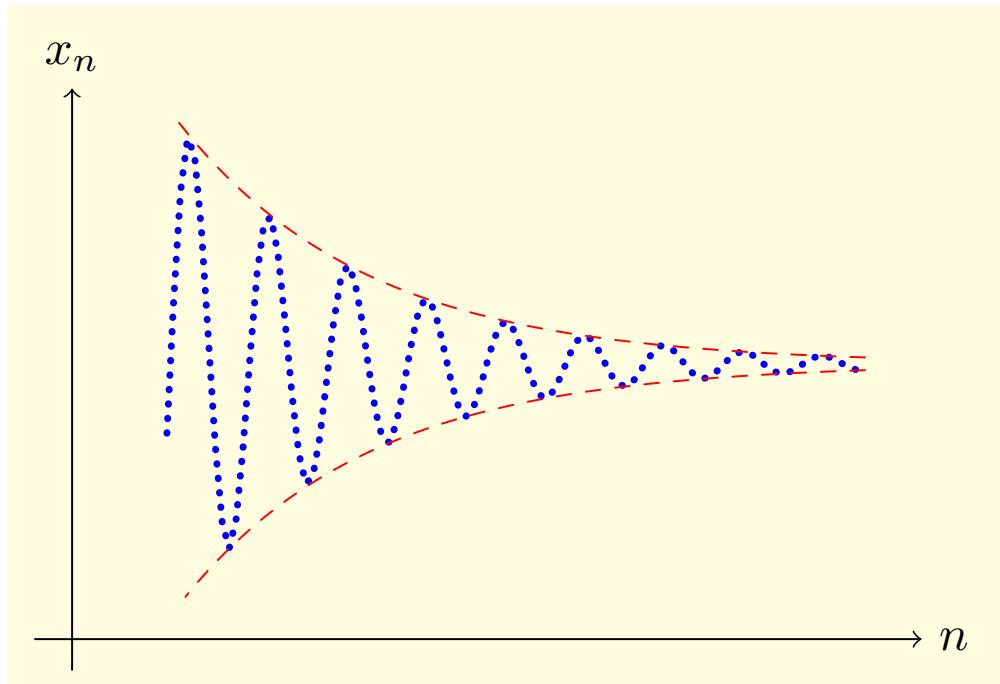


Math 147
Honours Calculus 1
Advanced
Course Notes

Barbara A. Forrest and Brian E. Forrest



Fall 2020/ Version 1.0

Copyright © Barbara A. Forrest and Brian E. Forrest.

All rights reserved.

August 1, 2019

All rights, including copyright and images in the content of these course notes, are owned by the course authors Barbara Forrest and Brian Forrest. By accessing these course notes, you agree that you may only use the content for your own personal, non-commercial use. You are not permitted to copy, transmit, adapt, or change in any way the content of these course notes for any other purpose whatsoever without the prior written permission of the course authors.

Author Contact Information:

Barbara Forrest (baforres@uwaterloo.ca)

Brian Forrest (beforres@uwaterloo.ca)

QUICK REFERENCE PAGE 1

Right Angle Trigonometry

$$\sin \theta = \frac{\text{opposite}}{\text{hypotenuse}}$$

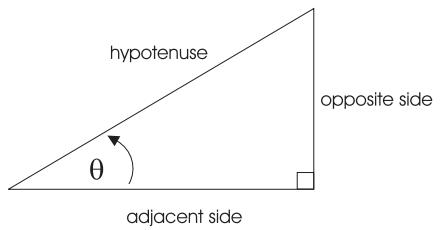
$$\cos \theta = \frac{\text{adjacent}}{\text{hypotenuse}}$$

$$\tan \theta = \frac{\text{opposite}}{\text{adjacent}}$$

$$\csc \theta = \frac{1}{\sin \theta}$$

$$\sec \theta = \frac{1}{\cos \theta}$$

$$\cot \theta = \frac{1}{\tan \theta}$$

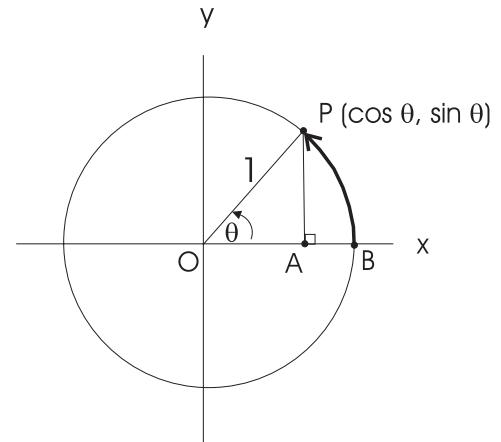


Radians

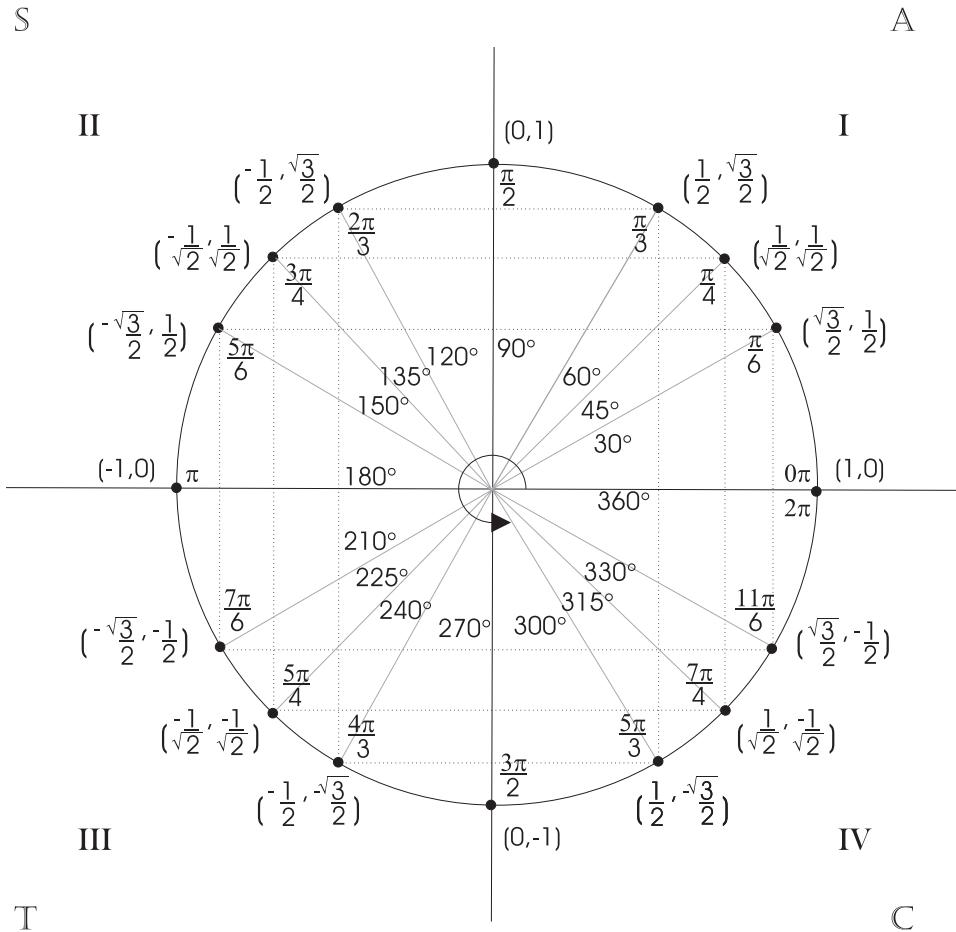
The angle θ in radians equals the length of the directed arc BP , taken positive counter-clockwise and negative clockwise. Thus, π radians = 180° or $1 \text{ rad} = \frac{180}{\pi}$.

Definition of Sine and Cosine

For any θ , $\cos \theta$ and $\sin \theta$ are defined to be the x - and y -coordinates of the point P on the unit circle such that the radius OP makes an angle of θ radians with the positive x -axis. Thus $\sin \theta = AP$, and $\cos \theta = OA$.



The Unit Circle



QUICK REFERENCE PAGE 2

Trigonometric Identities

<i>Pythagorean Identity</i>	$\cos^2 \theta + \sin^2 \theta = 1$
<i>Range</i>	$-1 \leq \cos \theta \leq 1$ $-1 \leq \sin \theta \leq 1$
<i>Periodicity</i>	$\cos(\theta \pm 2\pi) = \cos \theta$ $\sin(\theta \pm 2\pi) = \sin \theta$
<i>Symmetry</i>	$\cos(-\theta) = \cos \theta$ $\sin(-\theta) = -\sin \theta$

Sum and Difference Identities

$$\begin{aligned}\cos(A + B) &= \cos A \cos B - \sin A \sin B \\ \cos(A - B) &= \cos A \cos B + \sin A \sin B \\ \sin(A + B) &= \sin A \cos B + \cos A \sin B \\ \sin(A - B) &= \sin A \cos B - \cos A \sin B\end{aligned}$$

Complementary Angle Identities

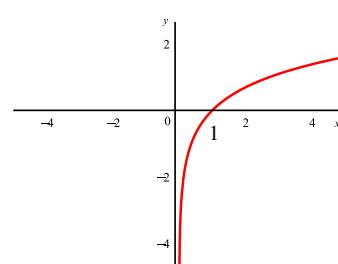
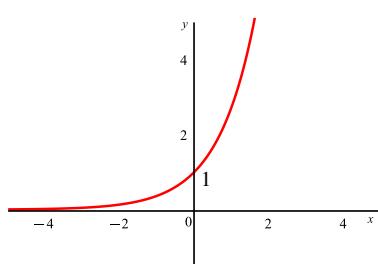
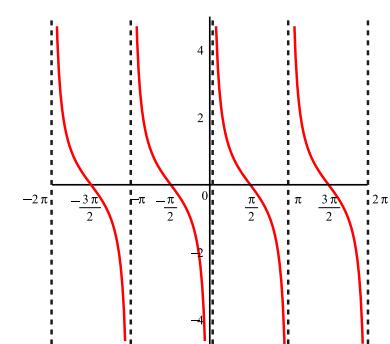
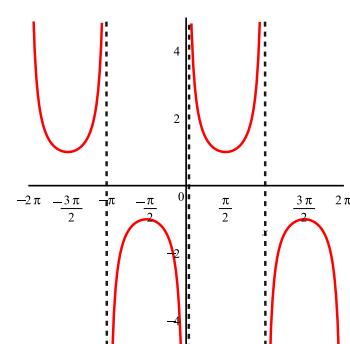
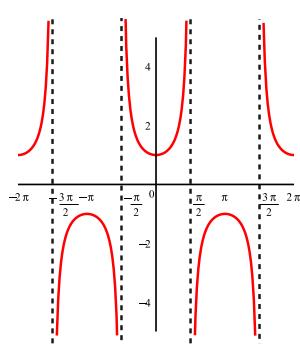
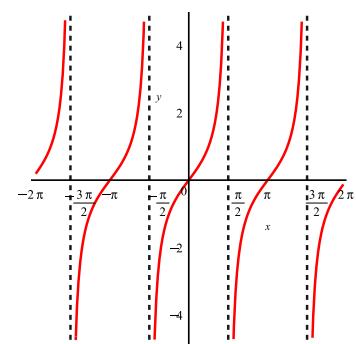
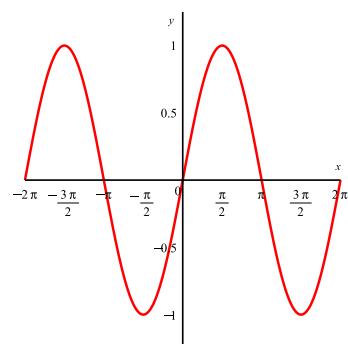
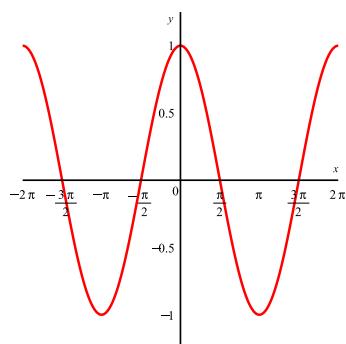
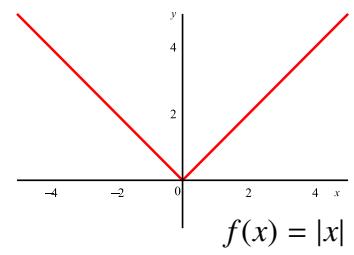
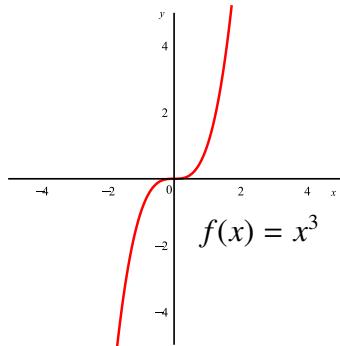
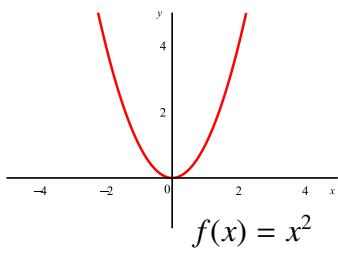
$$\begin{aligned}\cos\left(\frac{\pi}{2} - A\right) &= \sin A \\ \sin\left(\frac{\pi}{2} - A\right) &= \cos A\end{aligned}$$

<i>Double-Angle Identities</i>	$\cos 2A = \cos^2 A - \sin^2 A$
	$\sin 2A = 2 \sin A \cos A$

<i>Half-Angle Identities</i>	$\cos^2 \theta = \frac{1+\cos 2\theta}{2}$
	$\sin^2 \theta = \frac{1-\cos 2\theta}{2}$

<i>Other</i>	$1 + \tan^2 A = \sec^2 A$
--------------	---------------------------

QUICK REFERENCE PAGE 3



QUICK REFERENCE PAGE 4

Differentiation Rules

Function	Derivative
$f(x) = cx^a$, $a \neq 0$, $c \in \mathbb{R}$	$f'(x) = cax^{a-1}$
$f(x) = \sin(x)$	$f'(x) = \cos(x)$
$f(x) = \cos(x)$	$f'(x) = -\sin(x)$
$f(x) = \tan(x)$	$f'(x) = \sec^2(x)$
$f(x) = \sec(x)$	$f'(x) = \sec(x) \tan(x)$
$f(x) = \arcsin(x)$	$f'(x) = \frac{1}{\sqrt{1-x^2}}$
$f(x) = \arccos(x)$	$f'(x) = -\frac{1}{\sqrt{1-x^2}}$
$f(x) = \arctan(x)$	$f'(x) = \frac{1}{1+x^2}$
$f(x) = e^x$	$f'(x) = e^x$
$f(x) = a^x$ with $a > 0$	$f'(x) = a^x \ln(a)$
$f(x) = \ln(x)$ for $x > 0$	$f'(x) = \frac{1}{x}$

Table of Antiderivatives

$\int x^n dx = \frac{x^{n+1}}{n+1} + C$
$\int \frac{1}{x} dx = \ln(x) + C$
$\int e^x dx = e^x + C$
$\int \sin(x) dx = -\cos(x) + C$
$\int \cos(x) dx = \sin(x) + C$
$\int \sec^2(x) dx = \tan(x) + C$
$\int \frac{1}{1+x^2} dx = \arctan(x) + C$
$\int \frac{1}{\sqrt{1-x^2}} dx = \arcsin(x) + C$
$\int \frac{-1}{\sqrt{1-x^2}} dx = \arccos(x) + C$
$\int \sec(x) \tan(x) dx = \sec(x) + C$
$\int a^x dx = \frac{a^x}{\ln(a)} + C$

n -th degree Taylor polynomial for f centered at $x = a$

$$\begin{aligned}
 T_{n,a}(x) &= \sum_{k=0}^n \frac{f^{(k)}(a)}{k!} (x-a)^k \\
 &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n
 \end{aligned}$$

Linear Approximations ($L_0(x)$) and Taylor Polynomials ($T_{n,0}(x)$)

$f(x) = e^x$	$L_0(x) = T_{1,0}(x) = f(0) + f'(0)(x-0) = e^0 + e^0(x) = 1 + x$ $T_{2,0}(x) = f(0) + f'(0)(x-0) + \frac{f''(0)}{2!}(x-0)^2 = e^0 + e^0(x) + \frac{e^0}{2!}(x-0)^2 = 1 + x + \frac{x^2}{2}$ $T_{3,0}(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6}$ $T_{4,0}(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24}$
$f(x) = \sin(x)$	$L_0(x) = T_{1,0}(x) = x$ $T_{2,0}(x) = x$ $T_{3,0}(x) = x - \frac{x^3}{6}$ $T_{4,0}(x) = x - \frac{x^3}{6}$
$f(x) = \cos(x)$	$L_0(x) = T_{1,0}(x) = 1$ $T_{2,0}(x) = 1 - \frac{x^2}{2}$ $T_{3,0}(x) = 1 - \frac{x^2}{2}$ $T_{4,0}(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24}$

Table of Contents

	Page
1 A Short Introduction to Mathematical Logic and Proof	1
1.1 Basic Notions of Mathematical Logic and Truth Tables	1
1.2 Variables and Quantifiers	4
1.3 Rules of Inference and the Foundations of Proof	7
2 Sets, Relations and Functions	11
2.0 Notation	11
2.1 Sets, Products and Relations	12
2.2 Products of Sets	14
2.3 Relations and Functions	19
2.3.1 Relations	19
2.3.2 Functions	22
2.4 Composition of Functions	26
2.5 Transformations of Functions	28
2.5.1 Translations	28
2.5.2 Scaling	33
2.5.3 Reflections	35
2.5.4 Symmetries: Even and Odd Functions	37
2.6 Inverse Functions	40
2.6.1 One-to-one and Onto Functions	40
2.6.2 Inverse Functions	44
2.6.3 Graphing Inverse Functions	50
2.6.4 Inverse Trigonometric Functions	51
2.7 Pullback	54
2.8 Boolean Algebra and Sets: Enrichment	55
2.9 Principle of Mathematical Induction	58
2.10 Mathematical Induction	58
2.10.1 The Tower of Hanoi: Enrichment	66
3 Real Numbers	67
3.1 Absolute Values	67
3.1.1 Inequalities Involving Absolute Values	69
3.2 Least Upper Bound Property	73
3.3 Archimedean Property	77
4 Sequences and Convergence	80
4.1 Sequences and Their Limits	80
4.1.1 Introduction to Sequences	80
4.1.2 Recursively Defined Sequences	82

4.1.3	Subsequences and Tails	89
4.1.4	Limits of Sequences	90
4.1.5	Divergence to $\pm\infty$	100
4.1.6	Arithmetic for Limits of Sequences	101
4.2	Squeeze Theorem	115
4.3	Monotone Convergence Theorem	117
4.4	Introduction to Series	122
4.4.1	Geometric Series	126
4.4.2	Divergence Test	128
4.5	Bolzano-Weierstrass Theorem	132
4.6	Limit Points	137
4.7	Cauchy Sequences	138
5	Limits and Continuity	144
5.1	Introduction to Limits for Functions	144
5.2	Sequential Characterization of Limits	154
5.2.1	Three More Strange Functions	159
5.3	Arithmetic Rules for Limits of Functions	165
5.4	One-sided Limits	170
5.5	The Squeeze Theorem	173
5.6	The Fundamental Trigonometric Limit	177
5.7	Limits at Infinity and Asymptotes	181
5.7.1	Asymptotes and Limits at Infinity	182
5.7.2	Fundamental Log Limit	188
5.7.3	Vertical Asymptotes and Infinite Limits	193
5.8	Continuity	198
5.8.1	Types of Discontinuities	200
5.8.2	Continuity of Polynomials, $\sin(x)$, $\cos(x)$, e^x and $\ln(x)$. . .	203
5.8.3	Arithmetic Rules for Continuous Functions	206
5.8.4	Continuity on an Interval	209
5.9	Intermediate Value Theorem	210
5.9.1	Approximate Solutions of Equations	216
5.9.2	The Bisection Method	220
5.10	Extreme Value Theorem	223
5.11	Uniform Continuity	228
5.11.1	Sequential Characterization of Uniform Continuity . . .	231
5.11.2	Uniform Continuity on $[a, b]$	234
5.12	Curve Sketching: Part 1	236
6	Derivatives	239
6.1	Instantaneous Velocity	239
6.2	Definition of the Derivative	241
6.2.1	The Tangent Line	243
6.2.2	Differentiability versus Continuity	244
6.3	The Derivative Function	247
6.4	Derivatives of Elementary Functions	249
6.4.1	The Derivative of $\sin(x)$ and $\cos(x)$	251
6.4.2	The Derivative of e^x	253
6.5	Tangent Lines and Linear Approximation	256

6.5.1	The Error in Linear Approximation	260
6.5.2	Applications of Linear Approximation	263
6.6	Newton's Method	267
6.7	Arithmetic Rules of Differentiation	272
6.8	The Chain Rule	276
6.8.1	Proof of the Chain Rule	281
6.9	Derivatives of Other Trigonometric Functions	282
6.10	Derivatives of Inverse Functions	284
6.10.1	The Proof of the Inverse Function Theorem	290
6.11	Derivatives of Inverse Trigonometric Functions	295
6.12	Implicit Differentiation	301
6.13	Local Extrema	308
6.13.1	The Local Extrema Theorem	310
6.14	Related Rates	314
7	The Mean Value Theorem	320
7.1	The Mean Value Theorem	320
7.2	Applications of the Mean Value Theorem	325
7.2.1	Antiderivatives	325
7.2.2	Increasing Function Theorem	330
7.2.3	Functions with Bounded Derivatives	332
7.2.4	Comparing Functions Using Their Derivatives	334
7.2.5	Interpreting the Second Derivative	338
7.2.6	Formal Definition of Concavity	339
7.2.7	Classifying Critical Points: The First and Second Derivative Tests	343
7.2.8	Finding Maxima and Minima on $[a, b]$	347
7.3	L'Hôpital's Rule	350
7.4	Cauchy's Mean Value Theorem	359
7.4.1	Geometric Interpretation of the Cauchy Mean Value Theorem	360
7.5	The Proof of L'Hôpital's Rule	361
7.6	Curve Sketching: Part 2	364
8	Taylor Polynomials and Taylor's Theorem	376
8.1	Introduction to Taylor Polynomials and Approximation	376
8.2	Taylor's Theorem and Errors in Approximations	388
8.3	Big-O	397
8.3.1	Calculating Taylor Polynomials	403

Chapter 1

A Short Introduction to Mathematical Logic and Proof

In this course, we are going to take a rigorous approach to the main concepts in single variable Differential Calculus. This means that rather than simply asserting mathematical statements as facts, we will attempt whenever possible, to provide proofs of the validity of these statements. To do so, we will begin with a set of notions and statements that we will take as being given. For example, we will assume the basic notions of set theory and the algebraic and arithmetic properties of the natural numbers, the integers, the rational numbers and the real numbers. We will introduce as *axioms* some of the perhaps less well-known properties of these objects such as the Principle of Mathematical Induction for the natural numbers and the Least Upper Bound Property for the real numbers. It is important to note that while there is some value in rigour for its own sake and that it is even possible for proofs to be fun, our motivation in this course for including rigour is the hope that we will gain a deeper understanding of the fundamental concepts of Calculus as well as an appreciation for their limitations. In this respect, we will begin with a very brief, and admittedly incomplete, introduction to the formalities of mathematical logic and to the rules of inference that we will use in constructing our proofs.

1.1 Basic Notions of Mathematical Logic and Truth Tables

DEFINITION

Statement

A *statement* is a (mathematical) sentence that can be determined to be either true or false.

For example, “all differentiable functions are continuous”, and “all prime numbers are odd” are two examples of mathematical statements. The first statement will be later shown to be true and, since 2 is a natural number that is both prime and even, the latter statement is false. Sometimes we are not able to determine whether or not a statement is true or false but we can see that it must be one or the other. For example,

the Twin Prime Conjecture says that there are infinitely primes p such that $p + 2$ is also prime. Despite a great deal of effort that has been exerted to try and prove this statement, we still do not know that it is true. (A *conjecture* is a statement for which there is evidence or strong speculation that it is true but no known proof). However, it should be obvious that this statement is either true or it is false. A mathematical sentence such as $x > 0$ is not a statement since it can be either true or false depending on the value assigned to the variable x .

Throughout the rest of this chapter we will use italicized lower case letters to denote statements.

Given a statement p , we can also talk about the *negation* of p which we denote by $\neg p$ and which we call *not p*. The negation of a statement is exactly what one would expect from the name. For example, if the statement p is “the sky is blue”, then the negation $\neg p$ is simply the statement that “the sky is not blue”. When a statement p is true, its negation $\neg p$ is false and vice versa. We illustrate this fact by the use of a truth table.

p	$\neg p$
T	F
F	T

Notice that it is never the case that both p and $\neg p$ are simultaneously both true nor is it ever the case that both p and $\neg p$ are simultaneously false. This is known as *the Law of the Excluded Middle*.

In this course, we will be asked to prove statements that rely on hypotheses. For example, If $f(x)$ is differentiable, then $f(x)$ is continuous. If we let the statement p be “ $f(x)$ is differentiable”, and the statement q be “ $f(x)$ is continuous”, then we are asserting that the truth of p implies the truth of q . That is, p implies q . We will denote this by

$$p \Rightarrow q.$$

In p implies q , the statement p is called the *antecedent* and q is called the *consequence*.

We can construct a truth table for $p \Rightarrow q$. To do so, we ask ourselves how such a statement could be false. We would conclude that the only way for this to happen would be if p is true but q is false. This leads to the following truth table.

p	q	$p \Rightarrow q$
T	T	T
T	F	F
F	T	T
F	F	T

A close look at the truth table above yields a rather strange consequence, namely that **something false will imply anything**. For example if p is the statement “all animals

are dogs" and q is the statement that "the sky is blue", then the fact that p is false means that we can conclude that $p \Rightarrow q$ or that

All animals are dogs **implies** that the sky is blue.

The words "if ... then" are referred to as a logical *connective* as they join two statements together to form a compound statement. Two other common connectives are *and* and *or*. (Curiously, *not* is also a connective though it is only applied to a single statement.) Given two statements p and q , we can form two new statements p *and* q , and p *or* q , which we denote respectively by $p \wedge q$ and $p \vee q$.

It should be quite clear that for p *and* q to be true, it must be the case that both statements are true. This means that the truth table looks like:

p	q	$p \wedge q$
T	T	T
T	F	F
F	T	F
F	F	F

The situation for the word *or* is a little more ambiguous. In fact, in common speech *or* has two possible interpretations. We could say that p *or* q is satisfied if at least one of p and q is true. We could also say that p *or* q is satisfied if one of p and q is true, but not both. The first case is referred to as the *inclusive or*. This is the interpretation of the connective *or* that we use in mathematical logic. As we indicated above, it is denoted by $p \vee q$ and its truth table is:

p	q	$p \vee q$
T	T	T
T	F	T
F	T	T
F	F	F

The second case is called the *exclusive or*. It is logically equivalent to

$$(p \vee q) \wedge \neg(p \wedge q).$$

We can use truth tables to see how this works

p	q	$p \vee q$	$p \wedge q$	$\neg(p \wedge q)$	$(p \vee q) \wedge \neg(p \wedge q)$
T	T	T	T	F	F
T	F	T	F	T	T
F	T	T	F	T	T
F	F	F	F	T	F

We have suggested that the *exclusive or* is logically equivalent to the compound statement $(p \vee q) \wedge \neg(p \wedge q)$. Generally, we will say that statements p and q are *equivalent* if the truth of one implies the truth of the other. In other words, we say that p holds if and only if q holds. In terms of our connectives, we can interpret equivalence as the statement $(p \Rightarrow q) \wedge (q \Rightarrow p)$. The truth table is

p	q	$p \Rightarrow q$	$q \Rightarrow p$	$(p \Rightarrow q) \wedge (q \Rightarrow p)$
T	T	T	T	T
T	F	F	T	F
F	T	T	F	F
F	F	T	T	T

The truth table confirms that equivalence happens when p and q have the same truth values. We will denote equivalence by $p \Leftrightarrow q$.

Contrapositive:

In this example, we will use truth tables to show that the statements $p \Rightarrow q$ and $\neg q \Rightarrow \neg p$ are logically equivalent. The statement $\neg q \Rightarrow \neg p$ is called the *contrapositive* of the statement $p \Rightarrow q$. We will see later that the equivalence of an implication with its contrapositive leads to a method of proof called *proof by contradiction*.

We can consider the following truth table:

p	q	$\neg p$	$\neg q$	$p \Rightarrow q$	$\neg q \Rightarrow \neg p$	$(p \Rightarrow q) \Leftrightarrow (\neg q \Rightarrow \neg p)$
T	T	F	F	T	T	T
T	F	F	T	F	F	T
F	T	T	F	T	T	T
F	F	T	T	T	T	T

A quick look at the table shows that $p \Rightarrow q$ and $\neg q \Rightarrow \neg p$ have the same truth values. This is what we wanted for the statements to be equivalent. Alternatively, we see that $(p \Rightarrow q) \Leftrightarrow (\neg q \Rightarrow \neg p)$ is always true regardless what truth values p and q are assigned. A compound statement is called a *tautology* if it is always true regardless of the truth values assigned to the basic statements.

For example, given a function $f(x)$ defined on the real numbers, if p is the statement that “ $f(x)$ is differentiable” and q is the statement that “ $f(x)$ is continuous”, then $p \Rightarrow q$ is the statement that differentiability implies continuity. The contrapositive, $\neg q \Rightarrow \neg p$ represents the statement that if $f(x)$ is not continuous, it cannot be differentiable”. We will see later in this course that the first statement is true, and hence the second statement must also be true.

1.2 Variables and Quantifiers

We saw before that there are mathematical sentences which could be either true or false depending upon additional parameters. For example the statement “ $x > 0$ ” may be true or it may be false as the value of x is allowed to vary. For this reason, we call x a variable. The “truth value” of “ $x > 0$ ” will be determined once we assign a value to x . This reminds us of a function and as such we will use the functional notation

$$p(x) : x > 0$$

to represent the sentence $x > 0$. The potential values for the variable x will either be specified or they will be determined by the context of the sentence. For example, in

this course, the sentence $x > 0$ makes sense whenever x is assigned a value that is a real number. In this case, $p(4)$ is true but $p(-3)$ is false.

In this course, we will often want to show either that a sentence $p(x)$ is true for all possible values of x , that it is true for some values of x , or that it is false for every value of x . In the first case, we would say that **for every** x , $p(x)$ is true. The phrase *for every* is called the *universal quantifier*. It is denoted by the symbol \forall and we can write the above sentence symbolically as follows:

$$\forall x : p(x).$$

Here it is important to note that the scope of the variable x must be known for this to make sense. That is we must know the collection of all possible values of x that we are considering. For example, we may want x to be any real number, or any polynomial, or any dog. If the scope is known, then this sentence itself becomes a statement as it is either true or false.

To show that

$$\forall x : p(x)$$

is **true** we must have some way of confirming the statement $p(x)$ for all possible values of x . For example, the statement:

“For all natural numbers $n \geq 2$, n factors as a product of primes.”

requires us to show that for each such n , n can indeed be factored as a product of primes.

To show that the statement

$$\forall x : p(x)$$

is false we need only find one example of a value for x for which $p(x)$ is false. This example is often called a *counterexample* to the statement $p(x)$.

For example, the statement

“For all natural numbers n , n factors as a product of primes.”

can be shown to be false, because $n = 1$ is a natural number that cannot be factored as a product of primes.

Other English phrases that express the universal quantifier are “for each...” and “for all...”.

We say that

there exists an x such that $p(x)$ is true,

if we can find at least one value of x_0 such that when substituted into the sentence, $p(x_0)$ becomes true. The phrase *there exists* is called the *existential quantifier*. Symbolically, it is denoted by \exists . We also use the symbol \exists to represent the phrase *such*

that. Therefore, we can express the sentence “there exists an x such that $p(x)$ is true symbolically by

$$\exists x \ni p(x).$$

In this case, to prove that the statement

$$\exists x \ni p(x)$$

is true we need only find one example of a value for x that makes $p(x)$ true. For example, to prove the statement

“There exists a natural numbers n , such that n is divisible by both 2 and by 3.”

we can do so by presenting 6 as an example of a natural number that is divisible by both 2 and 3.

To show that the statement

$$\exists x \ni p(x)$$

is false, requires us to have some way to show that for every x , the statement $p(x)$ is false.

We will often require the use of more than one quantifier in a sentence. In this case, the order of the quantifiers is very important. For example the sentences “for every x there exists a z such that $x \leq z$ ” and the sentence

“there exists an x such that for every z , $x \leq z$ ” are very different. The first statement is clearly true for the set of all real numbers since once we have determined a value of x we may simply choose z to be x as well. The second statement is false for the real numbers since its truth would imply that the real numbers have a least element which is clearly not true.

It is important to know how to negate sentences with quantifiers. When, for example, would it not be true that “for every x the sentence $p(x)$ is true”. This happens precisely when “there exists at least one such x_0 such that **not** $p(x_0)$ is true”. Symbolically this means that

$$\neg(\forall x : p(x)) \Leftrightarrow \exists x \ni \neg p(x).$$

The value x_0 that negates $\forall x : p(x)$ is called a *counterexample* for the statement $\forall x : p(x)$. We emphasize that to prove the that the statement “ $\forall x : p(x)$ ” is true, we need some procedure or argument to exhaust all choices of x , whereas to show that “ $\forall x : p(x)$ ” is false, **we need only find one counterexample**.

Similarly, the statement “there exists an x such that $p(x)$ is true” is itself false precisely when for every x , the statement $p(x)$ is false. That is,

$$\neg(\exists x \ni p(x)) \Leftrightarrow \forall x : \neg p(x).$$

1.3 Rules of Inference and the Foundations of Proof

In some ways, constructing proofs is like a complex game. Just as in games, in constructing proofs there are certain basic rules that can be applied. These rules are called the *Rules of Inference*.

Modus Ponens:

Our first rule of inference is called *Modus Ponens*. Simply stated, this rule tells us that if we assume that p is true and we know that p implies q , then we can conclude that q is also true. Symbolically, this is expressed as

$$p \wedge (p \Rightarrow q) \Rightarrow q$$

We can validate Modus Ponens by looking at the truth table below

p	q	$p \Rightarrow q$
T	T	T
T	F	F
F	T	T
F	F	T

We see that the only situation in which both p and $p \Rightarrow q$ are true occurs when q is also true.

For example, we know that Fido is a dog and that if Fido is a dog, then Fido is an animal. From this we can conclude that Fido is an animal.

Modus Tollens:

A variant on Modus Ponens is the rule *Modus Tollens*. This rule tells us that if we know that $p \Rightarrow q$ is true and we also know that q is false, then we could conclude that p must also be false. This is represented symbolically by

$$[(p \Rightarrow q) \wedge \neg q] \Rightarrow \neg p$$

Again, this can be verified from the truth table for $p \Rightarrow q$. An example of Modus Tollens would be: If Fido is a dog, then Fido is an animal. However, Fido is not an animal and hence we conclude that Fido is not a dog.

Hypothetical Syllogism:

The next rule of inference that we will introduce is *Hypothetical Syllogism*. This is like an associativity law for implication. It says that if p implies q and if q implies s , then we can conclude that p implies s . Symbolically, this is

$$(p \Rightarrow q) \wedge (q \Rightarrow s) \Rightarrow (p \Rightarrow s)$$

We can see in the truth table below that when ever both $p \Rightarrow q$ and $q \Rightarrow s$ are true then so is $p \Rightarrow s$.

p	q	s	$p \Rightarrow q$	$q \Rightarrow s$	$p \Rightarrow s$
T	T	T	T	T	T
T	T	F	T	F	F
T	F	T	F	T	T
T	F	F	F	T	F
F	T	T	T	T	T
F	T	F	T	F	T
F	F	T	T	T	T
F	F	F	T	T	T

An example of Hypothetical Syllogism is: If Fido is a dog, then Fido is an animal and if Fido is an animal, Fido must eat, therefore, if Fido is a dog, Fido must eat.

Disjunctive Syllogism:

The next rule is *Disjunctive Syllogism*. Simply stated this rule says that if we know that p or q are true, and we can show that p is false, then we can conclude that q is true. This becomes

$$[(p \vee q) \wedge \neg p] \Rightarrow q.$$

An example of Disjunctive Syllogism is: We are told that Fido is either a cat or a dog. We know Fido is not a cat. Therefore, Fido is a dog.

Additional Rules are:

Constructive Dilemma:

This rule states that if $p \Rightarrow q$ and $r \Rightarrow s$, and if we know that either p or r is true, then we can conclude that either q or s must be true. That is

$$[(p \Rightarrow q) \wedge (r \Rightarrow s)] \wedge (p \vee r) \Rightarrow (q \vee s).$$

Destructive Dilemma:

This rule states that if $p \Rightarrow q$ and $r \Rightarrow s$, and if we know that either q or s is false, then we can conclude that either p or r must also be false. That is

$$[(p \Rightarrow q) \wedge (r \Rightarrow s)] \wedge (\neg q \vee \neg s) \Rightarrow (\neg p \vee \neg r).$$

The last three rules are very straight forward:

Simplification:

This rule says that if we know that both p and q are true, we can conclude that p is true. That is

$$(p \wedge q) \Rightarrow p$$

Addition:

This rule states that if p is true, then either p or q must be true for any q .

$$p \Rightarrow (p \vee q)$$

Conjunction:

This states that if we can establish the truth of p and the truth of q separately, then we have established the truth of the statement “ p and q ”.

We have seen that if we know that all dogs are animals, and we know that Fido is a dog, then Modus Ponens allows us to conclude that Fido is an animal. In this case, we are able to apply our general knowledge about dogs to conclude something about a specific dog Fido. This is an example of what is known as *deductive reasoning*. Most of the proofs in mathematics employ deductive reasoning. Generally, we will start with a hypothesis or something we know to be true, and then apply rules such as those above to reach our desired conclusion.

Inductive Reasoning:

There is another type of reasoning called *inductive reasoning*. In inductive reasoning, we begin with some specific observations and then try to draw a more general conclusion. For example, if we knew that the first few terms of an infinite sequence were $\{2, 4, 6, 8, 10, \dots\}$, then we might guess from the pattern of these five terms that this was the sequence of all even natural numbers. From this we could speculate that the next term in the sequence would be 12. Unlike, most instances of deductive reasoning that we will see in this course, inductive reasoning most often does **not** result in a proof. Indeed, it is possible that if we were to be told a few more terms in the sequence above we might find that we have $\{2, 4, 6, 8, 10, 0, 2, 4, 6, 8, 10, \dots\}$ where the general formula for the n -th term is $a_n = 2n \pmod{12}$. We see that our inductive conclusion was wrong. This does not make inductive reasoning useless. In fact, inductive reasoning is the foundation for much of science, particularly experimental science. Even in mathematics, inductive reasoning often leads us to an understanding of what is actually going on. It helps us to formulate conjectures, mathematical statements that we believe to be true, and for which we might later find proofs. It is also a key element in problem solving. Moreover, in the next chapter we will see how to employ an important formal technique of proofs that is based on inductive reasoning called *Proof by Induction*.

In closing out this chapter, we will briefly remark on a somewhat indirect technique of proof called *proof by contradiction*. In this technique we will use the fact that if we know a statement q to be true and that if we can show that $\neg p \Rightarrow \neg q$, then we could conclude that p must also be true. Formally, this follows from our rules of inference and from our understanding of the contrapositive. To see why this is the case, we observe that knowing that $\neg p \Rightarrow \neg q$ to be true also tells us that the contrapositive statement $q \Rightarrow p$ is true. However, since we have as a hypothesis that q is true, Modus Ponens tells us that p must also be true.

EXAMPLE 1 Prove that there are infinitely many prime numbers.

PROOF

We will not provide all of the details of the proof of this statement at this time. In particular, to prove this statement, we will need to know that every natural number greater than or equal to 2 has a prime factor, something that we will leave as an exercise following the next chapter. We then begin by assuming that there are not infinitely many prime numbers or equivalently that the list

$$p_1, p_2, \dots, p_n$$

of prime numbers is finite. Next we let

$$p = p_1 p_2, \dots, p_n + 1$$

Since p is larger than 1, it must have a prime factor. However, it is easy to see that none of the listed primes p_1, p_2, \dots, p_n could be a factor of p . This contradicts the assumption that the list p_1, p_2, \dots, p_n includes all primes and shows that our original assumption that there are not infinitely many prime numbers must be false. Hence we conclude that there are infinitely many prime numbers. ■

Finally, we note that the material of this chapter has been included to encourage the reader to think about what it means to formulate a proof of a mathematical statement. We will use the ideas introduced here in this course but we will in general not make full use the formal symbolism.

Chapter 2

Sets, Relations and Functions

In this chapter, we will introduce some basic material that will be used throughout the rest of the course.

2.0 Notation

We will use the following notation:

- \mathbb{N} will denote the set of natural numbers $\{1, 2, 3, \dots\}$.
- \mathbb{Z} will denote the set of integers $\{\dots, -2, -1, 0, 1, 2, \dots\}$.
- \mathbb{Q} will denote the set of rational numbers $\left\{\frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{N}\right\}$.
- \mathbb{R} will denote the set of real numbers.

Intervals. We will use the notation (a, b) to denote the set $\{x : a < x < b\}$. This is called an *open interval*. We will use $[a, b]$ to denote the set $\{x : a \leq x \leq b\}$. This is called a *closed interval*. Additionally, we will use the notation $(-\infty, b)$, (a, ∞) , $(-\infty, b]$, $[a, \infty)$ to mean the open intervals $\{x : x < b\}$, $\{x : x > a\}$, and the closed intervals $\{x : x \leq b\}$, $\{x : x \geq a\}$, respectively. Finally, we will use $[a, b)$ and $(a, b]$ to denote the half-open intervals $\{x : a \leq x < b\}$ and $\{x : a < x \leq b\}$, respectively.

Formally, we make the following definition.

DEFINITION

Intervals

A set $S \subset \mathbb{R}$ is an *interval* if for every $x, y \in S$, if $x \leq z \leq y$ then we must have $z \in S$.

It is easy to see that the *singleton set* $\{a\}$ is an interval for any $a \in \mathbb{R}$. It is somewhat less obvious that the *empty set*, denoted by \emptyset , is also an interval. To see why this is so we first ask what would it mean if the empty set was “not an interval”. In this case, we would have to be able to find a pair $x, y \in \emptyset$ and an element $z \in \mathbb{R}$ such that

$x \leq z \leq y$ but $z \notin \emptyset$. This is clearly impossible because no such x, y exist in \emptyset . As such, we have shown that the statement, “ \emptyset is not an interval” is false, and as such we have proved that \emptyset is an interval.

DEFINITION**Degenerate Intervals**

An interval I is said to be *degenerate* if $I = \{c\}$ for some $c \in \mathbb{R}$ or if $I = \emptyset$. Otherwise, we say that it is *non-degenerate*.

2.1 Sets, Products and Relations

In this section we will introduce some of the basic notation from set theory that we will use through out the rest of the course notes.

Subsets and Complements:

We will use the notation

$$A \subset B \quad \text{and} \quad A \subseteq B$$

interchangeably to mean that A is a subset of B with the possibility that $A = B$. That is, every element of A is also contained in B .

When we explicitly wish to emphasize that $A = B$ is a possibility, we will generally use $A \subseteq B$. When we wish to express that A is a *proper subset* of B , then we can either specify further that $A \neq B$, or we can use the notation

$$A \subsetneq B.$$

REMARK

The empty set, which we will denote by \emptyset , is a set with no elements.

The empty set may seem at first glance to be quite unremarkable. However, this is definitely not the case. For example, it has the unique property that it is a subset of every set. In fact, as we shall see later the empty set is really quite a mysterious object. ◀

DEFINITION**Power Set**

Given a set X we define the *power set of X* to be the set

$$\mathbb{P}(X) = \{A \mid A \subseteq X\}.$$

That is $\mathbb{P}(X)$ consists of all subsets of X including \emptyset and X itself.

Assume that A and B are subsets of some universal set X . We will let

$$B \setminus A = \{x \in B \mid x \notin A\}.$$

The set $B \setminus A$ is called the *set difference* of B minus A .

In the special case when $B = X$, we call the set $X \setminus A$ the *complement* of A in X and denote this set by

$$A^c.$$

Unions and Intersections:

Two of the most fundamental operations on sets are *union* and *intersection*. We define these as follows:

DEFINITION

Union

Let $A, B \subseteq X$. The *union* of A and B is the set

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}.$$

More generally, if for each $\alpha \in I$ we have $A_\alpha \subseteq X$, then

$$\bigcup_{\alpha \in I} A_\alpha = \{x \mid x \in A_\alpha \text{ for some } \alpha \in I\}.$$

DEFINITION

Intersection

Let $A, B \subseteq X$. The *intersection* of A and B is the set

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

More generally, if for each $\alpha \in I$ we have $A_\alpha \subseteq X$, then

$$\bigcap_{\alpha \in I} A_\alpha = \{x \mid x \in A_\alpha \text{ for all } \alpha \in I\}.$$

The operations of union, intersection and complementation are linked by the following important theorem:

THEOREM 1

De Morgan's Laws

$$1) (\bigcup_{\alpha \in I} A_\alpha)^c = \bigcap_{\alpha \in I} A_\alpha^c$$

$$2) (\bigcap_{\alpha \in I} A_\alpha)^c = \bigcup_{\alpha \in I} A_\alpha^c$$

The proof of first of De Morgan's Laws will be left as an exercise. The second law follows immediately from the first by replacing A_α with A_α^c and observing that for any subset A of X , we have that

$$(A^c)^c = A.$$

2.2 Products of Sets

In this section we will introduce *products* of sets. We will see that defining the product of an arbitrary collection of sets is a rather complex process. However, we will begin with the simplest case involving two sets. To do so we note that if X, Y are two sets, then we call (x, y) where $x \in X$ and $y \in Y$ an *ordered pair* in X and Y .

DEFINITION

Product of Two Sets

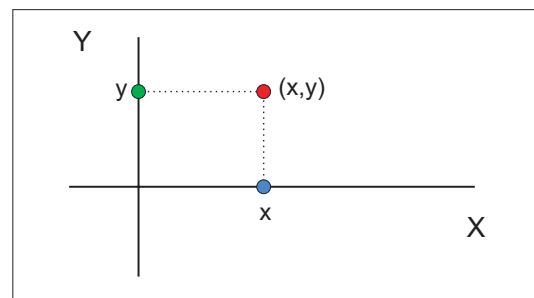
Given two sets X, Y , define the *product of X and Y* by

$$X \times Y = \{(x, y) \mid x \in X \text{ and } y \in Y\}.$$

x is called the *x-coordinate* of (x, y) .

y is called the *y-coordinate* of (x, y) .

We can visualize the product of two functions by plotting the points on a pair of axes:



Given finitely many sets X_1, X_2, \dots, X_n we can easily generalize the notion of a product of these n sets by replacing ordered pairs with ordered n -tuples (x_1, x_2, \dots, x_n) where each $x_i \in X_i$. This leads us to the following definition.

DEFINITION**Product of n Sets**

Given n sets $\{X_1, X_2, X_3, \dots, X_n\}$, define the *product* of $\{X_1, X_2, X_3, \dots, X_n\}$ by

$$X_1 \times X_2 \times \dots \times X_n = \prod_{i=1}^n X_i = \{(x_1, x_2, x_3, \dots, x_n) \mid x_i \in X_i\}.$$

$(x_1, x_2, \dots, x_n) \in \prod_{i=1}^n X_i$ is called an *n -tuple* and x_i is called the *i th coordinate*.

If $X_i = X$ for all i , we write X^n for $\prod_{i=1}^n X_i$.

REMARK

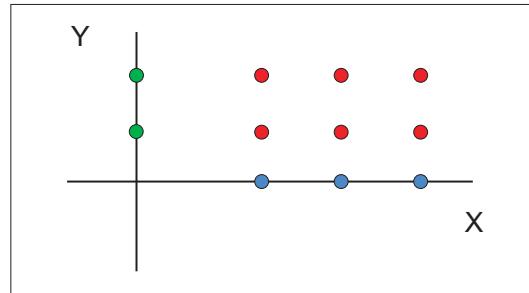
Given a finite set X , we say that X has *cardinality n* , if X contains n elements. In this case, we write

$$|X| = n.$$

Assume that we have a set X with three elements and a set Y with two elements. Then it is easy to see that the product $X \times Y$ has

$$3 \times 2 = 6$$

elements.



After a little thought, it would seem reasonable to speculate that if $\{X_1, X_2, \dots, X_n\}$ was a finite collection of finite sets, then the cardinality of the product of these n sets would be the product of their individual cardinalities. In fact, this is the case as the next theorem shows. ◀

THEOREM 2**Cardinality of Products**

Let $\{X_1, X_2, \dots, X_n\}$ be a finite collection of finite sets. Then

$$|\prod_{i=1}^n X_i| = \prod_{i=1}^n |X_i| = |X_1| \cdot |X_2| \cdots |X_n|.$$

In particular, if $X_i = X$ for all i , then

$$|\prod_{i=1}^n X_i| = |X|^n.$$

So far we have seen how to define the product of a collection of n sets as the set of all ordered n -tuples (x_1, x_2, \dots, x_n) with $x_i \in X_i$ for each $i = 1, 2, \dots, n$. What happens if our collection of sets is not finite? Specifically, we ask the following question:

Fundamental Question: Suppose for example that we had some infinite set I and for each $\alpha \in I$ we have a set X_α . How do we define the product of sets in this collection $\{X_\alpha\}_{\alpha \in I}$? ◀

To see how we could do this we will first take another look at our products of a finite collection of sets.

REMARK

To answer the previous question we will take a closer look at how we obtain our product for finite collections.

We begin by noting that each $(x_1, x_2, \dots, x_n) \in \prod_{i=1}^n X_i$ determines a function

$$f_{(x_1, x_2, \dots, x_n)} : \{1, 2, \dots, n\} \rightarrow \bigcup_{i=1}^n X_i$$

by

$$f_{(x_1, x_2, \dots, x_n)}(i) = x_i.$$

For here we make three key observations.

Key Observation 1: Each such function satisfies:

$$f_{(x_1, x_2, \dots, x_n)}(i) \in X_i.$$

Key Observation 2: Given $f : \{1, 2, \dots, n\} \rightarrow \bigcup_{i=1}^n X_i$ with $f(i) \in X_i$, we can define

$$(x_1, x_2, \dots, x_n) \in \prod_{i=1}^n X_i,$$

by setting

$$x_i = f(i).$$

Key Observation 3: The identification

$$(x_1, x_2, \dots, x_n) \iff f_{(x_1, x_2, \dots, x_n)}$$

establishes a one-to-one correspondence between the product $\prod_{i=1}^n X_i$ and the set

$$\{f : \{1, 2, \dots, n\} \rightarrow \bigcup_{i=1}^n X_i \mid f(i) \in X_i\}.$$

In this case, we write

$$\prod_{i=1}^n X_i \cong \{f : \{1, 2, \dots, n\} \rightarrow \bigcup_{i=1}^n X_i \mid f(i) \in X_i\}.$$

Since f allows us to choose one element from each of our sets, we call f a *choice function*. ◀

We have just seen that the product of finitely many sets can be identified with a collection of special functions called *choice functions*. Since it is possible to define a choice function for an arbitrary collection of sets this is the concept that will allow us to extend our definition of a product to even infinitely many sets.

DEFINITION

Product of Sets: The General Case

Given a collection $\{X_\alpha\}_{\alpha \in I}$ of sets, define

$$\prod_{\alpha \in I} X_\alpha = \{f : I \rightarrow \bigcup_{\alpha \in I} X_\alpha \mid f(\alpha) \in X_\alpha\}.$$

If $X_\alpha = X$ for all $\alpha \in I$, $\prod_{\alpha \in I} X_\alpha$ is written as

$$X^I.$$

A function $f \in \prod_{\alpha \in I} X_\alpha$ is called a *choice function* on $\{X_\alpha\}_{\alpha \in I}$.

REMARK

We have seen that if we have a finite collection of finite non-empty sets, then the cardinality of the product of these sets is the product of the cardinality of the individual sets. In particular, the product is certainly non-empty. In fact, this is easily seen to be true, that the product is non-empty, even if the individual sets are not finite. Given $\{X_1, X_2, \dots, X_n\}$ we simply chose one element x_1 from X_1 , the x_2 from X_2 and so on until finally, we choose x_n from X_n leaving us with the n -tuple (x_1, x_2, \dots, x_n) .

While it might be reasonable to assume that all products of non-empty sets must at least contain one point, if the collection of sets $\{X_\alpha\}_{\alpha \in I}$ is itself infinite, then things get more complicated. Can we actually choose one element simultaneously from each X_α if our collection is infinite? In other words does a choice function even exist for such an arbitrary collection?

It turns out that the answer to the above question is extremely profound. In 1938, the logician Kurt Gödel showed that using the standard axioms of set theory that you cannot prove the negation of the claim that all such products are non-empty. Then in 1963, Paul Cohen showed that using the standard axioms of set theory you cannot prove that the claim is true either. These two remarkable results tell us that we literally have a choice to accept the claim or not. This leads us to add the following axiom to our rules of set theory: 

AXIOM 3**Axiom of Choice**

If $\{X_\alpha\}_{\alpha \in I}$ is a non-empty collection of non-empty sets, then is

$$\prod_{\alpha \in I} X_\alpha \neq \emptyset.$$

REMARK

There is an equivalent formulation of the axiom of choice that is perhaps more useful. This version of the Axiom of Choice provides us with a rule that will allow us to simultaneously choose one element from each non-empty subset of X . We will see later that if $X = \mathbb{N}$, then there is an easy way to describe an explicit rule to allow us to do just that. Given a non-empty set $A \subseteq \mathbb{N}$, we let $f(A)$ be the least element in A , which we will soon show always exists. However, if $X = \mathbb{R}$, then this rule will not work since $A = (0, 1)$ has no least element. In fact, it is the case that no such explicit function can ever be constructed for \mathbb{R} . 

AXIOM 4**Axiom of Choice: Version 2**

Given any non-empty set X , there exists a function $f : \mathbb{P}(X) \setminus \{\emptyset\} \rightarrow X$ such that

$$f(A) \in A$$

for every $A \in \mathbb{P}(X) \setminus \{\emptyset\}$

Why should we choose to accept choice? It turns out that we will use the axiom, or some weaker form of it many times through out this course . Most often we will not even be aware of doing so. Moreover, in the majority of cases where we use the axiom, without it it would not be at all clear how we could proceed. So while the Axiom of Choice may be a rather abstract concept, it is none the less an important tool which will help us to establish many of the most fundamental from Calculus. As an exercise, see if you can spot the next time the axiom is used.

2.3 Relations and Functions

We have already had several instances where we encountered functions. But what is a function? We can give an informal heuristic or working definition of a function that should be quite familiar to you.

DEFINITION**Function: Heuristic Definition**

A function is a rule that assigns to each element x in a set X a unique value y in a set Y .

- Denote the function by a lowercase letter such as f and we write $y = f(x)$.
- We also use the notation $f : X \rightarrow Y$ to denote the function.

While this definition certainly works for most purposes, it does lack mathematical precision. In this section we will explore ways to make the definition of a function more precise. To do so we first need to look at the concept of a *relation*

2.3.1 Relations

Like functions, there is a heuristic meaning to the term *relation*. Informally, in mathematics, a *relation* is a collection of related numbers or objects.

The relations that you are probably most familiar with often arise as solutions of an *equation*, although as we will see you can also specify a relation by using *ordered pairs*, a *graphical representation*, or *descriptive sentences*.

EXAMPLE 1 Specifying a Relation

1. Equation

- $y = x + 1$

In this case, the assumption is that both x and y are in \mathbb{R} (the Real numbers) and that they are related if they satisfy the given equation. In this case, we could represent the relationship between x and y by specifying the solution set S to the equation.

- $S = \{(x, y) \mid y = x + 1; x, y \in \mathbb{R}\}$

We could modify the relation by restriction the set of point x and y that we would accept as solutions. For example, if we wish to have that both x and y are integers we get

- $K = \{(x, y) \mid y = x + 1; x, y \in \mathbb{Z}\}$

The relation statement for K is similar to the statement for relation S in that in both cases we are asking that $y = x + 1$. . However, the pairs that we identify as being related are different so they are different relations even though they come from the same mathematical expression.

2. Lists of Ordered Pairs

You can describe a relation by listing all of its ordered pairs. For example, ordered pairs that belong to relation S above include

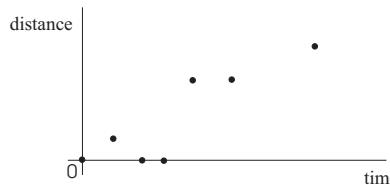
$$(-2, -1), (-1, 0), (0, 1), (1, 2), (2, 3)$$

Of course, this is not the complete list of ordered pairs that defines relation S . In fact, it is not possible to list all of the ordered pairs that define S and this is a drawback of using lists of ordered pairs to define a relation.

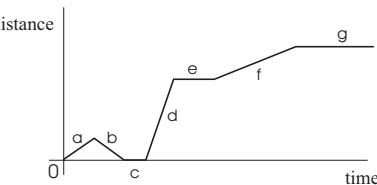
However, ordered pairs are quite helpful if you are trying to determine a relation from experimental data, or on any finite set.

3. Graphical Representation

A relation can be represented by graphing data on a set of axes.



Discrete graph of ordered pairs



Continuous graph of distance traveled

The first graph contains a set of ordered pairs that represents where a traveller stopped at a finite list of times. When ordered pairs are plotted on a set of axes, the plot is called a **discrete** graph.

In the second graph we plot the distance travelled at each moment between the discrete set of points. This gives us a **continuous** graph since it contains no breaks.

What is common in all of these representations of relations is that they involve pairs of elements that are designated to be related to one another. This suggests that ordered pairs play a key role here. In fact, this observation leads us to a natural formal definition for a relation.

DEFINITION

Relation

A *relation on X and Y* is a set $R \subseteq X \times Y$.

We say x is *R-related to y* if $(x, y) \in R$ and write xRy .

The *domain* of the relation R is the set

$$\{x \in X \mid \text{there exists a } y \in Y \text{ such that } (x, y) \in R\}$$

and is denoted by $\text{dom}(R)$.

Y is called the *codomain* of the relation R and is denoted by $\text{codom}(R)$.

The *range* of the relation R is the set

$$\{y \in Y \mid \text{there exists an } x \in X \text{ such that } (x, y) \in R\}$$

and is denoted by $\text{ran}(R)$.

NOTE

Generally we will have $X = Y = \mathbb{R}$ and we say that R is a relation on \mathbb{R} .

To further illustrate these definitions, consider the following examples.

EXAMPLE 2

(a) Determine the domain and range of the finite relation J , where

$$J = \{(3, 2), (3, 4), (-2, 2), (1, 4)\}$$

Answer:

Domain of J : $\{3, -2, 1\}$

Range of J : $\{2, 4\}$

(b) Determine the domain and range of the range of relation K , where

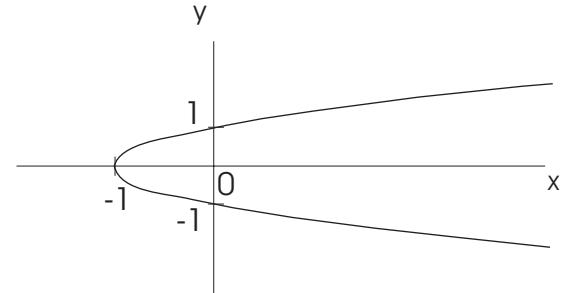
$$K = \{(x, y) \mid y^2 = x + 1, x \in \mathbb{R}, y \in \mathbb{R}\}$$

Answer:

It is always true that for each real number y , $y^2 \geq 0$. Since $y^2 = x + 1$ and $y^2 \geq 0$, it follows that $x + 1 \geq 0$. This simplifies to $x \geq -1$. Thus, the domain of $K = \{x \mid x \geq -1\}$.

Now for each Real number $x \geq -1$, we have $y^2 = x + 1$ is equivalent to

$$y = \sqrt{x + 1} \quad \text{and} \quad y = -\sqrt{x + 1}$$



This means that y can take on any positive real number, negative real number, or 0. Hence, the range of $K = \{\text{all real numbers}\}$, or more simply the range of $K = \{y \mid y \in \mathbb{R}\}$.

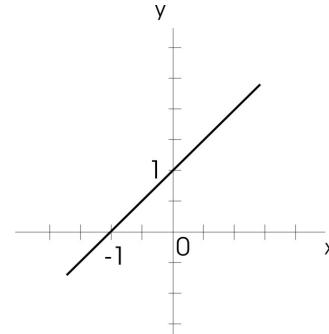
(c) The relation is specified by the graph shown. State the relation using set notation and then determine its domain and range.

Answer:

Since the graph is a straight line with slope 1 and y -intercept 1, the relation is

$$\{(x, y) \mid y = x + 1\}$$

Both the domain and the range are \mathbb{R} .



Now that you are familiar with the idea of a “relation”, we next consider functions.

2.3.2 Functions

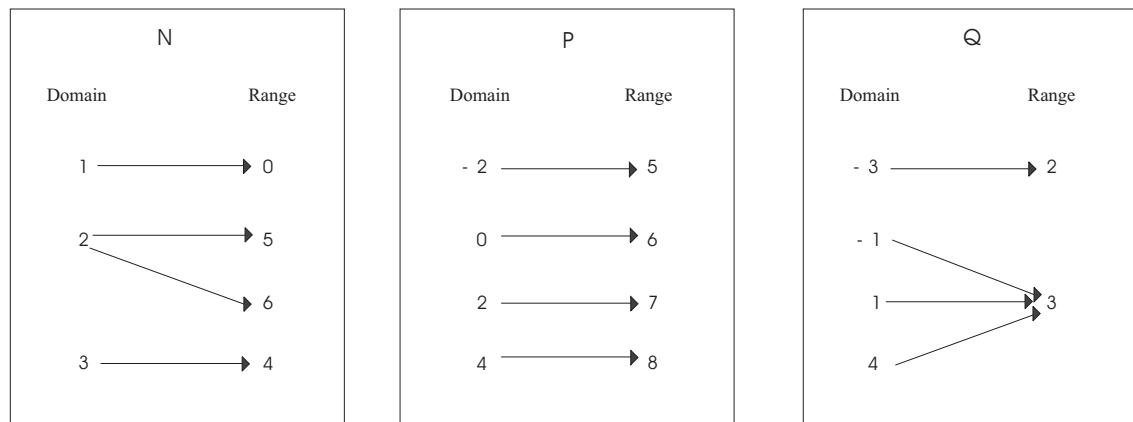
Formally, a *function* is a special type of relation. The key defining characteristic for a function that each input is *not* permitted to return more than one output.

For example, each student in a class is assigned an ID number. This assignment represents a function defined on the set of students in the class to integer values. In this case the nature of the rule that assigns students to integers may not be understood, but the class list gives us all the information we require to determine the value of the function for any member of the class, and each ID number is **unique**.

Consider the following relations:

$$\begin{aligned} N &= \{(1, 0), (2, 5), (2, 6), (3, 4)\} \\ P &= \{(-2, 5), (0, 6), (2, 7), (4, 8)\} \\ Q &= \{(-3, 2), (-1, 3), (1, 3), (4, 3)\} \end{aligned}$$

The following **arrow diagram** provides a picture of these relations:



Notice that in relation **N**, the number 2 in the domain is assigned (or “mapped”) to *two* different values in the range, that being 5 and 6. In this case, *N* is a relation, but it is **not** a function.

In relation **P**, each value in the domain of *P* is mapped to only one value in the range of *P*. We can then say that *P* is both a relation and a function.

In relation **Q**, again each value in the domain of *Q* is assigned to only one value in the range of *Q*. Note that some of the values in the domain of *Q* have the same range values (that of $(-1, 3)$, $(1, 3)$, and $(4, 3)$), but this is permissible as long as no domain value has more than one range value. Thus, *Q* is both a relation and a function.

These examples lead us to the following definition:

DEFINITION

Function: Version 1

A *function* f on X with values in Y , denoted by $f : X \rightarrow Y$, is a relation $f \subseteq X \times Y$ such that:

- For every $x \in X$ there exists exactly one $y \in Y$ for which (x, y) is in f .

In this case, we denote the value y by $f(x)$ and write $y = f(x)$.

Given $f : X \rightarrow Y$,

- X is called the *domain of f* and is denoted by $\text{dom}(f)$.
- Y is called the *co-domain of f* and is denoted by $\text{codom}(f)$.
- $\{y = f(x) \mid x \in X\}$ is called the *range of f* and is denoted by $\text{ran}(f)$.
- The *graph of f* is the set

$$\text{graph}(f) = \{(x, y) = (x, f(x)) \mid x \in X\} \subseteq X \times Y.$$

In the above definition we ask that the domain of our relation be the entire set X . In this course we will typically be looking at functions whose domain and co domain are both \mathbb{R} . However, if we view the function given by the formula

$$f(x) = \sqrt{x}$$

the natural domain for this function is $\{x \in \mathbb{R} \mid x \geq 0\}$ rather than all of \mathbb{R} . As such it will be convenient if we extend the definition of a function as follows:

DEFINITION

Function: Version 2

Let $R \subseteq X \times Y$ be a relation. Then R determines a function f if and only if

- whenever $(x, y_1) \in R$ and $(x, y_2) \in R$, we have $y_1 = y_2$.

In this case, we let

$$\text{dom}(f) = \{x \in X \mid (x, y) \in R \text{ for some } y \in Y\}$$

to get $f : \text{dom}(f) \rightarrow Y$ and write

$$y = f(x)$$

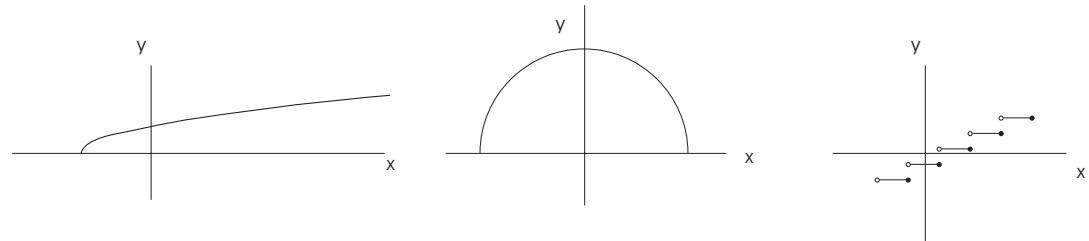
if $x \in \text{dom}(f)$ and $(x, y) \in R$.

Graphically, there exists an easy test to determine if a relation is a function.

The Vertical Line Test for Functions

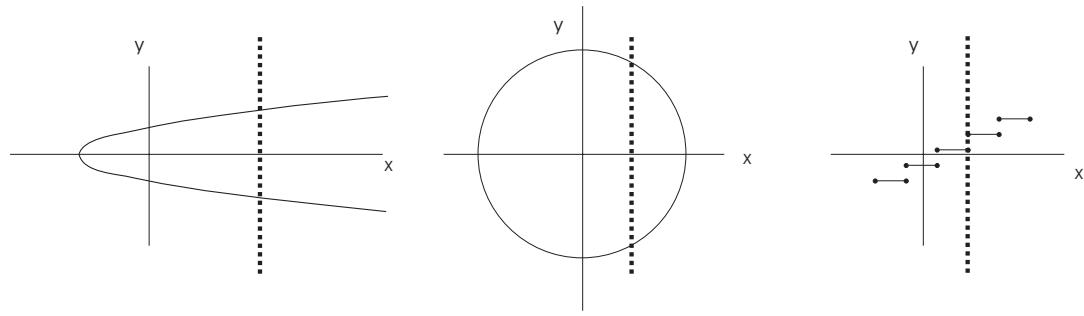
The vertical line test indicates whether or not a graph is a function. If any **vertical** line can be drawn that intersects a graph more than once, then the graph can **not** be a function. Why does this work? If the vertical line intersects the graph more than once, it means that two or more y values in the range are mapped to some x value in the domain.

Examples of functions:



Note: In the third graph, the left-hand points of each line segment are unfilled circles. These unfilled circles represent points that are NOT part of the graph. In a similar manner, the right-hand points of each line segment are filled. These filled circles represent points that ARE part of the graph.

Examples of relations that are NOT functions:



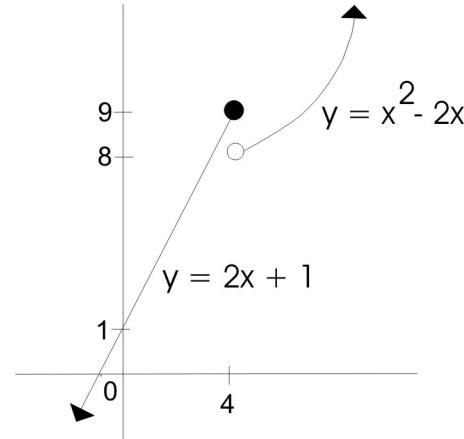
More often than not we will be given our function by specifying only the mathematical formula that determines its values. For example

$$g(x) = \sqrt{x}$$

We are not given the domain explicitly. In this case our convention will be to choose as the domain the *largest subset* of \mathbb{R} for which the formula makes sense. In the case of our “square root” function, the formula is valid for all $x \geq 0$. We will therefore assume that $\text{dom}(g) = [0, \infty)$. In other words, the domain is 0 along with all positive real numbers.

Frequently we will encounter functions that are defined by different formulae over various regions of \mathbb{R} . For example, we can define the function

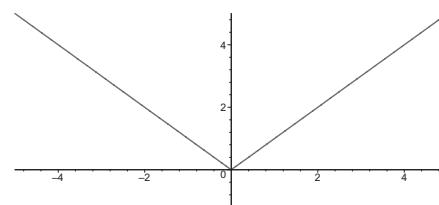
$$g(x) = \begin{cases} 2x + 1 & \text{if } x \leq 4 \\ x^2 - 2x & \text{if } x > 4 \end{cases}$$



Examples of evaluating these types of functions are thus $g(2) = 2(2) + 1 = 5$ and $g(7) = 7^2 - 2(7) = 35$.

Moreover, we have already seen another example of this type. Recall the definition of absolute value where if $f(x) = |x|$, then

$$f(x) = |x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}.$$



NOTE

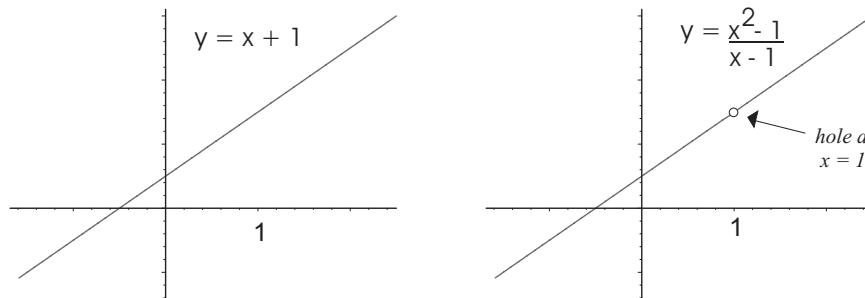
It is important to mention that when studying functions, we must always be conscious of the *domain on which it is defined*. To illustrate this point, consider two functions:

$$f(x) = x + 1 \quad \text{and} \quad g(x) = \frac{x^2 - 1}{x - 1}$$

Note that we can factor the numerator $x^2 - 1$ to get that

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

It is tempting to simply cancel the common factor of $(x - 1)$ to conclude that $\frac{x^2 - 1}{x - 1} = x + 1$ and then to conclude that f and g are really the same function in disguise. However, we note that the domain of $f(x)$ is all of \mathbb{R} whereas if we were to try and evaluate $g(x)$ at $x = 1$ we would be left with the *indeterminate value* $\frac{0}{0}$. Consequently, $\text{dom}(g) = \{x \in \mathbb{R} \mid x \neq 1\}$ is a **proper subset** of $\text{dom}(f)$ — their domains are not equal. Therefore, despite the fact that these functions are very similar in that they assign the same values to each real number other than 1, the fact that they have different domains means that *they are different functions*.



2.4 Composition of Functions

Many of the functions that will be used in this course can be built by combining two or more simple functions to create a more complex function. One way to do this is to use the operation of *composition*. To illustrate this process let's first consider the function

$$h(x) = \sqrt{x^2 + 1}.$$

To calculate the value of this function at a given value of x we would generally first calculate $x^2 + 1$ and then calculate the square root of this number.

Let us denote the function $x^2 + 1$ by $g(x)$ and write $f(w) = \sqrt{w}$. (Remember, the name of the variable is really not important.) Now to calculate $h(x)$ we first evaluate $g(x)$ and then we substitute this result for w in $f(w)$. Symbolically, this means that

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

For example, to calculate $h(2)$, let

$$w = g(2) = 2^2 + 1 = 5.$$

Then

$$h(2) = f(5) = \sqrt{5}.$$

This is a simple example that leads us to a very important method of constructing new functions from old functions. In fact, anytime we have two functions f and g such that

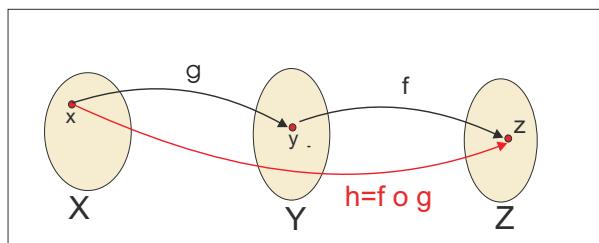
$$\text{ran}(g) \subset \text{dom}(f)$$

we can define a new function

$$h(x) \stackrel{\text{def}}{=} f(g(x)).$$

This new function is called the *composition of f by g* and is denoted by

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



DEFINITION

Composition of Functions

Let f and g be two functions such that $\text{ran}(g) \subset \text{dom}(f)$. The *composition of f by g* is the function $h(x) = f \circ g(x)$ defined by

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

There are a few important observations that must be made concerning the composition of two functions. Firstly, and most importantly, it is imperative that the condition $\text{ran}(g) \subset \text{dom}(f)$ be valid, otherwise the composition need *not* be defined. For example, if we were to compose the function $f(x) = \sqrt{x}$ with the function $g(x) = -1 - x^2$, we would be left with the function

$$h(x) = f \circ g(x) = f(g(x)) = \sqrt{-1 - x^2}.$$

However, it is easy to see that the range of $g(x) = (-\infty, -1]$. It follows that no matter what value of x we choose, $g(x)$ returns a negative number. This is a problem, because, f is not defined for negative values and hence, $f(g(x))$ is *impossible to calculate*. As such, this particular composition does not make sense.

The example above also leads us to the second important fact that we should know about compositions. **Order matters!** Indeed, we have seen that for the functions f and g above, $f \circ g(x)$ is not defined for any x . However, for any $x \geq 0$, that is for any $x \in \text{dom}(f)$, we get that $f(x) \in \text{dom}(g)$. Therefore, the composite function

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

makes sense for any $x \geq 0$. In fact,

$$\begin{aligned} k(x) &= g \circ f(x) = g(f(x)) = -1 - (\sqrt{x})^2 \\ &= -1 - x \end{aligned}$$

since for any $x \geq 0$, $(\sqrt{x})^2 = x$.

We must still be somewhat cautious in dealing with this composition. Looking at the function $k(x) = -1 - x$, we might be tempted to claim that the domain of this function is all of \mathbb{R} . However, we must always be aware that the first step in evaluating $k(x) = g \circ f(x)$ is to first calculate $f(x) = \sqrt{x}$. This can only be done if $x \geq 0$. It follows that $\text{dom}(k) = \{x \in \mathbb{R} \mid x \geq 0\}$.

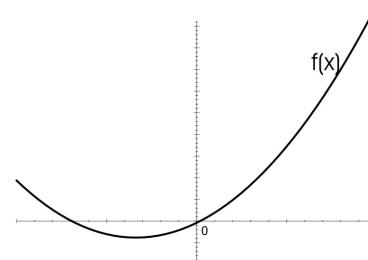
2.5 Transformations of Functions

We now turn our attention to a number of basic operations that can be performed to obtain new functions from a given function. These operations include translation, scaling, and reflections. We will also look at various types of symmetry that a graph of a function might exhibit.

2.5.1 Translations

In this section, we will see how to use translation to create new functions from a given function. Translations appear in many applications of mathematics and result from such phenomena as time delay and phase change.

Let's assume that we have a function $f : \mathbb{R} \rightarrow \mathbb{R}$ that has the following graph.

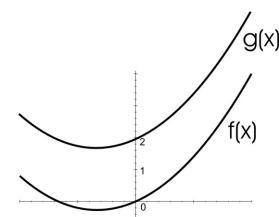


We can create a number of new functions from $f(x)$.

For example, we can define

$$g(x) = f(x) + 2$$

To evaluate $g(x)$ at a point x , we first evaluate $f(x)$ and then add 2. This has the effect of shifting every point on the graph of $f(x)$ up 2 units to obtain the graph of g . The diagram illustrates this shift.

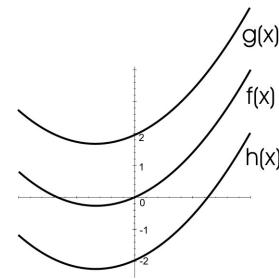


The new function g is a type of *translation* of f called a *vertical shift*.

If instead of adding 2 to $f(x)$, we *subtracted* 2, we would get the new function

$$h(x) = f(x) + (-2) = f(x) - 2$$

The graph of h is shown with the graph of the original function f and is contrasted with that of g .



We are led to the following rule for vertical shift translations.

Vertical Shift Rule:

Suppose that a is any real number and that

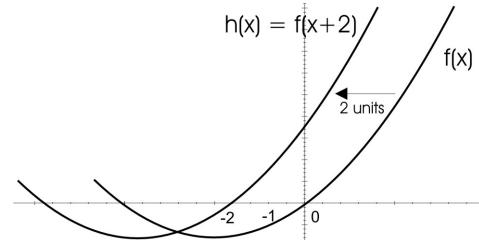
$$g(x) = f(x) + a.$$

Then,

- 1) If $a > 0$, then the graph of g is simply the graph of f shifted *upwards* by a units.
- 2) If $a < 0$, then the graph of g is simply the graph of f shifted *downwards* by $|a| = -a$ units.

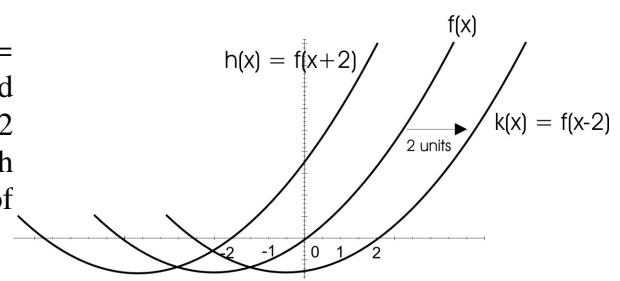
There is another type of translation that we will need to understand. Consider the function $h(x) = f(x+2)$ where f is the function from above. While this looks similar to the expression for $g(x) = f(x) + 2$, it is indeed a different function. In evaluating $g(x)$ at a point x , we first evaluate $f(x)$ and then add 2 to this number.

In the case of $h(x)$, we first add 2 to the value x and *then apply* the function f . This means that the value of $h(0)$ is the same as the value of $f(0+2) = f(2)$ and the value of $h(-2)$ is the same as the value of $f(-2+2) = f(0)$. If we were to plot $f(x)$ and $h(x)$ simultaneously, we see that they both have the same shape. Indeed, the graph of $h(x)$ is just the graph of $f(x)$ shifted 2 units to the *left*.



Consequently, we call this translation of $f(x)$ to $h(x)$ a *horizontal shift*.

If we were to define $k(x) = f(x+(-2)) = f(x-2)$, then the values of $k(x)$ at 4 and 2 would agree with the values of f at 2 and 0, respectively. In this case, the graph of $k(x)$ is the same shape as the graph of $f(x)$, but shifted to the *right* by 2 units.



We are now able to state the rule for *horizontal shift* translations.

Horizontal Shift Rule:

Suppose that a is any real number and that

$$h(x) = f(x + a).$$

Then,

- 1) If $a > 0$, then the graph of $h(x)$ is simply the graph of $f(x)$ shifted to the *left* by a units.
- 2) If $a < 0$, then the graph of $h(x)$ is simply the graph of $f(x)$ shifted to the *right* by $|a| = -a$ units.

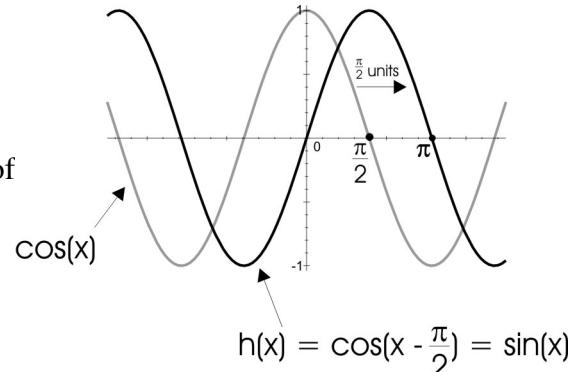
NOTE

Although the Vertical Shift Rule is intuitive in that $+$ represents a shift *up* and $-$ represents *down*, take notice that in the Horizontal Shift Rule, while we might **expect** $+$ to represent a shift to the *right* and $-$ to the *left*, in fact exactly the **opposite** occurs!

**EXAMPLE 3**

Consider the function $\cos(x)$ and translate the graph $\frac{\pi}{2}$ units to the right.

Solution: The graph of $\cos(x)$ and that of the new function $h(x)$ are shown.



You might notice that the graph of the new function $h(x)$ looks suspiciously like the graph of $\sin(x)$. In fact, the rule for horizontal shift translations tells us that

$$h(x) = \cos(x - \frac{\pi}{2}).$$

However, the rules for cosines of sums gives us that

$$\begin{aligned} h(x) &= \cos(x - \frac{\pi}{2}) \\ &= \cos(x) \cos(\frac{-\pi}{2}) - \sin(x) \sin(\frac{-\pi}{2}) \\ &= \cos(x) \cdot 0 - \sin(x) \cdot (-1) \\ &= \sin(x) \end{aligned}$$

just as we suspected.

This also shows that if we were to start with $\sin(x)$ and translate the graph $+\frac{\pi}{2}$ units to the *left*, we should get back $\cos(x)$. Using the rule for horizontal shifts, we have the identity

$$\cos(x) = \sin(x + \frac{\pi}{2})$$

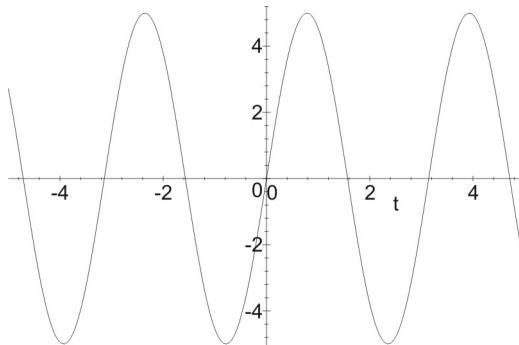
which you can verify by using the rule for the $\sin(\alpha + \beta)$ that we studied in the trigonometry section. ◀

EXAMPLE 4

A pool that is 1m deep everywhere has two identical wave machines located at either end. Consider a particle located at the point on the surface exactly in the middle of the pool. We want to plot the height of the particle as a function $h(t)$ of time as the waves generated by the two machines pass by. (Here $h(t)$ represents the height of the particle above a fixed reference point measured in centimetres.)

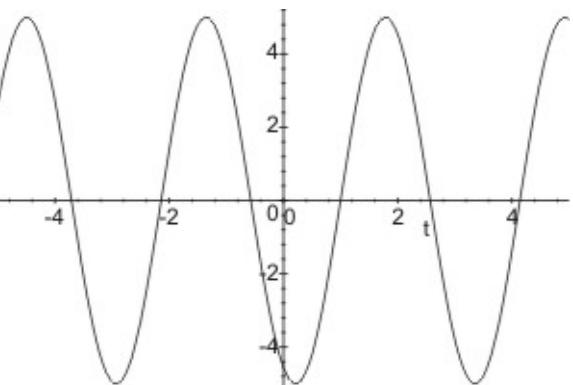
If both machines are turned off, the function $h(t)$ is constant. For our purposes we can assume that the height in this case is set to 0m.

Suppose, on the other hand, that we began measuring the height at exactly 12:00pm and that one of the wave machines had been running for some time. With this in mind we will set 12:00pm to be time zero and let the variable t represent the number of minutes that have passed. Let $H(t)$ represent the height of the particle if *only the first* wave machine was turned on. We would find that the graph of $H(t)$ would have the same basic shape as a translation of the graph of $\sin(t)$. (The movement of the particle up-and-down caused by the wave machine is called *simple harmonic motion*.)

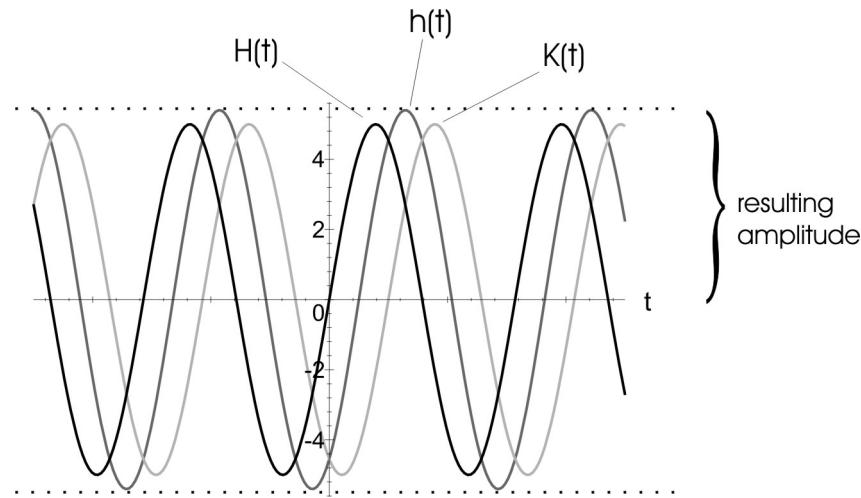


Suppose that one minute after the first machine was started, the operator started the second machine. Since the point is in the middle of the pool, the effect of the waves generated by the second machine would be identical to the first, but we would have a one minute delay in this case.

This means that the effect from the second wave at time t would be the same as the effect of the first machine at time $t - 1$. Thus, we get a new function, $K(t) = H(t - 1)$. $K(t)$ represents the height that the particle would reach t minutes after noon if only the second machine had been turned on. Its graph, which is a horizontal shift of the graph of H , is as follows.

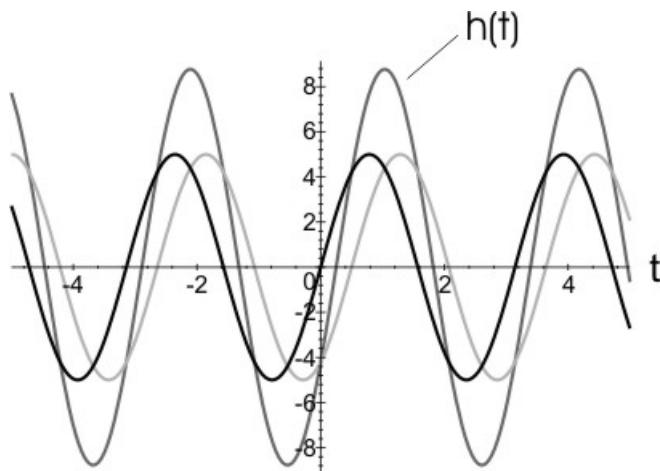


Now these two graphs represent the effect that the wave machines would each have on the particle separately. It turns out that if both machines were running, with the second starting 1 minute after the first, we could find the net effect on the particle by *adding* the two functions. In the picture below, $H(t)$ represents the effect that the *first* machine has on the particle, $K(t)$ represents the effect that the *second* machine has on the particle, and $h(t)$ represents the *net effect* that both machines have on the particle.

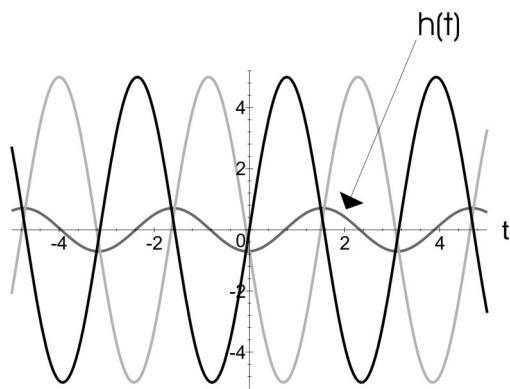


Note that the maximum height (called its *amplitude*) of the new wave $h(t)$ is slightly larger than that of either the two original waves.

It turns out that the size of the delay in starting the second machine is very important. The graph below represents the scenario with only a 30 second delay. Notice that the original waves are now a closer match to one another and the resulting net wave has much greater highs and lows (larger amplitude).



In contrast, if the delay was 1.5 minutes, we would have the following situation. Notice that in this case, the original waves are nearly mirror images of one another. This has the effect of causing the two waves to nearly *cancel* one another out. Consequently, the net wave has a rather small amplitude.



2.5.2 Scaling

We now consider the transformation called *scaling*. There are two types of scaling—in x and in y . Both can result in either compression or stretching in the graph of the function. Algebraically, this process involves multiplication by a constant c .

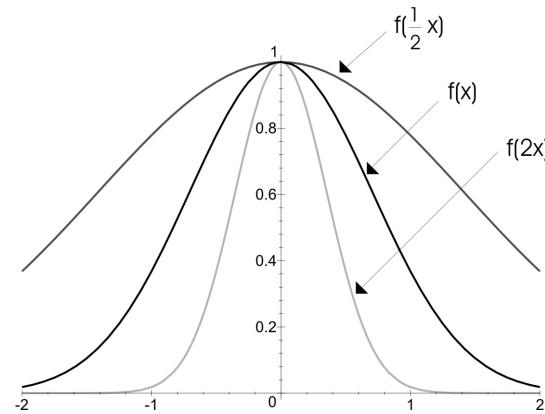
Scaling in x :

Given a positive constant c and a function $f(x)$, the new function $g(x)$ defined by

$$g(x) = f(cx)$$

is called an *x-scaling of $f(x)$* .

x-scaling causes a *horizontal stretching* away from the y axis of the graph of f if $0 < c < 1$ and a *horizontal compression* toward the y axis of f if $c > 1$.



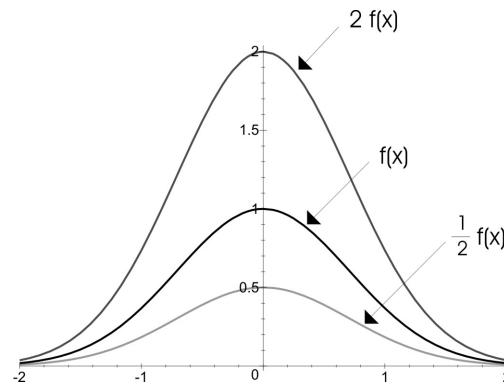
Scaling in y :

Given a positive constant c and a function f , the new function g defined by

$$g(x) = cf(x)$$

is called an *y-scaling of f* .

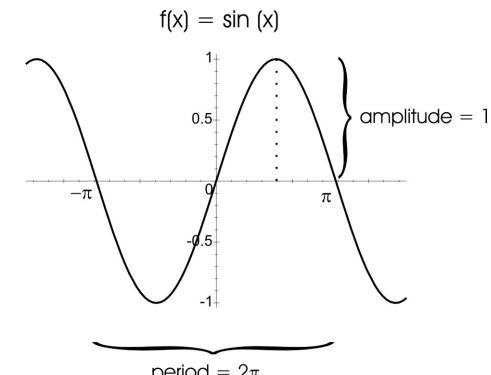
y-scaling causes a *vertical stretching* away from the x axis of the graph of f if $c > 1$ and a *vertical compression* toward the x axis of $f(x)$ if $0 < c < 1$.



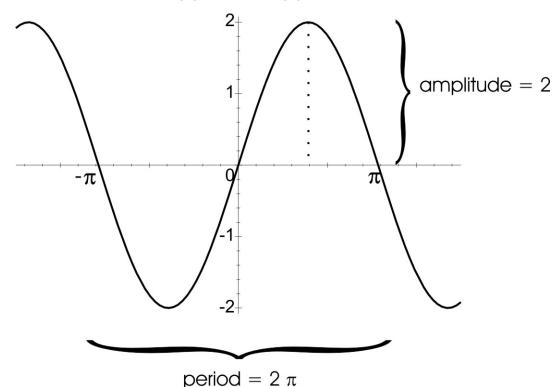
EXAMPLE 5

This example illustrates the transformation of scaling in the function $f(x) = \sin(x)$.

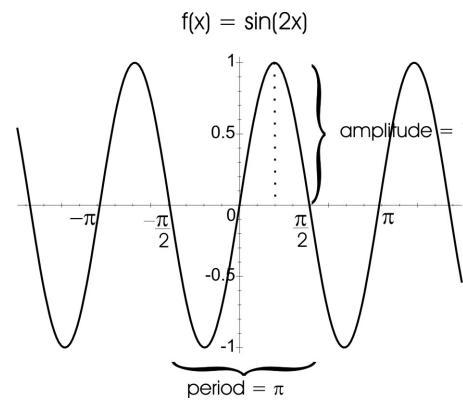
On the first set of axes, $f(x) = \sin(x)$ is plotted. Notice that $\sin(x)$ is 2π periodic and its amplitude is 1.



The second set of axes displays $f(x) = 2 \sin(x)$. Notice that $2 \sin(x)$ is also 2π periodic. However, its amplitude is 2 (i.e., twice that of $\sin(x)$), as this is the result of vertical stretching since the \sin function was multiplied by a constant $c > 1$.

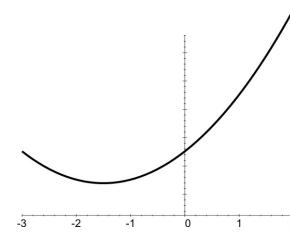


The third set of axes displays $f(x) = \sin(2x)$. Here, the original sin function is compressed horizontally (by half), since the variable x in the original function was multiplied by a constant $c > 1$ (actually $c = 2$ here). The period of $\sin(2x)$ is only π units (i.e., $\frac{1}{2}(2\pi)$), but its amplitude is still 1.



2.5.3 Reflections

Consider a function f with the following graph.



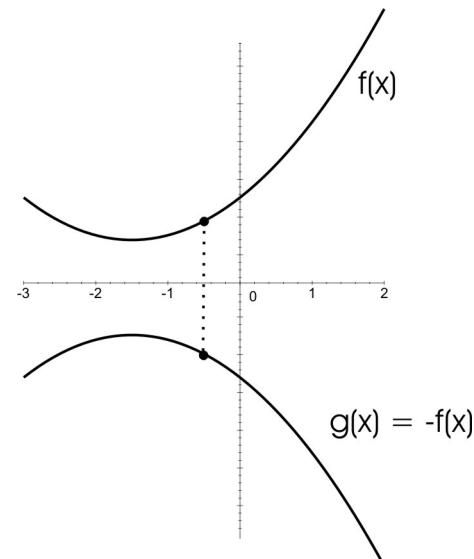
We can define two new functions

$$g(x) = -f(x)$$

and

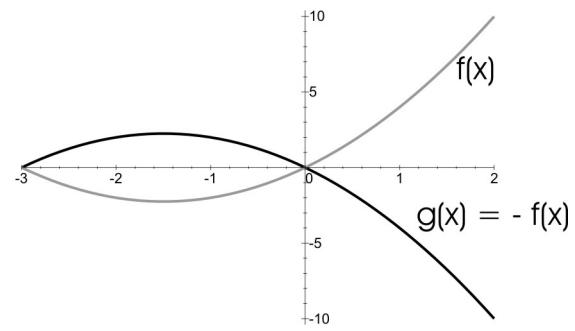
$$h(x) = f(-x).$$

Observe that for any x , the values $f(x)$ and $g(x)$ have exactly the same magnitude but have *opposite sign*. This means that the points on the graph of the two functions with the same x -coordinate are mirror reflections of one another through the x -axis. To illustrate this point, we have included the graph of $f(x)$ and $g(x)$ on the same axes.

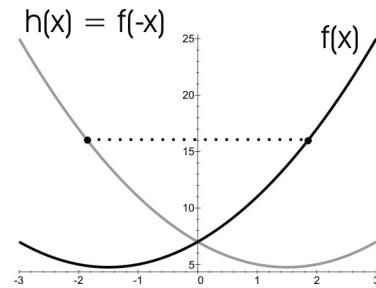


You can see that since we started with a function f that was always positive its graph is in the upper half plane. As such its reflection is entirely in the lower half plane just as we would expect since $f(x) > 0$ implies $g(x) < 0$.

Even if $f(x)$ is not positive, the graph of $g(x) = -f(x)$ will still be the reflection of the graph of $f(x)$ through the x -axis. Indeed, the picture illustrates that this reflection principle still applies even if $f(x)$ is modified so that $f(x)$ is no longer always positive.



In the case of the function, $h(x) = f(-x)$, we find that the value of the function h at 1 is the same as the value of the function f at -1 . Thus the points $(1, h(1))$ and $(-1, f(-1))$ have the same y -component and are equidistant from the y -axis. In fact this is true of any pair, $(x, h(x))$, $(-x, f(-x))$. This means that the graphs of f and h are again mirror reflections of one another, though in this case the reflection is through the y -axis.



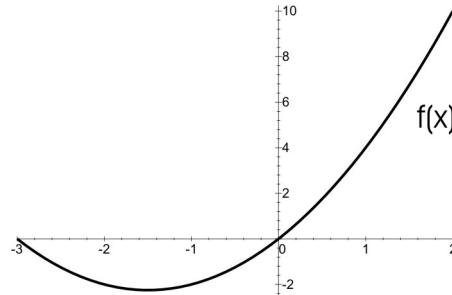
Notice that the graphs cross at $x = 0$. This is because

$$f(0) = f(-0) = h(0).$$

We can modify the operations in this section in an interesting way by introducing absolute values into our investigation. What would the function

$$k(x) = |f(x)|$$

look like if $f(x)$ is as given below?



We need to appeal to the definition of the absolute value to get a feel for what $k(x)$ will look like. Recall that for a real number a , we have

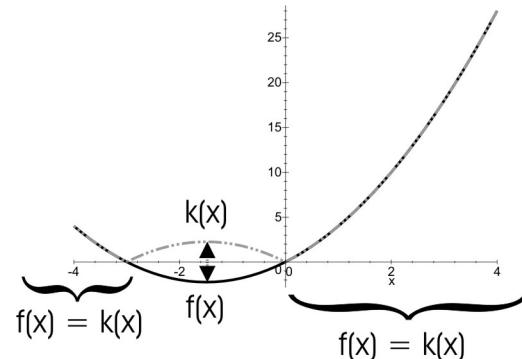
$$|a| = \begin{cases} a & \text{if } a \geq 0 \\ -a & \text{if } a < 0 \end{cases}.$$

It follows that

$$k(x) = |f(x)| = \begin{cases} f(x) & \text{if } f(x) \geq 0 \\ -f(x) & \text{if } f(x) < 0 \end{cases}$$

This means that $f(x)$ and

$k(x) = |f(x)|$ return exactly the same value whenever x is such that $f(x) \geq 0$ and $k(x)$ agrees with $-f(x)$ whenever $f(x) < 0$. Following our reasoning from earlier in this section, we see that in order to obtain the graph of $k(x)$ from that of $f(x)$, we do nothing when $f(x) \geq 0$ and for those x 's for which $f(x) < 0$, we reflect through the x -axis.



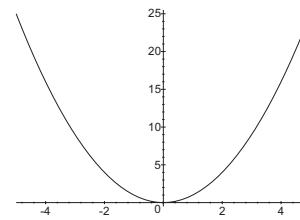
We see that the only difference between the graphs of $f(x)$ and $k(x)$ happens when $f(x) < 0$. Moreover, for these x 's, the new graph is indeed the reflection of the old one through the x -axis.

2.5.4 Symmetries: Even and Odd Functions

Consider the function $f(x) = x^2$. This function has the property that for any x ,

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

Moreover, the graph of $f(x)$ is symmetric about the y -axis. In fact, any function that satisfies $f(x) = f(-x)$ will always have a graph that is symmetric about the y -axis. Recall that the graph of $f(-x)$ is the reflection through the y -axis. Therefore, if $f(x) = f(-x)$, the graph of $f(x)$ is its own reflection and as such must be symmetric.



DEFINITION

Even Function

A function $f(x) : \mathbb{R} \rightarrow \mathbb{R}$ is called *even* if

$$f(x) = f(-x)$$

for all x .

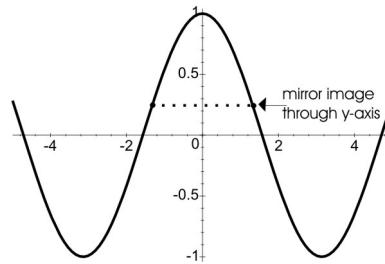
EXAMPLES

- 1) The function $f(x) = x^2$ is even. Indeed, the term “even” essentially comes from the fact that the polynomial

$$p(x) = x^n$$

is an even function if and only if n is an even integer. Consequently, x^4, x^6, \dots are all even functions.

- 2) The function $f(x) = |x|$ is even.
- 3) Constant functions of the form $g(x) = c$ are even functions.
- 4) The identity $\cos(x) = \cos(-x)$ means that cosine is also an even functions.



- 5) Assume that $f(x)$ and $g(x)$ are even functions. Let $\alpha, \beta \in \mathbb{R}$. Then the functions

$$h(x) = \alpha f(x) + \beta g(x)$$

and

$$k(x) = f(x)g(x)$$

are also even. To see this note that

$$\begin{aligned} h(-x) &= \alpha f(-x) + \beta g(-x) \\ &= \alpha f(x) + \beta g(x) \\ &= h(x) \end{aligned}$$

since $f(x) = f(-x)$ and $g(x) = g(-x)$. The case for the product is similar.

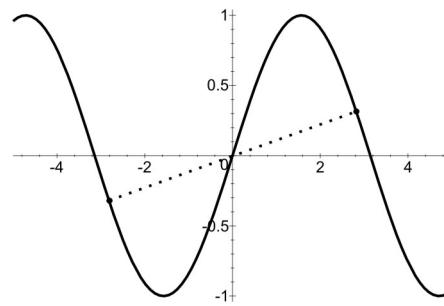
- 6) The function $\sin(x)$ is *not* even since $\sin(\frac{\pi}{2}) = 1$ while $\sin(-\frac{\pi}{2}) = -1$.

REMARK

We saw above that $\sin(x)$ is *not* an even function. In fact, we know from trigonometry that

$$\sin(-x) = -\sin(x)$$

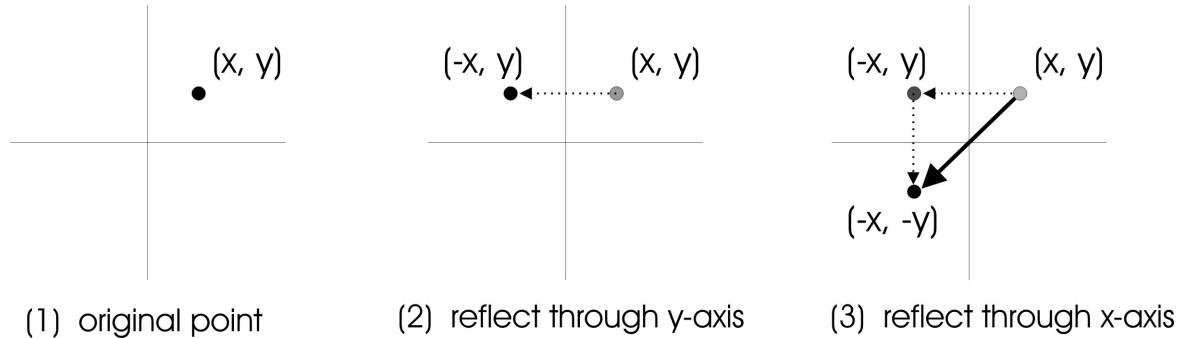
However, if we look at the graph of $\sin(x)$, it certainly does have symmetry, though not with respect to the y -axis as was the case for even functions. In fact, a close examination shows that the graph of $\sin(x)$ is actually *symmetric about the origin* in that $(-x, \sin(-x))$ and $(x, \sin(x))$ are mirror images of each other through the origin.



This symmetry is typical of any function $f(x)$ for which

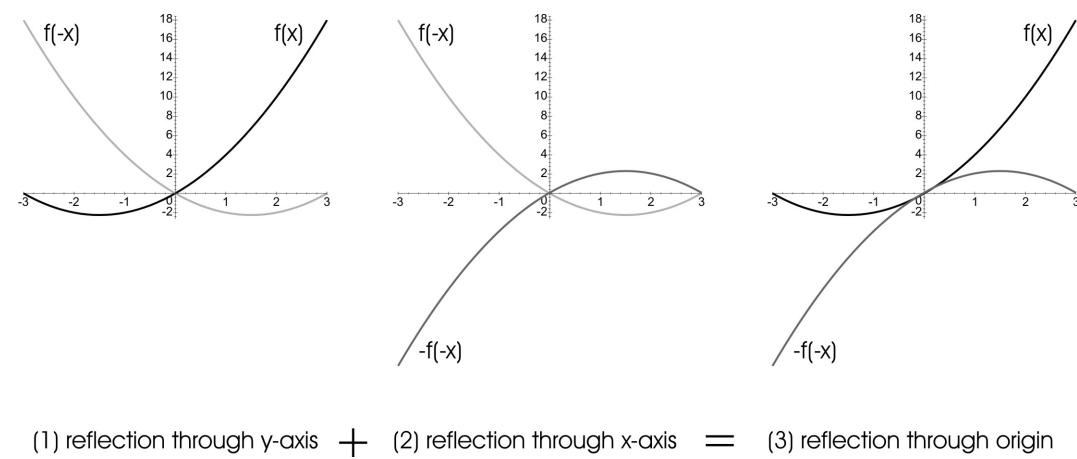
$$f(x) = -f(-x).$$

To see why this is true, first take any point (x, y) and reflect it through the y -axis to get $(-x, y)$. Then reflect $(-x, y)$ through the x -axis to get $(-x, -y)$. This point becomes the mirror image of (x, y) through the origin.



A reflection through the y -axis followed by a reflection through the x -axis = a reflection through the origin.

Now the function $-f(-x)$ is obtained from $f(x)$ by first constructing $f(-x)$ by reflecting the graph of $f(x)$ through the y -axis and then obtaining $-(f(-x))$ by reflecting the resulting graph through the x -axis.



This means that for any function $f(x)$, the graph of $-f(-x)$ is the *mirror image* of the original graph through the origin. It follows that $f(x) = -f(-x)$ if and only if the graph of $f(x)$ is the same as its *reflection through the origin*. In other words, it is symmetric about the origin.

The functions $f(x) = x$ and $g(x) = x^3$ are easily seen to have the property above as does any other polynomial of the form $p(x) = x^n$ where n is an *odd* integer. For this reason we are led to the following definition. 

DEFINITION**Odd Function**

A function $f(x) : \mathbb{R} \rightarrow \mathbb{R}$ is called an *odd* function if

$$f(x) = -f(-x)$$

for all x .

It is important to note that most functions are *neither* even nor odd. For example, if we let $f(x) = x^2 + 3x$, then $f(1) = 4$ while $f(-1) = -2$. It follows that $f(x)$ is neither even or odd.

Questions:

- (a) We have seen that the sum of two even functions is even (Example 5 above). Is the sum of two odd functions odd?
- (b) What can be said about the product of two odd functions?

2.6 Inverse Functions

There are times when we want to be able to reverse a process. Similarly, there are times when we want to be able to undo the effects of a function. To accomplish this task requires the existence of an *inverse function*. In this section, we will define what we mean by the inverse of a function and determine the conditions under which inverses exist.

2.6.1 One-to-one and Onto Functions

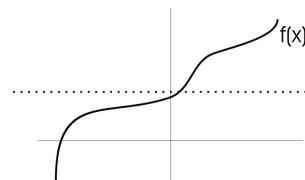
In this section we will assume that we have a function $f : X \rightarrow Y$. The definition of a function requires that we assign to each $x \in X$ a unique $y \in Y$, denoted by $f(x)$. The definition did not require that we assign different x 's to different y 's, nor did every y have to be the image of some x . This leads us to two special classes of functions.

DEFINITION**One-to-one and Onto Functions**

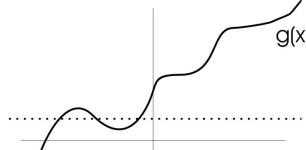
A function $f : X \rightarrow Y$ is said to be $1 - 1$ (reads “one to one”) if f assigns different x ’s to different y ’s. That is, whenever $x_1, x_2 \in X$ with $x_1 \neq x_2$, then $f(x_1) \neq f(x_2)$.

A function $f : X \rightarrow Y$ is *onto* if $\text{ran}(f) = Y$. That is, if for every $y_0 \in Y$, there exists some $x_0 \in X$ such that $f(x_0) = y_0$.

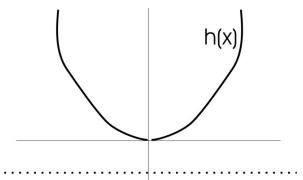
Visually, for a function $f(x)$ from \mathbb{R} to \mathbb{R} , f is $1 - 1$ if every *horizontal line* crosses the graph of $f(x)$ *at most once*. The function $f(x)$ is *onto* if every horizontal line crosses the graph *at least once*.



The function $f(x)$ is $1-1$ since a horizontal line crosses the graph of $f(x)$ at most once.



The function $g(x)$ is NOT $1-1$ since a horizontal line crosses the graph of $g(x)$ more than once. However, this function IS onto.



The function $h(x)$ is NOT $1-1$ and it is NOT onto.

EXAMPLE 6

Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be defined by $f(x) = x^2$. Then note that $\text{ran}(f) = \{y \in \mathbb{R} \mid y \geq 0\}$ is a proper subset of \mathbb{R} . Thus, $f(x)$ is *not onto*. Moreover, $f(1) = 1 = f(-1)$, so $f(x)$ is not $1 - 1$. ◀

EXAMPLE 7

Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be defined by $f(x) = x^3$. We can show that the function $f(x) = x^3$ is $1 - 1$ by using some basic algebra. Suppose that we have a fixed point $w \in \mathbb{R}$. We want to find all points x such that $f(x) = x^3 = w^3 = f(w)$. This means that

$$0 = x^3 - w^3 = (x - w)(x^2 + xw + w^2)$$

One way to find such values for x is to solve

$$0 = (x^2 + xw + w^2).$$

Since we have a fixed w , we can apply the quadratic formula to the polynomial $p(x) = x^2 + xw + w^2$ to look for

$$0 = x^2 + xw + w^2 = p(x).$$

However, the *discriminant* in this case is

$$w^2 - 4w^2$$

which is $-3w^2$. However, $-3w^2 < 0$ if $w \neq 0$ and we only get solutions if the discriminant is *nonnegative*. This means that we can only find an x so that

$$x^2 + xw + w^2 = 0$$

if $w = 0$. But then the equation becomes

$$x^2 = 0$$

and hence we also have that $x = 0 = w$. Otherwise, to get

$$x^3 - w^3 = 0$$

we must have

$$x - w = 0$$

or $x = w$. Consequently, if $x \neq w$, then $x^3 - w^3 \neq 0$. This means that if $x \neq w$, then $f(x) \neq f(w)$ and hence that f is 1-1.

