})^2 - 4^2 \\ &< \varepsilon \end{aligned}$$

4) Show that the limit of sin(1 / x) as x approaches 0 does not exist.

Suppose the limit exists and is *l*. We will proceed by contradiction. Assume that $l \neq 1$, the case for *l* = 1 is similar. Choose $\varepsilon = |l - 1|$, then for every δ > 0, there exists a large enough *n* such that $0 < x_0 = \frac{1}{\pi/2 + 2\pi n} < \delta$, but $|\sin(1/x_0) - l| = |1 - l| < \varepsilon$ a contradiction.

The function sin(1 / *x*) is known as the topologist's comb.

5) What is the limit of *x*sin(1 / *x*) as *x* approaches 0?

It is 0. For every $\varepsilon > 0$, choose $\delta = \varepsilon$ so that for all *x*, if 0 < | *x* | < δ, then $|x \sin(1/x) - 0| \leq |x| < \varepsilon$ as required.

Real numbers

Fields

You are probably already familiar with many different sets of numbers from your past experience. Some of the commonly used sets of numbers are

- Natural numbers, usually denoted with an **N**, are the numbers 0,1,2,3,...
- Integers, usually denoted with a **Z**, are the positive and negative natural numbers: ...-3,-2,-1,0,1,2,3...
- Rational numbers, denoted with a **Q**, are fractions of integers (excluding division by zero): -1/3, 5/1, 0, 2/7. etc.
- Real numbers, denoted with a **R**, are constructed and discussed below.

Note that different sets of numbers have different properties. In the set integers for example, any number always has an *additive inverse*: for any integer *x*, there is another integer *t* such that *x* + *t* = 0 This should not be terribly surprising: from basic arithmetic we know that *t* = − *x*. Try to prove to yourself that not all natural numbers have an additive inverse.

In mathematics, it is useful to note the important properties of each of these sets of numbers. The rational numbers, which will be of primary concern in constructing the real numbers, have the following properties:

There exists a number 0 such that for any other number *a*, 0+*a*=*a*+0=*a*
For any two numbers *a* and *b*, *a*+*b* is another number
For any three numbers *a,b*, and *c*, *a*+(*b*+*c*)=(*a*+*b*)+*c*
For any number *a* there is another number -*a* such that *a*+(-*a*)=0
For any two numbers *a* and *b*, *a*+*b*=*b*+*a*
For any two numbers *a* and *b*,*a***b* is another number

There is a number 1 such that for any number a, $a*1=1*a=a$
For any two numbers a and b, $a*b=b*a$
For any three numbers a,b and c, $a(bc)=(ab)c$
For any three numbers a,b and c, $a(b+c)=ab+ac$
For every number a there is another number a^{-1} *such that* $aa^{-1}=1$

As presented above, these may seem quite intimidating. However, these properties are nothing more than basic facts from arithmetic. Any collection of numbers (and operations + and * on those numbers) which satisfies the above properties is called a *field.* The properties above are usually called *field axioms.* As an exercise, determine if the integers form a field, and if not, which field axiom(s) they violate.

Even though the list of field axioms is quite extensive, it does not fully explore the properties of rational numbers. Rational numbers also have an *ordering.'* A *total ordering* must satisfy several properties: for any numbers *a, b,* and *c*

if $a \leq b$ and $b \leq a$ then $a = b$ (antisymmetry)
if $a \leq b$ and $b \leq c$ then $a \leq c$ (transitivity)
$a \leq b$ or $b \leq a$ (totality)

To familiarize yourself with these properties, try to show that (a) natural numbers, integers and rational numbers are all totally ordered and more generally (b) convince yourself that any collection of rational numbers are totally ordered (note that the integers and natural numbers are both collections of rational numbers).

Finally, it is useful to recognize one more characterization of the rational numbers: every rational number has a decimal expansion which is either repeating or terminating. The proof of this fact is omitted, however it follows from the definition of each rational number as a fraction. When performing long division, the remainder at any stage can only take on positive integer values smaller than the denominator, of which there are finitely many.