To see that f is *onto*, note that if $y \in \mathbb{R}$, then

$$y = (y^{\frac{1}{3}})^3 = f(y^{\frac{1}{3}})$$

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



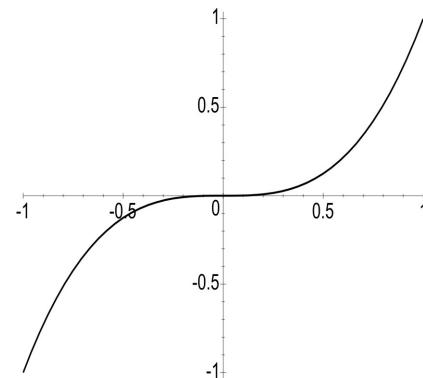
You may feel that the argument which showed that $f(x) = x^3$ is 1 – 1 was rather *complicated*. It would be nice if there was a better way to do this. In fact, there is!

EXAMPLE 8

If we look at the graph of $f(x) = x^3$, we see that if $x_1 < x_2$, then

$$f(x_1) < f(x_2).$$

Such a function is said to be *increasing*. We will soon show that *increasing functions are always 1 – 1*.



It is often the case that a function is *not* 1 – 1 on all of its domain, but it is if we restrict our attention to a *particular subset of the domain*. For example, if we have two different positive real numbers x_1 and x_2 , then $x_1^2 \neq x_2^2$. This means that the function $f(x) = x^2$ is 1 – 1 on the set $\{x \in \mathbb{R} \mid x \geq 0\}$.

DEFINITION**One-to-one Function on an Interval**

Let I be an interval contained in the domain of a function $f(x)$. We say that $f(x)$ is $1 - 1$ on I if whenever $x_1, x_2 \in I$ with $x_1 \neq x_2$, then $f(x_1) \neq f(x_2)$.

The previous example gives us a clue as to how to determine possible intervals on which $f(x)$ could be $1 - 1$.

We will now look at perhaps the most important classes of $1 - 1$ functions on \mathbb{R}

DEFINITION**Increasing and Decreasing Functions**

Suppose that f is defined on an interval I .

- i) We say that f is *increasing on I* if $f(x_1) < f(x_2)$ for all $x_1, x_2 \in I$ with $x_1 < x_2$.
- ii) We say that f is *decreasing on I* if $f(x_1) > f(x_2)$ for all $x_1, x_2 \in I$ with $x_1 < x_2$.
- iii) We say that f is *non-decreasing on I* if $f(x_1) \leq f(x_2)$ for all $x_1, x_2 \in I$ with $x_1 < x_2$.
- iv) We say that f is *non-increasing on I* if $f(x_1) \geq f(x_2)$ for all $x_1, x_2 \in I$ with $x_1 < x_2$.

If f satisfies any of these four conditions we say that f is *monotonic* on I . If f is either increasing or decreasing on I , we say that f is *strictly monotonic* on I .

The next Proposition tells us that functions that are either increasing or decreasing on a particular interval are $1 - 1$ on the interval.

PROPOSITION 5

Let I be an interval contained in the domain of a function f .

- a) If f is increasing on I , then it is $1 - 1$ on I .
- b) If f is decreasing on I , then it is $1 - 1$ on I .

PROOF

- a) Assume that $f(x)$ is increasing on I . Let $x_1, x_2 \in I$ with $x_1 \neq x_2$. Then we must show that $f(x_1) \neq f(x_2)$. To do this we can always assume that $x_2 > x_1$. Since $f(x)$ is increasing on I , we have that $f(x_2) > f(x_1)$. Hence $f(x_1) \neq f(x_2)$.

- b) This is essentially the same argument as in part a). Assume that $f(x)$ is decreasing on I . Let $x_1, x_2 \in I$ with $x_1 \neq x_2$. We can again assume that $x_2 > x_1$. Since $f(x)$ is decreasing on I , we have that $f(x_2) < f(x_1)$. Hence $f(x_1) \neq f(x_2)$.

■

2.6.2 Inverse Functions

Suppose that $f : X \rightarrow Y$ is one-to-one and onto. Then for every $y \in X$ there exists a unique $x \in X$ such that $f(x) = y$. This simple observation allows us to define the *inverse* function $g : Y \rightarrow X$ for f as follows:

DEFINITION

Inverse Function

If $f : X \rightarrow Y$ is 1 – 1 and onto, we can define a function $g : Y \rightarrow X$ by

$$g(y) = x \text{ if and only if } f(x) = y.$$

$g(y)$ is called the *inverse function of $f(x)$* and is often denoted by $f^{-1}(y)$.

REMARK

The inverse effectively *undoes* the action of f . To make this more precise we observe that if we start with an $x \in X$ and apply f to get $y = f(x)$, then if we apply our inverse function g to this value y we get

$$g(y) = g(f(x)) = x$$

That is

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

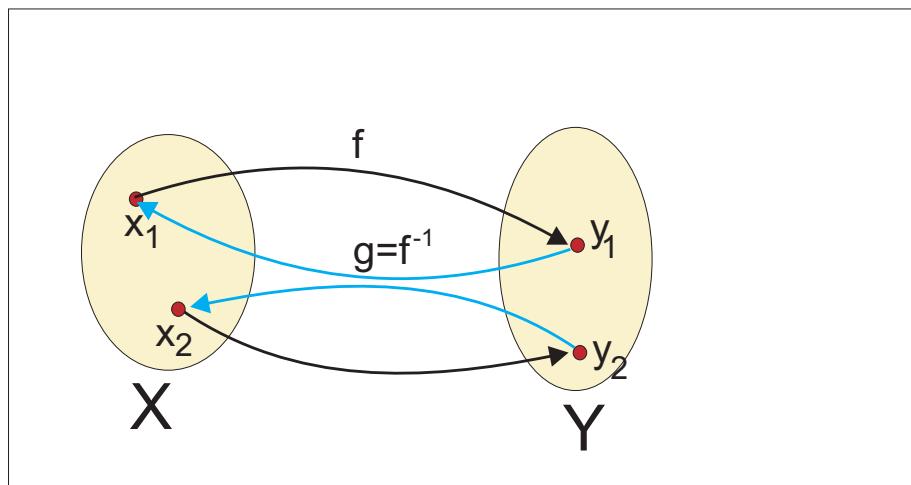
for every $x \in X$. A similar calculation also shows that for any $y \in Y$ we have

$$f(g(y)) = y$$

so that our original function f is also the inverse of g . That is

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

◀



The next example illustrates why less restrictive requirement for the existence of an inverse may be desirable.

EXAMPLE 9

Assume that you have a message that you want to send via satellite to a location on the other side of the world. Your message is text made up of words from the English alphabet.

Unfortunately, your transmitter is capable of sending only two different types of data, either a 0 or a 1. One way to send this message is to label all of the letters by a number from 1 – 26 and then to convert each number to a string of five 0's and 1's by converting the decimal numbers from 1-26 to their *binary representation* (i.e., *encode*).

For example, the first letter in the alphabet is “a”. It is assigned the decimal number 1 which has a five-digit binary representation 00001. Our last letter, “z”, corresponds to 26 which has a five-digit binary representation 11010.

The person who receives the message must be able to convert the string of 0's and 1's back into the original text (i.e., *decode*). For example, if the message is sent as 10011, the person who receives this must recognize that we are looking for the 19th letter of the alphabet, namely “s”.

Let X be the set of all 26 letters of the alphabet. The set Y will consist of all five digit strings of 0's and 1's. We can think of the *encoding* process as a function $f(x)$ from X with values in Y . The process of *decoding* the message requires us to take a value in $y \in Y$ and find the $x \in X$ such that $f(x) = y$.

You might notice an immediate problem with this. The string 11111 corresponds to the number 31 and we only have 26 letters. Hence, there is no such x for this y . We are saved from this by the fact that we will never receive the signal 11111, unless there is a transmission error so we really don't need worry about this. In fact, we are not interested in the elements of Y , but rather in the elements of the range of f .

Let $Y_1 = \text{ran}(f)$.

If we receive a signal y , to undo the process, we need to determine which x it came from. It turns out that we are able to do this because for each $y \in \text{ran}(f)$, there is *exactly one* letter of the alphabet such that $y = f(x)$. Indeed, the x is uniquely determined since each letter is encoded in a different way. This means that if $f(x_1) = y = f(x_2)$, then $x_1 = x_2$. This is just another way of saying that f is $1 - 1$.

Consequently, we get a new function $g : \text{ran}(f) \rightarrow X$ that can be defined by the statement that

$$x = g(y) \text{ if and only if } y = f(x)$$

Again, we can do this because each element in $\text{ran}(f)$ is the image of *one and only one* $x \in X$. That is, $f : X \rightarrow \text{ran}(f)$ is $1 - 1$ and *onto*.

As we have seen, the function $g(y)$ can be viewed as a way of undoing f . It also has the following interesting property. Suppose we start with an $x_1 \in X$. Let $y_1 = f(x_1) \in \text{ran}(f)$. If we ask which x is sent to y_1 , the answer is obviously x_1 . This means that

$$x_1 = g(f(x_1)) = g \circ f(x_1).$$

Indeed for any $x \in X$,

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

Therefore, $g \circ f(x)$ is the *identity function* of X .

Similarly, if we choose any $y \in \text{ran}(f)$, then the definition of g gives us that if $x = g(y)$, then $f(x) = y$. This means that

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

That is $f \circ g$ is the *identity function* on $\text{ran}(f)$. ◀

The process that we have identified above is quite general. In fact, it works whenever $f(x)$ is $1 - 1$. In this course, as is typical in Calculus, we will always take the view that a $1-1$ function is onto its range and that this will be sufficient for us to define our inverse. This will allow us for example to view $g(y) = \ln(y)$ as the *inverse* of $f(x) = e^x$ even though when viewed as a function $f : \mathbb{R} \rightarrow \mathbb{R}$ $f(x) = e^x$ is not onto and as such in the strictest sense is not invertible.

DEFINITION

Inverse of a Function: Version 2

Assume that $f : X \rightarrow \text{ran}(f)$ is $1 - 1$. We define a new function $g : \text{ran}(f) \rightarrow X$ by the statement that

$$x = g(y) \text{ if and only if } y = f(x).$$

This function is called the *inverse of f* and it is often denoted by f^{-1} . When the inverse exists, we say that f is *invertible* on X .

Our focus will be on real-valued functions. As you would expect, these functions are *not* generally invertible. For example, the function $f(x) = x^2$ is not invertible. To see

this we simply note that for $y = 1$ we get two x 's, 1 and -1 , such that $f(x) = 1$. This means that in defining the inverse function we don't know what to do with $y = 1$!

You probably guessed that the problem is that $f(x) = x^2$ is not $1 - 1$. However, if we restrict our attention to the interval $I = \{x \in \mathbb{R} \mid x \geq 0\}$, then $f(x) = x^2$ is $1 - 1$. This means that we can undo the effects of $f(x)$ on I . We note that the range of $f(x)$ over I is

$$f(I) = \{y \in \mathbb{R} \mid y \geq 0\}.$$

For each $y \geq 0$, there is indeed a unique $x \geq 0$ such that $y = x^2$. Given y , we denote this x by $g(y)$. Then $g(y)$ becomes a function from $f(I)$ back to I . In fact,

$$g(y) = \sqrt{y}.$$

You can verify again that if $x \geq 0$, then

$$g \circ f(x) = g(f(x)) = g(x^2) = \sqrt{x^2} = x$$

and if $y \geq 0$, then

$$f \circ g(y) = f(g(y)) = f(\sqrt{y}) = (\sqrt{y})^2 = y.$$

This shows that $g(y)$ has very similar properties to what we saw in the inverse of a function. This leads us to the following definition.

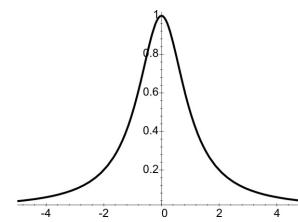
DEFINITION

Inverse of a Function over an Interval

Let f be $1 - 1$ on an interval $I \subseteq \mathbb{R}$. Then we say that f is invertible on I and define the *inverse of f* with respect to I to be the function $g(y) : f(I) \rightarrow I$ by $x = g(y)$ if and only if $x \in I$ and $y = f(x)$.

EXAMPLE 10

Let $f(x) = \frac{1}{1+x^2}$. The function is not $1 - 1$ on all of \mathbb{R} , but it is $1 - 1$ on the interval $I = \{x \in \mathbb{R} \mid x \geq 0\}$ since it is decreasing. We can also see that $f(x)$ takes on all values $J = \{y \mid 0 < y \leq 1\}$ as x ranges over I .



It follows that $f(x)$ is invertible on I and that its inverse is a function

$$g(y) : J \rightarrow I.$$

To find g , we can do as follows: First we write

$$y = \frac{1}{1+x^2}.$$

Then we try to solve for x in terms of y . We have

$$\begin{aligned} y &= \frac{1}{1+x^2} \\ (1+x^2)y &= 1 \\ yx^2 + y &= 1 \\ yx^2 &= 1-y \\ x^2 &= \frac{1-y}{y}. \end{aligned}$$

We can see that if $0 < y \leq 1$, then $1-y \geq 0$ and hence $\frac{1-y}{y} \geq 0$. This means that we can take square roots of both sides. Since $x \geq 0$, we have

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

The inverse function is given by

$$g(y) = \sqrt{\frac{1-y}{y}}$$

for $0 < y \leq 1$.

We can verify that $g(y)$ is the inverse by noting that

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

for $x \geq 0$ and

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

$$\begin{aligned}
 &= \frac{1}{\frac{y}{y} + \frac{1-y}{y}} \\
 &= \frac{1}{\frac{1}{y}} \\
 &= y
 \end{aligned}$$

just as we wanted.

You might also have noticed that $f(x)$ is also 1–1 on the interval $I_1 = (-\infty, 0]$ and that $f(I_1)$ is again the interval $J = \{y \mid 0 < y \leq 1\}$. It follows that $f(x)$ is also invertible on I_1 . This time the inverse is a function $h(y) : J \rightarrow I_1$. To find h we can again write

$$y = \frac{1}{1+x^2}$$

and try to solve for x in terms of y . As before, we have

$$\begin{aligned}
 y &= \frac{1}{1+x^2} \\
 (1+x^2)y &= 1 \\
 yx^2 + y &= 1 \\
 yx^2 &= 1-y \\
 x^2 &= \frac{1-y}{y}.
 \end{aligned}$$

However, we now have $x \leq 0$ so we get

$$-x = |x| = \sqrt{x^2} = \sqrt{\frac{1-y}{y}}.$$

This means that

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

so the new inverse is

$$h(y) = -\sqrt{\frac{1-y}{y}}.$$

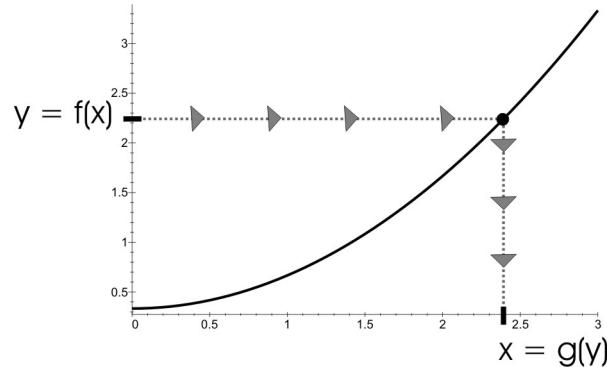


2.6.3 Graphing Inverse Functions

It turns out that if a function is invertible, you can find everything you need to know about the graph of the inverse function from the graph of the original function. To see this, let's begin with a function that we know is invertible—the function $f(x) = \frac{x^2}{3} + \frac{1}{3}$ on the interval $[0, \infty)$.

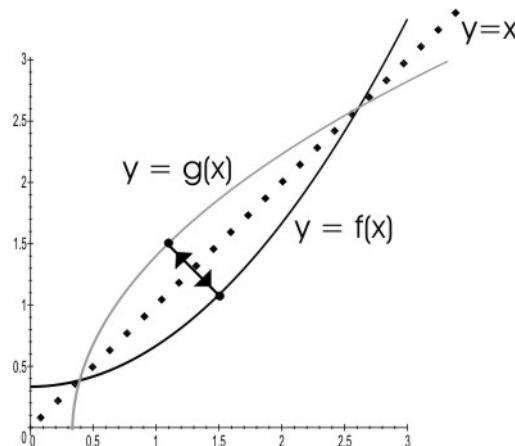
The arrows above indicate the procedure we would use to show how we would find the value of $f(x)$ from the graph of f .

Let $g(y)$ be the inverse function for $f(x)$. Since $f([0, \infty)) = [\frac{1}{3}, \infty)$, the domain of $g(y)$ is $[\frac{1}{3}, \infty)$ and its range is $[0, \infty)$. Moreover, we can still use the graph of $f(x)$ to calculate values of $g(y)$. We do this by simply reversing the direction of the arrows.



We have just seen that the graph of $x = g(y)$ and the graph of $y = f(x)$ are really the same object, but with the role of the x and y coordinates reversed. It follows that we can obtain the graph of g from the graph of f by simply exchanging the x and y coordinates.

We can do this geometrically by reflecting the graph of $f(x)$ through the line $y = x$. This gives us the graph of the inverse function with the independent variable on the horizontal axis.

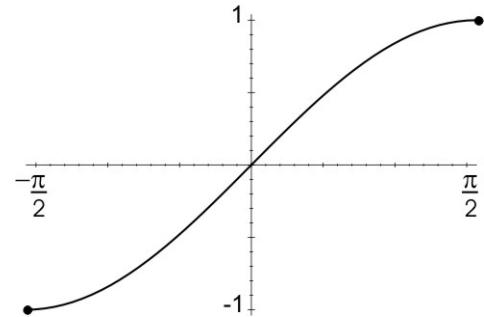


Summary: If $f(x)$ is invertible, then the graph of its inverse is the reflection of the graph of $f(x)$ through the line $y = x$.

2.6.4 Inverse Trigonometric Functions

At first look it might seem unreasonable to speak of the inverse for any of the standard trigonometric functions since none of these functions are $1 - 1$.

$$y = \sin(x)$$



However, if we look at $\sin(x)$, for example, we will see that it is increasing on the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and is therefore invertible on this interval.

On the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$, $\sin(x)$ takes on all values between -1 and 1 . Consequently, the inverse function is a function $g(y)$ with domain $[-1, 1]$ and range $[-\frac{\pi}{2}, \frac{\pi}{2}]$. We call this function the *arcsine* function and denote it by

$$g(y) = \arcsin(y).$$

From the definition of an inverse function, we get:

DEFINITION

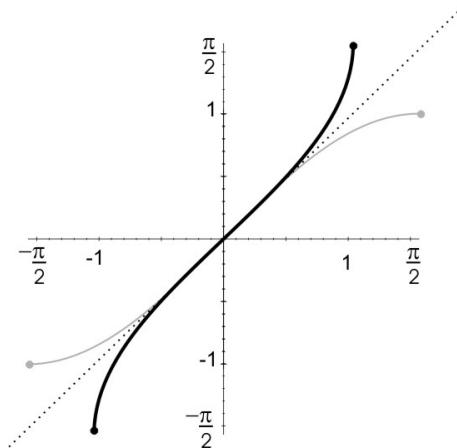
Arcsine Function

For each $y \in [-1, 1]$,

$$x = \arcsin(y) \text{ if and only if } y = \sin(x)$$

and $x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

$$y = \arcsin(x)$$



To obtain the graph of $\arcsin(x)$, we reflect the restricted graph of $\sin(x)$ through the line $y = x$, just as we did in the previous section.

You can see from the graph that $\arcsin(1) = \frac{\pi}{2}$. This is consistent since $\sin(\frac{\pi}{2}) = 1$. Similarly, $\arcsin(-1) = -\frac{\pi}{2}$ and $\arcsin(0) = 0$. Again, both of these results are consistent since $\sin(-\frac{\pi}{2}) = -1$ and $\sin(0) = 0$.

Observe that the composition $\sin(\arcsin(x))$ makes sense for any $x \in [-1, 1]$, which is the domain of $\arcsin(x)$. Furthermore, if we start with any $x \in [-1, 1]$ and let $y = \arcsin(x)$, then by definition $y = \arcsin(x)$ if and only if $\sin(y) = x$. Consequently, $\sin(\arcsin(x)) = \sin(y) = x$ for all $x \in [-1, 1]$.

It is also true that for all $y \in \mathbb{R}$, $\sin(y)$ belongs to $[-1, 1]$ which is the domain of $\arcsin(x)$. Hence the composition

$$\arcsin(\sin(y))$$

also makes sense for any $y \in \mathbb{R}$. The rules for inverse functions would tempt us to guess that

$$y = \arcsin(\sin(y))$$

for all $y \in \mathbb{R}$. Recall that $\arcsin(y)$ was constructed to be the inverse of $\sin(x)$ **only** for those x 's in the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Thus $y = \arcsin(\sin(y))$ is valid if $y \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, but not for other values of y .

For example, if $y = \frac{5\pi}{2} = \frac{\pi}{2} + 2\pi$, then

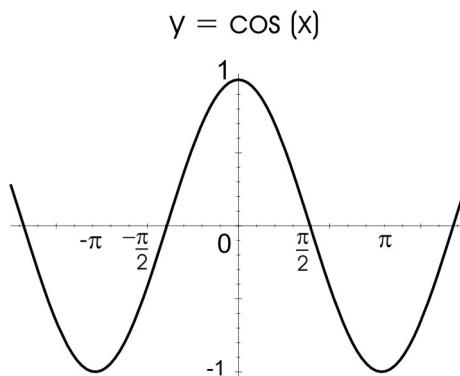
$$\sin(y) = \sin(\frac{5\pi}{2}) = \sin(\frac{\pi}{2} + 2\pi) = \sin(\frac{\pi}{2}) = 1.$$

Thus $\arcsin(\sin(\frac{5\pi}{2})) = \arcsin(1) = \frac{\pi}{2} \neq \frac{5\pi}{2} !!!$

There are two other important inverse trigonometric functions. They are derived from $\cos(x)$ and $\tan(x)$.

Like $\sin(x)$, neither $\cos(x)$ nor $\tan(x)$ are 1-1 on all of their domains. Consequently, we need to use the same trick as we did for $\sin(x)$ and find suitable intervals on which these functions are 1 – 1.

First lets look at $\cos(x)$. We could try $[-\frac{\pi}{2}, \frac{\pi}{2}]$, but $\cos(x)$ is *not* 1 – 1 on this interval as the graph shows.



However, the graph does suggest that if we use the interval $[0, \pi]$, then $\cos(x)$ is 1-1 and hence, invertible. We call the inverse of $\cos(x)$ on the interval $[0, \pi]$ the *arccosine* function and denote it by

$$y = \arccos(x)$$

The domain of $\arccos(x)$ is again the interval $[-1, 1]$ and the range is $[0, \pi]$. Formally we have:

DEFINITION

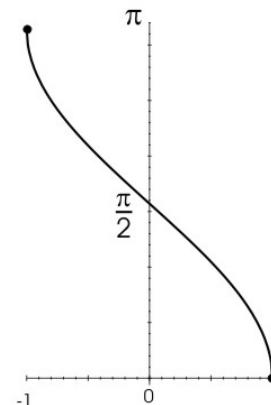
Arccosine Function

For each $y \in [-1, 1]$,

$$x = \arccos(y) \text{ if and only if } y = \cos(x)$$

and $x \in [0, \pi]$.

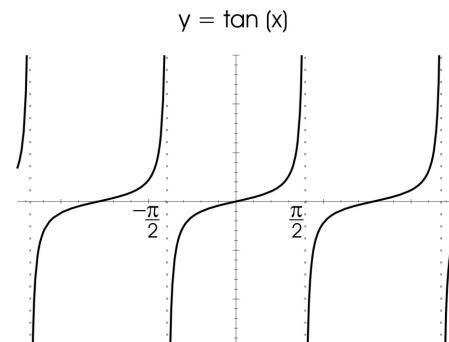
$$y = \arccos(x)$$



To obtain the graph of $\arccos(x)$, we reflect the restricted graph of $\cos(x)$ through the line $y = x$, just as we did for $\sin(x)$.

You may notice that it is not as easy to see that the graph of $\arccos(x)$ is the reflection of the graph of $\cos(x)$. You should verify the accuracy of this graph by calculating directly $\arccos(-1)$, $\arccos(0)$, and $\arccos(1)$. You should also verify that the domain and range are correct.

From the graph of $\tan(x)$ you will notice that while $\tan(x)$ is *not* 1-1 on its domain, like $\sin(x)$, it is 1-1 on the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$. **(Note:** The endpoints are *excluded* since $\cos(-\frac{\pi}{2}) = 0 = \cos(\frac{\pi}{2})$ and as such $\tan(x)$ is not defined on these points.



As x varies through $(-\frac{\pi}{2}, \frac{\pi}{2})$, $\tan(x)$ takes on all real values. Therefore, the inverse function of $\tan(x)$ on the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$, which we will denote by $\arctan(y)$, is defined for all $y \in \mathbb{R}$.

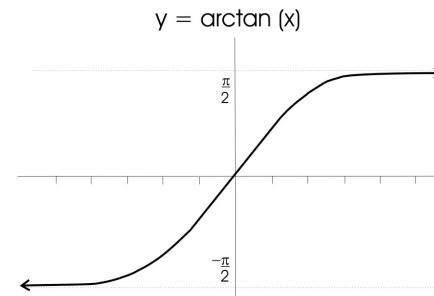
DEFINITION**Arctangent Function**

For each $y \in \mathbb{R}$,

$$x = \arctan(y) \text{ if and only if } y = \tan(x)$$

and $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$.

To obtain the graph of $\arctan(x)$, we reflect the restricted graph of $\tan(x)$ through the line $y = x$, just as we did before.

**2.7 Pullback**

We have just seen that if we have a function $f : X \rightarrow Y$ that is one-to-one and onto, then we can define the inverse function $g : Y \rightarrow X$ with the property that

$$g(y) = x \text{ if and only if } f(x) = y.$$

we have also seen that since f is always onto its range, we can effectively define the inverse function g from $\text{ran}(f)$ onto X provide that f is one-to-one. Moreover, if f is not one-to-one on all of X , but is one-to-one on a subset A of X , we can define the relative inverse g of X on A by

$$g(y) = x \text{ if and only if } f(x) = y \text{ and } x \in A.$$

In each of these cases it is common to use the notation f^{-1} for the inverse.

It turns out that there is another very common use for the notation f^{-1} other than the inverse of the function, or in the case of real-valued functions the reciprocal. In this case, if $f : X \rightarrow Y$, then f^{-1} is actually a function from $\mathbb{P}(Y)$ into $\mathbb{P}(X)$ which is defined as follows:

DEFINITION**Pullback**

Given a function $f : X \rightarrow Y$, we define the *pullback* of f to be the function $f^{-1} : \mathbb{P}(Y) \rightarrow \mathbb{P}(X)$ by

$$f^{-1}(B) = A = \{x \in X \mid f(x) \in B\}$$

for each $B \in \mathbb{P}(Y)$.

In other words, the pullback of a subset B of Y tells us all the elements in X that are mapped into B by f .

REMARK

Pullbacks play significant roles in many parts of mathematics. While we will not encounter pullbacks often in this course we can deduce some things immediately from the f^{-1} .

- 1) A function $f : X \rightarrow Y$ is one-to-one if and only for any $y \in Y$ the pullback $f^{-1}(\{y\})$ of the singleton $\{y\}$ is either a singleton $\{x\}$ or it is empty.
- 2) A function $f : X \rightarrow Y$ is onto if and only for any $y \in Y$ the pullback $f^{-1}(\{y\})$ of the singleton $\{y\}$ is non-empty.
- 3) A function $f : X \rightarrow Y$ is one-to-one and onto if and only for any $y \in Y$ the pullback $f^{-1}(\{y\})$ of the singleton $\{y\}$ is a singleton .
- 4) If $f : X \rightarrow Y$ is one-to-one and onto with inverse $g : Y \rightarrow X$, then $g(y) = x$ if and only if $f^{-1}(\{y\}) = \{x\}$



2.8 Boolean Algebra and Sets: Enrichment

Boolean algebra allows us to translate rules of logic and set theory into simple arithmetic. In this section we will briefly investigate how this can help us provide proofs of some of the most basic results in set theory. To begin with we will need the following definition:

DEFINITION**Characteristic Function**

Given $A \subseteq X$ define

$$\chi_A(x) := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

$\chi_A(x)$ is called the *characteristic function of A*.

NOTE

The characteristic function of a set completely determines the set A in the sense that we can see if an element x in our set A simply by looking at the value the characteristic function returns when evaluated at x . In particular, we know that $\chi_A = \chi_B$ if and only if $A = B$. This tells us that the map $\Gamma : \mathbb{P}(X) \rightarrow \{f : X \rightarrow \{0, 1\}\}$ given by

$$\Gamma(A) = \chi_A$$

is always one-to-one.

Conversely, given a function $f : X \rightarrow \{0, 1\}$, if we let

$$A = \{x \in X \mid f(x) = 1\},$$

then it turns out that

$$f = \chi_A.$$

This means that Γ is actually a one-to-one correspondence between the power set $\mathbb{P}(X)$ of X and the set of functions on X taking only values 0 or 1. In more sophisticated language

$$\mathbb{P}(X) = \prod_{x \in X} \{0, 1\}.$$

One immediate consequence of this observation is that if $X = \{x_1, x_2, \dots, x_n\}$ is a finite set with n elements, then

$$|\mathbb{P}(X)| = 2^n.$$

That is a set with n elements has exactly 2^n subsets.

This identification between subsets of X and their characteristic functions can be quite useful in terms of proving set theoretic identities. In this respect the following rules are useful: ◀

THEOREM 6

Boolean Arithmetic

Let $A, B \subseteq X$. Then

- 1) $\chi_{A \cap B} = \chi_A \cdot \chi_B$
- 2) $\chi_{A \cup B} = \chi_A + \chi_B - \chi_A \cdot \chi_B$
- 3) $\chi_{A^c} = 1 - \chi_A$

PROOF

We will prove the first statement only. The second and third are left as exercises.

To prove i), there are 4 cases to consider:

Cases:

- 1) $x \in A$ and $x \in B$,
- 2) $x \in A$ and $x \notin B$,
- 3) $x \notin A$ and $x \in B$,
- 4) $x \notin A$ and $x \notin B$.

We can use the following table to check these four cases and to verify our claim.

Case	χ_A	χ_B	$\chi_{A \cap B}$	$\chi_A \cdot \chi_B$
1	1	1	1	$1 \cdot 1 = 1$
2	1	0	0	$1 \cdot 0 = 0$
3	0	1	0	$0 \cdot 1 = 0$
4	0	0	0	$0 \cdot 0 = 0$

Since $\chi_{A \cap B}$ and $\chi_A \cdot \chi_B$ return the same value in all four cases we have established the validity of 1). ■

REMARK

You may have noticed a similarity between the table above and the truth tables we considered in the previous chapter. This is no coincidence. If we assign a true statement the value 1 and a false statement the value 0, and if we interpret \cap as *and*; \cup as *or*; and complementation as negation, we can translate the results in the previous theorem to give us the truth tables for *and*, *or*, and negation. ◀

NOTE

A key property of characteristic functions is that $\chi_A \cdot \chi_A = \chi_A$. ◀

We can use Boolean algebra to verify more complex statements as well.

EXAMPLE 11 Prove that

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C).$$

PROOF

$$\begin{aligned}
 (A \cup B) \cap C &\cong (\chi_A + \chi_B - \chi_A \cdot \chi_B) \cdot \chi_C \\
 &= (\chi_A \cdot \chi_C) + (\chi_B \cdot \chi_C) - \chi_A \cdot \chi_B \cdot \chi_C \\
 &= (\chi_A \cdot \chi_C) + (\chi_B \cdot \chi_C) - \chi_A \cdot \chi_B \cdot \chi_C^2 \\
 &= (\chi_A \cdot \chi_C) + (\chi_B \cdot \chi_C) - (\chi_A \cdot \chi_C) \cdot (\chi_B \cdot \chi_C) \\
 &\cong (A \cap C) \cup (B \cap C)
 \end{aligned}$$

2.9 Principle of Mathematical Induction

2.10 Mathematical Induction

As we have already seen, mathematics is built on *axioms*. Axioms are mathematical statements that we accept as being true without need for proof. The following axiom introduces one of the fundamental properties of the set \mathbb{N} of natural numbers. It will lead to an important method of proof called “*proof by induction*”.

AXIOM 7

(Principle of Mathematical Induction)

If a set $S \subseteq \mathbb{N}$ is such that the following two conditions hold,

1. $1 \in S$.
2. For each $k \in \mathbb{N}$, if $k \in S$, then $k + 1 \in S$.

then $S = \mathbb{N}$.

We can give an informal argument that illustrates why the Principle of Mathematical Induction is a reasonable axiom. Suppose $S \subseteq \mathbb{N}$ satisfies the two conditions given in the axiom. Then we know that $1 \in S$. Because $1 \in S$, and because S satisfies the second condition, 2 must also be in S . Because $1 + 1 = 2 \in S$, and because S satisfies the second condition, $2 + 1 = 3$ must also be in S . Because $3 \in S$, 4 must be in S . Because $4 \in S$, 5 must be in S , and so on. If we are given any natural number n , we can use this argument to show that $n \in S$; therefore, every natural number n can be shown to be in S by simply repeating this process enough times and so $S = \mathbb{N}$. It is important to note however, that this is **not** a proof of the validity of the Principle of Mathematical Induction. Indeed, since we have stated the principle as an axiom, no proof is needed.

Mathematical Induction: As mentioned before, this axiom leads to a method of proof called “*proof by induction*”. We begin by asserting a statement $P(n)$ for each natural number n . The goal is then to show that $P(n)$ is true for each $n \in \mathbb{N}$. For example, we could let $P(n)$ be the statement that

$$\sum_{j=1}^n j = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

If we let $S = \{n \in \mathbb{N} : P(n) \text{ is true}\}$, then $S \subseteq \mathbb{N}$. Moreover, to prove $P(n)$ for each n , it suffices to show that $S = \mathbb{N}$. To do this we first show that $1 \in S$ (*i.e.*, that $P(1)$ is true). We must then show that for any $k \in \mathbb{N}$, $k \in S$ implies that $k + 1 \in S$, that is that the truth of $P(k)$ forces $P(k + 1)$ to also be true, for any natural number k . This would show that S satisfies the two hypotheses of the Principle of Mathematical Induction and as such, we can conclude that $S = \mathbb{N}$ and hence that $P(n)$ is true for all $n \in \mathbb{N}$.

In practise, we divide inductive proofs into three distinct steps as outlined above.

Step 1: The first step is to clearly identify the statements $P(n)$ that we are trying to prove.

Step 2: The next step is usually (but not always) the easiest part of the argument: Show that $P(1)$ is true. This is called the initial case. This step is very important. There are many instances where induction has led to false proofs because this initial case was not properly established.

Step 3: In the third step, we are allowed to assume that we know that the truth of $P(k)$ for some k . This assumption is referred to as *the induction hypothesis*. We must then use the truth of $P(k)$ as a tool to show that given the induction hypothesis, it must also be the case that $P(k + 1)$ is also true, that is that $P(k)$ implies $P(k + 1)$, for any $k \in \mathbb{N}$. In most inductive proofs we will see in this course, this step is the most involved.

Once we have successfully concluded each of the steps above, we may appeal to the Principle of Mathematical Induction to conclude that $P(n)$ is in fact true for each $n \in \mathbb{N}$.

We will illustrate “proof by induction” with the following example:

EXAMPLE 12 Prove by induction that

$$\sum_{j=1}^n j = 1 + 2 + \cdots + n = \frac{n(n + 1)}{2}.$$

Step 1:

Let $P(n)$ be the statement that

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

Step 2:

We must show that $P(1)$ is true. However, $P(1)$ is the statement that

$$\sum_{j=1}^1 j = \frac{1(1 + 1)}{2},$$

which is obviously true because $\sum_{j=1}^1 j = 1 = \frac{2}{2} = \frac{1(1+1)}{2}$.

Step 3:

Next, we will verify that the truth of $P(k)$ implies the truth of $P(k + 1)$, for any $k \in \mathbb{N}$. Assume that $P(k)$ is true. This means that for some fixed k

$$\sum_{j=1}^k j = \frac{k(k + 1)}{2}.$$

With this assumption in hand, we need to show that $P(k + 1)$ is true, or that

$$\sum_{j=1}^{k+1} j = \frac{(k+1)((k+1)+1)}{2}.$$

We know that

$$\sum_{j=1}^{k+1} j = \left(\sum_{j=1}^k j \right) + (k+1).$$

Separating out the last term $k + 1$ allows us to make use of the induction hypothesis. In fact, by the assumed truth of $P(k)$, the following is true:

$$\begin{aligned} \sum_{j=1}^{k+1} j &= \left(\sum_{j=1}^k j \right) + (k+1) \\ &= \frac{k(k+1)}{2} + (k+1) \\ &= \left(\frac{k}{2} + \frac{2}{2} \right) (k+1) \\ &= \frac{(k+2)(k+1)}{2} \\ &= \frac{(k+1)((k+1)+1)}{2}. \end{aligned}$$

We have succeeded in showing that $\sum_{j=1}^{k+1} j = \frac{(k+1)((k+1)+1)}{2}$, which is precisely the assertion $P(k + 1)$.

Finally, by the Principle of Mathematical Induction, we have that $P(n)$ is true for all $n \in \mathbb{N}$. 

Strong Induction and the Well Ordering Principle. There are a number of equivalent formulations of the Principle of Mathematical Induction. Below we present the first of these which we called the Principle of Strong Mathematical Induction. It is so named because it seems to be formally stronger than the Principle of Mathematical Induction in the sense that you can assume the presence of $1, 2, 3, \dots, k$ in S to show that $k + 1$ is in S . We could prove the Principle of Strong Mathematical Induction as a consequence of the Principle of Mathematical Induction. In fact, it is an interesting exercise to show that the two statements are in fact logically equivalent in the sense that they each imply the other.

After presenting the Principle of Strong Induction, we will show that the Principle of Strong Induction can be used to establish another important property of the natural numbers known as the Well Ordering Principle.

THEOREM 8**Principle of Strong Induction**

If a set $S \subseteq \mathbb{N}$ is such that the following two conditions hold, then $S = \mathbb{N}$.

1. $1 \in S$.
2. For each $k \in \mathbb{N}$, if $1, 2, \dots, k \in S$, then $k + 1 \in S$.

PROOF

To prove the Principle of Strong Induction we will use Induction. The key to doing so is to choose our statement $P(n)$ cleverly!

Let $P(n)$ be the statement that $\{1, 2, 3, \dots, n\} \subseteq S$. We note that if $P(n)$ is true, then it also follows that $n \in S$. Hence if we can show that $P(n)$ is true for all $n \in \mathbb{N}$ we have indeed shown that $S = \mathbb{N}$.

Now we know by our first assumption that $1 \in S$. This means that $\{1\} \subseteq S$, and hence that $P(1)$ is true.

Next lets assume that $P(k)$ is true. This means that $\{1, 2, \dots, k\} \subseteq S$. We can now appeal to our second assumption about S to conclude that $k + 1 \in S$. But from this we can deduce that $\{1, 2, \dots, k, k + 1\} \subseteq S$ and therefore that $P(k + 1)$ is true.

Since we have shown that $P(1)$ is true and that if $P(k)$ is true so is $P(k + 1)$, we can apply Induction to conclude that $P(n)$ is true for each $n \in \mathbb{N}$. Finally, this gives us that $S = \mathbb{N}$.

■

In a similar manner as for the Principle of Mathematical Induction, the Principle of Strong Induction leads to a method of proof called *proof by strong induction* which we illustrate with the following example:

EXAMPLE 13

Let $f : \mathbb{N} \rightarrow \mathbb{R}$ be a function defined recursively by $f(1) = 3$, $f(2) = \frac{3}{2}$, and $f(n) = \frac{f(n-1) + f(n-2)}{2}$ for all $n \geq 3$. We will prove by strong induction that $f(n) = 2 + \left(\frac{-1}{2}\right)^{n-1}$.

Let $P(n)$ be the statement that

$$f(n) = 2 + \left(\frac{-1}{2}\right)^{n-1}.$$

Then for $n = 1$, $2 + \left(\frac{-1}{2}\right)^{n-1} = 2 + \left(\frac{-1}{2}\right)^0 = 3 = f(1)$. This shows that $P(1)$ is true. For $n = 2$, $2 + \left(\frac{-1}{2}\right)^{n-1} = 2 + \left(\frac{-1}{2}\right)^1 = \frac{3}{2} = f(2)$. This shows that $P(2)$ is true.

Now assume that $P(n)$ is true for all n satisfying $1 \leq n \leq k$, for some $k \in \mathbb{N}$. If $k = 1$, the $P(k + 1) = P(2)$ is true. If $k > 1$, then $k + 1 \geq 3$, and hence we have that

$$\begin{aligned}
 f(k + 1) &= \frac{f((k + 1) - 1) + f((k + 1) - 2)}{2} = \frac{f(k) + f(k - 1)}{2} \\
 &= \frac{2 + \left(\frac{-1}{2}\right)^{k-1} + 2 + \left(\frac{-1}{2}\right)^{k-2}}{2} = \frac{4 + \left(\frac{-1}{2}\right)^{k-2} \left[\frac{-1}{2} + 1\right]}{2} \\
 &= \frac{4 + \left(\frac{-1}{2}\right)^{k-2} \left[\frac{1}{2}\right]}{2} = 2 + \left[\frac{1}{2}\right] \left[\frac{1}{2}\right] \left(\frac{-1}{2}\right)^{k-2} \\
 &= 2 + \left[\frac{-1}{2}\right] \left[\frac{-1}{2}\right] \left(\frac{-1}{2}\right)^{k-2} = 2 + \left(\frac{-1}{2}\right)^k \\
 &= 2 + \left(\frac{-1}{2}\right)^{(k+1)-1}.
 \end{aligned}$$

This shows that the statement $P(k + 1)$ is true. By the Principle of Strong Induction, $P(n)$ is true for all $n \in \mathbb{N}$. 

REMARK

As we mentioned previously, on the surface it appears that Strong Induction is a more powerful technique when compared with Induction because in trying to prove that the statement $P(k + 1)$ is true you are able to not only assume that $P(k)$ is true but also that all of $P(1), P(2), \dots, P(k - 1)$ are also true. This gives you more tools to use in establishing $P(k + 1)$. However, as we will soon see this is not actually the case as any result which can be proved by Strong Induction could also be obtained by Induction. Toward this goal, we will now give another fundamental property of \mathbb{N} which can be deduced from the Principle of Strong Induction. 

Well-Ordering Principle. One way in which the natural numbers differ from the real numbers is that \mathbb{N} has a least element namely 1. In fact, since \mathbb{N} is only “infinite in one direction,” we would speculate that every nonempty subset of the natural numbers has a least element. In fact, we can use Strong Induction to show that this is true. This property is known as the Well-Ordering Principle and is stated below.

THEOREM 9

Well Ordering Principle

Every non-empty subset S of \mathbb{N} has a least element.

PROOF

Suppose there exists a subset S of \mathbb{N} such that S has no least element. We must show that $S = \emptyset$. To do so we let $T = \mathbb{N} \setminus S$. We will use the principle of strong induction on T to show that $T = \mathbb{N}$.

Let $P(n)$ be the statement that

$$n \in T.$$

Since S has no least element it must be the case that $1 \notin S$, and hence that $1 \in T$. This means that $P(1)$ holds.

Next assume that $P(j)$ holds for all $j \in \{1, 2, \dots, k\}$. That is that

$$\{1, 2, \dots, k\} \subseteq T.$$

In this case, if $k + 1 \in S$, then $k + 1$ would have to be the least element of S since all smaller natural numbers are assumed to be in S . However, this is impossible since S has no least element. We get that $k + 1 \in T$ and hence that $P(k + 1)$ holds.

We have just shown that $P(1), P(2), \dots, P(k)$ all being true implies that $P(k+1)$. Since we have also shown that $P(1)$ holds, the Principle of Strong Induction shows us that $P(n)$ holds for all $n \in \mathbb{N}$ and therefore that $T = \mathbb{N}$. This then gives us that $S = \emptyset$. This completes the proof since it means that a non-empty subset S of \mathbb{N} must have a least element. ■

NOTE

There is a strong link between the Well Ordering Principle and the Axiom of Choice. To see why this is the case we note that we can define a function $f : \mathbb{P}(\mathbb{N}) \setminus \{\emptyset\} \rightarrow \mathbb{N}$ by

$$f(A) = \text{the least element in } A.$$

Then f is a choice function for \mathbb{N} .

It is known that there is no way to construct an explicit *order* on \mathbb{R} , necessarily different from the usual way we usually order real numbers, so that with this new order every non-empty subset of \mathbb{R} has a least element. In fact the existence of such an abstract ordering of \mathbb{R} with this property not only requires the Axiom of Choice but it cannot be done without it. ◀

REMARK

Recall that a natural number n is prime if $n \geq 2$ and if n has no factors other than 1 and itself. One of the many standard applications of the Well Ordering Principle is the Fundamental Theorem of Arithmetic which we state below. The proof will be left as an exercise. ◀

THEOREM 10**The Fundamental Theorem of Arithmetic**

Let $n \in \mathbb{N}$ with $n \geq 2$. Then there exist primes $p_1 < p_2 < \cdots < p_k$ and natural numbers m_1, m_2, \dots, m_k such that

$$n = p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k}.$$

Moreover, if there exist primes $q_1 < q_2 < \cdots < q_l$ and natural numbers r_1, r_2, \dots, r_k such that

$$n = q_1^{r_1} q_2^{r_2} \cdots q_l^{r_l},$$

then $k = l$, $p_i = q_i$ for each $1 \leq i \leq k$ and $m_i = r_i$ for each $1 \leq i \leq k$.

PROOF

This proof that there is a decomposition into primes is left as an exercise. We will not address the uniqueness at this time. ■

Equivalence of three principles. It turns out that the three principles we have outlined in this section, the Principle of Mathematical Induction, the Principle of Strong Induction and the Well Ordering Principle are logically equivalent—*i.e.*, they imply each other.

THEOREM 11

The following are equivalent:

- i) Principle of Mathematical Induction;
- ii) Principle of Strong Induction;
- iii) Well-ordering Principle.

PROOF

We have already shown that $i) \Rightarrow ii)$ and that $ii) \Rightarrow iii)$.

To prove that $iii) \Rightarrow i)$ assume that N satisfies the Well Ordering Principle and that S is a subset of N satisfying

1. $1 \in S$.
2. For each $k \in \mathbb{N}$, if $k \in S$, then $k + 1 \in S$.

We let $T = \mathbb{N} \setminus S$. To prove the theorem we must show that $T = \emptyset$. Assume to the contrary that T is nonempty. Then by the Well Ordering Principle, T has a least element which we denote by k_0 . Since we have assumed already that $1 \in S$, we know immediately that $1 \neq k_0$. It follows that $k_0 - 1 \in \mathbb{N}$. However, since $k_0 - 1 < k_0$ and k_0 is the least element of T it must be that $k_0 - 1 \in S$. But if $k_0 - 1 \in S$, then by property 2 above we must also have that $k_0 = (k_0 - 1) + 1 \in S$. This is a contradiction since we know that $k_0 \in T$. Therefore, since the assumption that $T \neq \emptyset$ leads to a contradiction it must be the case that $T = \emptyset$. Finally, if $T = \emptyset$ and $T = \mathbb{N} \setminus S$, then it must be the case that $S = \mathbb{N}$. ■

The proof above uses the technique known as “proof by contradiction” that we mentioned briefly in the previous Chapter. As previously indicated, we first assume that the conclusion of the theorem is false and then proceed to derive a contradiction, thereby showing that conclusion cannot be false.

2.10.1 The Tower of Hanoi: Enrichment

We end this section with a famous problem that can be easily solved using Induction.



Tower of Hanoi

EXERCISE 1

(Tower of Hanoi)

You are given three pegs.

On one of the pegs is a tower made up of n rings placed on top of one another so that as you move down the tower each successive ring has a larger diameter than the previous ring.

The object of this puzzle is to reconstruct the tower on one of the other pegs by moving one ring at a time, from one peg to another, in such a manner that you never have a ring above any smaller ring on any of the three pegs.

Prove that for any $n \in \mathbb{N}$, if you begin with n rings, then the puzzle can be completed in $2^n - 1$ moves. Moreover, prove that for each n , this is the minimum number of moves necessary to complete the task.

Note: The key to this question is to formulate an appropriate description of the statement $P(n)$.

Chapter 3

Real Numbers

It is often the case that in order to solve complex mathematical problems we must first replace the problem with a simpler version for which we have appropriate tools to find a solution. In doing so our solution to the simplified problem may not work for the original question, but it may be close enough to provide us with useful information. Alternatively, we may be able to design an algorithm that will generate successive approximate solutions to the full problem in such a manner that if we apply the process enough times, the result will eventually be as close as we would like to the actual solution.

For example, there is no algebraic method to solve the equation

$$e^x = x + 2.$$

However, we can graphically show that there are two distinct solutions for this equation and that the two solutions are close to -2 and 1, respectively. One process we could use to solve this equation is a type of binary search algorithm that is based on the fact that the function $f(x) = e^x - (x + 2)$ is continuous. We could also use an alternate process which relies on the very useful fact that if a function is differentiable at a point $x = a$, then its tangent line is a very good approximation to the graph of a function near $x = a$.

In fact, *approximation* will be a theme throughout this course. But for any process that involves approximation, it is highly desirable to be able to control how far your approximation is from the true object. That is, to control the *error* in the process. To do so we need a means of measuring *distance*. In this course, we will do this by using the geometric interpretation of *absolute value*.

3.1 Absolute Values

One of the main reasons that mathematics is so useful is that the real world can be described using concepts such as geometry. Geometry means “earth measurement” and so at its heart is the notion of distance. We may view the number line as a ruler that extends infinitely in both directions. The point 0 is chosen as a reference point. We can think of the distance between 0 and 1 as our fixed unit of measure. It then

follows that the point π is located $3.141592\dots$ to the *right* of the reference point 0. The point $-\sqrt{2}$ is located $1.41421\dots$ units to the *left* of 0. Consequently, we can think of the non-zero real numbers as being quantities that are made up of two parts. First there is a *sign*, either positive or negative, depending on the point's *orientation* with respect to zero, and a *magnitude* that represents the *distance* that the point is from 0. This magnitude is also a real number, but it is always either positive or 0. This magnitude is called the *absolute value* of x and is denoted by $|x|$.

It is common to think of the absolute value as being a mechanism that simply drops negative signs. In fact, this is what it does provided that we are careful with how we represent a number. For example, we all know that $|5| = 5$ and that $|-3| = 3$. However, what if x was some unknown quantity, would $|-x| = x$? It is easy to see that this is not true if the mystery number x actually turned out to be -3 . Since the absolute value plays an important role for us in this course, we will take time to give it a careful definition that will remove any ambiguity.

DEFINITION

Absolute Value

For each $x \in \mathbb{R}$, define the *absolute value of x* by

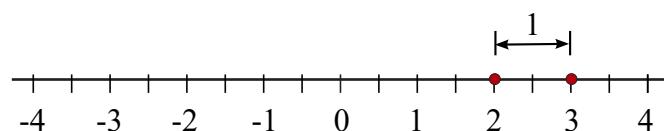
$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$$

REMARK

If we use this definition, then it is easy to see that

$$|x| = |-x|.$$

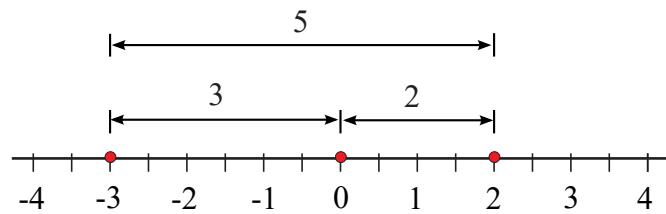
So far we have only considered the distance between a fixed number and 0. However, it makes perfect sense to consider the distance between any two arbitrary points. For example, we would assume that the distance between the points 2 and 3 should be 1.



$$\text{distance} = 1 = |2 - 3| = |3 - 2|$$

The distance between -3 and 2 is slightly more complicated. To get from -3 to 2 we can first travel from -3 to 0 , a distance of 3 units, and then travel from 0 to 2 , an additional 2 units. This makes for a total of 5 units. We observe that

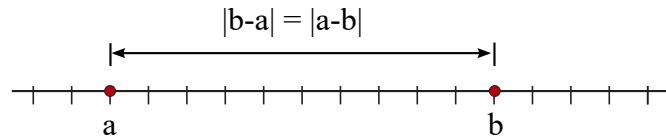
$$|3 - 2| = |2 - 3| = 1 \text{ (from the first example) and that } |-3 - 2| = |2 - (-3)| = 5.$$



$$\text{distance} = 5 = |2 - (-3)| = |(-3) - 2|$$

REMARK

These examples illustrate an important use of the absolute value, namely that given any two points a, b on the number line, the distance from a to b is given by $|b - a|$. Note that the distance is also $|a - b|$ since $|b - a| = |-(b - a)| = |a - b|$. Geometrically, this last statement corresponds to the fact that the distance from a to b should be equal to the distance from b to a . 

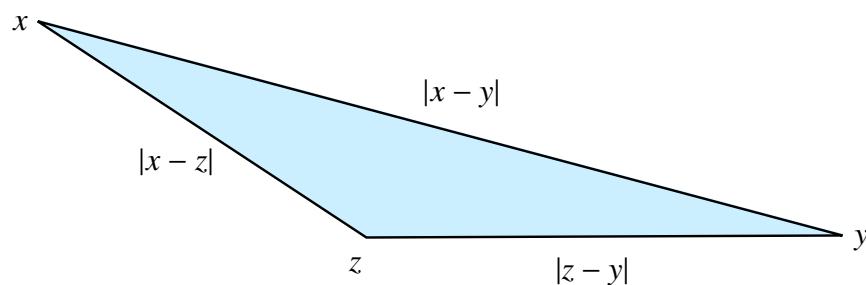


$$\text{distance} = |b - a| = |a - b|$$

3.1.1 Inequalities Involving Absolute Values

One of the fundamental concepts in Calculus is that of approximation. Consequently, we are often faced with the question of “*when is an approximation close enough to the exact value of the quantity?*” Mathematically, this becomes an inequality involving absolute values. These inequalities can look formidable. However, if you keep distances and geometry in mind, it will help you significantly.

One of the most fundamental inequalities in all of mathematics is the *Triangle Inequality*. In the two-dimensional world, this inequality reduces to the familiar statement that *the sum of the length of any two sides of a triangle exceeds the length of the third*. This means that if you have three points x, y and z , it is always at least as long to travel in a straight line from x to z and then from z to y as it is to go from x to y directly.



If we use this last statement as our guide, and recognize that the exact same principle applies on the number line, we are led to the following very important theorem:

THEOREM 1

Triangle Inequality

Let x, y and z be any real numbers. Then

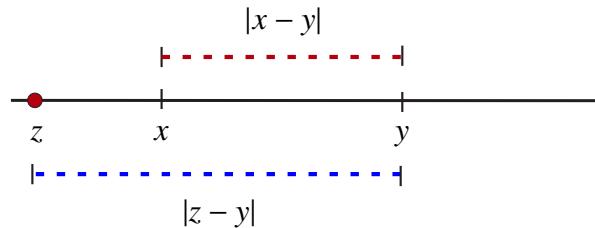
$$|x - y| \leq |x - z| + |z - y|$$

This theorem essentially says: The distance from x to y is less than or equal to the sum of the distance from x to z and the distance from z to y .

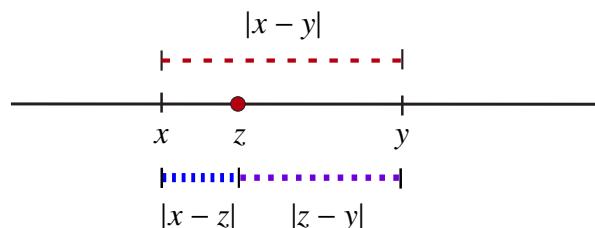
While we will try to avoid putting undue emphasis on formal proofs, it is often enlightening to convince ourselves of the truth of a mathematical assertion. To do this for the triangle inequality, we first note that since $|x - y| = |y - x|$, we could always rename the points so that $x \leq y$. With this assumption, we have three cases to consider:

PROOF

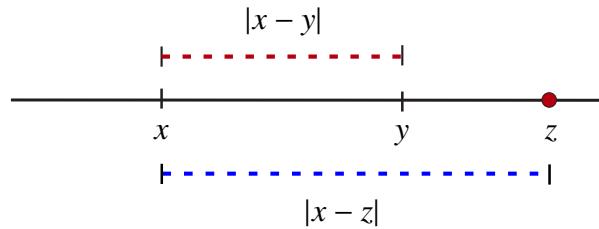
Case 1 : $z < x$. In this case, it is clear from the picture below that the distance from z to y exceeds the distance from x to y . That is, $|x - y| < |z - y|$ and therefore that $|x - y| \leq |x - z| + |z - y|$.



Case 2 : $x \leq z \leq y$. In this case, we can see that the distance from x to y is the sum of the distances from x to z and z to y . That is $|x - y| = |x - z| + |z - y|$.



Case 3 : $y < z$. In this case, it is clear from the picture below that the distance from x to z exceeds the distance from x to y . That is, $|x - z| < |x - y|$. Therefore, $|x - y| \leq |x - z| + |z - y|$.



Since we have exhausted all possible cases, we have verified the inequality. ■

There is one important variant of the Triangle Inequality that we will require. It can be derived as follows:

Let x and y be any real numbers. Then using the Triangle Inequality and the fact that $|-y| = |y|$ we have

$$\begin{aligned} |x + y| &= |(x - 0) - (0 - y)| \\ &\leq |x - 0| + |0 - y| \\ &= |x| + |-y| \\ &= |x| + |y| \end{aligned}$$

We will call the inequality we have just derived the *Triangle Inequality II*.

THEOREM 2

Triangle Inequality II

Let $x, y \in \mathbb{R}$. Then

$$|x + y| \leq |x| + |y|.$$

We will also need to be able to deal with inequalities of the form

$$|x - a| < \delta$$

where a is some fixed real number and $\delta > 0$. We can interpret this inequality geometrically. It asks for all real numbers x whose distance away from a is less than δ . This makes the inequality rather easy to solve. If we look at the number line we see that if we proceed δ units to the right from a , we reach $a + \delta$. For any x beyond this point, the distance from x to a exceeds δ and consequently our inequality is not valid. We can also move to the left δ units from a to get to $a - \delta$. We can again see that if $x < a - \delta$, then the distance from x to a again exceeds δ . We also note that since this is a *strict* inequality, $a - \delta$ and $a + \delta$ are excluded as solutions. Therefore, the solution set to the inequality

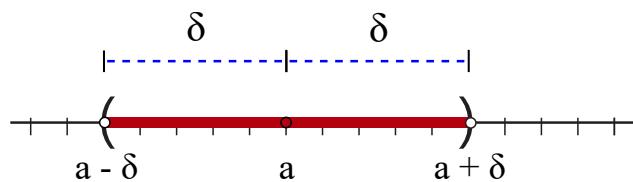
$$|x - a| < \delta$$

is the set

$$a - \delta < x < a + \delta$$

or

$$(a - \delta, a + \delta).$$



If we were to change the inequality to

$$|x - a| \leq \delta$$

then the endpoints $a - \delta$ and $a + \delta$ now satisfy the new inequality so that the new solution set would be $[a - \delta, a + \delta]$.

There is one more inequality that we will come across later. It is the inequality

$$0 < |x - a| < \delta.$$

In this case, the distance from x to a must be less than δ so $x \in (a - \delta, a + \delta)$ but it must also be greater than 0. This last condition means that $x \neq a$. So our solution is all points in $(a - \delta, a + \delta)$ except $x = a$. We will denote this set by

$$(a - \delta, a + \delta) \setminus \{a\}.$$

We can summarize our last three inequalities in the chart below:

Inequality	Solution
$ x - a < \delta$	$(a - \delta, a + \delta)$
$ x - a \leq \delta$	$[a - \delta, a + \delta]$
$0 < x - a < \delta$	$(a - \delta, a + \delta) \setminus \{a\}$

EXAMPLE 1 Find the solution to the inequality

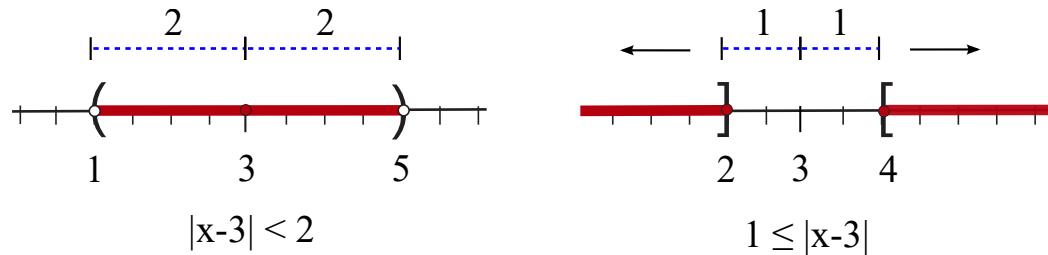
$$1 \leq |x - 3| < 2.$$

SOLUTION The inequality is asking for all points that are at least 1 unit from 3, but less than 2 units from 3. To the right of 3, the values of x that are at least one unit away are $x \geq 3 + 1 = 4$.

However, to also be less than 2 units away from 3, we must have $x < 3 + 2 = 5$.

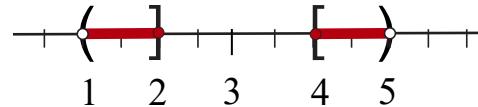
It follows that if $x \geq 3$ and x is a solution to the inequality, then $4 \leq x < 5$ or equivalently, $x \in [4, 5)$.

A similar argument shows that if $x \leq 3$ is a solution to the inequality, then $1 = 3 - 2 < x \leq 3 - 1 = 2$ or equivalently, $x \in (1, 2]$.



This leads us to the solution to the inequality, namely the set

$$(1, 2] \cup [4, 5) = (1, 5) \setminus (2, 4).$$



3.2 Least Upper Bound Property

In this section we will introduce one of the most fundamental defining characteristics of the real line, namely the *Least Upper Bound Property*. The Least Upper Bound Property will play a crucial role in much of the remainder of the course. But before we can state the property we need to introduce some terminology.

DEFINITION

Upper and Lower Bounds

Let $S \subset \mathbb{R}$. We say that α is an *upper bound* of S if

$$x \leq \alpha$$

for every $x \in S$. If S has an upper bound, we say that it is *bounded above*.

We say that β is a *lower bound* of S if

$$\beta \leq x$$

for every $x \in S$. If S has a lower bound, we say that it is *bounded below*.

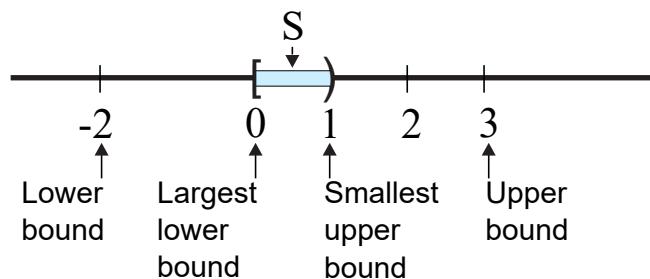
S is *bounded* if it is bounded both above and below. Note that S is bounded if there exists an M such that

$$S \subseteq [-M, M]$$

EXAMPLE 2 Let $S = [0, 1)$. Then $\alpha = 3$ is an upper bound for S . However, 1 is also an upper bound. Moreover, 1 has the property that amongst all of the upper bounds for S , it is the *smallest*.

Similarly, while -2 is a lower bound for our set, 0 has the property that it is the *largest* lower bound of S .

If $M = 2$, then clearly $S \subseteq [-2, 2]$, so S is bounded.



It turns out that the smallest or *least upper bound* for a set will play a key role in our investigation of convergence for monotone sequences.

DEFINITION

Least Upper Bound

Let $S \subseteq \mathbb{R}$. Then α is called the *least upper bound* of S if

1. α is an upper bound of S .
2. α is the smallest such upper bound. That is, if $x \leq \gamma$ for every $x \in S$, then $\alpha \leq \gamma$.

We write

$$\alpha = \text{lub}(S).$$

Note: The least upper bound is often called the *supremum* of S and is denoted by

$$\sup(S).$$

DEFINITION**Greatest Lower Bound**

Let $S \subseteq \mathbb{R}$. Then β is called the *greatest lower bound* of S if

1. β is a lower bound of S .
2. β is the largest such lower bound. That is, if $\gamma \leq x$ for every $x \in S$, then $\gamma \leq \beta$.

We write

$$\beta = \text{glb}(S).$$

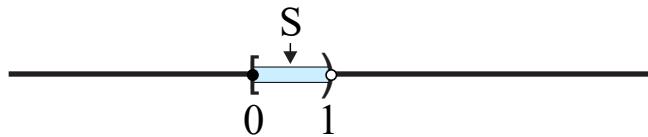
Note: The greatest lower bound is often called the *infimum* of S and is denoted by

$$\inf(S).$$

The next example shows that if the least upper bound or the greatest lower bound exist, they need *not* be in the set.

EXAMPLE 3

If $S = [0, 1)$, then $\text{lub}(S) = 1$ and $\text{glb}(S) = 0$.

**DEFINITION****Maxima and Minima**

1. If S contains $\alpha = \text{lub}(S)$, then α is called the *maximum* of S and is denoted by $\text{max}(S)$.
2. If S contains $\beta = \text{glb}(S)$, then β is called the *minimum* of S and is denoted by $\text{min}(S)$.

EXAMPLE 4

If S is a finite set with n elements

$$S = \{a_1 < a_2 < \cdots < a_n\},$$

then

- $a_n = \text{lub}(S) = \max(S)$, and
- $a_1 = \text{glb}(S) = \min(S)$.



The next property of the real numbers may seem obvious and perhaps quite unimportant. However, it is actually the theoretical foundation for all of the most fundamental results of calculus.

AXIOM 3

Least Upper Bound Property

Let $S \subset \mathbb{R}$ be nonempty and bounded above. Then S has a least upper bound.

Observe that in our statement of the Least Upper Bound Property we specifically excluded the empty set \emptyset . This leads to the following problem:

Problem: Is \emptyset bounded above or below?



We begin by asking whether there is an upper bound for \emptyset . Specifically, is 6 greater than or equal to every element in the empty set and hence an upper bound for \emptyset ? This might seem like a bit of an absurd question after all how can some number like 6 be greater than or equal to every element in a set that itself contains no element. However, the key is to ask, what would need to happen for it to be true that 6 was not an upper bound for \emptyset ? With this in mind, we have the following proposition:

PROPOSITION 4

Let $\alpha \in \mathbb{R}$. Then $\alpha \in \mathbb{R}$ is both an upper bound and a lower bound for \emptyset . In particular, \emptyset is bounded.

PROOF

If α were not an upper bound for \emptyset , then by definition there would exist some element $x \in \emptyset$ satisfying $x > \alpha$, a contradiction (nothing is in \emptyset). This shows that α must be an upper bound for \emptyset . Similarly, α must be a lower bound for \emptyset , as claimed.



The previous proposition yields a very unusual fact about the empty set. For this set and this set alone it is possible to have an upper bound α and a lower bound β such that $\alpha < \beta$. In fact, $\alpha = -1$ and $\beta = 5$ are such a pair. This proof again uses the general principle that a statement about elements in a set will be automatically true for \emptyset , because if it were not true, then there would exist elements in \emptyset to contradict the statement, which is impossible. The truth of a statement about \emptyset derived in this way is called a *vacuous truth* and the statement is said to be true *vacuously*.

The above proposition also shows that we must take care to apply the least upper bound property only to *nonempty* sets, because clearly, \emptyset fails to have a least upper bound even though it is bounded above!

The dual to the Least Upper Bound Property is the *Greatest Lower Bound Property*, stated below. Note that, as we show below, the Greatest Lower Bound Property can be derived as a theorem, using the Least Upper Bound Property as the key tool in the proof. However, had we assumed the Greatest Lower Bound Property as an axiom, we would be able to derive the Least Upper Bound Property as a theorem. That is the the two properties are equivalent.

THEOREM 5**Greatest Lower Bound Property**

A nonempty subset $S \subset \mathbb{R}$ that is bounded below always has a greatest lower bound.

PROOF

Let β be a lower bound for S . Therefore, if $x \in S$, we have that $\beta \leq x$. It follows that $-x \leq -\beta$ and hence that $-\beta$ is an upper bound for the set $T = -S = \{-s : s \in S\}$. Since S is nonempty, so is T . By the Least Upper Bound Property, T has a least upper bound α . Now if $x \in S$, then $-x \leq \alpha$, so that $-\alpha \leq x$, for every $x \in T$. This proves that $-\alpha$ is a lower bound for S . We also know that since $-\beta$ is an upper bound for T , we have $\alpha \leq -\beta$ and hence that $\beta \leq -\alpha$. However, since β was an arbitrary lower bound of S , $-\alpha$ is in fact the greatest lower bound for S . ■

3.3 Archimedean Property

We begin by posing the following question:

Question: Is \mathbb{N} bounded above? ◀

The fact that \mathbb{N} is not bounded above may seem trivial and obvious. Our first thought might be that \mathbb{N} has no largest element and as such would be unbounded. However, the set $(0, 1)$ does not have a largest element and yet it is definitely bounded above by 12. It is at least theoretically possible that there is some mysterious real number α such that $n < \alpha$ for all $n \in \mathbb{N}$ even if \mathbb{N} itself has no largest element. Yet is certainly seem evident that this cannot be.

One of the goals of the course is learn to proceed in mathematics rigorously and carefully, verifying every new statement with definitions, axioms, and previously proven results. This is important because there are many glaringly examples of statements that appear that they “must be true” that are either extremely difficult to prove or worse yet, for which there are even simple counterexamples showing them to be false. The good news is that in this case, our instincts are correct and we have the tools to verify this.

THEOREM 6**Archimedean Property**

The set $\mathbb{N} \subset \mathbb{R}$ has no upper bound in \mathbb{R} . That is, \mathbb{N} is not bounded above.

PROOF

Suppose on the contrary that \mathbb{N} is bounded above by some $M \in \mathbb{R}$. Then since $1 \in \mathbb{N}$, $\mathbb{N} \neq \emptyset$. By the least upper bound property, \mathbb{N} has a least upper bound $\alpha \in \mathbb{R}$. Then since α is the least upper bound, $\alpha - \frac{1}{2}$ is not an upper bound for \mathbb{N} . This shows that there exists an element $n \in \mathbb{N}$ with

$$\alpha - \frac{1}{2} < n \leq \alpha.$$

Since $n \in \mathbb{N}$, it follows that $n + 1 \in \mathbb{N}$, making

$$\begin{aligned} \alpha - \frac{1}{2} + 1 &< n + 1 \\ \Rightarrow \alpha &< \alpha + \frac{1}{2} < n + 1, \end{aligned}$$

which is a contradiction since no element of \mathbb{N} can be greater than α , an upper bound for \mathbb{N} . Consequently, the statement that \mathbb{N} is bounded above cannot be true and the theorem follows. ■

REMARK

The previous proof uses an observation that will be used repeatedly throughout this course, namely that if α is the least upper bound of a set S and if $\epsilon > 0$ is any positive number no matter how small, then $\alpha - \epsilon < \alpha$. Therefore, $\alpha - \epsilon$ is not an upper bound for S . It follows that there must be some $s \in S$ that makes $\alpha - \epsilon$ fail to be an upper bound; that is, this s satisfies $s > \alpha - \epsilon$. But since $s \in S$, it must be less than or equal to the upper bound α . Combining these two inequalities yields

$$\alpha - \epsilon < s \leq \alpha.$$



The following easy but important corollary is also known as the Archimedean Property. We will see later that this corollary gives us a formal proof that the sequence $\{\frac{1}{n}\}$ converges to 0.

COROLLARY 7

Archimedean Property II

For every $\epsilon > 0$, then there exists an $n \in \mathbb{N}$ satisfying $0 < \frac{1}{n} < \epsilon$.

PROOF

For any $\epsilon > 0$, we can find an $n \in \mathbb{N}$ with $n > \frac{1}{\epsilon} > 0$ by Archimedean Property I—for if such an n cannot be found, then $\frac{1}{\epsilon}$ would be an upper bound for \mathbb{N} , contradicting the theorem. Taking reciprocals yields $0 < \frac{1}{n} < \epsilon$. ■

REMARK

It might have been tempting to try and avoid using the Least Upper Bound Property to prove the Archimedean Property by instead arguing as follows: Assume that $\alpha \in \mathbb{R}$ is an upper bound for \mathbb{N} . Then we certainly know that $\alpha > 0$ and we also know that α has a decimal expansion

$$\alpha = m.a_1a_2a_3a_4\cdots$$

where $m \in \mathbb{N}$ and $0 \leq a_i \leq 9$. (The number m , which is the greatest integer less than α , is called the floor of α and it is denoted by $\lfloor \alpha \rfloor$.) From here we, we get that $m \leq \alpha < m + 1$. However $m + 1 \in \mathbb{N}$ contradicting the fact that α was an upper bound for \mathbb{N} .

It might well appear that we have succeeded in avoiding the use of the Least Upper Bound Property in the above “proof” until we ask ourselves: **how do we know that every real number has a decimal expansion?** It turns out that this familiar “fact” can actually be shown to be equivalent to the Least Upper Bound Property. Indeed, in an exercise later in the course, you will show one direction of this equivalence by showing that under the assumption that the real numbers satisfy the Least Upper Bound Property, then every real number does indeed have a decimal expansion.

What we want to do now is to show that every real number α can be approximated as closely as we would like by both rational and irrational numbers. This leads us to define the notion of *density*.

DEFINITION**Density in \mathbb{R}**

We say that a set $S \subseteq \mathbb{R}$ is dense in \mathbb{R} if for every $\alpha \in \mathbb{R}$ and every $\varepsilon > 0$, there exists an $s \in S$ such that $s \in (\alpha - \varepsilon, \alpha + \varepsilon)$.

An application of the previous corollary is the following corollary, which states something about the *density* of rational numbers and irrational numbers as subsets of \mathbb{R} .

COROLLARY 8**Density of \mathbb{Q} and $\mathbb{R} \setminus \mathbb{Q}$**

For every $a, b \in \mathbb{R}$ with $a < b$, there exists $r \in \mathbb{Q}$ and $s \notin \mathbb{Q}$ satisfying $a < r < b$ and $a < s < b$.

PROOF

This proof is left as a homework exercise. In this proof, you may assume that $\sqrt{2}$ is irrational. ■

Chapter 4

Sequences and Convergence

4.1 Sequences and Their Limits

In many applications of mathematics, *continuous* processes are modeled by lists of data points (*discrete* data). This leads naturally to the concept of a *sequence*. In this section, we will introduce sequences and define what is meant by the *limit of a sequence*.

4.1.1 Introduction to Sequences

A *sequence* is simply an *ordered list*. We encounter sequences everyday. For example a phone number can be thought of as a sequence. Indeed the phone number 519-555-1234 is read as a sequence of one digit *terms* 5-1-9-5-5-5-1-2-3-4 rather than as the integer 5,195,551,234. With phone numbers, we are very aware that the *order of the terms is important*.

A phone number is an example of a *finite* sequence in the sense that this ordered list contains only finitely many terms. For the remainder of this course we will only consider *infinite* sequences where the terms or elements of the sequence will be real numbers.

We write

$$\{a_1, a_2, a_3, \dots, a_n, \dots\} \quad \text{or} \quad \{a_n\}_{n=1}^{\infty} \quad \text{or simply} \quad \{a_n\}$$

to denote a sequence.¹ The real numbers a_n are called the *terms* or *elements* of the sequence. The natural number n is called the *index* of the term a_n .

There are many different ways that a sequence can be specified. The easiest way is to simply list the elements in a manner that gives the value of a_n as an explicitly defined function of n . For example, in the sequence

$$\{1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{n}, \dots\}$$

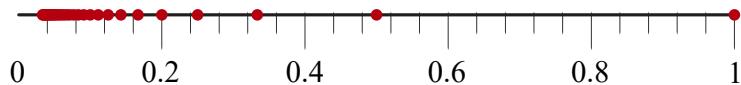
¹Some references use round brackets instead of curly brackets to denote sequences. Therefore you may encounter sequences written as $(a_1, a_2, a_3, \dots, a_n, \dots)$ or $(a_n)_{n=1}^{\infty}$ or (a_n) .

the n -th term in the sequence is $\frac{1}{n}$. The sequence can also be identified by writing $\{\frac{1}{n}\}$ or by giving the explicit function

$$a_n = \frac{1}{n}$$

that specifies the value of the terms.

We can also visualize a sequence by plotting its terms on the number line. For example, the sequence $\{1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{n}, \dots\}$ has a plot that looks like:



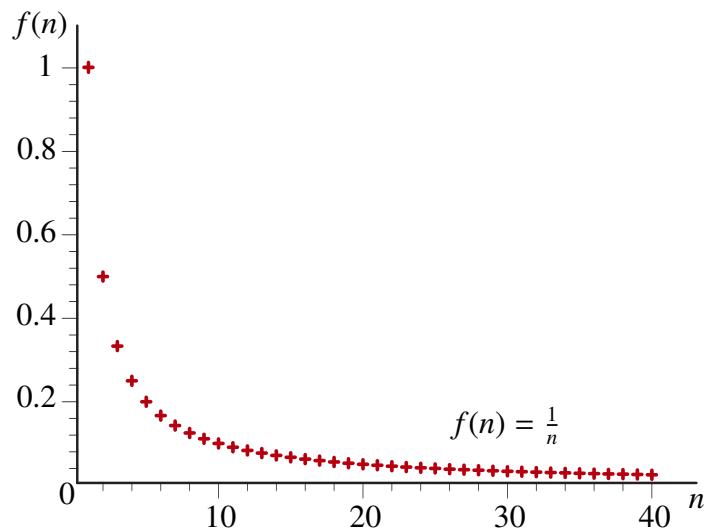
We can also consider a sequence as a function from the natural numbers \mathbb{N} with values in \mathbb{R} . Consequently, we can use function notation to identify a sequence as well. In this case, we will equate a_n with the value of the function f at the natural number n and write $f(n) = a_n$ as a means of designating the sequence. In particular, the sequence

$$\{1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{n}, \dots\}$$

can be identified with the function

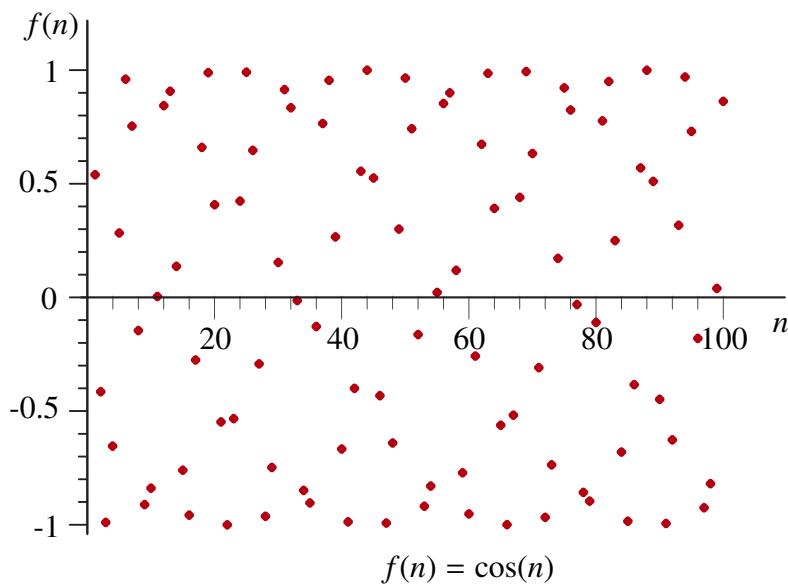
$$f(n) = \frac{1}{n}.$$

In this form, the graph of the function can give us some very useful visual information about the nature of the sequence.

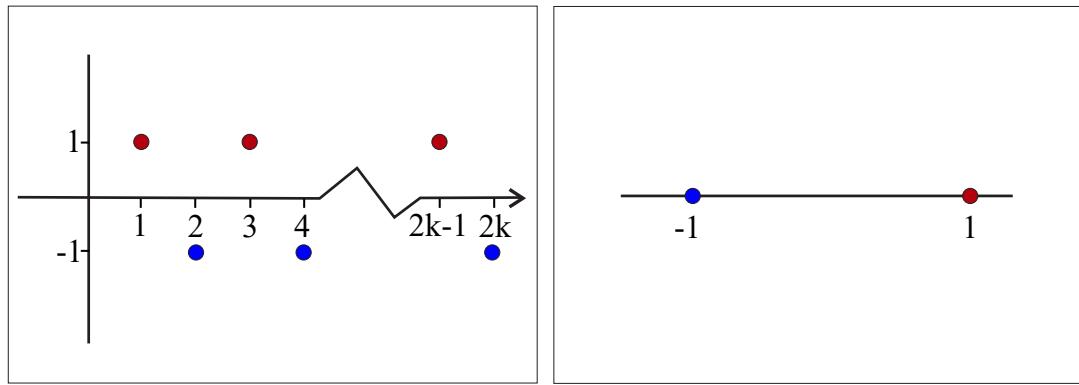


For $f(n) = \frac{1}{n}$, we see that as the index increases the terms decrease, and as the index gets large the values of the terms “approach” 0.

In contrast to the rather nice behavior of the sequence $\{\frac{1}{n}\}$, the sequence $\{\cos(n)\}$ has rather chaotic behavior as its graph shows:



Moreover, to see why the two-dimensional plot is often better as a way of representing a sequence than the simple one-dimensional, view let's look at these two plots for the sequence $\{(-1)^{n-1}\} = \{1, -1, 1, -1, \dots\}$.



The two-dimensional plot on the left representing the graph of the function $f(n) = (-1)^{n-1}$ clearly shows the sequence oscillating from 1 to -1 as the index increases. However, the one-dimensional plot actually looks very static. Moreover, the right-hand plot would essentially look identical to the plot of the sequence $\{(-1)^n\}$. This example, and the two plots above, also illustrates two important principles; firstly that the order of the terms does matter, and secondly, that a sequence always has infinitely many terms even if those terms may take on only finitely many different values.

4.1.2 Recursively Defined Sequences

In the last section we saw that to define a sequence we could either provide an explicit formula for the n -th term a_n , or we could present an ordered list from which this explicit formula may be derived. Another important method for specifying a sequence is to use *recursion*. When we define a sequence recursively we typically do not know

an explicit formula for the term a_n , but rather we are able to deduce its value from one or more previous terms in the sequence. To do so we generally need to know the first term of the sequence, or even the first few terms depending on the complexity of the recursive formula. For example, consider the sequence defined by

$$a_1 = 1 \text{ and } a_{n+1} = \frac{1}{1 + a_n}.$$

Suppose that we want to determine the value of a_5 . In the previous example with $a_n = \frac{1}{n}$, we could see by inspection that the value of a_5 was simply $\frac{1}{5}$. However, notice that in this new example you cannot determine a_5 directly from the given expression. Indeed, since $5 = 4 + 1$, the formula tells us that to find a_5 we must first calculate a_4 . But to do this we must know a_3 , which in turn depends upon the value of a_2 . At this point, we are explicitly told that $a_1 = 1$. Hence we can conclude that

$$a_2 = \frac{1}{1 + a_1} = \frac{1}{1 + 1} = \frac{1}{2}.$$

Then we get that

$$a_3 = \frac{1}{1 + a_2} = \frac{1}{1 + \frac{1}{2}} = \frac{2}{3}.$$

We can calculate $a_4 = \frac{3}{5}$ and then finally $a_5 = \frac{5}{8}$. If you are now asked to find a_{2573} , your first instinct would likely be to give up! Luckily though, it is often very easy to program a computer to evaluate the terms of a recursively defined sequence. In fact, a modern computer can calculate a_{2573} almost instantaneously.

Problem: Can you write a loop that calculates the terms of the recursively defined sequence above and stops at a_{2573} ? ◀

Despite the difficulty that we may have in identifying the terms of recursively defined sequences, such sequences are very important for practical applications. In this course, for example, we will use recursively defined sequences to find approximate solutions to equations that cannot be solved explicitly (see *Newton's Method*).

We end this section with three recursively defined sequences. The first, the Fibonacci sequence is one of the most famous sequences in mathematics. The last of the three sequences is of historical importance in that it arises from an algorithm that the Babylonians used to calculate square roots.

EXAMPLE 1 The Fibonacci Sequence

In his 1202 manuscript *Liber Abaci (Book of Abacus)* the Italian mathematician Leonardo Fibonacci posed the following problem: Assume that a newly born pair of breeding rabbits can mate at the age of 1 month and that each female will then produce exactly one more breeding pair one month later and that the pair will then mate again immediately. Assume also that rabbits never die. If you start with a single newly born pair of breeding rabbits, how many pairs will you have at the beginning of the n th month?

Let F_n denote the number of rabbit pairs at the beginning of month n . Then by assumption we start with one pair and as such $F_1 = 1$. Since this pair must wait one month before breeding, we also begin month 2 with the only this pair. That is $F_2 = 1$. At the end of the 2nd month the initial breeding pair has produced another breeding pair so we begin the third month with $F_3 = 2$.

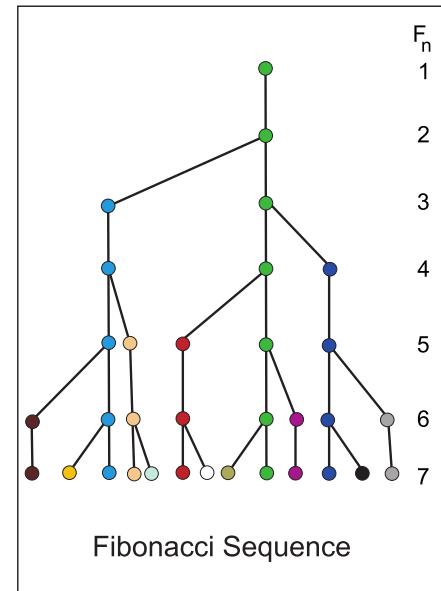
At this point, our initial pair will produce offspring each subsequent month. However, our newly born pair must wait one month to breed. As such at the beginning of month 4 we have our initial pair, their first pair of offspring and the new offspring the initial pair produces in month 3. That tells us that $F_4 = 3$.

To find F_n for a typical n , we make the following observation. Each pair that was alive at the beginning of month $n - 1$ will still be alive at the beginning of month n . Moreover, we will have one more additional pair for each pair alive at the beginning of month $n - 2$ as each such pair will be of breeding age at the beginning of month $n - 1$. This gives us a recursive formula for F_n of the form

$$F_n = F_{n-1} + F_{n-2}.$$

From this we can deduce that $F_5 = 3 + 2 = 5$, $F_6 = 5 + 3 = 8$, $F_7 = 8 + 5 = 13$, and so on.

This sequence, which was known in India centuries before Fibonacci, has many remarkable properties though curiously enough, we still do not know if there are infinitely many prime numbers in the sequence. ◀



Problem: Find the first 6 primes in the Fibonacci sequence. ◀

EXAMPLE 2

In this example we will look at the sequence defined by the recursive relation $a_1 = 1$ and $a_{n+1} = \sqrt{3 + 2a_n}$.

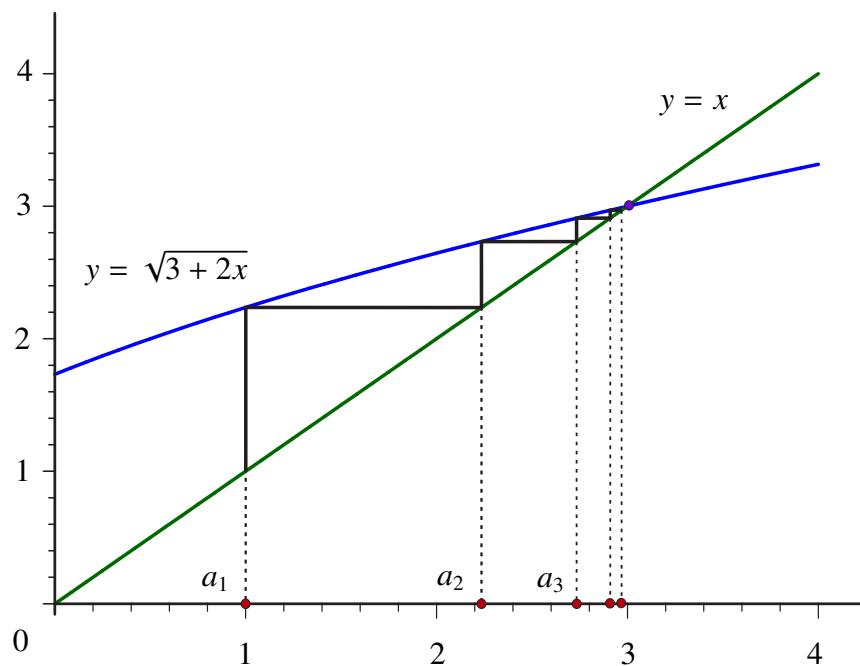
The table shows the first 10 terms in the sequence. You will notice an interesting pattern here. The terms are increasing, all of them are less than 3, but the terms do seem to be getting very close to 3. In fact, if we were to continue on we would find that $a_{30} = 2.999999999999996$. This is indeed no accident. It is a consequence of the nature of the function $f(x) = \sqrt{3 + 2x}$ which we use to generate the sequence. In particular, we note the fact that 3 is the unique solution to the equation.

$$L = \sqrt{3 + 2L}.$$

This is equivalent to saying that 3 is the x -coordinate of the unique intersection point of the graphs of the functions $y = x$ and $y = \sqrt{3 + 2x}$.

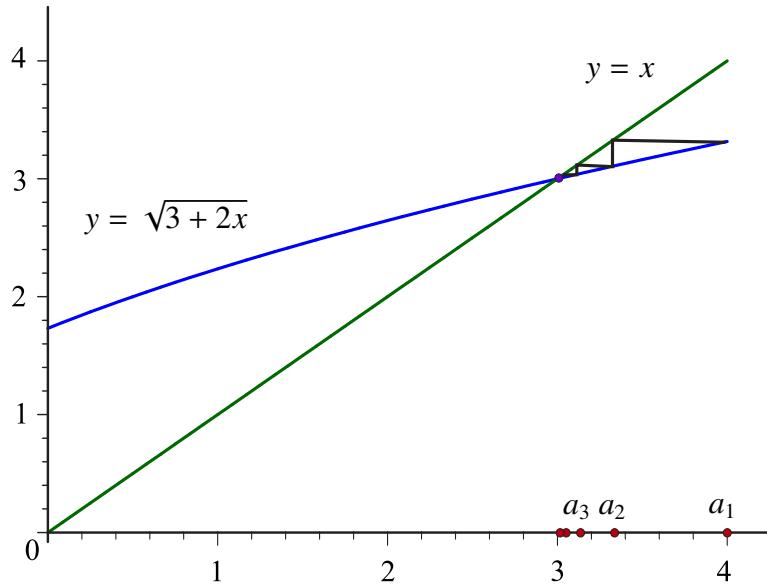
n	a_n
1	1
2	2.236067977
3	2.733520798
4	2.909818138
5	2.969787244
6	2.989912121
7	2.996635487
8	2.998878286
9	2.999626072
10	2.999875355

A graphical illustration of the comments above is given by the following diagram:



In this example, we might ask what happens if we were to change the value of a_1 ? For example, what if we let $a_1 = 4$? How would the sequence behave? The chart on the right and the graph below shows that the sequence still behaves in a similar manner despite the new initial value. Of course, the sequence now decreases rather than increases, but it still rapidly approaches 3.

n	a_n	a_n
1	4	1756
2	3.31662479	59.28743543
3	3.103747667	11.02609953
4	3.034385495	5.005217184
5	3.011440019	3.606997972
6	3.003810919	3.195934283
7	3.001270038	3.064615566
8	3.000423316	3.021461754
9	3.000141102	3.007145409
10	3.000047034	3.002380858



Note that if we start with $a_1 = 1756$, then the sequence again decreases and still rapidly moves towards 3. Moreover, if we let $a_1 = 23675382$, then $a_{10} = 3.009560453$. In fact it is actually possible to show that if a_1 is any real number that is greater than $\frac{-3}{2}$ (so that $\sqrt{3 + 2x} > 0$), then the sequence will increase if $a_1 < 3$, decrease if $a_1 > 3$, but in all cases the sequence will rapidly approach 3.

Problem: What happens if $a_1 = 3$? ◀

In the next example we will present an algorithm for calculating square roots that has its historical origins going back to the Babylonians. The algorithm itself was first presented explicitly by the Greek mathematician Heron of Alexandria (also known as Hero of Alexandria).

EXAMPLE 3 Heron's Algorithm for Finding Square Roots

Consider the following recursively defined sequence:

$$a_1 = 4 \quad \text{and} \quad a_{n+1} = \frac{1}{2}(a_n + \frac{17}{a_n}).$$

Let's see what this sequence looks like by calculating the first 10 terms.

You will notice that up to ten decimal places the sequence actually stabilizes from $n = 4$ onwards. In fact the terms do change as n increases but the difference between successive terms is so small as to be almost undetectable very quickly. In particular, like our previous example, the terms of this sequence seem to rapidly approach a certain limiting value α which we would guess to be very close to 4.1231056256. So it is now worth asking, what is the significance of this value α ?

n	a_n
1	4
2	4.125
3	4.1231060606
4	4.1231056256
5	4.1231056256
6	4.1231056256
7	4.1231056256
8	4.1231056256
9	4.1231056256
10	4.1231056256

In fact it turns out that we will later be able to show that $\alpha = \sqrt{17}$. In particular, we can show that $a_4 - \sqrt{17}$ is roughly 2.31×10^{-14} which means that a_4 is a remarkably accurate approximation to $\sqrt{17}$. Even a_3 is very close to $\sqrt{17}$ with $a_3 - \sqrt{17} \approx 4.35 \times 10^{-7}$. It is also worth noting that despite the fact that in the table above we represented the terms in the sequence with decimal expansions, the calculations certainly produce rational values for each a_n . In particular, $a_1 = 4$, $a_2 = \frac{33}{8}$, $a_3 = \frac{2177}{528}$ and $a_4 = \frac{9478657}{2298912}$. So this algorithm not only gives an approximate value for $\sqrt{17}$, but it also generates very accurate rational approximations to this irrational number.

More generally, if $\alpha > 0$ is any positive real number and we have a_1 chosen so that a_1 is a real number that is reasonably close to $\sqrt{\alpha}$, then the recursive sequence with initial term a_1 and

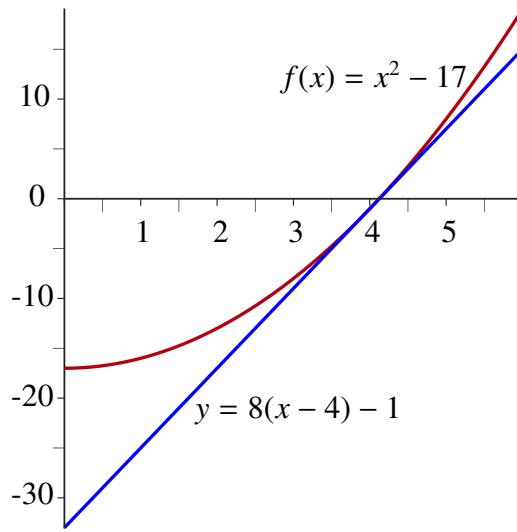
$$a_{n+1} = \frac{1}{2}(a_n + \frac{\alpha}{a_n})$$

will generate a sequence which will very rapidly approach the value $\sqrt{\alpha}$. If both α and a_1 are rational numbers, then so is a_n for each n .

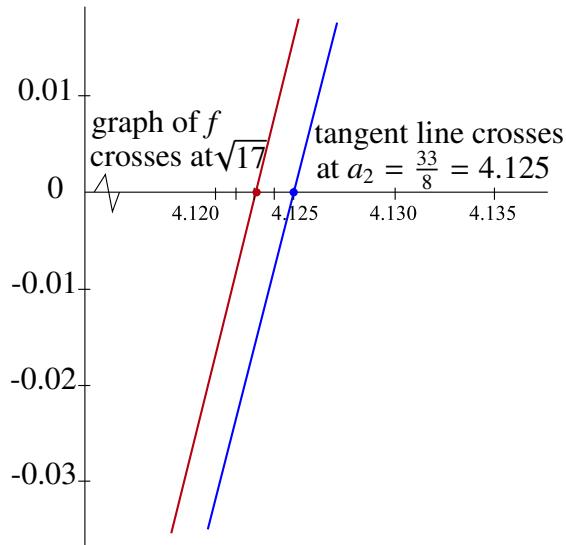
Problem: Let $\alpha = 198$ and $a_1 = 14$. Find the rational expressions for a_2 and a_3 using the algorithm above. Using a calculator calculate the decimal expression for a_3 and for $\sqrt{198}$ to 8 decimal places? Are they the same?

Before we leave this example, let's take a closer look at this algorithm. For example, suppose we want to find $\sqrt{17}$. This is the same as finding the unique positive solution to the equation $x^2 - 17 = 0$, or equivalently, finding the unique positive x -intercept

for the graph of the function $f(x) = x^2 - 17$. It turns out that if $a_1 = 4$, then a_2 is the x -coordinate of the intersection of the x -axis and the tangent line to the graph of f through $(4, -1)$.



To illustrate, we note that the tangent line above has equation $y = 8(x - 4) - 1$ and if we solve the equation $0 = 8(x - 4) - 1$, the solution is exactly $a_2 = \frac{33}{8}$ as claimed. From here you will note that the graph above shows that the tangent line provides a very close approximation to the function $f(x) = x^2 - 17$ near $x = 4$ and it crosses the x -axis very close to the place where the graph of f crosses the x -axis. The latter happens precisely at $x = \sqrt{17}$.



The geometric process we have outlined above is a special case of an algorithm called *Newton's Method* which we will look at in detail later in the course.

REMARK

Historical Note: The algorithm above that generates a sequence converging to \sqrt{a} is often called *Heron's algorithm*, in honor of the first century Greek mathematician Heron who is attributed with the first explicit description of the method. The process is also called the *Babylonian Square Root Method* since this method was essentially used by Babylonian mathematicians to calculate $\sqrt{2}$. 

4.1.3 Subsequences and Tails

Consider any sequence $\{a_n\}$. We can build many new sequences by extracting infinitely many terms of $\{a_n\}$ in such a manner that we preserve the order in which the extracted terms appeared in the original sequence. We call this extracted sequence a *subsequence* of our original sequence $\{a_n\}$.

For example, given the sequence

$$\{a_n\} = \left\{1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{n}, \dots\right\}$$

we might extract every second term beginning with the first. This gives us

$$\left\{1, \frac{1}{3}, \frac{1}{5}, \dots\right\}.$$

In this case, our new sequence has first term 1, second term $\frac{1}{3}$, third term $\frac{1}{5}$, and for each $k \in \mathbb{N}$, the k -th term is $\frac{1}{2k-1}$ so that we denote this new sequence by $\{a_{2k-1}\}$ or $\{\frac{1}{2k-1}\}$.

A more formal definition of a subsequence is:

DEFINITION**Subsequence**

Let $\{a_n\}$ be a sequence. Let $\{n_1, n_2, n_3, \dots, n_k, \dots\}$ be a sequence of natural numbers such that $n_1 < n_2 < n_3 < \dots < n_k < \dots$. A *subsequence* of $\{a_n\}$ is a sequence of the form

$$\{a_{n_k}\} = \{a_{n_1}, a_{n_2}, a_{n_3}, \dots, a_{n_k}, \dots\}.$$

In this course, particularly when we talk of limits of sequences, we will most often be interested in the terms of the sequence with indexes that are at least as large as some fixed k .

To end this section we introduce one more piece of terminology with respect to sequences. Given a sequence $\{a_n\}$ and a natural number k we consider all of the terms of the sequence with index $n \geq k$.

DEFINITION**Tail of a Sequence**

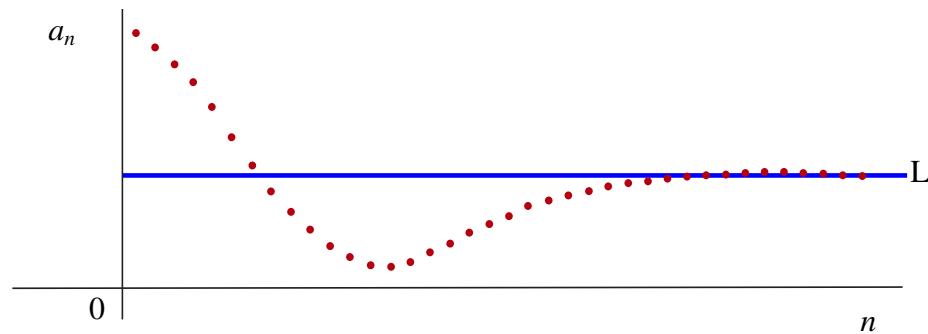
Given a sequence $\{a_n\}$ and $k \in \mathbb{N}$, the subsequence

$$\{a_k, a_{k+1}, a_{k+2}, \dots\}$$

is called the tail of $\{a_n\}$ with cutoff k .

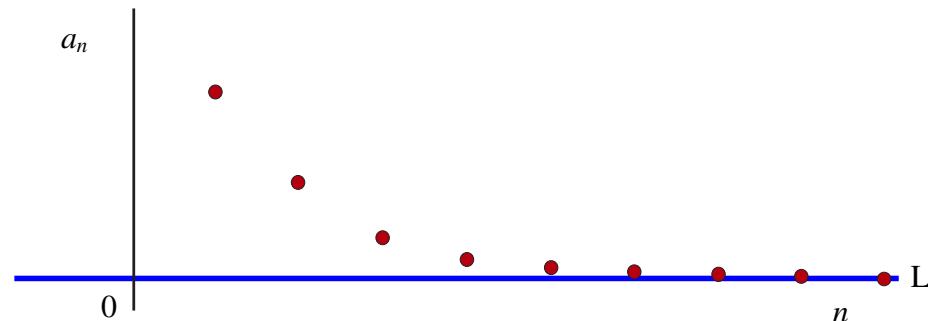
4.1.4 Limits of Sequences

The notion of a *limit* is fundamental to Calculus. Limits occur in various forms. In the previous section we saw examples of sequences with the property that as the index n became large, the terms of the sequences each approached or *converged* to a particular fixed value L .



We call such a sequence *convergent* and call the value L the *limit* of the sequence $\{a_n\}$. This is of course far from a precise mathematical definition of the limit. In the remainder of this section we will try to formulate just such a definition.

First let's consider the sequence $\{1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{n}, \dots\}$. Notice that as n gets larger and larger, the terms get closer and closer to the value 0.

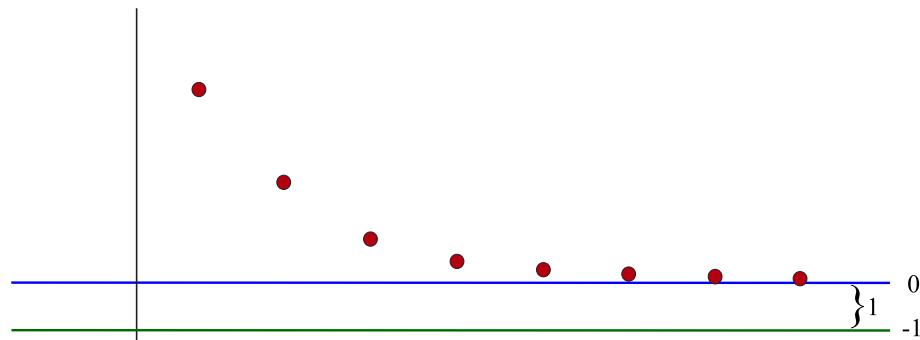


In particular, we can all agree that we should have that 0 is the limit of the sequence $\{\frac{1}{n}\}$. Moreover, this simple example leads us to our first attempt at a *heuristic* or *descriptive* definition of the limit of a sequence.

Heuristic Definition of the Limit of a Sequence I

Given a sequence $\{a_n\}$, we say that L is the limit of the sequence $\{a_n\}$ as n goes to infinity if, as n gets larger and larger, the terms of the sequence get closer and closer to L .

Unfortunately, there are several flaws associated with this definition. For example, if we take this definition exactly as it is written, then not only is 0 a limit of the sequence, but so is -1 because it is also true that as n gets larger and larger the terms in the sequence $\{\frac{1}{n}\}$ get closer and closer to -1 .



So what distinguishes 0 from -1 as a potential limit for $\{\frac{1}{n}\}$? In the case of 0, as n gets larger the terms of the sequence get as close as we would like to 0, whereas the distance from each term $\frac{1}{n}$ to -1 is always greater than 1. As such, the key observation we can take away from this example is that our terms should approximate the limit as accurately as we might wish so long as the index is large enough. Having made this observation, we are now in a position to present a much more precise heuristic definition for the limit.

Heuristic Definition of the Limit of a Sequence II

Given a sequence $\{a_n\}$ we say that L is the limit of the sequence $\{a_n\}$ as n goes to infinity if no matter what positive tolerance $\epsilon > 0$ we are given, we can find a cutoff $N \in \mathbb{N}$ such that the term a_n approximates L with an error less than ϵ provided that $n \geq N$.

This descriptive definition captures all of the properties we would like in a limit. We can also make this into a more formal mathematical definition as follows:

DEFINITION**Formal Definition of the Limit of a Sequence I**

We say that L is the limit of the sequence $\{a_n\}$ as n goes to infinity if for every $\epsilon > 0$ there exists an $N \in \mathbb{N}$ such that if $n \geq N$, then

$$|a_n - L| < \epsilon.$$

If such an L exists, we say that the sequence is convergent and write

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

We may also use the notation $a_n \rightarrow L$ to mean $\{a_n\}$ converges to L .

If no such L exists, then we say that the sequence diverges.

Let's see how the formal definition of a limit can be applied with a specific sequence in mind.

EXAMPLE 4

Use the formal definition of a limit to show that $\lim_{n \rightarrow \infty} \frac{1}{\sqrt{n}} = 0$.

It should be fairly obvious that as n grows so does \sqrt{n} . Moreover, since we can make \sqrt{n} as large as we like, and therefore $\frac{1}{\sqrt{n}}$ as close to 0 as we wish, it should be clear that 0 is in fact the limit. However, we still need to show that the formal definition can be satisfied. To start with, suppose that we are given a tolerance $\epsilon = \frac{1}{100}$. How big does our cutoff N have to be so that if $n \geq N$, then

$$\left| \frac{1}{\sqrt{n}} - 0 \right| = \frac{1}{\sqrt{n}} < \frac{1}{100} \quad (*)$$

whenever $n \geq N$? For $(*)$ to hold we would have

$$100 < \sqrt{n}$$

or equivalently

$$10000 < n.$$

So let's choose *any* cutoff $N_1 \in \mathbb{N}$ so that $N_1 > 10000$. In particular, $N_1 = 10001$ would work. Then if $n \geq N_1$, we would have

$$\left| \frac{1}{\sqrt{n}} - 0 \right| = \frac{1}{\sqrt{n}} \leq \frac{1}{\sqrt{N_1}} < \frac{1}{\sqrt{10000}} = \frac{1}{100}.$$

But of course we are still not done because we need to be able to find the appropriate cut off for *every* positive tolerance $\epsilon > 0$. Suppose instead of a tolerance of $\frac{1}{100}$ the tolerance we were given was $\frac{1}{10^{10}}$. This means we are significantly reducing our permissible error from our previous case and as such fewer terms in the sequence

may approximate our proposed limit within this new tolerance. In fact, if we look at $n = N_1 = 10001$, then

$$\frac{1}{\sqrt{10001}} > \frac{1}{10^{10}}$$

so our old cutoff N_1 is no longer good enough for our purposes. It actually turns out that with this new tolerance we need a cutoff N_2 which is actually greater than 10^{20} . This is a huge number but if we let $N_2 = 10^{20} + 1$, then for any $n \geq N_2$ we do indeed get that

$$|\frac{1}{\sqrt{n}} - 0| = \frac{1}{\sqrt{n}} \leq \frac{1}{\sqrt{N_2}} < \frac{1}{\sqrt{10^{20}}} = \frac{1}{10^{10}}$$

as required.

Now even though we have been able to manage this extremely small tolerance, we are not finished. Someone else could come along and give us an even smaller tolerance than $\frac{1}{10^{10}}$. As such what we really need is to be able to find a cutoff N regardless of what our tolerance $\epsilon > 0$ might be.