Constructing the Real Numbers

There are two additional tools which are needed for the construction of the real numbers: the upper bound and the least upper bound. **Definition** A collection of numbers E is bounded above if there exists a number m such that for all x in E $x \leq m$. Any number m which satisfies this condition is called an upper bound of the set E.

Definition If a collection of numbers E is bounded above with m as an upper bound of E, and all other upper bounds of E are bigger than m, we call m the *least upper bound* or *supremum* of E, denoted by sup E.

Many collections of rational numbers do not have a least upper bound which is also rational, although some do. Suppose the the numbers 5 and 10/3 are, together, taken to be E. The number 10/3 is not only an upper bound of E, it is a least upper bound. In general,

there are many upper bounds (12, for instance, is an upper bound of the collection above), but there can be at most one least upper bound.

Consider the collection of numbers $\{3, 3.1, 3.14, 3.141, 3.1415, \ldots\}$: You may recognize these decimals as the first few digits of pi. Since each decimal terminates, each number in this collection is a rational number. This collection has infinitely many upper bounds. The number 4, for instance, is an upper bound. There is no least upper bound, at least not in the rational numbers. Try to convince yourself of this fact by attempting to construct such a least upper bound: (a) why does pi not work as a least upper bound (hint: pi does not have a repeating or terminating decimal expansion), (b) what happens if the proposed supremum is equal to pi up to some decimal place, and zeros after (c) if the proposed supremum is bigger than pi, can you find a smaller upper bound which will work?

In fact, there are infinitely many collections of rational numbers which do not have a rational least upper bound. We define a real number to be any number that is the least upper bound of some collection of rational numbers.

Properties of Real Numbers

The reals are well ordered.

> For all reals; *a*, *b*, *c*
> Either *b*>*a*, *b*=*a', or* b<a.
> If *a*<*b* and *b*<*c* then *a*<*c*

Also

> *b*>*a* implies *b*+*c*>*a*+*c*
> *b*>*a* and *c*>*0* implies *bc*>*ac*
> *b*>*a* implies -*a*>-*b*

Upper bound axiom

> Every non-empty set of real numbers which is bounded above has a supremum.

The upper bound axiom is necessary for calculus. It is not true for rational numbers.

We can also define lower bounds in the same way.

Definition A set *E* is bounded below if there exists a real M such that for all $x \in E$ $x \geq M$
Any *M* which satisfies this condition is called an lower bound of the set *E*

Definition If a set, *E*, is bounded below, *M* is an lower bound of *E*, and all other lower bounds of *E* are less than *M*, we call *M* the *greatest lower bound* or *inifimum* of *E*, denoted by inf *E*

The supremum and infimum of finite sets are the same as their maximum and minimum.

Theorem

Every non-empty set of real numbers which is bounded below has an infimum.

Proof:

Let E be a non-empty set of of real numbers, bounded below
Let L be the set of all lower bounds of E
L is not empty, by definition of bounded below
Every element of E is an upper bound to the set L, by definition
Therefore, L is a non empty set which is bounded above
L has a supremum, by the upper bound axiom
1/ Every lower bound of E is ≤sup L, by definition of supremum
Suppose there were an *e*□E such that e<sup L
Every element of L is ≤e, by definition
Therefore e is an upper bound of L and e<sup L
This contradicts the definition of supremum, so there can be no such e.
If e□E then e≥sup L, proved by contradiction
2/ Therefore, sup L is a lower bound of E
inf E exists, and is equal to sup L, on comparing definition of infinum to lines 1 & 2

Bounds and inequalities, theorems:
$A \subseteq B \Rightarrow \sup A \leq \sup BA \subseteq B \Rightarrow \inf A \geq \inf B$
$\sup A \cup B = \max(\sup A, \sup B)\inf A \cup B = \min(\inf A, \inf B)$

Theorem: *(The triangle inequality)*

$$\forall a, b, c \in \mathbb{R} \quad |a - b| \leq |a - c| + |c - b|$$

Proof by considering cases

If a≤b≤c then |*a*-*c*|+|*c*-*b*| = (c-a)+(c-b) = 2(c-b)+(b-a)>b-a = |b-a|

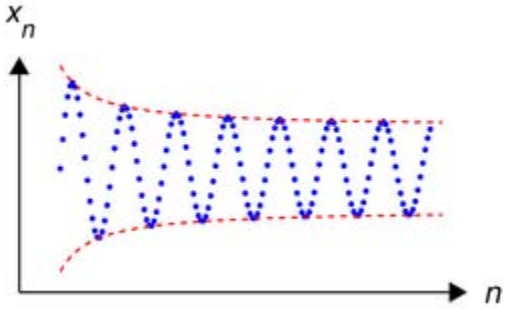
Exercise: Prove the other five cases.