The good news is that we can still handle a generic tolerance ϵ . The key is to observe that if we want

$$|\frac{1}{\sqrt{n}} - 0| = \frac{1}{\sqrt{n}} < \epsilon$$

then what we really need is for n to be large enough so that

$$\frac{1}{\epsilon} < \sqrt{n}$$

or equivalently that

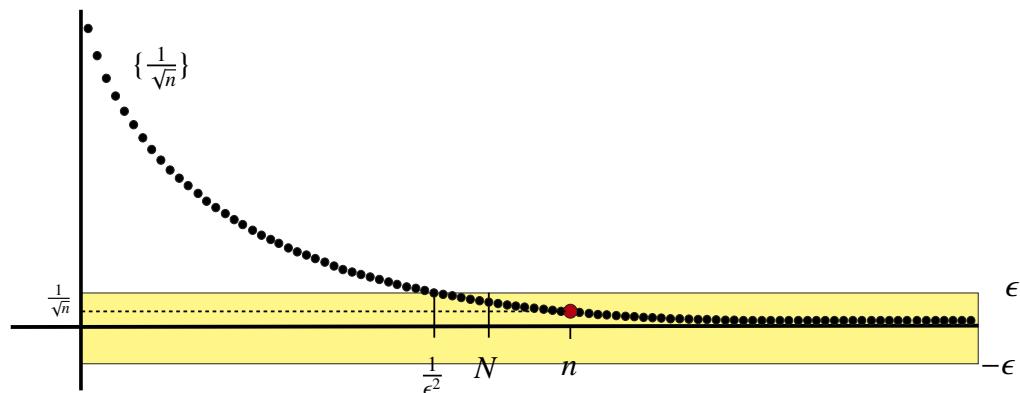
$$\frac{1}{\epsilon^2} < n.$$

But it is a basic property of the Natural numbers (called the *Archimedean Property*) that no matter what $\epsilon > 0$ we are given, we can always find a cutoff $N \in \mathbb{N}$ so that

$$\frac{1}{\epsilon^2} < N.$$

With this cutoff, we do have that if $n \geq N$, then

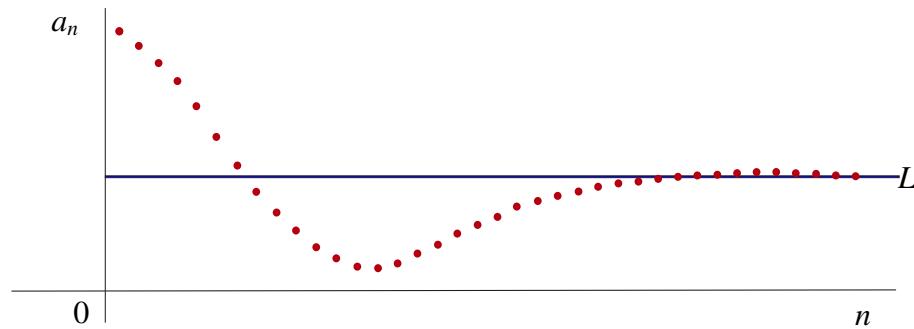
$$|\frac{1}{\sqrt{n}} - 0| = \frac{1}{\sqrt{n}} \leq \frac{1}{\sqrt{N}} < \epsilon.$$



Understanding the formal definition of the limit for a sequence, and later for a function, is perhaps one of the most challenging parts of this course. While we normally will not require that the formal definition of a limit be used to verify that a particular sequence has a particular value L as a limit, it is still worthwhile to develop a strong sense of how this process works. To do so it is perhaps easiest to think of this as a game:

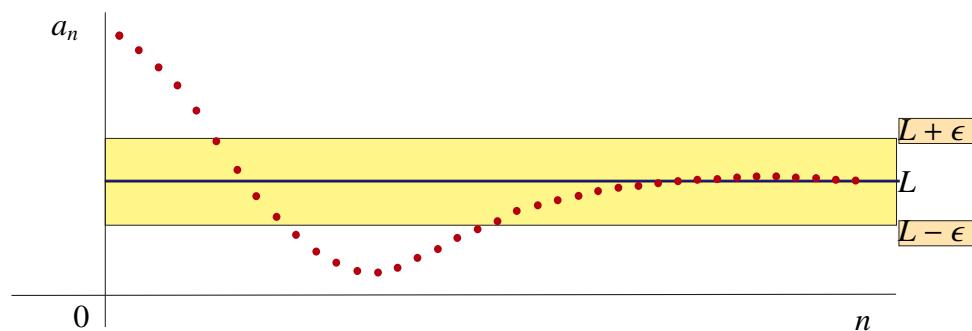
Let's say that the goal of the game is to show that a sequence $\{a_n\}$ converges. Here is how to proceed if you want to win the game:

Step 1: Your first task is to identify a possible candidate L for the limit. There is no absolute method to accomplish this task, and it may be extremely difficult to find the appropriate value. One method to try is to simply look at the terms of the sequence and guess.

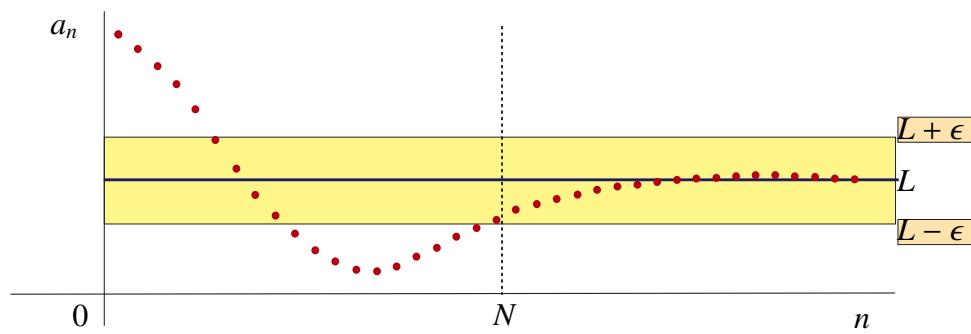


Once you have completed Step 1 you are ready to begin the game in earnest.

Step 2: This is your opponent's move. At this point your opponent presents you with a very small tolerance for you to manage. For example, let $\epsilon = .00001$. This tolerance creates an *error band* or target zone of the form $(L - \epsilon, L + \epsilon)$.

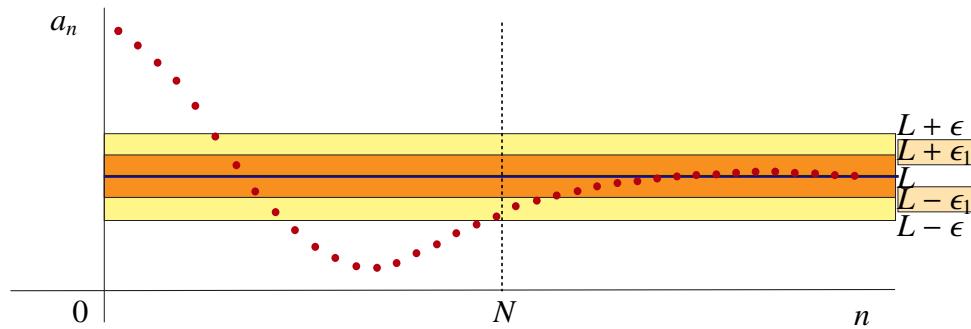


Step 3: To remain in the game with the tolerance in hand you must find a cutoff N so that if the index n is greater than or equal to N , then the term a_n approximates L with an error that is less than your given tolerance.

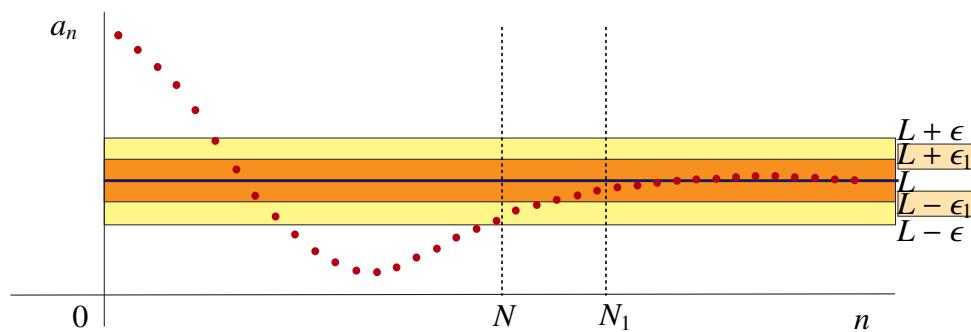


If you cannot find such a cutoff, then you lose and the L you chose is not the limit.

If you do find such a cutoff, then you are still in the game. Unfortunately, if you do find the cutoff for the tolerance $\epsilon = .00001$, you are not done. You simply go back to Step 2 and your opponent will provide you with a new tolerance ϵ_1 that is even smaller than the previous one.



You must now find a new cutoff N_1 , typically bigger than the last, or you lose and the game stops.



Given the rules above, it may appear that you can never win this game no matter how many times you are successful since once you find one cutoff your opponent is free to offer you another tolerance, smaller than the last, forcing you to start your search over again. But in fact you can win the game if you can present your opponent with an algorithm that will generate an appropriate cutoff no matter what tolerance ϵ you are given. We saw this in the example of the sequence $\{\frac{1}{\sqrt{n}}\}$ where given an $\epsilon > 0$,

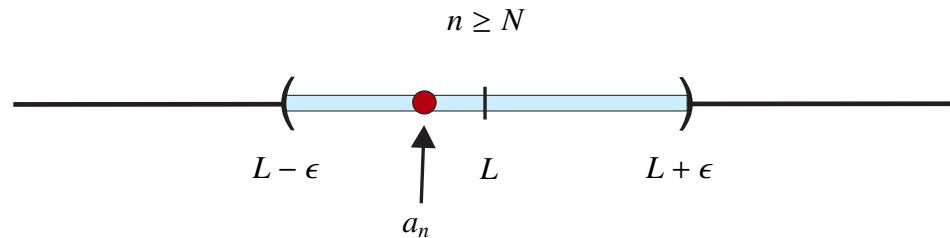
we simply present our opponent with a cutoff N which is greater than $\frac{1}{\epsilon^2}$. So here our algorithm could be

$$N = \lfloor \frac{1}{\epsilon^2} \rfloor + 1$$

where $\text{floor}(x) = \lfloor x \rfloor$ is the largest integer not greater than x .

Important Note: It is generally extremely hard or even impossible to show directly from the formal definition that a particular sequence has a limit. As such in this course, with very few exceptions, we will not try to do so. None the less, understanding the language of limits is something that can be mastered with a little patience and some perseverance. However, there are a few other equivalent formulations of the formal definition of convergence that may be easier to understand. 

Observation: Suppose that we want to show that L is the limit of the sequence $\{a_n\}$. Suppose also that we are given a tolerance $\epsilon > 0$. We must find a cutoff $N \in \mathbb{N}$ so that if $n \geq N$, then $|a_n - L| < \epsilon$. But we have already seen that $|a_n - L| < \epsilon$ is equivalent to having $a_n \in (L - \epsilon, L + \epsilon)$.



You will also recall that the collection of all the terms in $\{a_n\}$ with index $n \geq N$ is a tail of the sequence. So we can reformulate the definition of a limit as follows:

DEFINITION

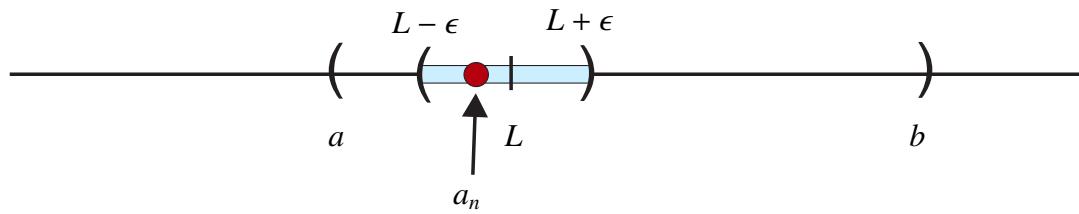
Formal Definition of the Limit of a Sequence II

We say that $L = \lim_{n \rightarrow \infty} a_n$ if for every tolerance $\epsilon > 0$, the interval $(L - \epsilon, L + \epsilon)$ contains a tail of the sequence $\{a_n\}$.

Moreover, if we have any open interval (a, b) containing L , we can find a small enough $\epsilon > 0$ so that

$$L \in (L - \epsilon, L + \epsilon) \subseteq (a, b).$$

But then if L is to be the limit, $(L - \epsilon, L + \epsilon)$ must contain a tail of $\{a_n\}$, and as a result so does (a, b) .



Finally, with the convention that ϵ represents an arbitrary positive tolerance, this observation allows us to present several new ways to view the formal definition of a limit which are conceptually easier to use.

THEOREM 1

The following statements can all be viewed as being equivalent:

1. $\lim_{n \rightarrow \infty} a_n = L$.
2. Every interval $(L - \epsilon, L + \epsilon)$ contains a **tail** of $\{a_n\}$.
3. Every interval $(L - \epsilon, L + \epsilon)$ contains **all but finitely many terms** of $\{a_n\}$.
4. Every interval (a, b) containing L contains a **tail** of $\{a_n\}$.
5. Every interval (a, b) containing L contains **all but finitely many terms** of $\{a_n\}$.

Important Note: Changing finitely many terms in $\{a_n\}$ does not affect convergence. ◀

So far in all of our examples of convergent sequences it appears that the sequence has had a unique limit. This leads us to ask:

Question: Can a sequence have more than one limit? ◀

EXAMPLE 5

Consider the sequence

$$\{1, -1, 1, -1, \dots, (-1)^{n+1}, \dots\} = \{(-1)^{n+1}\}.$$

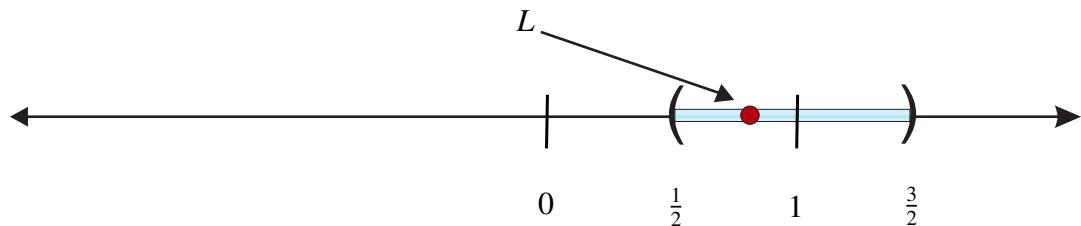
Since the number 1 appears as a term infinitely often, one might be tempted to guess that 1 is a limit of the sequence. Similarly, since -1 also appears infinitely often, we might be tempted to say that -1 is also a limit of this sequence. In fact, we will use what we have just learned to show that neither 1 or -1 are true limits of this sequence. Moreover, we can show that the sequence actually has no limit.

Suppose that we want to show that 1 was a limit of our sequence. Then since $1 \in (0, 2)$, we have to show that this open interval contains a tail of our sequence. But this interval does not contain any of the infinitely many terms with value -1 which means no such tail exists. This shows that 1 was not a limit at all.

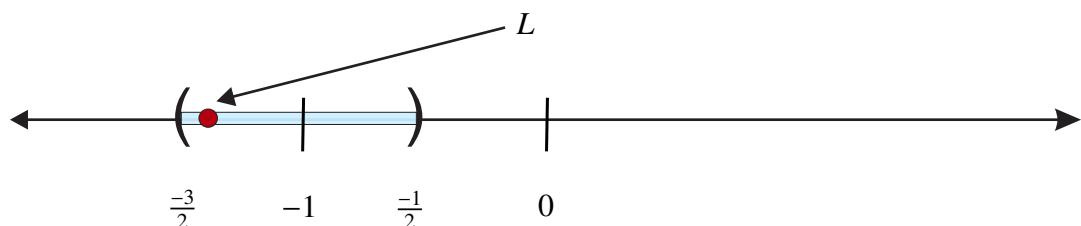
Similarly, suppose we want to show that -1 was a limit of our sequence. Then since $-1 \in (-2, 0)$, we now have to show that this open interval contains a tail of our sequence. But this interval does not contain any of the infinitely many terms with value 1 which means again no such tail exists. This shows that -1 was not a limit either.

Could some other L be a limit? Suppose it was. This would mean *every* open interval around L must contain a tail of our sequence and as such must contain both 1 and -1 since every tail has terms with value 1 and value -1 . But since the distance from -1 to 1 is 2 , it turns out that if we choose an interval around L of length 1 , that interval cannot contain both -1 and 1 simultaneously and as such cannot contain a tail.

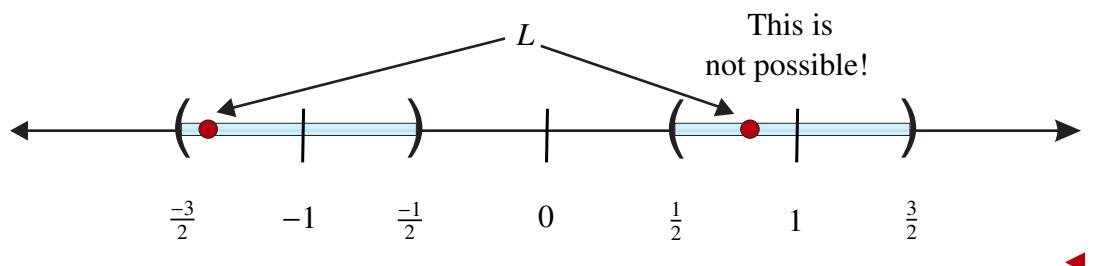
To make this more precise, we first let our tolerance be $\epsilon = \frac{1}{2}$. From the definition of a limit, the open interval $(L - \frac{1}{2}, L + \frac{1}{2})$ must contain a tail of our sequence. This would mean that the interval $(L - \frac{1}{2}, L + \frac{1}{2})$ would contain 1 , since every tail of our sequence contains terms with values equal to 1 . So now we know that the distance from 1 to L is less than $\frac{1}{2}$ (i.e., $|1 - L| < \frac{1}{2}$). Therefore we have $L \in (\frac{1}{2}, \frac{3}{2})$.



We also know that the tail contains terms with value -1 . So the distance from L to -1 is also less than $\frac{1}{2}$ and as such this time we have $L \in (-\frac{3}{2}, -\frac{1}{2})$.



This is a problem since there is no number that lies in both $(\frac{1}{2}, \frac{3}{2})$ and $(-\frac{3}{2}, -\frac{1}{2})$. Thus we have shown that our sequence $\{(-1)^{n+1}\}$ has no limit, so it diverges.



The argument in the previous example can also be modified to prove the following theorem that says that *limits of sequences are unique*.

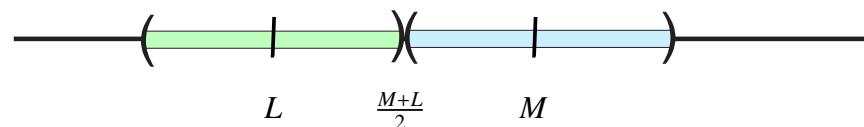
THEOREM 2

Uniqueness of Limits for Sequences

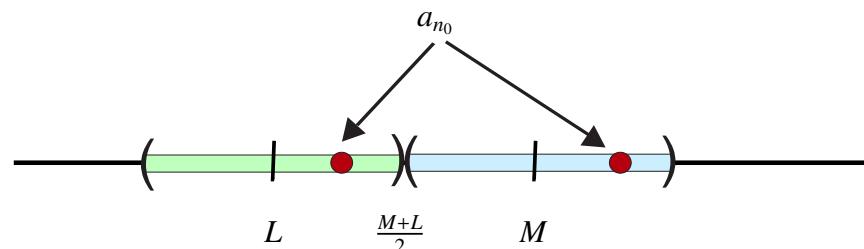
Let $\{a_n\}$ be a sequence. If it has a limit L , then the limit is unique.

PROOF

Suppose that the sequence had two different limits L and M . We can always assume that $L < M$, so let's do so. From here we will build two disjoint open intervals, one containing M and the other containing L . To do this we note that the midpoint between L and M is $\frac{M+L}{2}$. So let's consider the intervals $(L - 1, \frac{M+L}{2})$ which contains L and $(\frac{M+L}{2}, M + 1)$ which contains M .



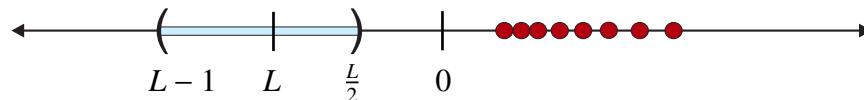
Since we are assuming that both L and M are limits, both of these intervals must contain a tail of the sequence. Since only finitely many terms are excluded from each tail, there are only finitely many terms that are not in each interval. But since we have infinitely many terms, at least one term a_{n_0} must be in both intervals.



However, our intervals are disjoint so this cannot happen. As such our assumption that the sequence had two different limits cannot be true. ■

We will end this section with one more useful observation.

Observation: Suppose we had a sequence $\{a_n\}$ consisting of only non-negative terms. That is $a_n \geq 0$ for all $n \in \mathbb{N}$. Then this sequence cannot converge to a negative value. To see why note that if $L < 0$, then the interval $(L - 1, \frac{L}{2})$ consists of only negative numbers. Therefore, this interval cannot contain any terms in the sequence. It follows that L could not be the limit of the sequence.



PROPOSITION 3

Let $\{a_n\}$ be a sequence with $a_n \geq 0$ for each $n \in \mathbb{N}$. Assume that $L = \lim_{n \rightarrow \infty} a_n$. Then $L \geq 0$.

EXERCISE 1

Let $\{a_n\}$ be a sequence with $a_n > 0$ for each $n \in \mathbb{N}$. Assume that $L = \lim_{n \rightarrow \infty} a_n$. Is it always the case that $L > 0$, or could it be that $L = 0$?

4.1.5 Divergence to $\pm\infty$

The sequence $\{1, 2^2, 3^2, \dots, n^2, \dots\}$ does not converge because as the index n increases the terms grow without bounds and therefore do not approach a fixed value L . As such we could say that the terms *approach* ∞ as the index n increases. For this reason we might want to write

$$\lim_{n \rightarrow \infty} n^2 = \infty.$$

The notation may be a little misleading because in fact, the sequence has no limit. But it does help to describe how the sequence behaves far out in the tail. In particular no matter how big $M > 0$ might be, so long as the index n is big enough, in this case bigger than \sqrt{M} , we have $n^2 > M$. This observation leads us to the following definition:

DEFINITION**Divergence to $+\infty$**

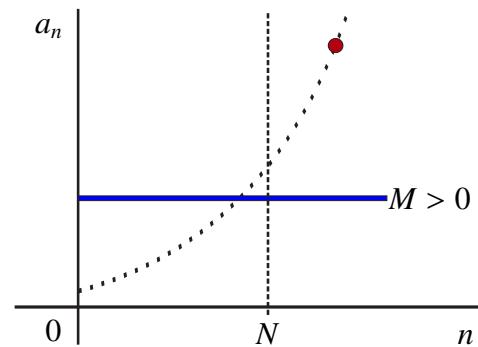
We say that a sequence $\{a_n\}$ diverges to ∞ if for every $M > 0$ we can find a cutoff $N \in \mathbb{N}$ so that if $n \geq N$, then

$$a_n > M.$$

In this case, we write

$$\lim_{n \rightarrow \infty} a_n = \infty.$$

Equivalently, we have that $\lim_{n \rightarrow \infty} a_n = \infty$ if every interval of the form (M, ∞) contains a tail of the sequence.



DEFINITION**Divergence to $-\infty$**

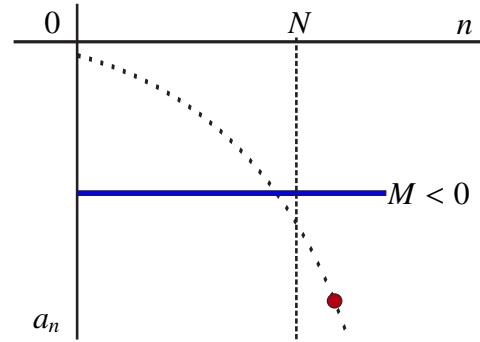
We say that a sequence $\{a_n\}$ diverges to $-\infty$ if for every $M < 0$ we can find a cutoff $N \in \mathbb{N}$ so that if $n \geq N$, then

$$a_n < M.$$

In this case, we write

$$\lim_{n \rightarrow \infty} a_n = -\infty.$$

Equivalently, we have that $\lim_{n \rightarrow \infty} a_n = -\infty$ if every interval of the form $(-\infty, M)$ contains a tail of the sequence.



Important Note: When we write $\lim_{n \rightarrow \infty} a_n = \infty$ or $\lim_{n \rightarrow \infty} a_n = -\infty$, we are *not* actually implying that the sequence has a limit. Despite the notation such sequences still diverge. The notation simply gives us a way of describing the behavior of the sequence far out in the tail. ◀

The following theorem is useful:

THEOREM 4

(i) If $\alpha > 0$, then

$$\lim_{n \rightarrow \infty} n^\alpha = \infty.$$

(ii) If $\alpha < 0$, then

$$\lim_{n \rightarrow \infty} n^\alpha = 0.$$

4.1.6 Arithmetic for Limits of Sequences

In this section, we will see that most of the usual rules of arithmetic hold for limits of sequences. To illustrate this statement assume that $\{a_n\}$ converges to 3 and that $\{b_n\}$ converges to 4. For large n , we would expect a_n to be very close 3 and b_n to be very close to 4. As such, we would expect $2a_n$ to be very close to $2 \times 3 = 6$, $a_n + b_n$ to be very close to $3 + 4 = 7$, and $a_n b_n$ to be very close to $3 \times 4 = 12$. This suggests that $\{2a_n\}$ should converge to 6, that $\{a_n + b_n\}$ should converge to 7 and that $\{a_n b_n\}$ should converge to 12. The next theorem shows that these statements are in fact true.

THEOREM 5**Arithmetic Rules for Limits of Sequences**

Let $\{a_n\}$ and $\{b_n\}$ be sequences. Assume that $\lim_{n \rightarrow \infty} a_n = L$ and $\lim_{n \rightarrow \infty} b_n = M$ where L and M are Real numbers. Then

- i) For any $c \in \mathbb{R}$, if $a_n = c$ for every n , then $c = L$.
- ii) For any $c \in \mathbb{R}$, $\lim_{n \rightarrow \infty} ca_n = cL$.
- iii) $\lim_{n \rightarrow \infty} (a_n + b_n) = L + M$.
- iv) $\lim_{n \rightarrow \infty} a_n b_n = LM$.
- v) $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \frac{L}{M}$ if $M \neq 0$.
- vi) If $a_n \geq 0$ for all n and if $\alpha > 0$, then $\lim_{n \rightarrow \infty} a_n^\alpha = L^\alpha$.
- vii) For any $k \in \mathbb{N}$, $\lim_{n \rightarrow \infty} a_{n+k} = L$.

PROOF

Proof of i): Given any $\epsilon > 0$, if $n \geq 1$, then

$$|a_n - c| = |c - c| = 0 < \epsilon.$$

Proof of ii): This proof splits into two cases. The first where $c = 0$. The second where $c \neq 0$.

If $c = 0$, then $b_n = ca_n = 0$ for all n . It then follows by i), that

$$\lim_{n \rightarrow \infty} ca_n = 0 = 0 \cdot L = c \cdot L.$$

Assume that $c \neq 0$. Let $\epsilon > 0$. Then since we know that $\{a_n\}$ converges to L , we can find a cutoff $N \in \mathbb{N}$ such that if $n \geq N$, then

$$|a_n - L| < \frac{\epsilon}{|c|}.$$

It follows that if $n \geq N$, then

$$|ca_n - cL| = |c| \cdot |a_n - L| < |c| \cdot \frac{\epsilon}{|c|} = \epsilon.$$

NOTE

While the calculation we just used to prove part ii) of our arithmetic rules is quite straight forward, it is none the less very useful. We will use variants of this argument to establish not only the remaining rules in this theorem, but many more times through out the course.

Proof of iii): The proof we give for Rule iii) illustrates an important point, namely that “the error in a sum is less than or equal to the sum of the errors.” More specifically, the triangle inequality shows that

$$|(a_n + b_n) - (L + M)| \leq |a_n - L| + |b_n - M| \quad (*).$$

So let's suppose that we are given a tolerance $\epsilon > 0$. If we can make both $|a_n - L| < \frac{\epsilon}{2}$ and $|b_n - M| < \frac{\epsilon}{2}$, Then (*) would tell us that

$$|(a_n + b_n) - (L + M)| \leq |a_n - L| + |b_n - M| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

But since $\lim_{n \rightarrow \infty} a_n = L$, we can find a cutoff $N_1 \in \mathbb{N}$ so that if $n \geq N_1$, then

$$|a_n - L| < \frac{\epsilon}{2}.$$

Similarly since, $\lim_{n \rightarrow \infty} b_n = M$, we can find a cutoff $N_2 \in \mathbb{N}$ so that if $n \geq N_2$, then

$$|b_n - M| < \frac{\epsilon}{2}.$$

Now we let $N_0 = \max\{N_1, N_2\}$, the largest of the two cutoffs. Then if $n \geq N_0$, we have that $n \geq N_1$ and $n \geq N_2$, simultaneously. This is the required cutoff.

Proof of iv): We want to show that $\lim_{n \rightarrow \infty} (a_n b_n) = \left(\lim_{n \rightarrow \infty} a_n\right) \cdot \left(\lim_{n \rightarrow \infty} b_n\right) = LM$. To do so we start with the following observations:

Observation:

$$\begin{aligned} |a_n b_n - LM| &= |(a_n b_n - L b_n) + (L b_n - LM)| \\ &\leq |a_n b_n - L b_n| + |L b_n - LM| \\ &= |a_n - L| |b_n| + |L| |b_n - M|. \end{aligned}$$

Let $\epsilon > 0$. We know from ii) that we can choose N_1 large enough so that if $n \geq N_1$,

$$|L| |b_n - M| < \frac{\epsilon}{2}.$$

Key Observation 1:

$|a_n - L| |b_n|$ is trickier because b_n varies. We could have

$$\begin{array}{ccc} |a_n - L| & |b_n| \\ \downarrow & \downarrow \\ \text{small} & \text{BIG} \end{array}$$

$$= ?$$

Conclusion: We must control the size of $|b_n|$.

Key Observation 2: The following theorem will allow us to control the size of our sequence $\{b_n\}$:

THEOREM 6

Assume that $\{b_n\}$ converges. Then $\{b_n\}$ is bounded when viewed as a subset of \mathbb{R} .

PROOF

We know that there exists a N so that if $n \geq N$, then

$$|b_n - M| < 1$$

The Triangle Inequality then shows that

$$|b_n| < |M| + 1.$$

for all $n \geq N$. Let

$$K = \max\{|b_1|, |b_2|, \dots, |b_{N-1}|, |M| + 1\}$$

then

$$|b_n| \leq K$$

for all $n \in \mathbb{N}$. ■

Proof of iv) Continued: There is a $K > 0$ such that,

$$|a_n - L| \leq K |a_n - L|.$$

We can find N_2 large enough so that if $n \geq N_2$, then

$$|a_n - L| \leq K |a_n - L| < \frac{\epsilon}{2}.$$

If $n \geq N_0 = \max\{N_1, N_2\}$, then

$$\begin{aligned} |a_n b_n - LM| &\leq |a_n - L| |b_n| + |L| |b_n - M| \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} \\ &= \epsilon. \end{aligned}$$
■

Proof of v): We want to show that $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \frac{L}{M}$ if $M \neq 0$.

Strategy: Start with a simpler case:

v') If $M \neq 0$, $\Rightarrow \lim_{n \rightarrow \infty} \frac{1}{b_n} = \frac{1}{M}$.

Observation:

$$\left| \frac{1}{b_n} - \frac{1}{M} \right| = \frac{|b_n - M|}{|b_n M|}$$

Problem: We can make $|b_n - M|$ small, but $|b_n M|$ might also be small, and

$$\frac{\text{small}}{\text{small}} = ?$$

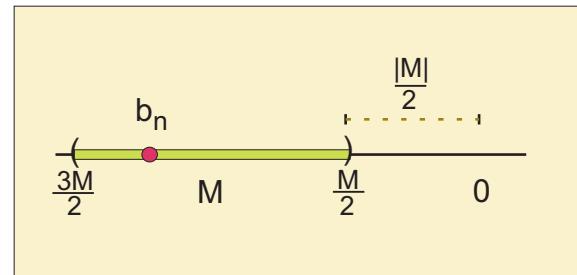
Challenge: We must make sure that b_n is not too small!

Letting $\epsilon = \frac{|M|}{2}$, we can find a cutoff $N_1 \in \mathbb{N}$ so that if $n \geq N_1$, then

$$|b_n - M| < \frac{|M|}{2}.$$

If $n \geq N_1$, we get that b_n will be in the open interval between $\frac{|M|}{2}$ and $\frac{3|M|}{2}$. In particular,

$$\frac{|M|}{2} < b_n.$$



If $n \geq N_1$ then

$$\begin{aligned} \frac{|M|}{2} &\leq |b_n| \Rightarrow \frac{1}{|b_n|} \leq \frac{1}{\frac{|M|}{2}} \\ &\Rightarrow \frac{1}{|b_n M|} \leq \frac{1}{\frac{M^2}{2}} \\ &\Rightarrow \frac{|b_n - M|}{|b_n M|} \leq \frac{|b_n - M|}{\frac{M^2}{2}} \end{aligned}$$

This shows that if $n \geq N_1$, then

$$\left| \frac{1}{b_n} - \frac{1}{M} \right| \leq \frac{|b_n - M|}{\frac{M^2}{2}}.$$

However, we know how to proceed given the last inequality. We choose a cutoff $N_2 \in \mathbb{N}$ so that if $n \geq N_2$, then

$$|b_n - M| < \left(\frac{|M|^2}{2}\right) \cdot \epsilon.$$

If we let $N_0 = \max\{N_1, N_2\}$, and if $n \geq N_0$, then

$$\left| \frac{1}{b_n} - \frac{1}{M} \right| \leq \frac{|b_n - M|}{\frac{|M|^2}{2}} \leq \frac{\left(\frac{|M|^2}{2}\right) \cdot \epsilon}{\frac{|M|^2}{2}} = \epsilon.$$

This proves v').

To prove v) we can now appeal to iv) and v') to get

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \left(\lim_{n \rightarrow \infty} a_n\right) \cdot \left(\lim_{n \rightarrow \infty} \frac{1}{b_n}\right) = \frac{L}{M}.$$

Proof of vi): Assume that $a_n \geq 0$ for all n and that $\lim_{n \rightarrow \infty} a_n = L$. If $\alpha > 0$, we want to show that $\lim_{n \rightarrow \infty} a_n^\alpha = L^\alpha$. There are two cases to consider: when $L = 0$ and when $L \neq 0$.

Case 1: Assume that $L = 0$.

Let $\epsilon > 0$. Since $L = 0$, we can find a cutoff $N \in \mathbb{N}$ such that if $n \geq N$, then

$$a_n < \epsilon^{\frac{1}{\alpha}}.$$

However, if $n \geq N$, this means that

$$a_n^\alpha < \epsilon$$

as desired.

Case 2: Assume that $L \neq 0$.

This case is much more complicated than the first case. We will prove the result for any rational number α . The irrational case is beyond the scope of this course. We will begin by assuming that $\alpha = \frac{1}{n}$ for some $\alpha \in \mathbb{N}$.

We will use the fact that for any $a, b \in \mathbb{R}$, with $a > 0$ and $b > 0$, we have

$$|a^n - b^n| = |a - b| \cdot [a^{n-1} + a^{n-2}b + a^{n-3}b^2 + \cdots + ab^{n-2} + b^{n-1}]$$

and hence

$$|a - b| = \frac{|a^n - b^n|}{[a^{n-1} + a^{n-2}b + a^{n-3}b^2 + \cdots + ab^{n-2} + b^{n-1}]} < \frac{|a^n - b^n|}{b^{n-1}}.$$

Let $a = (a_k)^{\frac{1}{n}}$ and $b = L^{\frac{1}{n}}$. we get

$$|(a_k)^{\frac{1}{n}} - L^{\frac{1}{n}}| < \frac{|a_k - L|}{L^{\frac{n-1}{n}}}.$$

As such, given $\epsilon > 0$, choose a cutoff N so that if $k \geq N$, then

$$|a_k - L| < L^{\frac{n-1}{n}} \cdot \epsilon.$$

From here we get that if $k \geq N$, then

$$|(a_k)^{\frac{1}{n}} - L^{\frac{1}{n}}| < \frac{|a_k - L|}{L^{\frac{n-1}{n}}} < \frac{L^{\frac{n-1}{n}} \cdot \epsilon}{L^{\frac{n-1}{n}}} = \epsilon.$$

This shows that

$$\lim_{k \rightarrow \infty} a_k^\alpha = L^\alpha$$

when $\alpha = \frac{1}{n}$.

Let $\alpha = \frac{m}{n}$ where $m, n \in \mathbb{N}$. Since

$$(a_k)^{\frac{m}{n}} = (a_k^m)^{\frac{1}{n}}$$

we can appeal to the above calculation as well as Rule iv) to get that

$$\lim_{k \rightarrow \infty} a_k^\alpha = L^\alpha$$

when α is a positive rational number and then to Rule v) to extend this to all rational numbers.

The case where α is irrational is beyond the scope of this course.

■

Important Note: It is worth devoting special attention to Rule v) concerning quotients. It states that $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \frac{L}{M}$ if $M \neq 0$. But what happens if $\lim_{n \rightarrow \infty} b_n = M = 0$? It turns out that in this case $\lim_{n \rightarrow \infty} \frac{a_n}{b_n}$ sometimes exists and sometimes it does not. For example, if

$$a_n = \frac{1}{n} = b_n$$

for each $n \in \mathbb{N}$, then if $c_n = \frac{a_n}{b_n}$, we get that $c_n = 1$ for each n . Consequently,

$$\lim_{n \rightarrow \infty} c_n = \lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 1.$$

However if $a_n = \frac{1}{n}$ and $b_n = \frac{1}{n^2}$, then in this case

$$c_n = \frac{a_n}{b_n} = n$$

for each $n \in \mathbb{N}$. Consequently,

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \lim_{n \rightarrow \infty} n = \infty.$$

There is an additional observation that we can make. If $\lim_{n \rightarrow \infty} \frac{a_n}{b_n}$ exists and $\lim_{n \rightarrow \infty} b_n = 0$, **then we must have that $\lim_{n \rightarrow \infty} a_n = 0$ as well**. To see this we let

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = K.$$

Then since $a_n = b_n \cdot \frac{a_n}{b_n}$, Rule iv) shows that

$$\begin{aligned}\lim_{n \rightarrow \infty} a_n &= \lim_{n \rightarrow \infty} b_n \cdot \frac{a_n}{b_n} \\ &= \lim_{n \rightarrow \infty} b_n \times \lim_{n \rightarrow \infty} \frac{a_n}{b_n} \\ &= 0 \cdot K \\ &= 0.\end{aligned}$$



Let's think about why the observation above is true. Suppose that $L \neq 0$, but $M = 0$. Then as n becomes large, b_n becomes very small. However when we divide a_n by a very small number, namely b_n , the quotient tends to have very large magnitude unless a_n is also very small. But if $\lim_{n \rightarrow \infty} a_n \neq 0$, then eventually the magnitude of b_n is much smaller than that of a_n , so the ratio $\frac{a_n}{b_n}$ grows very large. Therefore, the limit *cannot* exist.

THEOREM 7

Assume that $\{a_n\}$ and $\{b_n\}$ are two sequences and that $\lim_{n \rightarrow \infty} b_n = 0$. Assume also that $\lim_{n \rightarrow \infty} \frac{a_n}{b_n}$ exists. Then

$$\lim_{n \rightarrow \infty} a_n = 0.$$

The previous theorem will actually play an important role in this course. We will be able to use it to establish a similar result for limits of functions, and then we will rely on this new result to show that differentiability implies continuity.

REMARK

Rule vii) states that if $\lim_{n \rightarrow \infty} a_n = L$, then for any $k > 0$, $\lim_{n \rightarrow \infty} a_{n+k} = L$ as well. In particular, we will later use the fact that if $\lim_{n \rightarrow \infty} a_n = L$, then

$$\lim_{n \rightarrow \infty} a_{n+1} = L$$

to help us find the limit of a recursively defined sequence.



This rule is really just an observation that convergence is about the behaviour of the tail of a sequence. We can in fact change any finite number of the terms in a sequence without impacting convergences.

EXAMPLE 6 Find the limit for the sequence $\{\frac{n+3}{n+1}\}$.

SOLUTION

If we let $n = 1000$, we get $a_{1000} = \frac{1003}{1001} = 1.001998\dots$, while if we let $n = 10000000$, we get $a_{10000000} = \frac{10000003}{10000001} = 1.0000002\dots$. As you might guess, as n gets larger $\{\frac{n+3}{n+1}\}$ gets closer and closer to 1. Hence, we would expect $\lim_{n \rightarrow \infty} \frac{n+3}{n+1} = 1$. Let's see how we can use the Arithmetic Rules for Limits of Sequences to verify this result.

First notice that we can remove a factor of n from both the numerator and denominator to get

$$\begin{aligned}\frac{n+3}{n+1} &= \left(\frac{n}{n}\right)\left(\frac{1 + \frac{3}{n}}{1 + \frac{1}{n}}\right) \\ &= \frac{1 + \frac{3}{n}}{1 + \frac{1}{n}}\end{aligned}$$

From this we see that $\lim_{n \rightarrow \infty} \frac{n+3}{n+1} = \lim_{n \rightarrow \infty} \frac{1 + \frac{3}{n}}{1 + \frac{1}{n}}$.

We can now apply our limit rules to get that

$$\begin{aligned}\lim_{n \rightarrow \infty} \frac{n+3}{n+1} &= \lim_{n \rightarrow \infty} \frac{1 + \frac{3}{n}}{1 + \frac{1}{n}} \\ &= \frac{\lim_{n \rightarrow \infty} 1 + \lim_{n \rightarrow \infty} \frac{3}{n}}{\lim_{n \rightarrow \infty} 1 + \lim_{n \rightarrow \infty} \frac{1}{n}} \\ &= \frac{\lim_{n \rightarrow \infty} 1 + 3 \lim_{n \rightarrow \infty} \frac{1}{n}}{\lim_{n \rightarrow \infty} 1 + \lim_{n \rightarrow \infty} \frac{1}{n}} \\ &= \frac{1 + 3(0)}{1 + 0} \\ &= 1\end{aligned}$$

just as we expected. 

The next example is similar to the previous one. You may want to try to do this question before reading the solution.

EXAMPLE 7 Find the limit for the sequence $\{\frac{3n^2+6n-3}{n^2+2n}\}$.

SOLUTION

First notice that we can again factor out the largest power of n , namely n^2 , from each of the numerator and denominator to get

$$\begin{aligned}\frac{3n^2 + 6n - 3}{n^2 + 2n} &= \left(\frac{n^2}{n^2}\right) \left(\frac{3 + \frac{6}{n} - \frac{3}{n^2}}{1 + \frac{2}{n}}\right) \\ &= \frac{3 + \frac{6}{n} - \frac{3}{n^2}}{1 + \frac{2}{n}}\end{aligned}$$

Applying the rules of limits as in the previous example gives us

$$\begin{aligned}\lim_{n \rightarrow \infty} \frac{3n^2 + 6n - 3}{n^2 + 2n} &= \lim_{n \rightarrow \infty} \frac{3 + \frac{6}{n} - \frac{3}{n^2}}{1 + \frac{2}{n}} \\ &= 3\end{aligned}$$



EXAMPLE 8 Consider the recursively defined sequence

$$a_1 = 1 \text{ and } a_{n+1} = \frac{1}{1 + a_n}.$$

At this point it is likely impossible for you to determine if this sequence converges or diverges. To help you gain an understanding for how the sequence behaves, consider the following table that contains the exact values of a_n and their 7-decimal place equivalents for n from 1 to 25:

n	a_n	Decimal
1	1	1.0000000
2	$\frac{1}{2}$.5000000
3	$\frac{2}{3}$.6666666
4	$\frac{3}{5}$.6000000
5	$\frac{5}{8}$.6250000
6	$\frac{8}{13}$.6153846
7	$\frac{13}{21}$.6190476
8	$\frac{21}{34}$.6176470
9	$\frac{34}{55}$.6181818
10	$\frac{55}{89}$.6179775
11	$\frac{89}{144}$.6180555
12	$\frac{144}{233}$.6180257
13	$\frac{233}{377}$.6180371

n	a_n	Decimal
14	$\frac{377}{610}$.6180327
15	$\frac{610}{987}$.6180344
16	$\frac{987}{1597}$.6180338
17	$\frac{1597}{2584}$.6180340
18	$\frac{2584}{4181}$.6180339
19	$\frac{4181}{6765}$.6180340
20	$\frac{6765}{10946}$.6180339
21	$\frac{10946}{17711}$.6180339
22	$\frac{17711}{28657}$.6180339
23	$\frac{28657}{46368}$.6180339
24	$\frac{46368}{75025}$.6180339
25	$\frac{75025}{121393}$.6180339

Notice that the terms of this sequence do seem to be getting closer together. In fact, the values of a_n agree up to the first 7 decimal places for n from 20 to 25. This is strong evidence to suggest that this sequence converges to a value near 0.6180339, but this is still *not* a proof of convergence. Indeed, the sequence *does actually converge*, but we will not provide a proof of this statement here. However, once we know it converges, we can use the rules of limits to find the value of the limit. How do we do this?

Assume that $\lim_{n \rightarrow \infty} a_n = L$. We begin with the recursive definition:

$$a_{n+1} = \frac{1}{1 + a_n}.$$

First note that since $a_1 = 1 > 0$, all of the subsequent terms will also be positive. This shows that $L \geq 0$. From Rule vii) we get that $\lim_{n \rightarrow \infty} a_{n+1} = L$ as well. But then the rules of limits give us

$$\begin{aligned} L &= \lim_{n \rightarrow \infty} a_{n+1} \\ &= \lim_{n \rightarrow \infty} \frac{1}{1 + a_n} \\ &= \frac{1}{1 + \lim_{n \rightarrow \infty} a_n} \\ &= \frac{1}{1 + L}. \end{aligned}$$

This means that

$$L = \frac{1}{1 + L}.$$

Cross-multiplication shows that

$$L(L + 1) = 1$$

and hence that

$$L^2 + L - 1 = 0.$$

We use the *quadratic formula* to get that

$$\begin{aligned} L &= \frac{-1 \pm \sqrt{(1)^2 - 4(1)(-1)}}{2(1)} \\ &= \frac{-1 \pm \sqrt{5}}{2} \end{aligned}$$

To determine L , recall that we must have $L \geq 0$. This means that

$$L = \frac{-1 + \sqrt{5}}{2}.$$

Finally, you can use your calculator to see that

$$L = \frac{-1 + \sqrt{5}}{2} = .6180339\dots$$

exactly as the table of values predicted. 

EXAMPLE 9 Evaluate $\lim_{n \rightarrow \infty} \frac{3n^2+2n-1}{4n^2+2}$.

SOLUTION

Note that

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{3n^2 + 2n - 1}{4n^2 + 2} &= \lim_{n \rightarrow \infty} \frac{n^2}{n^2} \cdot \frac{3 + \frac{2}{n} - \frac{1}{n^2}}{4 + \frac{2}{n^2}} \quad (*) \\ &= \lim_{n \rightarrow \infty} \frac{3 + \frac{2}{n} - \frac{1}{n^2}}{4 + \frac{2}{n^2}} \\ &= \frac{3}{4} \end{aligned}$$

Observation: If we look at (*), then we see that for large values of n the terms $\frac{2}{n}$, $\frac{-1}{n^2}$ and $\frac{2}{n^2}$ are all very close to 0 and as a result when n is large $\frac{3n^2+2n-1}{4n^2+2}$ behaves like

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


EXAMPLE 10 Evaluate $\lim_{n \rightarrow \infty} \frac{3n^2+2n-1}{4n^3+2}$.

SOLUTION

Note that

$$\begin{aligned}
 \lim_{n \rightarrow \infty} \frac{3n^2 + 2n - 1}{4n^3 + 2} &= \lim_{n \rightarrow \infty} \frac{n^2}{n^3} \cdot \frac{3 + \frac{2}{n} - \frac{1}{n^2}}{4 + \frac{2}{n^3}} \\
 &= \lim_{n \rightarrow \infty} \frac{1}{n} \cdot \frac{3 + \frac{2}{n} - \frac{1}{n^2}}{4 + \frac{2}{n^3}} \\
 &= \lim_{n \rightarrow \infty} \frac{1}{n} \cdot \lim_{n \rightarrow \infty} \frac{3 + \frac{2}{n} - \frac{1}{n^2}}{4 + \frac{2}{n^3}} \\
 &= 0 \cdot \frac{3}{4} \\
 &= 0
 \end{aligned} \tag{**}$$

Observation: If we look at (**), then we see that for large values of n the terms $\frac{2}{n^2}$, $\frac{-1}{n^2}$ and $\frac{2}{n^3}$ are all very close to 0 and as a result when n is large $\frac{3n^2+2n-1}{4n^3+2}$ behaves like

$$\frac{3n^2}{4n^3} = \frac{3}{4n}$$

which converges to 0. ◀

EXAMPLE 11 Consider the sequence $\left\{ \frac{3n^2+5}{n^{3/2}+2} \right\}$. We know that

$$\frac{3n^2+5}{n^{3/2}+2} = \frac{n^{3/2}}{n^{3/2}} \cdot \frac{3n^{1/2} + \frac{5}{n^{3/2}}}{1 + \frac{2}{n^{3/2}}} \geq \frac{3n^{1/2}}{1+2} = \sqrt{n},$$

which tells us that the sequence grows without bound. This shows that $\lim_{n \rightarrow \infty} \frac{3n^2+5}{n^{3/2}+2} = \infty$.

Alternatively, by factoring out the highest power from both the numerator and denominator we get

$$\begin{aligned}
 \lim_{n \rightarrow \infty} \frac{3n^2+5}{n^{3/2}+2} &= \lim_{n \rightarrow \infty} \frac{n^2}{n^{3/2}} \cdot \frac{3 + \frac{5}{n^2}}{1 + \frac{2}{n^{3/2}}} \quad (***) \\
 &= \lim_{n \rightarrow \infty} \sqrt{n} \cdot \frac{3 + \frac{5}{n^2}}{1 + \frac{2}{n^{3/2}}} \\
 &= \infty
 \end{aligned}$$

Observation: If we look at (**), then we see that for large values of n the terms

$\frac{5}{n^2}$ and $\frac{2}{n^{3/2}}$ are both very close to 0 and as a result when n is large $\frac{3n^2+5}{n^{3/2}+2}$ behaves like

$$\frac{n^2}{n^{\frac{3}{2}}} \cdot 3 = 3n^{\frac{1}{2}}$$

which approaches ∞ . ◀

EXAMPLE 12 Let

$$a_n = \frac{b_0 + b_1 n + b_2 n^2 + b_3 n^3 + \cdots + b_j n^j}{c_0 + c_1 n + c_2 n^2 + \cdots + c_k n^k}.$$

Consider the sequence $\{a_n\}$. By factoring out n^j from the numerator and n^k from the denominator and rewriting the sequence as

$$a_n = \frac{n^j}{n^k} \left[\frac{\frac{b_0}{n^j} + \frac{b_1}{n^{j-1}} + \frac{b_2}{n^{j-2}} + \frac{b_3}{n^{j-3}} + \cdots + b_j}{\frac{c_0}{n^k} + \frac{c_1}{n^{k-1}} + \frac{c_2}{n^{k-2}} + \cdots + c_k} \right],$$

we can show that

$$\lim_{n \rightarrow \infty} a_n = \begin{cases} \frac{b_j}{c_k} & \text{if } j = k \\ 0 & \text{if } j < k \\ \infty & \text{if } j > k \text{ and } \frac{b_j}{c_k} > 0 \\ -\infty & \text{if } j > k \text{ and } \frac{b_j}{c_k} < 0. \end{cases}$$

Then we get that for large n

$$\frac{b_0 + b_1 n + b_2 n^2 + b_3 n^3 + \cdots + b_j n^j}{c_0 + c_1 n + c_2 n^2 + \cdots + c_k n^k} \sim \frac{b_j n^j}{c_k n^k} = \frac{b_j}{c_k} n^{j-k}.$$
◀

EXAMPLE 13

Consider the sequence $\{a_n\} = \{\sqrt{n^2 + n} - n\}$. We could try to evaluate this sequence by separating the components and arguing that

$$\lim_{n \rightarrow \infty} \sqrt{n^2 + n} - n = \lim_{n \rightarrow \infty} \sqrt{n^2 + n} - \lim_{n \rightarrow \infty} n = \infty - \infty.$$

However, $\infty - \infty$ has no mathematical meaning.

We could try evaluating a few terms for large values of the index n . For example, $a_{100} = 0.498756211$ and $a_{10^6} = 0.499999875$. These calculations suggest that the limit might be $\frac{1}{2}$. But how do we show this? It turns out that we can use the following

trick (the conjugate) and write

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

It follows that

$$\lim_{n \rightarrow \infty} \sqrt{n^2 + n} - n = \lim_{n \rightarrow \infty} \frac{1}{\sqrt{1 + \frac{1}{n}} + 1} = \frac{1}{2}.$$



4.2 Squeeze Theorem

We have seen that there are natural rules of arithmetic for sequences, but some care must be taken to meet all of the underlying conditions. We illustrate this with the following example:

EXAMPLE 14 Let $a_n = \frac{\sin(n)}{n}$. We want to find the limit of the sequence $\{a_n\}$. We would like to argue that:

$$\lim_{n \rightarrow \infty} a_n = \left(\lim_{n \rightarrow \infty} \sin(n) \right) \cdot \left(\lim_{n \rightarrow \infty} \frac{1}{n} \right) = \left(\lim_{n \rightarrow \infty} \sin(n) \right) \cdot 0 = 0.$$

However, it can be shown that the sequence $\{\sin(n)\}$ does not converge, and so we cannot use our Product Rule iv) for Sequences to conclude that the limit is 0.

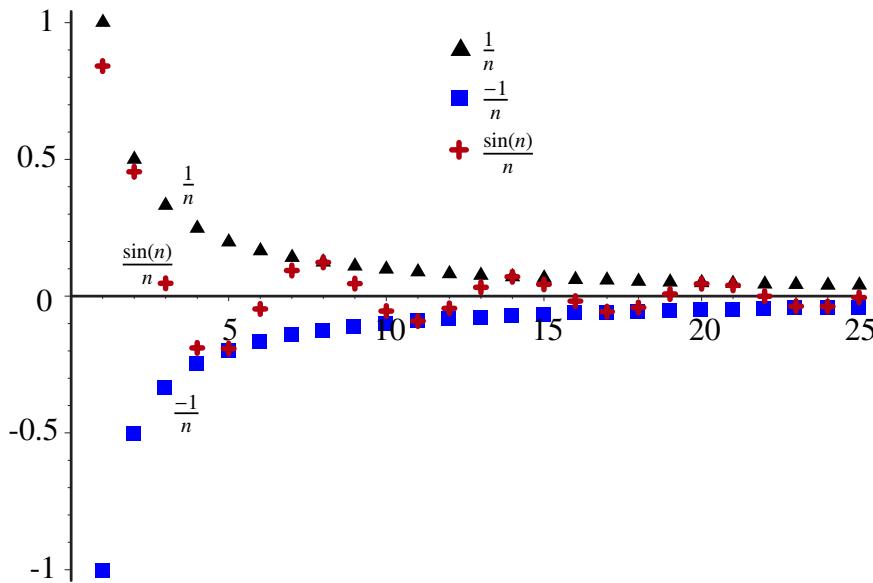
Intuitively, we can see that the limit is indeed 0 because the sine function is never larger than 1 in absolute value. Dividing by n will cause $\frac{\sin(n)}{n}$ to converge to 0 at least as quickly as $\frac{1}{n}$. Alternatively, since $|\sin(n)| \leq 1$ for all $n \in \mathbb{N}$, we have that

$$|a_n| = \left| \frac{\sin(n)}{n} \right| \leq \frac{1}{n}$$

for all n . But this means that

$$\frac{-1}{n} \leq \frac{\sin(n)}{n} \leq \frac{1}{n}.$$

That is, $\{\frac{\sin(n)}{n}\}$ is “squeezed” between two sequences that converge to the same limit: 0.



The diagram above suggests that the sequence $\{\frac{\sin(n)}{n}\}$ does indeed converge and that the limit is 0 as we expected. But must $\{\frac{\sin(n)}{n}\}$ actually converge? And if $\{\frac{\sin(n)}{n}\}$ converges, is its limit actually 0? ◀

The answer to these questions can be obtained from the following very useful rule called the *Squeeze Theorem for Sequences*.

THEOREM 8

Squeeze Theorem for Sequences

Assume that $a_n \leq b_n \leq c_n$ for all $n \in \mathbb{N}$, and

$$\lim_{n \rightarrow \infty} a_n = L = \lim_{n \rightarrow \infty} c_n.$$

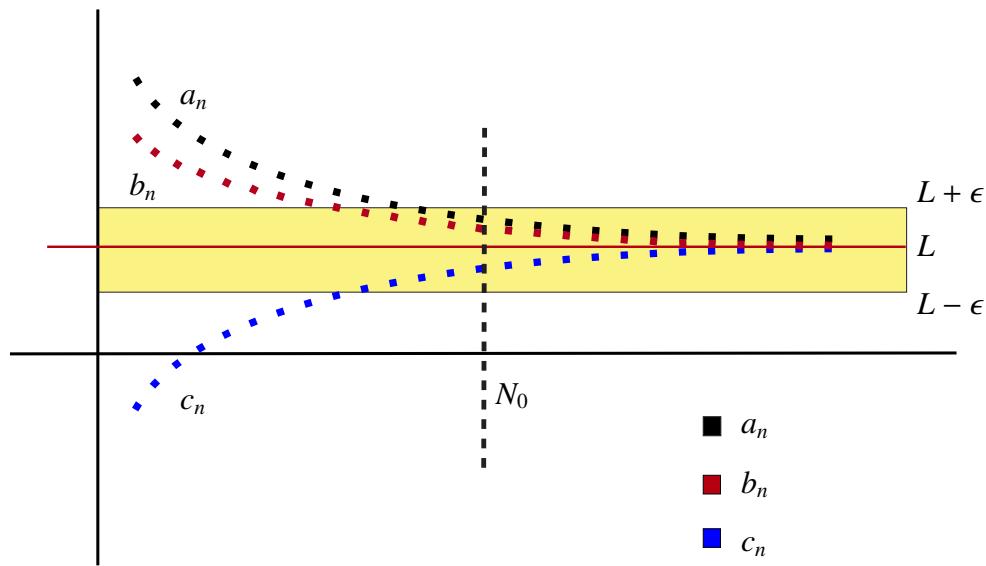
Then $\{b_n\}$ converges and $\lim_{n \rightarrow \infty} b_n = L$.

PROOF

Choose a tolerance $\epsilon > 0$. Since $\lim_{n \rightarrow \infty} a_n = L = \lim_{n \rightarrow \infty} c_n$, we can find an $N_0 \in \mathbb{N}$ such that if $n \geq N_0$, then $a_n \in (L - \epsilon, L + \epsilon)$ and $c_n \in (L - \epsilon, L + \epsilon)$. But then if $n \geq N_0$, this would mean that

$$L - \epsilon < a_n \leq b_n \leq c_n < L + \epsilon,$$

implying that $b_n \in (L - \epsilon, L + \epsilon)$. This shows that $\{b_n\}$ converges and $\lim_{n \rightarrow \infty} b_n = L$, as required. ■



We can now address the problems that we encountered in previous example.

EXAMPLE 15

We will show that $\left\{ \frac{\sin(n)}{n} \right\}$ converges and find its limit. Since $|\sin(n)| \leq 1$ for all $n \in \mathbb{N}$, $\left| \frac{\sin(n)}{n} \right| \leq \frac{1}{n}$ for all n . Then

$$\frac{-1}{n} \leq \frac{\sin(n)}{n} \leq \frac{1}{n},$$

for all $n \in \mathbb{N}$. Since $\lim_{n \rightarrow \infty} \frac{-1}{n} = 0 = \lim_{n \rightarrow \infty} \frac{1}{n}$, the Squeeze Theorem shows that $\left\{ \frac{\sin(n)}{n} \right\}$ converges and has limit 0. ◀

4.3 Monotone Convergence Theorem

In this section, we will study the convergence of an important class of sequences called *monotonic sequences*. We will discover that for sequences in this class, there is a simple rule for determining whether the sequence converges or diverges, and in the case of convergence, the limit of the sequence.

DEFINITION

Monotonic Sequences

We say that a sequence $\{a_n\}$ is:

- *increasing* if $a_n < a_{n+1}$, for all $n \in \mathbb{N}$.
- *non-decreasing* if $a_n \leq a_{n+1}$, for all $n \in \mathbb{N}$.
- *decreasing* if $a_n > a_{n+1}$, for all $n \in \mathbb{N}$.
- *non-increasing* if $a_n \geq a_{n+1}$, for all $n \in \mathbb{N}$.
- *monotonic* if $\{a_n\}$ is either non-decreasing or non-increasing.

The next example will give us a powerful clue to the nature of convergence for monotonic sequences.

EXAMPLE 16 Let S be the terms in the increasing sequence $\{1 - \frac{1}{n}\}$. That is,

$$S = \{0, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \dots\}.$$

Notice that each term is less than 1, but we can get as close to 1 as we like so long as the index n is large enough. This means that 1 is an upper bound for S and no other number that is smaller than 1 could be an upper bound. Hence $1 = \text{lub}(S)$. ◀

We also know that

$$1 = \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right).$$

It turns out that the fact that the limit and the least upper bound agree is no accident.

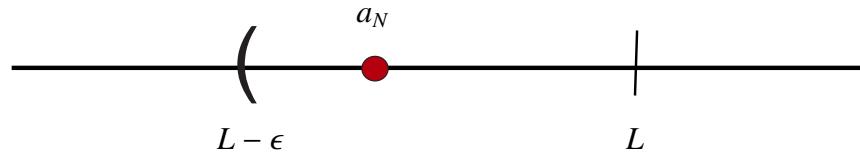
Important Observation: Let $\{a_n\}$ be a non-decreasing sequence that is bounded above. Let

$$L = \text{lub}(\{a_n\}).$$

We claim that $\{a_n\}$ converges to L . ◀

To see why this is the case, we choose a tolerance $\epsilon > 0$. Then the number $L - \epsilon$ is strictly smaller than L . This means that $L - \epsilon$ *cannot be an upper bound of the sequence*. Therefore, there must be an index N such that

$$L - \epsilon < a_N$$

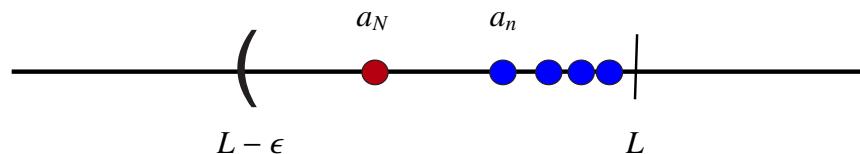


Let $n \geq N$. Since $\{a_n\}$ is non-decreasing

$$L - \epsilon < a_N \leq a_n.$$

However, since L is an upper bound, we have actually shown that for all $n \geq N$,

$$L - \epsilon < a_N \leq a_n \leq L.$$



It follows immediately from this inequality that if $n \geq N$, a_n approximates L with an error that is less than ϵ . This is exactly what we required for us to conclude that

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

If $\{a_n\}$ is non-decreasing and not bounded and M is any number, then since M is not an upper bound for our sequence there must be a N such that

$$M < a_N.$$

However, if $n \geq N$, we have

$$M < a_N \leq a_n.$$

This shows that $\{a_n\}$ diverges to ∞ or equivalently,

$$\lim_{n \rightarrow \infty} a_n = \infty.$$

We have just established a simple test for the convergence of a non-decreasing sequence.

THEOREM 9

Monotone Convergence Theorem (MCT)

Let $\{a_n\}$ be a non-decreasing sequence.

1. If $\{a_n\}$ is bounded above, then $\{a_n\}$ converges to $L = \text{lub}(\{a_n\})$.
2. If $\{a_n\}$ is not bounded above, then $\{a_n\}$ diverges to ∞ .

In particular, $\{a_n\}$ converges if and only if it is bounded above.

NOTE

A similar statement can be made about non-increasing sequences by replacing the least upper bound with the greatest lower bound and ∞ by $-\infty$. 

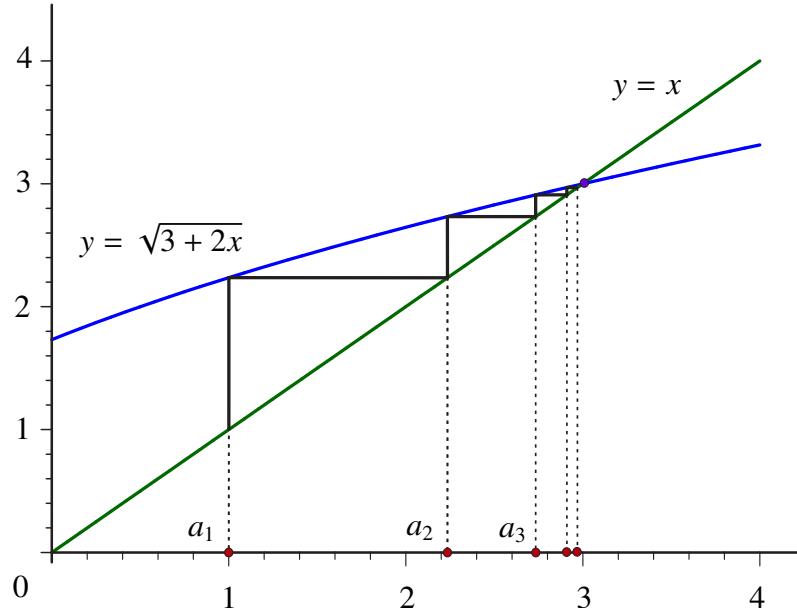
EXAMPLES

1. The sequence $\{\frac{-1}{n}\} = \{-1, -\frac{1}{2}, -\frac{1}{3}, -\frac{1}{4}, -\frac{1}{5}, \dots\}$ is increasing and bounded above by 0. In fact, 0 is the least upper bound of the sequence since

$$\lim_{n \rightarrow \infty} \frac{-1}{n} = 0.$$

2. The sequence $1, 2, 3, 4, 5, \dots$ is the simplest example of an increasing sequence that is not bounded above. It obviously diverges to ∞ .
3. Consider the recursively defined sequence $a_1 = 1$ and $a_{n+1} = \sqrt{3 + 2a_n}$.

We have seen this sequence before in Chapter 2, Example 2. As is suggested in diagram below can show by induction that $\{a_n\}$ is increasing and bounded above by 3.



Let $P(n)$ be the statement that

$$a_n < a_{n+1} < 3.$$

Since $a_1 = 1$ and $a_2 = \sqrt{5} < 3$ it follows that $P(1)$ is true.

Assume that $P(k)$ holds so that

$$a_k < a_{k+1} < 3.$$

Then we also have that

$$2a_k < 2a_{k+1} < 2 \cdot 3 = 6$$

and hence that

$$3 + 2a_k < 3 + 2a_{k+1} < 3 + 2 \cdot 3 = 9.$$

Since taking square roots preserves order we get that

$$\sqrt{3 + 2a_k} < \sqrt{3 + 2a_{k+1}} < \sqrt{3 + 2 \cdot 3} = \sqrt{9}.$$

This gives us that

$$a_{k+1} < a_{k+2} < 3$$

so that $P(k + 1)$ holds. By induction we have that

$$a_n < a_{n+1} < 3$$

for all $n \in \mathbb{N}$.

We have shown that $\{a_n\}$ is increasing and bounded above by 3. Therefore, the Monotone Convergence Theorem shows us that the sequence converges. However, while it is tempting to conclude that the limit of the sequence is 3 we do not yet know this for sure. As of this point we know that 3 is an upper bound for the sequence but not necessarily the least upper bound.

Let

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

We know from our arithmetic rules that if $a_n \geq 0$ and $\lim_{n \rightarrow \infty} a_n = L$, then $\lim_{n \rightarrow \infty} \sqrt{a_n} = \sqrt{L}$. We can use this property to show that

$$\begin{aligned} L &= \lim_{n \rightarrow \infty} a_{n+1} \\ &= \lim_{n \rightarrow \infty} \sqrt{3 + 2a_n} \\ &= \sqrt{3 + 2 \lim_{n \rightarrow \infty} a_n} \\ &= \sqrt{3 + 2L} \end{aligned}$$

Given that $L = \sqrt{3 + 2L}$, squaring both sides shows that

$$L^2 = 3 + 2L$$

or

$$L^2 - 2L - 3 = 0.$$

Factoring the left-hand side gives

$$(L + 1)(L - 3) = 0$$

so $L = -1$ or $L = 3$. However, all of the terms in the sequence are positive so the limit must be greater than or equal to 0. Therefore, $L = 3$ as expected.

4. Let

$$S_n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n}.$$

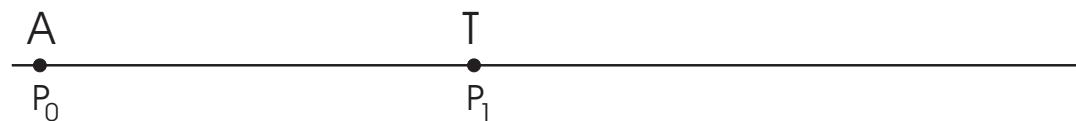
$\{S_n\}$ is called the sequence of partial sums. Since $S_{n+1} - S_n = \frac{1}{n+1} > 0$, this sequence is increasing. Unfortunately, it is not at all obvious whether or not the sequence $\{S_n\}$ is bounded. In fact we will soon see that it is not.

REMARK

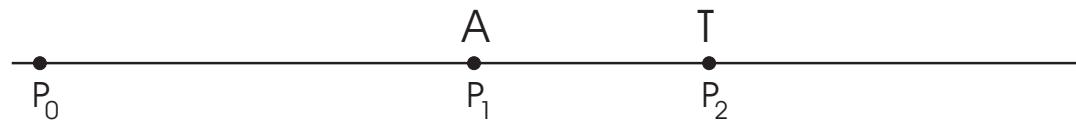
We have seen that the Monotone Convergence Theorem follows as a consequence of the Least Upper Bound Property. What is perhaps less obvious is that if we were to assume the Monotone Convergence Theorem as an axiom for the real numbers, then from this we could derive the Least Upper Bound Property. 

4.4 Introduction to Series

The Greek philosopher Zeno, who lived from 490-425 BC, proposed many paradoxes. The most famous of these is the *Paradox of Achilles and the Tortoise*. In this paradox, the great warrior Achilles is to race a tortoise. To make the race fair, Achilles (A) gives the tortoise (T) a substantial head start.



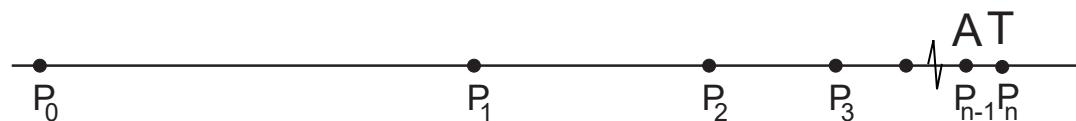
Zeno would argue that before Achilles could catch the tortoise, he must first go from his starting point at P_0 to that of the tortoise at P_1 . However, by this time the tortoise has moved forward to P_2 .



This time, before Achilles could catch the tortoise, he must first go from P_1 to where the tortoise was at P_2 . However, by the time Achilles completes this task, the tortoise has moved forward to P_3 .



Each time Achilles reaches the position that the tortoise had been, the tortoise has moved ahead further.



This process of Achilles trying to reach where the tortoise was going ad infinitum led Zeno to suggest that Achilles could *never catch the tortoise*.

Zeno's argument seems to be supported by the following observation:

Let t_1 denote the time it would take for Achilles to get from his starting point P_0 to P_1 . Let t_2 denote the time it would take for Achilles to get from P_1 to P_2 , and let t_3 denote the time it would take for Achilles to get from P_2 to P_3 . More generally, let t_n denote the time it would take for Achilles to get from P_{n-1} to P_n . Then the time it would take to catch the tortoise would be at least as large as the sum

$$t_1 + t_2 + t_3 + t_4 + \cdots + t_n + \cdots$$

of all of these infinitely many time periods.

Since each $t_n > 0$, Achilles is being asked to complete infinitely many tasks (each of which takes a positive amount of time) in a finite amount of time. It may seem that this is impossible. However, this is certainly a paradox because we know from our own experience that someone as swift as Achilles will eventually catch and even pass the tortoise. Hence, the sum

$$t_1 + t_2 + t_3 + t_4 + \cdots + t_n + \cdots$$

must be *finite*.

This statement brings into question the following very fundamental problem:

Problem: Given an *infinite* sequence $\{a_n\}$ of real numbers, what do we mean by the sum

$$a_1 + a_2 + a_3 + a_4 + \cdots + a_n + \cdots ?$$

To see why this is an issue, consider the following example:

EXAMPLE 17 Let $a_n = (-1)^{n-1}$. Consider

$$1 + (-1) + 1 + (-1) + 1 + (-1) + 1 + (-1) + \cdots .$$

If we want to find this sum, we could try to use the associative property of finite sums and group the terms as follows:

$$[1 + (-1)] + [1 + (-1)] + [1 + (-1)] + [1 + (-1)] + \cdots .$$

This would give

$$0 + 0 + 0 + 0 + \cdots$$

which must be 0. Therefore, we might expect that

$$1 + (-1) + 1 + (-1) + 1 + (-1) + 1 + (-1) + \cdots = 0.$$

This makes sense since *there appears to be the same number of 1's and -1's*, so cancellation should make the sum 0.

However, if we choose to group the terms differently,

$$1 + [(-1) + 1] + [(-1) + 1] + [(-1) + 1] + [(-1) + 1] + \cdots ,$$

then we get

$$1 + 0 + 0 + 0 + 0 + \cdots = 1.$$

Both methods seem to be equally valid so we cannot be sure of what the real sum should be. It seems that *the usual rules of arithmetic do not hold for infinite sums*. We must look for an alternate approach.



Since finite sums behave very well, we might try adding up all of the terms up to a certain cut-off k and then see if a pattern develops as k gets very large. This is in fact what we will do.

DEFINITION

Series

Given a sequence $\{a_n\}$, the *formal* sum

$$a_1 + a_2 + a_3 + a_4 + \cdots + a_n + \cdots$$

is called a *series*. (The series is called *formal* because we have not yet given it a meaning numerically.)

The a_n 's are called the *terms* of the series. For each term a_n , n is called the *index* of the term.

We will denote the series by

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

For each k , we define the k -th *partial sum* S_k by

$$S_k = \sum_{n=1}^k a_n.$$

We say that the series $\sum_{n=1}^{\infty} a_n$ *converges* if the sequence $\{S_k\}$ of partial sums converges. In this case if $L = \lim_{k \rightarrow \infty} S_k$, then we write

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

and assign the sum this value. Otherwise, we say that the series $\sum_{n=1}^{\infty} a_n$ *diverges*.

Note that all of the series we have listed so far have started with the first term indexed by 1. This is not necessary. In fact, it is quite common for a series to begin with the initial index being 0. In fact, the series can start at any initial point.

final index

$$\sum_{n=j}^{\infty} a_n = a_j + a_{j+1} + a_{j+2} + a_{j+3} + \dots$$

initial index

We can apply the previous definitions to the first series that we considered in this section.

EXAMPLE 18 Let $a_n = (-1)^{n-1}$. Then

$$\begin{aligned} S_1 &= a_1 & &= 1 \\ S_2 &= a_1 + a_2 &= S_1 + a_2 &= S_1 - 1 &= 0 \\ S_3 &= a_1 + a_2 + a_3 &= S_2 + a_3 &= S_2 + 1 &= 1 \\ S_4 &= a_1 + a_2 + a_3 + a_4 &= S_3 + a_4 &= S_3 - 1 &= 0 \\ S_5 &= a_1 + a_2 + a_3 + a_4 + a_5 &= S_4 + a_5 &= S_4 + 1 &= 1 \\ &\vdots \end{aligned}$$

Therefore,

$$S_k = \begin{cases} 0 & \text{if } k \text{ is even} \\ 1 & \text{if } k \text{ is odd} \end{cases}$$

This clearly shows that $\{S_k\}$ diverges, and hence so does $\sum_{n=1}^{\infty} (-1)^{n-1}$.



EXAMPLE 19 Determine if the series

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

converges or diverges.

Observe that

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

Moreover, we can write

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

Therefore the series becomes

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

To calculate S_k note that

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

If we re-group the terms in the last expression, we get

$$\begin{aligned} S_k &= \left(1 - \frac{1}{2} \right) + \left(\frac{1}{2} - \frac{1}{3} \right) + \left(\frac{1}{3} - \frac{1}{4} \right) + \left(\frac{1}{4} - \frac{1}{5} \right) + \cdots + \left(\frac{1}{k} - \frac{1}{k+1} \right) \\ &= 1 - \left(\frac{1}{2} - \frac{1}{2} \right) - \left(\frac{1}{3} - \frac{1}{3} \right) - \left(\frac{1}{4} - \frac{1}{4} \right) - \left(\frac{1}{5} - \frac{1}{5} \right) - \cdots - \left(\frac{1}{k} - \frac{1}{k} \right) - \frac{1}{k+1} \\ &= 1 - 0 - 0 - 0 - \cdots - 0 - \frac{1}{k+1} \\ &= 1 - \frac{1}{k+1} \end{aligned}$$

Then

$$\lim_{k \rightarrow \infty} S_k = \lim_{k \rightarrow \infty} \left(1 - \frac{1}{k+1} \right) = 1.$$

This shows that the series $\sum_{n=1}^{\infty} \frac{1}{n^2+n}$ converges and that

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



What is actually remarkable about the series above is not that we were able to show that it converges, but rather that we could find its value so easily. Generally, this will not be the case. In fact, even if we know a series converges, it may be very difficult or even impossible to determine the exact value of its sum. In most cases, we will have to be content with either showing that a series converges or that it diverges and, in the case of a convergent series, estimating its sum.

The next section deals with an important class of series known as *geometric series*. Not only can we easily determine if such a series converges, but we can easily find the sum.

4.4.1 Geometric Series

Perhaps the most important type of series are the geometric series.

DEFINITION**Geometric Series**

A *geometric series* is a series of the form

$$\sum_{n=0}^{\infty} r^n = 1 + r + r^2 + r^3 + r^4 + \dots$$

The number r is called the *ratio* of the series.

If $r = (-1)$, the series is

$$\sum_{n=0}^{\infty} (-1)^n = 1 + (-1) + 1 + (-1) + 1 + (-1) + 1 + (-1) + \dots$$

which we have already seen diverges.

If $r = 1$, the series is

$$\sum_{n=0}^{\infty} 1^n = 1 + 1 + 1 + 1 + 1 + \dots$$

which again diverges since $S_k = \sum_{n=0}^k 1^n = k + 1$ diverges to ∞ .

Question: Which if any of the geometric series converge?

Assume that $r \neq 1$. Let

$$S_k = 1 + r + r^2 + r^3 + r^4 + \dots + r^k.$$

Then

$$\begin{aligned} rS_k &= r(1 + r + r^2 + r^3 + r^4 + \dots + r^k) \\ &= r + r^2 + r^3 + r^4 + \dots + r^{k+1}. \end{aligned}$$

Therefore

$$\begin{aligned} S_k - rS_k &= (1 + r + r^2 + r^3 + r^4 + \dots + r^k) - (r + r^2 + r^3 + r^4 + \dots + r^k + r^{k+1}) \\ &= 1 - r^{k+1}. \end{aligned}$$

Hence

$$(1 - r)S_k = S_k - rS_k = 1 - r^{k+1}$$

and since $r \neq 1$,

$$S_k = \frac{1 - r^{k+1}}{1 - r}.$$

The only term in this expression that depends on k is r^{k+1} , so $\lim_{k \rightarrow \infty} S_k$ exists if and only if $\lim_{k \rightarrow \infty} r^{k+1}$ exists. However, if $|r| < 1$, then r^{k+1} becomes very small for large k . That is $\lim_{k \rightarrow \infty} r^{k+1} = 0$.

If $|r| > 1$, then $|r^{k+1}|$ becomes very large as k grows. That is, $\lim_{k \rightarrow \infty} |r^{k+1}| = \infty$. Hence, $\lim_{k \rightarrow \infty} r^{k+1}$ does not exist.

Finally, if $r = -1$, then r^{k+1} alternates between 1 and -1 , so it again diverges. This shows that r^{k+1} , and hence the series $\sum_{n=0}^{\infty} r^n$, will converge if and only if $|r| < 1$. Moreover, in this case,

$$\begin{aligned}\lim_{k \rightarrow \infty} S_k &= \lim_{k \rightarrow \infty} \frac{1 - r^{k+1}}{1 - r} \\ &= \frac{1 - \lim_{k \rightarrow \infty} r^{k+1}}{1 - r} \\ &= \frac{1}{1 - r}\end{aligned}$$

THEOREM 10

Geometric Series Test

The geometric series $\sum_{n=0}^{\infty} r^n$ converges if $|r| < 1$ and diverges otherwise.

If $|r| < 1$, then

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

EXAMPLE 20

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

SOLUTION This is a geometric series with ratio $r = \frac{1}{2}$. Since $0 < \frac{1}{2} < 1$, the Geometric Series Test shows that $\sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n$ converges. Moreover,

$$\begin{aligned}\sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n &= \frac{1}{1 - \frac{1}{2}} \\ &= 2.\end{aligned}$$

4.4.2 Divergence Test

It makes sense that if we are to add together infinitely many positive numbers and get something finite, then the terms must eventually be small. We will now see that this statement holds for any convergent series.

THEOREM 11**Divergence Test**

Assume that $\sum_{n=1}^{\infty} a_n$ converges. Then

$$\lim_{n \rightarrow \infty} a_n = 0.$$

Equivalently, if $\lim_{n \rightarrow \infty} a_n \neq 0$ or if $\lim_{n \rightarrow \infty} a_n$ does not exist, then $\sum_{n=1}^{\infty} a_n$ diverges.

The Divergence Test gets its name because it can identify certain series as being divergent, but **it cannot show that a series converges**.

EXAMPLE 21

Consider the geometric series $\sum_{n=0}^{\infty} r^n$ with $|r| \geq 1$. Then $\lim_{n \rightarrow \infty} r^n = 1$ if $r = 1$ and it does not exist for all other r with $|r| \geq 1$ (i.e., if $r = -1$ or if $|r| > 1$). The Divergence Test shows that if $|r| \geq 1$, then $\sum_{n=0}^{\infty} r^n$ diverges. 

The Divergence Test works for the following reason. Assume that $\sum_{n=1}^{\infty} a_n$ converges to L . This is equivalent to saying that

$$\lim_{k \rightarrow \infty} S_k = L.$$

By the basic properties of convergent sequences, we get that

$$\lim_{k \rightarrow \infty} S_{k-1} = L$$

as well.

However, for $k \geq 2$,

$$\begin{aligned} S_k - S_{k-1} &= \sum_{n=1}^k a_n - \sum_{n=1}^{k-1} a_n \\ &= (a_1 + a_2 + a_3 + a_4 + \cdots + a_{k-1} + a_k) - (a_1 + a_2 + a_3 + a_4 + \cdots + a_{k-1}) \\ &= a_k \end{aligned}$$

Therefore,

$$\begin{aligned} \lim_{k \rightarrow \infty} a_k &= \lim_{k \rightarrow \infty} (S_k - S_{k-1}) \\ &= \lim_{k \rightarrow \infty} S_k - \lim_{k \rightarrow \infty} S_{k-1} \\ &= L - L \\ &= 0. \end{aligned}$$

EXAMPLES

1. Consider the sequence $\{\frac{n}{n+1}\}$. Then

$$\lim_{n \rightarrow \infty} \frac{n}{n+1} = 1.$$

Therefore, the Divergence Test shows that

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

diverges.

2. While it is difficult to do so, it is possible to show that

$$\lim_{n \rightarrow \infty} \sin(n)$$

does not exist. Therefore, the Divergence Test shows that the series

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

diverges.

3. The Divergence Test shows that if either $\lim_{n \rightarrow \infty} a_n \neq 0$ or if $\lim_{n \rightarrow \infty} a_n$ does not exist, then $\sum_{n=1}^{\infty} a_n$ diverges. It would seem natural to ask if the *converse* statement holds. That is:

Question: If $\lim_{n \rightarrow \infty} a_n = 0$, does this mean that $\sum_{n=1}^{\infty} a_n$ converges?

Let $a_n = \frac{1}{n}$. Let

$$S_k = \sum_{n=1}^k \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots + \frac{1}{k}.$$

Then

$$\begin{aligned}
 S_1 &= 1 \\
 S_2 &= 1 + \frac{1}{2} \\
 S_4 &= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} \\
 &= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) \\
 &> 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) \\
 &= 1 + \frac{1}{2} + \frac{1}{2} \\
 &= 1 + \frac{2}{2} \\
 S_8 &= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} \\
 &= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) \\
 &> 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) \\
 &= 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} \\
 &= 1 + \frac{3}{2} \\
 &\vdots
 \end{aligned}$$

We have seen that

$$\begin{aligned}
 S_1 &= S_{2^0} = 1 + \frac{0}{2} \\
 S_2 &= S_{2^1} = 1 + \frac{1}{2} \\
 S_4 &= S_{2^2} > 1 + \frac{2}{2} \\
 S_8 &= S_{2^3} > 1 + \frac{3}{2} \\
 &\vdots
 \end{aligned}$$

A pattern has emerged. In general, we can show that for any m

$$S_{2^m} \geq 1 + \frac{m}{2}.$$

However, the sequence $1 + \frac{m}{2}$ grows without bounds. It follows that the partial sums of the form S_{2^m} also grow without bound. This shows that the series $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges to ∞ as well. 

This example shows that even if $\lim_{n \rightarrow \infty} a_n = 0$, it is still possible for $\sum_{n=1}^{\infty} a_n$ to diverge!!!

Note:

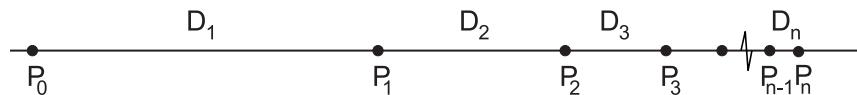
1. The sequence $\{\frac{1}{n}\}$ was first studied in detail by Pythagoras who felt that these ratios represented *musical harmony*. For this reason the sequence $\{\frac{1}{n}\}$ is called *the Harmonic Progression* and the series $\sum_{n=1}^{\infty} \frac{1}{n}$ is called *the Harmonic Series*.

We have just shown that the Harmonic Series $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges to ∞ . However, the argument to do this was quite clever. Instead, we might ask if we could use a computer to add up the first k terms for some large k and show that the sums are getting large? In this regard, we may want to know how many terms it would take so that

$$S_k = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots + \frac{1}{k} > 100?$$

The answer to this question is very surprising. It can be shown that k must be at least 10^{30} , which is an enormous number. No modern computer could ever perform this many additions!!!

2. Recall that in Zeno's paradox, Achilles had to travel infinitely many distances in a finite amount of time to catch the tortoise. If D_n represents the distance between points P_{n-1} (where Achilles is after $n-1$ steps) and P_n (where the tortoise is currently located), then the D_n 's are becoming progressively smaller.



If t_n is the time it takes Achilles to cover the distance D_n , then the t_n 's are also becoming progressively smaller. In fact, they are so small that $\lim_{n \rightarrow \infty} t_n = 0$ and indeed it is reasonable to assume that

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

converges! This is how we can resolve Zeno's paradox.

4.5 Bolzano-Weierstrass Theorem

We have seen that every convergent sequence is bounded and that for monotonic sequences boundedness will also imply convergence. However, it is easy to see that boundedness will not imply convergence without some additional conditions since the sequence

$$\{(-1)^{n+1}\} = \{1, -1, 1, -1, \dots\}$$

is clearly bounded but does not converge. It is worth noting however that this sequence does have a convergent subsequence namely

$$\{a_{2k-1}\} = \{1, 1, 1, \dots\}.$$

In fact this sequence has infinitely many convergent subsequences.

In this section we will consider the following question:

Question: Under what conditions can we guarantee that a sequence $\{a_n\}$ will have a convergent subsequence. ◀

Of course it would also make sense to ask if all sequences have convergent subsequences. The first of the two examples below shows this is not the case.

EXAMPLE 22

The sequence $a_n = n$ has no convergent subsequences since every subsequence $\{n_k\}$ is unbounded. ◀

EXAMPLE 23

Since convergent sequences are bounded we might be tempted to think that only bounded sequences could have convergent subsequences. However, the next sequence

$$\{n^{(-1)^{n+1}}\} = \{1, \frac{1}{2}, 3, \frac{1}{4}, 5, \frac{1}{6}, \dots\}$$

is unbounded since the odd index terms grow without bound, but

$$\{a_{2k}\} = \{\frac{1}{2}, \frac{1}{4}, \dots, \frac{1}{2k}, \dots\}$$

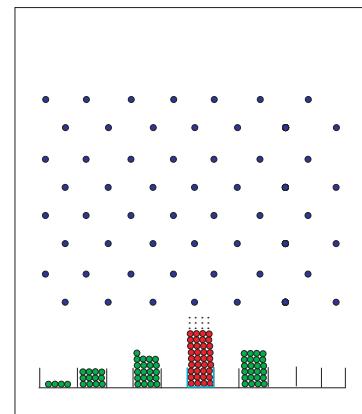
converges to 0. ◀

At this point the best that we can say is that for $\{a_n\}$ to have a convergent subsequence, it must at the very least have a bounded subsequence. We will show that in fact that having a bounded subsequence is sufficient to insure that $\{a_n\}$ also has a convergent subsequence by showing that every bounded sequence has a convergent subsequence.

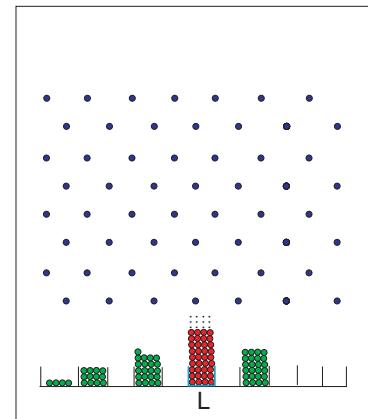
REMARK

Why is claim that every bounded sequence has a convergent subsequence reasonable? To see why this might be lets assume that L is the limit of such a subsequence. Then for every open interval I containing L , no matter how small there must be infinitely many terms from our sequence within I . It is this *clustering* of any term in the sequence that is characteristic of a convergent subsequence.

Suppose that $\{a_n\}$ is bounded and that $\{a_n\} \subset [-M, M]$. If we cut the interval $[-M, M]$ up into a large number of small subintervals, then an infinite version of the pigeon hole principle tells us that one of these subintervals must contain infinitely many terms of our sequence.



If we take one such subinterval and break it up into even finer subintervals we must again have one of these that contains infinitely many terms in our sequence. Repeating this process over and over will eventually lead to us finding a point L with the property that no matter how small the subinterval around L might be it still contains infinitely many terms in our sequence. This point L is a candidate for the limit of a subsequence.

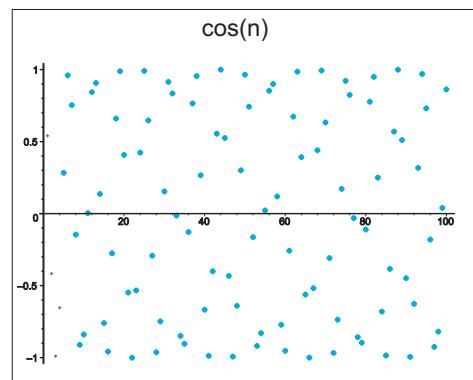


How can we show analytically that every bounded sequence has a convergent subsequence?

Strategy: Suppose we could show that every sequence $\{a_n\}$ has a monotonic subsequence $\{a_{n_k}\}$. If $\{a_n\}$ is assumed to be bounded, then $\{a_{n_k}\}$ would be both bounded and monotonic. We could then apply the Monotone Convergence Theorem to show that $\{a_{n_k}\}$ converges.

REMARK

Looking at the graph of the sequence $\{\cos(n)\}$ it is not at all obvious that this sequence has a monotonic subsequence.



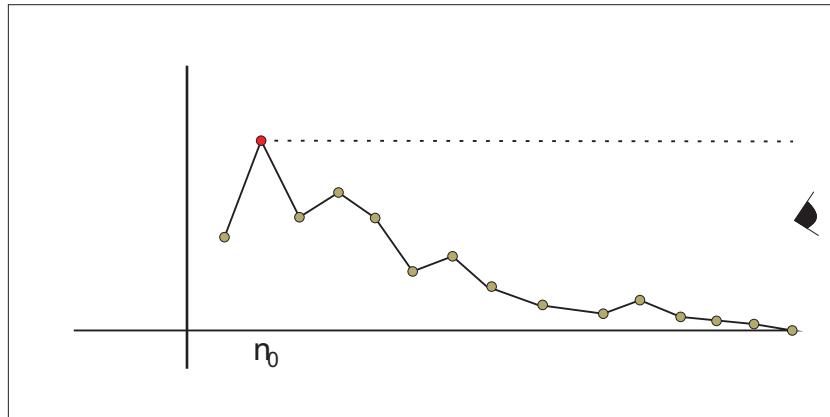
However, we will in fact show that this is the case. To do so we must first introduce some terminology.

DEFINITION**Peak Points**

Let $\{a_n\}$ be a sequence. Then the index $n_0 \in \mathbb{N}$ is called a *peak point* for the sequence $\{a_n\}$ if

$$a_k < a_{n_0}$$

for every $k > n_0$.

**THEOREM 12****The Peak Point Lemma**

Every sequence $\{a_n\}$ has a monotonic subsequence $\{a_{n_k}\}$.

PROOF

Let \mathbb{P} be the collection of peak points.

Case 1: \mathbb{P} is infinite.

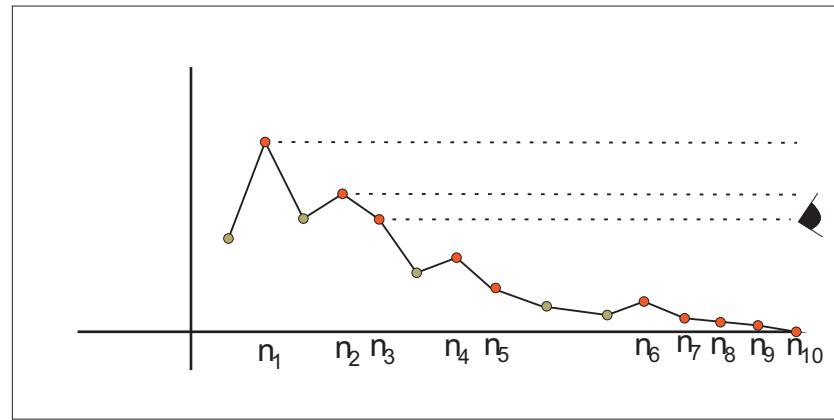
Let n_1 be the least element of \mathbb{P} . Since \mathbb{P} is infinite $\mathbb{P} \setminus \{n_1\}$ is non-empty. Let n_2 be the least element of $\mathbb{P} \setminus \{n_1\}$. Next let n_3 be the least element of $\mathbb{P} \setminus \{n_1, n_2\}$.

Given peak points $n_1 < n_2 < \dots < n_k$, let n_{k+1} be the least element of $\mathbb{P} \setminus \{n_1, n_2, \dots, n_k\}$. This process allows us to recursively define a sequence

$$n_1 < n_2 < \dots < n_k < n_{k+1} < \dots$$

of peak points for $\{a_n\}$. It follows that $\{a_{n_k}\}$ is decreasing because n_k being a peak point implies that

$$a_{n_k} > a_{n_{k+1}}.$$

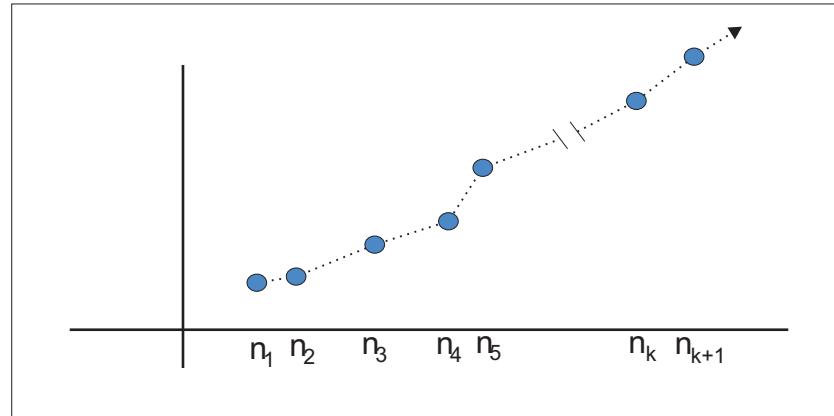


Case 2: \mathbb{P} is finite.

Let n_1 be larger than any peak point. Since n_1 is not a peak point there is an $n_1 < n_2$ such that $a_{n_1} \leq a_{n_2}$. Since n_2 is not a peak point there is an $n_1 < n_2 < n_3$ such that $a_{n_1} \leq a_{n_2} \leq a_{n_3}$.

Proceeding this way gives $n_1 < n_2 < n_3 < \dots < n_k < n_{k+1} < \dots$ with

$$a_{n_1} \leq a_{n_2} \leq a_{n_3} \leq \dots \leq a_{n_k} \leq a_{n_{k+1}} \leq \dots$$



Summary: If $\{a_n\}$ has infinitely many peak points, then $\{a_n\}$ has a decreasing subsequence while if it has only finitely many peak points there will be a non-decreasing subsequence. This proves that every sequence $\{a_n\}$ does indeed have a monotonic subsequence. ■

THEOREM 13

Bolzano-Weierstrass Theorem (BWT)

Every bounded sequence has a convergent subsequence.

PROOF

Assume that $\{a_n\}$ is a bounded sequence. By the Peak point Lemma, $\{a_n\}$ has a monotonic subsequence $\{a_{n_k}\}$. However, $\{a_{n_k}\}$ is also bounded and as a result the Monotone Convergence Theorem shows that $\{a_{n_k}\}$ converges. ■

4.6 Limit Points

We know that if a sequence converges to L , then every subsequence converges to L as well. We have also just seen that even non-convergent sequences can have subsequences which converge. This leads us to make the following definition.

DEFINITION**Limit Points of a Sequence**

An $\alpha \in \mathbb{R}$ is called a *limit point of* $\{a_n\}$ if there is a subsequence $\{a_{n_k}\}$ of $\{a_n\}$ such that

$$\lim_{k \rightarrow \infty} a_{n_k} = \alpha.$$

We denote the set of limit points of $\{a_n\}$ by $LIM(\{a_n\})$.

REMARK

The Bolzano-Weierstrass Theorem shows that every bounded sequence has at least one limit point. ◀

EXAMPLE 24

i) If $\lim_{n \rightarrow \infty} a_n = L$, then $LIM(\{a_n\}) = \{L\}$.

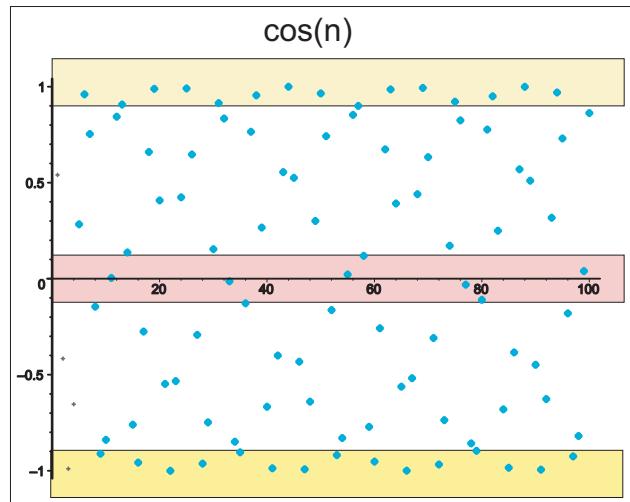
ii) $LIM(\{1, \frac{1}{2}, 3, \frac{1}{4}, \dots, n^{(-1)^{n+1}}, \dots\}) = \{0\}$

Despite having a unique limit point this sequence diverges! This shows that having a unique limit point is not sufficient to prove the sequence converges. However, it can be shown that if the sequence is bounded it is indeed the case that $\{a_n\}$ converges if and only if it has a unique limit point.

iii) $LIM(\{1, -1, 1, -1, \dots\}) = \{-1, 1\}$.

iv) $LIM(\{1, 2, 3, 4, \dots\}) = \emptyset$.

v) Consider the sequence $\{\cos(n)\}$. what are the limit points?



A look at the graph of the sequence $\{\cos(n)\}$ shows a significant amount of clustering around -1 and 1 suggesting that both values could be limit points of the sequence. The clustering around 0 is less pronounced. But it turns out that 0 is indeed a limit point. In fact with some work one can show that

$$LIM(\{\cos(n)\}) = [-1, 1].$$



We end this section with the following question:

Question: Does there exist a sequence $\{a_n\}$ such that

$$LIM(\{a_n\}) = \mathbb{R}?$$

That is, can we find a sequence so that no matter what $\alpha \in \mathbb{R}$ we choose, there will be a subsequence $\{a_{n_k}\}$ of $\{a_n\}$ converging to α ? ◀

4.7 Cauchy Sequences

Aside from monotonic sequences at this point if we want to show that a sequence converges we first guess what its limit might be, and then proceed with an ϵ - δ argument to show that the sequence does indeed converge to this value. One of the most useful aspects of the Monotone Convergence Theorem is that it often allows us to show that a sequence converges without explicitly knowing what the value of the limit might be. This means that we can determine convergence *intrinsically* from the data contained within the sequence alone. We do not have to look outside and first identify a candidate for the limit before concluding convergence. In this section we will try to find an intrinsic characterization of convergence that can be used for all sequences rather than just monotonic ones. To see what such a characterization

might look like let's make the following observation about the nature of a convergent sequence:

Key Observation: Assume that $\{a_n\}$ converges with limit L . We know then that if we are far enough out in the tail of the sequence, then all of the terms will be close to L . If we take two such terms, a_n and a_m , then since both are close to L , they must also be close to one another. To make this more precise we let $\epsilon > 0$. We can now find a cutoff $N \in \mathbb{N}$ so that if $k \geq N$, then

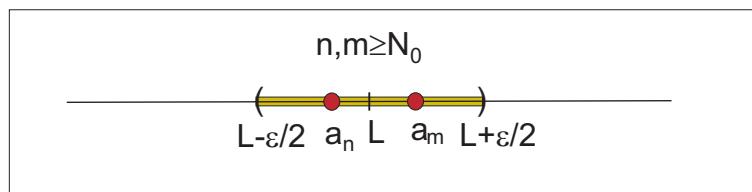
$$|a_k - L| < \frac{\epsilon}{2}.$$

Now let $n, m \geq N$. Then the Triangle Inequality shows that

$$|a_n - a_m| \leq |a_n - L| + |L - a_m| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

That is, if $\{a_n\}$ converges, then for every $\epsilon > 0$, we can find a cutoff $N \in \mathbb{N}$ such that if $n, m \geq N$, we have

$$|a_n - a_m| < \epsilon.$$



We have just seen that convergent sequences have tails that cluster together as closely as we would like. This clustering of the tail gives us a candidate for our intrinsic characterization of convergence.

DEFINITION

Cauchy Sequence

We say that a sequence $\{a_n\}$ is *Cauchy* if for every $\epsilon > 0$, there exists some $N \in \mathbb{N}$ such that if $m, n \geq N$, then $|a_n - a_m| < \epsilon$.

We have established the following proposition:

PROPOSITION 14

Every convergent sequence $\{a_n\}$ is Cauchy.

We are left to ask:

Fundamental Question: Does every Cauchy sequence converge? ◀

Strategy: We will answer this question in two stages. In fact we will show:

- 1) Every Cauchy sequence $\{a_n\}$ is bounded.

This will allow us to apply the Bolzano-Weierstrass Theorem to conclude that $\{a_n\}$ has a convergent subsequence $\{a_{n_k}\}$. Of course for a generic sequence having a convergent subsequence is not enough to show that the sequence converges. However, we will also show that Cauchy sequences have the following rather remarkable property:

- 2) If $\{a_n\}$ is a Cauchy sequence with a subsequence $\{a_{n_k}\}$ converging to L , then $\{a_n\}$ also converges to L .

THEOREM 15

Boundedness of Cauchy Sequences

Every Cauchy sequence is bounded.

PROOF

The proof of this result is very similar to the proof that convergent sequences are bounded, except we cannot focus our attention on the limit point L . So we need to manufacture a focal point by making use of the Cauchy criteria.

Choose an N_0 such that if $n, m \geq N_0$ then

$$|a_n - a_m| < 1.$$

If $n \geq N_0$, then

$$|a_n - a_{N_0}| < 1$$

and this in turn implies that

$$|a_n| < |a_{N_0}| + 1.$$

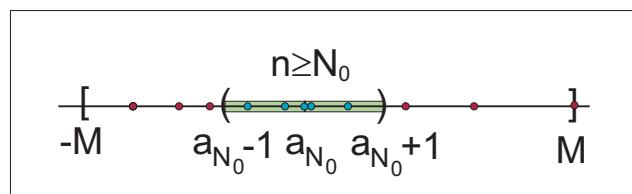
Let

$$M = \max\{|a_1|, |a_2|, \dots, |a_{N_0-1}|, |a_{N_0}| + 1\}.$$

Hence

$$|a_n| \leq M$$

for all $n \in \mathbb{N}$.



We now show that if a Cauchy sequence has a convergent subsequence, then it must itself be convergent.

PROPOSITION 16

Let $\{a_n\}$ be Cauchy. Assume that $\{a_n\}$ has a subsequence $\{a_{n_k}\}$ that converges to L . Then $\{a_n\}$ also converges to L .

PROOF

Outline of the Proof:

1. Choose N_0 so that $n, m \geq N_0$ implies that $|a_n - a_m| < \frac{\epsilon}{2}$.
2. Show that for this choice of cutoff N_0 that if $n \geq N_0$, then we have

$$|a_n - L| < \epsilon$$

as desired.

Let's assume that $\{a_n\}$ is Cauchy and that it has a subsequence $\{a_{n_k}\}$ that converges to L . Given $\epsilon > 0$, we can choose $N_0 \in \mathbb{N}$ so that $n, m \geq N_0$ implies that

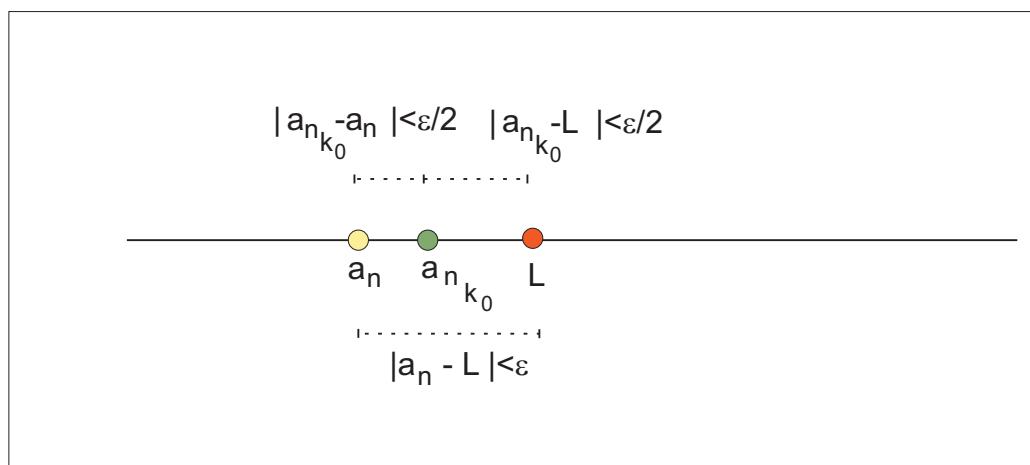
$$|a_n - a_m| < \frac{\epsilon}{2}.$$

We now use the fact that we have a convergent subsequence. Since $\{a_{n_k}\}$ converges to L , we can find a k_0 so that $n_{k_0} > N_0$ and

$$|a_{n_{k_0}} - L| < \frac{\epsilon}{2}.$$

Finally, choose any $n \geq N_0$. The key is that any such n , we have that a_n must be very close to $a_{n_{k_0}}$ which is in turn very close to L . More precisely, if $n \geq N_0$, then

$$|a_n - L| \leq |a_n - a_{n_{k_0}}| + |a_{n_{k_0}} - L| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$



This shows that

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

We are now in a position to show that all Cauchy Sequences converge.

THEOREM 17

Completeness Theorem for \mathbb{R}

Every Cauchy sequence $\{a_n\}$ converges.

PROOF

If $\{a_n\}$ is Cauchy, then $\{a_n\}$ is bounded. By the Bolzano-Weierstrass Theorem, $\{a_n\}$ has a convergent subsequence $\{a_{n_k}\}$. Therefore, $\{a_n\}$ also converges. ■

REMARK

The Completeness Theorem shows that a sequence in \mathbb{R} converges if and only if it is Cauchy. This theorem is far more profound than it might seem. In fact the following are logically equivalent.

1. LUBP: Every non-empty bounded subset of \mathbb{R} has a least upper bound.
2. MCT: Every bounded monotonic sequence converges.
3. BWT: Every bounded sequence has a convergent subsequence.
4. Every real number has a decimal expansion.
5. Completeness Theorem: Every Cauchy sequence converges.

All of these are often referred to as the **Completeness Property** for \mathbb{R} . ◀

REMARK

A common mistake is to assume that if $\{a_n\}$ is such that

$$\lim_{n \rightarrow \infty} a_{n+1} - a_n = 0,$$

then $\{a_n\}$ is Cauchy. While every Cauchy sequence has this property, this does not imply that the sequence is Cauchy as the next example shows.

EXAMPLE 25

Let

$$a_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}.$$

Then

$$\begin{aligned}a_{n+1} - a_n &= (1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} + \frac{1}{n+1}) - (1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}) \\&= \frac{1}{n+1} \\&\rightarrow 0\end{aligned}$$

but $\{a_n\}$ diverges to ∞ so $\{a_n\}$ is not Cauchy.



Chapter 5

Limits and Continuity

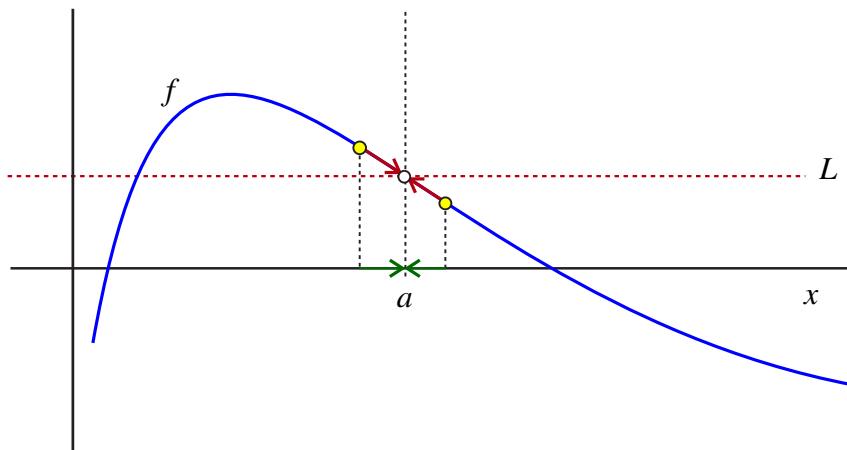
Previously we studied what was meant by the limit of a sequence. Now we will focus on *limits for functions*, something that you should be familiar with from your previous Calculus course. It is important to note that these ideas are actually very similar in flavour and we will be able to use what we have learned about limits of sequences to give us a better understanding of limits for functions.

5.1 Introduction to Limits for Functions

Let's begin by stating the common heuristic definition of a limit.

Heuristic Definition of the Limit of a Function at a Point $x = a$

We say that L is the limit of a function f as x approaches a if, as x gets closer and closer to a without ever reaching a , $f(x)$ gets closer and closer to L . 