This theorem is a special case of the triangle inequality theorem from geometry: The sum of two sides of a triangle is greater than or equal to the third side. It is useful whenever we need to manipulate inequalities and absolute values.

Theory of Sequences

A **sequence** is an ordered list of objects (or events). Like a set, it contains members (also called *elements* or *terms*), and the number of terms (possibly infinite) is called the *length* of the sequence. Unlike a set, order matters, and the exact same elements can appear multiple times at different positions in the sequence.

For example, (C, R, Y) is a sequence of letters that differs from (Y, C, R), as the ordering matters. Sequences can be *finite*, as in this example, or *infinite*, such as the sequence of all even positive integers (2, 4, 6,...).



An infinite sequence of real numbers (in blue). This sequence is neither increasing, nor decreasing, nor convergent. It is however bounded.

Examples and notation

There are various and quite different notions of sequences in mathematics, some of which (e.g., exact sequence) are not covered by the notations introduced below.

A sequence may be denoted $(a_1, a_2, ...)$. For shortness, the notation (a_n) is also used.

A more formal definition of a **finite sequence** with terms in a set S is a function from $\{1, 2, ..., n\}$ to S for some $n \geq 0$. An **infinite sequence** in S is a function from $\{1, 2, ...\}$ (the set of natural numbers without 0) to S.

Sequences may also start from 0, so the first term in the sequence is then a_0.

A finite sequence is also called an n-tuple. Finite sequences include the *empty sequence* ( ) that has no elements.

A function from all integers into a set is sometimes called a **bi-infinite sequence**, since it may be thought of as a sequence indexed by negative integers grafted onto a sequence indexed by positive integers.

Types and properties of sequences

A subsequence of a given sequence is a sequence formed from the given sequence by deleting some of the elements without disturbing the relative positions of the remaining elements.

If the terms of the sequence are a subset of an ordered set, then a *monotonically increasing* sequence is one for which each term is greater than or equal to the term before it; if each term is strictly greater than the one preceding it, the sequence is called *strictly monotonically increasing*. A monotonically decreasing sequence is defined similarly. Any sequence fulfilling the monotonicity property is called monotonic or *monotone*. This is a special case of the more general notion of monotonic function.

The terms *non-decreasing* and *non-increasing* are used in order to avoid any possible confusion with strictly increasing and strictly decreasing, respectively. If the terms of a sequence are integers, then the sequence is an integer sequence. If the terms of a sequence are polynomials, then the sequence is a polynomial sequence.

If *S* is endowed with a topology, then it becomes possible to consider *convergence* of an infinite sequence in *S*. Such considerations involve the concept of the limit of a sequence.

Sequences in analysis

In analysis, when talking about sequences, one will generally consider sequences of the form

$$(x_1, x_2, x_3, ...)_{\text{or}}\,(x_0, x_1, x_2, ...)$$

which is to say, infinite sequences of elements indexed by natural numbers. (It may be convenient to have the sequence start with an index different from 1 or 0. For example, the sequence defined by $x_n = 1/\log(n)$ would be defined only for $n \geq 2$. When talking about such infinite sequences, it is usually sufficient (and does not change much for most considerations) to assume that the members of the sequence are defined at least for all indices large enough, that is, greater than some given *N*.)

The most elementary type of sequences are numerical ones, that is, sequences of real or complex numbers.

Summation notation

Summation notation allows an expression that contains a sum to be expressed in a simple, compact manner. The uppercase Greek letter sigma, Σ, is used to denote the sum of a set of numbers.