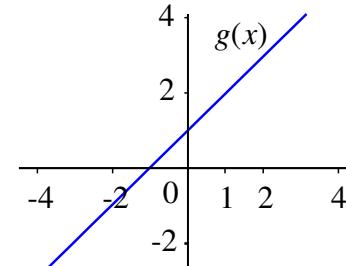
We will examine this definition with the following example:

EXAMPLE 1 Consider the two functions $f(x) = \frac{x^2-1}{x-1}$ and $g(x) = x + 1$. Factoring $x^2 - 1$, we get $x^2 - 1 = (x + 1)(x - 1)$. It might be tempting to use a little algebra to write

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

Does this mean that f and g are actually the same function? The answer is *almost*, but not quite. It is true that provided $x \neq 1$, the two functions will assign x to the same value. However, you will notice that $f(x)$ is not defined at $x = 1$, whereas $g(x)$ is defined at this point. This means that the two functions have *different domains* and that is enough to make them *different functions*.

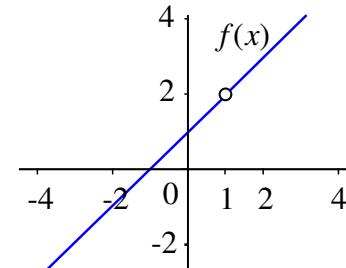
The graph of $g(x) = x + 1$ is a straight line with slope 1.



What happens if we graph $f(x) = \frac{x^2-1}{x-1}$?

We would get the same picture as the graph of $g(x) = x + 1$, except there is a *hole* in the graph corresponding to where $x = 1$.

(**Note:** When drawing a graph of this type, it is common practise to exaggerate the hole with a hollow circle.)



We want to focus on the values of $f(x)$ when x is very close to, but not equal to, 1.

The following is a table of some select values of $f(x)$ with x near 1.

x	$f(x)$
0	1
0.1	1.1
0.5	1.5
0.75	1.75
0.9	1.9
0.99	1.99
0.999	1.999
0.9999	1.9999
0.99999999	1.99999999

x	$f(x)$
2	3
1.9	2.9
1.5	2.5
1.25	2.25
1.1	2.1
1.01	2.01
1.001	2.001
1.00001	2.00001
1.00000001	2.00000001

We can see from the table of values that as x gets very close to 1, then $f(x)$ gets very close to 2. So we would like to say that 2 is the limit of $f(x)$ as x approaches 1. 

As was the case with sequences, our heuristic definition of limits for functions lacks precision. In fact, in our previous example, it is important to note that we can actually get as close as we like to 2 provided that we choose x close enough to, but not equal to, 1. (Remember, $f(x)$ is not defined at $x = 1$.) How can we quantify this statement?

To answer this question we will proceed in a manner similar to our study of limits of sequences. We would like to be able to specify a tolerance $\epsilon > 0$ and show that as x gets close enough to 1, the values of $f(x)$ approximate our limit 2 within our tolerance. In the case of sequences, we had to present a cutoff $N \in \mathbb{N}$ so that for any $n \geq N$, our term a_n was within ϵ of our limit. This time we need to present a distance $\delta > 0$ so that if the distance from x to 1 is less than the distance δ , and if $x \neq 1$, then we would have that $f(x)$ approximates our limit 2 within our tolerance.

That is, if $0 < |x - 1| < \delta$, then $|f(x) - 2| < \epsilon$.

In fact, in this section we will show that for any tolerance $\epsilon > 0$, if we let $\delta = \epsilon$, then for any x with $0 < |x - 1| < \delta$, we would have $|f(x) - 2| < \epsilon$. (This works because $y = x + 1$ represents a line with slope $m = 1$.) As such the function $f(x) = \frac{x^2 - 1}{x - 1}$ has a limit of 2 as x approaches 1.

NOTE

In this course, δ is the Greek letter *delta* and it will be used to represent a cutoff distance in the definition of limits. It plays a similar role as the cutoff number N in the definition of the limit of a sequence. 

With the previous example in mind, we will now present a more precise definition of the notion of a *limit of a function at a point $x=a$* .

DEFINITION

Limit of a Function at a Point $x = a$

Let f be a function and let $a \in \mathbb{R}$. We say that f has a limit L as x approaches a , or that L is the limit of $f(x)$ at $x = a$, if for any positive tolerance $\epsilon > 0$, we can find a cutoff distance $\delta > 0$ such that if the distance from x to a is less than δ , and if $x \neq a$, then $f(x)$ approximates L with an error less than ϵ .

That is, if $0 < |x - a| < \delta$, then $|f(x) - L| < \epsilon$.

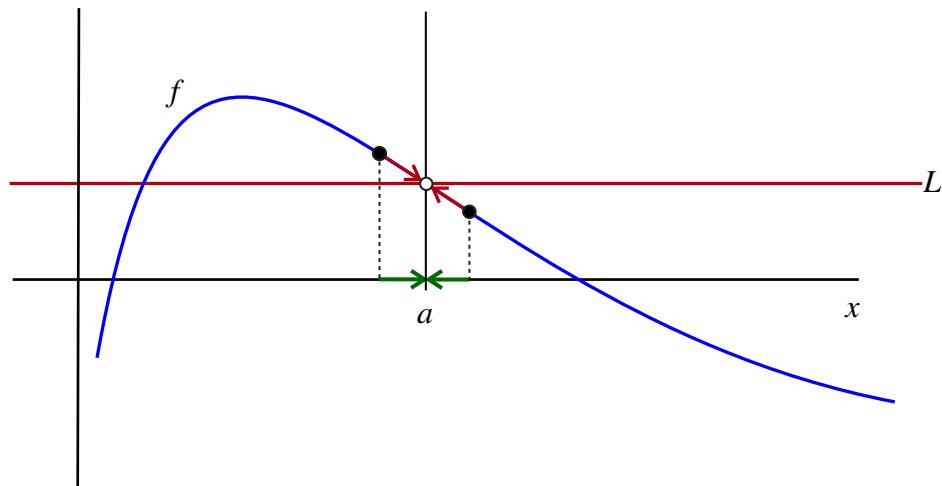
In this case, we write

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

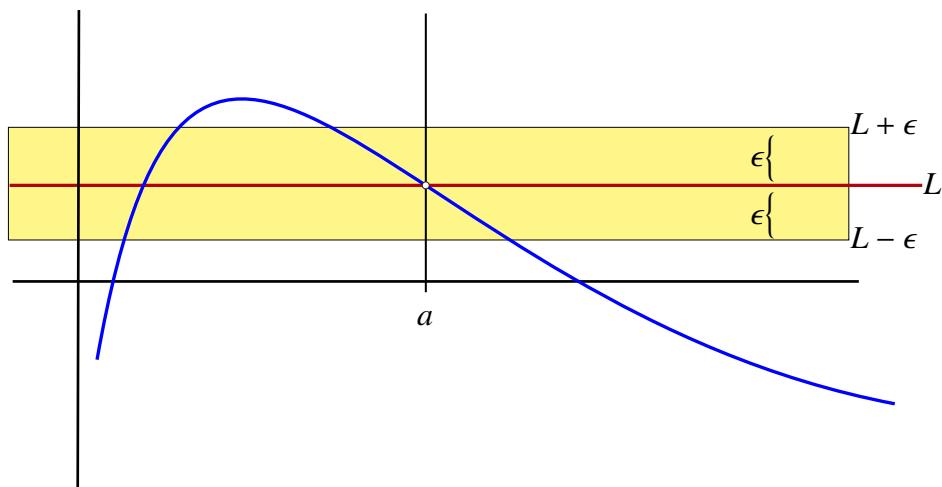
(Note: We sometimes write $x \rightarrow a$ as shorthand for “ x approaches a ” and $f(x) \rightarrow L$ as shorthand for “ $\lim_{x \rightarrow a} f(x) = L$.”)

Just as we did for sequences, we can illustrate how this works step by step.

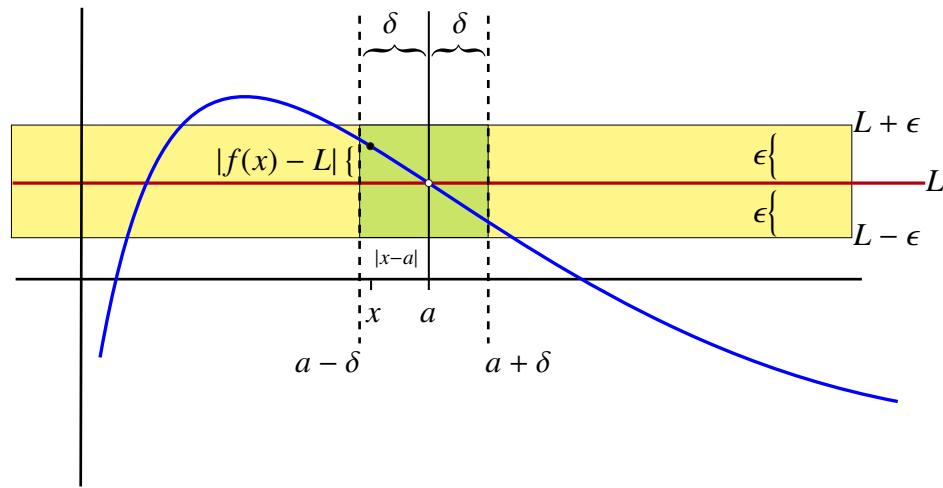
Start with a function f that we suspect has a limit of L as $x \rightarrow a$.



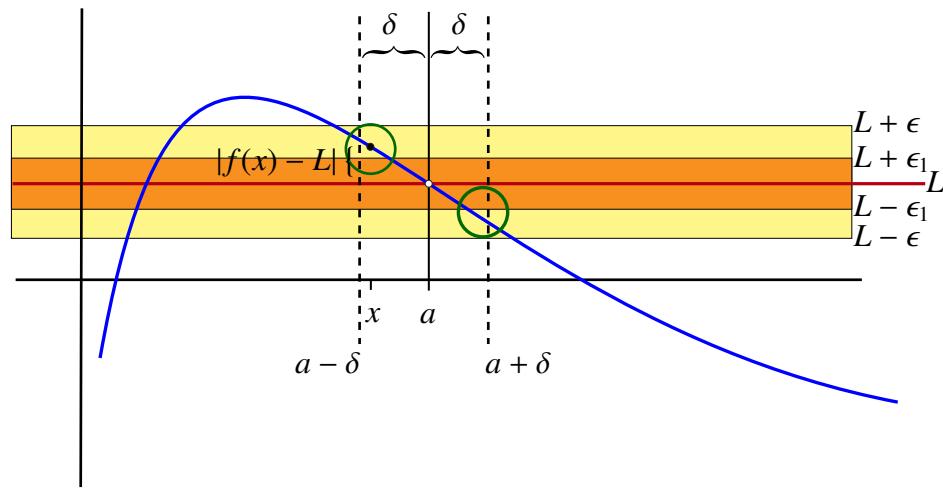
First choose some positive number ϵ to act as the tolerance. The horizontal lines $y = L + \epsilon$ and $y = L - \epsilon$ are called the *error bounds*.



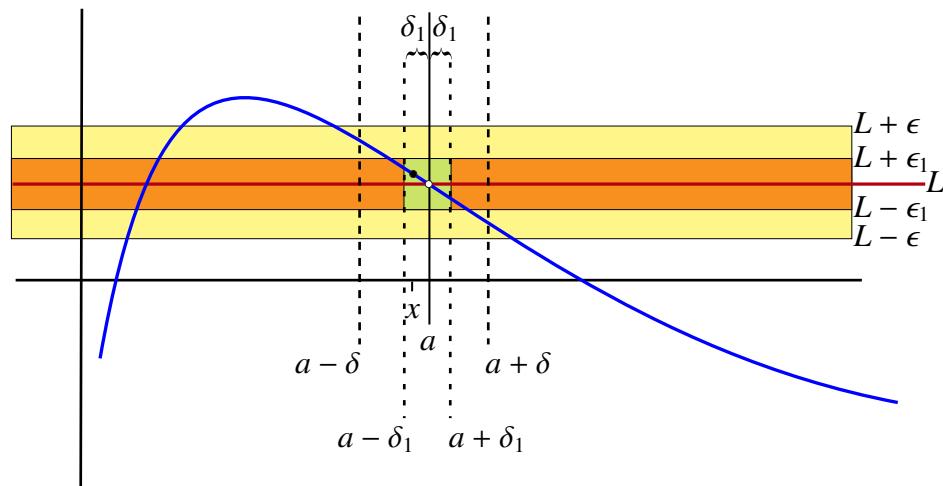
Our task is to find a $\delta > 0$ such that if $0 < |x - a| < \delta$, then $|f(x) - L| < \epsilon$. This means that on the interval $(a - \delta, a + \delta)$, excluding $x = a$, the values of $f(x)$ all lie between the lines of the error bounds. That is, if $0 < |x - a| < \delta$, then the graph of the function lies entirely within these bounds. To do this we can choose δ as in the following diagram.



Next, someone else comes along with a smaller tolerance ϵ_1 for us to use. Our task is again to find a new distance δ_1 such that if $0 < |x - a| < \delta_1$, then $|f(x) - L| < \epsilon_1$. Note that the old value δ may no longer work since there could be portions of the graph of f within $(a - \delta, a + \delta)$ that lie outside the new error bounds.



Consequently, we may have to pick a smaller δ_1 as illustrated in the next diagram.



You may notice that the choice of δ_1 is not the largest possible value that would serve our purposes. Finding the largest possible δ_1 is not important. We simply need to find *one* value that will work.

As in the case with sequences, this process continues with even smaller tolerances ϵ being presented, forcing us to find new values of δ . We would be able to conclude that $\lim_{x \rightarrow a} f(x) = L$ if we could ensure that no matter how small the ϵ we are given, we can find an appropriate δ . Unfortunately, to do this explicitly is often extremely difficult, if not impossible. In fact, in this example, since we only have a picture of the function, we would never really be able to find an appropriate δ explicitly for every given ϵ .

We end this section by demonstrating how the definition can sometimes be used to show that certain functions do *not* have limits at a particular point. We illustrate this with an example that will be useful to us later. It is an analog for functions of the sequence

$$\{1, -1, 1, -1, \dots, (-1)^{n+1}, \dots\} = \{(-1)^{n+1}\}.$$

EXAMPLE 2 Consider the function $f(x) = \frac{|x|}{x}$. Recall the definition of the absolute value:

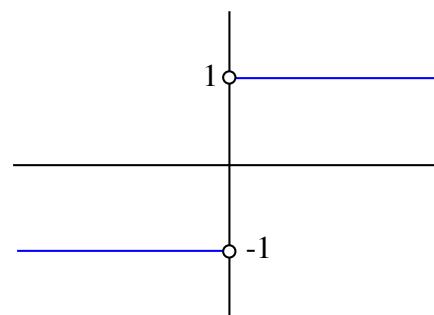
$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}.$$

Hence,

$$f(x) = \begin{cases} \frac{x}{x} & \text{if } x > 0 \\ \frac{-x}{x} & \text{if } x < 0 \end{cases}.$$

Consequently,

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



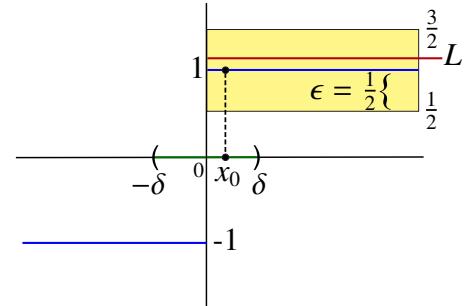
The graph of f appears as:

We want to know if this function has a limit at $x = 0$.

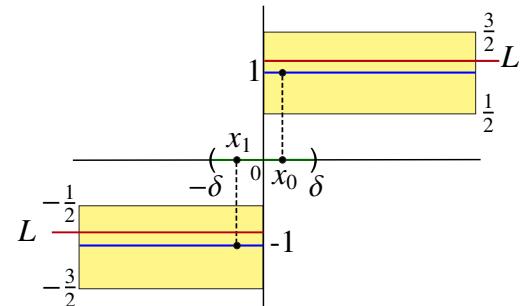
Notice from the graph that as x approaches 0 from the left (or negative side), the function has a constant value of -1 . We might guess that if this function had a limit as x approaches 0, then it should be -1 . On the other hand, if we allow x to approach 0 from the right (or positive side), we see that $f(x)$ always has its value equal to 1. This suggests that 1 should be the limit. However, just as in the case for sequences, the limit of a function should be *uniquely* defined. Since we are torn between a limit of -1 and 1, we will try to show that *no limit exists*.

Let's suppose that there is a limit and let $\lim_{x \rightarrow 0} f(x) = L$. Choose the tolerance ϵ to be $\frac{1}{2}$. (Note that any choice of $\epsilon < 1$ will work.) The definition of the limit of a function tells us that we can find a cutoff distance $\delta > 0$ such that if $0 < |x - 0| < \delta$, then $|f(x) - L| < \frac{1}{2}$.

First consider $x_0 \in (0, \delta)$. Then since $0 < |x_0 - 0| < \delta$, we get $|f(x_0) - L| < \frac{1}{2}$. But $x_0 > 0$, so $f(x_0) = 1$. This means that $|1 - L| < \frac{1}{2}$. That is, the distance from 1 to L is less than $\frac{1}{2}$. This shows us that $L \in (\frac{1}{2}, \frac{3}{2})$.



Now let $x_1 \in (-\delta, 0)$. Again we have that $0 < |x_1 - 0| < \delta$, so $|f(x_1) - L| < \frac{1}{2}$. But $x_1 < 0$ so $f(x_1) = -1$. This means that $|-1 - L| < \frac{1}{2}$. That is, the distance from -1 to L is less than $\frac{1}{2}$. This shows us that $L \in (-\frac{3}{2}, -\frac{1}{2})$.

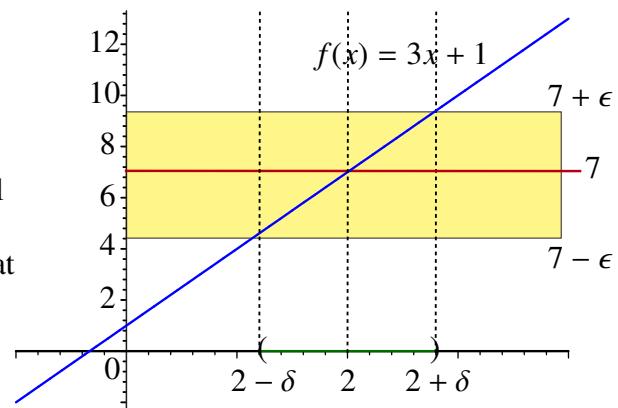


If we combine these two results we see that L must be simultaneously in both the interval $(-\frac{3}{2}, -\frac{1}{2})$ and in the interval $(\frac{1}{2}, \frac{3}{2})$. Since these intervals are disjoint, this is impossible. Thus, we have shown that this function *does not have a limit* at $x = 0$. \blacktriangleleft

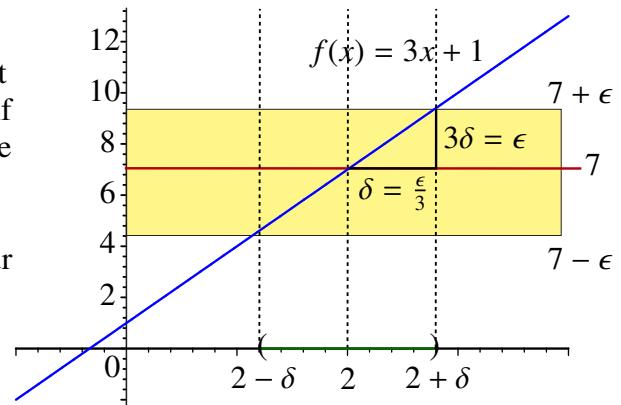
EXAMPLE 3

Show that $\lim_{x \rightarrow 2} 3x + 1 = 7$.

SOLUTION The fact that the limit should be 7 is easy to see. As x approaches 2 it makes sense that $3x$ approaches $3 \cdot 2 = 6$ and hence that $3x + 1$ should approach $6 + 1 = 7$. But we can also show that the formal definition of a limit is satisfied as well. To see how to do this, recall that given a tolerance $\epsilon > 0$ we must be able to find a cutoff distance $\delta > 0$ such that if x is within δ of 2, then $|3x + 1 - 7| < \epsilon$.



We observe that the graph of f is just a line with slope 3. This means that if we deviate from 2 by δ units, then the value of $f(x)$ can either increase or decrease by at most $3 \cdot \delta$. So if we want this deviation to be less than our tolerance ϵ , we should ask that $3 \cdot \delta \leq \epsilon$ or equivalently that $\delta \leq \frac{\epsilon}{3}$.



Alternatively, we can work backwards. If we want

$$|(3x + 1) - 7| < \epsilon,$$

then this is equivalent to

$$|3x - 6| = 3|x - 2| < \epsilon$$

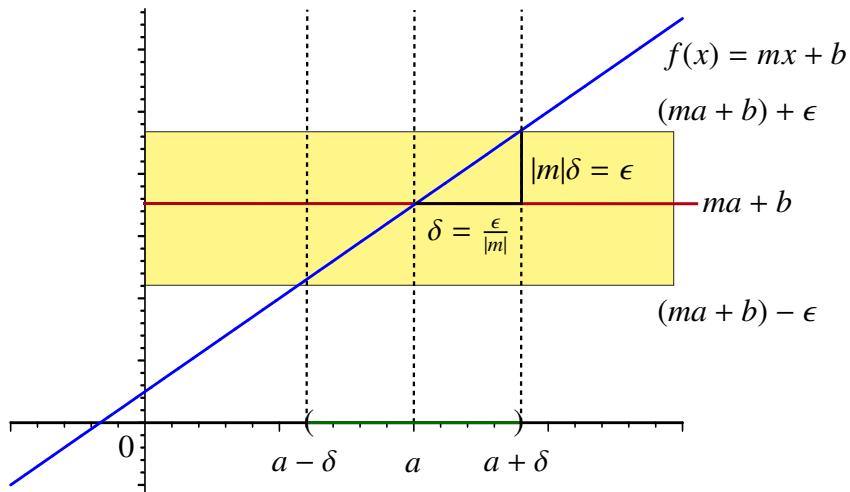
or

$$|x - 2| < \frac{\epsilon}{3}.$$

Therefore, we get that if $0 < |x - 2| < \frac{\epsilon}{3}$, then $|(3x + 1) - 7| < \epsilon$, so $\delta = \frac{\epsilon}{3}$ satisfies the definition. ◀

REMARK

If $f(x) = mx + b$, where $m \neq 0$. Then $\lim_{x \rightarrow a} f(x) = m \cdot a + b$.



In particular, given $\epsilon > 0$, and if $\delta = \frac{\epsilon}{|m|}$, then if $0 < |x - a| < \delta = \frac{\epsilon}{|m|}$, we have

$$|f(x) - (m \cdot a + b)| < \epsilon.$$
◀

We have just seen that for a function of the form $f(x) = mx + b$ with $m \neq 0$, the task of applying the definition of the limit to find an appropriate δ given an $\epsilon > 0$ is actually straightforward. In fact, any δ satisfying

$$0 < \delta < \frac{\epsilon}{|m|}$$

will suffice. However, even for relatively simple functions, the process can become quite complicated. To illustrate this we will consider the following example.

EXAMPLE 4 Show that

$$\lim_{x \rightarrow 3} x^2 = 9.$$

SOLUTION This result is intuitively clear since as x approaches 3 it makes sense that x^2 should approximate 9. However, to apply the definition we first choose $\epsilon > 0$. We want to choose a $\delta > 0$ so that if $0 < |x - 3| < \delta$, then $|x^2 - 9| < \epsilon$. In this case, we have

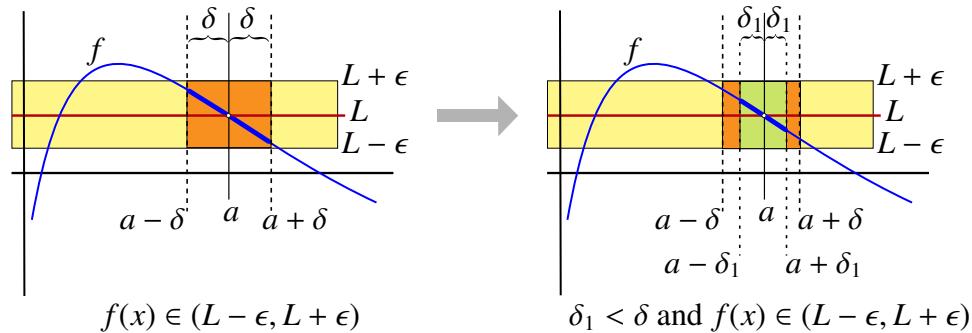
$$|x^2 - 9| = |(x - 3)(x + 3)| = |x - 3||x + 3|.$$

If we proceed as we did in the linear case, we might be tempted to choose $\delta = \frac{\epsilon}{|x+3|}$ because if $0 < |x - 3| < \frac{\epsilon}{|x+3|}$, then

$$\begin{aligned} |x^2 - 9| &= |x - 3||x + 3| \\ &< \frac{\epsilon}{|x + 3|}|x + 3| \\ &= \epsilon \end{aligned}$$

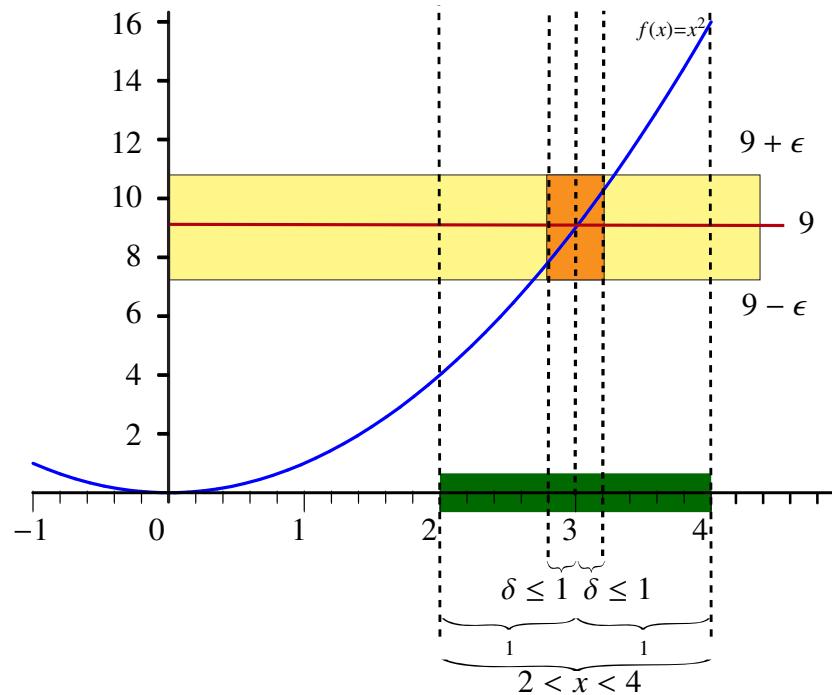
exactly as required. But the problem with this choice is that $|x + 3|$ is **not a constant** since as x moves toward 3, its value changes; thus, this is not a valid choice for δ . So how do we get around this? The key is the following trick that allows us to control the size of $|x + 3|$.

Trick: If we find a δ that works for a particular ϵ , then any smaller δ will also satisfy the definition of the limit of a function for the same ϵ .



Therefore, **we can always assume that the δ we are looking for is less than or equal to 1**. So for the rest of this example, we will assume that the strategy is to choose a $\delta \leq 1$.

Once we have assumed that $\delta \leq 1$, it follows that if $0 < |x - 3| < \delta \leq 1$, we must have that $2 < x < 4$ and these are the only values of x that we need to consider.



Now we can return to the task of finding a fixed $\delta \leq 1$ knowing that $2 < x < 4$. But if $2 < x < 4$, then

$$|x + 3| < |4 + 3| = 7.$$

This information is helpful because we can now refine our previous calculation so that

$$|x^2 - 9| = |(x - 3)(x + 3)| = |x - 3||x + 3| < 7|x - 3|$$

provided that $|x - 3| < 1$. Now we are in a situation that is similar to the linear case, and if we ask that

$$\delta = \min\{1, \frac{\epsilon}{7}\},$$

then when $0 < |x - 3| < \delta$ we would have both

$$|x + 3| < 7$$

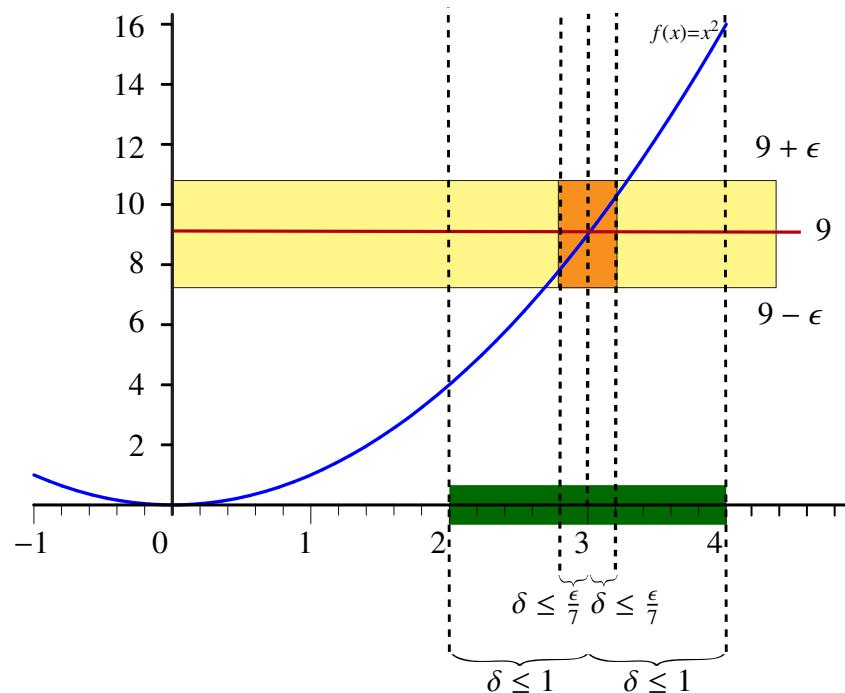
because $\delta \leq 1$, and

$$|x - 3| < \frac{\epsilon}{7}$$

as well. It follows that if $0 < |x - 3| < \delta$, then

$$\begin{aligned} |x^2 - 9| &= |x - 3||x + 3| \\ &< 7|x - 3| \\ &< 7\left(\frac{\epsilon}{7}\right) \\ &= \epsilon \end{aligned}$$

exactly as required.



The previous example shows that, even for simple functions, completing the ϵ - δ game successfully can be a challenge. For this reason we will soon develop various arithmetic properties to help us avoid having to explicitly do such calculations whenever possible.

We end this section with three important remarks about the existence of limits.

REMARKS

1. For $\lim_{x \rightarrow a} f(x)$ to exist, f must be defined on a **open** interval (α, β) containing $x = a$, except possibly at $x = a$.
 2. The value of $f(a)$, if it is defined at all, does not affect the existence of the limit or its value.
 3. If two functions are equal, except possibly at $x = a$, then their limiting behavior at a is identical.

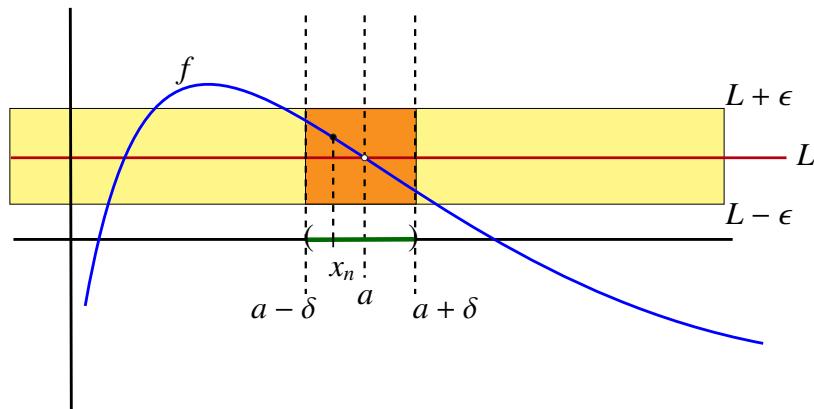
5.2 Sequential Characterization of Limits

We have just seen that there is a close connection between limits of sequences and limits of functions. This leads us to ask the following:

Question: Can we characterize limits of functions in terms of limits of sequences?

To see why we might be able to do so let's assume that $\lim_{x \rightarrow a} f(x) = L$. Now assume that we have a sequence $\{x_n\}$ such that $x_n \rightarrow a$ and $x_n \neq a$. Then for large n , x_n should be very close to a , and as such we should have that $f(x_n)$ should be very close to L , leading us to believe that perhaps $\lim_{n \rightarrow \infty} f(x_n) = L$. In fact, we can show that this is the case.

Let $\epsilon > 0$. Since $\lim_{x \rightarrow a} f(x) = L$, we can find a $\delta > 0$ such that if $x \in (a - \delta, a + \delta)$ and $x \neq a$, then $f(x) \in (L - \epsilon, L + \epsilon)$. But since $x_n \rightarrow a$, we can find a cutoff N so if $n \geq N$, we have $x_n \in (a - \delta, a + \delta)$. Since we also know that $x_n \neq a$, we have that if $n \geq N$, then $f(x_n) \in (L - \epsilon, L + \epsilon)$. So the interval $(L - \epsilon, L + \epsilon)$ contains a tail of the sequence $\{f(x_n)\}$ and as such we have shown that $f(x_n) \rightarrow L$.



In fact, we can say more as the next theorem completes the connection between limits of sequences and limits of functions.

THEOREM 1

Sequential Characterization of Limits

Let f be defined on an open interval containing $x = a$, except possibly at $x = a$. Then the following two statements are equivalent:

- i) $\lim_{x \rightarrow a} f(x)$ exists and equals L .
- ii) If $\{x_n\}$ is a sequence with $x_n \neq a$ and $x_n \rightarrow a$, then

$$\lim_{n \rightarrow \infty} f(x_n) = L.$$

PROOF

We have already shown that i) implies ii). To prove that ii) implies i) we will show that if L is not the limit of f as x approaches a , then ii) fails. That is, we will show

that we can construct a sequence $\{x_n\}$ with $x_n \neq a$ and $x_n \rightarrow a$ but $\{f(x_n)\}$ does not converge to L .

Assume that L is not the limit of f as x approaches a . Then there must exist some $\epsilon_0 > 0$ so that there is no $\delta > 0$ where $0 < |x - a| < \delta$ implies that

$$|f(x) - L| < \epsilon_0.$$

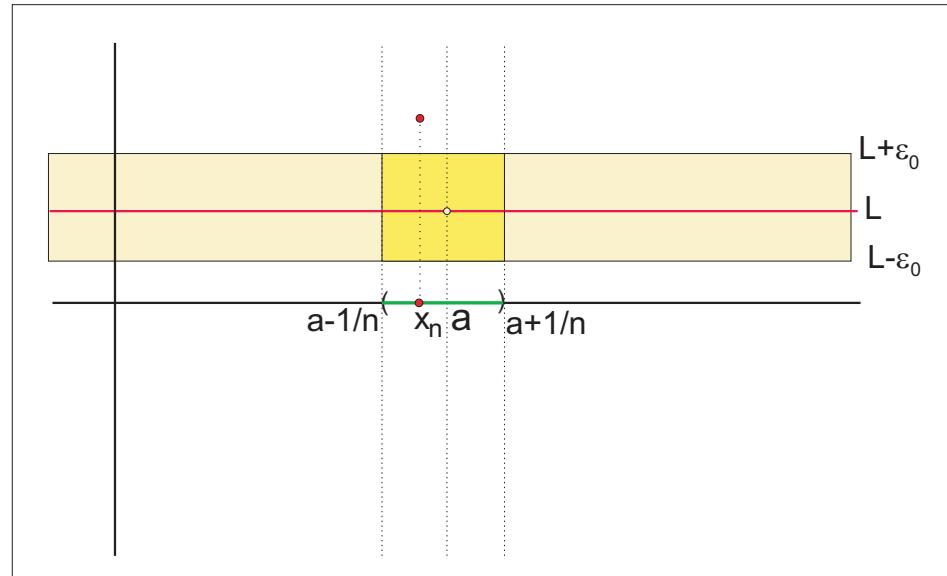
Therefore, for every $\delta > 0$ there is a $x_\delta \in (a - \delta, a + \delta) \setminus \{a\}$ with

$$|f(x_\delta) - L| \geq \epsilon_0.$$

In particular, for each $n \in \mathbb{N}$, this would be true for $\delta = \frac{1}{n}$.

For each $n \in \mathbb{N}$, consider $\delta = \frac{1}{n}$. Then there is an x_n with $0 < |x_n - a| < \frac{1}{n}$, but

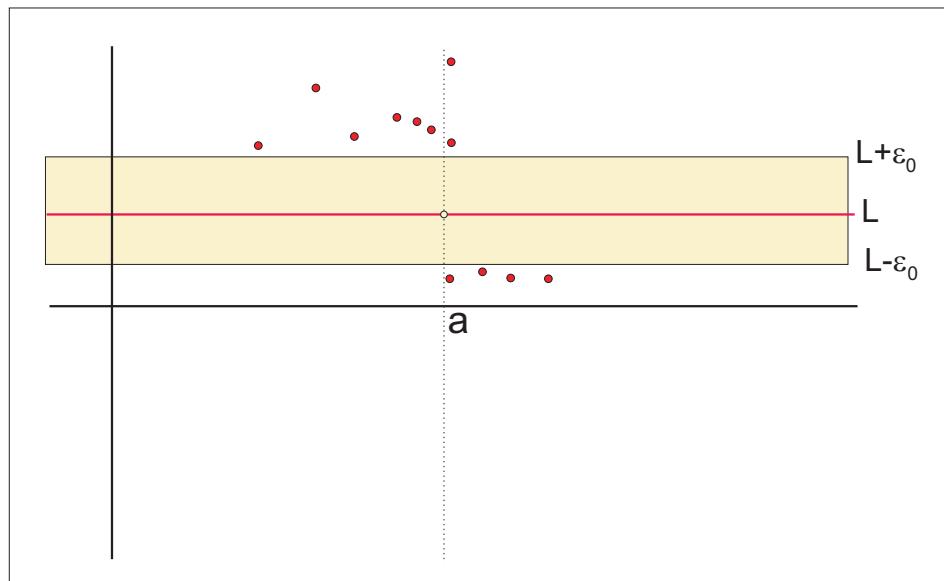
$$|f(x_n) - L| \geq \epsilon_0.$$



This gives us a sequence $\{x_n\}$ with $x_n \neq a$ and $x_n \rightarrow a$, while for each $n \in \mathbb{N}$, we have

$$|f(x_n) - L| \geq \epsilon_0.$$

It follows that $\{f(x_n)\}$ does not converge to L .



■

Since we know that convergent sequences can have only one limit, an immediate consequence of the Sequential Characterization of Limits Theorem is the fact that limits for functions must also be unique.

THEOREM 2

Uniqueness of Limits for Functions

Assume that $\lim_{x \rightarrow a} f(x) = L$ and that $\lim_{x \rightarrow a} f(x) = M$. Then $L = M$. That is, the limit of a function is unique.

We will later see that the Sequential Characterization of Limits Theorem allows us to carry over all of the arithmetic rules we developed for sequences, as well as the Squeeze Theorem, to limits of functions. It can also be a useful tool to show that certain functions do not have limits at certain points.

EXAMPLE 5

Recall that we showed the function

$$f(x) = \frac{|x|}{x} = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}$$

failed to have a limit at $x = 0$. We can use sequences to show this result as well.

Assume that $\lim_{x \rightarrow 0} f(x)$ did exist and was equal to L . Let $x_n = \frac{1}{n}$. Then $x_n \rightarrow 0$. So by the sequential characterization of limits we have $f(x_n) \rightarrow L$. But $f(x_n) = 1$ for each n , so the constant sequence $\{f(x_n)\}$ converges to 1. As such, we must have $L = 1$.

On the other hand, if $y_n = -\frac{1}{n}$, then again $y_n \rightarrow 0$ as well. Similar to before, this means that $f(y_n) \rightarrow L$. But $f(y_n) = -1$ for each n , so $L = -1$. However, since a function cannot have two different limits, it must not have any. ◀

This leads us to the following strategy for showing that certain limits do not exist:

Strategy [Showing Limits Do Not Exist]:

If you want to show that $\lim_{x \rightarrow a} f(x)$ does not exist you can do so by either of the following:

- 1) Find a sequence $\{x_n\}$ with $x_n \rightarrow a$, $x_n \neq a$ for which $\lim_{n \rightarrow \infty} f(x_n)$ does not exist.
- 2) Find two sequences $\{x_n\}$ and $\{y_n\}$ with $x_n \rightarrow a$, $x_n \neq a$ and $y_n \rightarrow a$, $y_n \neq a$ for which $\lim_{n \rightarrow \infty} f(x_n) = L$ and $\lim_{n \rightarrow \infty} f(y_n) = M$ but $L \neq M$.

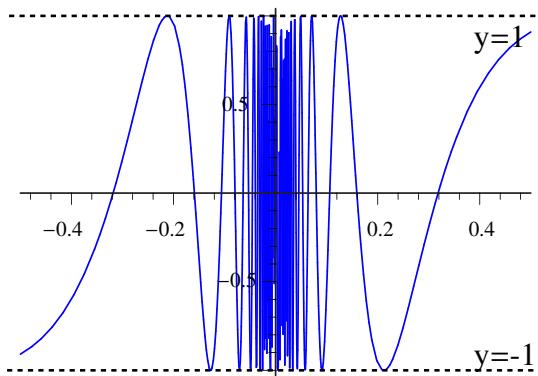
As another illustration of this strategy, we next consider a rather badly behaved function that will provide us with some interesting examples going forward.

EXAMPLE 6

In this example we will look at the exotic function given by the formula

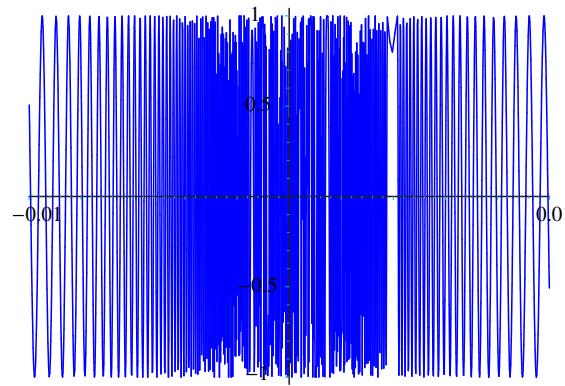
$$f(x) = \sin\left(\frac{1}{x}\right).$$

The following is the graph of this function on the interval $[-0.5, 0.5]$.



Notice that as x approaches 0 the function oscillates wildly between 1 and -1. (This happens as we approach 0 from the right because the function $g(x) = \frac{1}{x}$ maps the the interval $[\frac{1}{2(n+1)\pi}, \frac{1}{2n\pi}]$ onto the interval $[2n\pi, 2(n+1)\pi]$. As we approach 0 from the left $g(x) = \frac{1}{x}$ maps the interval $[\frac{-1}{2n\pi}, \frac{-1}{2(n+1)\pi}]$ onto the interval $[-2(n+1)\pi, -2n\pi]$.)

In fact this rapid oscillation can be seen even more clearly if we look at the graph on the interval $[-0.01, 0.01]$.



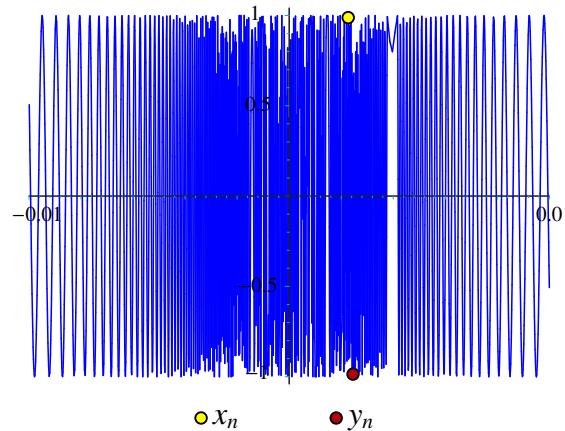
(Note: If you look closely you can see that this rendering of the plot illustrates sampling errors in the graphing routine. A more accurate graph may be created by using more sampling points in the mathematical software used to render the plot.)

Behavior such as what we have seen above suggests that $\lim_{x \rightarrow 0} \sin\left(\frac{1}{x}\right)$ should not exist. We can use the Sequential Characterization of Limits Theorem to show this explicitly.

First observe that for each $n \in \mathbb{N}$ we have $\sin(\frac{\pi}{2} + 2n\pi) = 1$ and $\sin(\frac{3\pi}{2} + 2n\pi) = -1$. Let

$$x_n = \frac{1}{\frac{\pi}{2} + 2n\pi} \quad \text{and} \quad y_n = \frac{1}{\frac{3\pi}{2} + 2n\pi}.$$

Then $x_n \rightarrow 0$ and $y_n \rightarrow 0$, however since $f(x_n) = \sin(\frac{\pi}{2} + 2n\pi) = 1$ and $f(y_n) = \sin(\frac{3\pi}{2} + 2n\pi) = -1$ for each $n \in \mathbb{N}$, we get that $f(x_n) \rightarrow 1$ while $f(y_n) \rightarrow -1$.



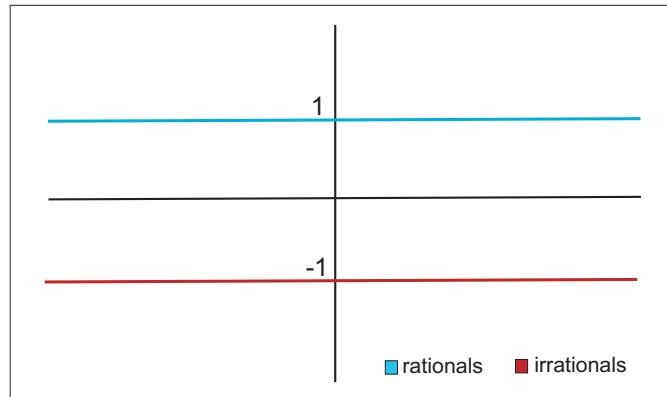
This is sufficient to show that $\lim_{x \rightarrow 0} \sin\left(\frac{1}{x}\right)$ does not exist. ◀

5.2.1 Three More Strange Functions

In the previous section we saw that the Sequential Characterization of Limits could be used to show that certain limits do not exist. In this section we will continue on

from where we left off and introduce three more rather strange functions for which the Sequential Characterization of Limits can be a useful tool to help us understand their behaviour.

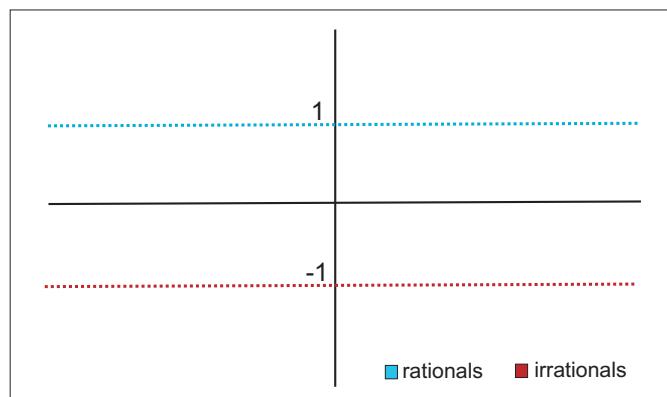
EXAMPLE 7



Consider

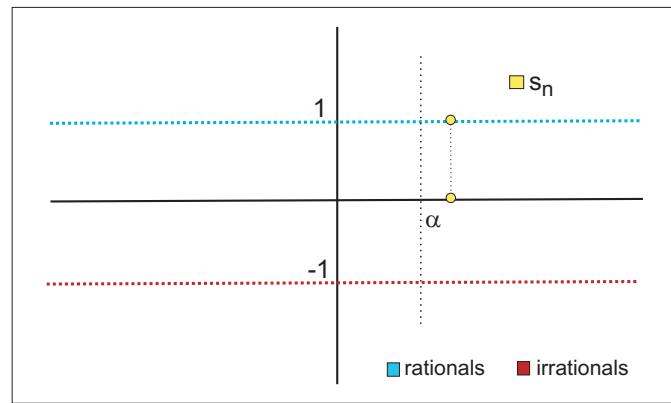
$$f(x) := \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ -1 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

The graph of f looks almost identical to the two lines $y = 1$ and $y = -1$. But in fact there are infinitely many gaps through out the domain so to emphasize this we will use two dotted lines rather than what would appear to be solid lines.

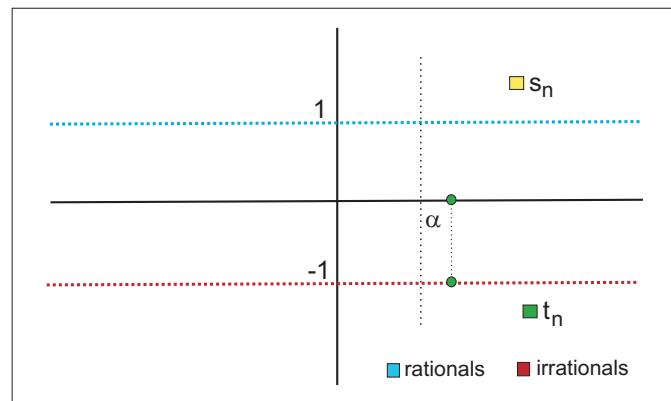


Let $\alpha \in \mathbb{R}$. We would like to investigate the possibility that a limit might exist at $x = a$. Our instincts probably tell us that the limit should not exist because there are x 's arbitrarily close to a where $f(x) = 1$ and x 's arbitrarily close to a where $f(x) = -1$. This would mean that the limit should be both 1 and -1 , which we know is not permitted. We can however make this more rigorous.

We begin by choosing a sequence $\{s_n\} \subset \mathbb{Q}$, with $s_n \rightarrow \alpha$.



Then $f(s_n) \rightarrow 1$ since $f(s_n) = 1$ for each $n \in \mathbb{N}$. the Sequential Characterization of Limits would imply that if $\lim_{x \rightarrow \alpha} f(x)$ exists it must be equal to 1.

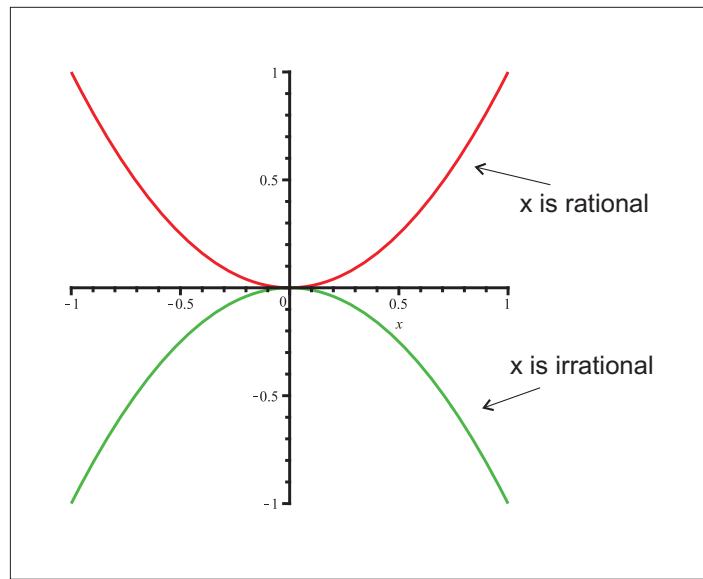


Next we choose $\{t_n\} \subset \mathbb{R} \setminus \mathbb{Q}$ with $t_n \rightarrow \alpha$. Then $f(t_n) \rightarrow -1$. However, the Sequential Characterization of Limits would then tell us that $\lim_{x \rightarrow \alpha} f(x)$ does not exist.

This function has the unusual property that the limit fails to exist at every point even though it is defined everywhere.



Our next strange function is a variant of the previous function.

EXAMPLE 8

Let

$$f(x) = \begin{cases} x^2 & \text{if } x \in \mathbb{Q}, \\ -x^2 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q}, \end{cases}$$

We can again investigate the limiting behaviour of this function. However, unlike the previous example, in this case there are two cases to consider.

Case 1: $\alpha \neq 0$.

A little thought might lead us to believe that the split nature of this function would again lead to our limit failing to exist. To make this clear we must answer the following questions which are left as exercises.

Questions: . Let $\{r_n\} \subset \mathbb{Q}$ and $\{s_n\} \subset \mathbb{R} \setminus \mathbb{Q}$ with $r_n, s_n \neq \alpha$ and $r_n \rightarrow \alpha$ and $s_n \rightarrow \alpha$.

- What is $\lim_{n \rightarrow \infty} f(r_n)$?
- What is $\lim_{n \rightarrow \infty} f(s_n)$?
- What does the sequential characterization of limits tell us about $\lim_{x \rightarrow a} f(x)$?

If the answers to *i*) and *ii*) are different, then the limit does not exist. If they are the same, then the limit exists. Why?

Case 2: $\alpha = 0$.

Here we will see that the situation is different from our previous example. We let $\{x_n\}$ be a sequence of non-zero real numbers converging to 0. We know that $f(x_n) = x_n^2$ if x_n is rational and $f(x_n) = -x_n^2$ if x_n is irrational. It follows that

$$-x_n^2 \leq f(x_n) \leq x_n^2$$

for all $n \in \mathbb{N}$. However, since $\lim_{n \rightarrow \infty} x_n = 0$, the rules of arithmetic for limits of sequences tells us that

$$\lim_{n \rightarrow \infty} -x_n^2 = 0 = \lim_{n \rightarrow \infty} x_n^2.$$

We can now appeal to the Squeeze Theorem to conclude that $\lim_{n \rightarrow \infty} f(x_n) = 0$ and hence to the Sequential Characterization of Limits to show that

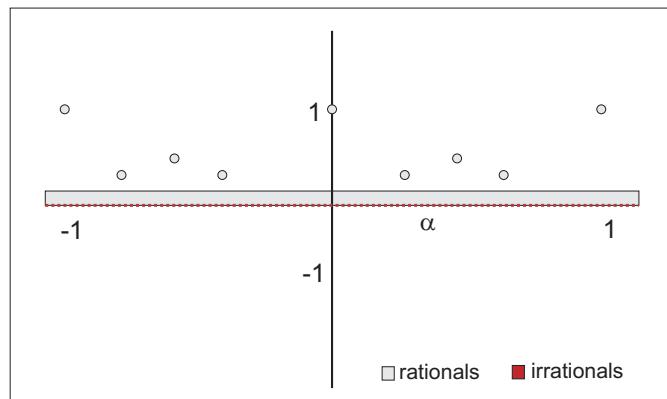
$$\lim_{x \rightarrow 0} f(x) = 0.$$



The last of the three rather unusual functions we will consider in this section is Thomae's function, which is often referred to as the popcorn function due to its unusual graph resembling the behaviour of popcorn being cooked.

EXAMPLE 9

The definition of Thomae's function is more complex than the previous two functions.



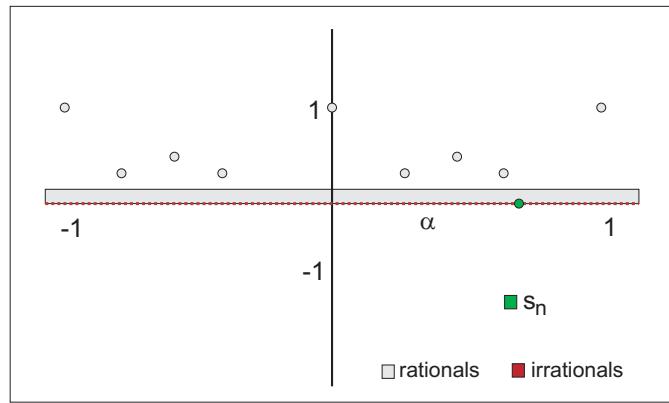
DEFINITION

Thomae's Function

The function f defined by

$$f(x) = \begin{cases} 1 & \text{if } x = 0 \\ \frac{1}{n} & \text{if } x = \frac{k}{n} \in \mathbb{Q} \text{ with } k \in \mathbb{Z} \setminus \{0\}, n \in \mathbb{N}, \gcd(k, n) = 1. \\ 0 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

is called *Thomae's Function*.



The first thing you should notice about f is that since $f(x) = 0$ for each irrational number. Moreover, for every $\alpha \in \mathbb{R}$ there exists a sequence $\{s_n\} \subset \mathbb{R} \setminus \mathbb{Q}$ with $s_n \neq \alpha$ and $s_n \rightarrow \alpha$. As a consequence the Sequential Characterization of Limits tells us that if $\lim_{x \rightarrow \alpha} f(x)$ exists it must be 0. As such we ask:

Question: Is $\lim_{x \rightarrow \alpha} f(x) = 0$?

To see why it actually is lets let $\epsilon > 0$ and choose $N \in \mathbb{N}$ so that

$$\frac{1}{N} < \epsilon.$$

Consider the interval $I = [\alpha - 1, \alpha + 1]$. For any natural number n there are only finitely many rational numbers of the form $\frac{m}{n}$ in the interval I . This means that there are also only finitely many rationals $r = \frac{m}{n}$ in I for which the denominator n is smaller than N . Let $\{r_1, r_2, \dots, r_k\}$ be the collection of rationals different from α itself. Let

$$\delta = \min\{|\alpha - r_1|, |\alpha - r_2|, \dots, |\alpha - r_k|\}.$$

Then if $0 < |x - \alpha| < \delta$, then either x is irrational and $f(x) = 0$, or x is a rational number different from any of the r_i 's. But in this case $x = \frac{m}{n}$ where $n \geq N$. It then follows that

$$f(x) = \frac{1}{n} \leq \frac{1}{N} < \epsilon.$$

This shows that if $0 < |x - \alpha| < \delta$, then

$$|f(x) - 0| < \epsilon$$

and hence that $\lim_{x \rightarrow \alpha} f(x) = 0$.

This argument works because most rational numbers have very large denominators. In fact what we have shown is that if you have any $\alpha \in \mathbb{R}$ and a sequence $\{r_n\}$ for rationals distinct from α which converges to α , the denominators of these rationals must diverge to ∞ .

REMARK

Later in this chapter we will introduce the notion of continuity. Thomae's Function has the unusual property that it is continuous at each irrational number but discontinuous at each rational number.

5.3 Arithmetic Rules for Limits of Functions

In this section, we will see that most of the usual rules of arithmetic hold for limits of functions just as they did for sequences. In fact, we have the following theorem:

THEOREM 3

Arithmetic Rules for Limits of Functions

Let f and g be functions and let $a \in \mathbb{R}$. Assume that $\lim_{x \rightarrow a} f(x) = L$ and that $\lim_{x \rightarrow a} g(x) = M$. Then

- i) Assume that $f(x) = c$ for every $x \in \mathbb{R}$. Then $\lim_{x \rightarrow a} f(x) = c$.
- ii) For any $c \in \mathbb{R}$, $\lim_{x \rightarrow a} cf(x) = cL$.
- iii) $\lim_{x \rightarrow a} f(x) + g(x) = L + M$.
- iv) $\lim_{x \rightarrow a} f(x)g(x) = LM$.
- v) $\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \frac{L}{M}$ if $M \neq 0$.
- vi) $\lim_{x \rightarrow a} (f(x))^\alpha = L^\alpha$ for all $\alpha > 0$, $L > 0$.

Similar to the case for sequences, in rule (v) we did not mention what happens if $M = 0$. Again, this is because there are examples of this type where the limit exists and examples where it does not. However, if we apply the Sequential Characterization of Limits Theorem, we immediately obtain the following theorem:

THEOREM 4

Assume that $\lim_{x \rightarrow a} \frac{f(x)}{g(x)}$ exists and $\lim_{x \rightarrow a} g(x) = 0$. Then

$$\lim_{x \rightarrow a} f(x) = 0.$$

REMARK

Similar to the case with sequences, if $\lim_{x \rightarrow a} g(x) = 0$ but $\lim_{x \rightarrow a} f(x) \neq 0$, the quotient will be unbounded near $x = a$. This point can be illustrated by using the example

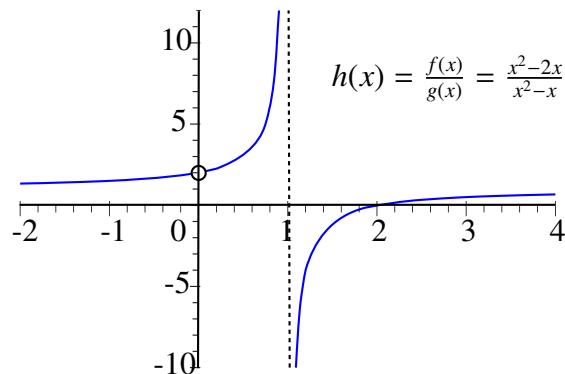
$h(x) = \frac{f(x)}{g(x)}$, where $f(x) = x^2 - 2x$ and $g(x) = x^2 - x$. The numerator and denominator are each polynomials and so (see the following example on *Limits of Polynomials* for a justification of this calculation)

$$\lim_{x \rightarrow 1} f(x) = -1$$

while

$$\lim_{x \rightarrow 1} g(x) = 0.$$

The graph of this function shows that it is unbounded near $x = 1$, exactly as we predicted.



EXAMPLE 10

Limits of Polynomials: We see from the Arithmetic Rules, rule i), that if $f(x) = \alpha_0 = f(a)$ is a constant function, then for any $a \in \mathbb{R}$, $\lim_{x \rightarrow a} f(x) = \alpha_0$.

We have already seen that if $f(x) = x$ for all $x \in \mathbb{R}$, then

$$\begin{aligned} \lim_{x \rightarrow a} f(x) &= \lim_{x \rightarrow a} x \\ &= a \\ &= f(a). \end{aligned}$$

We also get that if $g(x) = x^2$, then by rule (iv)

$$\begin{aligned} \lim_{x \rightarrow a} g(x) &= \lim_{x \rightarrow a} x^2 \\ &= \lim_{x \rightarrow a} x \times \lim_{x \rightarrow a} x \\ &= a \times a \\ &= a^2 \\ &= g(a). \end{aligned}$$



In fact, it can be shown that $\lim_{x \rightarrow a} x^n = a^n$ for any $n \in \mathbb{N}$.

Consequently, by using all of the limit rules, we get the following:

THEOREM 5
Limits of Polynomials

If $p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_n x^n$ is any *polynomial*, then

$$\lim_{x \rightarrow a} p(x) = p(a).$$

EXAMPLE 11

Limits of Rational Functions: Recall that a *rational* function is a function $f(x) = \frac{P(x)}{Q(x)}$, where P and Q are polynomials. Let's see how to calculate $\lim_{x \rightarrow a} f(x)$.

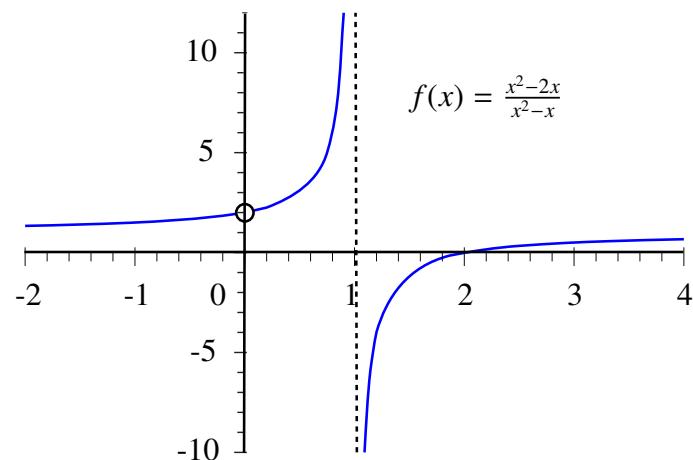
The first step is to note that from arithmetic limit rule v), we get that if

$$\lim_{x \rightarrow a} Q(x) = Q(a) \neq 0,$$

then

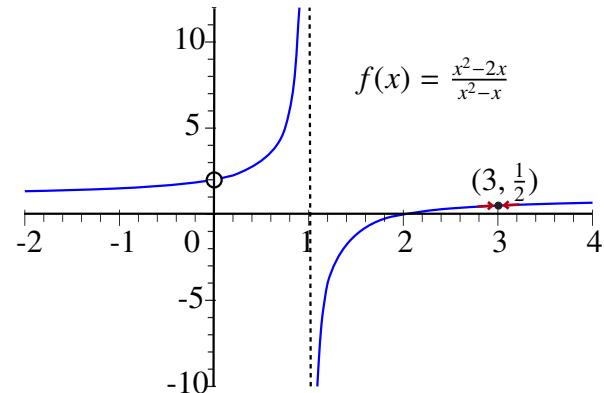
$$\begin{aligned} \lim_{x \rightarrow a} f(x) &= \lim_{x \rightarrow a} \frac{P(x)}{Q(x)} \\ &= \frac{\lim_{x \rightarrow a} P(x)}{\lim_{x \rightarrow a} Q(x)} \\ &= \frac{P(a)}{Q(a)} \\ &= f(a). \end{aligned}$$

For example, let $f(x) = \frac{x^2-2x}{x^2-x}$. The graph of f looks as follows:



If we want to find $\lim_{x \rightarrow 3} f(x)$, we get that

$$\begin{aligned}
 \lim_{x \rightarrow 3} f(x) &= \lim_{x \rightarrow 3} \frac{x^2 - 2x}{x^2 - x} \\
 &= \frac{\lim_{x \rightarrow 3} x^2 - 2x}{\lim_{x \rightarrow 3} x^2 - x} \\
 &= \frac{3^2 - 2(3)}{3^2 - 3} \\
 &= \frac{3}{6} \\
 &= \frac{1}{2}.
 \end{aligned}$$



This is consistent with what we see from the graph of f .

Suppose on the other hand that

$$\lim_{x \rightarrow a} Q(x) = Q(a) = 0,$$

but

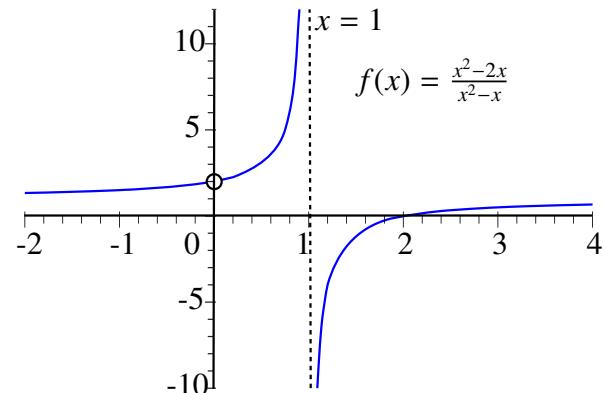
$$\lim_{x \rightarrow a} P(x) = P(a) \neq 0.$$

Then we have seen from the limit rules that $\lim_{x \rightarrow a} \frac{P(x)}{Q(x)}$ cannot exist.

To illustrate this point, consider

$$\lim_{x \rightarrow 1} \frac{x^2 - 2x}{x^2 - x}.$$

We have that $\lim_{x \rightarrow 1} x^2 - 2x = -1$, but $\lim_{x \rightarrow 1} x^2 - x = 0$. Therefore $f(x) = \frac{x^2 - 2x}{x^2 - x}$ does not have a limit as x approaches 1.



Once again, we can see this clearly from the graph of f which shows that the function is unbounded near $x = 1$.

The last case to deal with is

$$\lim_{x \rightarrow a} P(x) = P(a) = 0 = \lim_{x \rightarrow a} Q(x) = Q(a).$$

But if the polynomial P is such that $P(a) = 0$, then $P(x)$ must have $(x - a)$ as a factor. This means that there is a new polynomial $P^*(x)$ such that

$$P(x) = (x - a)P^*(x).$$

Similarly, there is a new polynomial $Q^*(x)$ such that

$$Q(x) = (x - a)Q^*(x).$$

But then

$$\begin{aligned} \frac{P(x)}{Q(x)} &= \frac{(x - a)P^*(x)}{(x - a)Q^*(x)} \\ &= \frac{P^*(x)}{Q^*(x)} \end{aligned}$$

for all $x \neq a$. Hence $\frac{P(x)}{Q(x)}$ will have the same limit as x approaches a as $\frac{P^*(x)}{Q^*(x)}$. So we can replace $\frac{P(x)}{Q(x)}$ with $\frac{P^*(x)}{Q^*(x)}$ and start again.

Let's consider

$$\lim_{x \rightarrow 0} \frac{x^2 - 2x}{x^2 - x}.$$

Now $P(x) = x^2 - 2x$ while $Q(x) = x^2 - x$. It is easy to see that

$$P(0) = 0 = Q(0).$$

This means that both $P(x)$ and $Q(x)$ contain a factor $(x - 0) = x$. In fact,

$$\begin{aligned} \frac{P(x)}{Q(x)} &= \frac{(x)(x - 2)}{(x)(x - 1)} \\ &= \frac{x - 2}{x - 1} \end{aligned}$$

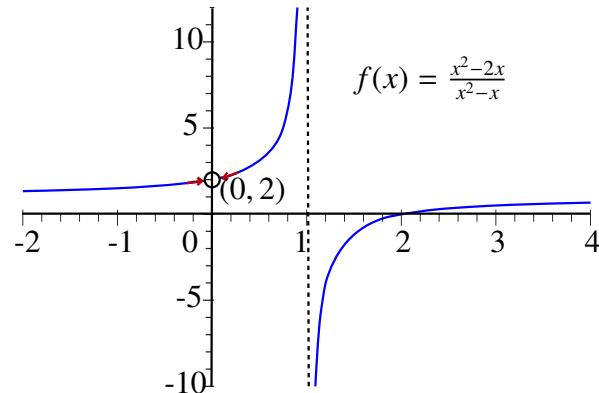
for all $x \neq 0$. But

$$\lim_{x \rightarrow 0} \frac{x - 2}{x - 1} = \frac{-2}{-1} = 2.$$

It follows that

$$\lim_{x \rightarrow 0} \frac{x^2 - 2x}{x^2 - x} = 2.$$

Once more, the graph agrees with our calculation. Notice that even though the graph has a *hole* at $(0, 2)$, the limit can still exist there!



This suggests the following algorithm for finding limits of rational functions:

Strategy [Finding Limits for Rational Functions]:

Let $f(x) = \frac{P(x)}{Q(x)}$.

Step 1: If $Q(a) \neq 0$, then

$$\lim_{x \rightarrow a} \frac{P(x)}{Q(x)} = \frac{P(a)}{Q(a)}.$$

Otherwise go to Step 2.

Step 2: If $P(a) \neq 0$ but $Q(a) = 0$, then the *limit does not exist*. Otherwise, go to Step 3.

Step 3: If $P(a) = 0$ and $Q(a) = 0$, write

$$\frac{P(x)}{Q(x)} = \frac{(x-a)P^*(x)}{(x-a)Q^*(x)}$$

and return to Step 1 using the new function

$$f^*(x) = \frac{P^*(x)}{Q^*(x)}$$

since $\lim_{x \rightarrow a} f^*(x) = \lim_{x \rightarrow a} f(x)$.

It is worth noting that for some examples there may be a more efficient method to find the limit, but this strategy will always work.

5.4 One-sided Limits

In a previous section, we encountered the function $f(x) = \frac{|x|}{x}$. We showed that this function did *not* have a limit as x approached 0. However, if we consider only positive values of x , then the function has constant value 1. Thus as x approaches 0 from the right, $f(x)$ approaches 1, and in fact is equal to 1. In this case, we say that 1 is the limit of $f(x)$ as x approaches 0 from the right (or from above). In general, we say that L is the limit of a function f as x approaches a from the right, or from above, if $f(x)$ approximates L as closely as we wish by choosing $x > a$ but close enough to a .

Similarly, we can consider limits from the left, or from below. We say that L is the limit of a function f as x approaches a from the left, or from below, if $f(x)$ approximates L as closely as we wish by choosing $x < a$ but close enough to a . In the case $f(x) = \frac{|x|}{x}$, this function has a limit as x approaches 0 from the left (or from below) of -1 .

The limits we have described above are called *one-sided limits*, while the limits we have been looking at up until now are called *two-sided limits*. We can make the description of one-sided limits more precise by mimicking the definition of a two-sided limit that we developed earlier.

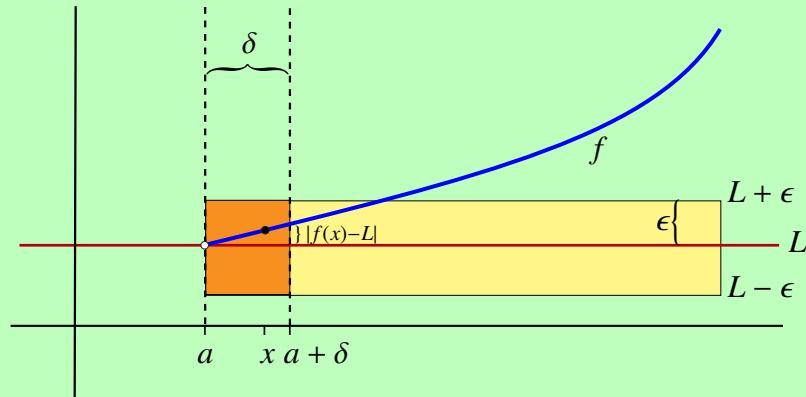
We begin with the *limit from the right*.

DEFINITION

Limit from the Right

Let f be a function and let $a \in \mathbb{R}$.

We say that f has a limit L as x approaches a from the *right*, or from above, if for any positive tolerance $\epsilon > 0$, we can find a cutoff distance $\delta > 0$ such that if the distance from x to a is less than δ , and if $x > a$, then $f(x)$ approximates L with an error less than ϵ . That is, if $0 < x - a < \delta$, then $|f(x) - L| < \epsilon$.



In this case, we write

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

Note that the positive superscript, “+”, indicates that this is a limit from the *right*.

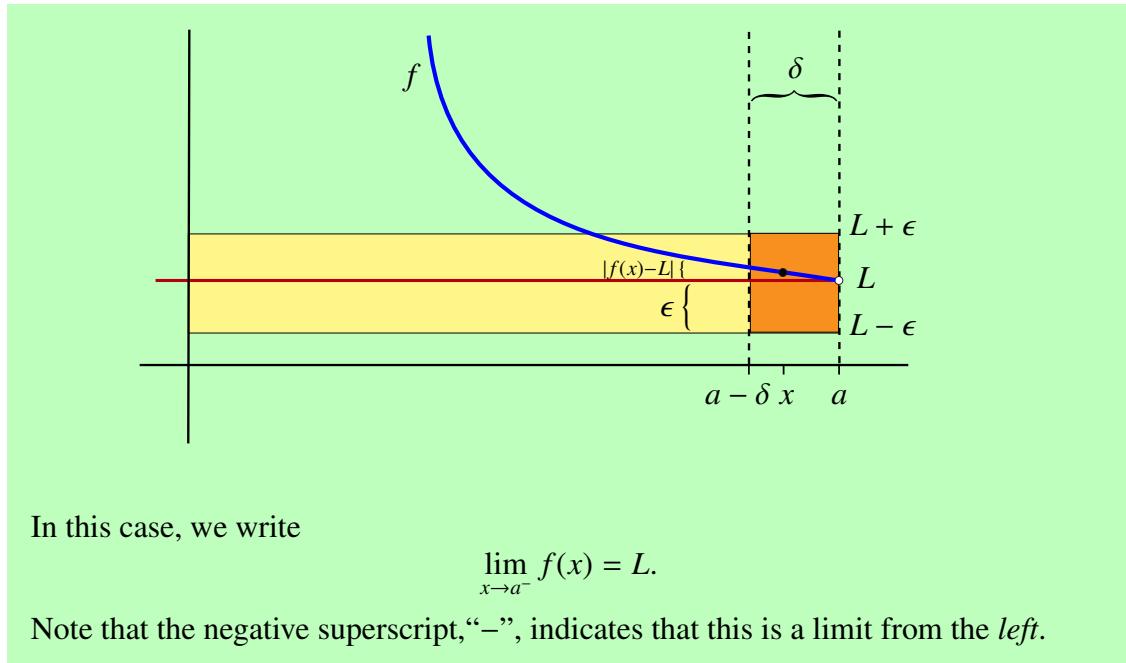
We can now present a similar definition for *limits from the left*.

DEFINITION

Limit from the Left

Let f be a function and let $a \in \mathbb{R}$.

We say that f has a limit L as x approaches a from the *left*, or from below, if for any positive tolerance $\epsilon > 0$, we can find a cutoff distance $\delta > 0$ such that if the distance from x to a is less than δ , and if $x < a$, then $f(x)$ approximates L with an error less than ϵ . That is, if $0 < a - x < \delta$, then $|f(x) - L| < \epsilon$.



You may have noticed an obvious connection between one-sided limits and two-sided limits. In particular, if a function f has a two-sided limit at a point $x = a$, then it is easy to see that both one-sided limits exist as well and that they have the same value as the two-sided limit. (Formally, given a tolerance $\epsilon > 0$, the cutoff distance $\delta > 0$ that works for the two-sided limit also works for both one-sided limits simultaneously).

The converse is a little bit more subtle. None the less, we could show that if both one-sided limits exist and if they are equal, then the two-sided limit also exists with this common value.

In fact, the following theorem summarizes the relationship between these two concepts.

THEOREM 6

One-sided versus Two-sided Limits

Let f be a function defined on an open interval containing $x = a$ except possibly at $x = a$. Then the following two statements are logically equivalent:

- 1) $\lim_{x \rightarrow a} f(x)$ exists and equals L .
- 2) Both one-sided limits exist, and

$$\lim_{x \rightarrow a^-} f(x) = L = \lim_{x \rightarrow a^+} f(x).$$

Finally, we note that there are also valid sequential versions for both one-sided limits and also that all of our arithmetic rules also hold for these limits.

5.5 The Squeeze Theorem

We had previously seen a version of the Squeeze Theorem for *sequences*. In this section, we introduce the Squeeze Theorem for *limits of functions*. We will then show how it can be used to calculate limits of some rather exotic functions.

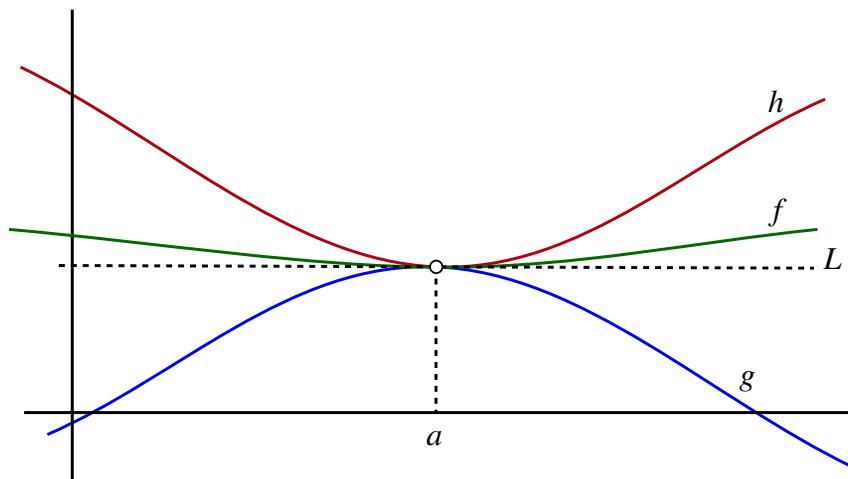
Assume that we have three functions, f , g and h , defined on an open interval I containing $x = a$, except possibly at $x = a$. Also assume that for each $x \in I$, except possibly $x = a$, we have

$$g(x) \leq f(x) \leq h(x)$$

and that

$$\lim_{x \rightarrow a} g(x) = L = \lim_{x \rightarrow a} h(x).$$

The picture illustrates these assumptions.



Notice that since

$$\lim_{x \rightarrow a} g(x) = L = \lim_{x \rightarrow a} h(x)$$

as x approaches a , both $g(x)$ and $h(x)$ get very close to L . But the graph of f is *squeezed* between the graphs of g and h near a . As such the values of $f(x)$ *must also be close to L near a*. In other words, the picture suggests that f also has a limit as x approaches a and that

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

In fact, like the case for sequences, this is also the case for functions as the next theorem shows. For obvious reasons we will also call this theorem the *Squeeze Theorem*.

THEOREM 7

Squeeze Theorem for Functions

Assume that three functions, f , g and h , are defined on an open interval I containing $x = a$, except possibly at $x = a$. Assume also that for each $x \in I$, except possibly $x = a$, that

$$g(x) \leq f(x) \leq h(x)$$

and that

$$\lim_{x \rightarrow a} g(x) = L = \lim_{x \rightarrow a} h(x).$$

Then $\lim_{x \rightarrow a} f(x)$ exists and

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

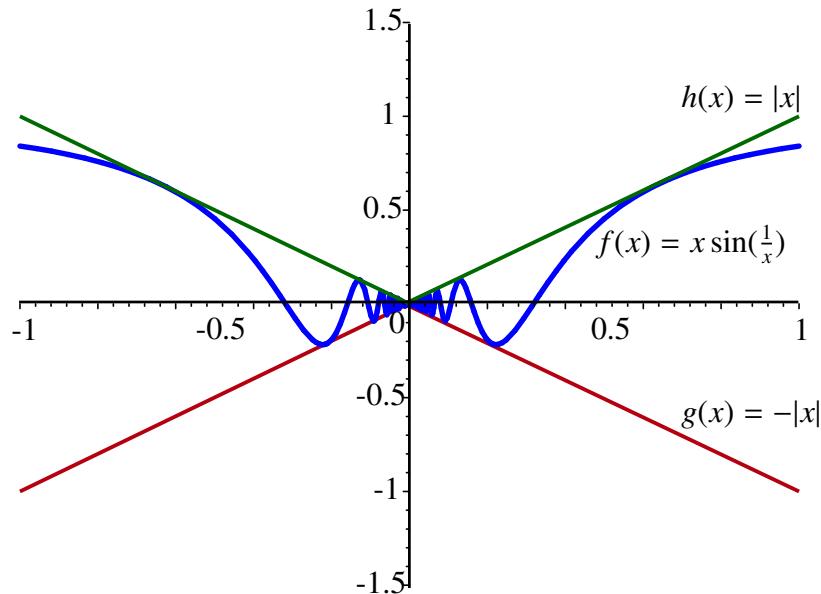
EXAMPLE 12 Earlier we looked at the rather unusual function $f(x) = \sin\left(\frac{1}{x}\right)$. We also saw that $\lim_{x \rightarrow 0} \sin\left(\frac{1}{x}\right)$ does not exist. However, if we let $f(x) = x \sin\left(\frac{1}{x}\right)$, then the result is different. In fact, since $\left|\sin\left(\frac{1}{x}\right)\right| \leq 1$ for all $x \neq 0$, we get that

$$\begin{aligned} \left|x \sin\left(\frac{1}{x}\right)\right| &= |x| \times \left|\sin\left(\frac{1}{x}\right)\right| \\ &\leq |x| \times 1 \\ &= |x| \end{aligned}$$

This gives us that

$$-|x| \leq x \sin\left(\frac{1}{x}\right) \leq |x|$$

for all $x \neq 0$ as shown in the following diagram:



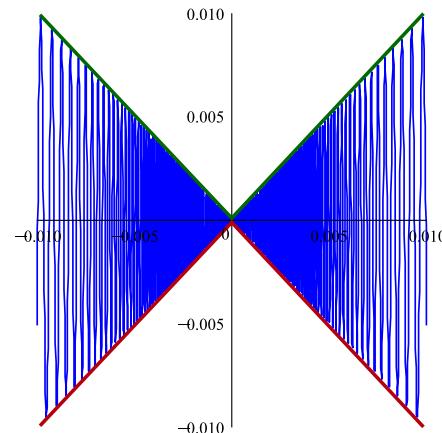
If we let $g(x) = -|x|$ and $h(x) = |x|$, then

$$\lim_{x \rightarrow 0} -|x| = 0 = \lim_{x \rightarrow 0} |x|.$$

Therefore the hypotheses for the Squeeze Theorem are satisfied and we can conclude that $\lim_{x \rightarrow 0} x \sin\left(\frac{1}{x}\right)$ exists and that

$$\lim_{x \rightarrow 0} x \sin\left(\frac{1}{x}\right) = 0.$$

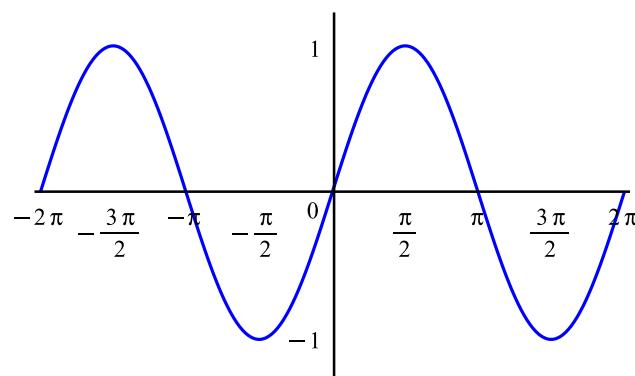
Before we end this example, let's take a look at the graph of f on the interval $[-.01, .01]$. This graph clearly supports our assertion that $\lim_{x \rightarrow 0} x \sin\left(\frac{1}{x}\right) = 0$.



It should be clear that there are corresponding versions of the Squeeze Theorem for both one-sided limits as well. As an illustration of this we have the following example:

EXAMPLE 13 A quick look at the graph of $f(\theta) = \sin(\theta)$ suggests that

$$\lim_{\theta \rightarrow 0} \sin(\theta) = 0.$$



We can use the Squeeze Theorem to justify this claim.

To see how this can be done we let $0 < \theta < \frac{\pi}{2}$. Recall that on the unit circle, the radian measure of an angle θ is equal to the length of the arc subtended by the angle. Then using the unit circle, we get that

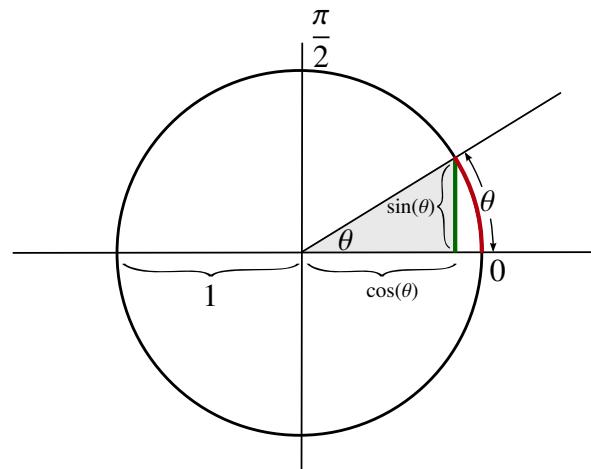
$$0 < \sin(\theta) < \theta.$$

However since

$$\lim_{\theta \rightarrow 0^+} 0 = 0 = \lim_{\theta \rightarrow 0^+} \theta$$

the Squeeze Theorem shows that

$$\lim_{\theta \rightarrow 0^+} \sin(\theta) = 0.$$



Now let $-\frac{\pi}{2} < \theta < 0$. We note that $\sin(\theta)$ is an *odd* function. That is

$$\sin(-\theta) = -\sin(\theta).$$

Moreover, as $\theta \rightarrow 0^-$, we have $-\theta \rightarrow 0^+$. From this we can deduce that

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

Since

$$\lim_{\theta \rightarrow 0^-} \sin(\theta) = 0 = \lim_{\theta \rightarrow 0^+} \sin(\theta)$$

it follows that

$$\lim_{\theta \rightarrow 0} \sin(\theta) = 0.$$

We observe that for $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$, we have

$$\cos(\theta) = \sqrt{1 - \sin^2(\theta)}.$$

Using the Arithmetic Rules for Limits we get

$$\begin{aligned} \lim_{\theta \rightarrow 0} \cos(\theta) &= \lim_{\theta \rightarrow 0} \sqrt{1 - \sin^2(\theta)} \\ &= \sqrt{1 - (\lim_{\theta \rightarrow 0} \sin(\theta))^2} \\ &= 1 \end{aligned}$$

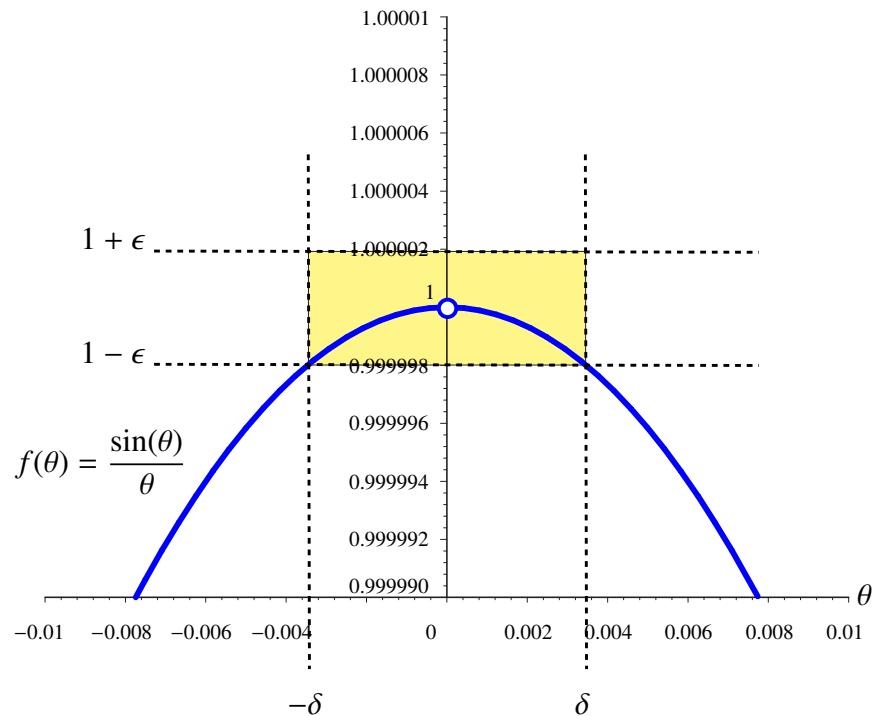
In the next section we will see how the Squeeze Theorem can be used to establish a very important trigonometric limit that we will use to evaluate the derivatives of the basic trigonometric functions.

5.6 The Fundamental Trigonometric Limit

One of the most important limits involving trigonometric functions is

$$\lim_{\theta \rightarrow 0} \frac{\sin(\theta)}{\theta}.$$

Recall that as θ approaches 0, so does $\sin(\theta)$. Consequently, it is difficult to determine how the function $\frac{\sin(\theta)}{\theta}$ behaves near 0. To help us understand, we can look at a plot of the function on the interval $[-.01, .01]$.



It appears from the graph that as θ approaches 0, $\frac{\sin(\theta)}{\theta}$ gets very close to 1. In fact, the graph demonstrates how to find the δ corresponding to a tolerance as small as 0.000002 if we applied the definition of the limit with $L = 1$. Still, this does not actually prove that 1 is the limit. It turns out that we can give a geometric argument that relies on a clever application of the Squeeze Theorem to verify that the limit is actually 1.

To simplify matters, we will only calculate

$$\lim_{\theta \rightarrow 0^+} \frac{\sin(\theta)}{\theta}.$$

Choose θ with $0 < \theta < \frac{\pi}{2}$. Consider the following diagram of a circle with radius 1 (the *unit circle*).

We have three distinct regions, a small triangle, a sector of the circle, and a larger triangle superimposed on one another. By simply comparing these areas we will be able to use the Squeeze Theorem to establish the desired limit.

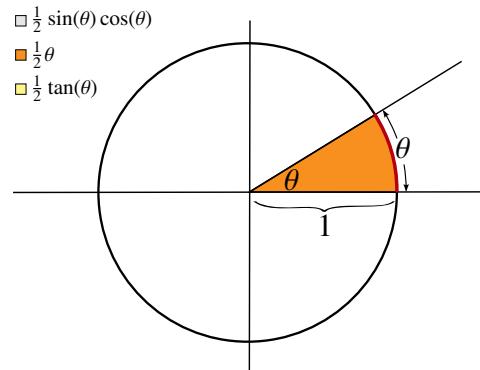
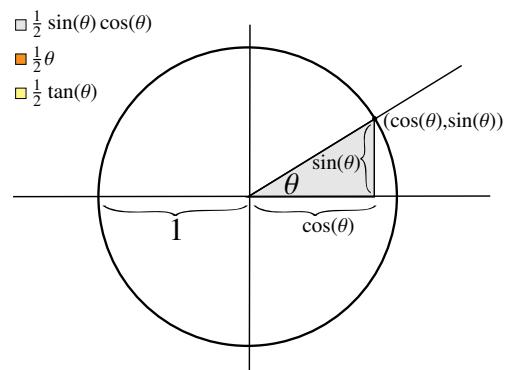
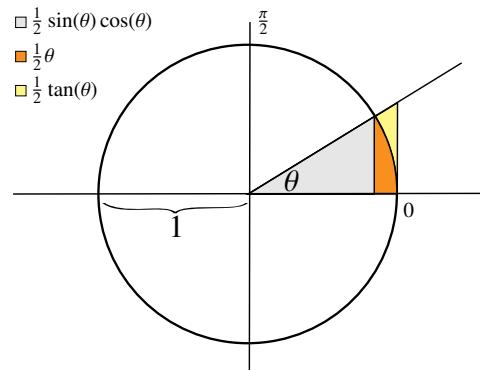
As noted, the previous diagram identifies three regions. The first region we will consider is the small triangle.

Recall from trigonometry that the triangle has a base of length $\cos(\theta)$ and height $\sin(\theta)$. Therefore, the area of this triangle is

$$\frac{1}{2} \text{ base} \times \text{height} = \frac{\sin(\theta) \cos(\theta)}{2}.$$

The second area is the sector shown the following diagram.

Since the circle has radius 1, the area of the circle is π . The area of the sector is found by multiplying the fraction of the circle represented by the sector by the total area of the circle. Since the full circle is made up of an arc with 2π radians and the sector has an arc that measures θ radians, the fraction of the circle taken up by the sector is $\frac{\theta}{2\pi}$. It follows that the area of the sector is

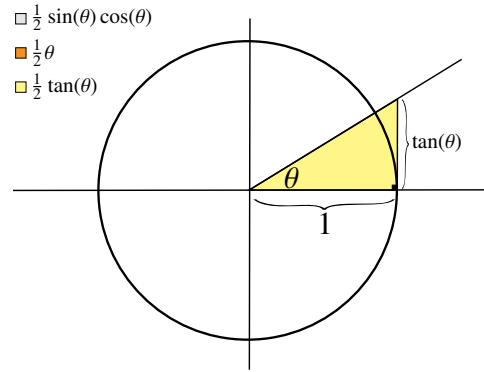


$$\begin{aligned} \text{Sector area} &= \text{fraction of circle} \times \text{area of circle} \\ &= \frac{\theta}{2\pi} \times \pi \\ &= \frac{\theta}{2} \end{aligned}$$

The third region is the largest of the three. It is the outside triangle as indicated in the diagram.

The triangle is a right triangle with base 1. Since $\tan(\theta) = \frac{\text{opposite}}{\text{adjacent}} = \frac{\text{opposite}}{1}$, the triangle must have height equal to $\tan(\theta)$. Therefore, the area of the triangle is equal to

$$\frac{\tan(\theta)}{2}.$$



These regions are listed in order of increasing area, so

$$\frac{\sin(\theta) \cos(\theta)}{2} < \frac{\theta}{2} < \frac{\tan(\theta)}{2}.$$

The next step is to multiply every term in this inequality by $\frac{2}{\sin(\theta)}$ to get

$$\cos(\theta) < \frac{\theta}{\sin(\theta)} < \frac{1}{\cos(\theta)}.$$

Then take reciprocals and reverse the order of the inequalities to get

$$\frac{1}{\cos(\theta)} > \frac{\sin(\theta)}{\theta} > \cos(\theta).$$

We know from the properties of the cosine function that

$$\lim_{\theta \rightarrow 0^+} \cos(\theta) = 1$$

and hence that

$$\lim_{\theta \rightarrow 0^+} \frac{1}{\cos(\theta)} = 1.$$

We can now use the Squeeze Theorem to conclude that

$$\lim_{\theta \rightarrow 0^+} \frac{\sin(\theta)}{\theta} = 1.$$

A similar calculation can be done to show that

$$\lim_{\theta \rightarrow 0^-} \frac{\sin(\theta)}{\theta} = 1.$$

This gives us the Fundamental Trigonometric Limit.

THEOREM 8

The Fundamental Trigonometric Limit

$$\lim_{\theta \rightarrow 0} \frac{\sin(\theta)}{\theta} = 1.$$

The Fundamental Trigonometric Limit tells us is that if “ θ is small”, then

$$\sin(\theta) \cong \theta.$$

This principle is actually quite useful and can be valuable in calculating other limits.

EXAMPLE 14 Find

$$\lim_{\theta \rightarrow 0} \frac{\sin(3\theta)}{\sin(\theta)}.$$

SOLUTION As $\theta \rightarrow 0$, we have that $3\theta \rightarrow 0$. This means that if “ θ is small”, so is 3θ . We know that if θ is small, then

$$\sin(\theta) \cong \theta.$$

Similarly, if 3θ is small, then we would expect that

$$\sin(3\theta) \cong 3\theta.$$

Putting these two statements together leads us to the possibility that if θ is small, then

$$\frac{\sin(3\theta)}{\sin(\theta)} \cong \frac{3\theta}{\theta} = 3.$$

We might guess that

$$\lim_{\theta \rightarrow 0} \frac{\sin(3\theta)}{\sin(\theta)} = 3.$$

This is in fact the case.

To see how we can make this rigorous, first note that the Fundamental Trigonometric Limit also shows that

$$\lim_{\theta \rightarrow 0} \frac{\sin(3\theta)}{3\theta} = 1$$

and that

$$\begin{aligned} \lim_{\theta \rightarrow 0} \frac{\theta}{\sin(\theta)} &= \lim_{\theta \rightarrow 0} \frac{1}{\frac{\sin(\theta)}{\theta}} \\ &= \frac{1}{1} \\ &= 1. \end{aligned}$$

Since

$$\frac{\sin(3\theta)}{\sin(\theta)} = 3 \left(\frac{\sin(3\theta)}{3\theta} \right) \left(\frac{\theta}{\sin(\theta)} \right),$$

we get

$$\begin{aligned} \lim_{\theta \rightarrow 0} \frac{\sin(3\theta)}{\sin(\theta)} &= \lim_{\theta \rightarrow 0} 3 \left(\frac{\sin(3\theta)}{3\theta} \right) \left(\frac{\theta}{\sin(\theta)} \right) \\ &= 3 \left(\lim_{\theta \rightarrow 0} \frac{\sin(3\theta)}{3\theta} \right) \left(\lim_{\theta \rightarrow 0} \frac{\theta}{\sin(\theta)} \right) \\ &= 3(1)(1) \\ &= 3. \end{aligned}$$

EXAMPLE 15 Show

$$\lim_{\theta \rightarrow 0} \frac{\tan(\theta)}{\theta} = 1.$$

SOLUTION Observe that

$$\begin{aligned} \lim_{\theta \rightarrow 0} \frac{\tan(\theta)}{\theta} &= \lim_{\theta \rightarrow 0} \left(\frac{\sin(\theta)}{\theta \cos(\theta)} \right) \\ &= \lim_{\theta \rightarrow 0} \left(\frac{1}{\cos(\theta)} \right) \left(\frac{\sin(\theta)}{\theta} \right) \\ &= \lim_{\theta \rightarrow 0} \left(\frac{1}{\cos(\theta)} \right) \lim_{\theta \rightarrow 0} \left(\frac{\sin(\theta)}{\theta} \right) \\ &= (1)(1) \\ &= 1. \end{aligned}$$



EXAMPLE 16 Find

$$\lim_{\theta \rightarrow 0} \frac{\tan(\theta)}{\sin(2\theta)}.$$

SOLUTION Observe that

$$\begin{aligned} \lim_{\theta \rightarrow 0} \frac{\tan(\theta)}{\sin(2\theta)} &= \lim_{\theta \rightarrow 0} \left(\frac{\sin(\theta)}{\cos(\theta)} \right) \left(\frac{1}{\sin(2\theta)} \right) \\ &= \lim_{\theta \rightarrow 0} \left(\frac{1}{\cos(\theta)} \right) \left(\frac{\sin(\theta)}{\sin(2\theta)} \right) \\ &= \lim_{\theta \rightarrow 0} \left(\frac{1}{\cos(\theta)} \right) \left(\frac{\sin(\theta)}{\theta} \right) \left(\frac{2\theta}{\sin(2\theta)} \right) \left(\frac{1}{2} \right) \\ &= \left(\frac{1}{2} \right) \lim_{\theta \rightarrow 0} \left(\frac{1}{\cos(\theta)} \right) \lim_{\theta \rightarrow 0} \left(\frac{\sin(\theta)}{\theta} \right) \lim_{\theta \rightarrow 0} \left(\frac{2\theta}{\sin(2\theta)} \right) \\ &= \left(\frac{1}{2} \right) (1)(1)(1) \\ &= \frac{1}{2}. \end{aligned}$$



5.7 Limits at Infinity and Asymptotes

In this section we extend the concept of a limit in two ways. In particular, we will define:

- *limits at infinity*, where x becomes arbitrarily large, either positive or negative; and
- *infinite limits*, where the function grows without bound near a particular point.

Note: It is important to recognize that the symbol “ ∞ ” is not a real number. When we say that “the limit of a function is infinity”, we are not saying that the limit exists in the proper sense. Instead, this expression simply provides useful information about the behavior of functions whose values become arbitrarily large, either positive or negative.

5.7.1 Asymptotes and Limits at Infinity

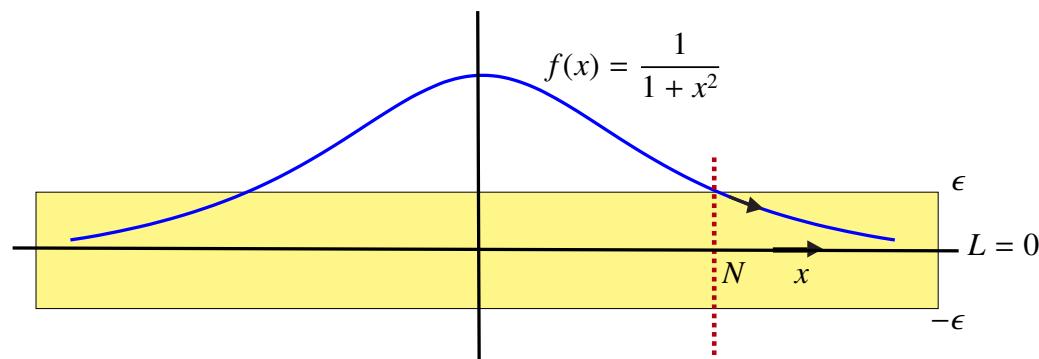
So far, whenever we have considered limits we have always focused on the behavior of a function near a particular point. In this section we will be concerned with what happens when we allow the variable to approach either ∞ or $-\infty$. For example, if we let

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

then as x gets very large, $1+x^2$ also gets very large. Consequently, $\frac{1}{1+x^2}$ becomes very small or very close to 0. That is, as “ x approaches ∞ ”, $f(x)$ tends to zero. This leads us to say that 0 is the limit of $f(x)$ as x goes to ∞ .

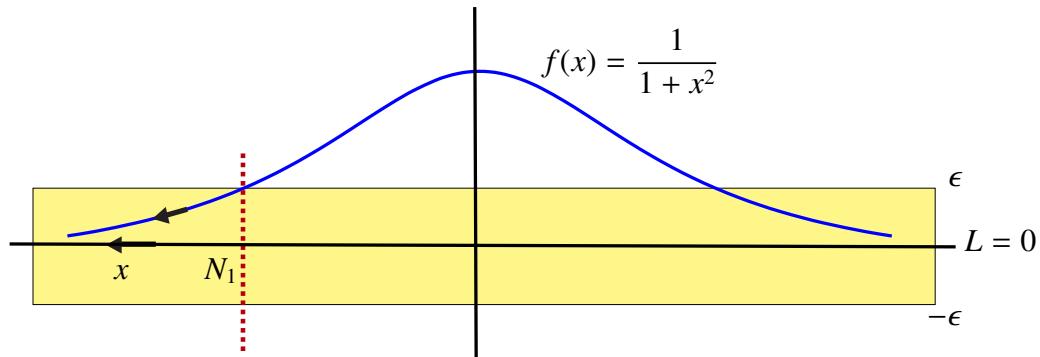
Similarly, as x approaches $-\infty$, we again have that $1+x^2$ gets very large. It follows that $\frac{1}{1+x^2}$ also becomes very small. We say that 0 is the limit of $f(x)$ as x goes to $-\infty$.

We want to make the notion of limits at $\pm\infty$ precise. To do this we mimic what we did for ordinary limits and take our lead from what we did for sequences. In the example above, if we are given a positive tolerance $\epsilon > 0$, we can always find a cutoff N such that if $x > N$, then $f(x)$ approximates 0 with an error less than ϵ .



For limits at infinity, the cutoff N plays the role of our cutoff distance δ in our previous definition of a limit. It tells us how far out we must be so that $f(x)$ approximates the limit within the given tolerance. Generally speaking, the smaller the tolerance ϵ , the larger N must be.

Similarly, given a tolerance $\epsilon > 0$, we can find a cutoff N_1 such that if $x < N_1$, then $f(x)$ approximates 0 with an error less than ϵ .



This leads us to the following definition:

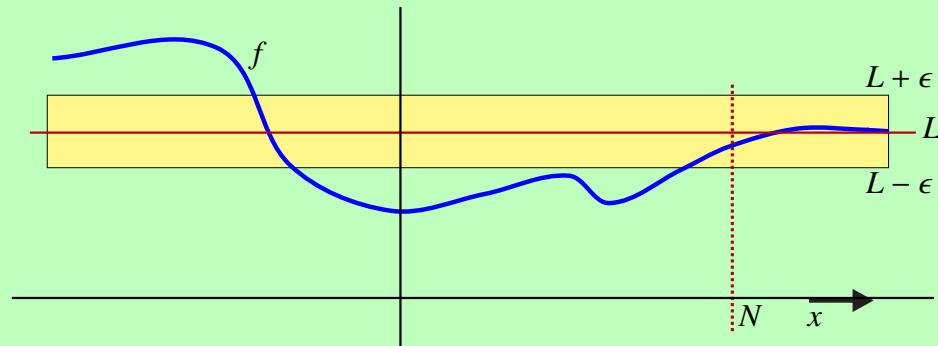
DEFINITION

Limits at Infinity

We say that a function f has a limit L as x approaches ∞ if for every positive tolerance $\epsilon > 0$, we can always find a cutoff $N > 0$ such that if $x > N$, then $f(x)$ approximates L with an error less than ϵ .

That is,

$$\text{if } x > N, \text{ then } |f(x) - L| < \epsilon.$$



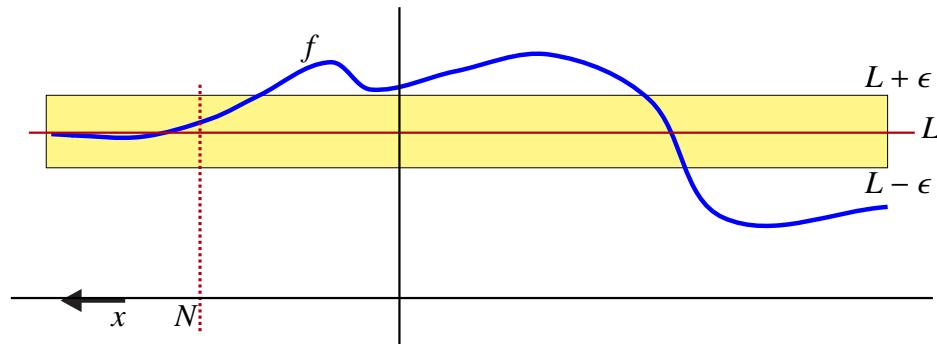
In this case, we write

$$\lim_{x \rightarrow \infty} f(x) = L.$$

We can also define limits at $-\infty$ in a similar manner. In particular, we say that a function f has a limit L as x approaches $-\infty$ if for every positive tolerance $\epsilon > 0$, we can always find a cutoff N such that if $x < N$, then $f(x)$ approximates L with an error less than ϵ .

That is,

$$\text{if } x < N, \text{ then } |f(x) - L| < \epsilon.$$



This time we write

$$\lim_{x \rightarrow -\infty} f(x) = L.$$

Assume that $\lim_{x \rightarrow \infty} f(x) = L$. Then for large enough values of x , the graph of f is as near as we would like to the line $y = L$.

Similarly, assume that $\lim_{x \rightarrow -\infty} f(x) = L$. Then again for large enough negative values of x , the graph of f is as near as we would like to the line $y = L$.

This leads us to the following definition:

DEFINITION

Horizontal Asymptote

Assume that $\lim_{x \rightarrow \infty} f(x) = L$ or $\lim_{x \rightarrow -\infty} f(x) = L$.

Then in either case, we say that the line $y = L$ is a *horizontal asymptote* of f .

Like the case of sequences, we can define the divergence of a function to $\pm\infty$ as x approaches either ∞ or $-\infty$.

DEFINITION

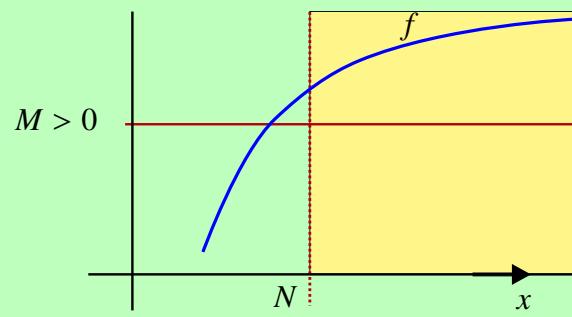
Infinite Limits at ∞

We say that the limit of $f(x)$ as x approaches ∞ is ∞ if for every $M > 0$ there exists a cutoff $N > 0$ such that if $x > N$, then

$$f(x) > M.$$

We write

$$\lim_{x \rightarrow \infty} f(x) = \infty.$$



Note: Similarly, we can define $\lim_{x \rightarrow \infty} f(x) = -\infty$ and $\lim_{x \rightarrow -\infty} f(x) = \pm\infty$.

It is both useful and important to note that all of the usual rules for the arithmetic of limits hold for limits at $\pm\infty$. In fact, the Squeeze Theorem also holds with the proper modifications.

THEOREM 9

Squeeze Theorem for Limits at $\pm\infty$

Assume that $g(x) \leq f(x) \leq h(x)$ for all $x \geq N$. If

$$\lim_{x \rightarrow \infty} g(x) = L = \lim_{x \rightarrow \infty} h(x)$$

then $\lim_{x \rightarrow \infty} f(x)$ exists and it equals L .