Example

$$\sum_{i=3}^{7} i^2 = 3^2 + 4^2 + 5^2 + 6^2 + 7^2$$

Let *f* be a function and *N,M* are integers with $N < M$. Then

$$\sum_{i=N}^{M} f(i) = f(N) + f(N+1) + f(N+2) + \cdots + f(M).$$

We say N is the lower limit and M is the upper limit of the sum.

We can replace the letter i with any other variable. For this reason i is referred to as a *dummy variable*. So

$$\sum_{i=1}^{4} i = \sum_{j=1}^{4} j = \sum_{\alpha=1}^{4} \alpha = 1 + 2 + 3 + 4$$

Conventionally we use the letters i, j, k, m for dummy variables.

Example

$$\sum_{i=1}^{5} i = 1 + 2 + 3 + 4 + 5$$

Here, the *dummy variable* is i, the *lower limit* of summation is 1, and the *upper limit* is 5.

Example

Sometimes, you will see summation signs with no dummy variable specified, e.g.,

$$\sum_{1}^{4} i^3 = 100$$

In such cases the correct dummy variable should be clear from the context.

You may also see cases where the limits are unspecified. Here too, they must be deduced from the context.

Common summations

$$\sum_{i=1}^{n} c = c + c + \ldots + c = nc, c \in \mathbb{R}$$

$$\sum_{i=1}^{n} i = 1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}$$

$$\sum_{i=1}^{n} i^2 = 1^2 + 2^2 + 3^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{i=1}^{n} i^3 = 1^3 + 2^3 + 3^3 + ... + n^3 = \frac{n^2(n+1)^2}{4}$$

Tables of Integrals

Rules

- $\int cf(x)\,dx = c\int f(x)\,dx$
- $\int f(x) + g(x)\,dx = \int f(x)\,dx + \int g(x)\,dx$
- $\int f(x) - g(x)\,dx = \int f(x)\,dx - \int g(x)\,dx$
- $\int u\,dv \ = uv - \int v\,du$

Powers

- $\int dx = x + C$
- $\int a\,dx = ax + C$
- $\int x^n\,dx = \frac{1}{n+1}x^{n+1} + C \qquad \text{if } n \neq -1$
- $\int x^{-n}\,dx = \frac{1}{-n+1}x^{-n+1} + C \qquad \text{if } n \neq 1$
- $\int \frac{1}{x}\,dx = \ln|x| + C$
- $\int \frac{1}{ax+b}\,dx = \frac{1}{a}\ln|ax+b| + C \qquad \text{if } a \neq 0$

Trigonometric Functions

Basic Trigonometric Functions

- $\int \sin x\,dx = -\cos x + C$
- $\int \cos x\,dx = \sin x + C$
- $\int \tan x\,dx = \ln|\sec x| + C$

- $\int \sin^2 x\,dx = \frac{1}{2}x - \frac{1}{4}\sin 2x + C$
- $\int \cos^2 x\,dx = \frac{1}{2}x + \frac{1}{4}\sin 2x + C$
- $\int \tan^2 x\,dx = \tan(x) - x + C$

Reciprocal Trigonometric Functions

- $\int \sec x\,dx = \ln|\sec x + \tan x| + C = \ln\left|\tan\left(\frac{1}{2}x + \frac{1}{4}\pi\right)\right| + C$
- $\int \csc x\,dx = -\ln|\csc x + \cot x| + C = \ln\left|\tan\left(\frac{1}{2}x\right)\right| + C$
- $\int \cot x\,dx = \ln|\sin x| + C$

- $\int \sec^2 kx\,dx = \frac{1}{k}\tan kx + C$
- $\int \csc^2 kx\,dx = -\frac{1}{k}\cot kx + C$
- $\int \cot^2 kx\,dx = -x - \frac{1}{k}\cot kx + C$

- $\int \sec x\tan x\,dx = \sec x + C$
- $\int \csc^n kx\cot kx\,dx = -\frac{1}{ka}\csc^n ax + C$
- $\int \sec x\csc x\,dx = \ln|\tan x| + C$

Inverse Trigonometric Functions

- $\int \frac{1}{\sqrt{1-x^2}}\,dx = \arcsin(x) + C$
- $\int \frac{1}{\sqrt{a^2-x^2}}\,dx = \arcsin(x/a) + C \qquad \text{if } a \neq 0$
- $\int \frac{1}{1+x^2}\,dx = \arctan(x) + C$

- $\int \frac{1}{a^2+x^2}\,dx = \frac{1}{a}\arctan(x/a) + C \qquad \text{if } a \neq 0$

Exponential and Logarithmic Functions

- $\int e^x\,dx = e^x + C$
- $\int e^{ax}\,dx = \frac{1}{a}e^{ax} + C \qquad \text{if } a \neq 0$
- $\int a^x\,dx = \frac{1}{\ln a}a^x + C \qquad \text{if } a > 0, a \neq 1$
- $\int \ln x\,dx = x\ln x - x + C$

Inverse Trigonometric Functions

- $\int \arcsin(x)\,dx = x\arcsin(x) + \sqrt{1-x^2} + C$
- $\int \arccos(x)\,dx = x\arccos(x) - \sqrt{1-x^2} + C$
- $\int \arctan(x)\,dx = x\arctan(x) - \frac{1}{2}\ln(1+x^2) + C$

Tables of Derivatives

General Rules

$$\frac{d}{dx}(f+g) = \frac{df}{dx} + \frac{dg}{dx}$$

$$\frac{d}{dx}(cf) = c\frac{df}{dx}$$

$$\frac{d}{dx}(fg) = f\frac{dg}{dx} + g\frac{df}{dx}$$

$$\frac{d}{dx}\left(\frac{f}{g}\right) = \frac{g\frac{df}{dx} - f\frac{dg}{dx}}{g^2}$$

Powers and Polynomials

- $\frac{d}{dx}(c) = 0$
- $\frac{d}{dx}x = 1$
- $\frac{d}{dx}x^n = nx^{n-1}$
- $\frac{d}{dx}\sqrt{x} = \frac{1}{2\sqrt{x}}$
- $\frac{d}{dx}\frac{1}{x} = -\frac{1}{x^2}$
- $\frac{d}{dx}(c_n x^n + c_{n-1}x^{n-1} + c_{n-2}x^{n-2} + \cdots + c_2 x^2 + c_1 x + c_0) = nc_n x^{n-1} + (n-1)c_{n-1}x^{n-}$

Trigonometric Functions

$$\frac{d}{dx}\sin(x) = \cos(x)$$

$$\frac{d}{dx}\cos(x) = -\sin(x)$$

$$\frac{d}{dx}\tan(x) = \sec^2(x)$$

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

$$\frac{d}{dx}\sec(x) = \sec(x)\tan(x)$$

$$\frac{d}{dx}\csc(x) = -\csc(x)\cot(x)$$

Exponential and Logarithmic Functions

- $\frac{d}{dx}e^x = e^x$

- $\frac{d}{dx} a^x = a^x \ln(a) \qquad \text{if } a > 0$
- $\frac{d}{dx} \ln(x) = \frac{1}{x}$
- $\frac{d}{dx} \log_a(x) = \frac{1}{x \ln(a)} \qquad \text{if } a > 0, a \neq 1$
- $(f^g)' = \left(e^{g \ln f}\right)' = f^g \left(f' \frac{g}{f} + g' \ln f\right), \qquad f > 0$
- $(c^f)' = \left(e^{f \ln c}\right)' = f' c^f \ln c$

Inverse Trigonometric Functions

- $\frac{d}{dx} \arcsin \mathrm{x} = \frac{1}{\sqrt{1 - x^2}}$
- $\frac{d}{dx} \arccos \mathrm{x} = -\frac{1}{\sqrt{1 - x^2}}$
- $\frac{d}{dx} \arctan \mathrm{x} = \frac{1}{1 + x^2}$
- $\frac{d}{dx} \operatorname{arcsec} x = \frac{1}{|x| \sqrt{x^2 - 1}}$
- $\frac{d}{dx} \operatorname{arccot} x = \frac{-1}{1 + x^2}$
- $\frac{d}{dx} \operatorname{arccsc} x = \frac{-1}{|x| \sqrt{x^2 - 1}}$