Assume that $g(x) \leq f(x) \leq h(x)$ for all $x \leq N$. If

$$\lim_{x \rightarrow -\infty} g(x) = L = \lim_{x \rightarrow -\infty} h(x)$$

then $\lim_{x \rightarrow -\infty} f(x)$ exists and it equals L .

Let's look at some examples of limits at $\pm\infty$.

EXAMPLE 17

Evaluate

$$\lim_{x \rightarrow \infty} \frac{2x^2 - 3x + 4}{x^2 + x - 5}.$$

SOLUTION Observe that when dealing with polynomials, for large values of x the highest power terms dominate. This means that we might expect that if x is very large then

$$\frac{2x^2 - 3x + 4}{x^2 + x - 5} \cong \frac{2x^2}{x^2} = 2.$$

From this we might guess that

$$\lim_{x \rightarrow \infty} \frac{2x^2 - 3x + 4}{x^2 + x - 5} = 2.$$

The limit rules can be used to show that this guess is correct. First factor out the highest power of x from both the numerator and the denominator. In this case, factor out x^2 from both the numerator and the denominator to get

$$\begin{aligned} \frac{2x^2 - 3x + 4}{x^2 + x - 5} &= \frac{x^2(2 - \frac{3}{x} + \frac{4}{x^2})}{x^2(1 + \frac{1}{x} - \frac{5}{x^2})} \\ &= \frac{2 - \frac{3}{x} + \frac{4}{x^2}}{1 + \frac{1}{x} - \frac{5}{x^2}} \end{aligned}$$

for all $x > 0$. But then

$$\begin{aligned}\lim_{x \rightarrow \infty} \frac{2x^2 - 3x + 4}{x^2 + x - 5} &= \lim_{x \rightarrow \infty} \frac{x^2(2 - \frac{3}{x} + \frac{4}{x^2})}{x^2(1 + \frac{1}{x} - \frac{5}{x^2})} \\ &= \lim_{x \rightarrow \infty} \frac{2 - \frac{3}{x} + \frac{4}{x^2}}{1 + \frac{1}{x} - \frac{5}{x^2}} \\ &= \frac{2 - 0 + 0}{1 + 0 - 0} \\ &= 2\end{aligned}$$

exactly as we had predicted. 

EXAMPLE 18 Evaluate

$$\lim_{x \rightarrow \infty} \frac{3x^3 + x}{x^2 + 1}.$$

SOLUTION We follow the same procedure as was outlined in the previous example and write

$$\begin{aligned}\frac{3x^3 + x}{x^2 + 1} &= \frac{x^3(3 + \frac{1}{x^2})}{x^2(1 + \frac{1}{x^2})} \\ &= x \left(\frac{3 + \frac{1}{x^2}}{1 + \frac{1}{x^2}} \right)\end{aligned}$$

Now, as x approaches ∞ , $\frac{3 + \frac{1}{x^2}}{1 + \frac{1}{x^2}}$ approaches 3. This means that for very large values of x , the function behaves much like $y = 3x$. In particular, the function $f(x)$ grows without bound and as such does *not* have a limit. We conclude that $\lim_{x \rightarrow \infty} \frac{3x^3 + x}{x^2 + 1}$ does not exist. 

EXAMPLE 19 Evaluate

$$\lim_{x \rightarrow -\infty} \frac{3x^3 + x}{x^4 + 1}.$$

SOLUTION This is very similar to the previous example. We write

$$\begin{aligned}\frac{3x^3 + x}{x^4 + 1} &= \frac{x^3(3 + \frac{1}{x^2})}{x^4(1 + \frac{1}{x^4})} \\ &= \frac{1}{x} \left(\frac{3 + \frac{1}{x^2}}{1 + \frac{1}{x^4}} \right)\end{aligned}$$

Just as before, we have that as x approaches $-\infty$, $\frac{3x^3 + x}{1 + x^4}$ approaches 3. This means that for large negative values of x , we have $f(x) \approx \frac{3}{x}$. From this we can conclude that $f(x)$ approaches 0 as x approaches $-\infty$. That is,

$$\lim_{x \rightarrow -\infty} \frac{3x^3 + x}{x^4 + 1} = 0.$$



In general, for a rational function

$$f(x) = \frac{p(x)}{q(x)} = \frac{a_n x^n + \dots + a_1 x + a_0}{b_m x^m + \dots + b_1 x + b_0}$$

the existence of the limit at $\pm\infty$ depends on the relative degrees of the polynomials. If $n > m$, then the numerator grows much faster than the denominator and the function will eventually grow without bounds. This means that *no limit exists*.

If $n < m$, then the denominator will grow faster than the numerator. This means that the limit of the function tends towards 0. That is if $n < m$, then

$$\lim_{x \rightarrow \pm\infty} \frac{a_n x^n + \dots + a_1 x + a_0}{b_m x^m + \dots + b_1 x + b_0} = 0.$$

The most interesting situation occurs when $n = m$. In this case, you can factor out $x^n = x^m$ from both the numerator and the denominator, and then follow the procedure we used in our previous examples to show that

$$\lim_{x \rightarrow \pm\infty} \frac{a_n x^n + \dots + a_1 x + a_0}{b_n x^n + \dots + b_1 x + b_0} = \frac{a_n}{b_n}.$$

We stated that the Squeeze Theorem holds for limits at $\pm\infty$. The next example illustrates how it can be used.

EXAMPLE 20

Evaluate

$$\lim_{x \rightarrow \infty} \frac{\sin(x)}{x}.$$

SOLUTION We know that for any x ,

$$|\sin(x)| \leq 1.$$

It follows that if $x \neq 0$, then

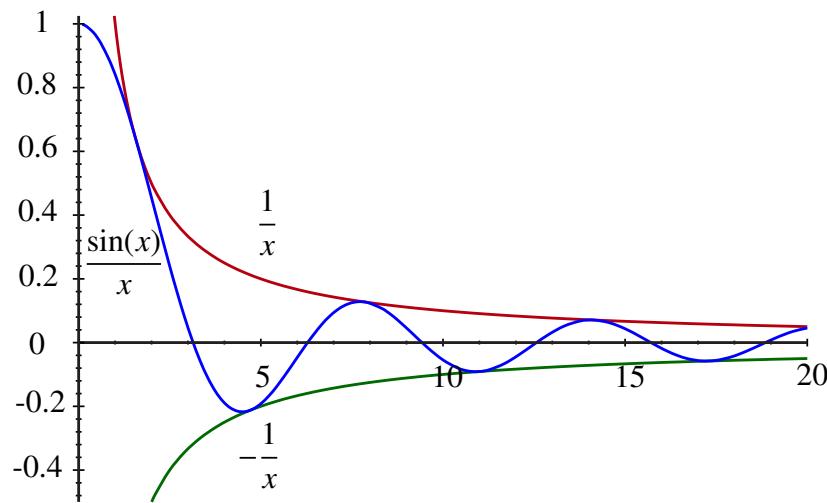
$$\left| \frac{\sin(x)}{x} \right| \leq \frac{1}{|x|}$$

and hence that

$$\frac{-1}{|x|} \leq \frac{\sin(x)}{x} \leq \frac{1}{|x|}.$$

For $x > 0$, this inequality becomes

$$\frac{-1}{x} \leq \frac{\sin(x)}{x} \leq \frac{1}{x}.$$



This expression is exactly what we require to apply the Squeeze Theorem. In fact, we know that

$$\lim_{x \rightarrow \infty} \frac{-1}{x} = 0 = \lim_{x \rightarrow \infty} \frac{1}{x}.$$

We can now apply the Squeeze Theorem to get that $\lim_{x \rightarrow \infty} \frac{\sin(x)}{x}$ exists and that

$$\lim_{x \rightarrow \infty} \frac{\sin(x)}{x} = 0.$$

A similar argument shows that

$$\lim_{x \rightarrow -\infty} \frac{\sin(x)}{x} = 0.$$



In the next section, we will again use the Squeeze Theorem to establish some fundamental results concerning the growth rate of logarithmic functions.

5.7.2 Fundamental Log Limit

In this section, we consider the limit

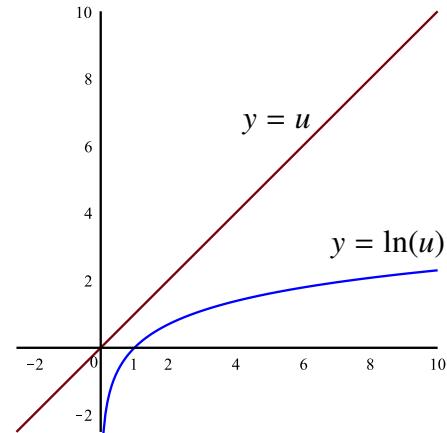
$$\lim_{x \rightarrow \infty} \frac{\ln(x)}{x}.$$

This limit is called the *Fundamental Log Limit*. From this limit we will be able to derive a great deal of information about the relative growth rates of logarithmic, polynomial, and exponential functions.

First, note that $\ln(x)$ grows much more slowly than x . For example, $\ln(10000) = 9.210340468$. This would lead us to guess that

$$\lim_{x \rightarrow \infty} \frac{\ln(x)}{x} = 0.$$

To make this idea more precise, we begin with the observation that for each $u > 0$, we have $\ln(u) < u$. This is illustrated in the diagram.

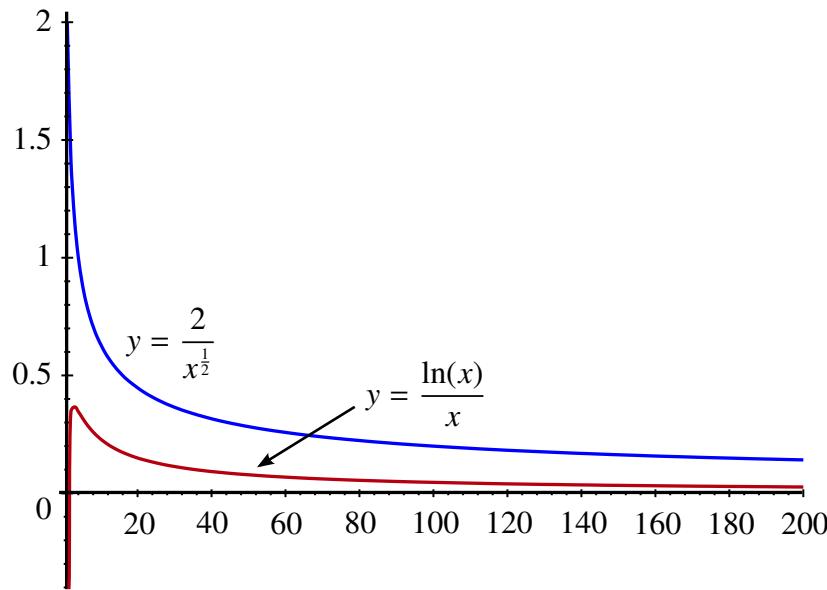


Now if $u \geq 1$, then $\ln(u) \geq 0$. We can use an algebraic trick to get an inequality that will help us find the limit. In particular, if we let $u = x^{\frac{1}{2}}$ and let $x \geq 1$, recognizing that $x = x^{\frac{1}{2}} \cdot x^{\frac{1}{2}}$, we get that

$$0 \leq \frac{\ln(x)}{x} = \frac{2 \ln(x^{\frac{1}{2}})}{x} = \frac{2}{x^{\frac{1}{2}}} \left(\frac{\ln(x^{\frac{1}{2}})}{x^{\frac{1}{2}}} \right) \leq \frac{2}{x^{\frac{1}{2}}}.$$

In summary, we have

$$0 \leq \frac{\ln(x)}{x} \leq \frac{2}{x^{\frac{1}{2}}}.$$



This inequality is exactly what we require since

$$\lim_{x \rightarrow \infty} 0 = 0 = \lim_{x \rightarrow \infty} \frac{2}{x^{\frac{1}{2}}}.$$

From this result we can use the Squeeze Theorem to get:

THEOREM 10
Fundamental Log Limit

$$\lim_{x \rightarrow \infty} \frac{\ln(x)}{x} = 0.$$

The Fundamental Log Limit shows that $\ln(x)$ grows more slowly than x . What about the growth rate of $\ln(x)$ versus that of \sqrt{x} or $x^{\frac{1}{100}}$?

For example, we have already seen that $\ln(10000) = 9.210340468$ and we have

$$10000^{\frac{1}{100}} = 1.096478196.$$

We might guess that $x^{\frac{1}{100}}$ actually grows more slowly than $\ln(x)$. However, the results for $x = 10000$ are deceptive. While 10000 may seem like a large number, in the big scheme of things, it is not. While it may take some time to get going, eventually the function $x^{\frac{1}{100}}$ surpasses $\ln(x)$. In fact, we can show that

$$\lim_{x \rightarrow \infty} \frac{\ln(x)}{x^{\frac{1}{100}}} = 0$$

so $\ln(x)$ is eventually dominated by $x^{\frac{1}{100}}$. Moreover, as the next example shows, for any $p > 0$, no matter how small, x^p eventually dominates $\ln(x)$.

EXAMPLE 21

If $p > 0$, then

$$\lim_{x \rightarrow \infty} \frac{\ln(x)}{x^p} = 0.$$

To show this, we use the Fundamental Log Limit and a small trick. First, write

$$\frac{\ln(x)}{x^p} = \frac{\frac{1}{p} \ln(x^p)}{x^p}.$$

Since p is a constant, it follows that

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{\ln(x)}{x^p} &= \lim_{x \rightarrow \infty} \frac{\frac{1}{p} \ln(x^p)}{x^p} \\ &= \frac{1}{p} \lim_{x \rightarrow \infty} \frac{\ln(x^p)}{x^p}. \end{aligned}$$

However, if $p > 0$ and $x \rightarrow \infty$, then $x^p \rightarrow \infty$. Replacing x by x^p in the Fundamental Log Limit gives us

$$\lim_{x \rightarrow \infty} \frac{\ln(x^p)}{x^p} = 0.$$

From this, we get

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{\ln(x)}{x^p} &= \frac{1}{p} \lim_{x \rightarrow \infty} \frac{\ln(x^p)}{x^p} \\ &= 0. \end{aligned}$$



EXAMPLE 22 Evaluate

$$\lim_{x \rightarrow \infty} \frac{\ln(x^p)}{x}.$$

SOLUTION This is a simple variant of the Fundamental Log Limit since

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{\ln(x^p)}{x} &= \lim_{x \rightarrow \infty} \frac{p \ln(x)}{x} \\ &= p \lim_{x \rightarrow \infty} \frac{\ln(x)}{x} \\ &= 0. \end{aligned}$$



EXAMPLE 23 Evaluate

$$\lim_{x \rightarrow \infty} \frac{\ln(x^{40})}{x^{\frac{1}{1000}}}.$$

SOLUTION We know that

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{\ln(x^{40})}{x^{\frac{1}{1000}}} &= \lim_{x \rightarrow \infty} \frac{40 \ln(x)}{x^{\frac{1}{1000}}} \\ &= 40 \lim_{x \rightarrow \infty} \frac{\ln(x)}{x^{\frac{1}{1000}}} \\ &= 0 \end{aligned}$$

by using the result from a previous example.



The limit in the last example follows easily from the Fundamental Log Limit, but it might not be something that we would guess from numerical testing. For example, if $f(x) = \frac{\ln(x^{40})}{x^{\frac{1}{1000}}}$, then $f(1000000) = 545.0381916$. While this function eventually drops off to 0, it does take quite a while to do so.

So far, we have seen that $\ln(x)$ grows at a rate that is at least an *order of magnitude less* than a polynomial function. We will now show that exponential functions grow at a rate that is an *order of magnitude greater* than that of a polynomial function.

EXAMPLE 24 Let $p > 0$. Evaluate

$$\lim_{x \rightarrow \infty} \frac{x^p}{e^x}.$$

SOLUTION To find the limit, we first transform what we have into one of our previous limits by letting $u = e^x$. This means that $x = \ln(u)$. In this case, $\frac{x^p}{e^x}$ becomes

$$\frac{(\ln(u))^p}{u} = \left(\frac{\ln(u)}{u^{\frac{1}{p}}} \right)^p.$$

Note that if $x \rightarrow \infty$, then $u = e^x \rightarrow \infty$. From this we get

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{x^p}{e^x} &= \lim_{u \rightarrow \infty} \left(\frac{\ln(u)}{u^{\frac{1}{p}}} \right)^p \\ &= \left(\lim_{u \rightarrow \infty} \frac{\ln(u)}{u^{\frac{1}{p}}} \right)^p \\ &= 0^p \\ &= 0. \end{aligned}$$



We know that as x goes to 0 from above, $\ln(x)$ goes to $-\infty$. The next limit shows us that this happens rather slowly.

EXAMPLE 25 Let $p > 0$. Evaluate

$$\lim_{x \rightarrow 0^+} x^p \ln(x).$$

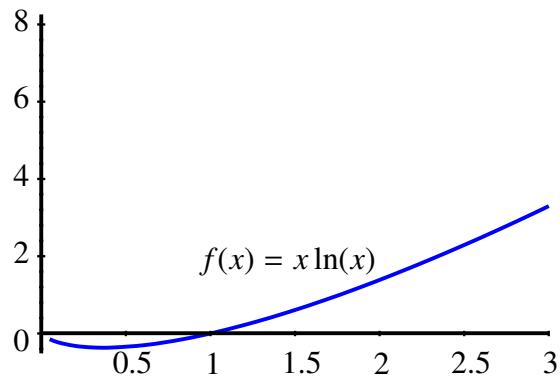
SOLUTION To find this limit, let $x = \frac{1}{u}$. Then

$$x^p \ln(x) = \frac{\ln(\frac{1}{u})}{u^p} = \frac{-\ln(u)}{u^p}.$$

If $x \rightarrow 0^+$, then $u \rightarrow \infty$. This gives us

$$\begin{aligned} \lim_{x \rightarrow 0^+} x^p \ln(x) &= \lim_{u \rightarrow \infty} \frac{-\ln(u)}{u^p} \\ &= 0. \end{aligned}$$

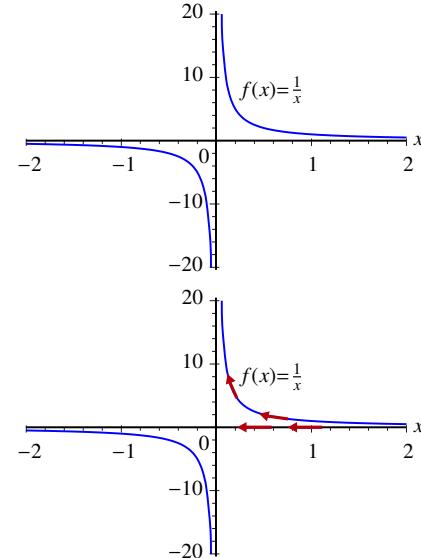
The diagram shows the graph of $f(x) = x \ln(x)$ near 0. It confirms our calculation by showing that as $x \rightarrow 0^+$, $f(x) \rightarrow 0$.



5.7.3 Vertical Asymptotes and Infinite Limits

Consider the function $f(x) = \frac{1}{x}$. We can see that as $x \rightarrow 0$, this function does not have a limit.

However, we can actually say more. As $x \rightarrow 0$ from above, the function is positive and it grows without bound. That is, $f(x)$ approaches ∞ .



It may be tempting to write

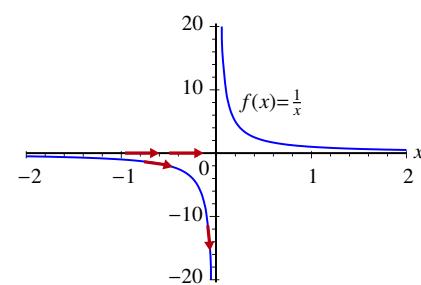
$$\lim_{x \rightarrow 0^+} \frac{1}{x} = \infty.$$

Similarly, as $x \rightarrow 0$ from below, the function is negative and it again grows without bound. This time we have that $f(x)$ approaches $-\infty$ so we might write

$$\lim_{x \rightarrow 0^-} \frac{1}{x} = -\infty.$$

Up until now the expressions

$$\lim_{x \rightarrow 0^+} \frac{1}{x} = \infty$$



and

$$\lim_{x \rightarrow 0^-} \frac{1}{x} = -\infty$$

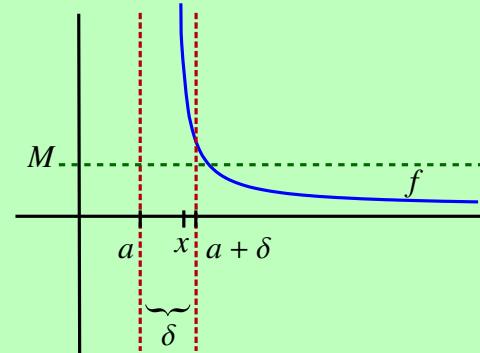
have had no formal meaning. However, they do tell us something important about the behaviour of the function $f(x) = \frac{1}{x}$ near $x = 0$.

In this section, we will see how to quantify the statements above. The key observation is that to approach ∞ we must eventually exceed any fixed positive number, and to approach $-\infty$ we must eventually be less than any fixed negative number.

DEFINITION

Right-Hand Infinite Limits

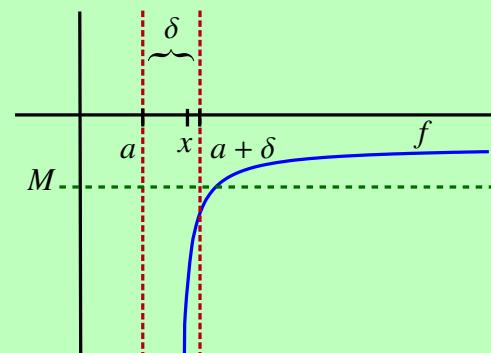
We say that f has a limit of ∞ as x approaches a from *above* if for every cutoff $M > 0$, we can find a cutoff distance $\delta > 0$ such that if $x > a$ and if the distance from x to a is less than δ , then $f(x) > M$. That is, if $a < x < a + \delta$, then $f(x) > M$.



In this case, we write

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

We say that f has a limit of $-\infty$ as x approaches a from *above* if for every cutoff $M < 0$, we can find a cutoff distance $\delta > 0$ such that if $x > a$ and if the distance from x to a is less than δ , then $f(x) < M$. That is, if $a < x < a + \delta$, then $f(x) < M$.

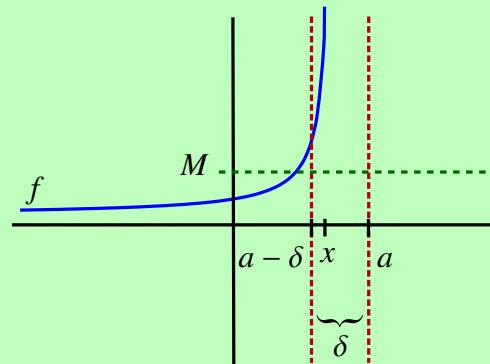


In this case, we write

$$\lim_{x \rightarrow a^+} f(x) = -\infty.$$

DEFINITION**Left-Hand Infinite Limits**

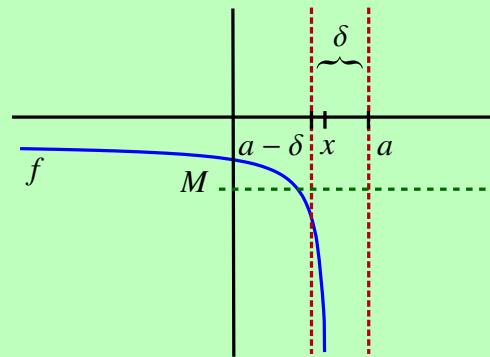
We say that f has a limit of ∞ as x approaches a from *below* if for every cutoff $M > 0$, we can find a cutoff distance $\delta > 0$ such that if $x < a$ and if the distance from x to a is less than δ , then $f(x) > M$. That is, if $a - \delta < x < a$, then $f(x) > M$.



In this case, we write

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

We say that f has a limit of $-\infty$ as x approaches a from *below* if for every cutoff $M < 0$, we can find a cutoff distance $\delta > 0$ such that if $x < a$ and if the distance from x to a is less than δ , then $f(x) < M$. That is, if $a - \delta < x < a$, then $f(x) < M$.



In this case, we write

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

DEFINITION**Infinite Limits**

We say that

$$\lim_{x \rightarrow a} f(x) = \infty$$

if

$$\lim_{x \rightarrow a^-} f(x) = \infty = \lim_{x \rightarrow a^+} f(x).$$

We say that

$$\lim_{x \rightarrow a} f(x) = -\infty$$

if

$$\lim_{x \rightarrow a^-} f(x) = -\infty = \lim_{x \rightarrow a^+} f(x).$$

DEFINITION**Vertical Asymptote**

If any of

$$\lim_{x \rightarrow a^\pm} f(x) = \pm\infty$$

occur, we say that the line $x = a$ is a *vertical asymptote* for the function f .

NOTE

It is important to note that despite our terminology and notation, when we write expressions such as

$$\lim_{x \rightarrow a} f(x) = \infty,$$

we do **not** mean to imply that the function f has a limit at the point $x = a$. The symbol “ ∞ ” is not a real number. In fact, what this expression actually tells us is that the limit of the function **does not exist** precisely because the function **grows without bounds** near a . A similar statement can be made for all of the other cases we have encountered in this section. This is a subtle point but one of which you must be aware. 

EXAMPLE 26

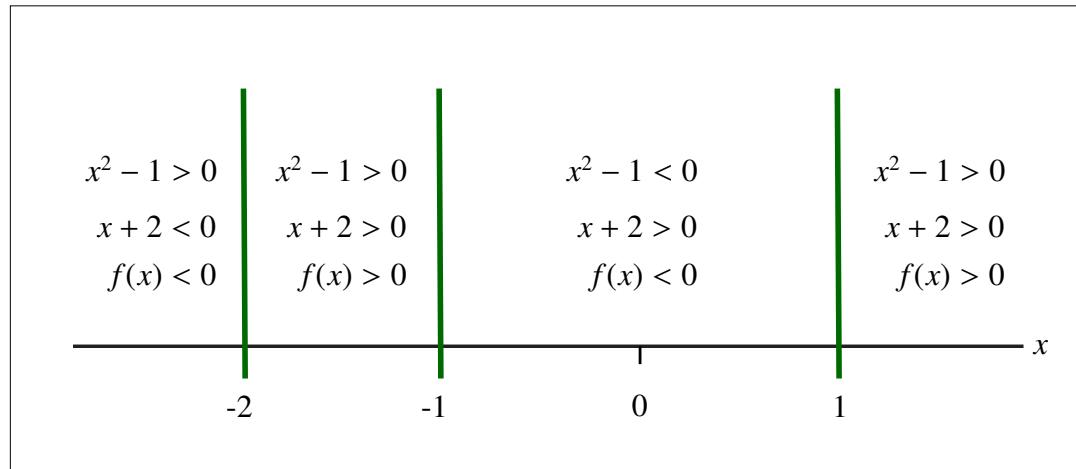
Let $f(x) = \frac{x^2-1}{x+2}$. Find

$$\lim_{x \rightarrow -2^+} f(x).$$

SOLUTION Let $p(x) = x^2 - 1$ and $q(x) = x + 2$. Since this function is a rational function of the form $\frac{p(x)}{q(x)}$, test to see if the limit exists by first evaluating the denominator $q(x)$ at $x = -2$. We get that $q(-2) = -2 + 2 = 0$. Next evaluate the numerator $p(x)$ at $x = -2$ to get that $p(-2) = (-2)^2 - 1 = 3 \neq 0$. Our previous work on limits tells us that when the denominator goes to 0, but the numerator does not, the limit does not exist. We also know that the magnitude of the quotient approaches ∞ . To say more we must consider the sign of the function near $x = -2$.

A rational function can only change sign if either the numerator or the denominator changes sign. This occurs at $x = \pm 1$ for the numerator and at $x = -2$ for the denominator. These points divide our domain into four regions; $x > 1$, $-1 < x < 1$, $-2 < x < -1$ and $x < -2$.

In the region $x > 1$, both the numerator and denominator are positive, so $f(x) > 0$. Moving to the left, when we cross $x = 1$ the numerator becomes negative while the denominator remains positive. This means that $f(x) < 0$ if $-1 < x < 1$. Crossing $x = -1$ returns the numerator to a positive value and as a result the function is also positive on the interval $-2 < x < -1$. Finally, when we cross $x = -2$, the denominator becomes negative and the numerator is positive. Hence, $f(x) < 0$ if $x < -2$. This information is summarized in the following diagram:



Since we are interested in the behaviour of $f(x)$ as x approaches -2 from above, we will focus on the region $-2 < x < -1$. In this region, the function is positive. We can conclude that

$$\lim_{x \rightarrow -2^+} \frac{x^2 - 1}{x + 2} = \infty.$$

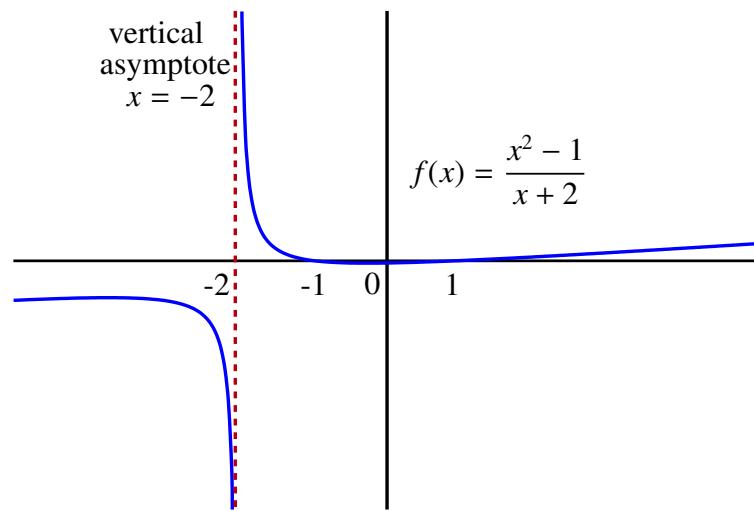
If we wanted to know

$$\lim_{x \rightarrow -2^-} f(x),$$

then our analysis would be similar. However, in this case we would be interested in the region $x < -2$. Since $f(x) < 0$ when $x < -2$, we have

$$\lim_{x \rightarrow -2^-} \frac{x^2 - 1}{x + 2} = -\infty.$$

The following is a graph of this function and it confirms our calculations.



5.8 Continuity

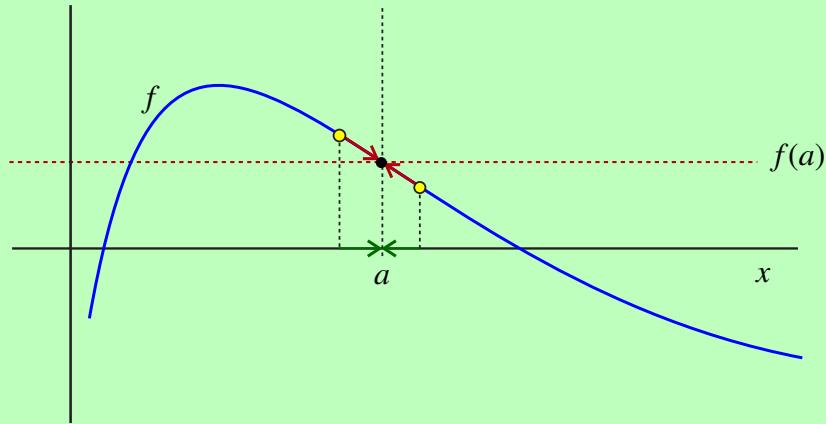
One of the fundamental concepts in Calculus is *continuity*. Roughly speaking, continuity means that the *value of a function* at a fixed point $x = a$ is determined uniquely by the *behavior of the function near and at the point $x = a$* . Since the behavior near $x = a$ is central to the concept of a limit, this would suggest the following definition.

DEFINITION

Formal Definition of Continuity I

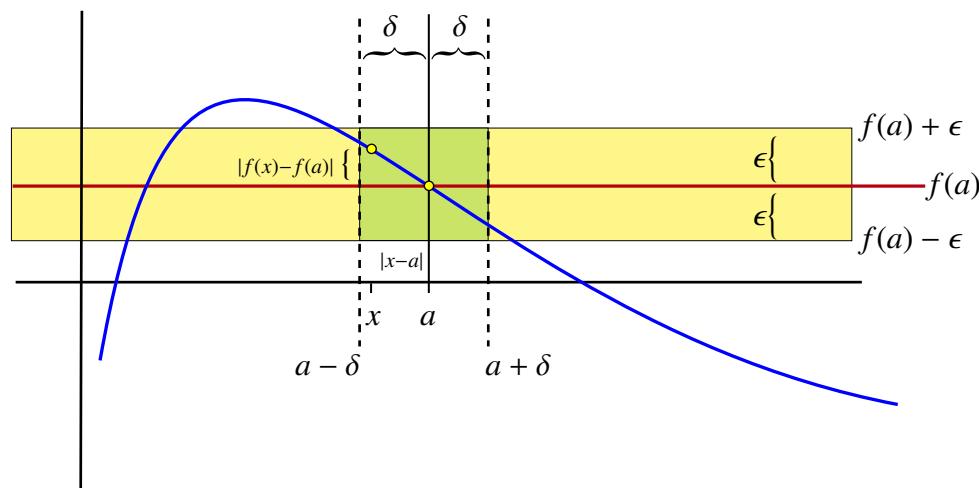
We say that a function f is *continuous at a point $x = a$* if

- i) $\lim_{x \rightarrow a} f(x)$ exists, and
- ii) $\lim_{x \rightarrow a} f(x) = f(a)$.



Otherwise, we say that f is *discontinuous at $x = a$* or that $x = a$ is a *point of discontinuity* for the function f .

The formal definition of a limit tells us that a function f is continuous at a point $x = a$ precisely when given any positive tolerance ϵ , we can find a cutoff distance $\delta > 0$ such that if x is within δ units of a , then $f(x)$ approximates $f(a)$ with an error less than ϵ . That is, if $|x - a| < \delta$, then $|f(x) - f(a)| < \epsilon$.



This leads us to our second formal definition for continuity:

DEFINITION

Formal Definition of Continuity II

We say that a function f is continuous at $x = a$ if for every positive tolerance $\epsilon > 0$, there is a cutoff distance $\delta > 0$ such that if $|x - a| < \delta$, then

$$|f(x) - f(a)| < \epsilon.$$

We note that as was the case for limits, there is a sequential characterization of continuity at a point.

THEOREM 11

Sequential Characterization of Continuity

A function f is continuous at $x = a$ if and only if whenever $\{x_n\}$ is a sequence with $\lim_{n \rightarrow \infty} x_n = a$, we must have that

$$\lim_{n \rightarrow \infty} f(x_n) = f(a).$$

You will notice that this is essentially the sequential characterization of limits with L replaced by $f(a)$ and without the restriction that $x_n \neq a$.

We end this section with a useful observation that is really nothing more than a notational trick.

Observation: Suppose that we want to consider $\lim_{x \rightarrow a} f(x)$. Assume $x \neq a$. Then we can write

$$x = a + h$$

where $h \neq 0$. In particular, $f(x) = f(a + h)$. We already know that if $h \rightarrow 0$, then $x = a + h \rightarrow a + 0 = a$. As a result we obtain the following

$$\lim_{x \rightarrow a} f(x) = \lim_{h \rightarrow 0} f(a + h)$$

in the sense that if either limit exists, then both exist and they are equal. If one fails to exist, so does the other. From this we can deduce an alternative way of stating that f is continuous at $x = a$.

Fact:

A function f is continuous at $x = a$ if and only if

$$\lim_{h \rightarrow 0} f(a + h) = f(a).$$

5.8.1 Types of Discontinuities

You may notice that the second requirement in the definition of continuity (that $\lim_{x \rightarrow a} f(x) = f(a)$) actually implies the first (that $\lim_{x \rightarrow a} f(x)$ exists). Why then did we write the definition in this way rather than simply requiring that $\lim_{x \rightarrow a} f(x) = f(a)$?

The answer to this question comes from the observation that to really understand what it means for a function to be continuous at a point you need to first see what makes a function discontinuous. This can occur in two ways. Either (i) holds and (ii) fails, or (i) fails and as a consequence so must (ii).

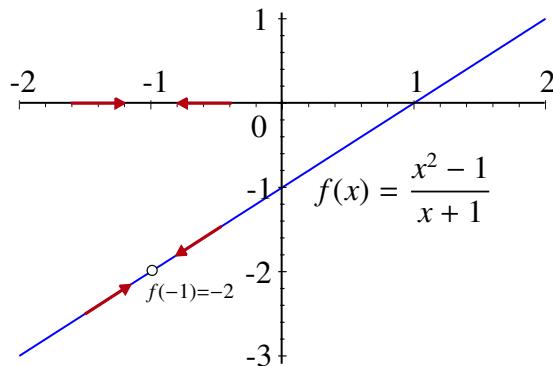
The first type of discontinuity we want to discuss happens when (i) holds, but (ii) fails. In this case, since the limit as x approaches a exists, we might conclude that the function is well-behaved near $x = a$, but it is either not defined at $x = a$ or it was defined in some sense “incorrectly.” An example of this type of discontinuity happens when we consider the function $f(x) = \frac{x^2 - 1}{x + 1}$ at $x = -1$. We know

$$\frac{x^2 - 1}{x + 1} = x - 1$$

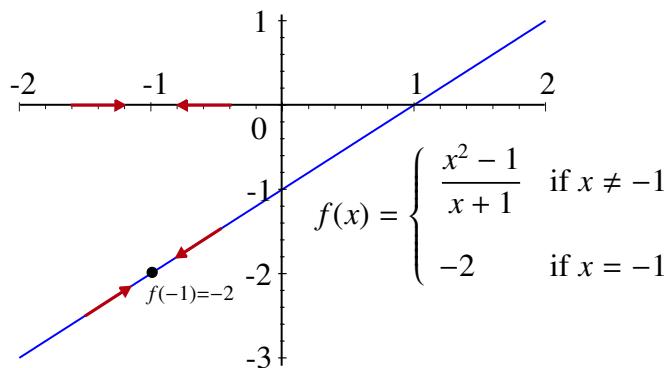
for all $x \neq -1$. It follows that

$$\lim_{x \rightarrow -1} \frac{x^2 - 1}{x + 1} = -1 - 1 = -2.$$

However, since $f(x)$ is not defined at $x = -1$, the graph of f has a *hole* at the point $(-1, -2)$.



This type of discontinuity is called a *removable discontinuity* because it can be removed by simply defining or redefining $f(a)$ to be the value of the limit at $x = a$. This would fill the hole in the graph. In the case of our current example, we could simply define $f(-1) = -2$. The new graph of f would look as follows:

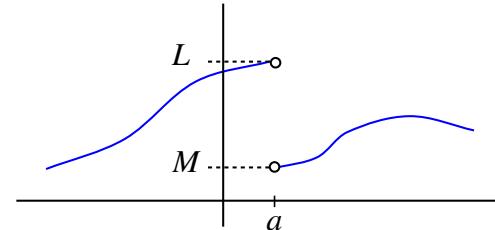


The second type of discontinuity happens when $\lim_{x \rightarrow a} f(x)$ fails to exist. These discontinuities are called *essential discontinuities* since they cannot be repaired by simply defining or redefining $f(a)$. We will now look at three ways that essential discontinuities can happen. The first type is called a *finite jump discontinuity*.

We know that $\lim_{x \rightarrow a} f(x)$ exists if and only if both $\lim_{x \rightarrow a^-} f(x)$ and $\lim_{x \rightarrow a^+} f(x)$ exist and the two one-sided limits are equal.

Suppose on the other hand that

$\lim_{x \rightarrow a^-} f(x) = L$ and $\lim_{x \rightarrow a^+} f(x) = M$, but that $L \neq M$. We then have that $\lim_{x \rightarrow a} f(x)$ does not exist, so f is discontinuous at $x = a$.

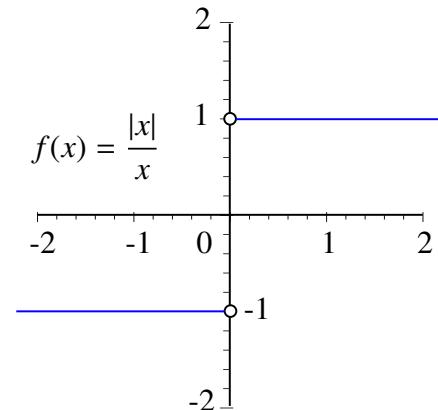


Notice the *gap* or *jump* of length $|L - M|$ on the graph at $x = a$. It is this *finite jump* that gives the discontinuity its name. It is also clear that the gap cannot be filled by defining $f(a)$ in some appropriate manner.

EXAMPLE 27

Let

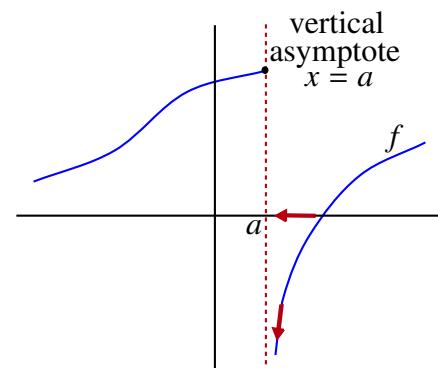
$$f(x) = \frac{|x|}{x} = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}.$$



The graph of f looks as follows:

Since $\lim_{x \rightarrow 0^+} f(x) = 1$ and $\lim_{x \rightarrow 0^-} f(x) = -1$, the function f has a jump discontinuity of length 2 at $x = 0$. ◀

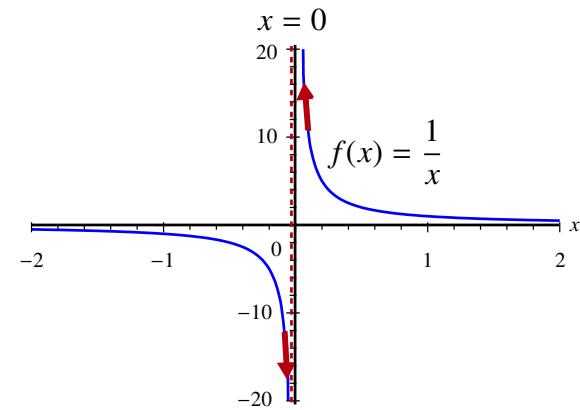
The second type of essential discontinuity happens when the graph has a vertical asymptote at $x = a$. This means that at least one of the one-sided limits is infinite. In the picture, we have $\lim_{x \rightarrow a^+} f(x) = -\infty$.



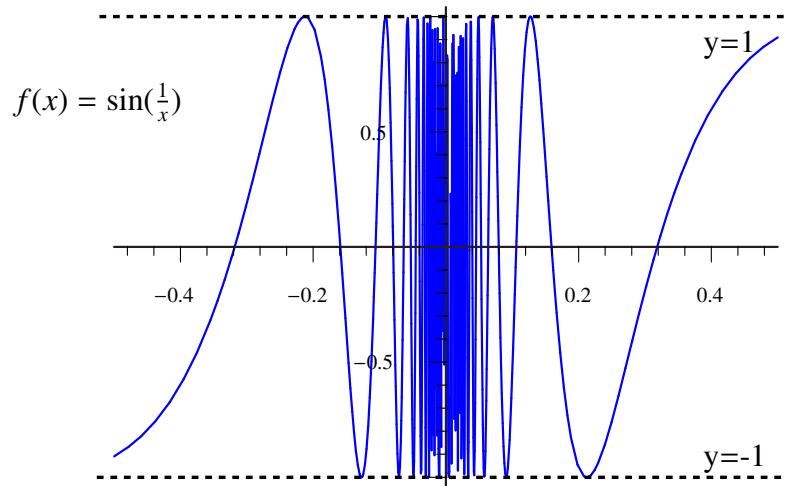
Just as in the previous case, the graph of f is broken by a gap at $x = a$, but this time the gap or jump is infinite in length. For this reason it is known as an *infinite jump*.

discontinuity. Like the finite jump discontinuity, it cannot be removed by defining or redefining $f(x)$ at $x = a$.

An example of this type of infinite jump discontinuity is $f(x) = \frac{1}{x}$ at $x = 0$.



The third type of essential discontinuity is the *oscillatory discontinuity*. This happens when f is bounded near $x = a$ but it does not have a limit because of *infinitely many oscillations* near a . The standard example of this phenomenon is the function $f(x) = \sin\left(\frac{1}{x}\right)$ at $x = 0$. We remind you that the graph of $\sin\left(\frac{1}{x}\right)$ looks as follows:



Unlike the previous two types of essential discontinuities, the graph of $\sin\left(\frac{1}{x}\right)$ does not exhibit any obvious break at $x = 0$. However, a break still exists in the sense that the y -axis divides the part of the graph to the left of 0 from the part to the right of 0 and there is no way to define the function at 0 so that these two parts become “connected.” That is, suppose you would like to trace the graph with a pencil. There is no way to define $f(0)$ so that you can get from a point on the graph located to the left of 0 to a point on the graph to the right of 0 without the pencil leaving the graph. In fact, all discontinuities result in “breaks” in the graph of the function at a point $x = a$.

REMARK

Earlier in this chapter we introduced the Thomae Function:

$$f(x) = \begin{cases} 1 & \text{if } x = 0 \\ \frac{1}{n} & \text{if } x = \frac{k}{n} \in \mathbb{Q} \text{ with } k \in \mathbb{Z} \setminus \{0\}, n \in \mathbb{N}, \gcd(k, n) = 1. \\ 0 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

We also showed that $\lim_{x \rightarrow a} f(x) = 0$ for every $a \in \mathbb{R}$. This means that f has the unusual property that f is continuous at each irrational and has a removable discontinuous at each rational.

Question: Does there exist a function $g : \mathbb{R} \rightarrow \mathbb{R}$ which is continuous at each rational but discontinuous at each irrational?

.

5.8.2 Continuity of Polynomials, $\sin(x)$, $\cos(x)$, e^x and $\ln(x)$

We have just seen what can happen when a function is discontinuous. Under what circumstances can we expect continuity?

EXAMPLE 28 Polynomial Functions

We already know that for a polynomial function p , we can find the limit at $x = a$ by simply evaluating $p(x)$ at $x = a$. This is another way of saying that if

$$p(x) = a_0 + a_1x + \cdots + a_nx^n,$$

then p is continuous at each point $a \in \mathbb{R}$.

EXAMPLE 29 Continuity of $\sin(x)$ and $\cos(x)$

We have seen that

$$\lim_{x \rightarrow 0} \sin(x) = 0 = \sin(0) \quad \text{and} \quad \lim_{x \rightarrow 0} \cos(x) = 1 = \cos(0).$$

This shows that both $\sin(x)$ and $\cos(x)$ are continuous at $x = 0$. We can also use these results to show continuity at any point. To see why this is the case observe that

$$\begin{aligned} \lim_{x \rightarrow a} \sin(x) &= \lim_{h \rightarrow 0} \sin(a + h) \\ &= \lim_{h \rightarrow 0} \sin(a) \cos(h) + \sin(h) \cos(a) \\ &= \sin(a) \cdot 1 + 0 \cdot \cos(a) \\ &= \sin(a) \end{aligned}$$

and

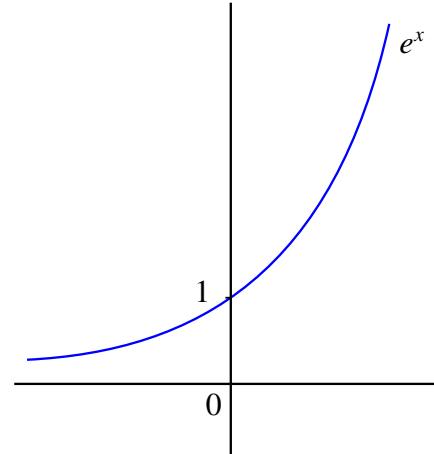
$$\begin{aligned} \lim_{x \rightarrow a} \cos(x) &= \lim_{h \rightarrow 0} \cos(a + h) \\ &= \lim_{h \rightarrow 0} \cos(a) \cos(h) - \sin(a) \sin(h) \\ &= \cos(a) \cdot 1 - \sin(a) \cdot 0 \\ &= \cos(a) \end{aligned}$$

EXAMPLE 30 Continuity of e^x and $\ln(x)$

Unlike the previous examples, it is actually not an easy task to prove the continuity of either the function $f(x) = e^x$ or the function $g(x) = \ln(x)$ at a particular point $x = a$. In fact, the easiest way to show that e^x is continuous at each real number is to realize that it can be defined by a special type of series construction known as a **power series**. The proof of this is beyond the scope of this course. However, we can show that if e^x is continuous at $x = 0$, then it is continuous everywhere.

To see why this is so, we first observe that if e^x is continuous at $x = 0$, then

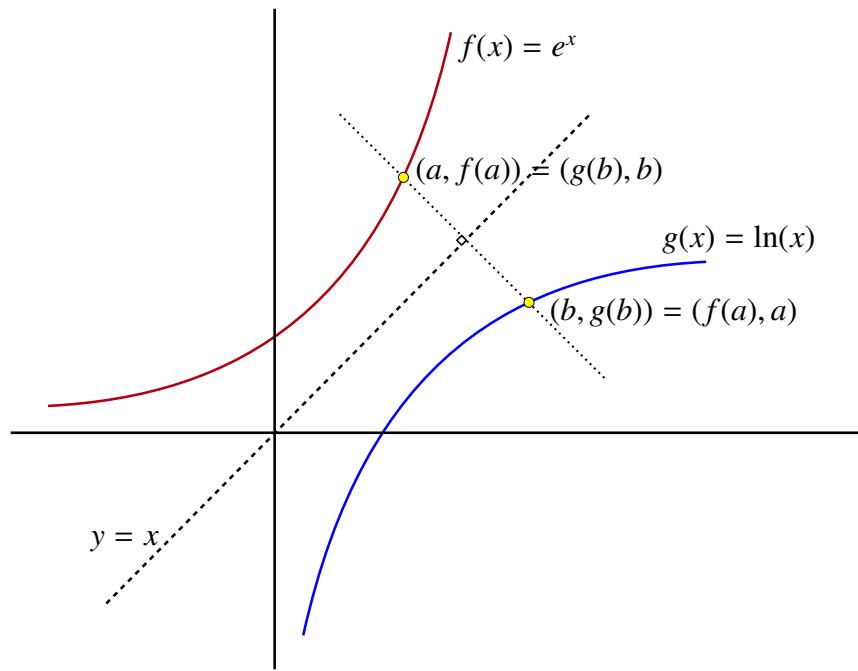
$$\lim_{h \rightarrow 0} e^h = e^0 = 1.$$



Therefore

$$\begin{aligned} \lim_{x \rightarrow a} e^x &= \lim_{h \rightarrow 0} e^{a+h} \\ &= \lim_{h \rightarrow 0} e^a e^h \\ &= e^a \lim_{h \rightarrow 0} e^h \\ &= e^a \end{aligned}$$

Once we have that e^x is continuous at every point $a \in \mathbb{R}$, we can give a geometric argument to show that $\ln(x)$ is also continuous at each point in its domain. Consider that $f(x) = e^x$ is invertible with inverse $g(x) = \ln(x)$. We know that the graph of g is simply the reflection of the graph of f through the line $y = x$.



Assume that $f(a) = b$. Continuity of f at $x = a$ means that there are no *breaks* in the graph at $(a, f(a))$. Since reflection does not create breaks that were not there before, the graph of g would have no breaks at the point $(b, g(b))$. This leads us to the following theorem:

THEOREM 12

Continuity of Inverses

Assume that $y = f(x)$ is invertible with inverse $x = g(y)$. If $f(a) = b$ and if f is continuous on an open interval containing $x = a$, then g is continuous at $y = b = f(a)$.

As a consequence of the previous theorem we immediately get that $\ln(x)$ is continuous at each point in its domain. ◀

REMARK

The previous theorem concerning continuity of the inverse function seems quite obvious given our geometric interpretation of the inverse function since it should be the case that there are no breaks in the graph of g after reflection if there were no breaks in the graph of f to begin with. However, it turns out that an analytic proof of this result is surprisingly complex and at this point we do not as yet have the tools to give the proof. As such we will delay the formal proof of this theorem until the next chapter. ◀

5.8.3 Arithmetic Rules for Continuous Functions

In this section, we see how to build further examples of continuous functions. The first two theorems follow immediately from the corresponding results for limits.

THEOREM 13

Continuity of Sums and Products

Let f and g be continuous at $x = a$, then

- 1) $f + g$ is continuous at $x = a$.
- 2) fg is continuous at $x = a$.

THEOREM 14

Continuity of Quotients

Let f and g be continuous at $x = a$. If $g(a) \neq 0$, then $\frac{f}{g}$ is continuous at $x = a$.

Let $f(x) = \frac{p(x)}{q(x)}$ be a rational function. Recall that p and q are polynomials. Let $a \in \mathbb{R}$. Then p and q are both continuous at $x = a$. It follows from the theorem on continuity for quotients that f is continuous at $x = a$ if and only if $q(a) \neq 0$.

EXAMPLE 31 Let $p(x) = x^2 - 1$ and $q(x) = x^2 + x - 2$. Then

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

is continuous precisely when $x^2 + x - 2 \neq 0$. But

$$x^2 + x - 2 = (x - 1)(x + 2)$$

so $x^2 + x - 2 = 0$ if $x = 1$ or $x = -2$. This means that f is continuous everywhere except at $x = 1$ and $x = -2$.

To test the nature of the discontinuities at $x = 1$ and $x = -2$, we must see if the limits exist at either of these points. In fact, since

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

if $x \neq 1$ we get that

$$\begin{aligned}\lim_{x \rightarrow 1} f(x) &= \lim_{x \rightarrow 1} \frac{x+1}{x+2} \\ &= \frac{2}{3}.\end{aligned}$$

This shows that the function has a *removable discontinuity* at $x = 1$.

At $x = -2$, the situation is different. Since $p(-2) = (-2)^2 - 1 = 3 \neq 0$, the limit of this rational function f does not exist at $x = -2$. In fact, the function is unbounded near $x = -2$. Moreover, since $f(x) < 0$ on the interval $(-2, -1)$ and $f(x) > 0$ when $x < -2$, we have

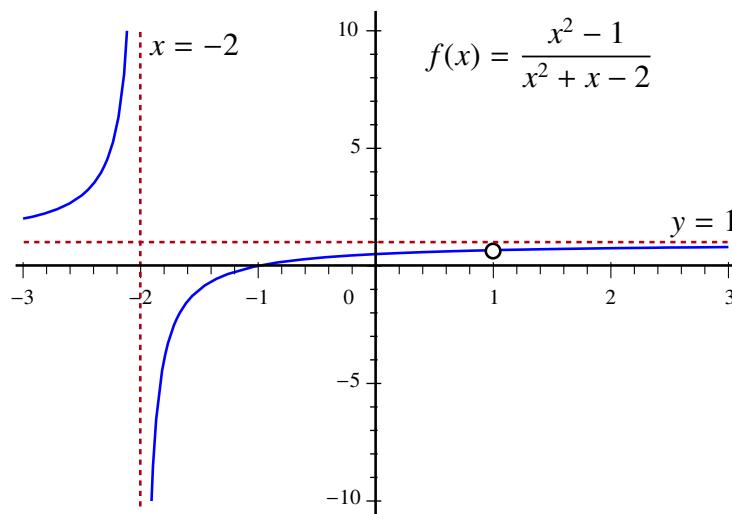
$$\lim_{x \rightarrow -2^-} f(x) = \infty$$

and

$$\lim_{x \rightarrow -2^+} f(x) = -\infty.$$

This means that there is an *essential discontinuity* at $x = -2$.

This analysis of the function f is confirmed by the plot of its graph.



Though it is not relevant for our discussion of continuity, the picture also shows a horizontal asymptote at $y = 1$. This is because

$$\lim_{x \rightarrow -\infty} f(x) = 1 = \lim_{x \rightarrow \infty} f(x).$$

The next theorem is a key tool in our quest to find continuous functions.

THEOREM 15**Continuity of Compositions**

Let f be continuous at $x = a$ and g be continuous at $x = f(a)$. Then $h = g \circ f$ is continuous at $x = a$.

PROOF

We can give a simple proof that composition preserves continuity using the sequential characterization of continuity.

Let f be continuous at $x = a$ and g be continuous at $x = f(a)$. Let $h(x) = (g \circ f)(x) = g(f(x))$. Let $\{x_n\}$ be a sequence such that $x_n \rightarrow a$. Since f is continuous at $x = a$ the sequential characterization of continuity shows that

$$\lim_{n \rightarrow \infty} f(x_n) = f(a).$$

But now since $f(x_n) \rightarrow f(a)$ and since g is continuous at $f(a)$ this time the sequential characterization of continuity shows that

$$\lim_{n \rightarrow \infty} g(f(x_n)) = g(f(a)).$$

This means that

$$\lim_{n \rightarrow \infty} h(x_n) = \lim_{n \rightarrow \infty} g(f(x_n)) = g(f(a)) = h(a)$$

which is exactly what we require to show that the composition is continuous. ■

EXAMPLE 32 Show that $h(x) = e^{x^2 \sin(x)}$ is continuous at each $a \in \mathbb{R}$.

SOLUTION To establish the continuity of this complicated function directly from the definition would be extremely difficult. However, the arithmetic rules will make the task much simpler. We first observe that

- 1) x^2 is continuous at each $a \in \mathbb{R}$.
- 2) $\sin(x)$ is continuous at each $a \in \mathbb{R}$.

From this the product rule implies that

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

is continuous at each $a \in \mathbb{R}$. Next, we know that

$$g(x) = e^x$$

is continuous at each $a \in \mathbb{R}$. Finally, from the rule for compositions, we can conclude that

$$h(x) = e^{x^2 \sin(x)} = g(f(x))$$

is also continuous at each $a \in \mathbb{R}$. ■

5.8.4 Continuity on an Interval

So far whenever we have looked at limits or continuity for a function we have focused our attention on a single point. In both cases, we were looking at the behaviour of the function f at or very near to $x = a$. For this reason we call limits and continuity at a point *local* properties of a function. However, we will often want to study the behavior of a function over an interval or over the entire real line \mathbb{R} . In this case, we are looking at the *global* nature of f . In particular, it will be useful to define what we mean by continuity over an entire interval I rather than just at a single point. To do so we will need to treat open intervals and closed intervals somewhat differently. For open intervals of the form (a, b) or for \mathbb{R} there is a very simple way to accomplish our goal:

DEFINITION

Continuity on (a, b) or \mathbb{R}

We say that a function f is *continuous on the open interval* (a, b) if it is continuous at each $x \in (a, b)$.

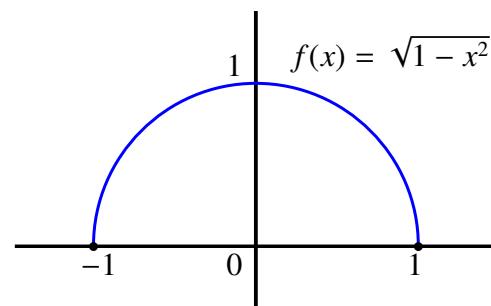
We say that a function f is *continuous on \mathbb{R}* , or just *continuous* for short, if it is continuous at each $x \in \mathbb{R}$.

This definition means that the graph of f has no breaks in the open interval (a, b) or anywhere on the Real line, respectively. However, if we consider closed intervals, the situation becomes more complicated. For example, consider the function $f(x) = \sqrt{1 - x^2}$ which is defined on the closed interval $[-1, 1]$. Not surprisingly, the function f can be shown to be continuous at each $x \in (-1, 1)$ and as such the definition above would mean that f is continuous on the open interval $(-1, 1)$. But what happens at the end points?

Technically, $\lim_{x \rightarrow -1} f(x)$ does not exist. Since the function f is not defined for $x < -1$, we get that $\lim_{x \rightarrow -1^-} f(x)$ cannot possibly exist. But for $\lim_{x \rightarrow -1} f(x)$ to exist, both one-sided limits must exist as well. Consequently, f is not continuous at $x = -1$ in the traditional sense. Similarly, $f(x)$ is not defined for $x > 1$ and hence $\lim_{x \rightarrow 1^+} f(x)$ does not exist. This would mean that this function is not continuous at $x = 1$ under the previous definition. To see why this is troubling, look at the graph of f .

The graph of f appears to have all of the characteristics that we expect to see in a continuous function. In particular, there are no breaks in the graph. Why is this so?

Firstly, because f is continuous on $(-1, 1)$ there are no breaks in the open interval $(-1, 1)$. Then, as we approach -1 from within the interval, that is from *above*, the



value of the function approaches 0 which is $f(-1)$. Finally, as we approach 1 from within the interval, or from *below*, we again see that the function approaches 0 which is $f(1)$. This means that instead of the two-sided limits at $x = -1$ and $x = 1$ agreeing with the values of the function at the endpoints, we have

$$\lim_{x \rightarrow -1^+} \sqrt{1 - x^2} = 0 = f(-1)$$

and

$$\lim_{x \rightarrow 1^-} \sqrt{1 - x^2} = 0 = f(1).$$

Since we are really only interested in what happens *inside* the closed interval $[-1, 1]$, this is essentially the best that we could expect. In summary, at the end points it is the appropriate one-sided limit that tells us whether or not the value of the function is properly reflected by the behavior of the function near that point, **but within the closed interval**. This leads us to define continuity on a closed interval more liberally as follows:

DEFINITION

Continuity on $[a, b]$

A function f is continuous on the *closed* interval $[a, b]$ if

- i) it is continuous at each $x \in (a, b)$,
- ii) $\lim_{x \rightarrow a^+} f(x) = f(a)$, and
- iii) $\lim_{x \rightarrow b^-} f(x) = f(b)$.

The function $f(x) = \sqrt{1 - x^2}$ satisfies all of these conditions on the interval $[-1, 1]$. As such, we would say that $\sqrt{1 - x^2}$ is continuous on $[-1, 1]$ even though technically it is not continuous at either $x = 1$ or $x = 2$ using our existing definition.

5.9 Intermediate Value Theorem

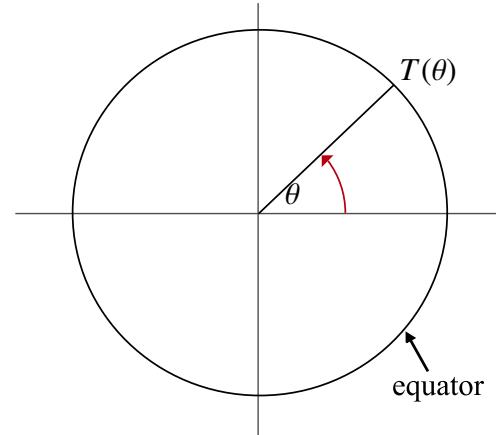
In the previous section, we introduced the notion of continuity and looked at some of the basic properties of continuous functions. In this section, we will look at a very important consequence of a function being continuous on an interval, namely the *Intermediate Value Theorem*. To motivate this theorem, we will begin with a rather curious fact concerning temperature at points on the equator.

Fact:

At any given time there will always be a pair of diametrically opposite points on the equator with *exactly the same temperature*!

If you give the statement some thought you will find that it is not easy to see why this must be true. In fact, it is rather surprising. It is also not immediately clear what this has to do with continuity.

To understand this situation more clearly, we will first assume that the earth's equator can be viewed as a circle and that each point on the equator can be identified by an angle θ in standard position. We will then let $T(\theta)$ denote the temperature at the given point on the equator.

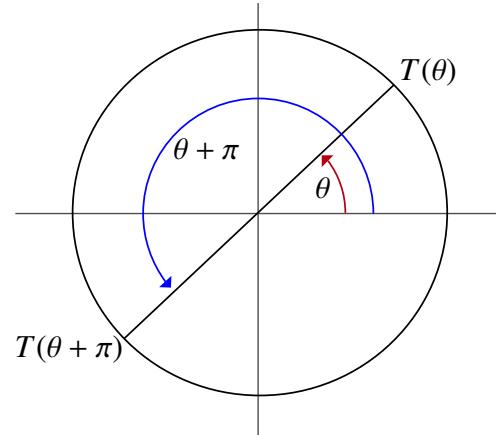


If a point has a standard position angle of θ , then the point diametrically opposite to it has a standard position angle of $\theta + \pi$. We want to show that we can always find an angle θ so that $T(\theta) = T(\theta + \pi)$, or equivalently, that

$$H(\theta) = 0$$

where

$$H(\theta) = T(\theta + \pi) - T(\theta).$$



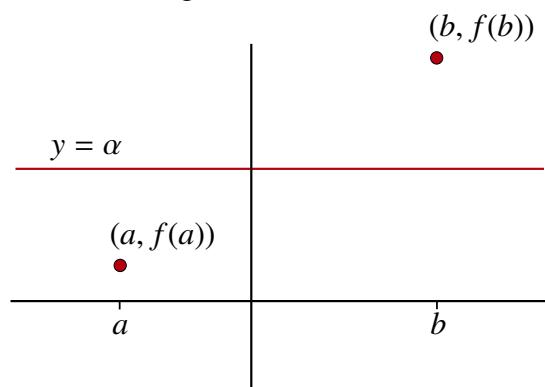
It is consistent with our understanding of the physical world to assume that temperature varies continuously with position. For us this means that T is a continuous function and hence so is H . Finally, we can achieve all diametrically opposite pairs if we let θ range from 0 to π . Hence, we have reduced the problem to one of showing that the function H , which is continuous on the closed interval $[0, \pi]$, must take on the value 0 at some point in this interval. Why must this happen?

To answer this latest question, we consider the following situation.

Suppose that we have a function f that is continuous on a closed interval $[a, b]$ and that α is such that

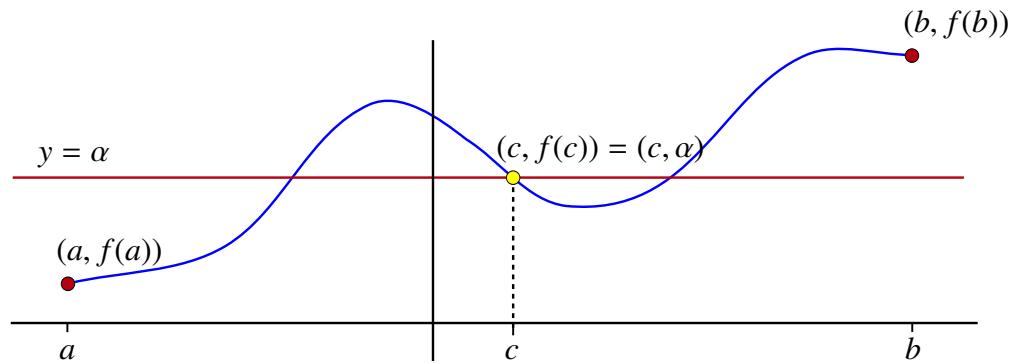
$$f(a) < \alpha < f(b).$$

This means that the graph of f starts out below the line $y = \alpha$ at $x = a$ and eventually rises above the line $y = \alpha$ as we move towards $x = b$.



However, we have seen that when f is continuous on $[a, b]$, its graph has no breaks in this region. It would then seem rather obvious that to get from below the line $y = \alpha$

to above the line $y = \alpha$ without creating such a break, we must cross the line $y = \alpha$ at least once. This means that there will be some point c in (a, b) with $f(c) = \alpha$.



This is the essence of the Intermediate Value Theorem.

THEOREM 16

The Intermediate Value Theorem (IVT)

Assume that f is continuous on the closed interval $[a, b]$, and either

$$f(a) < \alpha < f(b) \text{ or } f(a) > \alpha > f(b).$$

Then there exists a $c \in (a, b)$ such that $f(c) = \alpha$.

NOTE

- 1) There may be many points where $f(x) = \alpha$.
- 2) In our proof c will be the last one.



PROOF

We will first assume that $\alpha = 0$ and that

$$f(a) < 0 < f(b).$$

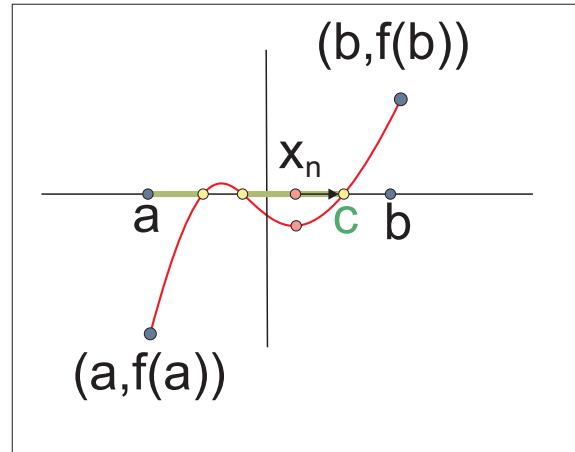
Let

$$S = \{x \in [a, b] \mid f(x) \leq 0\}.$$

Note that $S \neq \emptyset$ since $a \in S$ and S is bounded above by b . Let

$$c = \text{lub}(S).$$

We claim that $f(c) = 0$. However, we will first show that $f(c) \leq 0$. Since $c = \text{lub}(S)$, there exists a sequence $\{x_n\} \subseteq S$ with $x_n \rightarrow c$.



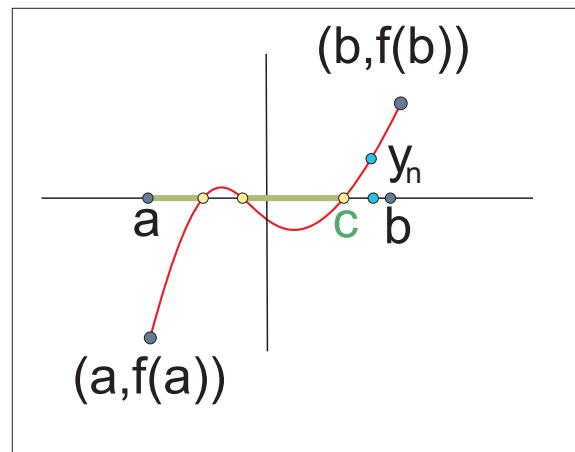
Since f is continuous on $[a, b]$ and since $f(x_n) \leq 0$ for each $n \in \mathbb{N}$, the Sequential Characterization of Continuity shows that

$$f(c) = \lim_{n \rightarrow \infty} f(x_n) \leq 0.$$

Next we let

$$y_n = c + \frac{b - c}{n}.$$

Then $c < y_n \leq b$ so $y_n \notin S$ and $f(y_n) > 0$.



Since $y_n \rightarrow c$, we have

$$f(c) = \lim_{n \rightarrow \infty} f(y_n) \geq 0.$$

This tells us that $f(c) = 0$ as claimed.

To obtain the more general result we consider the functions

$$g(x) = f(x) - \alpha \text{ or } h(x) = \alpha - f(x)$$

respectively depending on whether $f(a) < \alpha < f(b)$ or if $f(b) < \alpha < f(a)$. ■

REMARK

It is worth noting that despite the fact that the Intermediate Value Theorem seems to be rather obvious, it is actually quite difficult to prove rigorously. The proof, as we have seen relies on the Completeness Property of \mathbb{R} via the Least Upper Bound Property. In fact, the completeness of the real line is necessary. To see why suppose that our universe was the rationals \mathbb{Q} . Continuity still makes sense and the function $f(x) = x^2 - 2$ is continuous on $[0, 2]$. We also have that $f(0) = -2 < 0$ and $f(2) = 2 > 0$. However, there is no $c \in (0, 2) \cap \mathbb{Q}$ such that $f(c) = 0$. This shows that the IVT fails for the rationals. ◀

The Intermediate Value Theorem, denoted by IVT, is exactly the tool we require to complete the investigation of our temperature problem for the equator. Recall that we need only show that the function H is 0 at some point in the closed interval $[0, \pi]$. To see that this is the case, there are three possibilities that we must consider.

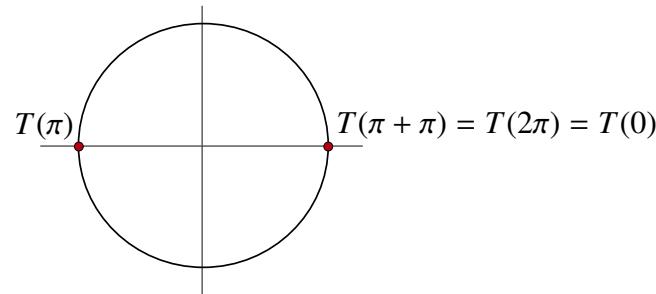
Firstly, it is possible that $H(0) = 0$. But then $T(\pi) = T(0)$ and we are done.

Secondly, we may have that $H(0) < 0$. This means $T(\pi) < T(0)$. In this case, if we could show that $H(\pi) > 0$, then the Intermediate Value Theorem would give us a point $\theta_0 \in (0, \pi)$ such that $H(\theta_0) = 0$ and hence that $T(\theta_0 + \pi) = T(\theta_0)$.

The key here is that

$$T(\pi + \pi) = T(2\pi) = T(0)$$

since the angles $\theta = 0$ and $\theta = 2\pi$ both represent the same point on the circle. what does that tell us about $H(\pi)$? Is $H(\pi) > 0$ as desired?



The third possible option is that $H(0) > 0$. In this case, we must show that $H(\pi) < 0$.

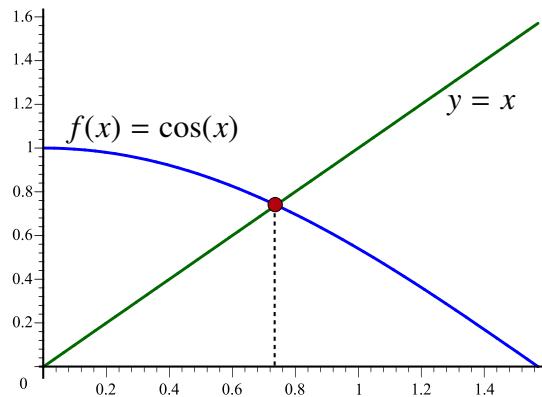
We claim that if $H(0) < 0$, then $H(\pi) > 0$ and if $H(0) > 0$, then $H(\pi) < 0$. Consequently, the Intermediate Value Theorem ensures us that *there will always be a pair of diametrically opposite points such that the temperatures at the two points are identical*. These claims are left as an exercise.

EXAMPLE 33

Show that there exists a $c \in (0, 1)$ such that

$$\cos(c) = c$$

SOLUTION We can see that this is true by plotting the graphs of the two functions on the same set of axes.

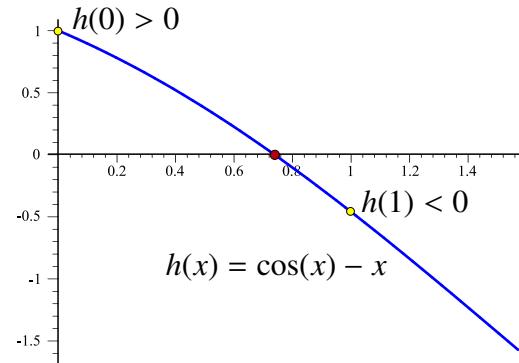


Can we do better than a *picture proof*? In fact we can, but to do so we must recognize that finding a $c \in (0, 1)$ with $\cos(c) = c$ is equivalent to finding a $c \in (0, 1)$ for $h(c) = 0$ when

$$h(x) = \cos(x) - x.$$

We also know that h is continuous on the closed interval $[0, 1]$. Moreover as the graph suggests

$$h(0) = \cos(0) - 0 = 1 > 0$$



and

$$h(1) = \cos(1) - 1 < 0.$$

Since we have satisfied the conditions of the IVT, we can conclude that there exists $0 < c < 1$ such that $h(c) = 0$ or equivalently that

$$\cos(c) = c.$$

NOTE

The above argument shows us that there is a point $0 < c < 1$ such that $\cos(c) = c$, however it does not tell us the value of c nor whether c is unique. That said, we will soon see that the IVT does present us with a relatively simple algorithm to find the approximate value of c with an error in the approximation that is as small as we might choose. ◀

5.9.1 Approximate Solutions of Equations

The Intermediate Value Theorem has a number of important applications. Perhaps the most important application for us will be the algorithm we can derive from the IVT to find accurate approximations to solutions of equations that cannot easily be solved exactly.

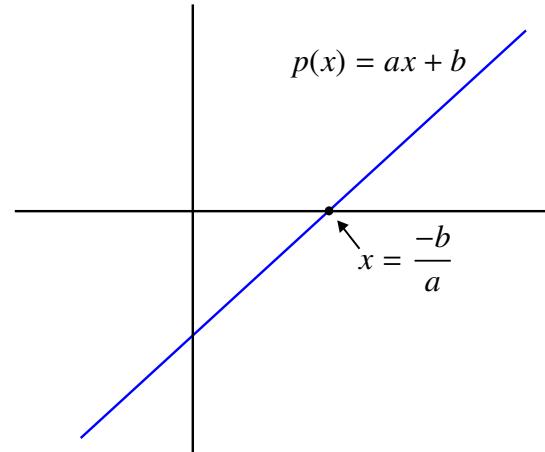
Approximating Roots of a Polynomial

Recall that if $p(x) = a_0 + a_1x + \cdots + a_nx^n$ is a polynomial, then a root of p is any number c such that $p(c) = 0$. For first-degree polynomials of the form $p(x) = ax + b$, where $a \neq 0$, it is easy to find the root. We require

$$ax + b = 0.$$

If $a \neq 0$, the equation is satisfied when $x = \frac{-b}{a}$.

This is exactly the point where the line $y = ax + b$ crosses the x -axis.



If $a = 0$ and $b \neq 0$, then there is no real root since the horizontal line $y = b$ does not cross the x -axis.

For a quadratic polynomial of the form $p(x) = ax^2 + bx + c$, the quadratic formula tells us that $p(x) = 0$ if and only if

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

There are formulae like the quadratic formula that calculate the roots of third and fourth degree polynomials. However, it is possible to use sophisticated ideas from algebra to prove that there is no general formula for finding the roots of all fifth degree polynomials or indeed for any degree above four. For example, we may want to know if $p(x) = x^5 + x - 1$ has any real roots and if so how can we find one?

It turns out that the Intermediate Value Theorem can provide a definitive answer to the first part of this question and can give us a very useful tool to find an approximate answer to the second part.

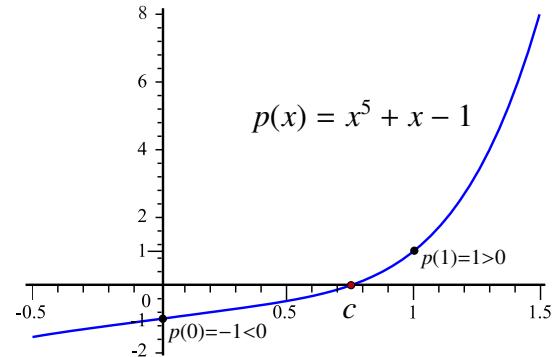
EXAMPLE 34 Does $p(x) = x^5 + x - 1$ have any real roots?

SOLUTION To see that p does have at least one real root, we note that $p(0) = 0^5 + 0 - 1 = -1 < 0$ and $p(1) = 1^5 + 1 - 1 = 1 > 0$. Since the polynomial $p(x)$ is continuous on the closed interval $[0, 1]$, the IVT implies that there will be a point c with $0 < c < 1$ such that $p(c) = 0$.

We also know from high school calculus that:

- (i) the derivative of the polynomial p is $p'(x) = 5x^4 + 1$ and that in this case, $p'(x) > 0$ for all x , and
- (ii) that a function which has a strictly positive derivative at each x is *increasing*.

This implies that the point c is the *only* real root.



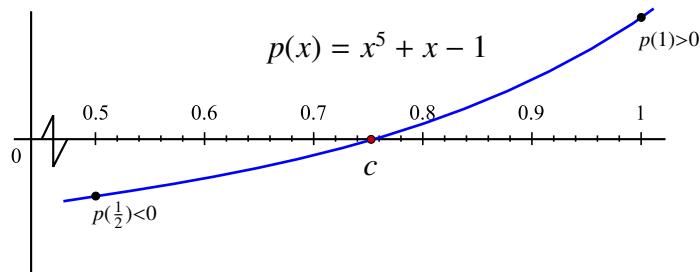
The IVT is an *existence theorem*. This means that it tells you that a point c with $f(c) = \alpha$ exists, but it does not tell you exactly how to find it. Nonetheless, we can often use the IVT to find a very good *approximation* for such a point c . In fact, we can use the IVT to approximate the root of the polynomial $p(x) = x^5 + x - 1$.

EXAMPLE 35 Let $p(x) = x^5 + x - 1$. Find the approximate value of its root.

SOLUTION To begin with, we know that the root c lies somewhere between 0 and 1. If we want to narrow the search, we could test the midpoint of this interval. In this case, our midpoint is $\frac{1}{2}$ and

$$\begin{aligned} p\left(\frac{1}{2}\right) &= \left(\frac{1}{2}\right)^5 + \frac{1}{2} - 1 \\ &= \frac{1}{32} + \frac{16}{32} - \frac{32}{32} \\ &= -\frac{15}{32} \\ &< 0. \end{aligned}$$

We now have that $p(0) < 0$, $p(\frac{1}{2}) < 0$ and $p(1) > 0$. Since we have a sign change between $x = \frac{1}{2}$ and $x = 1$, the IVT tells us that c is between $\frac{1}{2}$ and 1. This means that we are now looking for c in an interval that is half the length of our original interval.



To refine the search even further we repeat this process with the new interval $[0.5, 1]$. The new midpoint is

$$d = \frac{1 + .5}{2} = 0.75$$

We now test $p(0.75)$. If $p(0.75) > 0$, then since $p(0.5) < 0$, the root would be in the interval $[0.5, 0.75]$. If $p(0.75) < 0$, the root would be in the interval $[0.75, 1]$ since we know that $p(1) > 0$. In fact,

$$p(0.75) = -0.0126953 < 0,$$

so the root is in the interval $[0.75, 1]$.

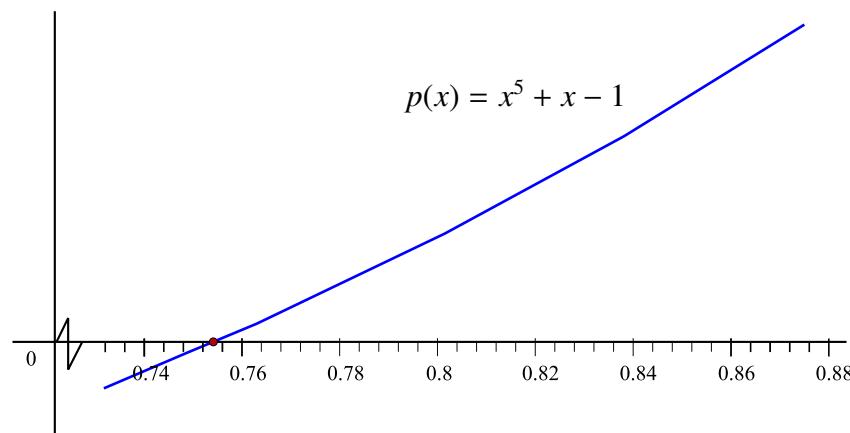
Notice that we are now searching for c in an interval that is $\frac{1}{4} = \frac{1}{2^2}$ times the length of the original interval.

We can continue by finding the new midpoint

$$d = \frac{1 + 0.75}{2} = 0.875$$

Test this new midpoint to find that

$$p(0.875) = 0.3879089 > 0$$



Since the sign change now occurs between 0.75 and 0.875, the next interval of interest is $[.75, .875]$.

Again, we find the new midpoint

$$d = \frac{.75 + .875}{2} = .8125$$

As the previous diagram suggested, we see that

$$p(.8125) = .1665926 > 0$$

so we know that the root lies in the interval $[.75, .8125]$

We continue by finding the new midpoint

$$d = \frac{.75 + .8125}{2} = .78125.$$

and then determine that

$$p(.78125) = .0722883 > 0$$

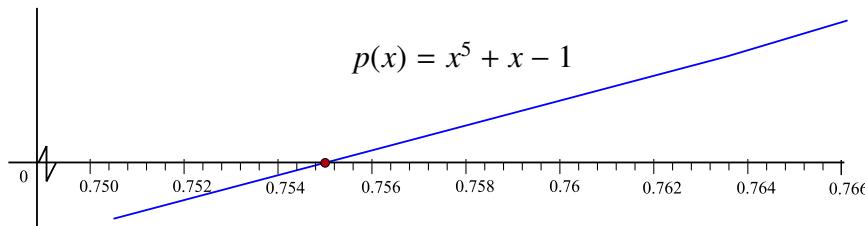
Since $p(.78125)$ is the same sign as $p(.8125)$, we replace 0.8125 with 0.78125 giving us the new interval $[.75, .78125]$. The next midpoint becomes

$$d = \frac{.75 + .78125}{2} = .765625$$

We have

$$p(.765625) = .0287006 > 0$$

so the root is in the interval $[.75, .765625]$.



The next midpoint is

$$d = \frac{.75 + .765625}{2} = .7578125$$

Evaluating $p(x)$ at this point gives

$$p(.7578125) = .007737 > 0$$

This means that the sign change happens between $x = .75$ and $x = .7578125$

One more iteration of the procedure gives us a new midpoint

$$d = \frac{.75 + .7578125}{2} = .75390625$$

with

$$p(0.75390625) = -0.002544 < 0.$$

As such, we replace 0.75 as the new left-hand endpoint with 0.75390625. We have now shown that

$$0.75390625 < c < 0.7578125$$

Notice that the length of the interval containing the root c is now

$$\begin{aligned} 0.7578125 - 0.75390625 &= \frac{1}{256} \\ &= \frac{1}{2^8} \end{aligned}$$

This is what we would expect since our original interval had length 1 and we have run through 8 iterations of the procedure with each iteration producing a new interval exactly $\frac{1}{2}$ the length of the previous interval. If we want a final estimate of the root, we can take the midpoint of the last two endpoints to get that

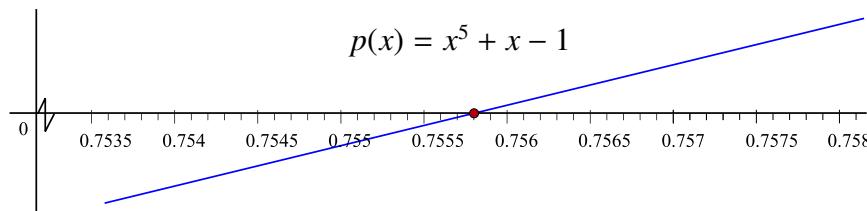
$$c \cong \frac{0.75390625 + 0.7578125}{2} = 0.755859375$$

The error in the estimate is at most the maximum distance from the final estimate to each of the two endpoints in the final interval. But since the estimate is the midpoint of this interval, this maximum difference is half the length of the interval. That is,

$$|0.755859375 - c| \leq \frac{1}{2^9} = \frac{1}{512}$$

Since the estimate for c is quite good, we would expect that the value of the function at the estimate should be close to 0. In fact,

$$p(0.755859375) = 0.002579752$$



5.9.2 The Bisection Method

The method that we have just outlined is called the *Bisection Method* because at each stage we bisect the previous interval. If we wanted to increase the accuracy of the

estimate, we would simply perform additional iterates of the procedure. Each additional iterate cuts the potential error in half. Since $\frac{1}{2^4} = \frac{1}{16} < \frac{1}{10}$, each block of four additional iterations gives us at least one additional decimal place of accuracy. Since $\frac{1}{2^{10}} = \frac{1}{1024} < \frac{1}{1000}$, each block of ten additional iterations gives us three additional decimal places of accuracy. This is useful because while the procedure is tedious to carry out manually, it is very easy to program on a computer. The only cautionary point we should make is that at some point the round-off error that arises when a computer performs inexact arithmetic will become a limiting factor to the accuracy achieved with this method. However, the IVT still provides us with a rather easy method of obtaining a very accurate estimate of the root of a polynomial p .

The procedure we used for estimating the root of a polynomial works in much more generality. For example, suppose that we wanted to show that the equation

$$e^x = -3x^2 + 4$$

has a solution in the interval $[0, 1]$. We could still apply the IVT. To do so we first introduce the function

$$F(x) = e^x + 3x^2 - 4$$

obtained by subtracting the function on the right-hand side of our equation from the function on the left-hand side. We then note that a point c is a solution of the equation if and only if $F(c) = 0$. This new function F is a continuous function and

$$F(0) = e^0 + 3(0)^2 - 4 = -3 < 0$$

while

$$F(1) = e + 3 - 4 > 0.$$

Therefore, the IVT guarantees that there is at least one point $c \in [0, 1]$ such that $F(c) = 0$. That is,

$$e^c = -3c^2 + 4.$$

To gain a better understanding about where such a c might be located, we could bisect the original interval at the midpoint $\frac{1}{2}$ and evaluate $F(\frac{1}{2})$. If $F(\frac{1}{2}) < 0$, then since $F(1) > 0$ we would have a solution in the new interval $[\frac{1}{2}, 1]$. Otherwise, if $F(\frac{1}{2}) > 0$, the new focus would be on the interval $[0, \frac{1}{2}]$. We can then bisect the new interval, test the midpoint and proceed as we have outlined previously.

In general, we now have an algorithm for using the *Bisection Method* to find approximate solutions of equations that works as follows:

Suppose that we want to find an approximate solution to the equation

$$f(x) - g(x) = 0,$$

where both f and g are continuous functions of x , with an error of at most a fixed tolerance ϵ .

Step 1: Form the new continuous function

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

Step 2: Find two points $a_0 < b_0$ such that either $F(a_0) > 0$ and $F(b_0) < 0$, or $F(a_0) < 0$ and $F(b_0) > 0$.

The IVT now guarantees us that there is a point c between a_0 and b_0 such that $F(c) = 0$.

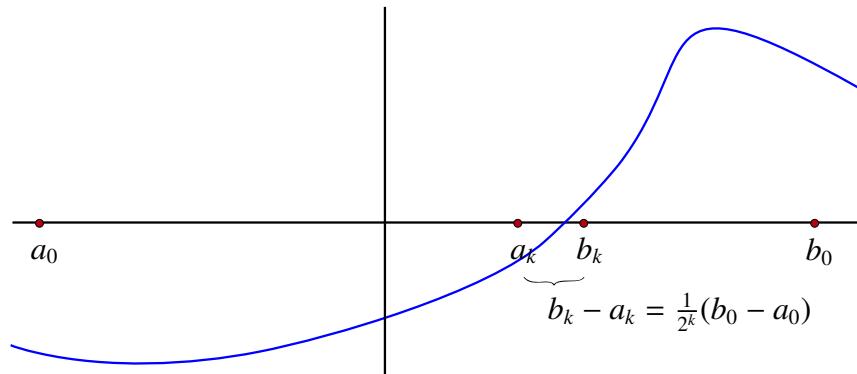
Step 3: Find the midpoint of the interval $[a_0, b_0]$ by using

$$d = \frac{a_0 + b_0}{2}$$

and evaluate $F(d)$.

Step 4: If $F(a_0)$ and $F(d)$ have the same sign, then let $a_1 = d$ and $b_1 = b_0$. Otherwise, let $a_1 = a_0$ and $b_1 = d$ to obtain a new interval $[a_1, b_1]$ which will contain a solution to the equation. We also have that $b_1 - a_1 = \frac{1}{2}(b_0 - a_0)$.

Step 5: Repeat steps 3 and 4 to obtain a new intervals $[a_2, b_2]$, $[a_3, b_3]$, \dots , $[a_n, b_n]$, each of which contains a solution to the equation. Moreover, for each $k = 1, 2, \dots, n$, $b_k - a_k = \frac{1}{2^k}(b_0 - a_0)$.



Step 6: Stop if

$$\frac{1}{2^{n+1}}(b_0 - a_0) < \epsilon.$$

Let

$$d = \frac{a_n + b_n}{2}.$$

Then there is a c such that $F(c) = 0$ and $|d - c| < \epsilon$.

NOTE

Suppose that you are given a function f . Consider two distinct points a and b . To test to see if $f(a)$ and $f(b)$ have the same sign we can simply calculate the product $f(a)f(b)$. If $f(a)f(b) > 0$, the two values have the same sign. If $f(a)f(b) < 0$, the two values have the opposite sign. ◀

With this in mind we present a summary of the algorithm for the Bisection Method.

Summary [Bisection Method]

Problem: Given a continuous function f and a positive tolerance $\epsilon > 0$, find a point d so that there exists a point c with $f(c) = 0$ and $|c - d| < \epsilon$.

Algorithm:

Step 1: Find two points $a < b$ with $f(a)f(b) < 0$.

Step 2: Set $\ell = b - a$.

Step 3: Set counter n to equal 0.

Step 4: Let $d = \frac{a+b}{2}$.

Step 5: If $\frac{\ell}{2^{n+1}} < \epsilon$, then STOP.

Step 6: If $f(d) = 0$, then STOP.

Step 7: If $f(a)f(d) < 0$, let $b = d$ and $n = n + 1$, then go to Step 4.

Step 8: Let $a = d$ and $n = n + 1$, then go to Step 4.

We have just seen that the Intermediate Value Theorem gives us a simple but effective method for finding approximate solutions to many equations. In the next chapter, we will introduce *Newton's Method*, which is also very easy to describe and to program, but is **much more efficient** as a means of finding approximate solutions to equations.

5.10 Extreme Value Theorem

In this section we present an important result that illustrates why continuity on a *closed* interval differs significantly from continuity on an *open* interval. To motivate our discussion we consider the following definition:

DEFINITION**Global Maxima and Global Minima**

Suppose that $f : I \rightarrow \mathbb{R}$, where I is an interval.

- We say that c is a *global maximum for f on I* if $c \in I$ and $f(x) \leq f(c)$ for all $x \in I$.
- We say that c is a *global minimum for f on I* if $c \in I$ and $f(x) \geq f(c)$ for all $x \in I$.
- We say that c is an *global extremum for f on I* if it is either a global maximum or a global minimum for f on I .

Note that a global maximum or minimum is sometimes called an **absolute** maximum or minimum.

In many practical applications of mathematics finding extrema is either the primary goal or it is a crucial step towards solving the problem being studied. However, before we try to find extrema it is helpful to know when they exist. With this in mind we ask the following question:

Question:

Given a function f defined on a non-empty interval I , do there exist points $c, d \in I$ such that $f(c) \leq f(x) \leq f(d)$ for all $x \in I$? That is, does f achieve both a global maximum and a global minimum on I ?

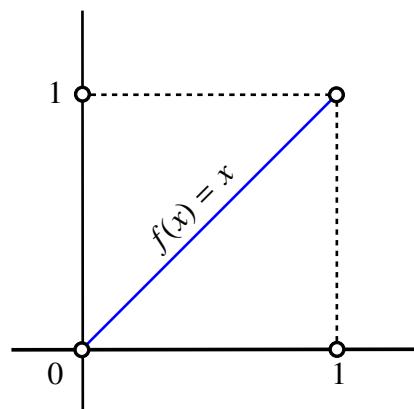
Unfortunately, the answer to the question above is generally—**No!**

The first example we will look at shows that if f is a continuous function on an open interval, then it is possible that neither a global maximum nor a global minimum exists.

EXAMPLE 36

Let $f(x) = x$ on the open interval $(0, 1)$.

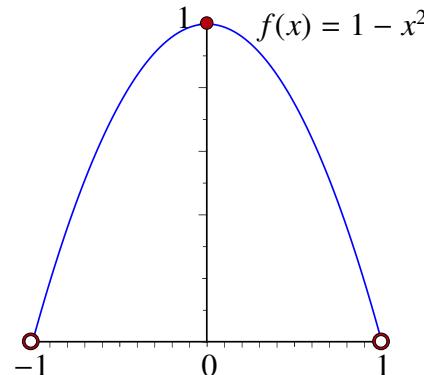
Since the open interval $(0, 1)$ has no largest or smallest value, f has no global maximum or global minimum on $(0, 1)$.



Key Observation: Notice that the function above seems to want to have a maximum and a minimum at the end points $x = 0$ and $x = 1$ of the open interval $(0, 1)$ but, unfortunately, those points are not available for us to use. ◀

The next example shows that it is possible that either a global maximum exists or a global minimum exists, but not both.

EXAMPLE 37 Let $f(x) = 1 - x^2$ on the open interval $(-1, 1)$.



Then f has no global minimum on $(-1, 1)$, but $f(x)$ does have a global maximum on the interval $(-1, 1)$ at $x = 0$.

As was the case in the previous example, the function does seem to want to achieve its minimum at the missing end points of the open interval $(-1, 1)$. This suggests that if we replace the open interval with a *closed* interval by adding the endpoints, we may have some hope to resolve our issues. In fact, this is the case as the next theorem shows.

THEOREM 17

The Extreme Value Theorem (EVT)

Suppose that f is continuous on $[a, b]$. There exist two numbers c and $d \in [a, b]$ such that

$$f(c) \leq f(x) \leq f(d)$$

for all $x \in [a, b]$.

PROOF

The proof will require three stages. In the first stage we will show that f is bounded on $[a, b]$.

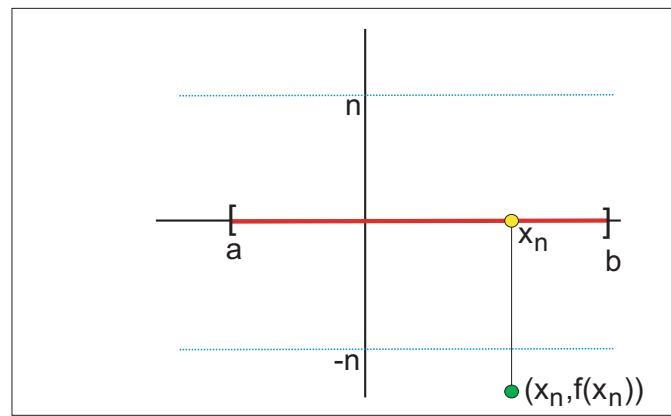
Stage 1: We claim that

$$f([a, b]) = \{f(x) \mid x \in [a, b]\}$$

is bounded.

Assume that this was not the case and that f is not bounded on $[a, b]$. Then for each $n \in \mathbb{N}$ there would exist an $x_n \in [a, b]$ such that

$$|f(x_n)| > n.$$



Since $\{x_n\} \subset [a, b]$ is bounded the Bolzano-Weierstrass Theorem implies that there exists a subsequence $\{x_{n_k}\}$ which converges to some point $t \in [a, b]$. The Sequential Characterization of Continuity tells us that $f(x_{n_k}) \rightarrow f(t)$. However, this is impossible since

$$|f(x_{n_k})| > n_k$$

so that $\{f(x_{n_k})\}$ is not bounded. It follows that f must be bounded on $[a, b]$.

Stage 2: Show that there exists $d \in [a, b]$ such that

$$f(x) \leq f(d)$$

for every $x \in [a, b]$.

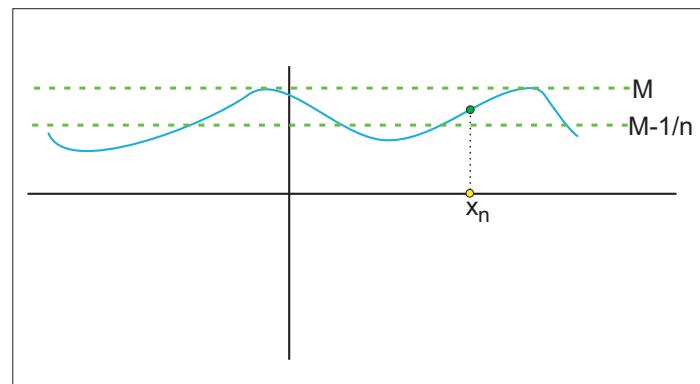
We begin by letting

$$M = \text{lub}(\{f(x) \mid x \in [a, b]\}).$$

It will suffice to show that there exists $d \in [a, b]$ such that $f(d) = M$.

For each $n \in \mathbb{N}$, the fact that $M - \frac{1}{n} < M$ means that there exists an $x_n \in [a, b]$ such that

$$M - \frac{1}{n} < f(x_n) \leq M.$$



By the BWT, $\{x_n\}$ has a subsequence $\{x_{n_k}\}$ with $x_{n_k} \rightarrow d \in [a, b]$. It then follows from the Sequential Characterization of Continuity and the Squeeze Theorem that

$$f(d) = \lim_{k \rightarrow \infty} f(x_{n_k}) = M.$$

Stage 3: Show that there exists $c \in [a, b]$ such that

$$f(c) \leq f(d)$$

for every $x \in [a, b]$.

We let

$$L = \text{glb}(\{f(x) \mid x \in [a, b]\}).$$

By modifying the argument in Stage 2, we can show that there exists a $c \in [a, b]$ such that $f(c) = L$.

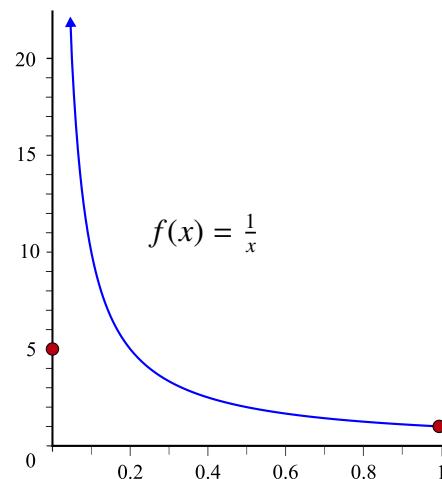
The next example shows that the assumption of continuity is essential in the statement of the EVT.

EXAMPLE 38

Let

$$f(x) = \begin{cases} \frac{1}{x} & \text{if } 0 < x \leq 1 \\ 5 & \text{if } x = 0 \end{cases}$$

Then f has a global minimum on $[0, 1]$ at $x = 1$, but it has no global maximum on $[0, 1]$.



This example does not contradict the EVT because f is **not** continuous on $[0, 1]$.

NOTE

At this point the EVT tells us that if f is a continuous function on a closed interval $[a, b]$, then the function always achieves its maximum and minimum value on the interval. Unfortunately, the theorem does not tell us how to find the global extrema.

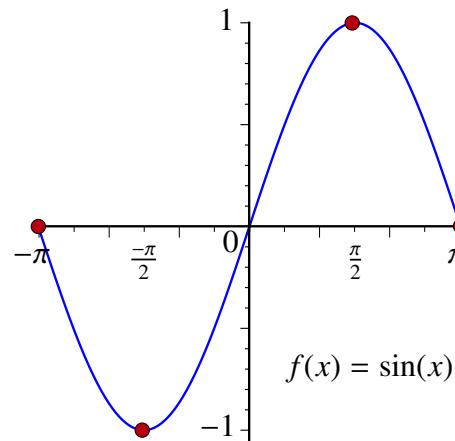


We have seen that the endpoints of an interval may play a role in locating maxima or minima. However, as the next example shows, it is possible that neither extrema occurs at an endpoint.

EXAMPLE 39

Consider the function $f(x) = \sin(x)$ on the interval $[-\pi, \pi]$.

The function $f(x) = \sin(x)$ attains its maximum and minimum values on $[-\pi, \pi]$ at $x = \frac{\pi}{2}$ and $x = -\frac{\pi}{2}$, respectively.



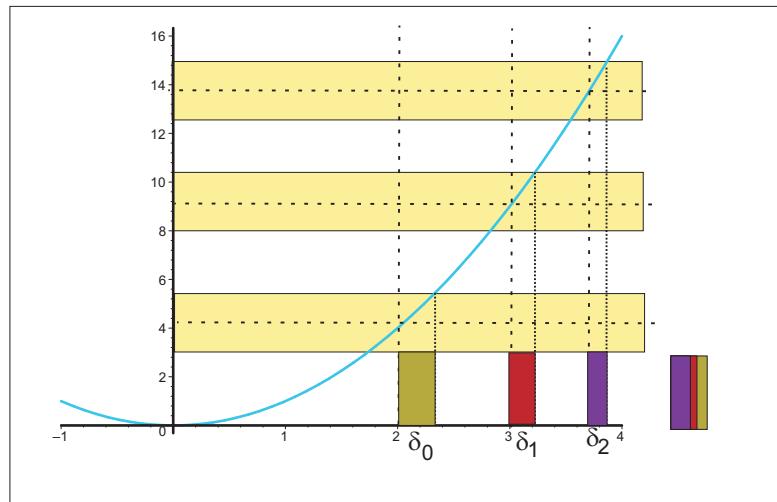
We will later show how to identify potential extrema if the function f is continuous on the closed interval $[a, b]$ and is differentiable on the open interval (a, b) .

5.11 Uniform Continuity

Assume that f is continuous at each point in an interval I given an $\epsilon > 0$ and a point $x = a \in I$ we know that we can find a $\delta > 0$ such that if $|x - a| < \delta$, then $|f(x) - f(a)| < \epsilon$. We know that the choice of δ depends on ϵ . However, it may also depend on the point a as well. To see this consider the following example.

EXAMPLE 40

Consider the function $f(x) = x^2$. This is continuous at each point in \mathbb{R} . The following diagram show the largest δ which will work in the definition of continuity given a particular fixed ϵ at three points $0 < x_0 < x_1 < x_2$.



If we denote the δ 's associated with x_0, x_1 and x_2 respectively by δ_0, δ_1 and δ_2 , then we see that

$$\delta_2 < \delta_1 < \delta_0$$

In fact, if $f(x) = x^2$, then for $x_0 > 0$ and $\epsilon > 0$ fixed, the δ we get at x_0 is

$$\delta_{x_0} = \sqrt{x_0^2 + \epsilon} - x_0 \rightarrow 0$$

as $x_0 \rightarrow \infty$.

This shows that we cannot find a $\delta > 0$ sufficiently small so that for the given fixed $\epsilon > 0$, this δ would work to satisfy the definition of continuity simultaneously at each point in \mathbb{R} .



It would of course be desirable if given a function f that is continuous on an interval I and any $\epsilon > 0$ that we could find a single $\delta > 0$, depending only on ϵ , which would work in the definition of continuity simultaneously at all points in I whenever this is possible. If we can do so, we will say that f is *uniformly continuous on I* .

DEFINITION

Uniform Continuity

We say that f is *uniformly continuous* on $S \subseteq \mathbb{R}$ if for every $\epsilon > 0$ there exists a $\delta > 0$ such that if $x, y \in S$ and

$$|x - y| < \delta,$$

then

$$|f(x) - f(y)| < \epsilon.$$

REMARK

It is an easy exercise to show that if f is uniformly continuous on S , then it is continuous on S . Moreover, if $T \subseteq S$ and if f is uniformly continuous on S , then it is also uniformly continuous on T with the same δ . ◀

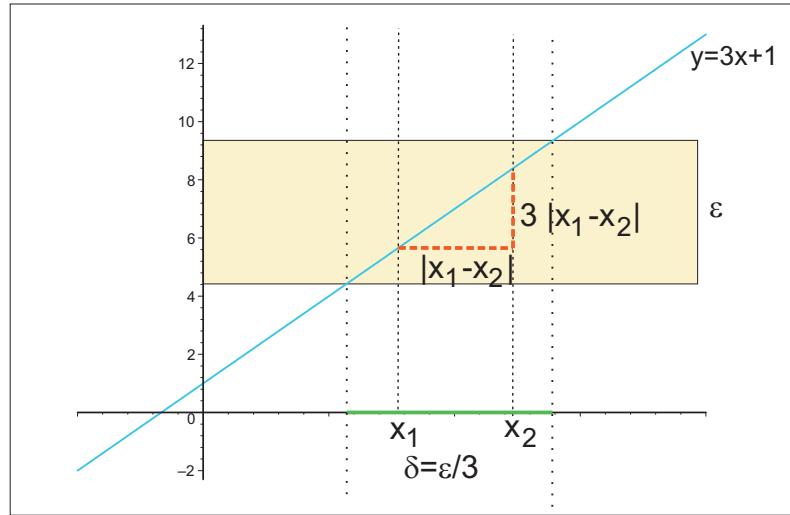
We have already seen that $f(x) = x^2$ is not uniformly continuous on \mathbb{R} . We will soon show that it is uniformly continuous on $I = [0, 1]$ however. But right now we give an example of a function that is uniformly continuous on \mathbb{R} .

EXAMPLE 41 Let $f(x) = 3x + 1$ and let $\epsilon > 0$. Given two points $x_1, x_2 \in \mathbb{R}$ we have that

$$\begin{aligned} |f(x_1) - f(x_2)| &= |(3x_1 + 1) - (3x_2 + 1)| \\ &= |3x_1 - 3x_2| \\ &= 3 \cdot |x_1 - x_2| \end{aligned}$$

It follows that if we let $\delta = \frac{\epsilon}{3}$ and if $|x_1 - x_2| < \delta$, then

$$\begin{aligned} |f(x_1) - f(x_2)| &= 3 \cdot |x_1 - x_2| \\ &< 3 \cdot \frac{\epsilon}{3} \\ &= \epsilon \end{aligned}$$



This shows that $f(x) = 3x + 1$ is uniformly continuous on \mathbb{R} . ◀

REMARK

The previous example can easily be modified to show that $f(x) = mx + b$ is uniformly continuous on \mathbb{R} .

If $m = 0$, f is a constant function and given $\epsilon > 0$, any $\delta > 0$ will satisfy the definition of uniform continuity for this ϵ .

If $m \neq 0$, then given $\epsilon > 0$ we simply let $\delta = \frac{\epsilon}{|m|}$. ◀

5.11.1 Sequential Characterization of Uniform Continuity

Just as was the case with limits and with continuity, it should not be surprising that there is a useful sequential characterization of uniform continuity.

THEOREM 18**Sequential Characterization of Uniform Continuity**

Assume that $f(x)$ is defined on $S \subseteq \mathbb{R}$. Then the following are equivalent:

- i) $f(x)$ is uniformly continuous on S .
- ii) If $\{x_n\}, \{y_n\} \subseteq S$ with $\lim_{n \rightarrow \infty} |x_n - y_n| = 0$, then
$$\lim_{n \rightarrow \infty} |f(x_n) - f(y_n)| = 0.$$

PROOF

i) implies ii):

Assume that f is uniformly continuous on S and that $\{x_n\}, \{y_n\} \subseteq S$ with

$$\lim_{n \rightarrow \infty} |x_n - y_n| = 0.$$

Let $\epsilon > 0$. Since f is uniformly continuous on S , we can find a $\delta > 0$ such that if $x, y \in S$ and $|x - y| < \delta$, then

$$|f(x) - f(y)| < \epsilon.$$

Moreover, because $\lim_{n \rightarrow \infty} |x_n - y_n| = 0$, we can find a cutoff $N \in \mathbb{N}$ so that if $n \geq N$, then

$$|x_n - y_n| < \delta.$$

It follows that if $n \geq N$, then

$$|f(x_n) - f(y_n)| < \epsilon$$

which shows that

$$\lim_{n \rightarrow \infty} |f(x_n) - f(y_n)| = 0.$$

ii) implies *i)*:

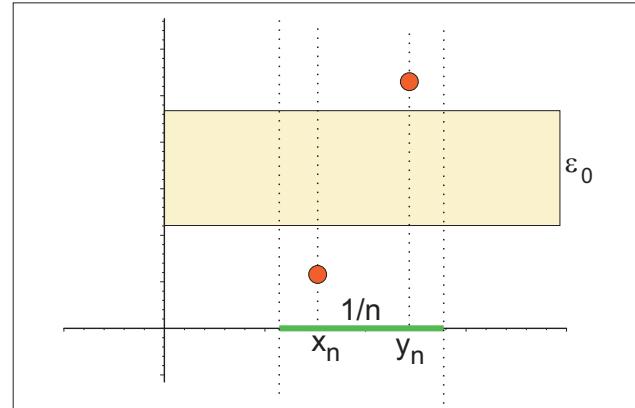
This time we will show that if *i)* fails, the *ii)* fails as well.

Assume that f is not uniformly continuous on S . Then there exists $\epsilon_0 > 0$ such that for each $\delta > 0$ there exists $x_\delta, y_\delta \in S$ with $|x_\delta - y_\delta| < \delta$ but

$$|f(x_\delta) - f(y_\delta)| \geq \epsilon_0.$$

In particular, if $n \in \mathbb{N}$ with $\delta = \frac{1}{n}$, we get a pair of points x_n and y_n with $|x_n - y_n| < \frac{1}{n}$ but

$$|f(x_n) - f(y_n)| \geq \epsilon_0.$$



It follows that $\lim_{n \rightarrow \infty} |x_n - y_n| = 0$, but

$$\lim_{n \rightarrow \infty} |f(x_n) - f(y_n)| \neq 0.$$

Since *ii)* fails whenever *i)* fails, this means that *ii)* implies *i)* completing the proof. ■

Just as was the case for the Sequential Characterization of Continuity, the Sequential Characterization of Uniform Continuity can be used to show that certain functions are **not** uniformly continuous on a given set f .

EXAMPLE 42

We have seen that $f(x) = x^2$ is not uniformly continuous on \mathbb{R} . The Sequential Characterization of Uniform Continuity can provide us with confirmation of this fact. Indeed let

$$x_n = n + \frac{1}{n} \quad \text{and} \quad y_n = n.$$

Then clearly $\lim_{n \rightarrow \infty} |x_n - y_n| = \lim_{n \rightarrow \infty} \frac{1}{n} = 0$. However,

$$\begin{aligned}\lim_{n \rightarrow \infty} |f(x_n) - f(y_n)| &= \lim_{n \rightarrow \infty} |(n + \frac{1}{n})^2 - n^2| \\ &= \lim_{n \rightarrow \infty} 2 + \frac{1}{n^2} \\ &= 2.\end{aligned}$$

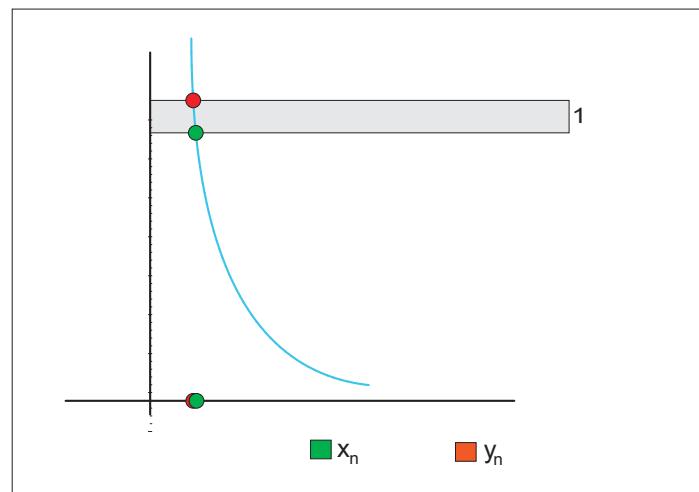
This shows that f is not uniformly continuous on \mathbb{R} as claimed. ◀

EXAMPLE 43 Show that $f(x) = \frac{1}{x}$ is not uniformly continuous on $(0, 1)$.

Solution: Let $x_n = \frac{1}{n}$ and $y_n = \frac{1}{n+1}$. Then $\lim_{n \rightarrow \infty} |x_n - y_n| = \lim_{n \rightarrow \infty} \frac{1}{n(n+1)} = 0$. However,

$$\begin{aligned}\lim_{n \rightarrow \infty} |f(x_n) - f(y_n)| &= \lim_{n \rightarrow \infty} |n - (n+1)| \\ &= \lim_{n \rightarrow \infty} 1 \\ &= 1.\end{aligned}$$

What this shows is that we can have two points arbitrarily close together yet when we apply f the results differ by the same constant 1. Visually, we can see this happening in the following diagram:



5.11.2 Uniform Continuity on $[a, b]$

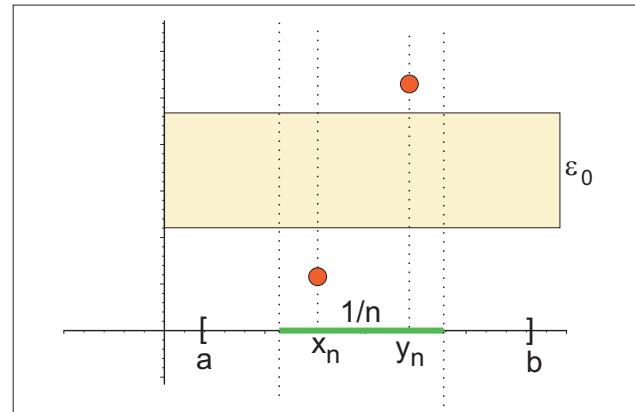
Given the examples we have seen so far it might seem that it is rare for a continuous function to be uniformly continuous on an interval. We will not show that thankfully this is not the case.

THEOREM 19

Uniform Continuity on $[a, b]$

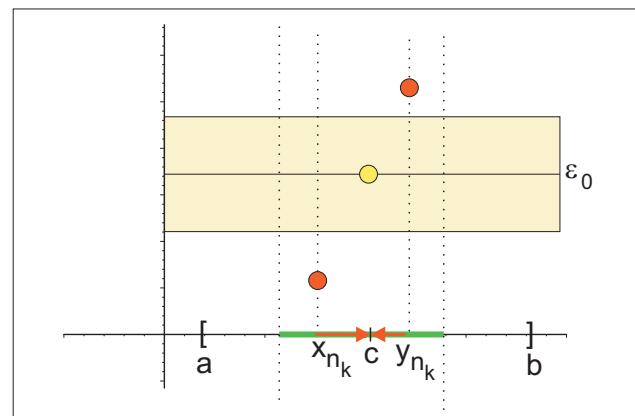
If f is continuous on $[a, b]$, then $f(x)$ is uniformly continuous on $[a, b]$.

PROOF



Assume that f is **not** uniformly continuous. Then there exists an $\epsilon_0 > 0$ such for each $n \in \mathbb{N}$ we can choose $x_n, y_n \in S$ with $|x_n - y_n| < \frac{1}{n}$, but

$$|f(x_n) - f(y_n)| \geq \epsilon_0.$$

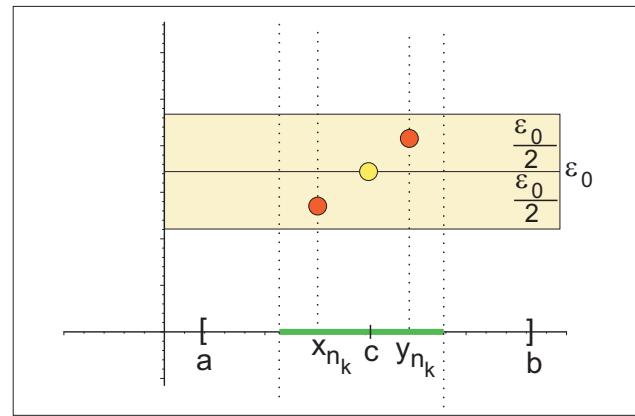


By the Bolzano-Weierstrass Theorem we can choose $\{x_{n_k}\}$ such that

$$x_{n_k} \rightarrow c \in [a, b].$$

Since $|x_{n_k} - y_{n_k}| < \frac{1}{n_k} \rightarrow 0$,

$$y_{n_k} \rightarrow c.$$



By continuity $f(x_{n_k}) \rightarrow f(c)$ and $f(y_{n_k}) \rightarrow f(c)$. It follows that if n_k is large enough, we have

$$|f(x_{n_k}) - f(c)| < \frac{\epsilon_0}{2},$$

and

$$|f(y_{n_k}) - f(c)| < \frac{\epsilon_0}{2},$$

and hence

$$\Rightarrow |f(x_{n_k}) - f(y_{n_k})| < \epsilon_0$$

which is a contradiction. Therefore f must be uniformly continuous on $[a, b]$. ■

REMARK

You will notice that the previous result also depends on the completeness of \mathbb{R} , this time via the Bolzano-Weierstrass Theorem.

The use of the Bolzano-Weierstrass Theorem also gives us a clue as to what went wrong in the case of $f(x) = x^2$ over \mathbb{R} and for $g(x) = \frac{1}{x}$ on $(0, 1)$.

In the first case, our sequences $\{x_n\}$ and $\{y_n\}$ need not be bounded so we may not have a convergent subsequence $\{x_{n_k}\}$ to work with.

In the second case, we would get a convergent subsequence $\{x_{n_k}\} \subset (0, 1)$, but this time we would have $x_{n_k} \rightarrow 0 \notin (0, 1)$. As such we could not appeal to the Sequential Characterization of Continuity to complete the proof. It follows that the theorem may well fail on either an open interval or an unbounded interval. But all is not lost as far as open intervals are concerned as the next example shows. ◀

EXAMPLE 44

Let $f(x) = x^2$. we know that f is not uniformly continuous on \mathbb{R} . However our most recent theorem shows us that f is uniformly continuous on $[0, 1]$. Moreover, we know that if $T \subset S$ and if f is uniformly continuous on S , then it is also uniformly continuous on T . It follows that $f(x) = x^2$ is also uniformly continuous on $(0, 1)$. In fact, it is uniformly continuous on any interval of finite length. ◀

5.12 Curve Sketching: Part 1

With modern computational tools available to help create precise plots of even rather complicated functions, it might seem that curve sketching is no longer a useful skill to learn. However, it is still a very valuable exercise since it forces you to really think about what the various concepts of Calculus tell you about the underlying function.

Usually in a first course in Calculus the derivative is the central tool in curve sketching. However, it is actually the case that many functions can be drawn with a fair degree of accuracy by determining only where the function is or is not continuous, and by taking limits at certain points of interest. As such, below are steps that should be followed when sketching the graph of f based on the ideas of this chapter.

Strategy [Basic Curve Sketching]

- Step 1:** Determine the domain of f .
- Step 2:** Determine any symmetries that the graph may have. In particular, test to see if the function is either even or odd.
- Step 3:** Determine, if possible, where the function changes sign and plot these points.
- Step 4:** Find any discontinuity points for f .
- Step 5:** Evaluate the relevant one-sided and two-sided limits at the points of discontinuity and identify the nature of the discontinuities. In particular, indicate any removable discontinuity with a small circle to denote the hole.
- Step 6:** From 5), draw any vertical asymptotes.
- Step 7:** Find any horizontal asymptotes by evaluating the limits of the function at $\pm\infty$, if applicable. Draw the horizontal asymptotes on your plot.
- Step 8:** Finally, use the information you have gathered above to construct as accurate a sketch as possible for the graph of the given function. It is often helpful to plot a few sample points as a guide.

EXAMPLE 45 Sketch the graph of

$$f(x) = \frac{xe^x}{x^3 - x}.$$

SOLUTION

Step 1: This function is defined everywhere except when the denominator is zero:

$$x^3 - x = x(x - 1)(x + 1) = 0.$$

That is, everywhere except when $x = 0$ and $x = \pm 1$.

Step 2: The function is not even since $f(-x) = \frac{-xe^{-x}}{-x^3+x} \neq f(x)$. The function is not odd since $-f(-x) = \frac{xe^{-x}}{x^3-x} \neq f(x)$. In fact, there are no obvious symmetries.

Step 3: We first observe that

$$f(x) = \frac{xe^x}{x^3 - x} = \frac{e^x}{x^2 - 1}$$

for all $x \neq 0$. Since e^x is never 0, the function is never 0. The IVT then tells us that we could only have a sign change at a point of discontinuity.

Step 4: The function is the ratio of two continuous functions. Therefore, f is discontinuous only at $x = 0$, $x = -1$ and $x = 1$ since these are the only points where the denominator is 0.

Step 5: Since e^x is always positive and since $f(x) = \frac{xe^x}{x^3-x} = \frac{e^x}{x^2-1}$ for all $x \neq 0$, there is no sign change at $x = 0$. However, the function goes from positive to negative as we move across $x = -1$ (moving left to right) and then from negative to positive as we cross $x = 1$ (moving left to right).

In evaluating the limits, we will again use the fact that

$$f(x) = \frac{xe^x}{x^3 - x} = \frac{e^x}{x^2 - 1}$$

for all $x \neq 0$. This gives us that

$$\lim_{x \rightarrow 0} f(x) = \lim_{x \rightarrow 0} \frac{e^x}{x^2 - 1} = -1.$$

Hence, $x = 0$ is a removable discontinuity for f .

Since $e^x > 0$, this means that $x = 1$ and $x = -1$ are both vertical asymptotes for f . Furthermore, since $f(x) > 0$ if $x > 1$ or $x < -1$, and $f(x) < 0$ if $x \neq 0$ and $-1 < x < 1$, we get that

$$\lim_{x \rightarrow 1^+} f(x) = \infty,$$

$$\lim_{x \rightarrow 1^-} f(x) = -\infty,$$

$$\lim_{x \rightarrow -1^+} f(x) = -\infty,$$

and

$$\lim_{x \rightarrow -1^-} f(x) = \infty.$$

Step 6: Draw these vertical asymptotes on the sketch of the plot.

Step 7: Since e^x grows much more rapidly than any polynomial for large positive values of x , we have

$$\lim_{x \rightarrow \infty} f(x) = \infty.$$

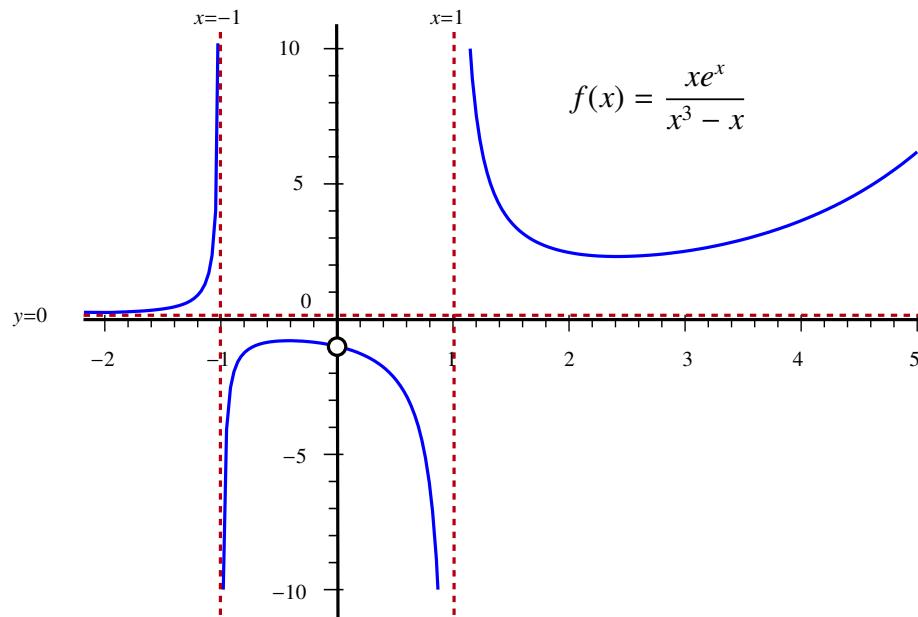
Thus, $f(x)$ grows without bound as $x \rightarrow \infty$.

But since e^x becomes very small for large negative values of x , we have

$$\lim_{x \rightarrow -\infty} f(x) = 0.$$

Thus, $y = 0$ is a horizontal asymptote as $x \rightarrow -\infty$. Draw this horizontal asymptote on your plot.

Step 8: Taking all of this information into consideration gives us the following sketch of the function.



Chapter 6

Derivatives

In this chapter we introduce and study the *derivative* of a function. Intuitively, derivatives can be viewed as *instantaneous* rates of change of a quantity. However, to make this statement more precise mathematically we will appeal to the theory of limits that we developed in the previous chapter.

6.1 Instantaneous Velocity

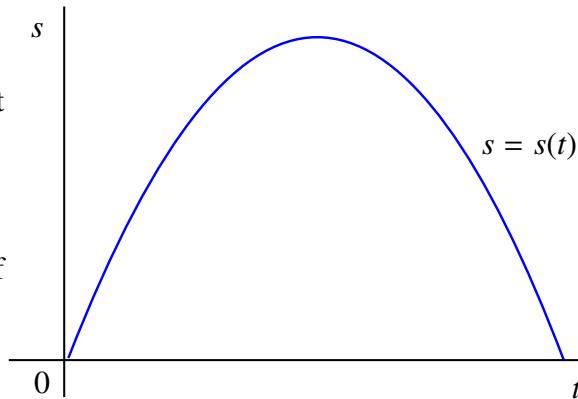
To motivate the concept of instantaneous rates of change, we begin by developing a definition of velocity by considering the following problem.

Problem:

A stone is thrown straight upward in the air and eventually falls back to the ground. How can we define the *instantaneous velocity* of the stone at any given time?

We begin by looking at the graph that represents the height s of the stone above the ground at time t .

Note: The graph represents the height function, *not* the actual path of the stone.



You will recall that the *average velocity* of the stone relative to the ground over the period from time $t = t_0$ to $t = t_1$ is given by the formula

$$\begin{aligned} V_{\text{ave}} &= \frac{\text{displacement (change in position)}}{\text{elapsed time}} \\ &= \frac{s(t_1) - s(t_0)}{t_1 - t_0} \\ &= \frac{\Delta s}{\Delta t} \end{aligned}$$

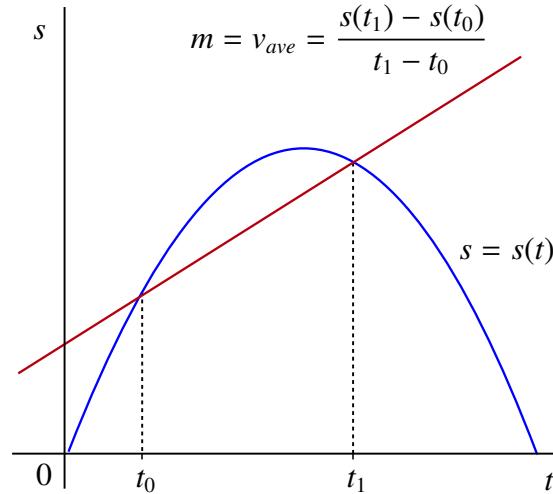
where

$$\Delta s = s(t_1) - s(t_0)$$

and

$$\Delta t = t_1 - t_0.$$

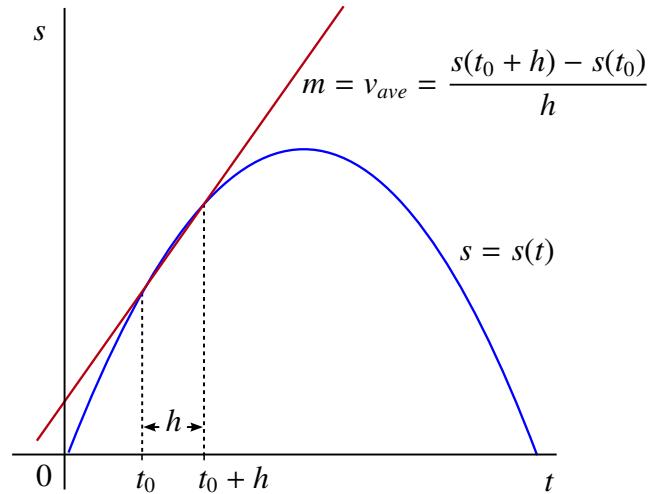
This can be realized geometrically as the slope m of the “secant line” to the graph of s through the points $(t_0, s(t_0))$ and $(t_1, s(t_1))$.



It makes sense that the velocity of the stone should not vary a great deal over very small intervals of time. Therefore, we should be able to use the average velocity over a small interval around t_0 to approximate $v(t_0)$, the instantaneous velocity at time t_0 .

As a first approximation, let h be a small number. We can calculate the average velocity for the time period between t_0 and $t_0 + h$ as follows:

$$\begin{aligned} v(t_0) &\cong v_{ave} \\ &= \frac{s(t_0 + h) - s(t_0)}{(t_0 + h) - t_0} \\ &= \frac{s(t_0 + h) - s(t_0)}{h} \end{aligned}$$



In general, it makes sense that the smaller h is, the better the estimate of $v(t_0)$. This leads us to define the instantaneous velocity to be the limit of the average velocities over smaller and smaller time intervals around t_0 . That is,

$$v(t_0) = \lim_{h \rightarrow 0} \frac{s(t_0 + h) - s(t_0)}{h}$$

provided this limit exists.

6.2 Definition of the Derivative

In the previous section, we saw how we could define the instantaneous velocity of a particle as the limit of average velocities. However, *velocity* is simply the instantaneous rate of change of displacement s with respect to time t and the *average velocity* is simply the average rate of change of displacement over a fixed interval of time. In this section, the same process is used to define the instantaneous rate of change for any quantity.

Given a function f , we can define the average rate of change as t goes from t_0 to t_1 to be the ratio

$$\frac{f(t_1) - f(t_0)}{t_1 - t_0}.$$

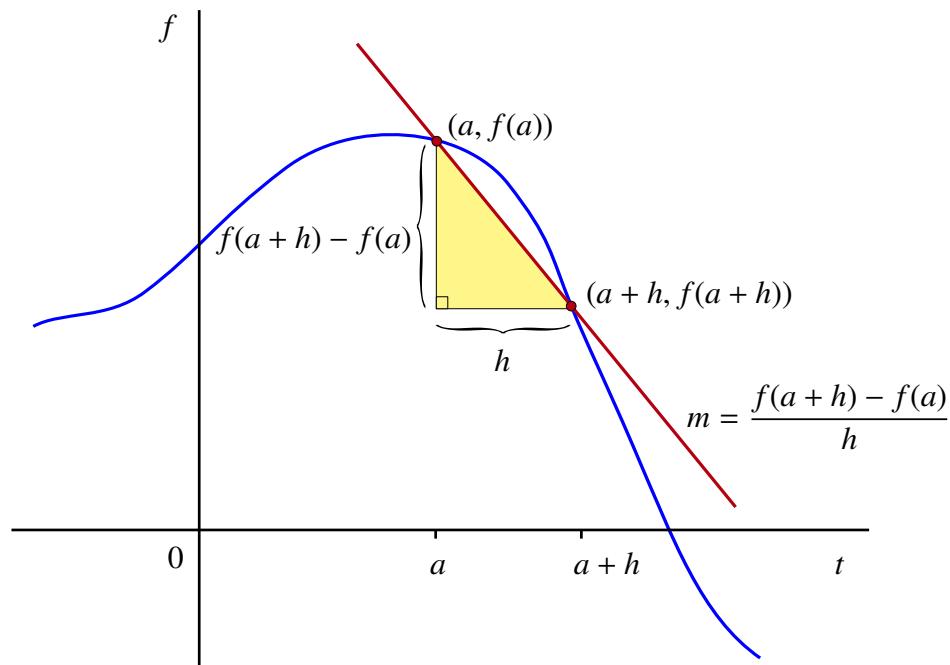
If we fix a point a and let h be small, then

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

again represents the *average* change in f over a small interval around a . The quotient

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

is called a *Newton Quotient* for f centered at a . Geometrically, the Newton Quotient represents the slope of the secant line to the graph of f through the points $(a, f(a))$ and $(a + h, f(a + h))$.



In the same manner that we defined velocity, we should be able to approximate the instantaneous rate of change of f at $t = a$ by calculating the average rate of change over smaller and smaller intervals around $t = a$. This leads us to the following familiar definition.

DEFINITION**The Derivative at $t = a$**

We say that the function f is differentiable at $t = a$ if

$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}$$

exists.

In this case, we write

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

and we call $f'(a)$ the derivative of f at $t = a$.

There is an alternate form for the definition of the derivative that is also quite useful. It can be obtained by noting that if $t = a + h$, then as $h \rightarrow 0$ we have $t \rightarrow a$. Furthermore, since $h = t - a$ we get

$$\begin{aligned} f'(a) &= \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \\ &= \lim_{t \rightarrow a} \frac{f(t) - f(a)}{t - a} \end{aligned}$$

provided the limits exist.

As we have already suggested, if $y = f(t)$, $\Delta y = f(t) - f(a)$ and $\Delta t = t - a$, then the Newton Quotient

$$\frac{\Delta y}{\Delta t} = \frac{f(t) - f(a)}{t - a}$$

is the *average* change in y . Letting $\Delta t \rightarrow 0$ gives us that

$$f'(a) = \lim_{\Delta t \rightarrow 0} \frac{\Delta y}{\Delta t}$$

is the limit of average rates of change over smaller and smaller intervals and as such represents the *instantaneous rate of change of y with respect to t* .

EXAMPLE 1

Let $s = s(t)$ represent the displacement of an object. Then

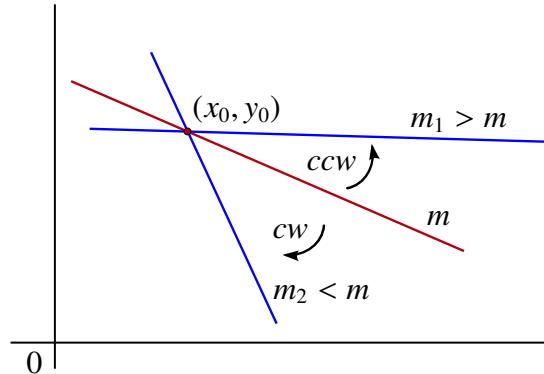
$$\begin{aligned} s'(t_0) &= \lim_{\Delta t \rightarrow 0} \frac{\Delta s}{\Delta t} \\ &= \lim_{t \rightarrow t_0} \frac{s(t) - s(t_0)}{t - t_0} \\ &= \lim_{h \rightarrow 0} \frac{s(t_0 + h) - s(t_0)}{h} \\ &= v(t_0) \end{aligned}$$

where $v(t_0)$ is the instantaneous velocity of the object at time t_0 . This shows us that *velocity is the derivative of displacement*. ◀

The existence of the derivative also has a very important geometric consequence.

6.2.1 The Tangent Line

Recall that given a fixed point (x_0, y_0) in the Real plane, all non-vertical lines that pass through (x_0, y_0) are determined by the slope m of the line. Increasing m corresponds to rotating the line in a counter-clockwise direction. Decreasing m corresponds to a clockwise rotation.



Assume that $f'(a)$ exists. Pick h_0, h_1, h_2 and h_3 with

$$h_0 > h_1 > h_2 > h_3 > 0.$$

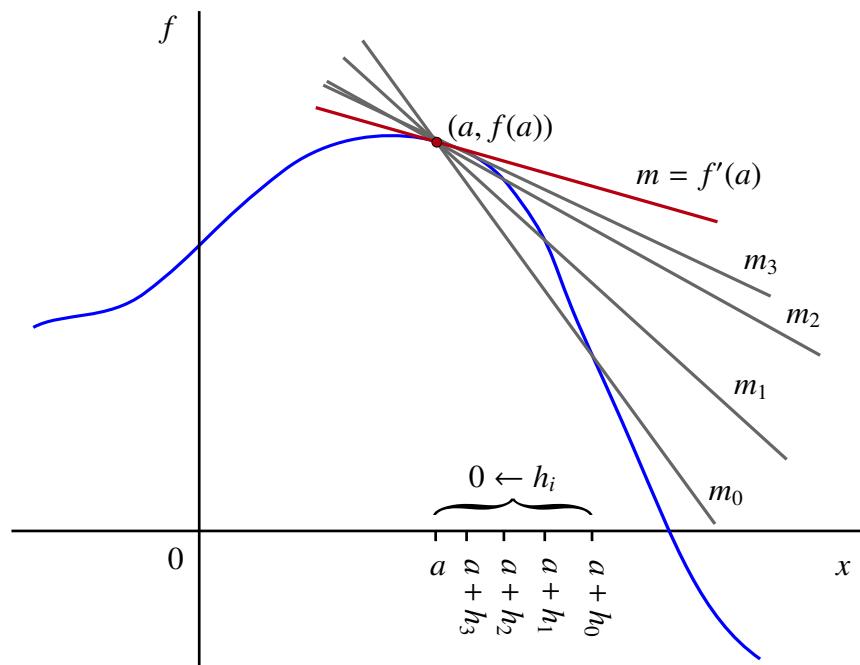
For $i = 0..3$, let

$$m_i = \frac{f(a + h_i) - f(a)}{h_i}$$

and

$$m = f'(a).$$

The following diagram shows the graph of f , the secant lines through $(a, f(a))$ and $(a + h_i, f(a + h_i))$ for $i = 0, 1, 2, 3$ with slope m_i , respectively, and the unique line passing through $(a, f(a))$ with slope $m = f'(a)$.



Notice that as $h \rightarrow 0$, the slopes of the secant lines m_0, m_1, m_2 , and m_3 are getting closer to m . In the diagram, this is represented by the secant lines visually “converging” to the line passing through $(a, f(a))$ with slope $m = f'(a)$. That is, we can view the line passing through $(a, f(a))$ with slope $m = f'(a)$ as a “limit” of secant lines passing through $(a, f(a))$. We call this line the *tangent line* to the graph of f at $x = a$.

DEFINITION

The Tangent Line

Assume that f is differentiable at $x = a$. The *tangent line* to the graph of f at $x = a$ is the line passing through $(a, f(a))$ with slope $m = f'(a)$.

It follows that the equation of the tangent line is

$$y = f(a) + f'(a)(x - a).$$

NOTE

It is often said that “the derivative is the slope of the tangent line.” While we have certainly just seen that this statement is true, it is *not* appropriate to use this statement as the definition of the derivative. In fact, without *first* defining the derivative as a limit of Newton Quotients, it is not at all obvious what we mean by the tangent line. This is a subtle point, but an important one to remember. 

Finally, we want to highlight an important relationship between continuity and differentiability.

6.2.2 Differentiability versus Continuity

Suppose that f is differentiable at $t = a$. Then

$$f'(a) = \lim_{t \rightarrow a} \frac{f(t) - f(a)}{t - a}.$$

Now since $t \rightarrow a$, we have $t - a \rightarrow 0$. But we know from our study of limits that if the limit of a quotient exists, and if the denominator approaches zero, then the numerator must also approach zero. This means that if f is differentiable at $t = a$, then

$$\lim_{t \rightarrow a} f(t) - f(a) = 0$$

or

$$\lim_{t \rightarrow a} f(t) = f(a).$$

However, this last statement implies that f is continuous at $t = a$. This establishes the following important theorem.

THEOREM 1 **Differentiability Implies Continuity**

Assume that f is differentiable at $t = a$. Then f is continuous at $t = a$.

EXAMPLE 2 Let's illustrate graphically why differentiability implies continuity. Consider

$$f(t) = \begin{cases} \frac{|t|}{t} & \text{if } t \neq 0 \\ 0 & \text{if } t = 0 \end{cases}.$$

Then we know that

$$f(t) = \begin{cases} 1 & \text{if } t > 0 \\ 0 & \text{if } t = 0 \\ -1 & \text{if } t < 0 \end{cases}.$$

In this case,

$$\lim_{t \rightarrow 0^-} f(t) = -1 \neq 1 = \lim_{t \rightarrow 0^+} f(t)$$

so f is certainly not continuous at $t = 0$. Therefore, it follows from the previous theorem, that f is not differentiable at $t = 0$.

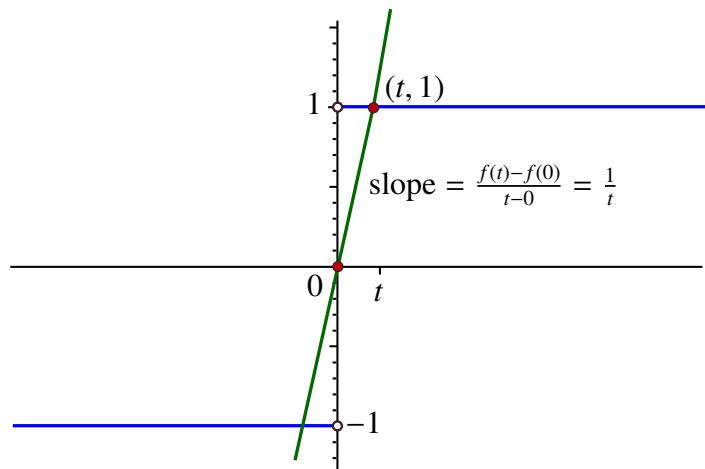
In fact, if we were to evaluate

$$\lim_{t \rightarrow 0^+} \frac{f(t) - f(0)}{t - 0}$$

we would get

$$\lim_{t \rightarrow 0^+} \frac{1}{t}$$

since $f(0) = 0$ and if $t > 0$, then $f(t) = 1$.



However, we have seen in our study of vertical asymptotes that

$$\lim_{t \rightarrow 0^+} \frac{1}{t} = \infty.$$

That is, the slopes of these secant lines approach ∞ as $t \rightarrow 0^+$. Moreover, since this one-sided limit does not exist,

$$\lim_{t \rightarrow 0} \frac{f(t) - f(0)}{t - 0}$$

does not exist, and hence f is not differentiable at $t = 0$. ◀

Now that we have established that differentiability implies continuity, it makes sense to ask if the converse also holds. That is, does continuity imply differentiability? We will see that this is *not* the case.

EXAMPLE 3 Let $f(x) = |x|$ and let $a = 0$. We can see from the graph of f that $|x|$ is continuous at 0.

We are interested in calculating

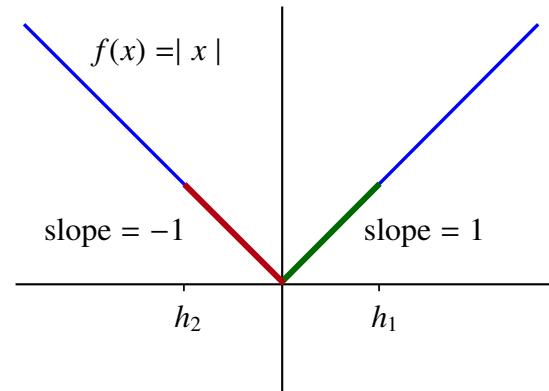
$$\lim_{x \rightarrow 0} \frac{f(x) - f(0)}{x - 0}.$$

But since $f(0) = 0$, we get that

$$\lim_{x \rightarrow 0} \frac{f(x) - f(0)}{x - 0} = \lim_{x \rightarrow 0} \frac{|x|}{x}$$

and we know that this last limit does not exist (see previous example). Therefore, f is not differentiable at 0. In fact, we can see this clearly from an examination of the graph of $f(x) = |x|$.

If we choose $h_1 > 0$, then the slope of the secant line through $(0, f(0)) = (0, 0)$ and $(h_1, f(h_1)) = (h_1, h_1)$ is 1. However, if we choose $h_2 < 0$, then the slope of the secant line through $(0, f(0)) = (0, 0)$ and $(h_2, f(h_2)) = (h_2, -h_2)$ is -1.



This means that

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

but

$$\begin{aligned} \lim_{x \rightarrow 0^-} \frac{f(x) - f(0)}{x - 0} &= \lim_{x \rightarrow 0^-} \frac{|x|}{x} \\ &= \lim_{x \rightarrow 0^-} \frac{-x}{x} \\ &= -1. \end{aligned}$$

We have just seen that when the two one-sided limits from the Newton Quotients exist but are different, the derivative fails to exist. Geometrically, because the two one-sided limits are different, we have a “sharp” point at 0 on the graph of $|x|$. These sharp points are the most common sign that a continuous function fails to be differentiable. ►

We have just seen that **continuity does *not* imply differentiability**. However, in this course, all of the continuous functions that we study will be differentiable at most points in their domain. We might be led to believe that this is always the case. Unfortunately, using ideas that are beyond the scope of this course, it is possible to build functions that are continuous at each point, but are not differentiable anywhere!

It would be interesting to know what such a *nowhere differentiable* function might look like. However, it turns out that it is impossible to draw such a function, but you can get an idea about what its graph might look like by comparing it to a rocky coastline. At a distance, the coastline has many visible nooks and crannies. Moreover, as you look more closely at a small piece of the coastline, you see even more jagged edges corresponding to the sharp corner that we saw in the previous example. We know that these sharp points indicate where the derivative does not exist. This phenomenon continues even if you inspect a single rock that composes part of the coastline, and then even at the microscopic level. Similar behavior can be observed if you look closely at the edge of a snow flake.

Note: Continuous, nowhere differentiable functions are rather strange. This unusual behavior leads to the study of *fractals*, an important area of modern mathematics with many real-life applications.

6.3 The Derivative Function

Up until now we have only considered the derivative at a fixed point a in the domain of a function. We will now consider the derivative on an interval.

DEFINITION

The Derivative Function

We say that a function f is differentiable on an interval I if $f'(a)$ exists for every $a \in I$. In this case, we define the derivative function, denoted by f' , as

$$f'(t) = \lim_{h \rightarrow 0} \frac{f(t+h) - f(t)}{h}.$$

That is, the value of the derivative function at t is simply the derivative of f at t for each $t \in I$.

There is an important alternative to the notation associated with the derivative. This notation is due to Gottfried Wilhelm Leibniz who, along with Isaac Newton, is often credited with having invented modern Calculus.

Leibniz Notation

Given a function $y = f(t)$, Leibniz wrote

$$\frac{dy}{dt}$$

for the derivative of y (or equivalently, of f) with respect to t . An alternate form of Leibniz's notation is to write

$$\frac{df}{dt} \quad \text{or} \quad \frac{d}{dt}(f)$$

to indicate that f is to be differentiated with respect to the variable t . The symbol

$$\frac{d}{dt}$$

is called a *differential operator*.

In Leibniz's notation, we denote $f'(a)$, the derivative at $t = a$, by

$$\frac{dy}{dt}\bigg|_a.$$

That is,

$$\frac{dy}{dt}\bigg|_a = f'(a).$$

Note that it also makes sense to differentiate the function f' . The derivative of this function, if it exists, is called the *second derivative* of f and is denoted by f'' . We could then differentiate f'' to get the third derivative f''' , and so on. This leads us to define the higher derivatives of f .

DEFINITION

Higher Derivatives

Let f be a differentiable function with respect to x with derivative f' . If f' is also differentiable, then its derivative

$$\frac{d}{dx}(f')$$

is called the *second derivative* of f and it is usually denoted by

$$f''.$$

It is also commonly denoted by either

$$f^{(2)} \text{ or } \frac{d^2}{dx^2}(f).$$

If f'' is also differentiable, then its derivative is called the *third derivative* of f and it is denoted by

$$f''' \text{ or } f^{(3)}.$$

In general, for any $n \geq 1$,

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

and $f^{(n)}$ is called the n -th derivative of f .

We will see later that f'' impacts the geometry of the graph of f . In particular, the larger the magnitude of f'' , the more *curved* the graph of f .

6.4 Derivatives of Elementary Functions

In this section the definition of the derivative and the theory of limits are used to determine the derivatives of some important functions. Using the definition of the derivative to calculate derivatives is often called *differentiation by first principles*.

EXAMPLE 4 The Derivative of a Constant Function

Assume that f is a constant function. That is, there exists some $c \in \mathbb{R}$ such that

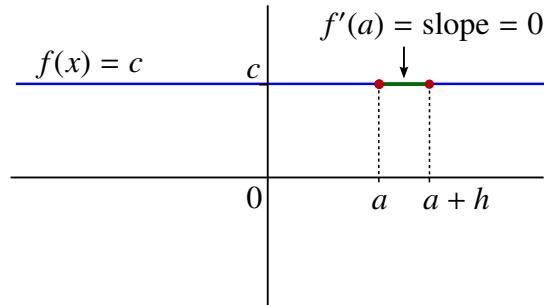
$$f(x) = c$$

for every $x \in \mathbb{R}$. Fix $a \in \mathbb{R}$. We want to find $f'(a)$ if it exists. To do this we evaluate

$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}.$$

However, since $f(a+h) = c = f(a)$ for each $h \neq 0$,

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

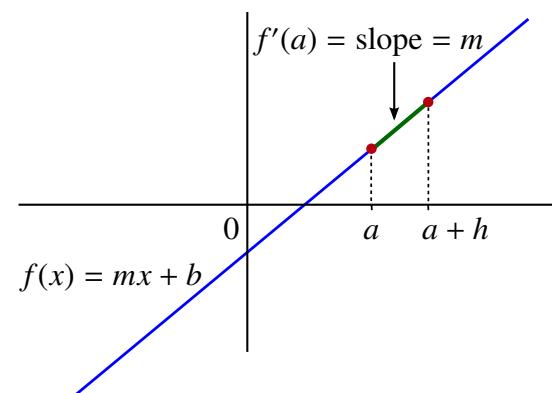


This result should be expected since $f'(a)$ represents the instantaneous rate of change of f at $x = a$. Since **constant functions do not change in value (horizontal line)**, the slope of a constant function is always 0 and it follows that the derivative at any point should also be 0. ►

EXAMPLE 5 The Derivative of a Linear Function

Let $f(x) = mx + b$. Then the graph of f is the straight line with slope m .

Choose $a \in \mathbb{R}$. For $h \neq 0$, the secant line joining $(a, f(a))$ and $(a + h, f(a + h))$ is coincident with the line that is the graph of f . Therefore, any secant line to $f(x)$ has slope m .



This suggests that $f'(a) = m$. We can verify this algebraically as follows:

$$\begin{aligned}
 f'(a) &= \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{(m(a+h) + b) - (ma + b)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{ma + mh + b - ma - b}{h} \\
 &= \lim_{h \rightarrow 0} \frac{mh}{h} \\
 &= m.
 \end{aligned}$$

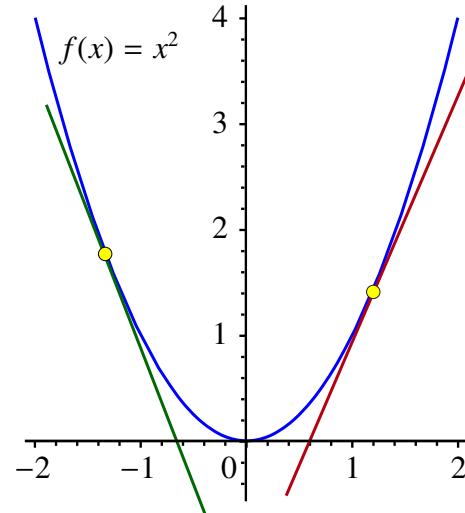
In particular, if $f(x) = x$ and if $g(x) = 7x + 4$, then $f'(x) = 1$ and $g'(x) = 7$ for all x .



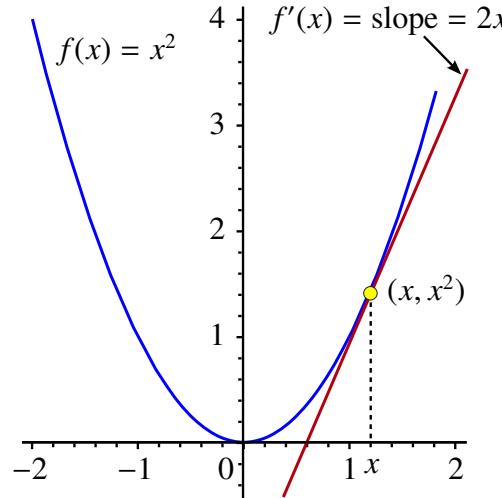
EXAMPLE 6 The Derivative of a Simple Quadratic Function

Calculate the derivative of $f(x) = x^2$.

Unlike the previous examples, the derivative is not constant for all x . We can see this by observing that the slopes of the tangent lines through $(x, f(x))$ vary as x varies.



We can use first principles to find the value of the derivative of x^2 at any point x :



$$f'(x) = \text{slope} = 2x$$

$$\begin{aligned}
 f'(x) &= \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{(x+h)^2 - x^2}{h} \\
 &= \lim_{h \rightarrow 0} \frac{x^2 + 2xh + h^2 - x^2}{h} \\
 &= \lim_{h \rightarrow 0} \frac{2xh + h^2}{h} \\
 &= \lim_{h \rightarrow 0} 2x + h \\
 &= 2x.
 \end{aligned}$$



The next example is a very important calculation of a derivative by first principles.

6.4.1 The Derivative of $\sin(x)$ and $\cos(x)$

To calculate the derivative of $\sin(x)$ we need to recall two very important facts. The first is the formula for the sine of a sum of angles. That is,

$$\sin(x + y) = \sin(x)\cos(y) + \cos(x)\sin(y).$$

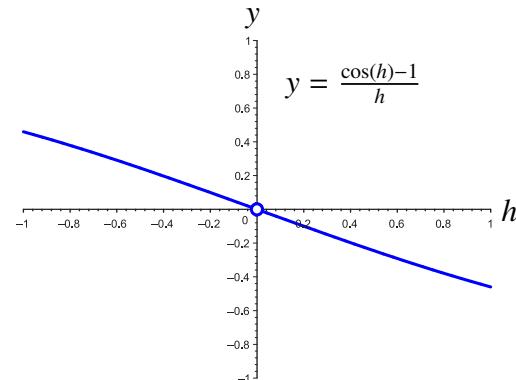
The second is the Fundamental Trig Limit,

$$\lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 1.$$

We also require another limit which can be derived from the Fundamental Trig Limit, namely that

$$\lim_{h \rightarrow 0} \frac{\cos(h) - 1}{h} = 0.$$

We can present graphical evidence to support our claim concerning this limit:



Let's derive this limit directly. To do so we first note that for any h near enough to 0, $\cos(h) \neq -1$, so

$$\frac{\cos(h) + 1}{\cos(h) + 1} = 1.$$

Therefore, we have

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

since $\lim_{h \rightarrow 0} \frac{\sin(h)}{h} = 1$, $\lim_{h \rightarrow 0} (-\sin(h)) = 0$ and $\lim_{h \rightarrow 0} (\cos(h) + 1) = 2$.

Now that we have established these facts, we can proceed directly to calculate the derivative (function) of $\sin(x)$.

The Derivative (Function) of $\sin(x)$

We want to consider

$$\lim_{h \rightarrow 0} \frac{\sin(x + h) - \sin(x)}{h}.$$

Using the rule for the sine of a sum of angles, this limit becomes

$$\lim_{h \rightarrow 0} \frac{(\sin(x) \cos(h) + \cos(x) \sin(h)) - \sin(x)}{h}.$$

Then

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

We have established the following very important theorem.

THEOREM 2

The Derivative of $\sin(x)$

Assume that $f(x) = \sin(x)$. Then

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

The Derivative (Function) of $\cos(x)$

To find the derivative of $\cos(x)$ we use a very similar calculation as above. This time we want to consider

$$\lim_{h \rightarrow 0} \frac{\cos(x + h) - \cos(x)}{h}.$$

Using the rule for the cosine of a sum of angles, this limit becomes

$$\lim_{h \rightarrow 0} \frac{(\cos(x) \cos(h) - \sin(x) \sin(h)) - \cos(x)}{h}.$$

Then

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{(\cos(x) \cos(h) - \sin(x) \sin(h)) - \cos(x)}{h} &= \lim_{h \rightarrow 0} \left(\cos(x) \frac{\cos(h) - 1}{h} - \sin(x) \frac{\sin(h)}{h} \right) \\ &= \cos(x) \lim_{h \rightarrow 0} \frac{\cos(h) - 1}{h} - \sin(x) \lim_{h \rightarrow 0} \frac{\sin(h)}{h} \\ &= \cos(x) \cdot 0 - \sin(x) \cdot 1 \\ &= -\sin(x). \end{aligned}$$

This result establishes the following theorem:

THEOREM 3

The Derivative of $\cos(x)$

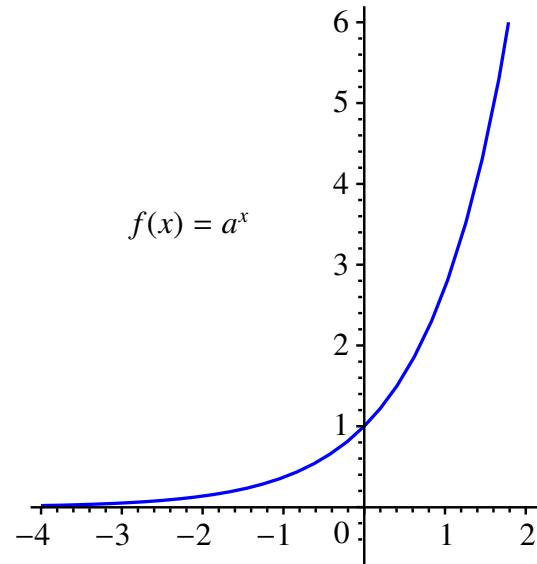
Assume that $f(x) = \cos(x)$. Then

$$f'(x) = -\sin(x).$$

6.4.2 The Derivative of e^x

Recall that for any $a > 1$, $a^0 = 1$ and the exponential function $f(x) = a^x$ produces the following type of graph through the point $(0, 1)$.

You can see that the graph has a very *smooth* appearance which is characteristic of a differentiable function. As such we can speculate that $f(x) = a^x$ should be differentiable everywhere.



Now under the assumption that $f(x) = a^x$ is differentiable, we can try to calculate the derivative. Then

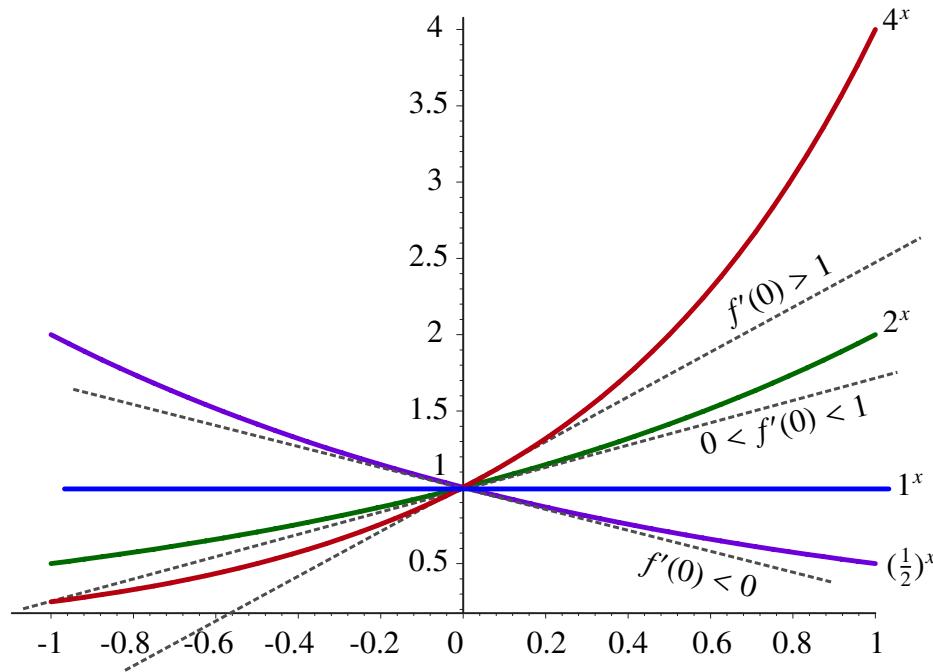
$$\begin{aligned} f'(x) &= \lim_{h \rightarrow 0} \frac{a^{x+h} - a^x}{h} \\ &= \lim_{h \rightarrow 0} \frac{a^x a^h - a^x}{h} \\ &= a^x \cdot \lim_{h \rightarrow 0} \frac{a^h - 1}{h} \\ &= a^x \cdot f'(0). \end{aligned}$$

This calculation tells us that

$$f'(x) = C_a f(x)$$

where the constant C_a is the value of the derivative at $x = 0$. In this way the derivative at $x = 0$ characterizes the function $f(x) = a^x$.

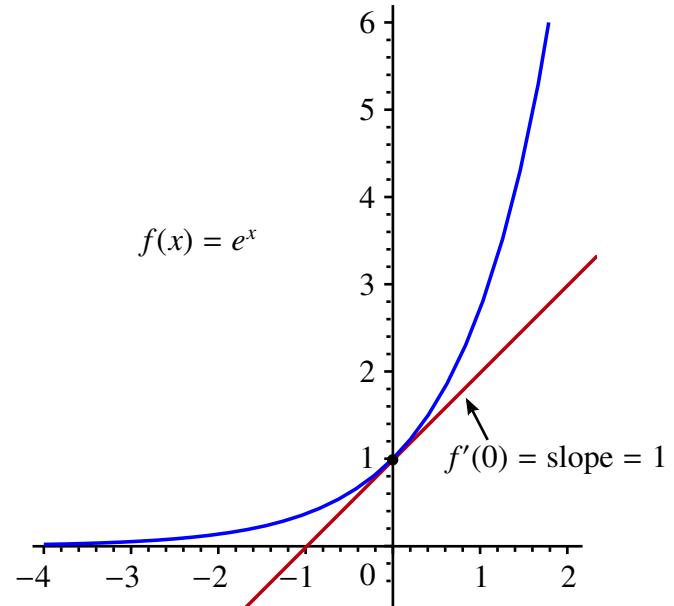
The following diagram gives us a sense of what the value of the derivative $f'(0)$ might be for various choices of the base a .



Notice in the diagram that if $0 < a < 1$, then $f'(0) < 0$. If $a = 1$, then $f'(0) = 0$. If $a = 2$, then $0 < f'(0) < 1$ and if $a = 4$, then $f'(0) > 1$. Moreover, as a increases so does $f'(0)$.

NOTE

Of all the possible choices for a , there is a unique value so that the slope of the tangent to the graph of a^x through $(0, 1)$ is 1. We call this number e . From the previous diagram we can see that $2 < e < 4$. In fact, e is known to be an irrational number that is approximately 2.718281828.



If $f(x) = e^x$, the slope of the tangent line at $x = 0$ is

$$\begin{aligned}
 1 &= f'(0) \\
 &= \lim_{h \rightarrow 0} \frac{f(0 + h) - f(0)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{e^h - e^0}{h} \\
 &= \lim_{h \rightarrow 0} \frac{e^h - 1}{h}.
 \end{aligned}$$

This gives us the important limit

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

This limit and the basic properties of exponential functions can be used to evaluate the derivative of $f(x) = e^x$ at any $x \in \mathbb{R}$.

Consider

$$\begin{aligned} f'(x) &= \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \\ &= \lim_{h \rightarrow 0} \frac{e^{x+h} - e^x}{h} \\ &= \lim_{h \rightarrow 0} \frac{e^x e^h - e^x}{h} \\ &= \lim_{h \rightarrow 0} \frac{e^x (e^h - 1)}{h} \\ &= e^x \cdot \lim_{h \rightarrow 0} \frac{e^h - 1}{h} \\ &= e^x \cdot 1 \\ &= e^x. \end{aligned}$$

Therefore, we have shown that e^x has the following remarkable property.

THEOREM 4

The Derivative of e^x

Assume that $f(x) = e^x$. Then

$$f'(x) = e^x.$$

We have just seen that the function $f(x) = e^x$ has the unusual property that it is equal to its own derivative. Later in the course we will be able to show that if g is any function such that $g(x) = g'(x)$ for all $x \in \mathbb{R}$, then there exists a constant $c \in \mathbb{R}$, such that $g(x) = ce^x$ for all $x \in \mathbb{R}$ (see *The Mean Value Theorem*).

For functions of the form $f(x) = a^x$, we have the following theorem.

THEOREM 5

The Derivative of a^x

Assume $a > 0$ and that $f(x) = a^x$. Then

$$f'(x) = \ln(a) a^x.$$

6.5 Tangent Lines and Linear Approximation

A central goal in many applications of mathematics is to approximate complicated objects by simpler ones in a way that the error can be kept very small. This is certainly one of the main themes in Calculus.

Perhaps the simplest types of functions are *linear* functions of the form $h(x) = mx + b$. In this section, we will see that if f is differentiable at a point $x = a$, then it is possible to approximate f by a linear function h with the properties that $h(a) = f(a)$, $h'(a) = f'(a)$, and that if x is close to a , then it is reasonable to expect that $h(x)$ will be very close to $f(x)$. To see why this might be so we make the following key observation.

Observation: Suppose that f is differentiable at $x = a$ with derivative $f'(a)$. Then by definition:

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

This tells us that *for values of x that are close to a we have*

$$\frac{f(x) - f(a)}{x - a} \cong f'(a). \quad (*)$$

If we rearrange $(*)$, we get

$$f(x) - f(a) \cong f'(a)(x - a),$$

and finally that

$$f(x) \cong f(a) + f'(a)(x - a).$$

So, in summary, if we define a new function $L_a^f(x)$ by

$$L_a^f(x) = f(a) + f'(a)(x - a),$$

then provided that x is close to a we have

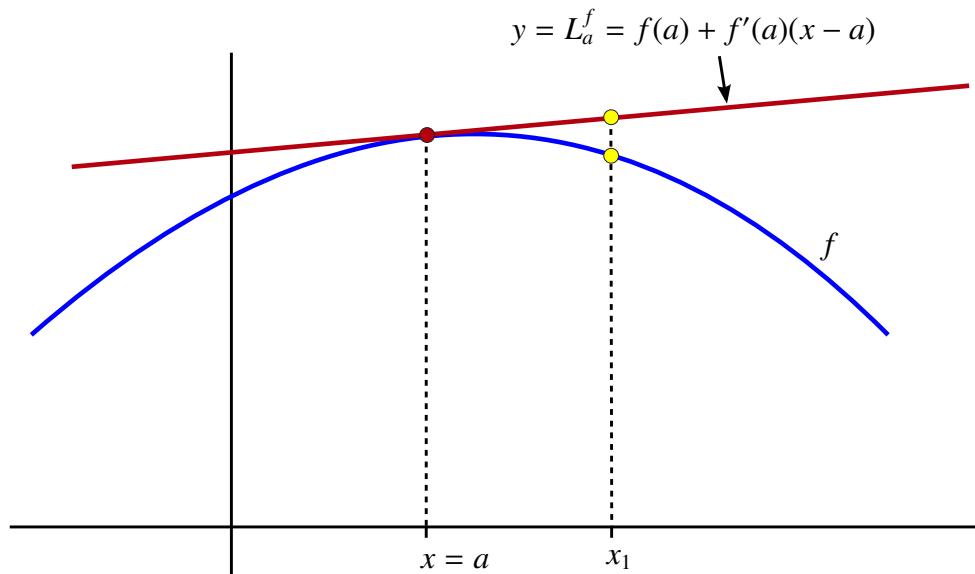
$$f(x) \cong L_a^f(x).$$

That is, $L_a^f(x)$ approximates $f(x)$ near $x = a$.

Also of interest is that the graph of L_a^f is actually a *line* with equation

$$y = f(a) + f'(a)(x - a).$$

In fact, it is not some arbitrary line, but rather the *tangent* line to the graph of f through $(a, f(a))$.



For this reason, we call L_a^f the *linear approximation to f centered at $x = a$* .

DEFINITION

Linear Approximation

Let $y = f(x)$ be differentiable at $x = a$. The linear approximation to f at $x = a$ is the function

$$L_a^f(x) = f(a) + f'(a)(x - a).$$

L_a^f is also called the *linearization* of f or the *tangent line approximation* to f at $x = a$.

Note: If f is clear from the context, then we will simplify this notation and write L_a to represent the linear approximation.

In summary, if f is differentiable at $x = a$ and if x is close to a , then we can approximate a complicated function f with the much simpler linear function L_a . That is, if x is sufficiently close to a , then

$$f(x) \cong L_a(x).$$

There are 3 very important properties of L_a that you should keep in mind. These are:

Three Properties of the Linear Approximation:

1. $L_a(a) = f(a)$.
2. L_a is differentiable and $L_a'(a) = f'(a)$.
3. L_a is the only first degree polynomial with properties (1) and (2).

Let's see how this works in the case of a familiar function.

EXAMPLE 7 The Fundamental Trig Limit [Revisited]

Let $f(x) = \sin(x)$ and $a = 0$. We know that $f(0) = \sin(0) = 0$ and $f'(x) = \cos(x)$, so $f'(0) = \cos(0) = 1$. Therefore, the linear approximation to $\sin(x)$ at $x = 0$ is

$$L_0(x) = f(0) + f'(0)(x - 0) = 0 + 1(x - 0) = x.$$

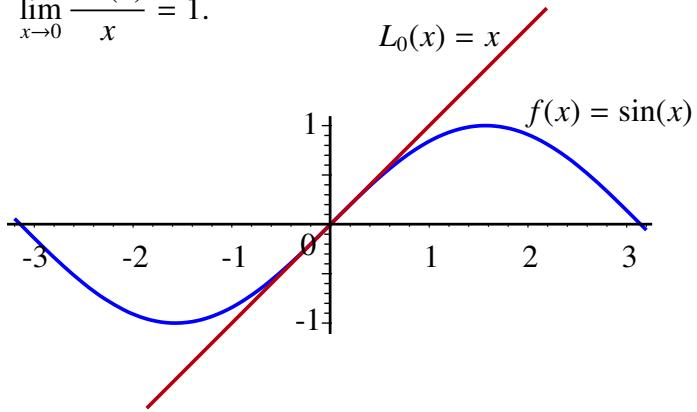
This means that if x is near 0, then

$$\sin(x) \cong L_0(x) = x.$$

This result is something that we already knew since it follows from the fact that

$$\lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 1.$$

The following diagram illustrates that $L_0(x) = x$ is a very good approximation to $\sin(x)$ near 0.



To investigate the accuracy of this estimate and to see how simple this process is to use, we will estimate $\sin(0.01)$.

$$\sin(0.01)$$

Since 0.01 is very close to 0 and we know that $\sin(0) = 0$, to find the estimate we simply write

$$\sin(0.01) \cong L_0(0.01) = x|_{0.01} = .01$$

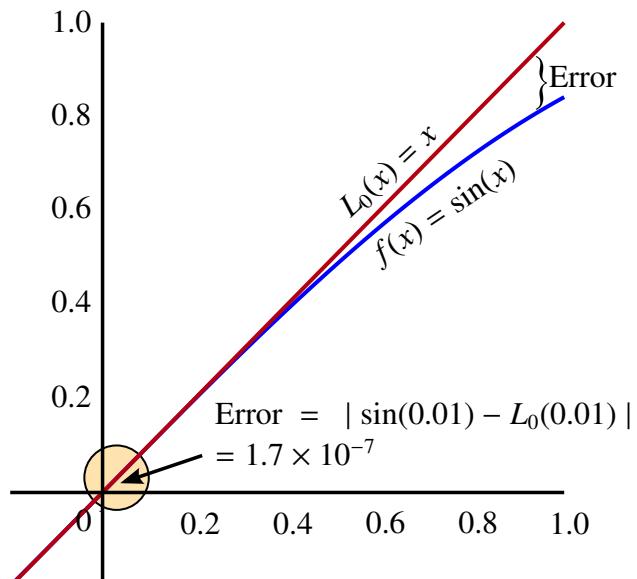
If we were to use a calculator (radian mode) to evaluate $\sin(0.01)$, we would find that

$$\sin(0.01) = .00999983$$

to eight decimal places. This means that the error is

$$\begin{aligned} \text{Error} &= |\sin(0.01) - L_0(0.01)| \\ &= .00000017 \\ &= 1.7 \times 10^{-7} \end{aligned}$$

which is remarkably accurate for such a simple process.



Moreover, we also know that the estimate is too *large* since the tangent line sits *above* the graph at $x = 0.01$.

EXAMPLE 8 The fact that if f is differentiable at $x = a$ implies that

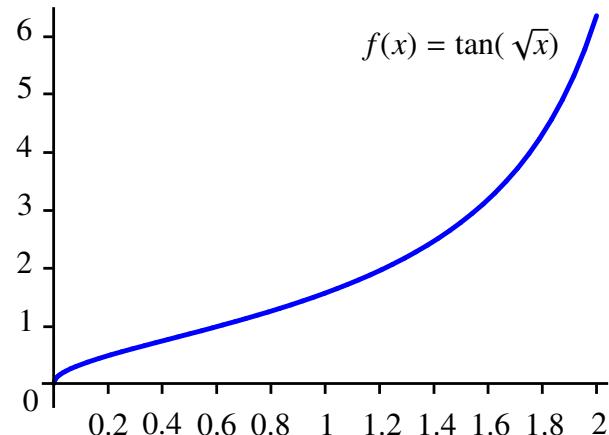
$$f(x) \cong L_a(x)$$

near $x = a$ can be interpreted to mean that *locally every differentiable function looks like a straight line*. This is not a very precise statement but it is somewhat akin to the fact that if you are in the middle of an ocean and look towards the horizon, the world appears to be very flat. In contrast, if you view the earth from the space station you can clearly see that it is not flat.

To further illustrate what we mean by this statement let's look at the function

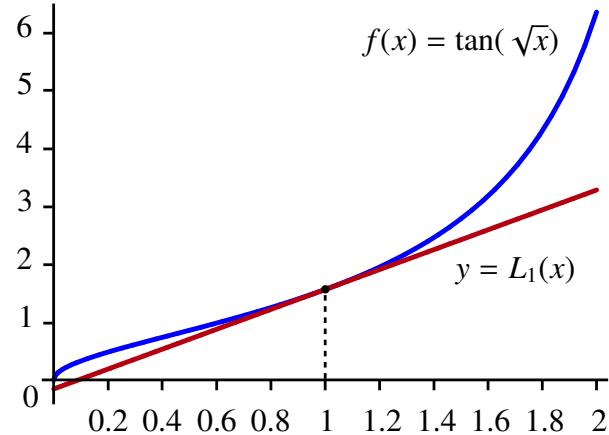
$$f(x) = \tan(\sqrt{x})$$

on the interval $[0, 2]$ centered around $x = 1$.



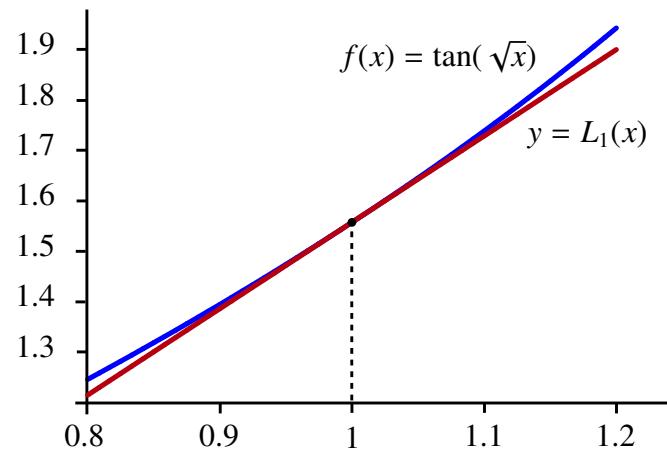
Now let's add in the tangent line at $x = 1$ which corresponds to the graph of L_1 .

The diagram illustrates that near $x = 1$, L_1 does a very good job of approximating the function $\tan(\sqrt{x})$. However, as we move away from $x = 1$, we can certainly see that the two functions are different.



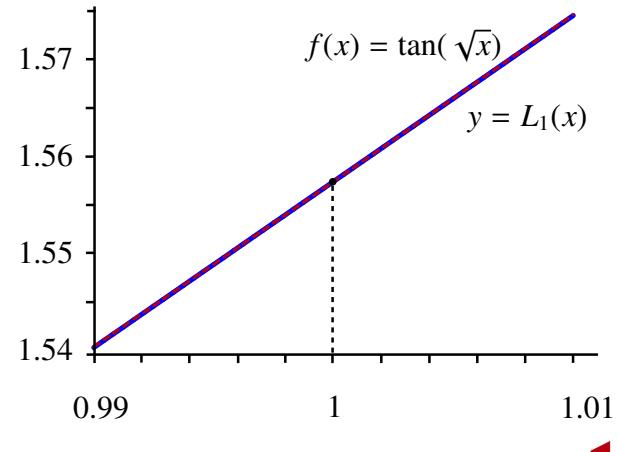
Next let's look at the graphs of f and L_1 on the interval $[0.8, 1.2]$ centered at $x = 1$.

We can still distinguish between the two functions but on this scale the graph of $\tan(\sqrt{x})$ looks very close to the line $y = L_1(x)$, and only deviates from it near the extremes of the interval.



Finally, let's focus our view on the interval $[0.99, 1.01]$.

In this case, within the accuracy of our graphing tool, it is essentially impossible to distinguish between $f(x) = \tan(\sqrt{x})$ and its linear approximation L_1 on this very small interval. In particular, the graph of $\tan(\sqrt{x})$ appears similar to a line over this interval.



6.5.1 The Error in Linear Approximation

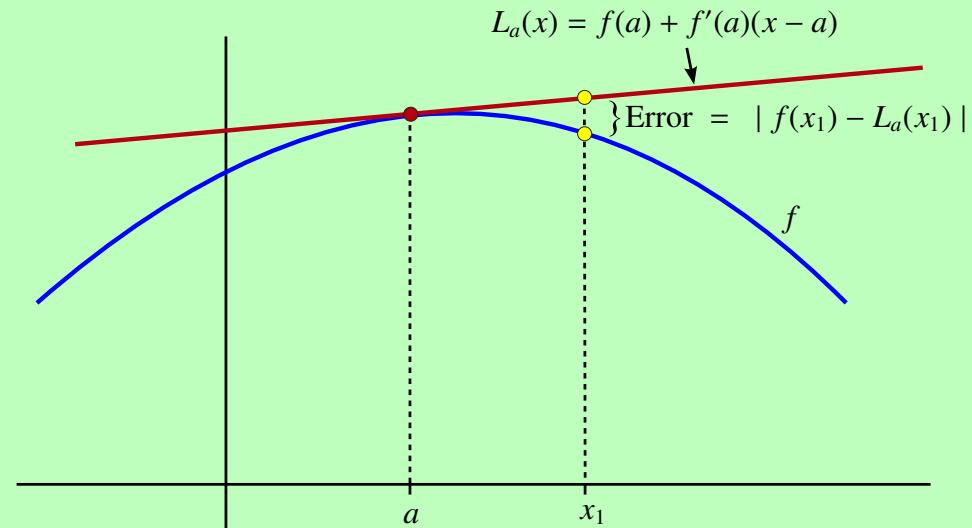
Anytime a process for approximation is used, it is always appropriate to have an understanding about the size of the error.

DEFINITION

The Error in Linear Approximation

Let $y = f(x)$ be differentiable at $x = a$. The error in using linear approximation to estimate $f(x)$ is given by

$$\text{Error} = |f(x) - L_a(x)|.$$



Notice that graphically, the error is represented by the vertical distance from the graph of $f(x)$ at $x = a$ to the tangent line $y = L_a(x)$.

Question: What major factors affect the error in linear approximation?

First Observation: Since the approximation

$$f(x) \cong f(a) + f'(a)(x - a) = L_a(x)$$

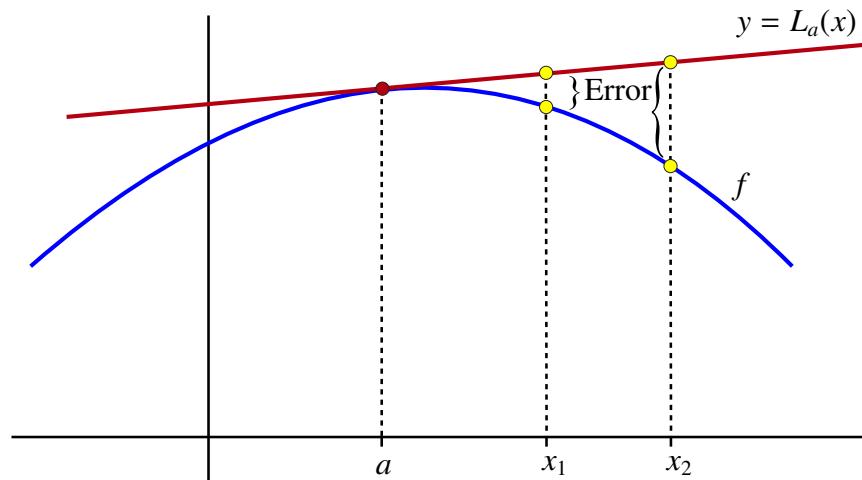
was obtained from the limit

$$f'(a) = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a}$$

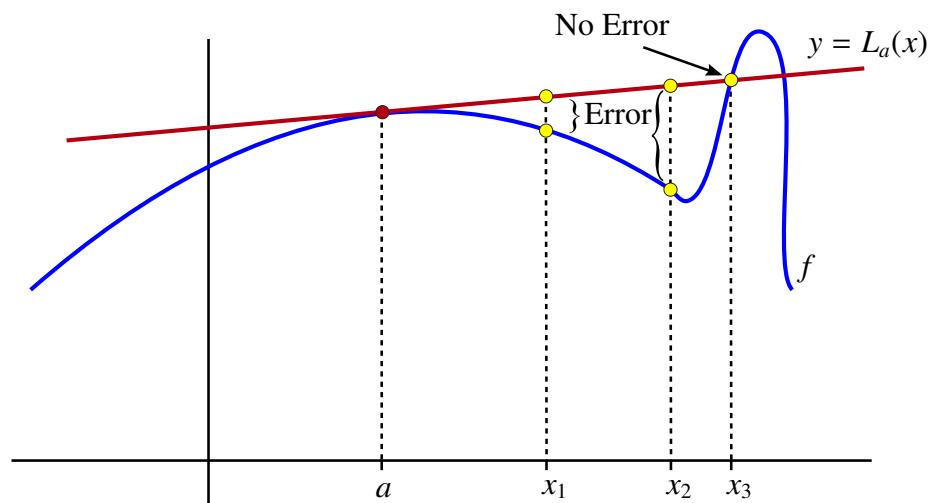
it would make sense that the farther we are from a , the further away $\frac{f(x) - f(a)}{x - a}$ might be from $f'(a)$, and hence the larger the potential error. That is, the larger the value of

$$|x - a|$$

the larger the possible error. In the following diagram we see that as we move away from a to x_1 and then to x_2 , the error does indeed grow.

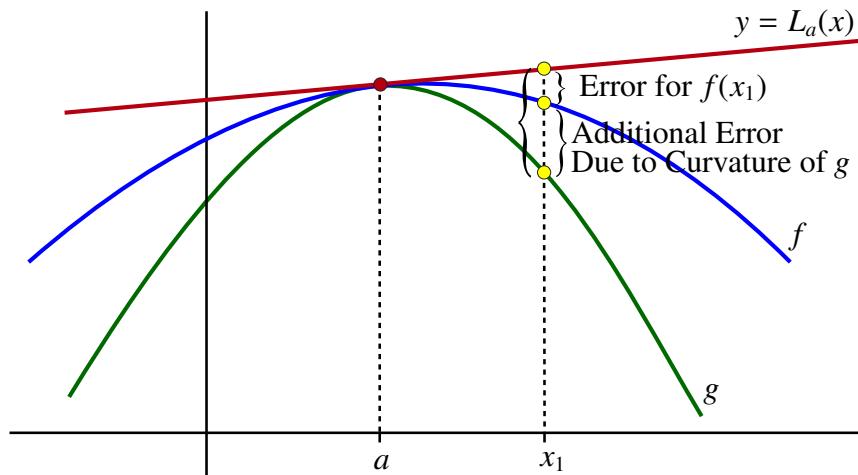


As this diagram illustrates, in principle, as the distance from x to a increases, so does the *potential* error in using $L_a(x)$ to approximate $f(x)$. However, it is not always true that the value of the error gets larger as the distance from x to a increases. Consider the following diagram:



In fact, at x_3 the linear approximation gives us the exact value of the function (no error). Situations such as this are uncommon. Generally speaking, the closer we are to a , the more confidence we will have in the accuracy of the estimate.

Second Observation: There is a second factor that affects the size of the potential error. Since we are using a line to approximate the function, the more *curved* the graph is near $x = a$, the greater the potential error. This situation is illustrated in the next diagram.

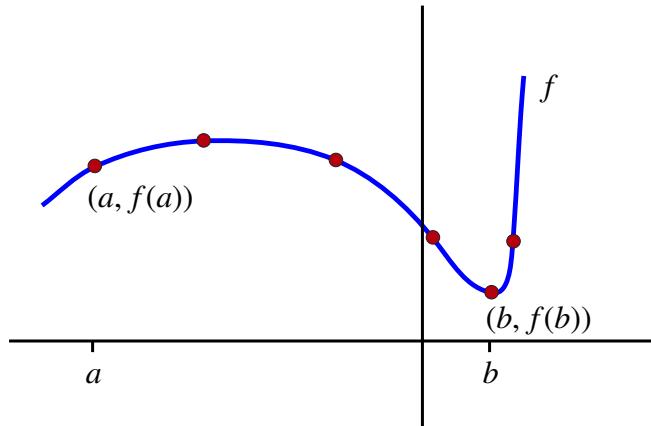


Question: How can we quantify the phrase “the more curved the graph is?”

Curvature arises from a change in the slope of the tangent lines. The more quickly these slopes change, the more curved the graph. (Note: Since the slope of the tangent line to a linear function never changes, lines have zero curvature.)

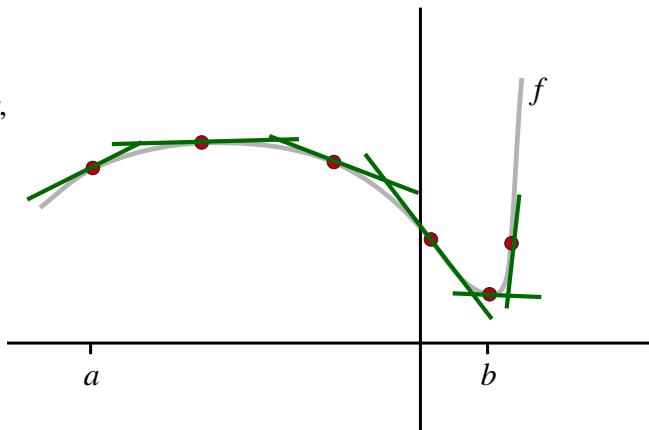
Consider the following graph of the function f .

Notice the two points labeled a and b . Near $(a, f(a))$ the graph curves slowly. Near $(b, f(b))$ the curvature is much more pronounced.



Now consider the tangent lines to the graph at selected points.

Near a the slopes of the tangent lines change very little as we move from left to right. However, near b , the slopes of the tangent lines change very quickly from steeply decreasing to flat, to steeply increasing. This rapid change in the slope is what is responsible for the significant curvature that is observed in the graph of f near $x = b$.



The slope of the tangent line is $f'(x)$. Hence, the rate at which $f'(x)$ is changing is given by $f''(x)$. It follows that the larger the magnitude of $f''(x)$ near $x = a$, the greater the curvature of the graph and the larger the potential error in using linear approximation. This means that the size of $|f''(x)|$ is the second factor affecting the potential size of the error in using linear approximation. This information leads us to the following theorem.

THEOREM 6

The Error in Linear Approximation

Assume that f is such that $|f''(x)| \leq M$ for each x in an interval I containing a point a . Then

$$|f(x) - L_a(x)| \leq \frac{M}{2}(x - a)^2$$

for each $x \in I$.

EXAMPLE 9

In our previous error estimate of $\sin(.01)$ we could have used the above theorem with $I = [0, .01]$ and $a = 0$. We know that if $f(x) = \sin(x)$, then $f''(x) = -\sin(x)$. We also know that on $I = [0, .01]$, the largest value for $|\sin(x)| = \sin(x)$ occurs at $x = .01$ and that

$$|-\sin(.01)| \leq .01$$

If we let $M = .01$, the Error in Linear Approximation Theorem tells us that

$$\begin{aligned} |\sin(.01) - L_a(.01)| &\leq \frac{M}{2}(.01 - 0)^2 \\ &= \frac{.01}{2}(.01)^2 \\ &= 5 \times 10^{-7}. \end{aligned}$$

In fact, our actual error was 1.7×10^{-7} which is less than 5×10^{-7} . ◀

6.5.2 Applications of Linear Approximation

There are many sophisticated applications of linear approximation. Some of these will be addressed later in this course. We will end this section with two simple but useful applications of linear approximation.

Application 1: Estimating Change

Assume that we know the value of $f(x)$ at a point a . We want to know what change we could expect in the value of $f(x)$ if we move to a point x_1 near a . That is, we want to know what

$$\Delta y = f(x_1) - f(a)$$

will be if we change the variable by

$$\Delta x = x_1 - a$$

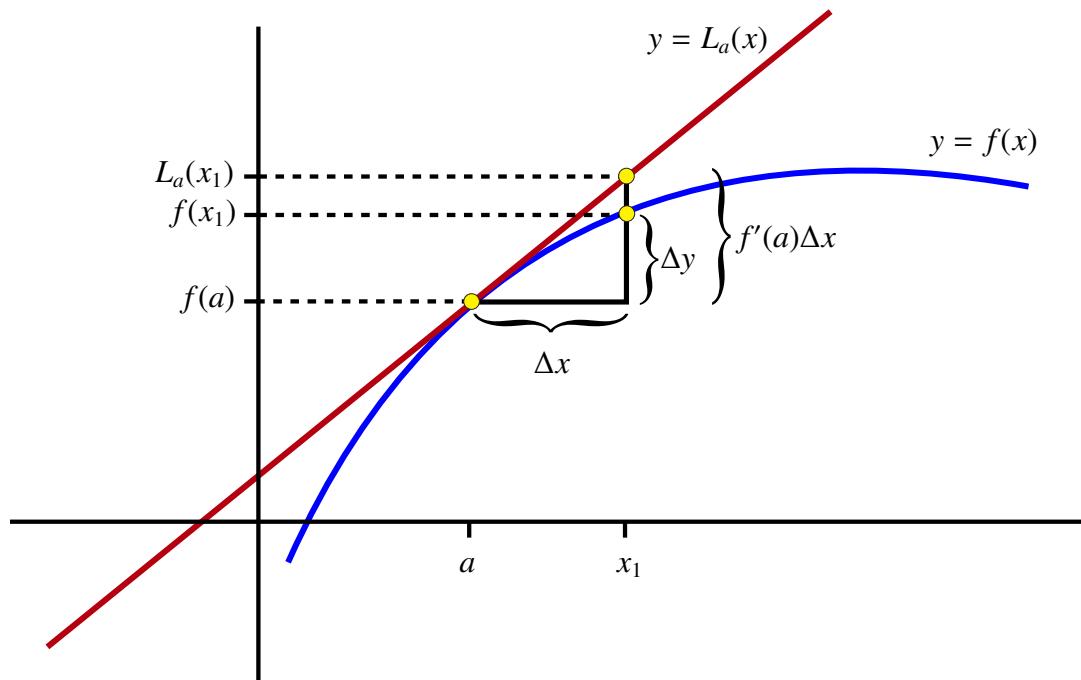
units. If we were to use the linear approximation L_a , we would find that

$$\begin{aligned}\Delta y &= f(x_1) - f(a) \\ &\cong L_a(x_1) - f(a) \\ &= (f(a) + f'(a)(x_1 - a)) - f(a) \\ &= f'(a)(x_1 - a) \\ &= f'(a)\Delta x.\end{aligned}$$

That is

$$\Delta y \cong f'(a)\Delta x.$$

This last statement is illustrated by the following diagram.



The next example uses the method we have just derived for estimating the change in a quantity.

EXAMPLE 10

A metal sphere of radius 10 cm expands when heated so that its radius increases by 0.01 cm. Estimate the change in the volume of the sphere.

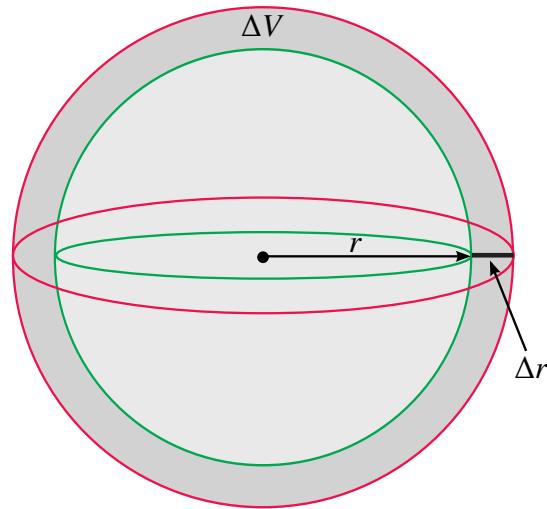
We know that the volume (V) of the sphere with radius r is given by

$$V = V(r) = \frac{4}{3}\pi r^3$$

and that $V'(r) = 4\pi r^2$. Our focal point is at $r = 10$ cm, so

$$V'(10) = 400\pi.$$

We also know that $\Delta r = .01$.



We want to know

$$\Delta V = V(10.01) - V(10)$$

If we use linear approximation our estimate becomes

$$\begin{aligned}\Delta V &= V(10.01) - V(10) \\ &\cong V'(10)\Delta r \\ &= 400\pi(.01) \\ &= 4\pi\end{aligned}$$

This means we should expect the volume to change by approximately 4π cm³. ◀

Application 2: Qualitative Analysis of Functions

The second application that we present in this section is an application of linear approximation to qualitative analysis of functions. In this case, our problem will be to study the behavior of the function

$$y = e^{-x^2}$$

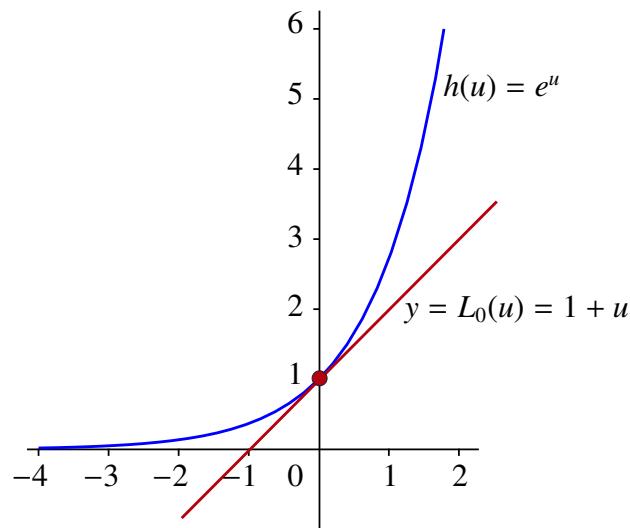
near $x = 0$. (This is a function that plays an important role in probability theory and statistics.)

Our first step is to begin with a simpler function, e^u . If $h(u) = e^u$, then we know that $h'(u) = e^u$ so

$$h(0) = h'(0) = e^0 = 1.$$

It follows that $y = 1 + u$ is the tangent line to $h(u) = e^u$ through $(0, 1)$. Linear approximation tells us that if u is near 0, then

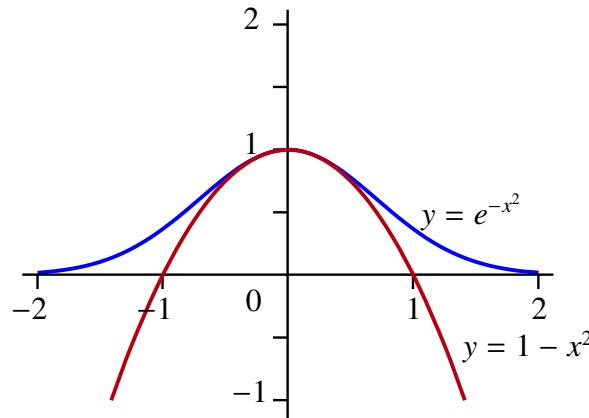
$$e^u \cong 1 + u.$$



However, if x is close to 0, then $-x^2$ is very close to 0. If we let $u = -x^2$, we get

$$y = e^{-x^2} \cong 1 + (-x^2) = 1 - x^2.$$

The next diagram illustrates the graphs of both $y = e^{-x^2}$ and $y = 1 - x^2$.



You can see that if x is close to 0, then $y = e^{-x^2}$ behaves like the much simpler function $y = 1 - x^2$. Consequently, if we were asked to sketch the graph of $y = e^{-x^2}$ near $x = 0$ we could simply draw the parabola associated with $y = 1 - x^2$.

REMARK

Despite the fact that we obtained the previous estimate by using linear approximation, the function $y = 1 - x^2$ is **not** the linear approximation to e^{-x^2} at $x = 0$. The simplest way to see this is to note that the graph of $y = 1 - x^2$ is a parabola and not a line. We will soon see that if $g(x) = e^{-x^2}$, then $g'(x) = -2xe^{-x^2}$ (see *The Chain Rule*). It follows that $g(0) = 1$ and $g'(0) = 0$, so the linear approximation to e^{-x^2} at $x = 0$ is the constant function $y = 1$. In fact, $y = 1 - x^2$ is the second degree analog of the linear approximation called the *second degree Taylor Polynomial*. We will study Taylor polynomials later in the course. ◀

6.6 Newton's Method

In the previous section we introduced the notion of the linear approximation to a differentiable function and studied some simple applications. In this section, we present a much more profound application called *Newton's Method*.

Recall that in the section on continuity we made use of the Intermediate Value Theorem to develop a bisection algorithm for approximating the solution to an equation of the type

$$f(x) = 0.$$

Newton's Method is another such algorithm, but it is in most cases much more efficient than the bisection technique. To see how this method works, we begin with the following simple case.

Assume that $f(x) = f(a) + m(x - a)$. Then f is a linear function whose graph passes through the point $(a, f(a))$. Suppose we wanted to find a point c such that $f(c) = 0$. In this case, provided that $m \neq 0$, there is no need to estimate c since we can calculate it explicitly.

We have

$$0 = f(c) = f(a) + m(c - a).$$

This implies that

$$-f(a) = m(c - a).$$

If $m \neq 0$, we can divide both sides of the equation by m to get

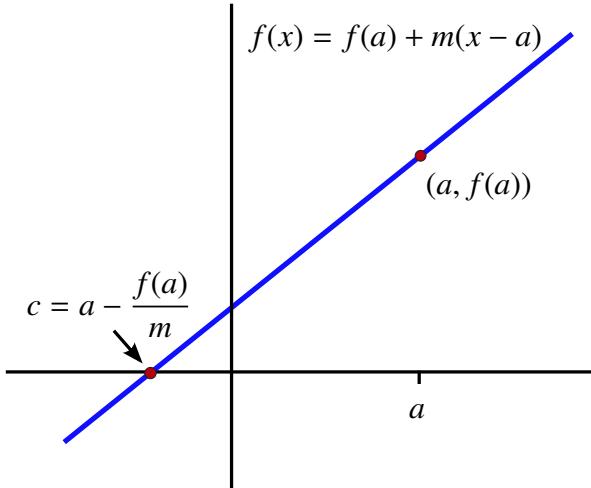
$$\frac{-f(a)}{m} = c - a.$$

Finally, adding a to both sides of the equation yields

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

Therefore, if $f(x) = f(a) + m(x - a)$ and $m \neq 0$, we can easily solve the equation

$$f(x) = 0.$$



(If $m = 0$, the graph of f is a horizontal line so if $f(a) \neq 0$, the graph does not cross the x -axis and no such c exists.)

What do we do if the graph of f is not a line? It is in this case where we can use linear approximation. The following steps outline a general method to find the linear approximation of a differentiable function.

Newton's Method [Steps]

Step 1: Pick a point x_1 that is reasonably close to a point c with $f(c) = 0$. (The IVT might be helpful in finding such an x_1 .)

Step 2: If f is differentiable at $x = x_1$, then we have seen that we can approximate f near x_1 by using its linear approximation $L_{x_1}(x) = f(x_1) + f'(x_1)(x - x_1)$. Since

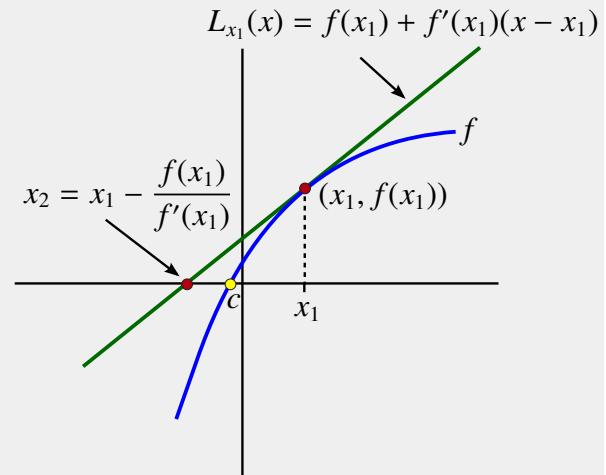
$$f(x) \cong L_{x_1}(x)$$

it would make sense that the graphs of f and L_{x_1} would cross the x -axis at roughly the same place. Therefore, if $f'(x_1) \neq 0$, we can approximate c by x_2 , where x_2 is such that

$$L_{x_1}(x_2) = 0.$$

But we have already seen that

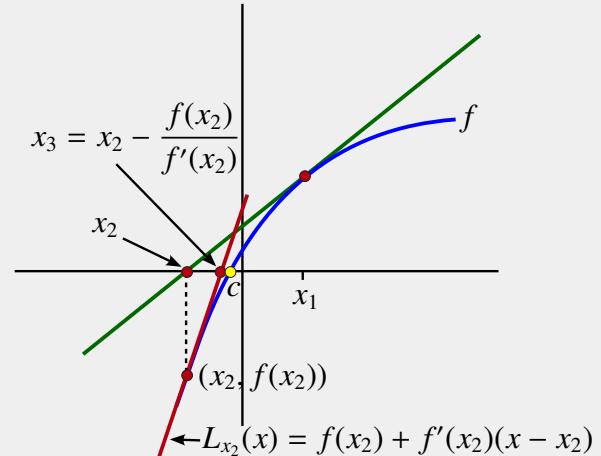
$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}.$$



Step 3: We now repeat the procedure, replacing x_1 by x_2 and using the linear approximation at x_2 , to get a new approximation

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}$$

for c .



The diagram shows that in this example x_3 is very close to c .

Continuing in this manner, we get a recursively defined sequence

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

where x_{n+1} is simply the point at which the tangent line to the graph of f through $(x_n, f(x_n))$ crosses the x -axis.

It can be shown that for most nice functions and reasonable choices of x_1 , the sequence $\{x_n\}$ converges very rapidly to a number c with $f(c) = 0$. However, it makes sense for us to ask: How accurate is the approximation?

Accuracy of Newton's Method

If we want to approximate c to k decimal places of accuracy, k decimal places must be carried throughout the calculations. The procedure stops when two successive terms, x_n and x_{n+1} , agree to k many decimal places. In most cases, this will happen after only a few iterations because each iteration usually doubles the number of decimal places of accuracy. Indeed, Newton's Method is much more efficient than the previous bisection method for finding approximate solutions to equations since the bisection method requires roughly 4 iterations to improve the accuracy of the estimate by just 1 decimal place.

EXAMPLE 11 Heron's Method Revisited

Use Newton's Method to estimate $\sqrt{2}$ to nine decimal places of accuracy.

In this case, to use Newton's Method we consider the function

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

The two solutions to the equation

$$f(x) = x^2 - 2 = 0$$

are $x = \sqrt{2}$ and $x = -\sqrt{2}$. Since $f(\sqrt{2}) = 0$, we can choose a point x_1 near $\sqrt{2}$, and then apply Newton's Method to f to estimate $\sqrt{2}$. In this case, we will begin at $x_1 = 1$.

Now since $f(x) = x^2 - 2$, we have $f'(x) = 2x$. The iterative sequence becomes $x_1 = 1$ and

$$\begin{aligned} x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \\ &= x_n - \frac{(x_n)^2 - 2}{2x_n} \\ &= \frac{2x_n^2}{2x_n} - \frac{(x_n)^2 - 2}{2x_n} \\ &= \frac{x_n^2 + 2}{2x_n} \\ &= \frac{1}{2}(x_n + \frac{2}{x_n}). \end{aligned}$$

That is,

$$x_{n+1} = \frac{1}{2}(x_n + \frac{2}{x_n}).$$

This last formula should be familiar. It is in fact the formula for generating the iterated sequence used to approximate $\sqrt{2}$ that we referred to as Heron's algorithm (or the Babylonian Square Root Method).

In fact, if we replace $f(x) = x^2 - 2$ by $f(x) = x^2 - \alpha$, then applying Newton's Method would generate the recursively defined sequence

$$x_{n+1} = \frac{1}{2}(x_n + \frac{\alpha}{x_n})$$

to estimate $\sqrt{\alpha}$ for any $\alpha > 0$.

With this formula we can calculate the successive approximations. Using $x_1 = 1$, we have

$$\begin{aligned} x_2 &= x_{1+1} \\ &= \frac{1}{2}(1 + \frac{2}{1}) \\ &= \frac{1}{2}(3) \\ &= \frac{3}{2} \\ &= 1.5 \end{aligned}$$

Next we have

$$\begin{aligned} x_3 &= x_{2+1} \\ &= \frac{1}{2}(\frac{3}{2} + \frac{2}{\frac{3}{2}}) \\ &= \frac{17}{12} \\ &= 1.416666667 \end{aligned}$$

We then get

$$\begin{aligned} x_4 &= x_{3+1} \\ &= \frac{1}{2}(\frac{17}{12} + \frac{2}{\frac{17}{12}}) \\ &= \frac{577}{408} \\ &= 1.414215686 \end{aligned}$$

The next iteration gives us

$$\begin{aligned} x_5 &= x_{4+1} \\ &= \frac{1}{2}(\frac{577}{408} + \frac{2}{\frac{577}{408}}) \\ &= \frac{665857}{470832} \\ &= 1.414213562 \end{aligned}$$

At this stage, the last two approximations agree to five decimal places. We should expect that the next iteration may very well agree with the previous estimate to all nine decimal places. In fact, we have

$$\begin{aligned} x_6 &= x_{5+1} \\ &= \frac{1}{2} \left(\frac{665857}{470832} + \frac{2}{\frac{665857}{470832}} \right) \\ &= 1.414213562 \end{aligned}$$

exactly as we expected. This means that Newton's Method gives us an estimate that

$$\sqrt{2} \approx 1.414213562$$

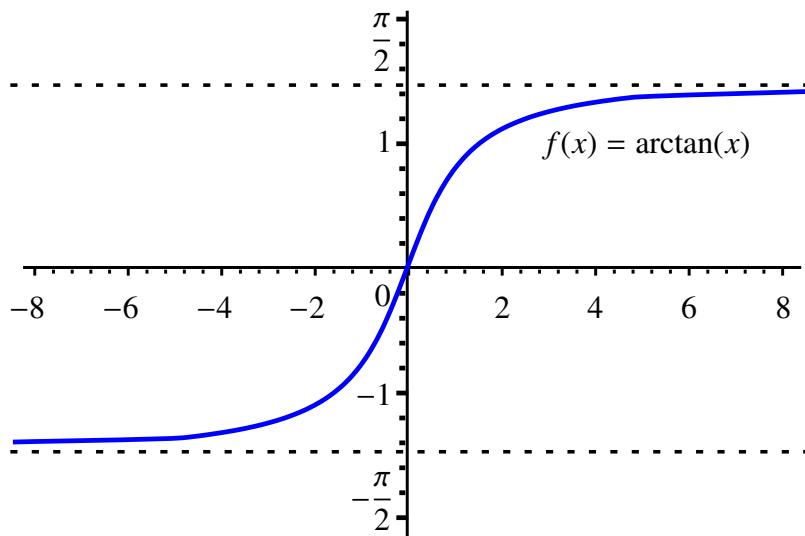
and this estimate is accurate to nine decimal places. ◀

EXAMPLE 12 Failure of Newton's Method

You will recall that given a continuous function f on $[a, b]$, if $f(a)$ and $f(b)$ are of opposite signs, then the IVT ensures that there must be a $c \in (a, b)$ for which $f(c) = 0$. Moreover, it gave us an algorithm to find such a c within an error that could be made as small as we wish. The problem with this algorithm is that while *it always works*, it can be a rather slow process.

In contrast, if the function f is known to be differentiable, and if we were to apply Newton's Method to approximate c , typically we can find an extremely good approximation with only a few iterations of this method. Recall that we expect the number of decimal places of accuracy to double with every iteration!

However, unlike the IVT based algorithm, Newton's Method can fail to find c if we are unlucky, even if we know it exists. It turns out that the most problematic situation occurs when we consider points where the tangent line is very *flat*. For example, consider the function $f(x) = \arctan(x)$.



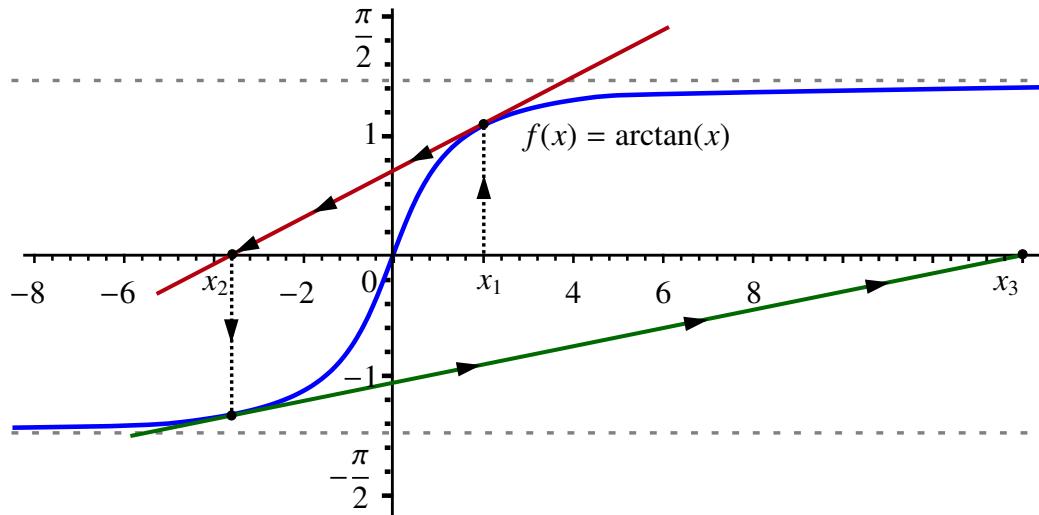
We know that

$$\arctan(x) = 0 \quad \text{if and only if } x = 0.$$

Notice that as x grows in magnitude, the graph of f becomes much flatter. This can also be seen by looking at the derivative

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

which approaches 0 as x grows. If we choose a point x_1 as in the following diagram and apply the iterative procedure, then we see that $|x_2| > |x_1|$ and that the point x_3 is much farther away from 0 than either x_1 or x_2 .



In fact, in the case of $\arctan(x)$ it can be shown that if we choose any starting point x_1 with $|x_1| > \frac{\pi}{4}$, then Newton's Method will *fail* with the iterates (i.e., points x_n) growing without bound.



6.7 Arithmetic Rules of Differentiation

In this section, we review the rules of differentiation that you learned in your high school Calculus class.

THEOREM 7**The Arithmetic Rules for Differentiation**

Assume that f and g are both differentiable at $x = a$.

1) The Constant Multiple Rule:

Let $h(x) = cf(x)$. Then h is differentiable at $x = a$ and

$$h'(a) = cf'(a).$$

2) The Sum Rule:

Let $h(x) = f(x) + g(x)$. Then h is differentiable at $x = a$ and

$$h'(a) = f'(a) + g'(a).$$

3) The Product Rule:

Let $h(x) = f(x)g(x)$. Then h is differentiable at $x = a$ and

$$h'(a) = f'(a)g(a) + f(a)g'(a).$$

4) The Reciprocal Rule:

Let $h(x) = \frac{1}{g(x)}$. If $g(a) \neq 0$, then h is differentiable at $x = a$ and

$$h'(a) = \frac{-g'(a)}{(g(a))^2}.$$

5) The Quotient Rule:

Let $h(x) = \frac{f(x)}{g(x)}$. If $g(a) \neq 0$, then h is differentiable at $x = a$ and

$$h'(a) = \frac{f'(a)g(a) - f(a)g'(a)}{(g(a))^2}.$$

We now present proofs of the Arithmetic Rules.

1) Proof of the Constant Multiple Rule:

Assume that $c \in \mathbb{R}$ and that f is differentiable at $x = a$. Then

$$\begin{aligned} (cf)'(a) &= \lim_{h \rightarrow 0} \frac{(cf)(a+h) - (cf)(a)}{h} \\ &= \lim_{h \rightarrow 0} \frac{cf(a+h) - cf(a)}{h} \\ &= c \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \\ &= cf'(a). \end{aligned}$$

■

2) **Proof of the Sum Rule:**

Assume that f and g are differentiable at $x = a$. Then

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

■

3) **Proof of the Product Rule:**

The justification for the Product Rule is a bit more complicated than for the previous two rules. It requires the following trick:

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

To evaluate the last two limits, we need to remember that since f is differentiable at $x = a$, it is also continuous. This means that $\lim_{h \rightarrow 0} f(a + h) = f(a)$. From this it follows that

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

The second limit is more straight forward since we can factor out the constant $g(a)$ to get

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

This gives us that

$$\begin{aligned}
 (fg)'(a) &= \lim_{h \rightarrow 0} \frac{f(a + h)(g(a + h) - g(a))}{h} + \lim_{h \rightarrow 0} \frac{g(a)(f(a + h) - f(a))}{h} \\
 &= f(a)g'(a) + f'(a)g(a).
 \end{aligned}$$

exactly as stated. ■

4) **Proof of the Reciprocal Rule:**

Assume that f is differentiable at $x = a$. Then

$$\begin{aligned}
 \left(\frac{1}{f}\right)'(a) &= \lim_{h \rightarrow 0} \frac{\frac{1}{f(a+h)} - \frac{1}{f(a)}}{h} \\
 &= \lim_{h \rightarrow 0} \frac{f(a) - f(a+h)}{f(a+h)f(a)h} \\
 &= -\lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \cdot \lim_{h \rightarrow 0} \frac{1}{f(a+h)f(a)} \\
 &= -f'(a) \cdot \frac{1}{(f(a))^2} \quad (\text{by continuity at } x = a) \\
 &= \frac{-f'(a)}{(f(a))^2}.
 \end{aligned}$$

■

5) **Proof of the Quotient Rule:**

The proof of the Quotient Rule is a combination of the Product Rule and the Reciprocal Rule. This proof is left as an exercise.

So far, we have seen that $\frac{d}{dx}(x) = 1$ and that $\frac{d}{dx}(x^2) = 2x$. It is not too difficult to show that if $n \in \mathbb{N}$, then

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

These can all be considered special cases of the next important rule.

THEOREM 8**The Power Rule for Differentiation**

Assume that $\alpha \in \mathbb{R}$, $\alpha \neq 0$, and $f(x) = x^\alpha$. Then f is differentiable and

$$f'(x) = \alpha x^{\alpha-1}$$

wherever $x^{\alpha-1}$ is defined.

NOTE

In the case where $\alpha \in \mathbb{N}$, the Power Rule can be derived from the Binomial Theorem. For $\alpha \in \mathbb{Q}$, the Power Rule can be obtained by using the *Chain Rule* and the *Inverse Function Theorem* (both of which will be discussed later in the course), and if necessary the Reciprocal Rule. Establishing differentiability in the case where α is irrational is beyond the scope of this course. However, if we assume differentiability, the Power Rule can be established using a technique known as *logarithmic differentiation*.

◀

EXAMPLE 13 Differentiating Polynomials and Rational Functions

Let $P(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ be a polynomial. Then the rules of differentiation can be used to show that P is always differentiable and that

$$P'(x) = a_1 + 2a_2x + 3a_3x^2 + \cdots + na_nx^{n-1}.$$

Using the Quotient Rule, we see that a rational function

$$R(x) = \frac{P(x)}{Q(x)}$$

is differentiable at any point where $Q(x) \neq 0$.

In particular, if

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

then R is differentiable provided that $x^2 - 1 \neq 0$. That is, when $x \neq \pm 1$. Moreover, the Quotient Rule shows that

$$\begin{aligned} R'(x) &= \frac{\left(\frac{d}{dx}(x+2)\right)(x^2-1) - (x+2)\left(\frac{d}{dx}(x^2-1)\right)}{(x^2-1)^2} \\ &= \frac{1 \cdot (x^2-1) - (x+2)(2x)}{(x^2-1)^2} \\ &= \frac{(x^2-1) - 2x^2 - 4x}{(x^2-1)^2} \\ &= \frac{-x^2 - 4x - 1}{(x^2-1)^2} \end{aligned}$$



6.8 The Chain Rule

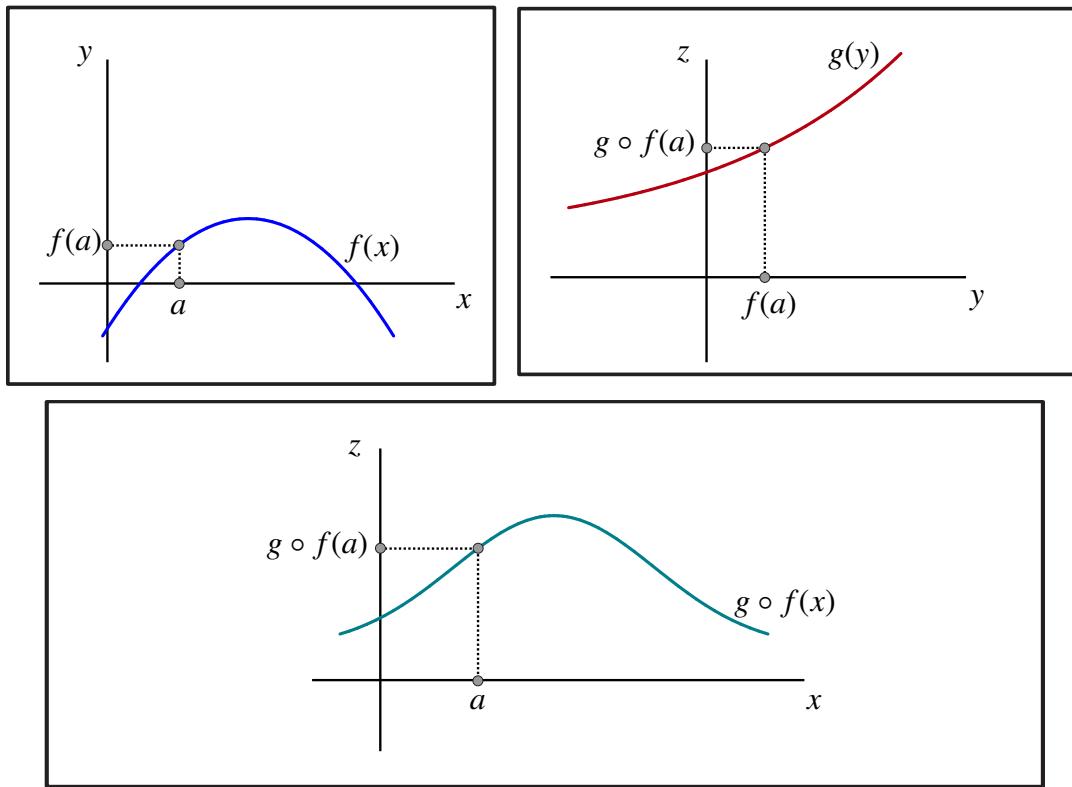
So far we have looked at various rules of differentiation. However, one of the most powerful rules of differentiation, the Chain Rule, shows us how to differentiate compositions of differentiable functions. In this section, we will use linear approximation to give a geometric derivation of this important rule.