Hyperbolic and Inverse Hyperbolic Functions

$$\frac{d}{dx} \sinh x = \cosh x$$
$$\frac{d}{dx} \cosh x = \sinh x$$
$$\frac{d}{dx} \tanh x = \operatorname{sech}^2 x$$
$$\frac{d}{dx} \operatorname{sech} x = -\tanh x \operatorname{sech} x$$
$$\frac{d}{dx} \coth x = -\operatorname{csch}^2 x$$

$$\frac{d}{dx}\operatorname{csch} x = -\coth x \operatorname{csch} x$$
$$\frac{d}{dx}\sinh^{-1} x = \frac{1}{\sqrt{x^2+1}}$$
$$\frac{d}{dx}\cosh^{-1} x = \frac{-1}{\sqrt{x^2-1}}$$
$$\frac{d}{dx}\tanh^{-1} x = \frac{1}{1-x^2}$$
$$\frac{d}{dx}\operatorname{sech}^{-1} x = \frac{1}{x\sqrt{1-x^2}}$$
$$\frac{d}{dx}\coth^{-1} x = \frac{-1}{1-x^2}$$
$$\frac{d}{dx}\operatorname{csch}^{-1} x = \frac{-1}{|x|\sqrt{1+x^2}}$$

Table of Trigonometry

- $\tan(x) = \dfrac{\sin x}{\cos x}$
- $\sec(x) = \dfrac{1}{\cos x}$
- $\cot(x) = \dfrac{\cos x}{\sin x} = \dfrac{1}{\tan x}$
- $\csc(x) = \dfrac{1}{\sin x}$

Pythagorean Identities

- $\sin^2 x + \cos^2 x = 1$
- $1 + \tan^2(x) = \sec^2 x$
- $1 + \cot^2(x) = \csc^2 x$

Double Angle Identities

- $\sin(2x) = 2\sin x \cos x$
- $\cos(2x) = \cos^2 x - \sin^2 x$
- $\tan(2x) = \dfrac{2\tan(x)}{1-\tan^2(x)}$

- $\cos^2(x) = \dfrac{1 + \cos(2x)}{2}$
- $\sin^2(x) = \dfrac{1 - \cos(2x)}{2}$

Angle Sum Identities

$$\sin(x + y) = \sin x \cos y + \cos x \sin y$$
$$\sin(x - y) = \sin x \cos y - \cos x \sin y$$
$$\cos(x + y) = \cos x \cos y - \sin x \sin y$$
$$\cos(x - y) = \cos x \cos y + \sin x \sin y$$
$$\sin x + \sin y = 2 \sin\left(\frac{x + y}{2}\right) \cos\left(\frac{x - y}{2}\right)$$
$$\sin x - \sin y = 2 \cos\left(\frac{x + y}{2}\right) \sin\left(\frac{x - y}{2}\right)$$
$$\cos x + \cos y = 2 \cos\left(\frac{x + y}{2}\right) \cos\left(\frac{x - y}{2}\right)$$
$$\cos x - \cos y = -2 \sin\left(\frac{x + y}{2}\right) \sin\left(\frac{x - y}{2}\right)$$
$$\tan x + \tan y = \frac{\sin(x + y)}{\cos x \cos y}$$
$$\tan x - \tan y = \frac{\sin(x - y)}{\cos x \cos y}$$
$$\cot x + \cot y = \frac{\sin(x + y)}{\sin x \sin y}$$
$$\cot x - \cot y = \frac{-\sin(x - y)}{\sin x \sin y}$$

Product-to-sum identities

$$\cos(x)\cos(y) = \frac{\cos(x + y) + \cos(x - y)}{2}$$
$$\sin(x)\sin(y) = \frac{\cos(x - y) - \cos(x + y)}{2}$$
$$\sin(x)\cos(y) = \frac{\sin(x + y) + \sin(x - y)}{2}$$
$$\cos(x)\sin(y) = \frac{\sin(x + y) - \sin(x - y)}{2}$$