Geometric Derivation of the Chain Rule

Preconditions: Suppose that we have two functions $y = f(x)$ and $z = g(y)$. Let

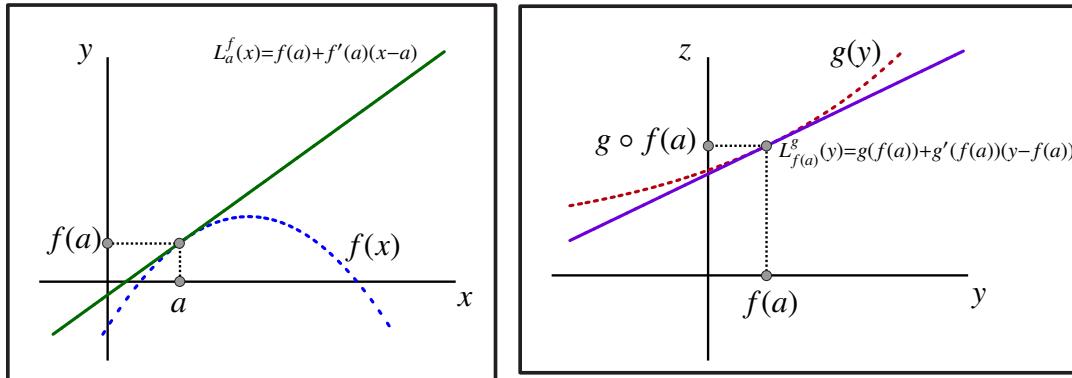
$$h(x) = g \circ f(x) = g(f(x))$$

be the composition function.



Assume now that $f(x)$ is differentiable at $x = a$ and $g(y)$ is differentiable at $y = f(a)$. We want to know if the composition function $h(x)$ will also be differentiable at $x = a$, and if so its derivative.

Geometric Derivation: Proving that the composition function is differentiable is a little tricky, but if we assume it is differentiable, we can use what we learned about linear approximation to derive its derivative. We will do this by building the linear approximation function $L_a^h(x)$ for our composition. To do so let's assume that we knew nothing about the functions $f(x)$ and $g(y)$ other than the values of $f(a)$, $f'(a)$, $g(f(a))$ and $g'(f(a))$. First recall that since $f(x)$ is differentiable at $x = a$ and $g(y)$ is differentiable at $y = f(a)$ we can approximate $f(x)$ near $x = a$ by $L_a^f(x)$ and we can approximate $g(y)$ near $y = f(a)$ by $L_{f(a)}^g(y)$.



Then since $f(x) \cong L_a^f(x)$ near $x = a$ and $g(y) \cong L_{f(a)}^g(y)$ near $y = f(a)$, we would hope that we can approximate the composition $g \circ f$ by composing the two linear

approximations. That is, we should have

$$h(x) = g \circ f(x) \cong L_{f(a)}^g \circ L_a^f(x)$$

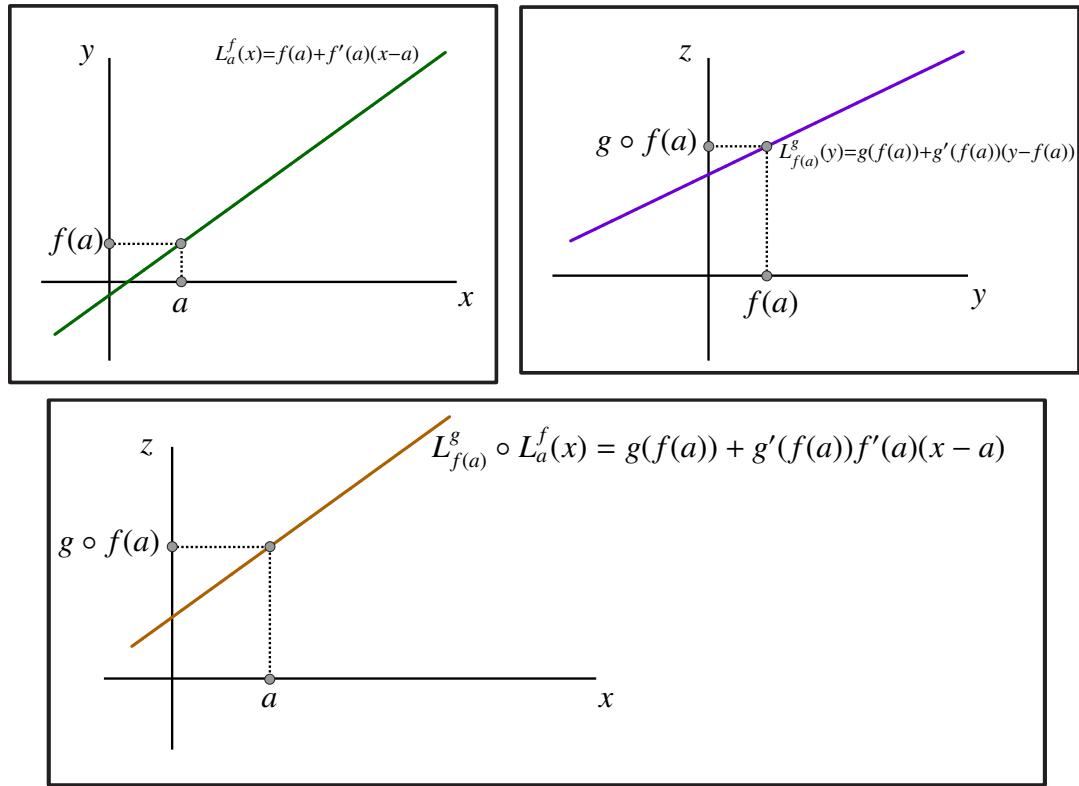
near $x = a$.

Moreover,

$$\begin{aligned} L_{f(a)}^g \circ L_a^f(x) &= L_{f(a)}^g(L_a^f(x)) \\ &= g(f(a)) + g'(f(a))(L_a^f(x) - f(a)) \\ &= g(f(a)) + g'(f(a))((f(a) + f'(a)(x - a)) - f(a)) \\ &= g(f(a)) + g'(f(a))f'(a)(x - a) \end{aligned}$$

Then we get that the composition of the two linear approximations yields another function whose graph is a line with equation

$$z = g(f(a)) + g'(f(a))f'(a)(x - a).$$



At this point we would have

$$h(x) \cong g(f(a)) + g'(f(a))f'(a)(x - a).$$

Question: Is

$$z = g(f(a)) + g'(f(a))f'(a)(x - a)$$

the tangent line to the graph of $z = h(x)$ through $(a, h(a))$? In other words, is

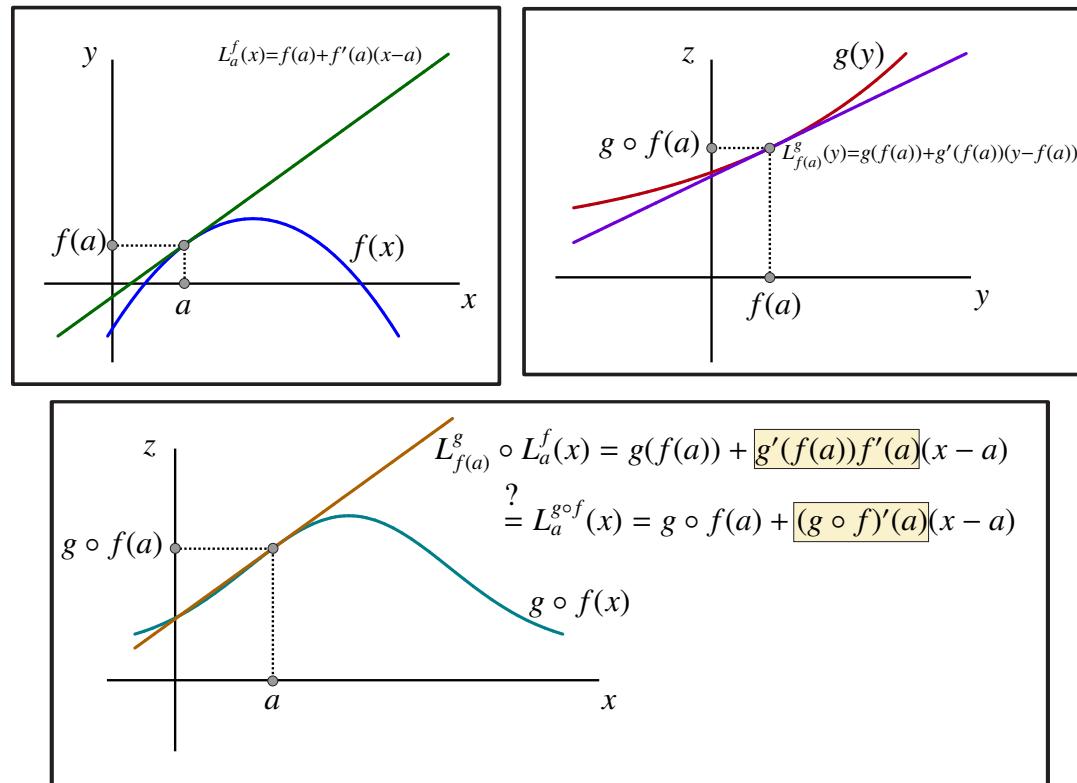
$$L_a^h(x) = g(f(a)) + g'(f(a))f'(a)(x - a)?$$

We do have

$$h(a) = g \circ f(a) = g(f(a))$$

so that all we would need for this to be the new linear approximation is that

$$(g \circ f)'(a) = g'(f(a)) \cdot f'(a).$$



One of the most profound results in Calculus is that the previous argument does indeed prove true. This is precisely the Chain Rule.

THEOREM 9

The Chain Rule

Assume that $y = f(x)$ is differentiable at $x = a$ and $z = g(y)$ is differentiable at $y = f(a)$. Then $h(x) = g \circ f(x) = g(f(x))$ is differentiable at $x = a$ and

$$h'(a) = g'(f(a))f'(a).$$

In particular,

$$L_a^h(x) = L_{f(a)}^g \circ L_a^f(x).$$

It turns out that Leibniz's notation is very convenient for describing the Chain Rule. Suppose that we have

$$z = g(y) \text{ and } y = f(x).$$

Then

$$\frac{dz}{dy} = g'(y) \text{ and } \frac{dy}{dx} = f'(x).$$

But

$$z = g(y) = g(f(x))$$

so the Chain Rule shows that

$$\begin{aligned}\frac{dz}{dx} &= g'(f(x))f'(x) \\ &= \frac{dz}{dy} \Big|_{f(x)} \frac{dy}{dx} \Big|_x\end{aligned}$$

Therefore, in Leibniz's notation, the Chain Rule simply becomes

$$\frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx}.$$

EXAMPLE 14 Find

$$\frac{d}{dx}(x^2 + 1)^3.$$

SOLUTION #1: Let $f(x) = x^2 + 1$ and $g(y) = y^3$. Let $h(x) = g(f(x))$. Then

$$\frac{d}{dx}(x^2 + 1)^3 = h'(x).$$

We also know that $f'(x) = 2x$ and $g'(y) = 3y^2$. From the Chain Rule we get that

$$\begin{aligned}h'(x) &= g'(f(x))f'(x) \\ &= 3(f(x))^2(2x) \\ &= 3(x^2 + 1)^2(2x) \\ &= 6x(x^2 + 1)^2.\end{aligned}$$

SOLUTION #2: Let $z = y^3$ and $y = x^2 + 1$. Then

$$z = y^3 = (x^2 + 1)^3$$

so that

$$\frac{d}{dx}(x^2 + 1)^3 = \frac{dz}{dx}.$$

The Chain Rule says that

$$\frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx}.$$

But

$$\frac{dz}{dy} = 3y^2 \text{ and } \frac{dy}{dx} = 2x.$$

From this it follows that

$$\begin{aligned} \frac{dz}{dx} &= \frac{dz}{dy} \frac{dy}{dx} \\ &= 3y^2(2x) \\ &= 3(x^2 + 1)^2(2x) \\ &= 6x(x^2 + 1)^2. \end{aligned}$$



6.8.1 Proof of the Chain Rule

So far we have given an important geometric interpretation of the Chain Rule. In this section, we will give a proof for this fundamental result. In fact, we will prove the following slightly upgraded version of the Chain Rule.

THEOREM 10

The Chain Rule: Upgraded Version

Assume that $f : I \rightarrow \mathbb{R}$, where $I \subseteq \mathbb{R}$, and that $g : J \rightarrow \mathbb{R}$, where $f(I) \subseteq J$ and I and J are open intervals such that I contains some $x = a$ and J contains $f(a)$. If $f(x)$ is differentiable at $x = a$ and $g(y)$ is differentiable at $y = f(a)$, then $h(x) := (g \circ f)(x)$ is differentiable at $x = a$ with $h'(a) = g'(f(a))f'(a)$.

PROOF

Let $\phi : J \rightarrow \mathbb{R}$ be defined by

$$\phi(y) = \begin{cases} \frac{g(y) - g(f(a))}{y - f(a)} & \text{if } y \neq f(a), \\ g'(f(a)) & \text{if } y = f(a). \end{cases}$$

Note that $f(a) \in J$, and so

$$\lim_{y \rightarrow f(a)} \phi(y) = \lim_{y \rightarrow f(a)} \frac{g(y) - g(f(a))}{y - f(a)} := g'(f(a)).$$

and $\phi(y)$ is continuous at $y = f(a)$.

Now we note that for all $y \in J$,

$$g(y) - g(f(a)) = \phi(y)[y - f(a)],$$

even when $y = f(a)$. Hence

$$g(f(x)) - g(f(a)) = \phi(f(x))[f(x) - f(a)]$$

for all $x \in I$, since $f(I) \subset J$. We get

$$\begin{aligned} \lim_{x \rightarrow a} \frac{g(f(x)) - g(f(a))}{x - a} &= \lim_{x \rightarrow a} \frac{\phi(f(x))[f(x) - f(a)]}{x - a} \\ &= \lim_{x \rightarrow a} \phi(f(x)) \left[\frac{f(x) - f(a)}{x - a} \right] \\ &= \left(\lim_{x \rightarrow a} \phi(f(x)) \right) \cdot \left(\lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a} \right) \\ &= \phi(f(a)) \cdot \left(\lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a} \right) \\ &:= g'(f(a))f'(a), \end{aligned}$$

6.9 Derivatives of Other Trigonometric Functions

Earlier we made use of the Fundamental Trig Limit to show that $\frac{d}{dx}(\sin(x)) = \cos(x)$. We can now use the Rules of Differentiation to calculate the derivatives of all of the other basic trigonometric functions.

EXAMPLE 15 Find $\frac{d}{dx}(\cos(x))$.

SOLUTION We have already shown from first principles that $\frac{d}{dx}(\cos(x)) = -\sin(x)$. In this example we will derive this result from the Chain Rule. To do so we use the identity

$$\cos(x) = \sin\left(x + \frac{\pi}{2}\right).$$

Let $y = \sin(u)$ and $u = x + \frac{\pi}{2}$. Substituting for u gives us that

$$y = y(x) = \sin\left(x + \frac{\pi}{2}\right) = \cos(x).$$

Therefore,

$$\frac{d}{dx}(\cos(x)) = \frac{dy}{dx}.$$

However, by the Chain Rule

$$\begin{aligned} \frac{dy}{dx} &= \frac{dy}{du} \frac{du}{dx} \\ &= \cos(u) \cdot (1) \\ &= \cos\left(x + \frac{\pi}{2}\right). \end{aligned}$$

Finally, using the addition identity for cosine, we get

$$\begin{aligned}\cos(x + \frac{\pi}{2}) &= \cos(x)\cos(\frac{\pi}{2}) - \sin(x)\sin(\frac{\pi}{2}) \\ &= \cos(x) \cdot (0) - \sin(x) \cdot (1) \\ &= -\sin(x).\end{aligned}$$

Therefore, using the Chain Rule we have shown that

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



EXAMPLE 16 Find $\frac{d}{dx}(\tan(x))$.

SOLUTION We begin by writing

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

and then apply the Quotient Rule to get

$$\begin{aligned}\frac{d}{dx}(\tan(x)) &= \frac{d}{dx}\left(\frac{\sin(x)}{\cos(x)}\right) \\ &= \frac{(\frac{d}{dx}\sin(x))\cos(x) - (\sin(x))(\frac{d}{dx}\cos(x))}{\cos^2(x)} \\ &= \frac{\cos(x)\cos(x) - (\sin(x))(-\sin(x))}{\cos^2(x)} \\ &= \frac{\cos^2(x) + \sin^2(x)}{\cos^2(x)} \\ &= \frac{1}{\cos^2(x)} \\ &= \sec^2(x).\end{aligned}$$

We have shown that

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



A similar calculation shows that

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

EXAMPLE 17 Find $\frac{d}{dx}(\sec(x))$.

SOLUTION Recall that $\sec(x) = \frac{1}{\cos(x)}$. We can again apply the Quotient Rule with

$$f(x) = 1 \text{ and } g(x) = \cos(x)$$

to get

$$\begin{aligned} \frac{d}{dx}\left(\frac{1}{\cos(x)}\right) &= \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2} \\ &= \frac{0 \cdot \cos(x) - 1(-\sin(x))}{\cos^2(x)} \\ &= \frac{\sin(x)}{\cos^2(x)} \\ &= \frac{\sin(x)}{\cos(x)} \frac{1}{\cos(x)} \\ &= \tan(x) \sec(x) \end{aligned}$$

That is,

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

A similar calculation shows that

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

6.10 Derivatives of Inverse Functions

In this section, the relationship between the derivative of an invertible function and that of its inverse is explored using linear approximation as a key tool. More specifically, if we assume that $y = f(x)$ is invertible on the interval $[a, b]$ and that it is differentiable on (a, b) , we want to know when will the inverse function $f^{-1}(y) = g(y)$ also be differentiable and what is its derivative?

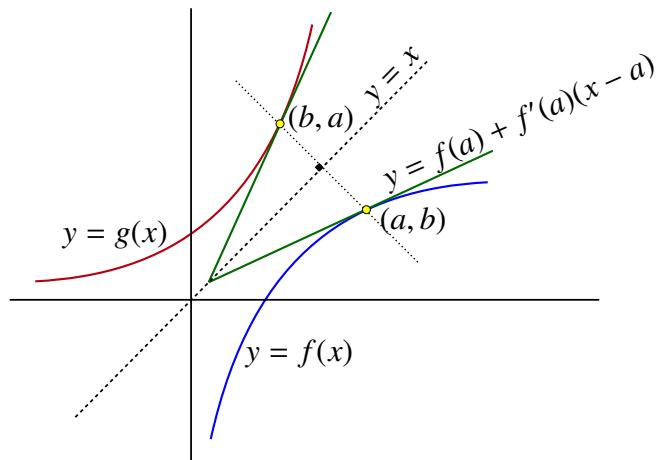
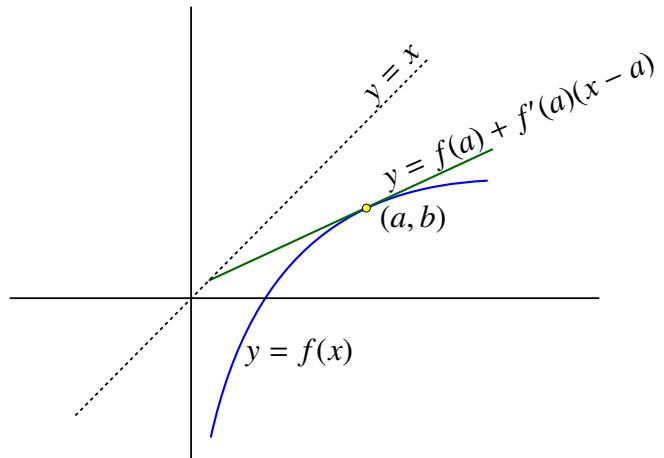
We will begin by looking at this problem geometrically.

Assume that we have a function $y = f(x)$ with an inverse $x = g(y)$. Assume also that $y = f(x)$ is differentiable at $x = a$ and that $f(a) = b$. This means that there is a tangent line to the graph of f through the point (a, b) . This line is actually the graph of the linear approximation

$$L_a^f(x) = f(a) + f'(a)(x - a).$$

Moreover, if $f'(a) \neq 0$, then $y = L_a^f(x)$ is also an invertible function.

We know that we can find the graph of the inverse function in its standard form $y = g(x)$ by reflecting the graph of f through the line $y = x$. If we also reflect the tangent line, the result is a new line that looks like a tangent line to the graph of the inverse function, except it passes through the point (b, a) .



The equation of the original tangent line is

$$y = f(a) + f'(a)(x - a).$$

To find the equation of the reflected line, take the tangent line equation and exchange the variables x and y to get

$$x = f(a) + f'(a)(y - a)$$

and then solve for y . To accomplish this, begin by subtracting $f(a)$ from both sides to get

$$x - f(a) = f'(a)(y - a).$$

Provided that $f'(a) \neq 0$, we can divide both sides by $f'(a)$. This gives us

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

Adding a to both sides yields

$$y = a + \frac{1}{f'(a)}(x - f(a)).$$

NOTE

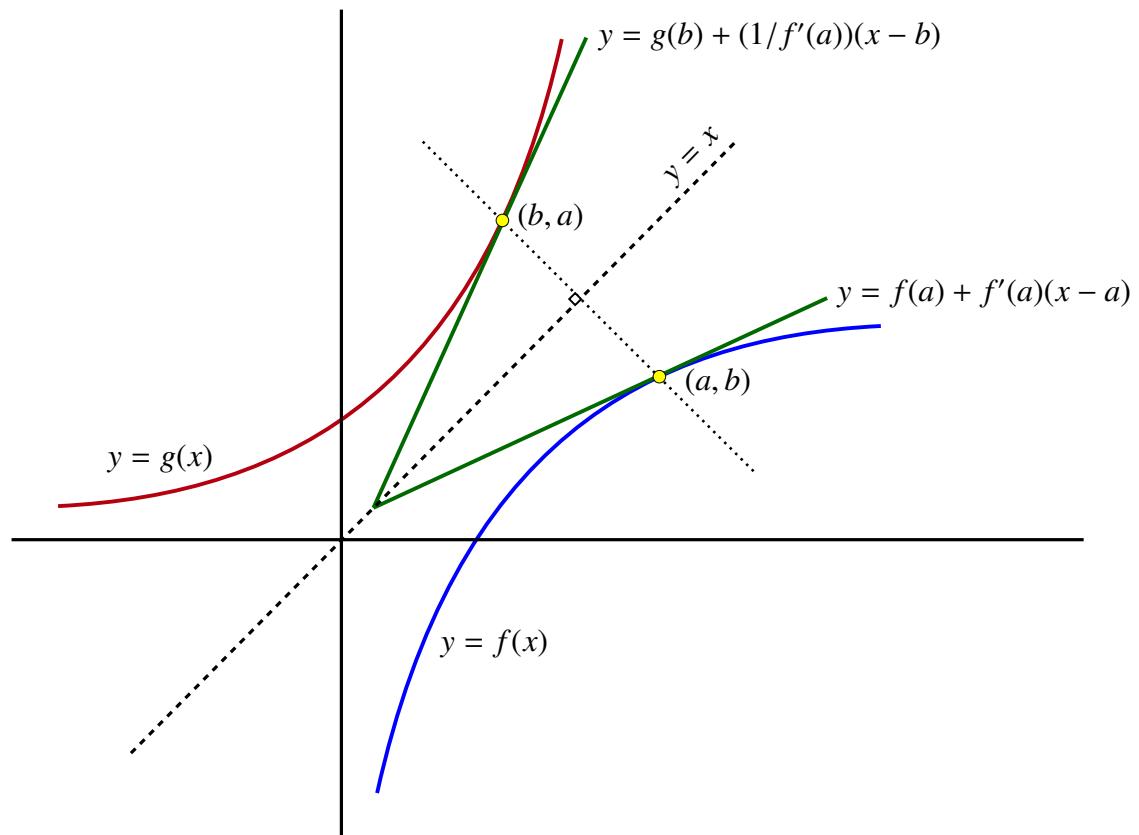
The procedure we used above is exactly how we would find the inverse of the function $y = L_a^f(x)$. That is,

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



The last step is to note that $a = g(b)$ and $b = f(a)$. Substituting these into the previous equation gives us

$$y = g(b) + \frac{1}{f'(a)}(x - b).$$

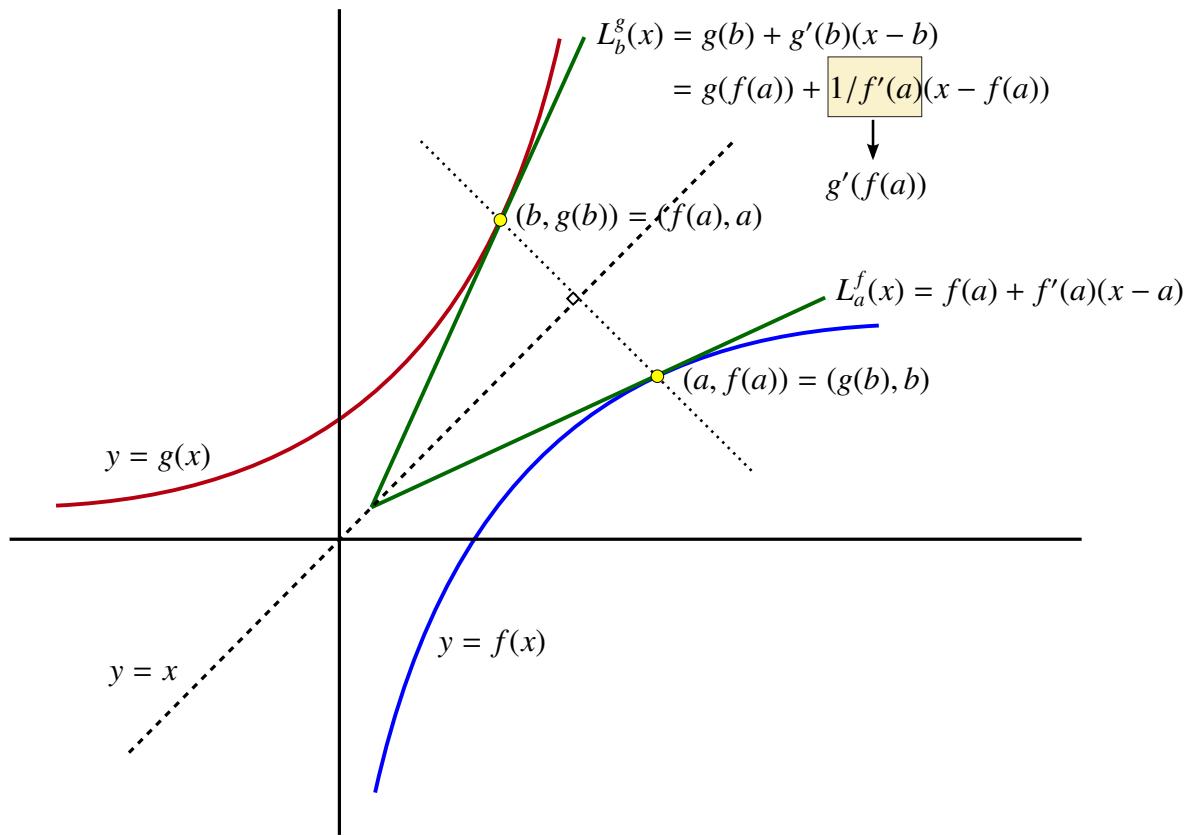


The picture suggests that this should be the equation of the tangent line to g through the point (b, a) . However, if g is differentiable at b , the equation of the tangent line would be

$$y = g(b) + g'(b)(x - b).$$

Comparing the last two equations shows that they agree provided that

$$g'(b) = g'(f(a)) = \frac{1}{f'(a)} = \frac{1}{f'(g(b))}.$$



It is worth noting again that this calculation can be done precisely when $f'(a) \neq 0$. We can summarize the previous discussion with the following important theorem.

THEOREM 11

The Inverse Function Theorem

Assume that $y = f(x)$ is continuous and invertible on $[c, d]$ with inverse $x = g(y)$, and f is differentiable at $a \in (c, d)$. If $f'(a) \neq 0$, then g is differentiable at $b = f(a)$, and

$$g'(b) = \frac{1}{f'(a)} = \frac{1}{f'(g(b))}.$$

Moreover, L_a^f is also invertible and

$$(L_a^f)^{-1}(x) = L_b^g(x) = L_{f(a)}^g(x).$$

EXAMPLE 18

Let $f(x) = x^3$. In this case, the inverse function is easy to calculate – it is $g(y) = y^{\frac{1}{3}}$.

To illustrate how the Inverse Function Theorem works, let $a = 2$. Then $f(a) = b = 2^3 = 8$. The Inverse Function Theorem states that

$$g'(b) = \frac{1}{f'(a)}.$$

Therefore, we expect that

$$g'(8) = \frac{1}{f'(2)}.$$

This can be easily verified. We note that $f'(x) = 3x^2$ so

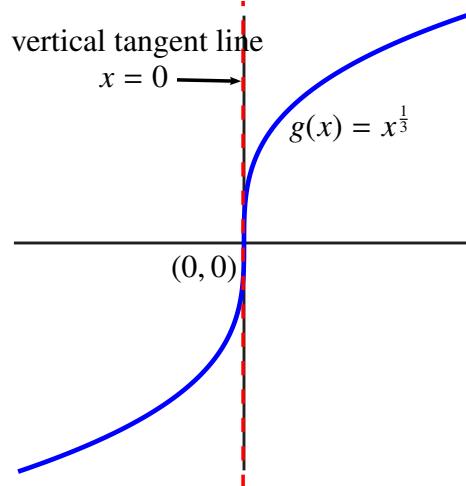
$$\frac{1}{f'(2)} = \frac{1}{12}.$$

We also know that $g'(y) = \frac{1}{3}y^{-\frac{2}{3}}$. It follows that

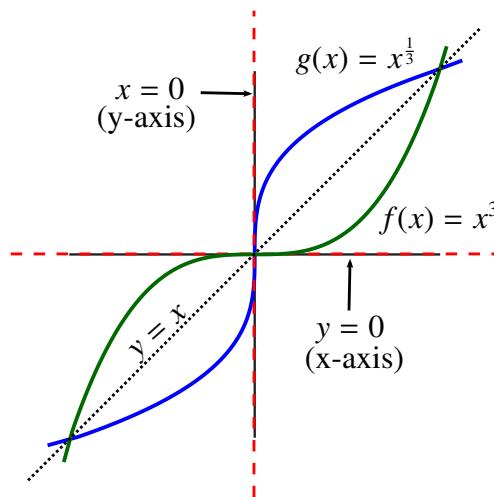
$$\begin{aligned} g'(8) &= \left(\frac{1}{3}\right)(8^{-\frac{2}{3}}) \\ &= \left(\frac{1}{3}\right)\left(\frac{1}{4}\right) \\ &= \frac{1}{12} \end{aligned}$$

exactly as expected. ◀

This example can also be used to show what happens when the assumption that $f'(a) \neq 0$ fails. If $a = 0$, then $b = f(a) = f(0) = 0$ as well. We have that $f'(a) = f'(0) = 0$. But $g'(y) = \frac{1}{3}y^{-\frac{2}{3}}$ is not defined at $b = f(0) = 0$. This can be seen from the graph of $g(x) = x^{\frac{1}{3}}$.



The graph of $g(x) = x^{\frac{1}{3}}$ has a “vertical tangent line” through $(0,0)$ which is not permitted. In fact, this “vertical tangent line” is simply the y -axis, or the line $x = 0$. Moreover, this line is the reflection through $y = x$ of the x -axis which is in turn the tangent line to $f(x) = x^3$ through $(0,0)$.



In summary, if the original function has a horizontal tangent line, that is if $f'(x) = 0$, then the reflection becomes a vertical line which is not permitted as the tangent line of a differentiable function. This explains why we do not allow $f'(x) = 0$ in the statement of the Inverse Function Theorem.

The previous example illustrates the Inverse Function Theorem, but it is artificial since we could just as easily have calculated the derivative of the inverse function directly. There is another way to view the Inverse Function Theorem that will turn out to be very useful.

Assume that we knew that $x = g(y)$ was the inverse of $y = f(x)$, and that both functions were differentiable. Then we know that

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

The Chain Rule shows that

$$\frac{d}{dx}(g(f(x))) = \frac{d}{dx}(x).$$

This means that

$$g'(f(x))f'(x) = 1$$

and hence that

$$g'(f(x)) = \frac{1}{f'(x)}$$

just as the Inverse Function Theorem suggested.

It is also useful to note that this equation can also be written as

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

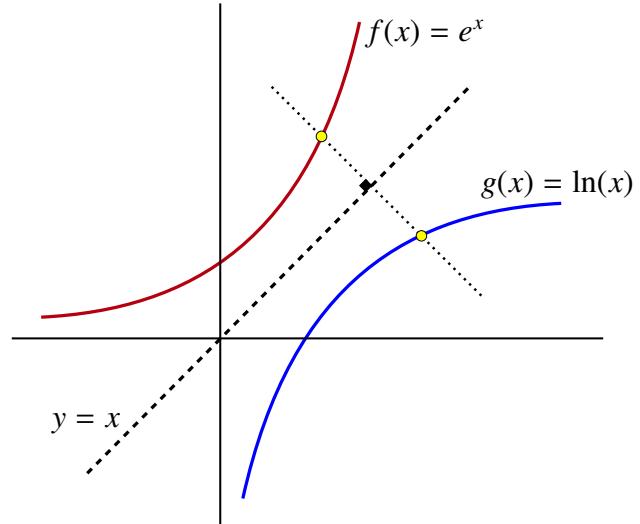
We can use these ideas to find the derivative of the natural logarithm function.

EXAMPLE 19 Derivative of $\ln(x)$

In this example, the derivative of $f(x) = \ln(x)$ is derived.

We know that $\ln(x)$ is invertible with inverse $g(y) = e^y$. Then for any $x > 0$ we have

$$g(f(x)) = e^{\ln(x)} = x.$$



The Chain Rule shows that

$$g'(f(x))f'(x) = 1$$

and hence that

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

But $g'(y) = g(y) = e^y$ for any y . This means that

$$\begin{aligned} f'(x) &= \frac{1}{g'(f(x))} \\ &= \frac{1}{e^{f(x)}} \\ &= \frac{1}{e^{\ln(x)}} \\ &= \frac{1}{x}. \end{aligned}$$

We have just shown that if $f(x) = \ln(x)$, then $f'(x) = \frac{1}{x}$. ◀

6.10.1 The Proof of the Inverse Function Theorem

In the previous section we gave a convincing geometric derivation of the Inverse Function Theorem. In this section we would like to give a proof of this very powerful tool.

Possible Strategy: We would like to say

$$\begin{aligned}
 \lim_{y \rightarrow b} \frac{g(y) - g(b)}{y - b} &= \lim_{y \rightarrow b} \frac{g(y) - a}{y - f(a)} \\
 &= \lim_{x \rightarrow a} \frac{x - a}{f(x) - f(a)} \\
 &= \lim_{x \rightarrow a} \frac{1}{\frac{f(x) - f(a)}{x - a}} \\
 &= \frac{1}{f'(a)}.
 \end{aligned}$$

Question: Why is this not a valid proof???

REMARK

It turns out that the main issue with the proposed proof lies in our second line. Transferring the limit from $y \rightarrow b$ over to $x \rightarrow a$ seems to make sense because $x = g(y)$ and $a = g(b)$, so we would assume that as $y \rightarrow b$, we would have $g(y) \rightarrow g(b)$. But in reality this is exactly what it means for the inverse function g to be continuous at $y = b$ and at this point we do not yet know that the continuity of g follows from that of f . If we can show that this is the case, then the calculation above is really a valid proof of the Inverse Function Theorem. As such we will devote the rest of this section to showing that if f is continuous, so is g . ◀

We tend to associate invertible functions with monotonic functions. Our next proposition will show that if f is continuous, they are one and the same.

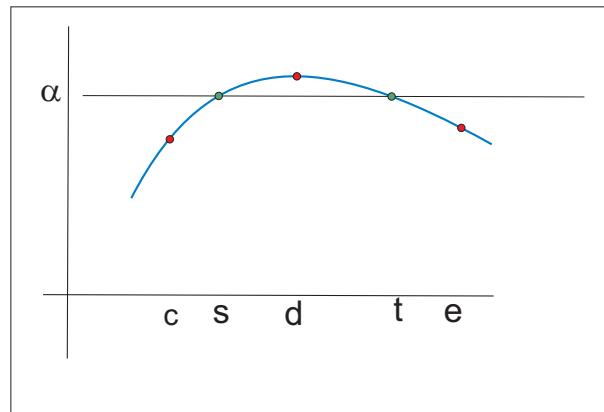
PROPOSITION 12

Suppose that f is continuous and one-to-one on $[a, b]$, then f is either increasing or decreasing on $[a, b]$.

PROOF

If f is neither increasing nor decreasing, then there exists points $c, d, e \in [a, b]$ with $c < d < e$ such that either:

- 1) $f(c) < f(d)$ and $f(d) > f(e)$ or
- 2) $f(c) > f(d)$ and $f(d) < f(e)$.



Case 1: If $f(c) < f(d)$ and $f(d) > f(e)$, then we can find $\alpha \in \mathbb{R}$ so that

$$f(c) < \alpha < f(d) \text{ and } f(e) < \alpha < f(d).$$

Since f is continuous on $[c, d]$ and on $[d, e]$ the IVT gives $s \in (c, d)$ and $t \in (d, e)$ such that

$$f(s) = \alpha = f(t)$$

which is impossible if f is one-to one.

Case 2: If $f(c) > f(d)$ and $f(d) < f(e)$, then we can argue in a manner similar to Case 1. ■

We will now prove a useful version of Monotone Convergence Theorem for Functions. The proof is remarkably similar to that of the MCT for sequences.

THEOREM 13

The Monotone Convergence Theorem for Functions

Suppose that f is increasing on $[a, b]$. Then

1) $\lim_{x \rightarrow c^+} f(x)$ exists for all $c \in [a, b)$ and $\lim_{x \rightarrow c^+} f(x) = \text{glb}(S)$ where

$$S = \{f(x) \mid x \in (c, b]\}.$$

2) $\lim_{x \rightarrow c^-} f(x)$ exists for all $c \in (a, b]$ and $\lim_{x \rightarrow c^-} f(x) = \text{lub}(T)$ where

$$T = \{f(x) \mid x \in [a, c)\}.$$

PROOF

We will only prove 1) as 2) follows in much the same manner.

Let

$$S = \{f(x) \mid x \in (c, b]\}.$$

Then S is bounded below by $f(c)$.

Let

$$L = \text{glb}(S).$$

Let $\epsilon > 0$. Then $L < L + \epsilon$ so $L + \epsilon$ is not a lower bound for S and hence there exists $d \in (c, b]$ such that

$$L \leq f(d) < L + \epsilon.$$

If $x \in (c, d)$, then

$$L \leq f(x) < f(d) < L + \epsilon.$$

This shows that

$$\lim_{x \rightarrow c^+} f(x) = L.$$

■

REMARK

The Monotone Convergence Theorem for functions tells us that if f is monotonic on $[a, b]$, then the only types of discontinuities that f could have would be finite jump discontinuities. Combining this observation with the Intermediate Value Theorem yields the following important characterization of continuity for monotonic functions:

◀

THEOREM 14**Continuity for Monotonic Functions**

Suppose that f is increasing on $[a, b]$. Then f is continuous on $[a, b]$ if and only if

$$f([a, b]) = \{f(x) \mid x \in [a, b]\} = [f(a), f(b)].$$

PROOF

Since f is increasing, we have for each $x \in [a, b]$ that

$$f(a) \leq f(x) \leq f(b).$$

It follows that

$$f([a, b]) \subseteq [f(a), f(b)].$$

⇒): Assume that f is continuous and that

$$f(a) < \alpha < f(b).$$

By the IVT there exists $c \in (a, b)$ such that

$$f(c) = \alpha.$$

Hence

$$f([a, b]) = [f(a), f(b)].$$

⇐): Assume that f is not continuous at some point $c \in (a, b)$. Then

$$\lim_{x \rightarrow c^-} f(x) = L < M = \lim_{x \rightarrow c^+} f(x).$$

However, we would then have that

$$[L, M] \cap f([a, b]) = \{f(c)\}.$$

We know that $[L, M]$ is infinite. It follows that

$$f([a, b]) \neq [f(a), f(b)].$$

If f is discontinuous at $x = a$, then

$$f(a) < M = \lim_{x \rightarrow a^+} f(x).$$

From here we note that

$$(f(a), M) \cap f([a, b]) = \emptyset.$$

Similarly, if f is discontinuous at $x = b$, then

$$\lim_{x \rightarrow b^-} f(x) = L < f(b).$$

This time we have that

$$(L, f(b)) \cap f([a, b]) = \emptyset.$$

Hence, if f is not continuous, then

$$f([a, b]) \neq [f(a), f(b)].$$

■

REMARK

Since the previous two results also hold for decreasing functions we now have an easy criterion for determining if a monotonic function is continuous on an interval I .

Moreover, this criterion provides us with the key tool to prove the following theorem which in turn completes our proof of the Inverse Function Theorem.



THEOREM 15

Continuity for Inverse Functions

Suppose that $f : [a, b] \rightarrow \mathbb{R}$, is continuous and one-to-one with $f([a, b]) = [c, d]$. Let $g : [c, d] \rightarrow [a, b]$ be the inverse function of f on $[c, d]$. Then g is continuous on $[c, d]$.

PROOF

Since f is either increasing or decreasing on $[a, b]$, it follows that g is also either increasing or decreasing. Since

$$g([c, d]) = [a, b]$$

g is continuous on $[c, d]$.



6.11 Derivatives of Inverse Trigonometric Functions

At the end of the previous section, the Inverse Function Theorem and the Chain Rule were used to calculate the derivative of the function $f(x) = \ln(x)$. In this section, we will use the same method to find the derivatives of $\arccos(x)$, $\arcsin(x)$ and $\arctan(x)$.

EXAMPLE 20

Derivative of $\arcsin(x)$

For any $x \in [-1, 1]$, if $y = f(x) = \arcsin(x)$ and if $x = g(y) = \sin(y)$ with $y \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, then

$$g(f(x)) = \sin(\arcsin(x)) = x.$$

Applying the Chain Rule gives us

$$g'(f(x))f'(x) = 1$$

and hence

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

But since $g'(y) = \cos(y)$, we get

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

To simplify this further, we again use the fact that $\cos^2(y) + \sin^2(y) = 1$ so that

$$\cos(y) = \pm \sqrt{1 - \sin^2(y)}.$$

However, $y \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ requires that $\cos(y) \geq 0$ so

$$\cos(y) = \sqrt{1 - \sin^2(y)}.$$

We now have

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

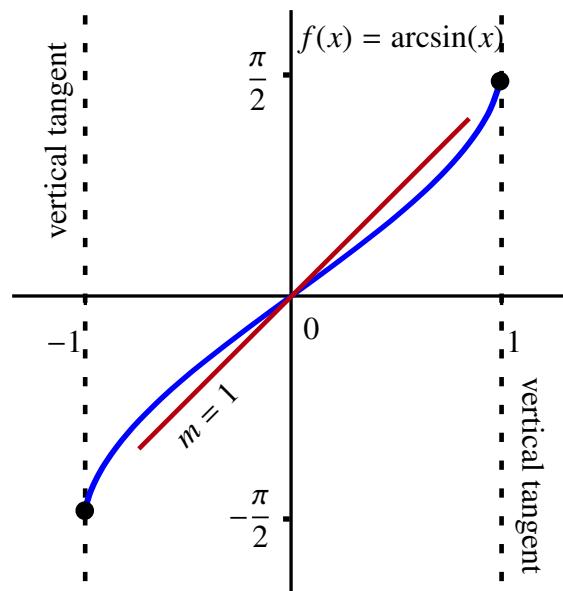
But $y = f(x) = \arcsin(x)$, so

$$\begin{aligned} f'(x) &= \frac{1}{\sqrt{1 - \sin^2(f(x))}} \\ &= \frac{1}{\sqrt{1 - \sin^2(\arcsin(x))}} \\ &= \frac{1}{\sqrt{1 - x^2}}. \end{aligned}$$

This shows that

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

The derivative calculation is consistent with the graph of $\arcsin(x)$.



Observe that the derivative

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

is always positive which we should expect since, as the graph shows, $\arcsin(x)$ is an increasing function.

As we approach $x = -1$ or $x = 1$, the graph suggests that the tangent lines become very steep and that there is a vertical tangent line at both $x = -1$ and $x = 1$. But we know that

$$\lim_{x \rightarrow -1^+} \frac{1}{\sqrt{1 - x^2}} = \infty$$

and

$$\lim_{x \rightarrow 1^-} \frac{1}{\sqrt{1 - x^2}} = \infty$$

so this behavior is expected.

Finally, since

$$\sin(0) = 0 = \arcsin(0)$$

the Inverse Function Theorem implies that

$$f'(0) = \frac{1}{\sqrt{1 - 0^2}} = 1.$$

This calculation is again consistent with the graph. ◀

A word of caution is required. Our calculation was based on the assumption that $\arcsin(x)$ was differentiable since we need this assumption to apply the Chain Rule. However, the Inverse Function Theorem tells us that $\arcsin(x)$ is differentiable when $x \in (-1, 1)$, so we need not worry!

EXAMPLE 21

Derivative of $\arctan(x)$

For any $x \in (-\infty, \infty)$, if $y = f(x) = \arctan(x)$ and if $x = g(y) = \tan(y)$ with $y \in (-\frac{\pi}{2}, \frac{\pi}{2})$, then

$$g(f(x)) = \tan(\arctan(x)) = x.$$

Applying the Chain Rule gives us that

$$g'(f(x))f'(x) = 1$$

and hence

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

But since $g'(y) = \sec^2(y)$ we get

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

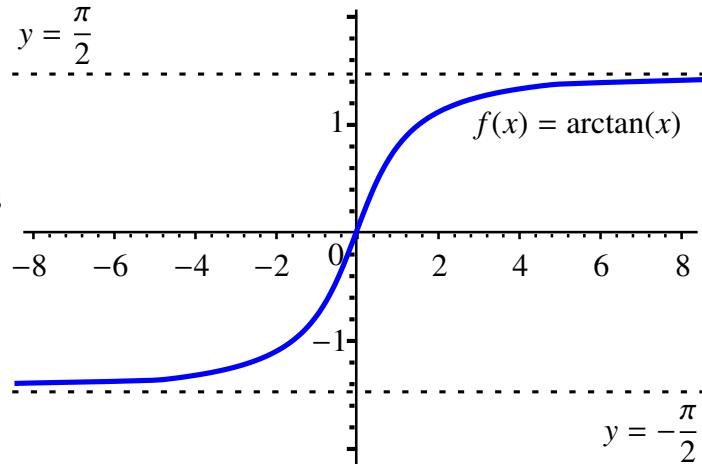
To simplify this equation, use the identity $\sec^2(y) = 1 + \tan^2(y)$ so that

$$\begin{aligned}
 f'(x) &= \frac{1}{1 + \tan^2(f(x))} \\
 &= \frac{1}{1 + \tan^2(\arctan(x))} \\
 &= \frac{1}{1 + x^2}.
 \end{aligned}$$

This shows us that

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

This derivative calculation is also consistent with the graph of $\arctan(x)$.



For example, we know that the derivative

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

is always positive which we expect since, as the graph shows, $\arctan(x)$ is an increasing function.

As we approach $-\infty$ or ∞ , the graph shows that the tangent lines become very flat and that $y = -\frac{\pi}{2}$ and $y = \frac{\pi}{2}$ are horizontal asymptotes. This is consistent with the fact that

$$\lim_{x \rightarrow \infty} \frac{d}{dx}(\arctan(x)) = \lim_{x \rightarrow \infty} \frac{1}{1 + x^2} = 0$$

and

$$\lim_{x \rightarrow -\infty} \frac{d}{dx}(\arctan(x)) = \lim_{x \rightarrow -\infty} \frac{1}{1 + x^2} = 0.$$



EXAMPLE 22 Derivative of $\arccos(x)$

Recall that for any $x \in [-1, 1]$, if $y = f(x) = \arccos(x)$ and if $x = g(y) = \cos(y)$ with

$y \in [0, \pi]$, then

$$g(f(x)) = \cos(\arccos(x)) = x.$$

Applying the Chain Rule gives

$$g'(f(x))f'(x) = 1$$

and hence that

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

But since $g'(y) = -\sin(y)$ we get

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

To simplify this further, remember that $\cos^2(y) + \sin^2(y) = 1$ so that

$$\sin(y) = \pm \sqrt{1 - \cos^2(y)}.$$

However, $y \in [0, \pi]$ means that $\sin(y) \geq 0$ and as such

$$\sin(y) = \sqrt{1 - \cos^2(y)}.$$

We now have

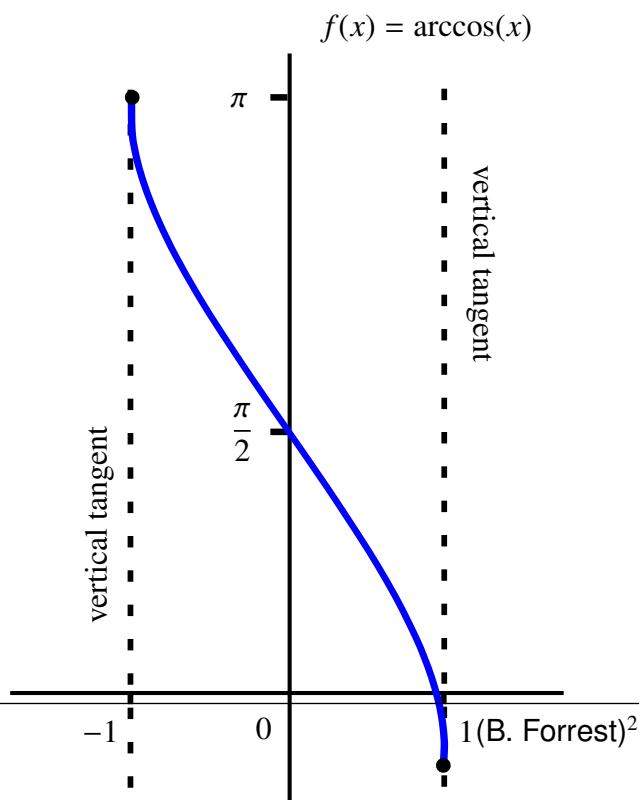
$$f'(x) = \frac{-1}{\sqrt{1 - \cos^2(f(x))}}.$$

But $y = f(x) = \arccos(x)$, so

$$\begin{aligned} f'(x) &= \frac{-1}{\sqrt{1 - \cos^2(f(x))}}. \\ &= \frac{-1}{\sqrt{1 - \cos^2(\arccos(x))}}. \\ &= \frac{-1}{\sqrt{1 - x^2}}. \end{aligned}$$

This shows that

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





EXAMPLE 23 Let $f(x) = \arctan(x^2)$. Find $f'(x)$.

SOLUTION The solution to this question is a simple application of the Chain Rule. Let $u = x^2$ and $y = \arctan(u)$. Then

$$\begin{aligned} f'(x) &= \frac{dy}{du} \frac{du}{dx} \\ &= \frac{1}{1+u^2}(2x) \\ &= \frac{2x}{1+x^4}. \end{aligned}$$

◀

EXAMPLE 24 Let

$$H(x) = \arcsin(x) + \arccos(x).$$

Show that $H'(x) = 0$.

SOLUTION Taking the derivative of $H(x)$, we get

$$\begin{aligned} H'(x) &= \frac{1}{\sqrt{1-x^2}} + \frac{-1}{\sqrt{1-x^2}} \\ &= 0 \end{aligned}$$

for all $x \in (-1, 1)$.

◀

This simple calculation reveals an interesting relationship between the functions $\arcsin(x)$ and $\arccos(x)$. To see this relationship, note that the function H is a continuous function on the interval $[-1, 1]$ with $H'(x) = 0$ on the open interval $(-1, 1)$. Such a function must be *constant*—that is, there exists some number $c \in \mathbb{R}$ such that

$$H(x) = \arcsin(x) + \arccos(x) = c$$

for all $x \in [-1, 1]$.

What is the value of c ? To answer this we could simply evaluate

$$H(0) = \arcsin(0) + \arccos(0).$$

But $\arcsin(0) = 0$ and $\arccos(0) = \frac{\pi}{2}$. It follows that

$$\arcsin(x) + \arccos(x) = \frac{\pi}{2},$$

and hence that $\arcsin(x) = \frac{\pi}{2} - \arccos(x)$ for all $x \in [-1, 1]$.

Question: Can you think of a trigonometric identity that might explain this result?

6.12 Implicit Differentiation

Up until now we have usually expressed functions in an explicit form. That is, we have written

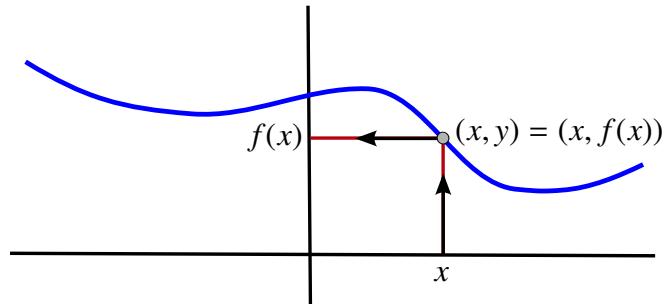
$$y = f(x)$$

to indicate that y is a function of the variable x with the rule for evaluation given explicitly and represented by the expression “ $f(x)$.” For example, if $y = x^2$ we know exactly how the rule works.

Once we have a function, its graph is all of the points of the form

$$\{(x, f(x)) \mid x \in \text{dom}(f)\}.$$

On the other hand, given the graph of a function, we can determine the value of the function at a point x in its domain by following the arrows as indicated.



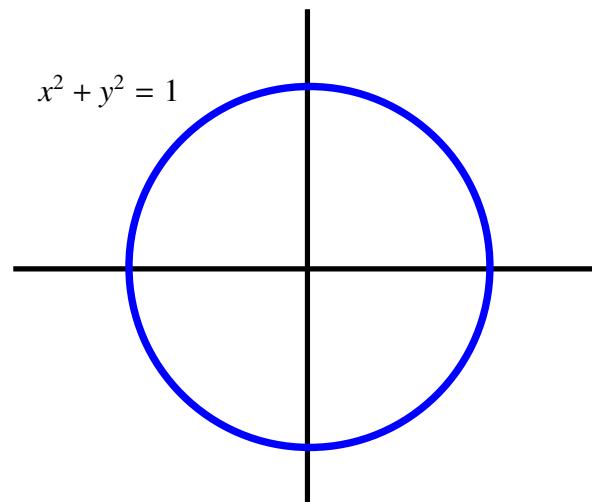
We can do this because for each x_0 in the domain of the function, the line $x = x_0$ cuts the graph at exactly one point (x_0, y_0) .

However, sometimes the functional relationship between x and y is *implicit* rather than explicit.

For example, consider the equation

$$x^2 + y^2 = 1.$$

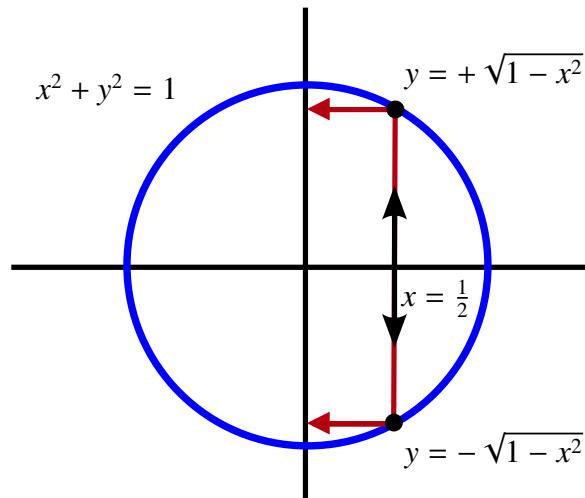
A relationship between x and y has certainly been specified, but in this form it does not look like a function.



However, if we “solve” this expression for y in terms of x , we get

$$y = \pm \sqrt{1 - x^2}.$$

This is still not a functional relation since, for example, when $x = \frac{1}{2}$ we have two choices for y , but functions must assign each point in their domain to one and only one value.



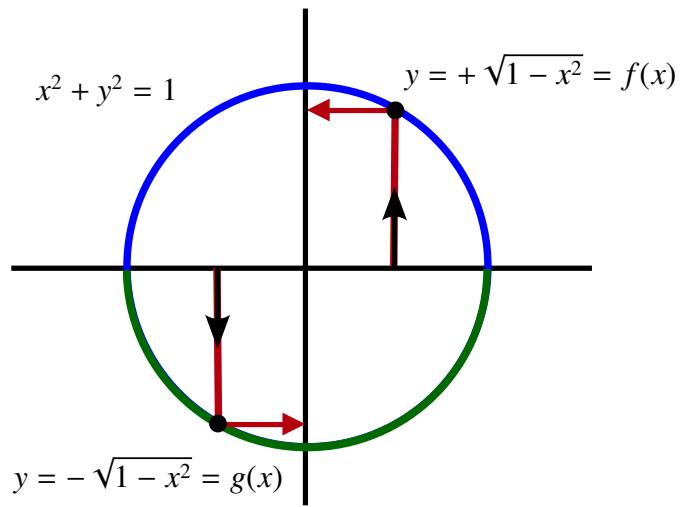
None the less, the relation actually “implies” at least two different functions, namely

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

and

$$y = g(x) = -\sqrt{1 - x^2}.$$

The graph of f is just the top half of the circle and the graph of g is the bottom half.



Both f and g are examples of functions that can be extracted implicitly from the expression $x^2 + y^2 = 1$.

Once we have an explicit function such as

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

we can treat it in the usual way. For example, we could differentiate to get

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

Notice that the denominator in this expression is just y , so we can substitute to get the equivalent expression

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

This suggests that the slope of the tangent line could be determined by only knowing the x and y coordinates of a point on the graph.

If we repeat this method with the second function

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

we get

$$\frac{dy}{dx} = g'(x) = \frac{x}{\sqrt{1 - x^2}}.$$

Since

$$y = -\sqrt{1 - x^2},$$

we still get that

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

These may be the most natural functions that are implicitly determined by the relation

$$x^2 + y^2 = 1$$

but there are many others. For example, we could let

$$y = h(x)$$

where

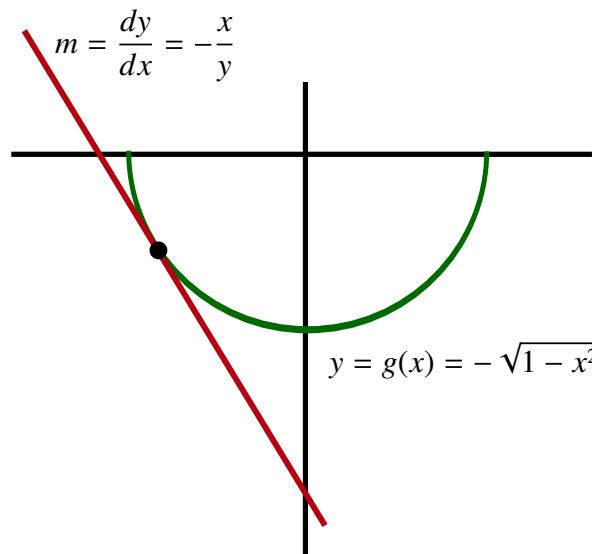
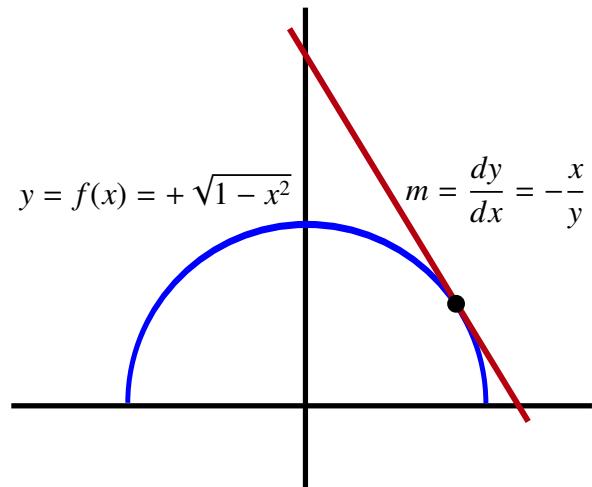
$$h(x) = \begin{cases} \sqrt{1 - x^2} & \text{if } x \in [0, 1] \\ -\sqrt{1 - x^2} & \text{if } x \in [-1, 0] \end{cases}.$$

This function is continuous except at $x = 0$. It is differentiable except at $x = \pm 1$ and $x = 0$. Moreover, wherever the derivative exists it also satisfies

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

In fact, if

$$y = h(x)$$



was any differentiable function that satisfied the equation

$$x^2 + (h(x))^2 = x^2 + y^2 = 1$$

then we would have

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

But

$$\begin{aligned}\frac{d}{dx}(x^2 + y^2) &= \frac{d}{dx}(x^2) + \frac{d}{dx}(y^2) \\ &= 2x + 2y\frac{dy}{dx}\end{aligned}$$

with the last equality following by the Chain Rule. Since $\frac{d}{dx}(1) = 0$ we have

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

Solving for $\frac{dy}{dx}$ gives us

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

This process of finding the derivative from the relation $x^2 + y^2 = 1$ without actually identifying the function is called *implicit differentiation*. Let's look at another example.

EXAMPLE 25 Folium of Descartes

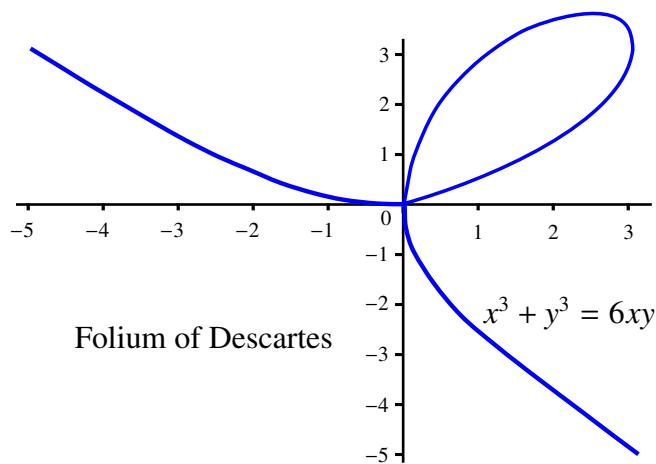
Consider the equation

$$x^3 + y^3 = 6xy$$

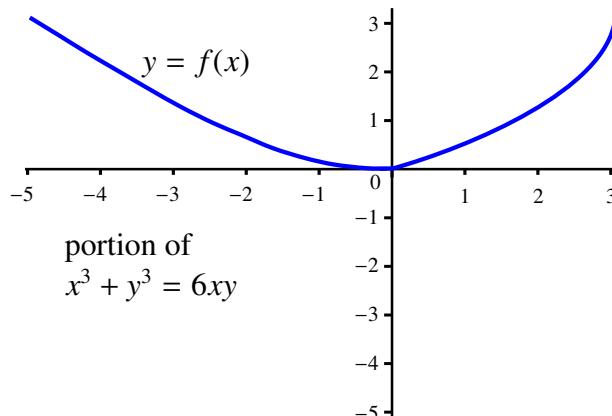
or equivalently

$$x^3 + y^3 - 6xy = 0.$$

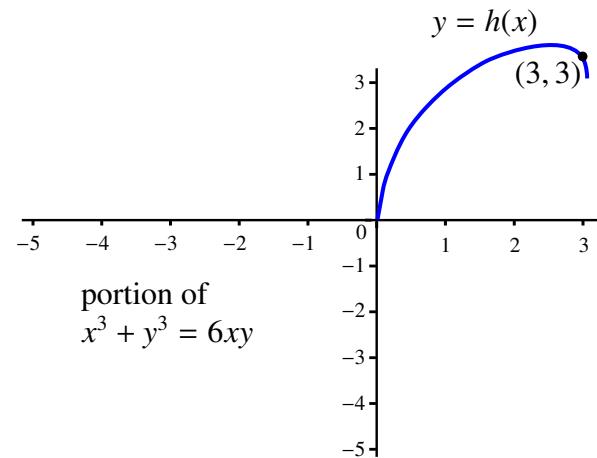
The set of points $\{(x, y) \mid x^3 + y^3 = 6xy\}$ forms a *curve* in the plane called the Folium of Descartes.



We can confirm from the picture that this is not the graph of a function. However, if we consider only a portion of the curve, we can get the graph of an implicitly defined function $y = f(x)$ that satisfies the equation $x^3 + (f(x))^3 = 6x \cdot f(x)$ on its domain. An example of such a function is given in the diagram.



This function is implicitly defined, but since it is very difficult to solve the equation for y in terms of x , we do not know the explicit rule for $f(x)$.



It is easy to verify that the point $(3, 3)$ is a solution to the equation and hence is also a point on the curve. Moreover, we can extract a different portion of the curve that represents the graph of a new function $y = h(x)$ with $h(3) = 3$.

Again, we do not know the explicit formula for $h(x)$, but the graph suggests that it is differentiable at $x = 3$. We can still proceed to find its derivative using

$$\frac{d}{dx}(x^3 + y^3) = \frac{d}{dx}(6xy)$$

to get

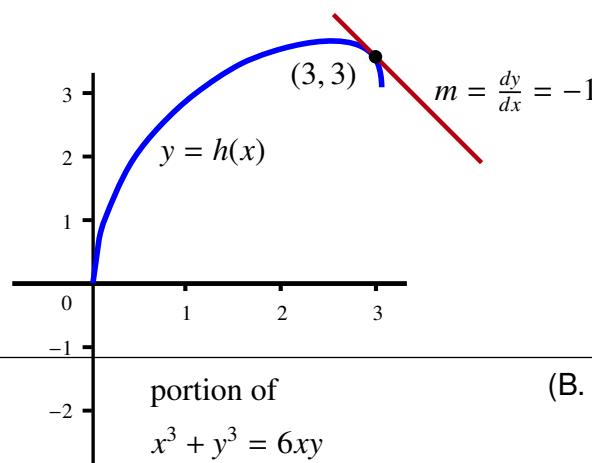
$$3x^2 + 3y^2 \frac{dy}{dx} = 6y + 6x \frac{dy}{dx}.$$

Solving this equation for $\frac{dy}{dx}$ gives us

$$\frac{dy}{dx} = \frac{6y - 3x^2}{3y^2 - 6x}.$$

However, when $x = 3$, $y = h(x) = 3$ by our choice of $h(x)$. We can substitute $x = 3$ and $y = 3$ to get

$$h'(3) = \frac{dy}{dx} \Big|_{(3,3)} = -1.$$



In fact, the following statement is true. If g is any differentiable function implicitly defined by the equation

$$x^3 + y^3 = 6xy,$$

that is if

$$x^3 + (g(x))^3 = 6xg(x),$$

and if (x, y) is any point with $y = g(x)$, then we would again have

$$g'(x) = \frac{dy}{dx} = \frac{6y - 3x^2}{3y^2 - 6x}.$$



EXAMPLE 26 Assume that

$$x^2y + y^2x = 6$$

defines an implicit function

$$y = f(x)$$

with $f(1) = 2$. (Note that the point $(1,2)$ is a solution to this equation.) Assume also that f is differentiable. Find $f'(1)$.

We know that

$$f'(1) = \frac{dy}{dx} \Big|_{(1,2)}.$$

Implicit differentiation gives us that

$$\frac{d}{dx}(x^2y + y^2x) = \frac{d}{dx}(6).$$

The Product Rule and Chain Rule give us

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

Rearranging terms this becomes

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

so

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

Finally

$$\begin{aligned} \frac{dy}{dx} \Big|_{(1,2)} &= \frac{-2xy - y^2}{x^2 + 2xy} \Big|_{(1,2)} \\ &= \frac{-2(1)(2) - 2^2}{1^2 + 2(1)(2)} \\ &= \frac{-8}{5}. \end{aligned}$$



The following example shows why **implicit differentiation requires some caution**.

EXAMPLE 27 Consider the equation

$$x^4 + y^4 = -1 - x^2y^2.$$

If we apply the method of implicit differentiation, you can verify that we would get

$$\frac{dy}{dx} = \frac{-2xy^2 - 4x^3}{4y^3 + 2x^2y}.$$

However, what does this mean? If you look closely at the left-hand side of the equation $x^4 + y^4 = -1 - x^2y^2$, you will see that for any pair (x, y) this expression will always be greater than or equal to 0 while the right-hand side is at most -1 . Therefore, **the equality is never satisfied, so there is no implicitly defined function**. We have in fact found the derivative of a ghost!! This is why the existence of the implicit function was always assumed before we applied this procedure. ◀

Logarithmic Differentiation

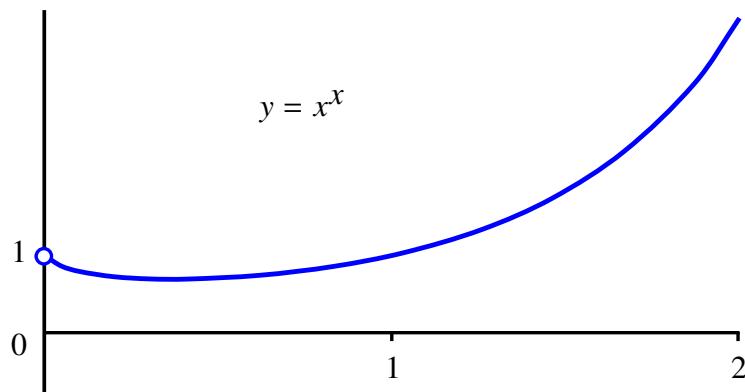
Implicit differentiation does have a very useful application, called *logarithmic differentiation*, that enables us to find derivatives of functions of the form

$$h(x) = g(x)^{f(x)}.$$

We will illustrate this method with an example.

EXAMPLE 28

Let $y = x^x$. The following is the graph of this unusual function. Notice that it is only defined for $x > 0$.



The graph looks quite smooth so we can assume that this function is differentiable. However, since both the base and the exponent vary with x , our current rules do not help us find the derivative. We can get around this problem by using a trick. Take the logarithm of both sides of the equation to get the following equality:

$$\ln(y) = x \ln(x).$$

Now differentiate this equation implicitly to get

$$\begin{aligned}\frac{1}{y} \frac{dy}{dx} &= \ln(x) + x\left(\frac{1}{x}\right) \\ &= \ln(x) + 1.\end{aligned}$$

Solving this for $\frac{dy}{dx}$ gives us

$$\begin{aligned}\frac{dy}{dx} &= y(\ln(x) + 1) \\ &= x^x(\ln(x) + 1).\end{aligned}$$



6.13 Local Extrema

In a previous section, we introduced the notion of *global extrema* for a function defined on an interval I . We also saw that the Extreme Value Theorem told us that if a function f is continuous on a closed interval $[a, b]$, then there is always both a global maximum and a global minimum located on $[a, b]$. Moreover, these extrema can either be located at the endpoints or inside the open interval (a, b) .

The fact that a global extrema for a continuous function f on a closed interval $[a, b]$ will often occur within the open interval (a, b) suggests that it would be very worthwhile to try and identify the characteristics of a point c at which f achieves either its maximum or minimum value on an *open* interval. In this section, we will see that for differentiable functions there is a simple criterion that can help us identify such points.

We begin by considering the following definition:

DEFINITION

Local Maxima and Local Minima

A point c is called a *local maximum* for a function f if there exists an open interval (a, b) containing c such that

$$f(x) \leq f(c)$$

for all $x \in (a, b)$.

A point c is called a *local minimum* for a function f if there exists an open interval (a, b) containing c such that

$$f(c) \leq f(x)$$

for all $x \in (a, b)$.

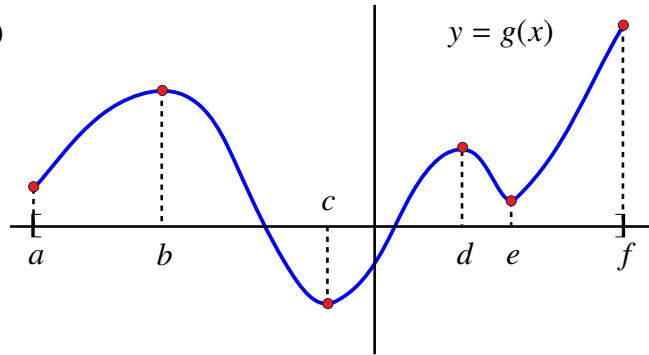
REMARK

The difference between a local maximum and a global maximum (or between a local minimum and a global minimum) is subtle. A *global* maximum is the point, if it exists, where the function takes on its *largest* value over the *entire* interval I . A *local* maximum is a point at which the function takes on its largest value on some **open subinterval** of I , but possibly not all of I .

A *local* maximum is a point c where the function takes on its *largest* value on a *potentially very small* portion of the interval and that portion must, by definition, contain points on *both sides* of c . A *local* minimum is a point at which the function takes on its *least* value on some **open subinterval** of I , but possibly not all of I . 

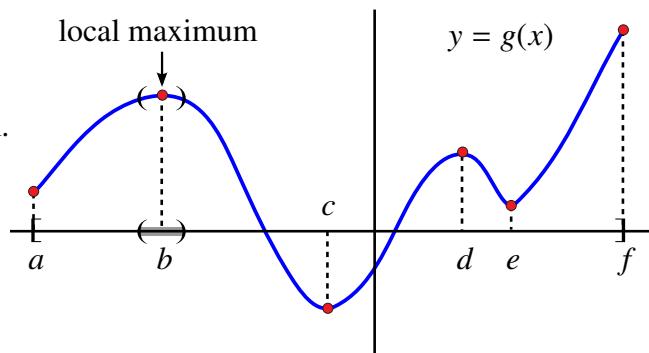
The following picture provides a better understanding about these points.

The picture represents the graph of a continuous function $y = g(x)$ defined on a closed interval $[a, f]$. The Extreme Value Theorem ensures us that there is both a global maximum and a global minimum for g on $[a, f]$. We have identified a number of interesting points which we have labeled a, b, c, d, e and f for consideration.

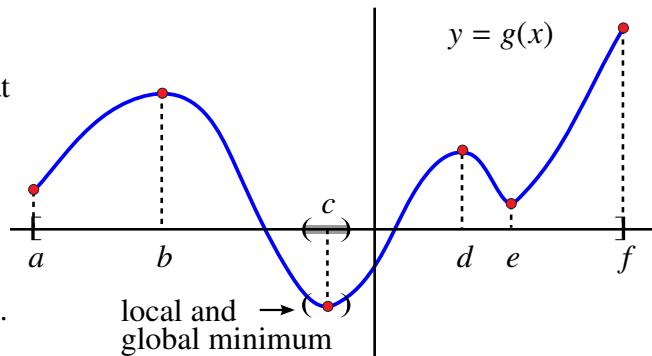


Let's begin by considering the left-hand end point a . This is *not* a global maximum nor a global minimum for g since $g(b) > g(a)$ and $g(c) < g(a)$. It is also *not* a local maximum nor a local minimum because a close look at the definition shows that to be either a local maximum or a local minimum, g must at least be defined a *little bit to the left* of a , which it is not since it is an end point of the interval.

The point b is neither a global maximum nor a global minimum. However, it *is* a local maximum. The diagram indicates an open interval on which $g(b)$ is the largest value.



The next point of interest is c . A close look at the graph shows that c is actually the global minimum of g on $[a, f]$. It is *also a local minimum* since as the diagram shows, an open interval exists around c on which c yields the minimum value of the function g .



It is worth noting that we could actually have chosen the interval (a, f) as the open interval in which c becomes a local minimum. This is an important observation because it demonstrates the following general fact.

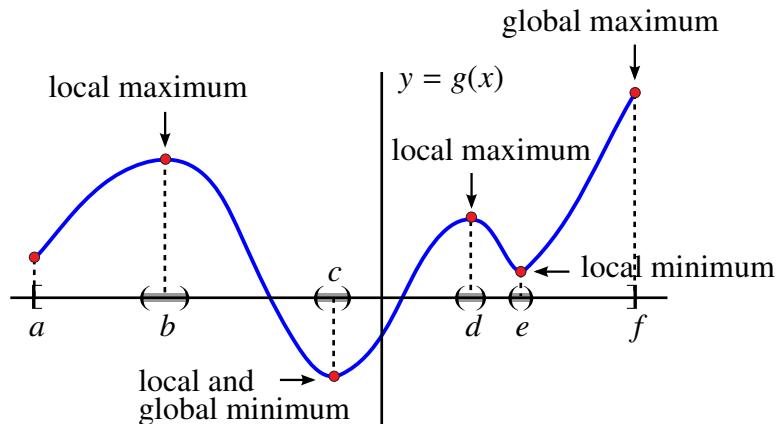
Fact

If g is a continuous function on a closed interval $[a, b]$ and if $a < c < b$ is a global maximum/minimum for g on $[a, b]$, then c is also a local maximum/minimum for g .

The next two points, d and e , are a local maximum and a local minimum, respectively. Neither is a global extremum.

The final point is f , the right-hand endpoint. Since $g(x)$ is not defined to the right of f , it follows that f *cannot* be a local maximum, but it is the global maximum for g on $[a, f]$.

The following diagram summarizes our analysis.



6.13.1 The Local Extrema Theorem

We have learned from the Extreme Value Theorem that a continuous function on a closed interval $[a, b]$ always attains its maximum and its minimum value at points in

the interval. Assume that $f(x) \leq f(c)$ for all $x \in [a, b]$. Either c is an endpoint, that is $c = a$ or $c = b$, or we have $a < c < b$. In the latter case, we have that $c \in (a, b)$ and $f(x) \leq f(c)$ for all $x \in (a, b)$. But this means that if $a < c < b$, then c satisfies the definition of a local maximum.

We have just seen that the maximum value of $f(x)$ on the closed interval $[a, b]$ either occurs at an endpoint or at a local maximum. Similarly, the minimum value of $f(x)$ on the closed interval $[a, b]$ either occurs at an endpoint or at a local minimum. Since the endpoints are easy to identify, we are forced to look at the problem of finding possible local maxima or minima. To see how to do this, begin by assuming that f is differentiable at the local maximum or local minimum.

Assume that f has a local maximum at $x = c$ and that $f'(c)$ exists. Since f has a local maximum at $x = c$, there exists an interval $a < c < b$ such that

$$f(x) \leq f(c)$$

for all $x \in (a, b)$. We also know that

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

However, this also means that

$$f'(c) = \lim_{h \rightarrow 0^+} \frac{f(c+h) - f(c)}{h}.$$

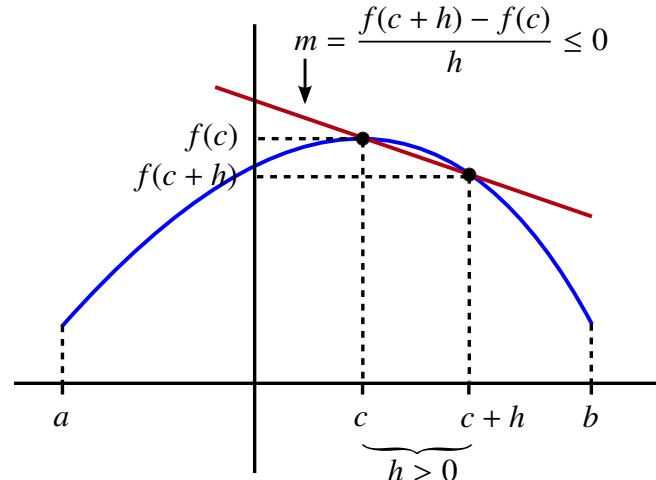
Choose $h > 0$ small enough so that $c < c+h < b$. Then because c is a local maximum, we have

$$f(c+h) - f(c) \leq 0.$$

Since $h > 0$ this means that the Newton Quotient

$$\frac{f(c+h) - f(c)}{h} \leq 0.$$

Geometrically, this says that the secant line slopes downward to the right.



We have shown that if $h > 0$ is small, then

$$\frac{f(c+h) - f(c)}{h} \leq 0.$$

The rules for limits give us that

$$f'(c) = \lim_{h \rightarrow 0^+} \frac{f(c+h) - f(c)}{h} \leq 0.$$

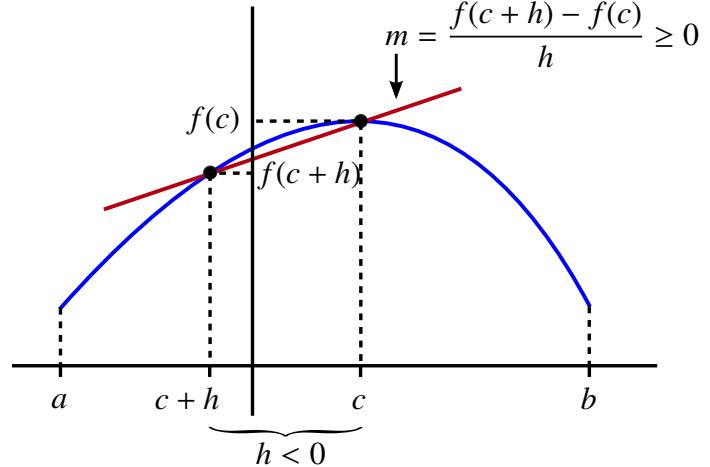
Now let $h < 0$ be chosen so that $a < c+h < c$. We again get that

$$f(c+h) - f(c) \leq 0.$$

But this time $h < 0$ so that

$$\frac{f(c+h) - f(c)}{h} \geq 0.$$

That is, the secant line slopes upward to the right.



We now have that if $h < 0$ is small, then

$$\frac{f(c+h) - f(c)}{h} \geq 0.$$

The rules for limits give us that

$$f'(c) = \lim_{h \rightarrow 0^-} \frac{f(c+h) - f(c)}{h} \geq 0.$$

To summarize, we have shown that

$$f'(c) \leq 0$$

and

$$f'(c) \geq 0.$$

The only way this can occur is if

$$f'(c) = 0.$$

Thus, if c is a local maximum for f and if $f'(c)$ exists, then we must have that

$$f'(c) = 0.$$

A similar argument shows that if c is a local minimum for f and if $f'(c)$ exists then we must have that

$$f'(c) = 0.$$

THEOREM 16**Local Extrema Theorem**

If c is a local maximum or local minimum for f and $f'(c)$ exists, then

$$f'(c) = 0.$$

Unfortunately, as the next two examples illustrate, we have not completed the story as far as finding local extrema is concerned. We will see that it is possible for $f'(c) = 0$ but that c is neither a local maximum or a local minimum. It is also possible that c is either a local maximum or a local minimum, but $f'(c)$ does not exist.

EXAMPLE 29

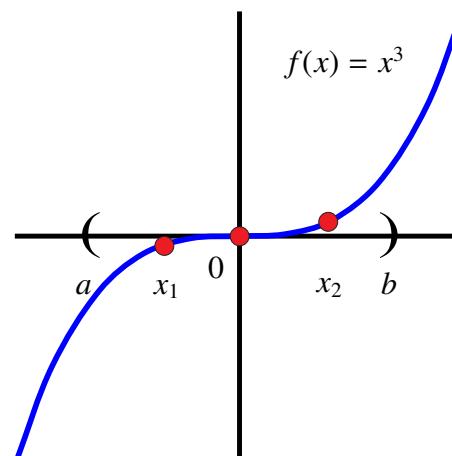
Let $f(x) = x^3$. Then $f'(x) = 3x^2$, so $f'(0) = 0$. However, 0 is neither a local maximum nor a local minimum for f . To see this we note that if we have any open interval (a, b) with $a < 0 < b$, then we can find two points x_1 and x_2 with

$$a < x_1 < 0 < x_2 < b.$$

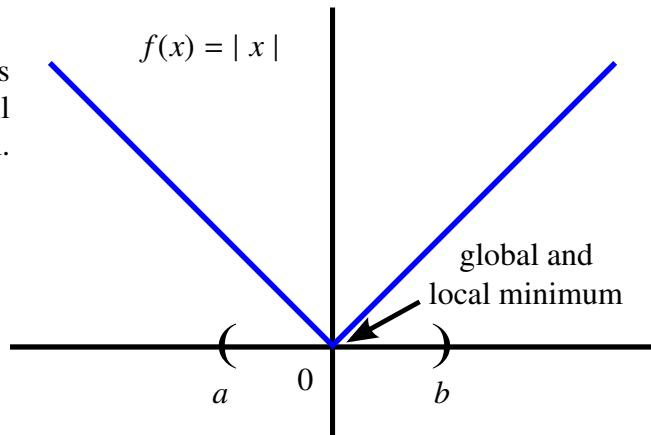
But then

$$f(x_1) < f(0) < f(x_2).$$

Since (a, b) was any open interval containing 0, this would be impossible if 0 was either a local maximum or a local minimum.

**EXAMPLE 30**

Let $f(x) = |x|$. Then $x = 0$ is a global minimum for f over all of \mathbb{R} . It is also a local minimum. However, we do not have $f'(0) = 0$ since $f'(0)$ does not exist.



We have just seen that in looking for local minima or maxima we should look for points where either $f'(x) = 0$ or where the function is not differentiable. This leads us to the following definition:

DEFINITION

Critical Point

A point c in the domain of a function f is called a *critical point* for f if either

$$f'(c) = 0$$

or

$$f'(c)$$

does not exist.

6.14 Related Rates

Many real world problems are concerned with the rate of change of a given quantity. In these situations we often start with a mathematical relationship between various quantities from which we can deduce a corresponding relationship between their respective rates of change. We call these *related rate* problems. To solve these problems we use *mathematical models*.

The basic idea behind mathematical modeling is to begin with a real world problem, interpret the problem by means of a mathematical expression which we call the *model*, manipulate the expression to gain information about the model and finally, use the information provided by the model to formulate a conclusion regarding the original problem.

In this section, we will look at some simple examples of related rate problems that can be solved by applying the ideas developed about derivatives and by using some simple mathematical modeling techniques.

EXAMPLE 31 The relationship between temperature T , pressure P and volume V for a gas is given by the formula

$$PV = kT$$

where k is a constant particular to the gas.

Assume that the gas is heated so that the temperature is increasing. Suppose also that the gas is allowed to expand so that pressure remains constant. If at a particular moment the temperature is 348 Kelvin, but is increasing at a rate of 2 Kelvin per second while the volume is increasing at a rate of 0.001 cubic meters per second, what is the volume of the gas?

SOLUTION It may not appear that we have enough information to solve this problem since we know nothing about k or about the pressure at that moment in time. Let's list what information is known.

We have the formula

$$PV = kT.$$

We can start by differentiating with respect to time t (remember k is a constant) to get

$$P \frac{dV}{dt} + V \frac{dP}{dt} = k \frac{dT}{dt}.$$

But we also know that the change in pressure remains constant, so $\frac{dP}{dt} = 0$.

Finally, we are told that at the instant when $T = 348$ Kelvin, we have that $\frac{dT}{dt} = +2$ Kelvin and $\frac{dV}{dt} = +0.001$ cubic meters per second.

Therefore, substituting we get

$$P \cdot (0.001) + V \cdot (0) = k \cdot (2)$$

or that

$$P = 2000k.$$

Substituting this expression for P back into the original formula and using the fact that $T = 348$ Kelvin gives us that

$$(2000k)V = k(348)$$

and hence that

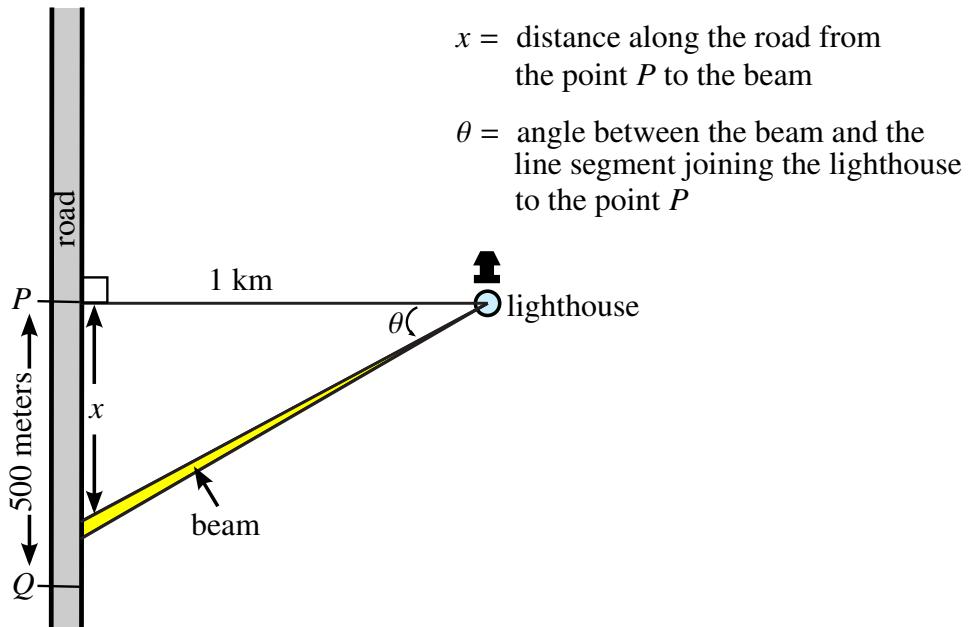
$$V = \frac{348}{2000} \cong 0.174 \text{ m}^3.$$

Notice that in this example we have been rather sloppy with respect to including units in the calculation. It is a valuable exercise to review this calculation and to verify the units. In particular, what are possible units for P and k ?

EXAMPLE 32

A lighthouse sends out a beam of light that rotates counter-clockwise 10 times every minute. The beam can be seen from a straight road that at its closest point P is 1 km away from the lighthouse. How fast is the beam moving along the road when it passes a point Q that is 500 m down the road from point P , with the beam moving away from P ?

SOLUTION The first step is to develop our mathematical model. For related rate questions it is often best to begin with a carefully labeled diagram and then try to identify the relationships that we have between our various quantities.



The problem asks us to determine how fast the beam is moving along the road at a particular point. In other words, how quickly is $x = x(t)$ changing? Mathematically, we are looking for

$$\frac{dx}{dt}.$$

We are also told that the light makes 10 complete rotations each minute. Since it is reasonable to assume that the light turns at a constant rate, then the light turns at the rate of $2\pi \cdot 10 = 20\pi$ radians per minute. We have just determined

$$\frac{d\theta}{dt} = 20\pi.$$

It is important to note that the derivative is positive since the beam is moving away from point P and so θ is increasing (since the beam is rotating counter-clockwise).

To determine $\frac{dx}{dt}$ we will make use of what we know about $\frac{d\theta}{dt}$. However, to do this we must first find a mathematical expression relating x and θ . For $0 < \theta < \frac{\pi}{2}$, the diagram shows that

$$x = x(\theta) = \tan(\theta(t)) = \tan(\theta).$$

Differentiating using the Chain Rule gives us

$$\frac{dx}{dt} = \sec^2(\theta) \frac{d\theta}{dt}.$$

We know that $\frac{d\theta}{dt} = 20\pi$. We also know that when the beam is 500 m = 0.5 km from point P , we have that

$$\tan(\theta) = \frac{1}{2}.$$

This means that

$$\theta = \arctan\left(\frac{1}{2}\right) \approx 0.4636 \text{ radians}$$

Substituting, we have

$$\begin{aligned} \frac{dx}{dt} &= \sec^2(\theta) \frac{d\theta}{dt} \\ &\approx \sec^2(0.4636)(20\pi) \\ &= 78.54 \text{ km/minute} \end{aligned}$$

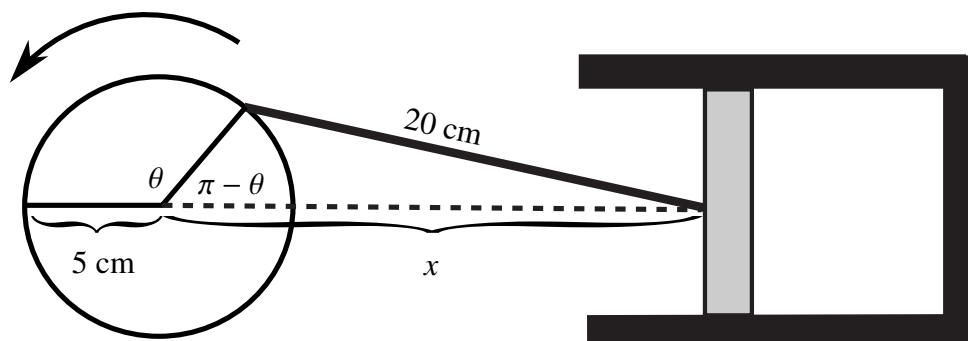
It follows that as the beam passes point Q , its speed along the road is

$$\sec^2(0.4636)(20\pi) = 78.54 \text{ km/minute.}$$

An interesting observation is that as $\theta \rightarrow \frac{\pi}{2}$, $\sec^2(\theta) \rightarrow \infty$. This means that as the point Q moves further away from P , we can expect to see that the speed at which the beam passes increases very rapidly despite the fact that the lighthouse rotates at a constant speed. ◀

EXAMPLE 33

A piston is attached to the exterior of a circular crankshaft with a radius of 5 cm by a steel rod of length 20 cm. The crankshaft is rotating counter-clockwise at a rate of 1000 revolutions per minute. Find the velocity of the piston when the angle $\theta = \frac{\pi}{2}$.



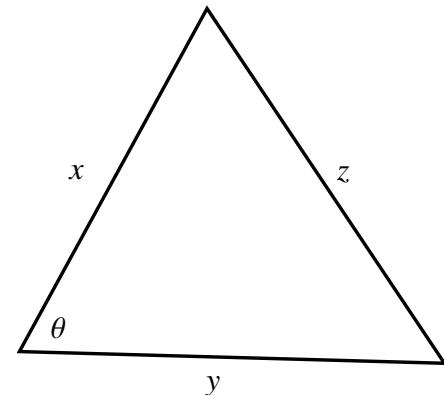
SOLUTION In the diagram the velocity of the piston is the rate of change of the quantity x . This means we are trying to find $\frac{dx}{dt}$. We also know that there are 1000 rpm with each revolution consisting of 2π radians. Therefore, we have

$$\frac{d\theta}{dt} = 1000 \cdot 2\pi$$

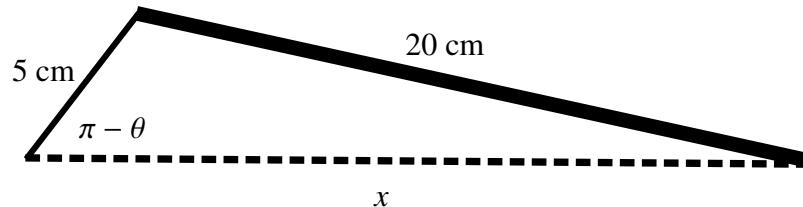
radians per minute. We must find an expression for the relationship between x and the angle θ . To do this we will use the *Cosine Law*.

Recall, that if we have a triangle with three sides labeled x, y, z and an angle θ opposite the side with length z , then the Cosine Law states that

$$z^2 = x^2 + y^2 - 2xy \cos(\theta)$$



In this case, we have a triangle with three sides that are $x, 20$ and 5 , respectively.



The angle $\pi - \theta$ is opposite the side of length 20 , so applying the Cosine Law leaves us with

$$20^2 = x^2 + 5^2 - 2(5)x \cos(\pi - \theta)$$

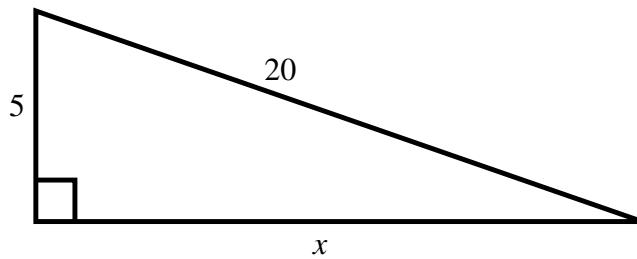
or

$$375 = x^2 - 10x \cos(\pi - \theta).$$

Differentiating both sides gives us

$$\begin{aligned} 0 &= \frac{d}{dt}(375) \\ &= \frac{d}{dt}(x^2 - 10x \cos(\pi - \theta)) \\ &= 2x \frac{dx}{dt} - 10 \cos(\pi - \theta) \frac{dx}{dt} - 10x \sin(\pi - \theta) \frac{d\theta}{dt}. \end{aligned}$$

We are interested in the case where $\theta = \frac{\pi}{2}$. In this case, the interior angle is $\pi - \frac{\pi}{2} = \frac{\pi}{2}$. This means that we actually are using a right triangle.



Moreover, since $\cos(\frac{\pi}{2}) = 0$, we have

$$x^2 = 375$$

so

$$x = \sqrt{375}.$$

Since $\cos(\frac{\pi}{2}) = 0$, $\sin(\frac{\pi}{2}) = 1$, and $\frac{d\theta}{dt} = 1000 \cdot 2\pi$, we can substitute into the previous expression and simplify to get

$$0 = 2\sqrt{375} \frac{dx}{dt} - 10\sqrt{375} (2000\pi).$$

Rearranging terms and solving for $\frac{dx}{dt}$ we get that

$$\frac{dx}{dt} = 10,000\pi \text{ cm/min.}$$



Chapter 7

The Mean Value Theorem

7.1 The Mean Value Theorem

Up until now when we have considered the derivative we have focused almost entirely on its *local* meaning. That is, on the information we can draw from the derivative about the nature of a function near the point at which we are differentiating. In this chapter we will study the *global* properties of differentiable functions over an interval. We will see that the derivative can tell us a great deal about the behavior of a function over an entire interval. For example, we will see that if $f'(x) > 0$ for all x in some interval I , we can conclude that the function is increasing on I .

The key to understanding the global implications of the derivative is the Mean Value Theorem (or MVT). This result states that the average rate of change for a differentiable function over an interval is equal to the instantaneous rate of change at some point in the interval. We will illustrate this idea by considering the following problem.

Problem:

A car travels forward a distance of 110 km on a straight road in a period of one hour. If the speed limit on the road is 100 km/hr, can you prove that the car must have exceeded the speed limit at some point?

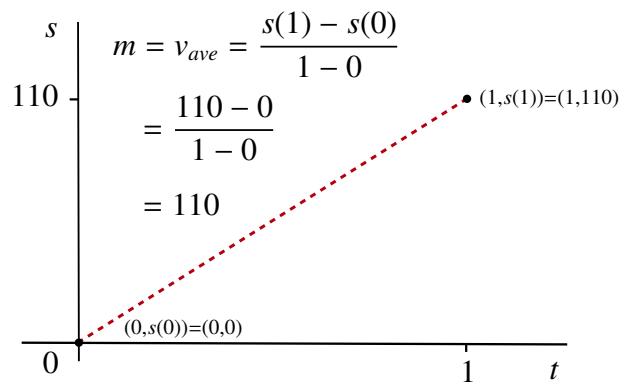
Using the information provided, we are led to the fact that the *average* velocity over the entire trip is

$$\frac{\text{displacement}}{\text{elapsed time}} = \frac{(110 - 0) \text{ km}}{(1 - 0) \text{ hr}} = 110 \text{ km/hr.}$$

It then seems reasonable that if the *average* velocity is 110 km/hr, the *instantaneous* velocity must have exceeded 100 km/hr at some point. Also, since the average velocity is 110 km/hr, it would make sense that at certain times the vehicle would be traveling in excess of 110 km/hour and at other times it would be traveling less than 110 km/hr. We might deduce that at some point the car would be traveling at exactly 110 km/hr. Unfortunately, this is *not* a proof! We must find a way to justify our intuition. To do this we will revisit what we have learned about the relationship between

average velocity and instantaneous velocity. In fact, we will actually show that there will be a time when the *instantaneous velocity was equal to the average velocity*.

Assuming the car is always driving forward, let $s(t)$ be the distance traveled from the starting point t hours from the beginning of the trip. The only information we currently know is that $s(0) = 0$ and $s(1) = 110$. The average velocity of 110 km/hr represents the *slope of the secant line* to the graph of s through the points $(0, s(0)) = (0, 0)$ and $(1, s(1)) = (1, 110)$.

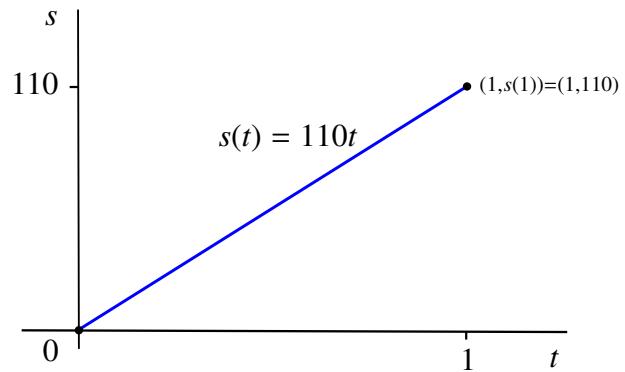


The simplest case for us to consider occurs if velocity was *constant* throughout the trip at 110 km/hour.

In this case, we would have

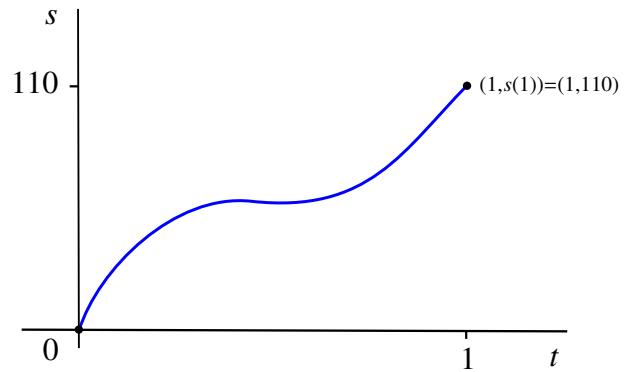
$$s(t) = 110t$$

and the graph would be a *straight line segment* that coincides with the previous secant line.



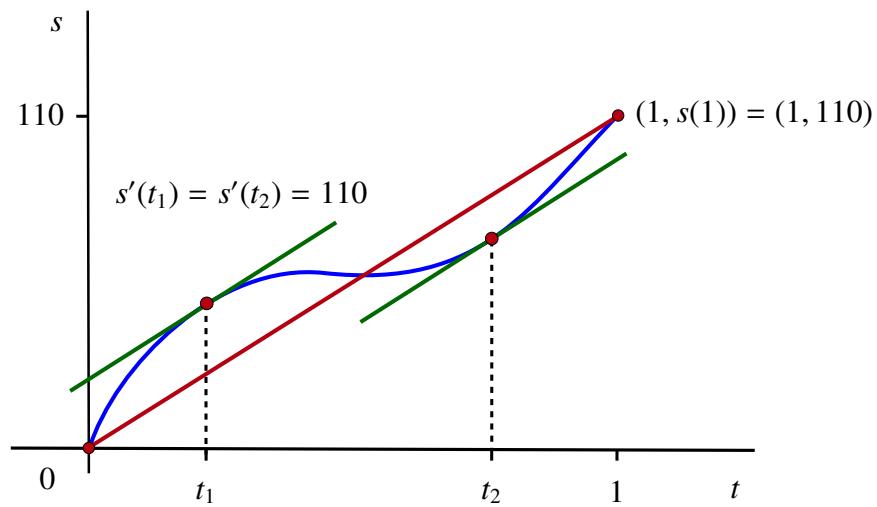
However, it would be unreasonable to believe that a constant velocity could be maintained throughout the entire trip.

It is more likely that the graph of distance versus time will appear as shown:



For this generic situation we want to know: Does there exist some point t_0 at which the *instantaneous velocity* is 110 km/hr?

Visually, a solution to this question occurs when the tangent line to the graph of s is parallel to the secant line through the points $(0, s(0)) = (0, 0)$ and $(1, s(1)) = (1, 110)$, since parallel lines have the same slope. On the graph this actually occurs at two distinct points as indicated on the following diagram.



The fact that we can always expect *at least one* such point is the main idea of the Mean Value Theorem.

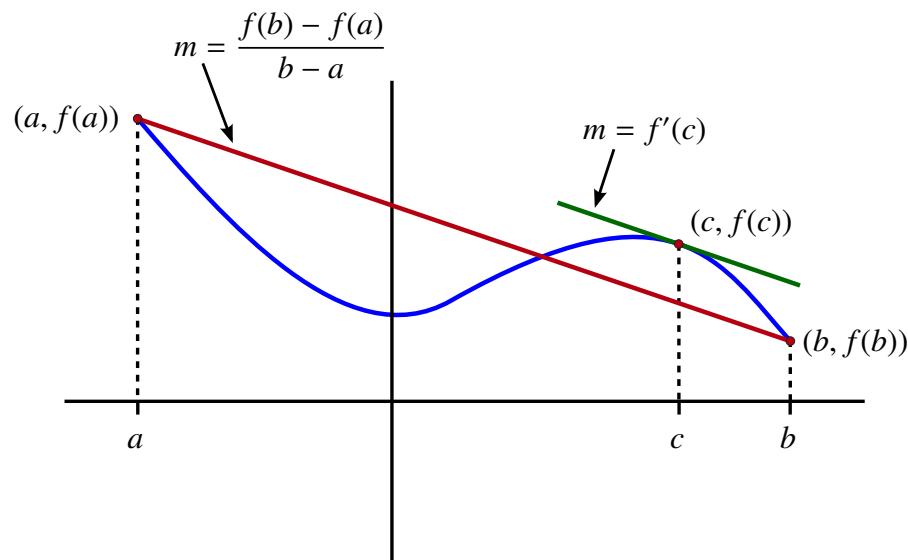
THEOREM 1

Mean Value Theorem

Assume that f is continuous on $[a, b]$ and f is differentiable on (a, b) . Then there exists $a < c < b$ such that

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

Essentially, the Mean Value Theorem states that for a differentiable function, the *average* rate of change over an interval will be the *same as the instantaneous* rate of change at some point c in the interval. Geometrically, this means that the tangent line to the graph of f through the point $(c, f(c))$ is *parallel* to the secant line through the points $(b, f(b))$ and $(a, f(a))$.



In order to justify the Mean Value Theorem, consider the following situation.

Assume that f is continuous on $[a, b]$, differentiable on (a, b) , and $f(a) = 0$ and $f(b) = 0$. Then

$$\frac{f(b) - f(a)}{b - a} = \frac{0 - 0}{b - a} = 0$$

so we want to show that there is point $a < c < b$ such that $f'(c) = 0$. There are three possible cases that we must address:

- 1) $f(x) = 0$ for all $x \in [a, b]$,
- 2) $f(x_0) > 0$ for some $x_0 \in [a, b]$, and
- 3) $f(x_0) < 0$ for some $x_0 \in [a, b]$.

In the first case, the function is constant on $[a, b]$, so it is easy to see that for any $a < c < b$, we have $f'(c) = 0$.

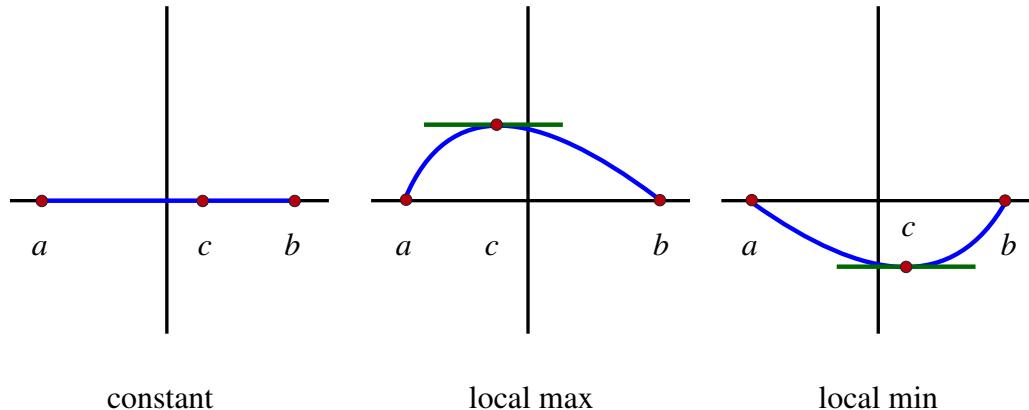
In both case 2 and case 3, we appeal to the Extreme Value Theorem to get that the function f must achieve both its maximum and minimum value on $[a, b]$.

In case 2, we have that the maximum value must be strictly greater than 0. As such the global maximum occurs at a point $x = c$ in the open interval (a, b) . But this means that c is also a local maximum. Finally, since f is differentiable at c , we get that $f'(c) = 0$ exactly as required.

In case 3, we have that the minimum value must be strictly less than 0. This time the global minimum occurs at a point $x = c$ in the open interval (a, b) and hence, c is also a local minimum. Just as before, we have that $f'(c) = 0$.

In all three cases, we have shown that there must be at least one point c in the interval (a, b) such that

$$f'(c) = \frac{f(b) - f(a)}{b - a} = 0.$$



The situation we have just looked at is important enough to be given its own theorem called *Rolle's Theorem*.

THEOREM 2**Rolle's Theorem**

Assume that f is continuous on $[a, b]$, that f is differentiable on (a, b) , and that $f(a) = 0 = f(b)$. Then there exists $a < c < b$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a} = 0.$$

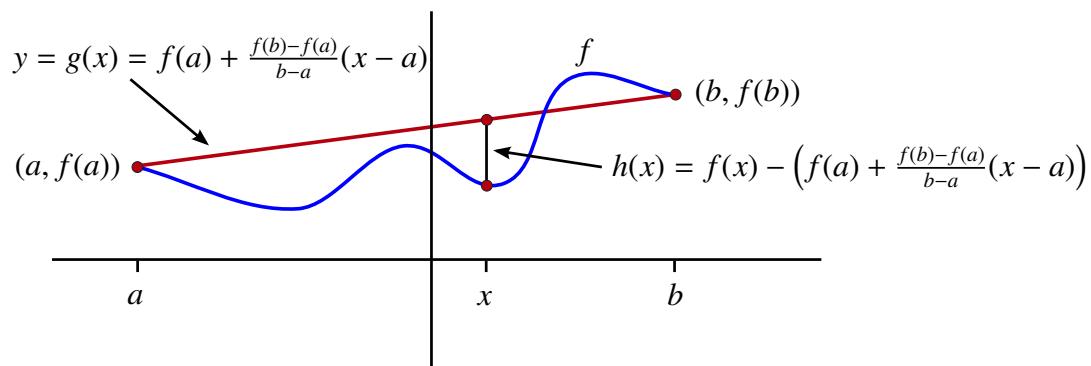
To see how the Mean Value Theorem follows from Rolle's Theorem, we introduce the following rather complicated looking function.

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

To understand what this function represents we first note that the function

$$y = g(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a)$$

is a linear function. Moreover, $g(a) = f(a)$ and $g(b) = f(b)$. This means that the graph of g passes through both $(a, f(a))$ and $(b, f(b))$. Therefore, the graph of g is the secant line through $(a, f(a))$ and $(b, f(b))$. But $h(x)$ is just $f(x) - g(x)$. This means that $h(x)$ is the vertical distance from the graph of f to the secant line joining $(a, f(a))$ and $(b, f(b))$ when the graph of f is above the secant line, and is the negative of this distance if the graph of f is below the secant line.



We also have that h is continuous on $[a, b]$, differentiable on (a, b) and is such that $h(a) = f(a) - f(a) = 0 = f(b) - f(b) = h(b)$. That is, h satisfies the conditions of Rolle's Theorem. It follows that there exists a point c with $a < c < b$ such that $h'(c) = 0$. But

$$h'(c) = f'(c) - g'(c)$$

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

since the graph of g is a line with slope equal to $\frac{f(b) - f(a)}{b - a}$.

We have shown that

$$0 = f'(c) - \frac{f(b) - f(a)}{b - a}$$

so

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

exactly as required.

7.2 Applications of the Mean Value Theorem

In this section, we present some direct consequences of the Mean Value Theorem.

7.2.1 Antiderivatives

We have already developed a number of techniques for calculating derivatives. In this section, we will see how we can sometimes “undo” differentiation. That is, given a function f , we will look for a new function F with the property that $F'(x) = f(x)$.

DEFINITION

Antiderivative

Given a function f , an *antiderivative* is a function F such that

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

If $F'(x) = f(x)$ for all x in an interval I , we say that F is an antiderivative for f on I .

EXAMPLE 1

Let $f(x) = x^2$. Let $F(x) = \frac{x^3}{3}$. Then

$$F'(x) = \frac{3x^{3-1}}{3} = x^2 = f(x),$$

so $F(x) = \frac{x^3}{3}$ is an antiderivative of $f(x) = x^2$. ◀

While the derivative of a function is always unique, this is *not* true of antiderivatives. In the previous example, if we let $G(x) = \frac{x^3}{3} + 2$, then we find that $G'(x) = x^2$. Therefore, both $F(x) = \frac{x^3}{3}$ and $G(x) = \frac{x^3}{3} + 2$ are antiderivatives of the same function $f(x) = x^2$.

This holds in greater generality as we shall soon see. That is, if F is an antiderivative of a given function f , then so is $G(x) = F(x) + C$ for every $C \in \mathbb{R}$. A question naturally arises – are these all of the antiderivatives of f ?

We will start with a very simple, yet very important result.

Recall that if a function f is constant on an open interval I , then $f'(x) = 0$ for all $x \in I$. The first application of the Mean Value Theorem that we will consider is the converse of this statement.

THEOREM 3**The Constant Function Theorem**

Assume that $f'(x) = 0$ for all x in an interval I , then there exists an α such that $f(x) = \alpha$ for every $x \in I$.

PROOF

Let x_1 be any point in I . Let $\alpha = f(x_1)$. Now choose any other x_2 in I . Since f is differentiable on all of I , it is also continuous. It follows that the Mean Value Theorem holds on the closed interval with endpoints x_1 and x_2 . This means that there exists a c between x_1 and x_2 such that

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

But $f'(c) = 0$, so we have

$$0 = \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

and hence

$$f(x_2) = f(x_1) = \alpha.$$

Since this holds for any $x_i \in I$, we have shown

$$f(x) = \alpha$$

for every $x \in I$. ■

EXAMPLE 2

We have seen that the function $f(x) = e^x$ has the unusual property that $f'(x) = f(x)$. This property also holds for the function $f_1(x) = Ce^x$ for any constant $C \in \mathbb{R}$. In this example we will see that these are the only functions with this property.

Assume that g is such that $g'(x) = g(x)$ for every $x \in \mathbb{R}$. Construct a new function by letting

$$h(x) = \frac{g(x)}{e^x}.$$

Differentiate h using the quotient rule to get

$$\begin{aligned} h'(x) &= \frac{e^x g'(x) - \frac{d}{dx}(e^x)g(x)}{(e^x)^2} \\ &= \frac{e^x g(x) - e^x g(x)}{e^{2x}} \\ &= 0 \end{aligned}$$

since $g'(x) = g(x)$ and $\frac{d}{dx}(e^x) = e^x$.

Then since $h'(x) = 0$ for all $x \in \mathbb{R}$, $h(x)$ is constant. Let $h(x) = C$ for all $x \in \mathbb{R}$. Then $g(x) = Ce^x$ for all $x \in \mathbb{R}$. ◀

REMARK

The Constant Function Theorem tells us that the family of all antiderivatives of the function $f(x) = 0$ consists of all constant functions. That is, functions of the form

$$F(x) = C$$

for some constant $C \in \mathbb{R}$. ◀

The next application of the Mean Value Theorem is a small variant of the first. It plays a very important role in the development of the theory of *integration*. In particular, it shows us that any two antiderivatives of the same function must differ only by a constant. Therefore, to find all of the antiderivatives of a given function it suffices to find just one.

THEOREM 4**The Antiderivative Theorem**

Assume that $f'(x) = g'(x)$ for all $x \in I$. Then there exists an α such that

$$f(x) = g(x) + \alpha$$

for every $x \in I$.

PROOF

To see that this theorem is true, consider the function

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

Since

$$h'(x) = f'(x) - g'(x) = 0$$

for every $x \in I$, the Constant Function Theorem tells us that there exists an α such that

$$h(x) = f(x) - g(x) = \alpha$$

for every $x \in I$. This means that

$$f(x) = g(x) + \alpha$$

for every $x \in I$. ■

Leibniz Notation:

We will denote the *family of antiderivatives* of a function f by

$$\int f(x) dx.$$

For example,

$$\int x^2 dx = \frac{x^3}{3} + C.$$

The symbol

$$\int f(x) dx$$

is called the *indefinite integral of f* and $f(x)$ is called the *integrand*.

Finding antiderivatives is generally much more difficult than differentiating. For example, if $f(x) = e^{x^2}$, then we can easily differentiate f using the Chain Rule to get

$$\begin{aligned} f'(x) &= e^{x^2} \frac{d}{dx}(x^2) \\ &= 2xe^{x^2}. \end{aligned}$$

However, it is not at all obvious how to find

$$\int e^{x^2} dx.$$

In fact, using sophisticated techniques from algebra, it is possible to prove that there is no “nice” function that we can identify as an antiderivative of e^{x^2} .

At this point, we will be content to find the antiderivatives of many of the basic functions that are used in this course. The next theorem tells us how to find the antiderivatives of one of the most important classes of functions, the powers of x . This will allow us to find antiderivatives for any polynomial.

THEOREM 5**Power Rule for Antiderivatives**

If $\alpha \neq -1$, then

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

To check that this theorem is correct we need only differentiate. Since

$$\frac{d}{dx} \left(\frac{x^{\alpha+1}}{\alpha+1} + C \right) = x^\alpha,$$

we have found all of the antiderivatives of x^α .

REMARK

Suppose that F is an antiderivative of f and that G is an antiderivative of g . Then for any real numbers α and β we have

$$\frac{d}{dx} (\alpha F(x) + \beta G(x)) = \alpha f(x) + \beta g(x).$$

As a consequence of this observation and the Power Rule for Antiderivatives, we get that

$$\int a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \, dx = C + a_0 x + \frac{a_1}{2} x^2 + \frac{a_2}{3} x^3 + \cdots + \frac{a_n}{n+1} x^{n+1}$$

for any polynomial $p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$. ◀

The next example identifies the antiderivatives of several basic functions. You can use differentiation to verify each one.

EXAMPLE 3

1)

$$\int \frac{1}{x} \, dx = \ln(|x|) + C.$$

2)

$$\int e^x \, dx = e^x + C.$$

3)

$$\int a^x \, dx = \frac{a^x}{\ln(a)} + C.$$

4)

$$\int \sin(x) \, dx = -\cos(x) + C.$$

5)

$$\int \cos(x) \, dx = \sin(x) + C.$$

6)

$$\int \sec^2(x) \, dx = \tan(x) + C.$$

7)

$$\int \frac{1}{1+x^2} dx = \arctan(x) + C.$$

8)

$$\int \frac{1}{\sqrt{1-x^2}} dx = \arcsin(x) + C.$$

9)

$$\int \frac{-1}{\sqrt{1-x^2}} dx = \arccos(x) + C.$$



7.2.2 Increasing Function Theorem

Assume that

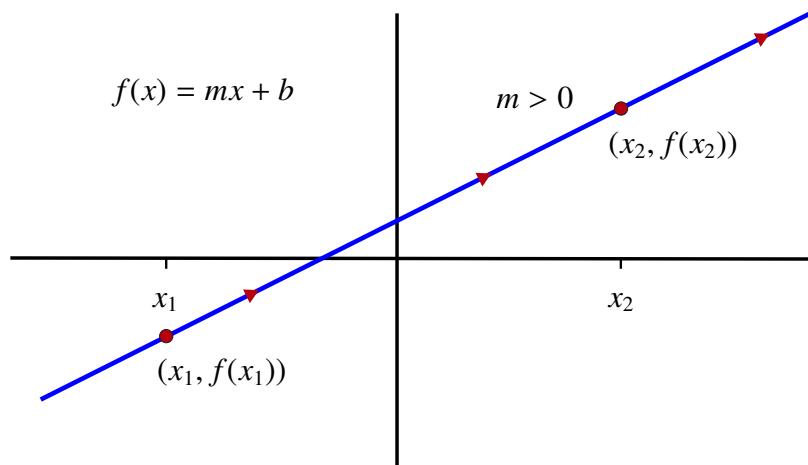
$$f(x) = mx + b.$$

If $m > 0$, the graph of the function slopes upwards as we move from left to right. In other words, if $x_1 < x_2$, then

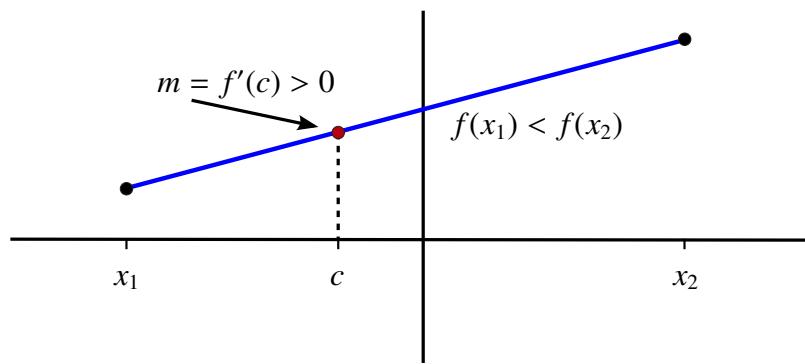
$$f(x_1) = mx_1 + b < mx_2 + b = f(x_2).$$

We have already seen that if a function is differentiable at a point it can be approximated by its tangent line. Consequently, it would make sense to suggest that if a function f was such that $f'(x) > 0$ at every point in an interval I then it should also be the case that if $x_1, x_2 \in I$ and $x_1 < x_2$, then we would expect that

$$f(x_1) < f(x_2).$$



In this section we will see that the Mean Value Theorem can be used to show that this is in fact the case. In fact, to see why this is the case for any differentiable function f we start by choosing two points $x_1 < x_2$ in I . If f is differentiable on I , then the Mean Value Theorem holds for the closed interval $[x_1, x_2]$.



Using the MVT, we get that there exists a c between x_1 and x_2 such that

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} = f'(c) > 0.$$

Since $x_2 - x_1 > 0$, we have

$$f(x_2) - f(x_1) > 0$$

or equivalently that

$$f(x_2) > f(x_1).$$

However, this is exactly what it means for f to be increasing on I .

A similar argument shows that if f is such that $f'(x) < 0$ for every x in an interval I , then f is *decreasing* on I . We can summarize this in the following important theorem.

THEOREM 6

The Increasing/Decreasing Function Theorem

- i) Let I be an interval and assume that $f'(x) > 0$ for all $x \in I$. If $x_1 < x_2$ are two points in I , then

$$f(x_1) < f(x_2).$$

That is, f is increasing on I .

- ii) Let I be an interval and assume that $f'(x) \geq 0$ for all $x \in I$. If $x_1 < x_2$ are two points in I , then

$$f(x_1) \leq f(x_2).$$

That is, f is non-decreasing on I .

- iii) Let I be an interval and assume that $f'(x) < 0$ for all $x \in I$. If $x_1 < x_2$ are two points in I , then

$$f(x_1) > f(x_2).$$

That is, f is decreasing on I .

- iv) Let I be an interval and assume that $f'(x) \leq 0$ for all $x \in I$. If $x_1 < x_2$ are two points in I , then

$$f(x_1) \geq f(x_2).$$

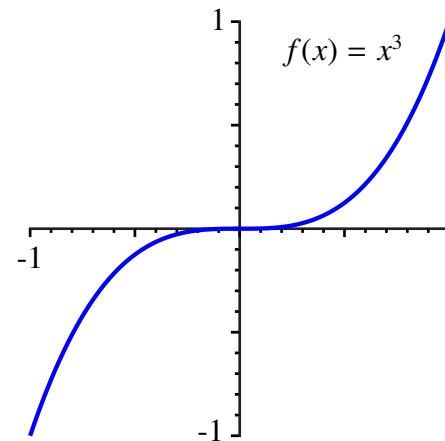
That is, f is non-increasing on I .

Question: If f is increasing and differentiable on $I = (a, b)$, must $f'(x) > 0$ for all $x \in I$?

It is easy to show that $f'(x) \geq 0$ (why?). However, the following example shows that the strict inequality, $f'(x) > 0$, is not always required.

Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be defined by $f(x) = x^3$. This function is increasing and differentiable on \mathbb{R} with $f'(x) = 3x^2$, and $f'(0) = 0$.

EXAMPLE 4



7.2.3 Functions with Bounded Derivatives

We motivated the introduction of the Mean Value Theorem by considering a problem involving a car that traveled a distance of 110 km in exactly one hour. Our task was to show that at some point the car exceeded the posted speed limit of 100 km/hr. Suppose we considered a different question:

Problem: If a car travels along a road and never exceeds a speed of 100 km/hr, what is the maximum distance that the car could travel in 1 hour?

Intuitively, the answer to this problem should be 100 km. If so, then the car in our original scenario could not have completed the 110 km trip without speeding. We will now show that the MVT can again be used to verify our intuition by giving us a direct relationship between the magnitude of the derivative and how much a function could possibly change over a given interval.

Observation: Let's assume that f is continuous on $[a, b]$ and is differentiable on (a, b) . Assume also that

$$m \leq f'(x) \leq M$$

for each $x \in (a, b)$. Pick some $x \in [a, b]$. The Mean Value Theorem is true on the interval $[a, x]$. This means that there exists a c between a and x such that

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

Since $m \leq f'(x) \leq M$, we get that

$$m \leq \frac{f(x) - f(a)}{x - a} \leq M.$$

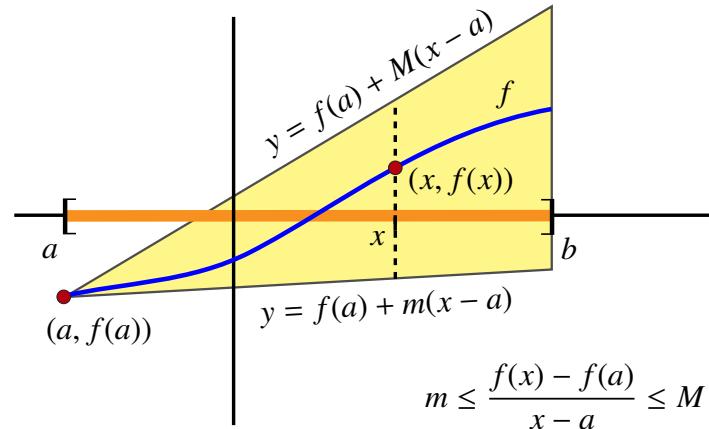
Geometrically this means that the slope of the secant line joining $(a, f(a))$ and $(x, f(x))$ has a value that sits between the maximum and minimum values of the derivative on the interval. Moreover, since $x - a > 0$, we get that

$$m(x - a) \leq f(x) - f(a) \leq M(x - a)$$

or equivalently that

$$f(a) + m(x - a) \leq f(x) \leq f(a) + M(x - a).$$

We have shown that the graph of f sits between the lines
 $y = f(a) + m(x - a)$ and
 $y = f(a) + M(x - a)$.



THEOREM 7

The Bounded Derivative Theorem

Assume that f is continuous on $[a, b]$ and differentiable on (a, b) with

$$m \leq f'(x) \leq M$$

for each $x \in (a, b)$. Then

$$f(a) + m(x - a) \leq f(x) \leq f(a) + M(x - a)$$

for all $x \in [a, b]$.

EXAMPLE 5 Assume that $f(0) = 3$ and that $1 \leq f'(x) \leq 2$ for all $x \in [0, 1]$. Show that

$$4 \leq f(1) \leq 5.$$

We know from the Mean Value Theorem that

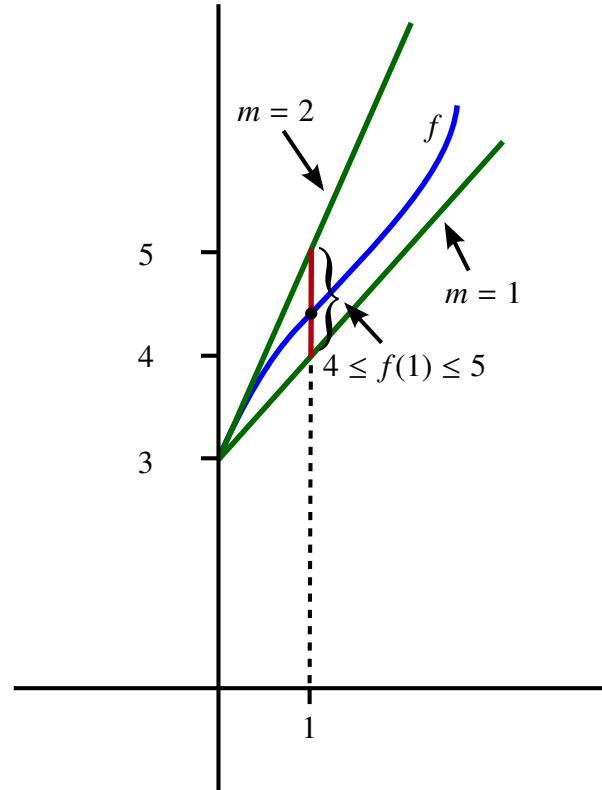
$$1 \leq \frac{f(1) - f(0)}{1 - 0} \leq 2.$$

Then

$$1 \leq f(1) - 3 \leq 2$$

and hence that

$$4 \leq f(1) \leq 5.$$

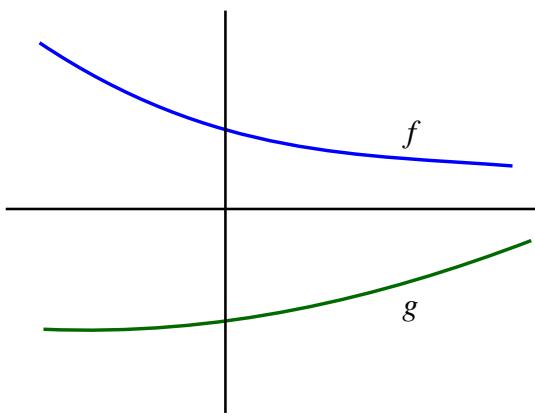


REMARK

If we return to the scenario of a car traveling one hour along a road without exceeding a speed of 100 km/hr, then the previous theorem tells us immediately that the maximum distance the car could have traveled in that time frame was in fact 100 km as we expected. ◀

7.2.4 Comparing Functions Using Their Derivatives

If we consider two functions, f and g , so that $f'(x) \leq g'(x)$ on an interval I , we cannot conclude that $f(x) \leq g(x)$. In fact, in the following diagram, notice that $f'(x) < 0$ (slopes of the tangent lines are negative) while $g'(x) > 0$ (slopes of the tangent lines are positive), yet $g(x) < f(x)$.



However, if we know that $f(a) = g(a)$ and that $f'(x) \leq g'(x)$, then we can conclude that $f(x) \leq g(x)$ for $x > a$, and also that $g(x) \leq f(x)$ for $x < a$. From these conclusions we will be able to derive some interesting inequalities.

THEOREM 8

Assume that f and g are continuous at $x = a$ with $f(a) = g(a)$.

- i) If both f and g are differentiable for $x > a$ and if $f'(x) \leq g'(x)$ for all $x > a$, then

$$f(x) \leq g(x)$$

for all $x > a$.

- ii) If both f and g are differentiable for $x < a$ and if $f'(x) \leq g'(x)$ for all $x < a$, then

$$f(x) \geq g(x)$$

for all $x < a$.

PROOF

- i) We will assume that f and g are continuous at $x = a$ with $f(a) = g(a)$, f and g are differentiable for $x > a$, and $f'(x) \leq g'(x)$ for all $x > a$. Let's build a new function

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

Then h is continuous at $x = a$ and differentiable for $x > a$ with

$$h'(x) = g'(x) - f'(x) \geq 0$$

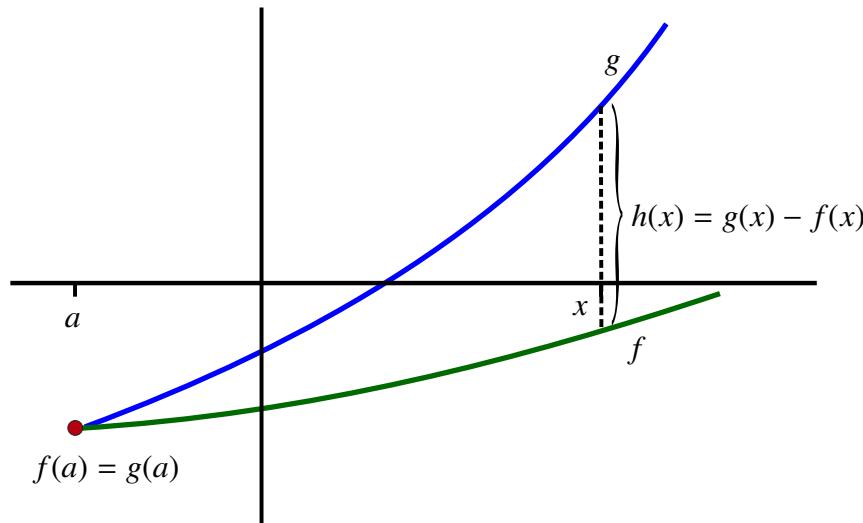
for all $x > a$. So if $x > a$, by the Mean Value Theorem, we can find $a < c < x$ so that

$$0 \leq h'(c) = \frac{h(x) - h(a)}{x - a}.$$

We know that $h(a) = 0$ and that $x - a > 0$, so this tells us that

$$h(x) = g(x) - f(x) \geq 0$$

exactly as we had hoped.



ii) The proof of part (ii) is similar. ■

REMARK

In the previous theorem, if we replace $f'(x) \leq g'(x)$ with the strict inequality $f'(x) < g'(x)$, then we can also show that $f(x) < g(x)$ if $x > a$ and that $f(x) > g(x)$ if $x < a$. ◀

EXAMPLE 6 We will now use what we have just learned to help us establish the following fundamental limit:

$$\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = e$$

To do so we first show that

$$x - \frac{1}{2}x^2 < \ln(1 + x) < x \quad (*)$$

for all $x > 0$. To see this let

$$\begin{aligned} f(x) &= x - \frac{1}{2}x^2 \\ g(x) &= \ln(1 + x), \text{ and} \\ h(x) &= x. \end{aligned}$$

Then

$$0 = f(0) = g(0) = h(0).$$

Moreover,

$$f'(x) = 1 - x,$$

$$g'(x) = \frac{1}{1+x}, \text{ and}$$

$$h'(x) = 1.$$

Therefore if $x > 0$, then

$$g'(x) = \frac{1}{1+x} < 1 = h'(x).$$

We also know that for any $x > 0$

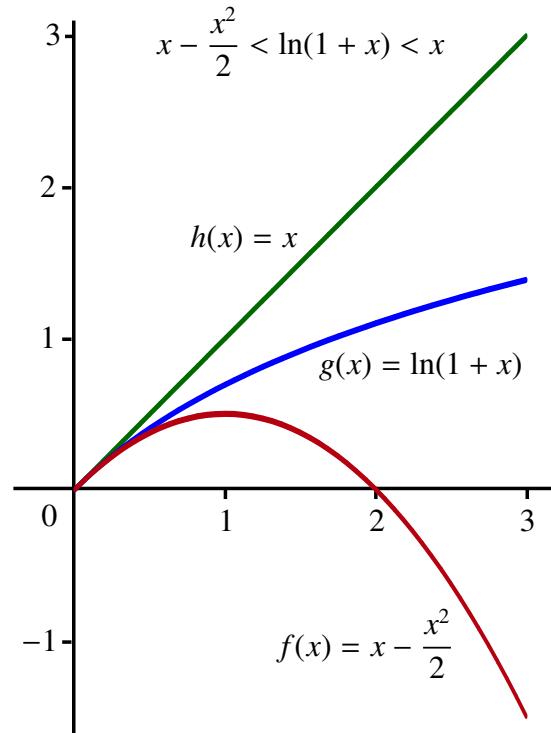
$$(1-x)(1+x) = 1 - x^2 < 1$$

so that if $x > 0$

$$f'(x) = 1 - x < \frac{1}{1+x} = g'(x).$$

It follows that for $x > 0$ we have

$$f'(x) < g'(x) < h'(x).$$



Applying the previous theorem twice gives us the inequality (*).

The next observation we can make is that if $x > 0$, we can divide all three terms in inequality (*) by x to get

$$1 - \frac{1}{2}x < \frac{\ln(1+x)}{x} < 1. \quad (**)$$

In particular, if $x = \frac{1}{n}$, we have

$$1 - \frac{1}{2n} < \frac{\ln\left(1 + \frac{1}{n}\right)}{\frac{1}{n}} = n \ln\left(1 + \frac{1}{n}\right) = \ln\left(1 + \frac{1}{n}\right)^n < 1. \quad (***)$$

Applying the Squeeze Theorem to (***) gives us that

$$\lim_{n \rightarrow \infty} \ln\left(1 + \frac{1}{n}\right)^n = 1.$$

Finally, since e^x is a continuous function, the Sequential Characterization of Continuity gives us that

$$e = e^1 = \lim_{n \rightarrow \infty} e^{\ln(1 + \frac{1}{n})^n} = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n.$$



The previous example can be modified to show the following:

THEOREM 9

Let $\alpha \in \mathbb{R}$. Then

$$e^\alpha = \lim_{n \rightarrow \infty} \left(1 + \frac{\alpha}{n}\right)^n.$$

7.2.5 Interpreting the Second Derivative

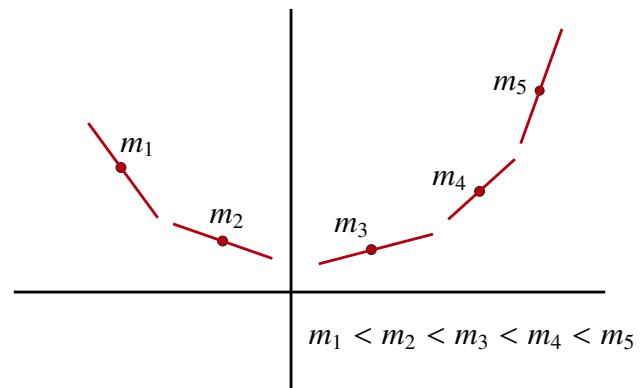
Since

$$f'' = \frac{d}{dx}(f')$$

$f''(x)$ represents the *instantaneous rate of change of $f'(x)$* . For example, if we let $s(t)$ denote the *displacement* of an object, we have already seen that $v(t) = s'(t)$ represents the *velocity* of the object. Then $v'(t) = s''(t)$ is the rate of change of velocity. That is, $s''(t) = a(t)$ is the *acceleration* of the object at time t .

Geometrically, $f'(x)$ represents the slope of the tangent line to the graph of f . Therefore, f'' measures *how quickly these slopes are changing*.

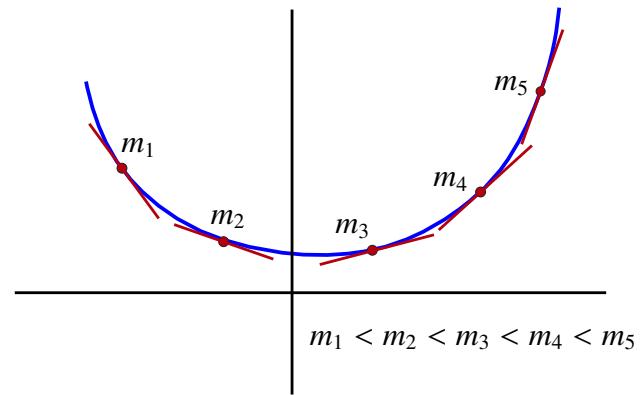
Increasing slopes correspond to the “counter-clockwise” rotations of the tangent lines.



We know that f' increases when

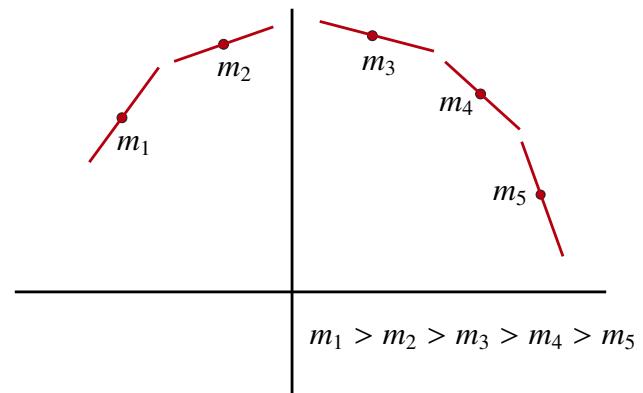
$$f''(x) > 0.$$

It follows that when $f''(x) > 0$ the graph of f has the following distinctive shape similar to a cup opening upwards.



In this case, we say that the graph of f is *concave upwards*.

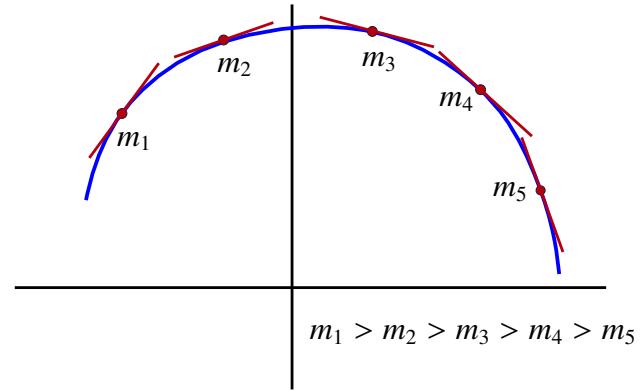
Similarly, decreasing slopes correspond to “clockwise” rotations of the tangent lines.



We know that f' decreases when

$$f''(x) < 0.$$

It follows that when $f''(x) < 0$, the graph of f has a shape similar to a cup opening downwards.



In this case, we say that the graph of f is *concave downwards*.

7.2.6 Formal Definition of Concavity

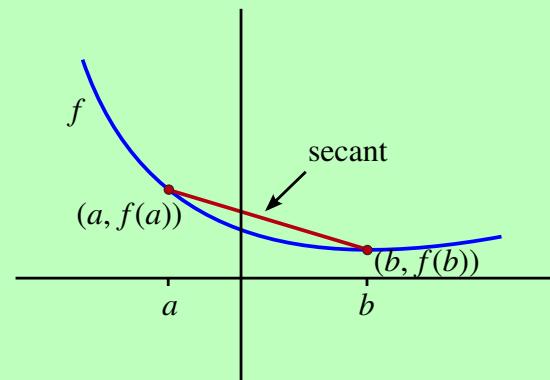
The previous section provided an informal definition of what it means for the graph of a function to be *concave upwards* or *concave downwards* on an interval. A more precise definition of *concavity* is now presented.

Recall that if $(a, f(a))$ and $(b, f(b))$ are two points on the graph of a function f , then the line segment joining these two points is called a *secant* to the graph of f .

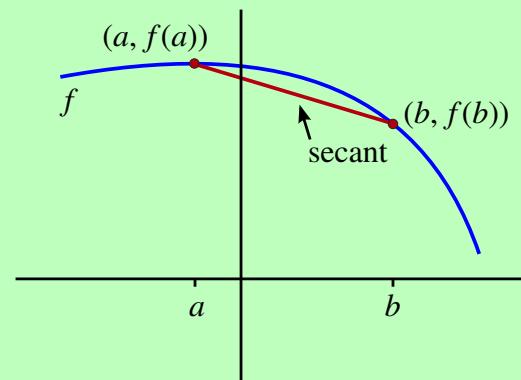
DEFINITION

Concavity

The graph of f is *concave upwards* on an interval I if for every pair of points a and b in I , the secant line joining $(a, f(a))$ and $(b, f(b))$ lies *above* the graph of f .



The graph of f is *concave downwards* on an interval I if for every pair of points a and b in I , the secant line joining $(a, f(a))$ and $(b, f(b))$ lies *below* the graph of f .



The next theorem summarizes what we have already observed about the relationship between concavity and the second derivative.

THEOREM 10

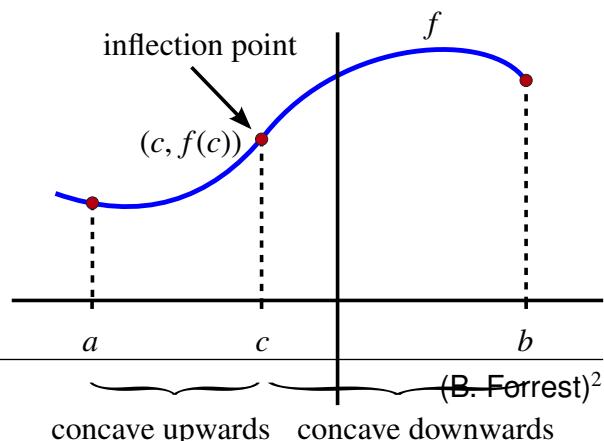
Second Derivative Test for Concavity

- i) If $f''(x) > 0$ for each x in an interval I , then the graph of f is concave upwards on I .
- ii) If $f''(x) < 0$ for each x in an interval I , then the graph of f is concave downwards on I .

EXAMPLE 7

In the diagram, the graph of the function f is concave upwards on the interval $[a, c]$ and concave downwards on $[c, b]$.

We say that f *changes its concavity* at $x = c$. The point $(c, f(c))$ is called an *inflection point*.



DEFINITION**Inflection Point**

A point $(c, f(c))$ is called an *inflection point* for the function f if

- i) f is continuous at $x = c$, and
- ii) the concavity of f changes at $x = c$.

Observation: Typically an inflection point at $x = c$ would occur when the second derivative changes from positive to negative, or vice versa. If f'' is continuous, the Intermediate Value Theorem requires that $f''(c) = 0$.

THEOREM 11**Test for Inflection Points**

If f'' is continuous at $x = c$ and $(c, f(c))$ is an inflection point for f , then $f''(c) = 0$.

WARNING This theorem shows us how to *locate candidates* for inflection points. However, $f''(c) = 0$ does *not* mean that an inflection point always occurs when $x = c$.

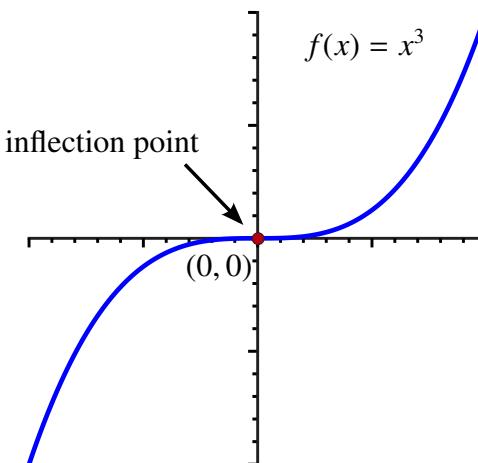
EXAMPLE 8

Let $f(x) = x^3$. Then $f'(x) = 3x^2$ and $f''(x) = 6x$. To find all possible candidates for an inflection point we solve

$$f''(x) = 6x = 0.$$

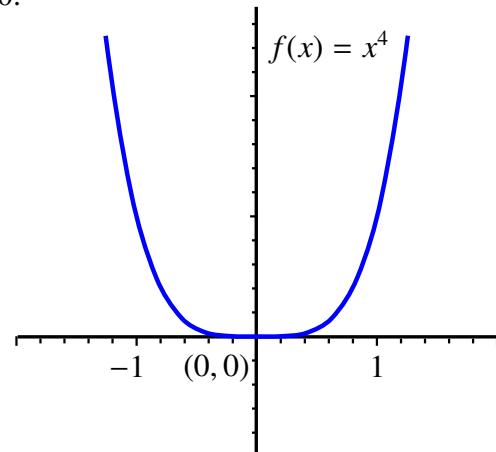
The only solution is $x = 0$. Therefore, $x = 0$ is a *candidate* for the location of a point of inflection for the function $f(x) = x^3$. To confirm whether f does indeed have a point of inflection at $x = 0$, we must check that the concavity of f changes from the interval $x < 0$ to the interval $x > 0$.

Since $f''(x) = 6x$, we know that if $x < 0$, then $f''(x) < 0$, so the graph of f is concave downwards on the interval $(-\infty, 0)$. On the other hand, $f''(x) > 0$ when $x > 0$, so the graph of f is concave upwards on the interval $(0, \infty)$. This shows that the concavity of f does indeed change at $x = 0$. Since f is clearly continuous at $x = 0$, we can now conclude that $(0, 0)$ is an inflection point.



EXAMPLE 9 Let $f(x) = x^4$. Then $f'(x) = 4x^3$ and $f''(x) = 12x^2$. To find all possible candidates for an inflection point, solve $f''(x) = 12x^2 = 0$.

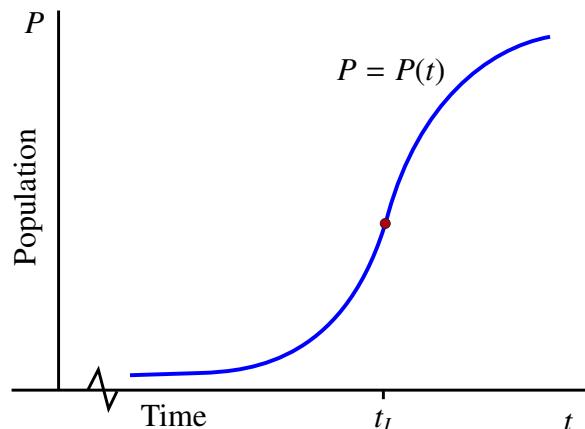
Once again, the only solution is $x = 0$. However, in this case, the second derivative does not change sign at $x = 0$. In fact, $f''(x) = 12x^2 \geq 0$ for all x . In particular, the graph of f is concave upwards on the interval $[-1, 1]$. (Note: f is actually concave upwards on all of \mathbb{R} .) As such, $(0, 0)$ is *not* an inflection point for $f(x)$ despite the fact that $f''(0) = 0$ because the concavity of f did not change.



We end this section with two applications. The first is an application in biology to population growth. The second is an analysis of customer satisfaction.

EXAMPLE 10

The diagram represents the graph of the population P of a particular bacteria over time t in a restricted environment.

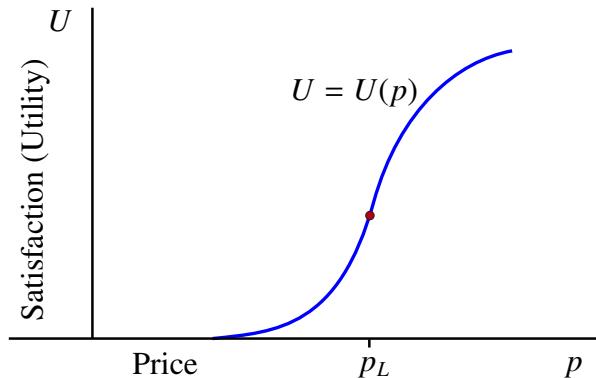


When the population is small, the number of bacteria available to multiply is very limited. As such, the *rate* at which the population increases is slow. As the population increases, so does the rate at which the bacteria multiply. Consequently, as long as the bacteria have ample food so that the death rate remains low, the overall population grows at an increasing rate. However, at some point, the bacteria will begin to exhaust the resources available for survival. At this point, the death rate due to starvation will increase and the growth rate in the overall population will start to slow down again.

Since the instantaneous rate of change in population is simply the derivative P' of P , and since we also know that the larger $P'(t)$ is the steeper the graph will be, the slow down in population growth occurs at the instant that the graph is at its steepest. However, on the graph this occurs at time t_I . Moreover, $(t_I, P(t_I))$ is an inflection point for the bacteria population versus time graph. At this point the scarcity of resources starts to restrict the rate at which the population can increase until it eventually flattens out due to mass starvation.

EXAMPLE 11

The diagram represents a graph of consumer satisfaction U (also known as utility), versus the price p spent on a type of good, such as an automobile.



At very low prices, this product is of very poor build quality and has few features. As the price rises, so does quality. The feature set available in the product improves with an increase in price. However, there is a phenomena called the “law of diminishing returns” that takes effect at some price point. This law says that after some price point the gain in satisfaction for each additional dollar spent will start to decrease.

Since the instantaneous rate of change in satisfaction per unit change in price is simply the derivative U' of U , and since we also know that the larger $U'(p)$ is the steeper the graph will be, “the law of diminishing returns” starts to take effect at the point at which the graph of U is at its steepest. On the graph, this occurs at price p_L . Moreover, $(p_L, U(p_L))$ is an inflection point for the utility versus price graph. This point represents the price at which each additional dollar spent on the product returns less in additional customer satisfaction. ◀

REMARK

You will notice that the graphs in the two previous example have a very similar shape. This is no accident. Both are examples of a phenomenon known as *Logistic Growth*. In both cases the inflection point represents the place where resources start to have a diminishing impact on the rate at which the quantity increases. ◀

7.2.7 Classifying Critical Points: The First and Second Derivative Tests

Previously we saw that if $x = c$ was either a local maximum or a local minimum for a function f , then either $f'(c) = 0$ or $f'(c)$ did not exist. That is, c is a *critical point*. However, we have also shown that *not all critical points are local extrema*. In this section we will look at how to determine if a critical point c is either a local maximum or a local minimum. We will present two such methods for solving this problem.

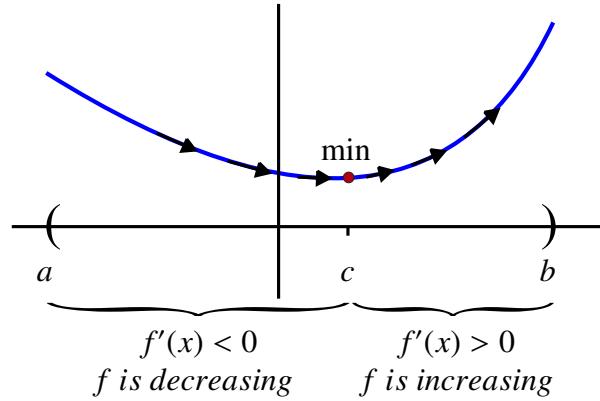
Method 1: Finding Maxima and Minima using the First Derivative Test

Assume $x = c$ is a critical point for f . As the name suggests, the first method for testing critical points involves a careful examination of the first derivative near c .

Assume that $a < c < b$ and that f is continuous at c . Let's also assume that

$$\begin{aligned} f'(x) &< 0 && \text{for all } x \in (a, c) \\ \text{and } f'(x) &> 0 && \text{for all } x \in (c, b). \end{aligned}$$

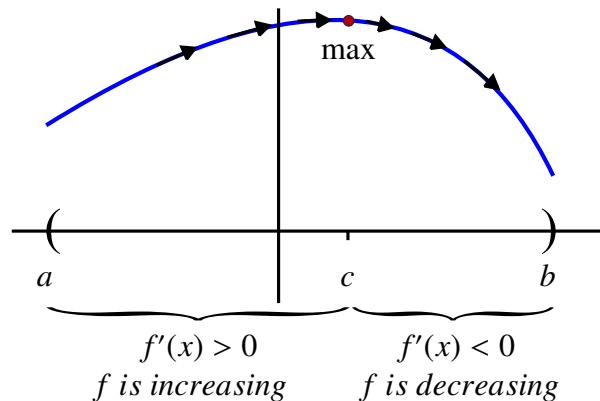
Since $f'(x) < 0$ on (a, c) , the function f is decreasing on (a, c) . We also know that if $f'(x) > 0$ on (c, b) , then f is increasing on that interval. This means that the function f *decreases* as we approach c from the *left* and f *increases* as we move away from c to the *right*. As the graph shows, this suggests that f has a local minimum at $x = c$.



Assume again that $a < c < b$ and that f is continuous at c . However, this time we will assume that

$$\begin{aligned} f'(x) &> 0 && \text{for all } x \in (a, c) \\ \text{and } f'(x) &< 0 && \text{for all } x \in (c, b). \end{aligned}$$

Since $f'(x) > 0$ on (a, c) , f is increasing on (a, c) . Also, since $f'(x) < 0$ on (c, b) , f is decreasing on (c, b) . In this case, f *increases* as we approach c from the *left* and f *decreases* as we move away from c to the *right*. Thus, f has a local maximum at $x = c$.



These observations are summarized in the following theorem.

THEOREM 12

First Derivative Test

Assume that c is a critical point of f , and f is continuous at c .

- i) If there is an interval (a, b) containing c such that

$$\begin{aligned} f'(x) &< 0 && \text{for all } x \in (a, c) \\ \text{and } f'(x) &> 0 && \text{for all } x \in (c, b), \end{aligned}$$

then f has a local minimum at c .

ii) If there is an interval (a, b) containing c such that

$$\begin{aligned} f'(x) &> 0 & \text{for all } x \in (a, c) \\ \text{and } f'(x) &< 0 & \text{for all } x \in (c, b), \end{aligned}$$

then f has a local maximum at c .

EXAMPLE 12 Find all of the critical points of the function

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

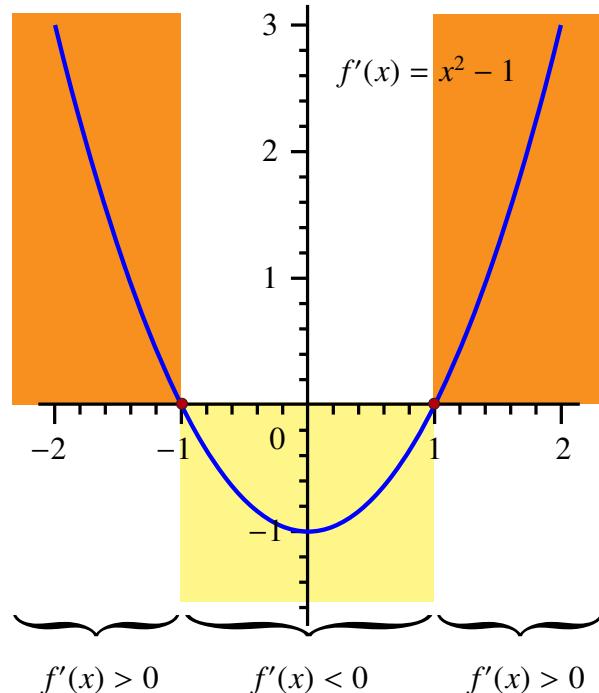
For each critical point, determine if it is a local maximum, a local minimum, or neither. Finally, graph f .

SOLUTION To find the critical points we need to solve the equation $f'(x) = 0$. But $f'(x) = x^2 - 1$. Hence $f'(x) = 0$ when $x = 1$ or $x = -1$. Since the function is a polynomial, it is always differentiable, so $x = 1$ and $x = -1$ are the only critical points.

We will use the First Derivative Test to determine the nature of the critical points. To apply the test we need to know about the sign of $f'(x)$ around the critical points.

Since $f'(x) = x^2 - 1$ is a second degree polynomial with a positive coefficient on its highest degree term, its graph is a parabola that opens upwards. Moreover, the only sign changes occur when $f'(x) = 0$ at $x = 1$ and $x = -1$.

We have that $f'(x) > 0$ if $x > 1$, $f'(x) < 0$ if $x \in (-1, 1)$, and $f'(x) > 0$ if $x < -1$.



Consider the critical point $x = 1$. We have just seen that $f'(x) < 0$ if $x \in (-1, 1)$, and $f'(x) > 0$ if $x > 1$. This means that f decreases as we approach $x = 1$ from the left and increases as we move away from $x = 1$ to the right. The First Derivative Test tells us that these conditions indicate a local *minimum*.

For $x = -1$ we know that $f'(x) > 0$ if $x < -1$, and $f'(x) < 0$ if $x \in (-1, 1)$. In this case, the First Derivative Test tells us that $x = -1$ is a local *maximum*.

To graph f we note that

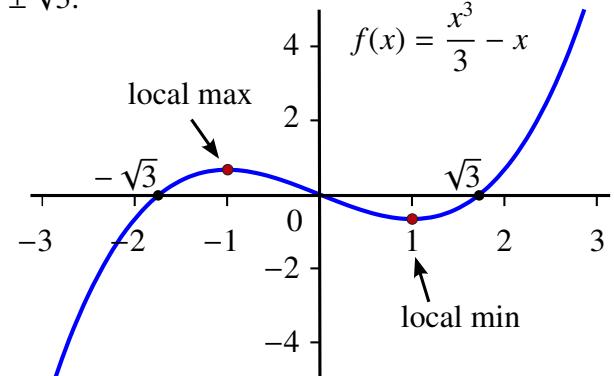
$$\lim_{x \rightarrow -\infty} \frac{x^3}{3} - x = -\infty \quad \text{and} \quad \lim_{x \rightarrow \infty} \frac{x^3}{3} - x = \infty.$$

Moreover, since

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

we have $f(x) = 0$ when $x = 0$ or $x = \pm\sqrt{3}$.

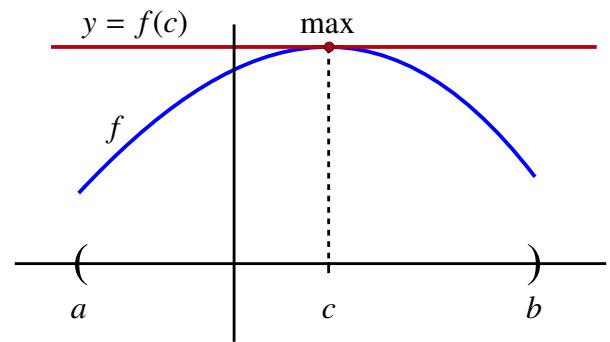
Combining this information with the results about the critical points gives us the following graph.



Method 2: Finding Maxima and Minima using the Second Derivative Test

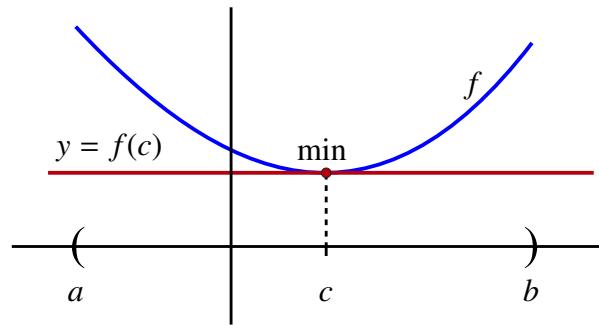
Assume that c is a critical point for f with $f'(c) = 0$. This means that the tangent line to the graph of f through $(c, f(c))$ is the *horizontal line* $y = f(c)$.

Suppose now that there is an open interval (a, b) containing c such that $f''(x) < 0$ for all $x \in (a, b)$. (This will happen, for example, if f'' is continuous at c and $f''(c) < 0$.) Then the graph of f is *concave downwards* on (a, b) . Therefore, the tangent line at $x = c$ sits *above the graph* as shown.



Our conclusion is that f has a *local maximum* at $x = c$.

If, instead of $f''(x) < 0$ for all $x \in (a, b)$, we assume that $f''(x) > 0$ for all $x \in (a, b)$, then the function is *concave upwards*. In this case, the tangent line sits *below the graph*.



This time we have that f has a *local minimum* at $x = c$.

We have just outlined a test for the nature of a critical point called the **Second Derivative Test**. The precise statement is as follows:

THEOREM 13

Second Derivative Test

Assume that $f'(c) = 0$ and that f'' is continuous at $x = c$.

- i) If $f''(c) < 0$, then f has a local maximum at c .
- ii) If $f''(c) > 0$, then f has a local minimum at c .

Let's revisit the previous example.

EXAMPLE 13

Let $f(x) = \frac{x^3}{3} - x$. We have seen that the critical points occur at $x = \pm 1$ since $f'(x) = x^2 - 1$. The first derivative test showed us that for f a local maximum occurs at $x = -1$ and a local minimum occurs at $x = 1$.

We can confirm this result again by applying the Second Derivative Test. In fact, $f''(x) = 2x$. Hence, $f''(-1) = -2 < 0$ so the test shows that f has a local maximum that occurs at $x = -1$. We also have that $f''(1) = 2 > 0$ so f has a local minimum at $x = 1$, exactly as we had concluded previously. ◀

7.2.8 Finding Maxima and Minima on $[a, b]$

Let's summarize what we have learned about local extrema for a continuous function f on a closed interval $[a, b]$.

The Extreme Value Theorem guarantees the existence of both a global maximum and a global minimum for f . Assume that the global maximum occurs at a point $x = c$. Then either

$$(1) \quad c = a \text{ or } c = b$$

or

$$(2) \quad c \in (a, b).$$

In the second case, c is also a local maximum. Consequently, c is a critical point. Therefore, $f'(c) = 0$ or $f'(c)$ does not exist. A similar statement holds for the global minimum. This leads to a simple algorithm for finding the global maximum and global minimum of a function f .

Summary [Finding the Global Maximum and Global Minimum]

To find the maximum and minimum for a continuous function f on $[a, b]$:

- **Step 1:** Evaluate $f(a)$ and $f(b)$.
- **Step 2:** Find all critical points c in (a, b) such that $f'(c) = 0$ and $f'(c)$ does not exist, where applicable.
- **Step 3:** Evaluate the function at each of the critical points.
- **Step 4:** The global maximum is at the point that produces the *largest* Real value from Steps 1 and 3. The global minimum is at the point that produces the *smallest* Real value from Steps 1 and 3.

EXAMPLE 14

Find the maximum and minimum value for the function $f(x) = \frac{x^3}{3} - x$ on the interval $[-3, 2]$.

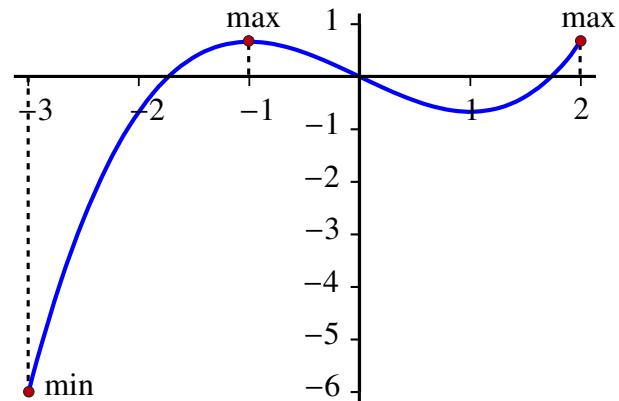
SOLUTION We first determine the values of the function $f(-3)$ and $f(2)$ at the endpoints of the interval. We have

$$f(-3) = -6 \quad \text{and} \quad f(2) = \frac{2}{3}.$$

We know that the only critical points in $[-3, 2]$ occur at $x = \pm 1$ since $f'(x) = x^2 - 1$. We now have

$$f(-1) = \frac{2}{3} \quad \text{and} \quad f(1) = -\frac{2}{3}.$$

Of the four values above, the maximum value of $f(x)$ is $\frac{2}{3}$ which occurs at both the interior critical point $x = -1$ and at the right-hand endpoint $x = 2$. The minimum value of $f(x)$ is -6 which occurs at the left-hand endpoint $x = -3$. The diagram confirms our analysis.



EXAMPLE 15 Find the maximum and minimum values for $f(x) = e^x + e^{-x}$ on the interval $[-1, 3]$. Sketch the graph of f on $[-1, 3]$.

SOLUTION We first evaluate the function at the endpoints to get

$$f(-1) = \frac{1}{e} + e \text{ and } f(3) = e^3 + \frac{1}{e^3}.$$

Next, we must find the critical points in the open interval $(-1, 3)$. Since f is differentiable, we need only solve

$$f'(x) = e^x - e^{-x} = 0.$$

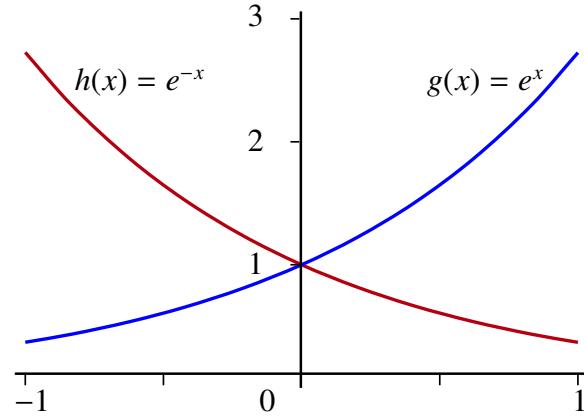
It may not be obvious how to solve this equation. To do so we use a trick and look at the graphs of $g(x) = e^x$ and $h(x) = e^{-x}$ simultaneously.

Since $f'(x) = g(x) - h(x)$, the solution to the equation

$$f'(x) = e^x - e^{-x} = 0$$

occurs when the graphs intersect.

This only occurs when $x = 0$.
Therefore, $x = 0$ is the only critical point.



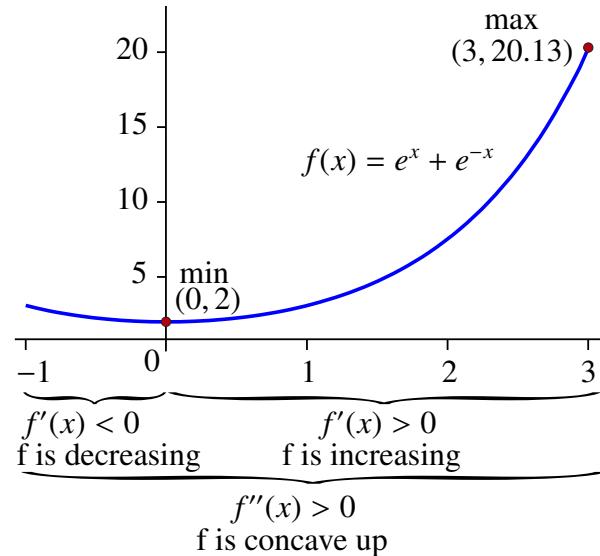
Now $f(0) = 1 + 1 = 2$. Moreover,

$$f(0) = 2 < f(-1) = \frac{1}{e} + e < f(3) = e^3 + \frac{1}{e^3}.$$

This means that the global minimum occurs at $x = 0$ and the minimum value on the interval is $f(0) = 2$. The global maximum is at $x = 3$ and its value is $f(3) = e^3 + \frac{1}{e^3}$ (or $f(3) \approx 20.13$).

In sketching the graph we can deduce from the previous diagram that $e^x < e^{-x}$ if $x < 0$, and $e^{-x} < e^x$ if $x > 0$. This means that $f'(x) < 0$ if $x < 0$, and $f'(x) > 0$ if $x > 0$. Therefore, the function is decreasing on $[-1, 0]$ and is increasing on $[0, 3]$.

We also note that since $f''(x) = e^x + e^{-x} > 0$ for all x , the graph is concave upwards.



7.3 L'Hôpital's Rule

In the section on limits, we saw that if we had a function $h(x) = \frac{f(x)}{g(x)}$ and if

$$\lim_{x \rightarrow a} f(x) = 0 = \lim_{x \rightarrow a} g(x),$$

then we could *not* say whether or not $\lim_{x \rightarrow a} h(x)$ exists. For this reason, we call such a situation an *indeterminate form of type $\frac{0}{0}$* .

Similarly, if

$$\lim_{x \rightarrow a} f(x) = \pm\infty = \lim_{x \rightarrow a} g(x),$$

we would not be able to determine immediately if the limit of the quotient exists. We call this situation an *indeterminate form of type $\frac{\infty}{\infty}$* .

L'Hôpital's Rule provides us with a tool for evaluating many of these indeterminate limits. To motivate the rule let's consider the following observation.

Observation: Let $h(x) = \frac{f(x)}{g(x)}$. Let's assume that

$$\lim_{x \rightarrow a} f(x) = 0 = \lim_{x \rightarrow a} g(x),$$

so that we have an indeterminate form of type $\frac{0}{0}$. Let's also assume that f and g have continuous derivatives with $g'(a) \neq 0$. We know from our work with linear approximations that for x near a we have that

$$\frac{f(x)}{g(x)} \cong \frac{f(a) + f'(a)(x - a)}{g(a) + g'(a)(x - a)} = \frac{f'(a)}{g'(a)}$$

since $f(a) = 0 = g(a)$. This might lead us to guess that if $\lim_{x \rightarrow a} \frac{f(x)}{g(x)}$ exists, then in fact

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \frac{f'(a)}{g'(a)}. \quad (*)$$

Moreover, since f' and g' are continuous with $g'(a) \neq 0$, we also have

$$\frac{f'(a)}{g'(a)} = \lim_{x \rightarrow a} \frac{f'(x)}{g'(x)}. \quad (**)$$

Combining $(*)$ and $(**)$ gives us

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \lim_{x \rightarrow a} \frac{f'(x)}{g'(x)}.$$

This leads us to the statement of L'Hôpital's Rule.

THEOREM 14**L'Hôpital's Rule**

Assume that $f'(x)$ and $g'(x)$ exist near $x = a$, $g'(x) \neq 0$ near $x = a$ except possibly at $x = a$, and that $\lim_{x \rightarrow a} \frac{f(x)}{g(x)}$ is an indeterminate form of type $\frac{0}{0}$ or $\frac{\infty}{\infty}$, then

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \lim_{x \rightarrow a} \frac{f'(x)}{g'(x)}$$

provided that the latter limit exists (or is ∞ or $-\infty$).

Moreover, this rule remains valid for one-sided limits and for limits at $\pm\infty$.

L'Hôpital's Rule can be derived from a rather sophisticated application of the Mean Value Theorem, further evidence of the importance of the MVT. However, the proof of L'Hôpital's Rule is beyond the scope of this discussion.

We will illustrate how the rule can be applied by looking at several examples.

EXAMPLE 16

Evaluate

$$\lim_{x \rightarrow 0} \frac{e^x - 1}{x}.$$

SOLUTION Let $f(x) = e^x - 1$ and $g(x) = x$. Then both functions are continuous and

$$\lim_{x \rightarrow 0} e^x - 1 = e^0 - 1 = 0 = \lim_{x \rightarrow 0} x.$$

Therefore, this is an indeterminate form of type $\frac{0}{0}$.

We also have that $f'(x) = e^x$ and $g'(x) = 1$. Hence, we have satisfied all of the conditions in the statement of L'Hôpital's Rule. Moreover,

$$\lim_{x \rightarrow 0} \frac{f'(x)}{g'(x)} = \lim_{x \rightarrow 0} \frac{e^x}{1} = 1$$

so L'Hôpital's Rule tells us that

$$\lim_{x \rightarrow 0} \frac{e^x - 1}{x} = \lim_{x \rightarrow 0} \frac{e^x}{1} = 1.$$

We can verify that this limit is correct by noting that

$$\lim_{x \rightarrow 0} \frac{e^x - 1}{x}$$

is by definition the derivative of the function $h(x) = e^x$ at the point $x = 0$. However, we know that $h'(x) = e^x$ so $h'(0) = e^0 = 1$, exactly as we expected. 

The next example is similar. However, it shows that we may need to apply L'Hôpital's Rule more than once to find the limit.

EXAMPLE 17 Evaluate

$$\lim_{x \rightarrow 0} \frac{e^{x^2} - \cos(x)}{x^2}.$$

SOLUTION Let $f(x) = e^{x^2} - \cos(x)$ and $g(x) = x^2$. We have

$$\begin{aligned}\lim_{x \rightarrow 0} f(x) &= \lim_{x \rightarrow 0} (e^{x^2} - \cos(x)) \\ &= e^0 - \cos(0) \\ &= 1 - 1 \\ &= 0\end{aligned}$$

and

$$\begin{aligned}\lim_{x \rightarrow 0} g(x) &= \lim_{x \rightarrow 0} x^2 \\ &= 0\end{aligned}$$

so the limit is indeterminate of type $\frac{0}{0}$.

Next we get that $f'(x) = 2xe^{x^2} + \sin(x)$ and $g'(x) = 2x$. It is easy to verify that

$$\lim_{x \rightarrow 0} (2xe^{x^2} + \sin(x)) = 0 = \lim_{x \rightarrow 0} 2x.$$

This means that we have another indeterminate form of type $\frac{0}{0}$, so we cannot yet determine the original limit. However, if we let $F(x) = 2xe^{x^2} + \sin(x)$ and $G(x) = 2x$, then we have all of the conditions satisfied to try and apply L'Hôpital's Rule again to find

$$\lim_{x \rightarrow 0} \frac{F(x)}{G(x)}.$$

This time, we have $F'(x) = 2e^{x^2} + 4x^2e^{x^2} + \cos(x)$ and $G'(x) = 2$. However,

$$\begin{aligned}\lim_{x \rightarrow 0} F'(x) &= \lim_{x \rightarrow 0} (2e^{x^2} + 4x^2e^{x^2} + \cos(x)) \\ &= 2e^0 + 4(0)e^0 + \cos(0) \\ &= 2 + 1 \\ &= 3\end{aligned}$$

and

$$\lim_{x \rightarrow 0} G'(x) = \lim_{x \rightarrow 0} 2 = 2.$$

Hence,

$$\begin{aligned}\lim_{x \rightarrow 0} \frac{F'(x)}{G'(x)} &= \lim_{x \rightarrow 0} \frac{2e^{x^2} + 4x^2e^{x^2} + \cos(x)}{2} \\ &= \frac{3}{2}.\end{aligned}$$

We can now apply L'Hôpital's Rule to get that

$$\begin{aligned}\lim_{x \rightarrow 0} \frac{2xe^{x^2} + \sin(x)}{2x} &= \lim_{x \rightarrow 0} \frac{f'(x)}{g'(x)} \\ &= \lim_{x \rightarrow 0} \frac{F(x)}{G(x)} \\ &= \lim_{x \rightarrow 0} \frac{F'(x)}{G'(x)} \\ &= \frac{3}{2}.\end{aligned}$$

A second application of L'Hôpital's Rule shows that

$$\begin{aligned}\lim_{x \rightarrow 0} \frac{e^{x^2} - \cos(x)}{x^2} &= \lim_{x \rightarrow 0} \frac{f(x)}{g(x)} \\ &= \lim_{x \rightarrow 0} \frac{f'(x)}{g'(x)} \\ &= \frac{3}{2}.\end{aligned}$$



The next example in this section is a blend of the previous two examples. However, it demonstrates how you may be tempted to use L'Hôpital's Rule incorrectly and therefore, calculate incorrect results.

EXAMPLE 18 Evaluate

$$\lim_{x \rightarrow 0} \frac{e^x - 1}{x^2}.$$

SOLUTION Let $f(x) = e^x - 1$ and $g(x) = x^2$. It is easy to verify that this limit is indeterminate of type $\frac{0}{0}$. Applying L'Hôpital's Rule we get $f'(x) = e^x$ and $g'(x) = 2x$.

Now

$$\lim_{x \rightarrow 0} 2x = 0.$$

So we might be tempted to try the method used in the previous example. Let

$$F(x) = f'(x) = e^x$$

and

$$G(x) = g'(x) = 2x.$$

Then

$$F'(x) = e^x$$

and

$$G'(x) = 2.$$

It follows that

$$\lim_{x \rightarrow 0} \frac{F'(x)}{G'(x)} = \lim_{x \rightarrow 0} \frac{e^x}{2} = \frac{1}{2}.$$

Using these limits we could apply L'Hôpital's Rule to get that

$$\lim_{x \rightarrow 0} \frac{f'(x)}{g'(x)} = \frac{1}{2}$$

and then apply the rule again to get that

$$\begin{aligned} \lim_{x \rightarrow 0} \frac{e^x - 1}{x^2} &= \lim_{x \rightarrow 0} \frac{f(x)}{g(x)} \\ &= \lim_{x \rightarrow 0} \frac{f'(x)}{g'(x)} \\ &= \frac{1}{2}. \end{aligned}$$

All of this looks very nice – the only problem is that *IT IS WRONG!*

What did we do that was incorrect? The mistake is that it is not true that

$$\begin{aligned} \lim_{x \rightarrow 0} \frac{f'(x)}{g'(x)} &= \lim_{x \rightarrow 0} \frac{e^x}{2x} \\ &= \frac{1}{2}. \end{aligned}$$

While we had

$$\lim_{x \rightarrow 0} g'(x) = \lim_{x \rightarrow 0} 2x = 0,$$

the numerator does not approach 0 since

$$\lim_{x \rightarrow 0} e^x = e^0 = 1.$$

The limit rules tell us that if the denominator approaches 0 but the numerator does not, then the limit of the quotient *does not exist*. In this case, we calculated the *wrong* answer because f' and g' do *not* satisfy the conditions for L'Hôpital's Rule.

However, since L'Hôpital's Rule is still valid if $\lim_{x \rightarrow 0} \frac{f'(x)}{g'(x)} = \infty$, we really should have stopped after the first stage and concluded that in fact

$$\lim_{x \rightarrow 0} \frac{e^x - 1}{x^2} = \infty.$$



This example illustrates the point that *care must always be taken* to ensure that all of

the conditions of the theorem are met before you apply L'Hôpital's Rule.

REMARK

In the previous example, we could have anticipated the fact that

$$\lim_{x \rightarrow 0} \frac{e^x - 1}{x^2} = \infty$$

because we know by using linear approximation that $e^x - 1 \approx x$ and so we might expect

$$\frac{e^x - 1}{x^2} \approx \frac{x}{x^2} = \frac{1}{x}.$$

Of course this is not a precise argument. However, in the next course will study a method that does make this simple argument much more rigorous.. 

EXAMPLE 19

Evaluate

$$\lim_{x \rightarrow \infty} \frac{\ln(x)}{x}.$$

SOLUTION This is the *Fundamental Log Limit* that we studied previously. In fact, we know that

$$\lim_{x \rightarrow \infty} \frac{\ln(x)}{x} = 0.$$

We can use L'Hôpital's Rule to verify this result. Let $f(x) = \ln(x)$ and $g(x) = x$. Then

$$\lim_{x \rightarrow \infty} f(x) = \infty = \lim_{x \rightarrow \infty} g(x).$$

This produces an indeterminate form of the type $\frac{\infty}{\infty}$.

Differentiating f and g gives us $f'(x) = \frac{1}{x}$ and $g'(x) = 1$. Therefore,

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{f'(x)}{g'(x)} &= \lim_{x \rightarrow \infty} \frac{\frac{1}{x}}{1} \\ &= \lim_{x \rightarrow \infty} \frac{1}{x} \\ &= 0. \end{aligned}$$

L'Hôpital's Rule implies that

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{\ln(x)}{x} &= \lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} \\ &= \lim_{x \rightarrow \infty} \frac{f'(x)}{g'(x)} \\ &= 0 \end{aligned}$$

exactly as we expected. 

Up until now we have dealt with two types of indeterminate forms which we have denoted by $\frac{0}{0}$ and $\frac{\infty}{\infty}$. There are five more standard indeterminate forms which we will denote by

$$0 \cdot \infty, \infty - \infty, 1^\infty, \infty^0, \text{ and } 0^0.$$

For example, an indeterminate form of type $0 \cdot \infty$ arises from the function $h(x) = f(x)g(x)$ when

$$\lim_{x \rightarrow a} f(x) = 0$$

and

$$\lim_{x \rightarrow a} g(x) = \infty.$$

Similarly, the function $(g(x))^{f(x)}$ would produce an indeterminate form of type ∞^0 .

EXAMPLE 20 Use L'Hôpital's Rule to evaluate

$$\lim_{x \rightarrow 0^+} x \ln(x)$$

with $g(x) = x$ and $f(x) = \ln(x)$.

SOLUTION This is an indeterminate form of type $0 \cdot \infty$ since

$$\lim_{x \rightarrow 0^+} x = 0$$

and

$$\lim_{x \rightarrow 0^+} \ln(x) = -\infty.$$

Using a trick we can turn this example into an indeterminate form of type $\frac{\infty}{\infty}$. To do so we write

$$x \ln(x) = \frac{\ln(x)}{\frac{1}{x}}.$$

With $F(x) = f(x) = \ln(x)$ and $G(x) = \frac{1}{g(x)} = \frac{1}{x}$, we now have an indeterminate form of type $\frac{0}{0}$.

Moreover, since $G'(x) = -\frac{1}{x^2}$ is never 0, we have satisfied all of the conditions required to use L'Hôpital's Rule. Then

$$\frac{F'(x)}{G'(x)} = \frac{\frac{1}{x}}{-\frac{1}{x^2}} = -x.$$

It follows that

$$\lim_{x \rightarrow 0^+} \frac{F'(x)}{G'(x)} = \lim_{x \rightarrow 0^+} -x = 0.$$

Therefore, L'Hôpital's Rule shows that

$$\begin{aligned} \lim_{x \rightarrow 0^+} x \ln(x) &= \lim_{x \rightarrow 0^+} \frac{F(x)}{G(x)} \\ &= \lim_{x \rightarrow 0^+} \frac{F'(x)}{G'(x)} \\ &= 0. \end{aligned}$$

The next example provides us with an alternative method for establishing a very important limit that can be used as another definition for the number e .

EXAMPLE 21 Evaluate

$$\lim_{x \rightarrow \infty} \left(1 + \frac{1}{x}\right)^x.$$

SOLUTION This is an indeterminate form of type 1^∞ .

The first step in evaluating this limit is to write the function as follows:

$$\left(1 + \frac{1}{x}\right)^x = e^{x \ln\left(1 + \frac{1}{x}\right)}.$$

Consider what happens to

$$x \ln\left(1 + \frac{1}{x}\right)$$

as $x \rightarrow \infty$. As $x \rightarrow \infty$, $\left(1 + \frac{1}{x}\right) \rightarrow 1$ so that

$$\ln\left(1 + \frac{1}{x}\right) \rightarrow 0.$$

This means that

$$x \ln\left(1 + \frac{1}{x}\right)$$

is indeterminate of type $0 \cdot \infty$ as $x \rightarrow \infty$. We can use the same trick as in the previous example and write $F(x) = \ln\left(1 + \frac{1}{x}\right)$ and $G(x) = \frac{1}{x}$ to turn this latter limit into type $\frac{\infty}{\infty}$.

Next, we have

$$\begin{aligned} \frac{F'(x)}{G'(x)} &= \lim_{x \rightarrow \infty} \frac{\frac{(-1)}{x^2}}{\frac{1+\frac{1}{x}}{x^2}} \\ &= \frac{1}{1 + \frac{1}{x}} \end{aligned}$$

so that

$$\lim_{x \rightarrow \infty} \frac{F'(x)}{G'(x)} = \lim_{x \rightarrow \infty} \frac{1}{1 + \frac{1}{x}} = 1.$$

Therefore L'Hôpital's Rule shows us that

$$x \ln\left(1 + \frac{1}{x}\right) \rightarrow 1$$

as $x \rightarrow \infty$.

Finally, since e^x is continuous we get that

$$\left(1 + \frac{1}{x}\right)^x = e^{x \ln(1 + \frac{1}{x})} \rightarrow e^1 = e$$

as $x \rightarrow \infty$.

In summary, we have shown that

$$\lim_{x \rightarrow \infty} \left(1 + \frac{1}{x}\right)^x = e.$$

A similar calculation can show that for any $a \in \mathbb{R}$,

$$\lim_{x \rightarrow \infty} \left(1 + \frac{a}{x}\right)^x = e^a.$$



The next problem is rather challenging. In theory the limit could be evaluated using L'Hôpital's Rule since it is indeterminate of the form $\frac{0}{0}$. However, the calculations involved become very messy, very quickly. Instead, and somewhat incredibly, we will soon study a method that could be used to evaluate the limit by inspection (see *Big-O Notation*) – you might want to wait until then to try it!

Problem: Show that

$$\lim_{x \rightarrow 0} \frac{4(e^{x^3} - 1 - x^3 - \frac{x^6}{2})^2}{x^6 \tan(x^7) \sin(2x^5)} = \frac{1}{18}.$$



Historical Note: L'Hôpital's Rule is named after G. F. A. L'Hôpital, a French nobleman, who lived from 1661-1704. The rule appeared in his book *Analyse des infiniment petits*, often regarded as the first ever Calculus text. While this book played a significant role in the development of the Calculus in the 18th century, the main ideas in the book were not developed by L'Hôpital. In fact, L'Hôpital's Rule is due to Johann Bernoulli whom L'Hôpital paid a rather substantial salary in exchange for Bernoulli's many mathematical discoveries. Indeed, the top mathematicians of the time often had patrons just like many of the world's great artists. (Reference: see the website *MacTutor History of Mathematics Archive*.)

We have already given a reasonable argument to why L'Hôpital's Rule should be valid using what we know about linear approximation. However, before we can give a formal proof of L'Hôpital's Rule we will need an upgraded version of the Mean Value Theorem.

7.4 Cauchy's Mean Value Theorem

In this section we will prove the following upgraded version of the Mean Value Theorem due to Cauchy:

THEOREM 15

Cauchy's Mean Value Theorem (CMVT)

Suppose that f and g are continuous on $[a, b]$ and differentiable on (a, b) with $g'(x) \neq 0$ for any $x \in (a, b)$. Then $g(a) \neq g(b)$ and there exists a $c \in (a, b)$ such that

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

NOTE

- 1) If $g(x) = x$, this is the MVT.
- 2) By the usual MVT we would get

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{\frac{f(b) - f(a)}{b-a}}{\frac{g(b) - g(a)}{b-a}} = \frac{f'(c_1)}{g'(c_2)}.$$

The Cauchy Mean Value Theorem tells us that we can replace c_1 and c_2 in the above equality with a single point c .

PROOF

Since $g'(x) \neq 0$, $g(a) \neq g(b)$.

Define

$$H(x) = \frac{f(b) - f(a)}{g(b) - g(a)}[g(x) - g(a)] - (f(x) - f(a)).$$

Then $H(x)$ is continuous on $[a, b]$ and differentiable on (a, b) with

$$H(a) = \frac{f(b) - f(a)}{g(b) - g(a)}[g(a) - g(a)] - (f(a) - f(a)) = 0$$

and

$$H(b) = \frac{f(b) - f(a)}{g(b) - g(a)}[g(b) - g(a)] - (f(b) - f(a)) = 0.$$

By Rolle's Theorem there is a $c \in (a, b)$ such that

$$0 = H'(c) = \frac{f(b) - f(a)}{g(b) - g(a)}g'(c) - f'(c).$$

It follows that

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

■

7.4.1 Geometric Interpretation of the Cauchy Mean Value Theorem

Recall that the Mean Value Theorem can be interpreted as saying that the secant line joining $(a, f(a))$ and $(b, f(b))$ is parallel to the tangent line through $(c, f(c))$ for some point $c \in (a, b)$. We will now give an analogous geometric interpretation of the Cauchy Mean Value Theorem. However, to do so we must first introduce the concept of a curve in \mathbb{R}^2 .

DEFINITION

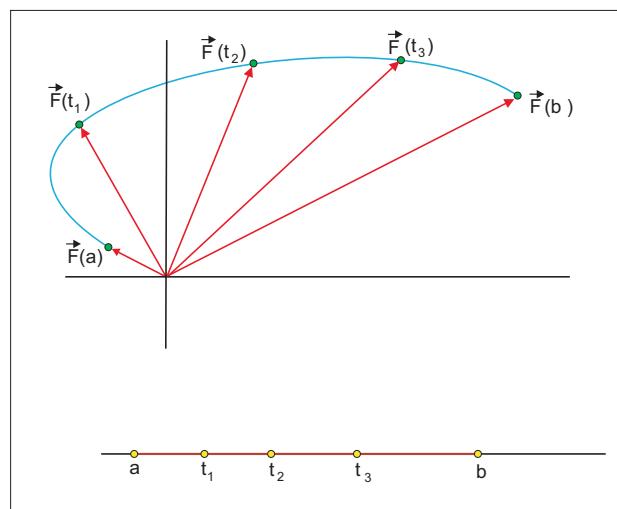
Curve in \mathbb{R}^2

A *curve* in \mathbb{R}^2 is a function $\vec{F}: [a, b] \rightarrow \mathbb{R}^2$ given by

$$\vec{F}(t) = (g(t), f(t))$$

$g(t)$ and $f(t)$ are called the *coordinate functions* of \vec{F} .

We can visualize a curve by plotting its range.



There is a notion of differentiability for functions such as \vec{F} . In fact it turns out that \vec{F} is differentiable precisely when its coordinate functions g and f are differentiable and the derivative in this case is a vector in \mathbb{R}^2 given by

$$\vec{F}'(t) = (g'(t), f'(t)).$$

Moreover, the line in the direction of $\vec{F}'(t) = (g'(t), f'(t))$ through the point $\vec{F}(t)$ is the *tangent line* to the curve at $\vec{F}(t)$. The slope of this line is simply

$$m = \frac{f'(t)}{g'(t)}.$$

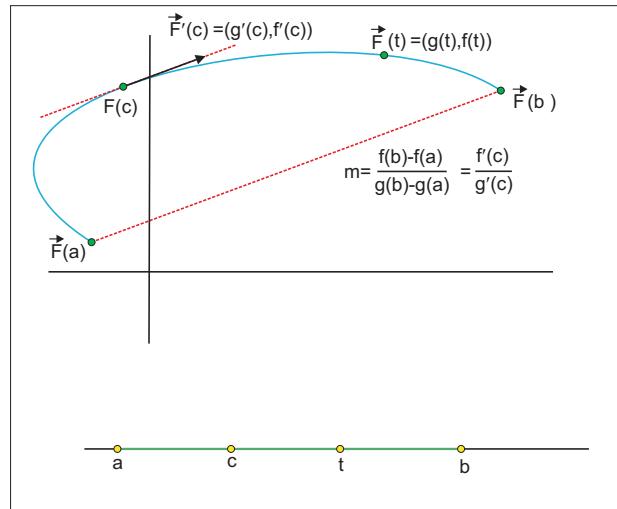
Consider the secant line through $\vec{F}(a) = (g(a), f(a))$ and $\vec{F}(b) = (g(b), f(b))$. The slope of this line is

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

Putting the last two observations together Cauchy Mean Value Theorem tells us that there exists a point $c \in (a, b)$ such that

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

In other words, the secant line through $\vec{F}(a) = (g(a), f(a))$ and $\vec{F}(b) = (g(b), f(b))$ is parallel to the tangent line to the curve through $\vec{F}(c)$.



7.5 The Proof of L'Hôpital's Rule

We now have the tools we need to prove L'Hôpital's Rule. In fact, we will prove a part of a slightly upgraded version of the rule. Before we do we will introduce some new notation.

DEFINITION**Indeterminate Forms**

$\mathbb{R}^* = \mathbb{R} \cup \{-\infty, \infty\}$ is called the set of extended real numbers.

Suppose that $f, g : I \rightarrow \mathbb{R}$, where I is an open interval containing some $a \in \mathbb{R}^*$ as an endpoint. Suppose that $g(x) \neq 0$ for all $x \in I$.

1) The limit $\lim_{x \rightarrow a^\pm} \frac{f(x)}{g(x)}$ is called an *indeterminate form of type $\frac{0}{0}$* if

$$\lim_{x \rightarrow a^\pm} f(x) = 0 = \lim_{x \rightarrow a^\pm} g(x).$$

2) The limit $\lim_{x \rightarrow a^\pm} \frac{f(x)}{g(x)}$ is called an *indeterminate form of type $\frac{\infty}{\infty}$* if

$$\lim_{x \rightarrow a^\pm} f(x) = \pm\infty \text{ and } \lim_{x \rightarrow a^\pm} g(x) = \pm\infty.$$

We will now state a slightly upgraded version of the $\frac{0}{0}$ version of L'Hôpital's Rule.

THEOREM 16**L'Hôpital's Rule Version $\frac{0}{0}$**

Assume that $f, g : (a, b) \rightarrow \mathbb{R}$, where $a, b \in \mathbb{R}^*$ with $a < b$. Also assume that f and g are differentiable on (a, b) and that both $g(x)$ and $g'(x) \neq 0$ for all $x \in (a, b)$.

1) Assume that $\lim_{x \rightarrow a^+} f(x) = 0 = \lim_{x \rightarrow a^+} g(x)$.

i) If $\lim_{x \rightarrow a^+} \frac{f'(x)}{g'(x)} = L \in \mathbb{R}$, then $\lim_{x \rightarrow a^+} \frac{f(x)}{g(x)} = L$.

ii) If $\lim_{x \rightarrow a^+} \frac{f'(x)}{g'(x)} = \pm\infty$, then $\lim_{x \rightarrow a^+} \frac{f(x)}{g(x)} = \pm\infty$.

2) Assume that $\lim_{x \rightarrow b^-} f(x) = 0 = \lim_{x \rightarrow b^-} g(x)$.

i) If $\lim_{x \rightarrow b^-} \frac{f'(x)}{g'(x)} = L \in \mathbb{R}$, then $\lim_{x \rightarrow b^-} \frac{f(x)}{g(x)} = L$.

ii) If $\lim_{x \rightarrow b^-} \frac{f'(x)}{g'(x)} = \pm\infty$, then $\lim_{x \rightarrow b^-} \frac{f(x)}{g(x)} = \pm\infty$.

PROOF

We will prove 1i). The remaining cases are very similar.

Assume that $f, g : (a, b) \rightarrow \mathbb{R}$, where $a, b \in \mathbb{R}^*$ with $a < b$. Also assume that f and g are differentiable on (a, b) and that both $g(x)$ and $g'(x) \neq 0$ for all $x \in (a, b)$. Assume

that $\lim_{x \rightarrow a^+} f(x) = 0 = \lim_{x \rightarrow a^+} g(x)$ and that

$$\lim_{x \rightarrow a^+} \frac{f'(x)}{g'(x)} = L$$

where $L \in \mathbb{R}$.

Let $\epsilon > 0$. Choose $a < \beta$ in I such that if $a < \xi < \beta$, then

$$\left| \frac{f'(\xi)}{g'(\xi)} - L \right| < \epsilon.$$

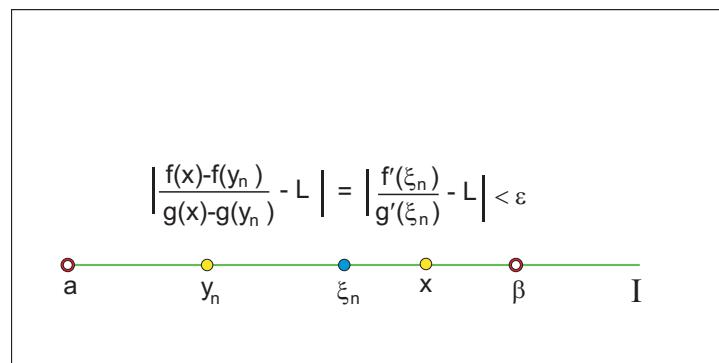
We can do this since $\lim_{x \rightarrow a^+} \frac{f'(x)}{g'(x)} = L$.

Let $a < x < \beta$. Next we choose a sequence $\{y_n\}$, with $a < y_n < x$ and $y_n \rightarrow a$. Then by the CMVT, for each $n \in \mathbb{N}$ there exists a point ξ_n between x and y_n such that

$$\left| \frac{f(x) - f(y_n)}{g(x) - g(y_n)} - L \right| = \left| \frac{f'(\xi_n)}{g'(\xi_n)} - L \right|.$$

Next we note that $a < \xi_n < \beta$. It then follows that

$$\left| \frac{f(x) - f(y_n)}{g(x) - g(y_n)} - L \right| = \left| \frac{f'(\xi_n)}{g'(\xi_n)} - L \right| < \epsilon.$$



We have shown that for each $n \in \mathbb{N}$ that

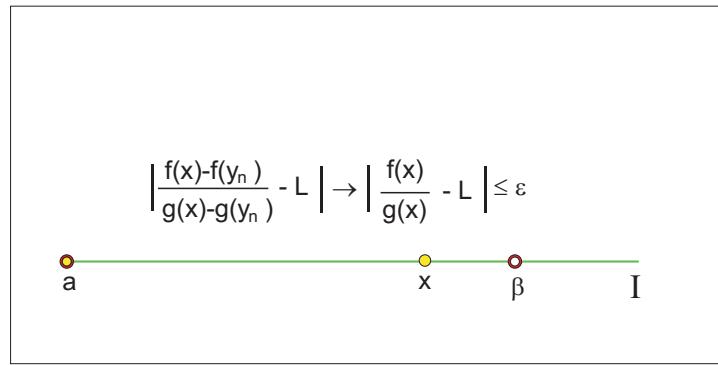
$$\left| \frac{f(x) - f(y_n)}{g(x) - g(y_n)} - L \right| < \epsilon.$$

Since

$$\lim_{n \rightarrow \infty} f(y_n) = 0 = \lim_{n \rightarrow \infty} g(y_n)$$

we have

$$\lim_{n \rightarrow \infty} \left| \frac{f(x) - f(y_n)}{g(x) - g(y_n)} - L \right| = \left| \frac{f(x)}{g(x)} - L \right| \leq \epsilon.$$



In summary, we have shown that if $a < x < \beta$, then

$$\left| \frac{f(x)}{g(x)} - L \right| \leq \epsilon.$$

This is exactly what we needed to do to prove that

$$\lim_{x \rightarrow a^+} \frac{f(x)}{g(x)} = L.$$

■

REMARK

As we mentioned earlier the proofs of all of the other cases of L'Hôpital's Rule with indeterminate forms of the type $\frac{0}{0}$ as quite similar. Moreover, the two-sided limit analogs follow immediately as well from our one-sided results.

The case where we have indeterminate forms of the type $\frac{\infty}{\infty}$ is slightly more complicated, involving a small trick. But since the basic argument is the same we will not address this case here.

◀

7.6 Curve Sketching: Part 2

Previously, we used the theory of limits and continuity to develop a procedure to aid in constructing basic sketches of graphs of functions. We can now use the information obtained from the derivative to refine the sketches.

To draw the graph of a function f you should first follow the steps for curve sketching outlined in Chapter 5 (see *Curve Sketching: Part 1*). Once you have completed those steps, you should try to obtain as much of the following information as possible.

Strategy [Basic Curve Sketching: Part 2]

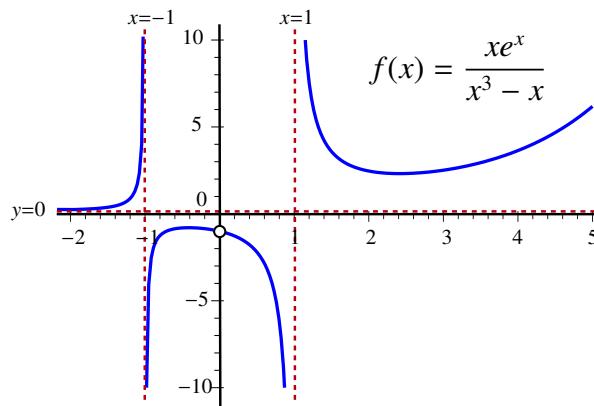
- **Step 1:** Complete the steps for *Curve Sketching: Part 1*.
- **Step 2:** Calculate $f'(x)$.
- **Step 3:** Identify any critical points: locate where $f'(x) = 0$ or where $f'(x)$ does not exist.
- **Step 4:** Determine where f is increasing or decreasing by analysing the sign of $f'(x)$ between the critical points.
- **Step 5:** Test the critical points to determine if they are local maxima, local minima, or neither.
- **Step 6:** Find $f''(x)$.
- **Step 7:** Locate where $f''(x) = 0$ or where $f''(x)$ does not exist. Use these points to divide the Real number line into intervals. Determine the concavity of f by analysing the sign of $f''(x)$ inside these intervals (if possible).
- **Step 8:** Identify any points of inflection.
- **Step 9:** Incorporate this information into your original sketch.

Ideally, we would follow each of these steps when we construct the graph of a function. However, in practice, some or all of these steps may be quite difficult to complete. Moreover, in many instances your initial sketch produced using the steps from *Curve Sketching Part 1* will be accurate enough to suggest the information that would be obtained using the derivative. We will illustrate these remarks by revisiting the example from our earlier look at curve sketching.

EXAMPLE 22 Sketch the graph of

$$f(x) = \frac{xe^x}{x^3 - x}.$$

SOLUTION Recall that our previous investigation of this function gave us a graph that appeared as follows:



The graph does indeed suggest some of the characteristics that we might find by following the steps in the *Basic Curve Sketching Strategy, Part 2*. For example, it suggests that the function has a local maximum between $x = -1$ and $x = 0$, and a local minimum somewhere after $x = 1$. Let's complete the curve sketching steps to confirm our suspicions.

Step 1: Calculate $f'(x)$. Taking the derivative of $f(x)$ we get

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

for $x \neq 0, x \neq \pm 1$.

Step 2: A critical point $x = c$ occurs when c is in the domain of f and either $f'(c)$ does not exist or $f'(c) = 0$. In step 1 we saw that $f'(x)$ does not exist when $x = 0$ or $x = \pm 1$. However, $x = 0$ and $x = \pm 1$ are not critical points according to the definition since they are not in the domain of the function f . Indeed, in our original sketch, we found that $x = 0$ was a removable discontinuity, and $x = 1$ and $x = -1$ were the locations of the vertical asymptotes. Now since $e^x > 0$ for all x and the denominator is always positive for all $x \neq \pm 1$, it follows that $f'(x) = 0$ when

$$x^2 - 2x - 1 = 0.$$

The quadratic formula tells us that these critical points occur when

$$x = \frac{2 \pm \sqrt{4 + 4}}{2} = 1 \pm \sqrt{2}.$$

Moreover,

$$1 + \sqrt{2} \approx 2.414213562$$

and

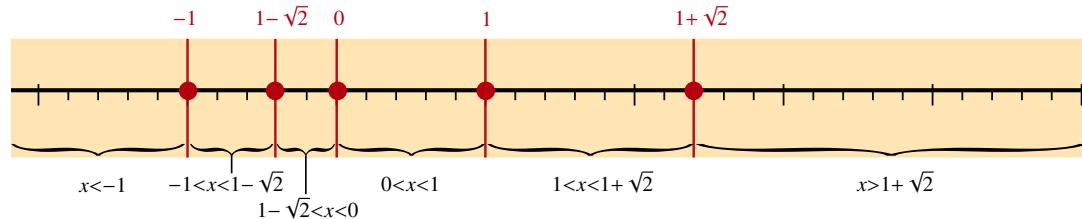
$$1 - \sqrt{2} \approx -0.414213562.$$

This suggests that the local minimum is located at $x \approx 2.414213562$ and the local maximum is located at $x \approx -0.414213562$. We will confirm these results in the next few steps.

Step 3: We can determine where f is increasing or decreasing by analysing the sign of $f'(x)$ in particular intervals and then applying the Increasing/Decreasing Function Theorem.

First, use the critical points and discontinuities we found in step 2 to divide the real line into intervals.

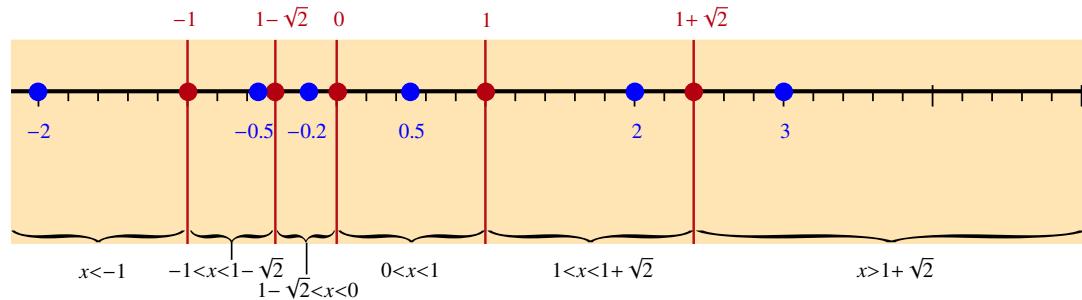
● $f(x)$ critical points and discontinuities



According to the Increasing/Decreasing Function Theorem, if $f'(x) > 0$ on some interval I , then f is increasing on I . Otherwise, f is decreasing. Using the intervals determined by the locations of the critical points and discontinuities, we now need to determine the sign of $f'(x)$ inside each of these intervals. Any test point within each interval will suffice, so we will choose numbers inside the intervals that make the calculation easier.

Since e^x is always positive and the denominator of $f'(x)$ is always positive when $x \neq \pm 1$, the sign of $f'(x)$ depends only on the sign of the numerator, $x^2 - 2x - 1$. Let's check $x = -2$, $x = -0.5$, $x = -0.2$, $x = 0.5$, $x = 2$, and $x = 3$ (these are random points, each found inside one of the intervals).

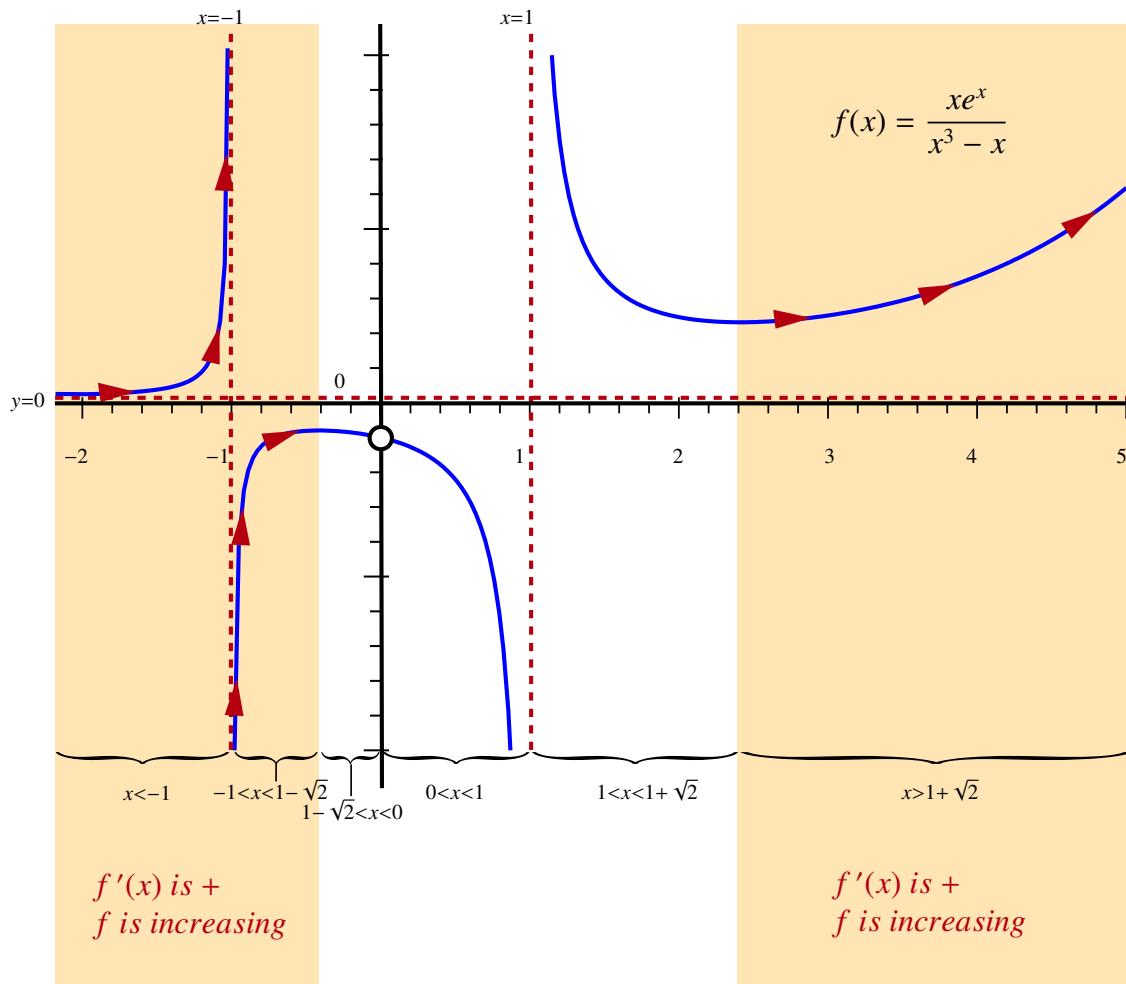
● $f(x)$ critical points and discontinuities
 ● $f'(x)$ test points



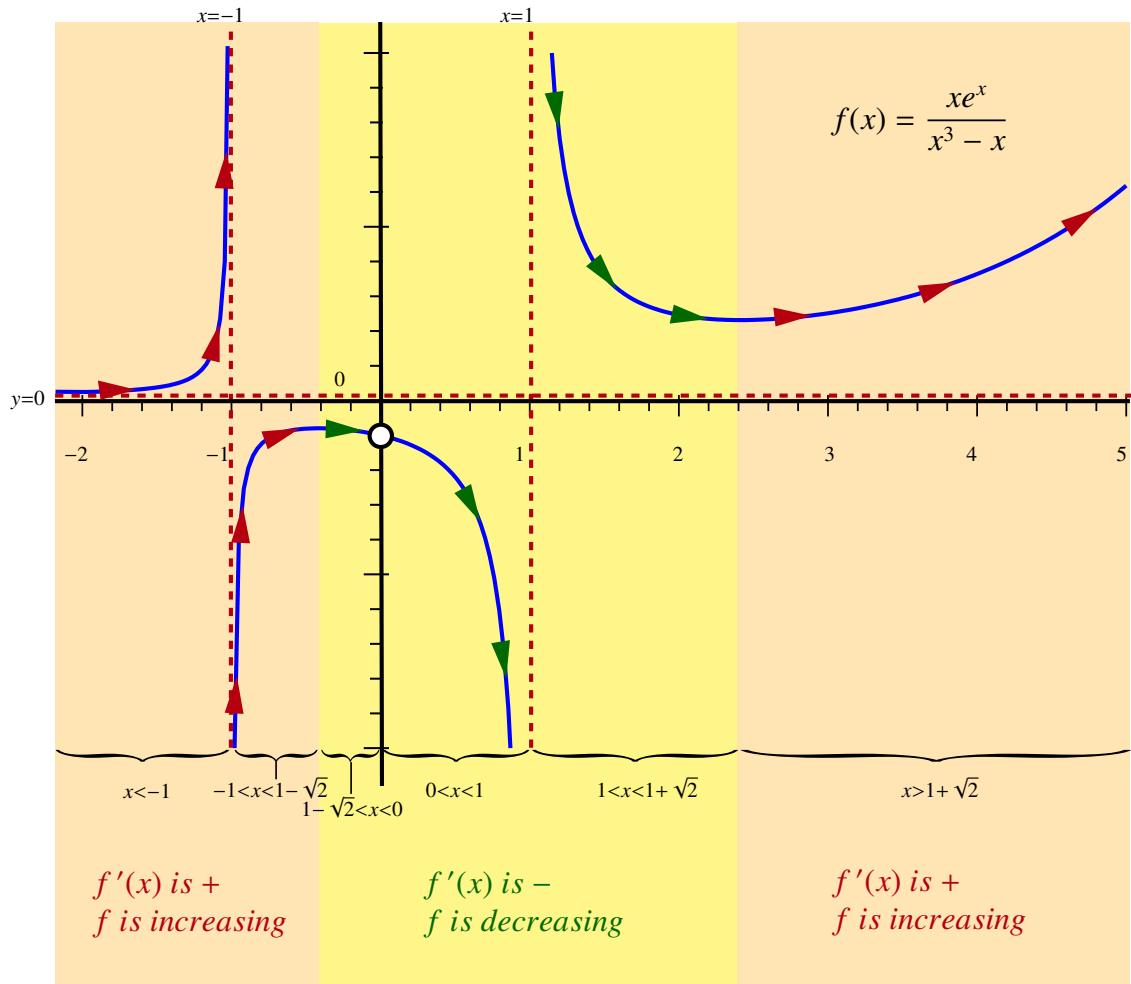
Then we get (you should verify these calculations):

Interval	Test Point	Calculate $x^2 - 2x - 1$	$f'(x)$ positive or negative	f increasing or decreasing
$x < -1$	$x = -2$	$(-2)^2 - 2(-2) - 1 = +7$	$f'(x) > 0$ (positive)	f increasing
$-1 < x < 1 - \sqrt{2}$	$x = -0.5$	$(-0.5)^2 - 2(-0.5) - 1 = +0.25$	$f'(x) > 0$ (positive)	f increasing
$1 - \sqrt{2} < x < 0$	$x = -0.2$	$(0.2)^2 - 2(0.2) - 1 = -1.36$	$f'(x) < 0$ (negative)	f decreasing
$0 < x < 1$	$x = 0.5$	$(0.5)^2 - 2(0.5) - 1 = -1.75$	$f'(x) < 0$ (negative)	f decreasing
$1 < x < 1 + \sqrt{2}$	$x = 2$	$(2)^2 - 2(2) - 1 = -1$	$f'(x) < 0$ (negative)	f decreasing
$x > 1 + \sqrt{2}$	$x = 3$	$(3)^2 - 2(3) - 1 = +2$	$f'(x) > 0$ (positive)	f increasing

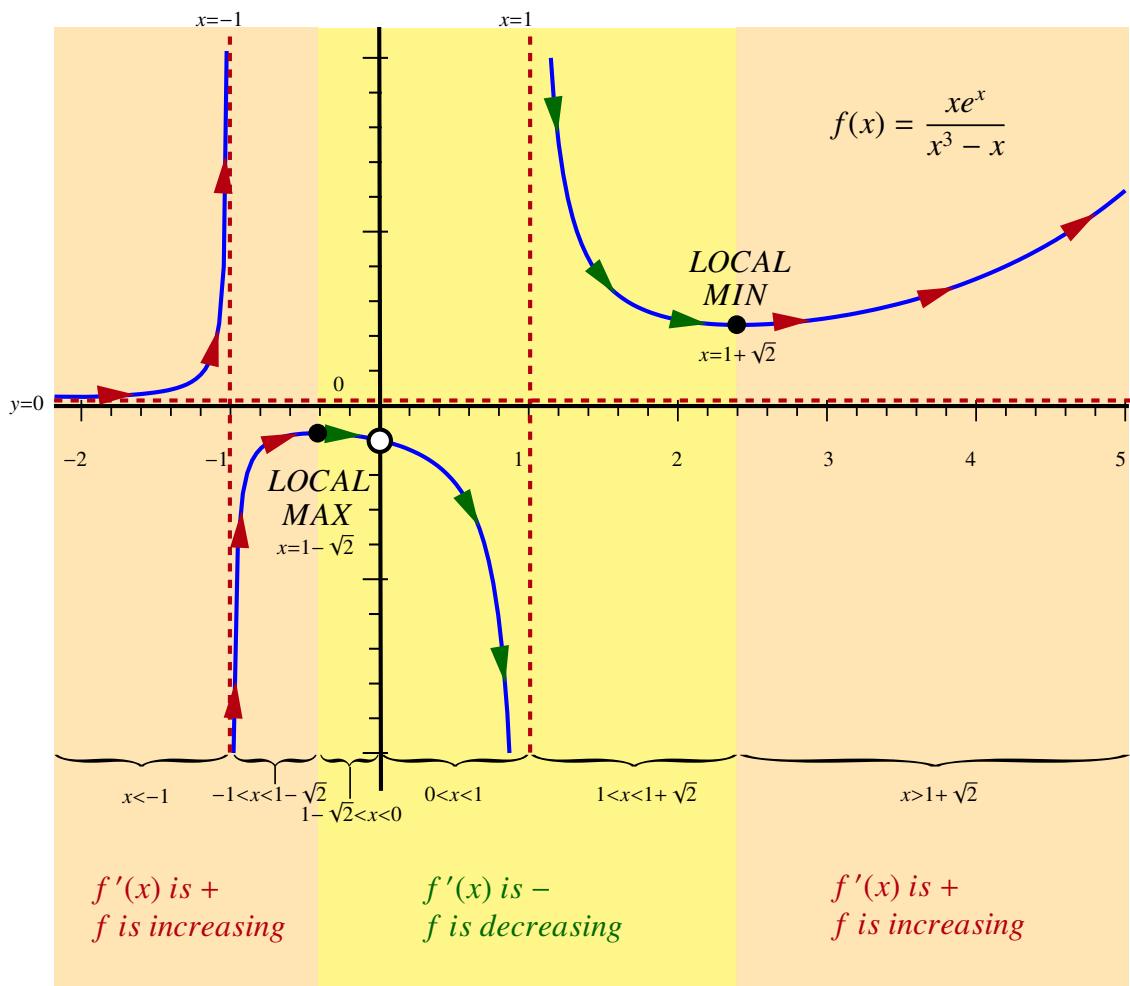
In summary, $x^2 - 2x - 1 > 0$ and therefore $f'(x) > 0$ in the intervals where $x > 1 + \sqrt{2}$ and if $-1 < x < 1 - \sqrt{2}$ and $x < -1$. This means $f(x)$ is increasing when $x > 1 + \sqrt{2}$ or when $x < 1 - \sqrt{2}$, except at $x \neq -1$ (vertical asymptote).



We also have that $x^2 - 2x - 1 < 0$ and therefore $f'(x) < 0$ in the interval where $1 - \sqrt{2} < x < 0$ or $0 < x < 1$ or $1 < x < 1 + \sqrt{2}$. Therefore, f is decreasing if $1 - \sqrt{2} < x < 1 + \sqrt{2}$, except at $x = 0$ (removable discontinuity) and at $x = 1$ (vertical asymptote).



Step 4: We can now use the *First Derivative Test* to verify the nature of the two critical points, $x = 1 - \sqrt{2}$ and $x = 1 + \sqrt{2}$. We know that the function is decreasing as we approach $x = 1 + \sqrt{2}$ from the left and the function is increasing as we move away from this critical point to the right. As such, the First Derivative Test guarantees that the local minimum is located at $x = 1 + \sqrt{2}$. Similarly, we can confirm that the local maximum is located at $x = 1 - \sqrt{2}$.



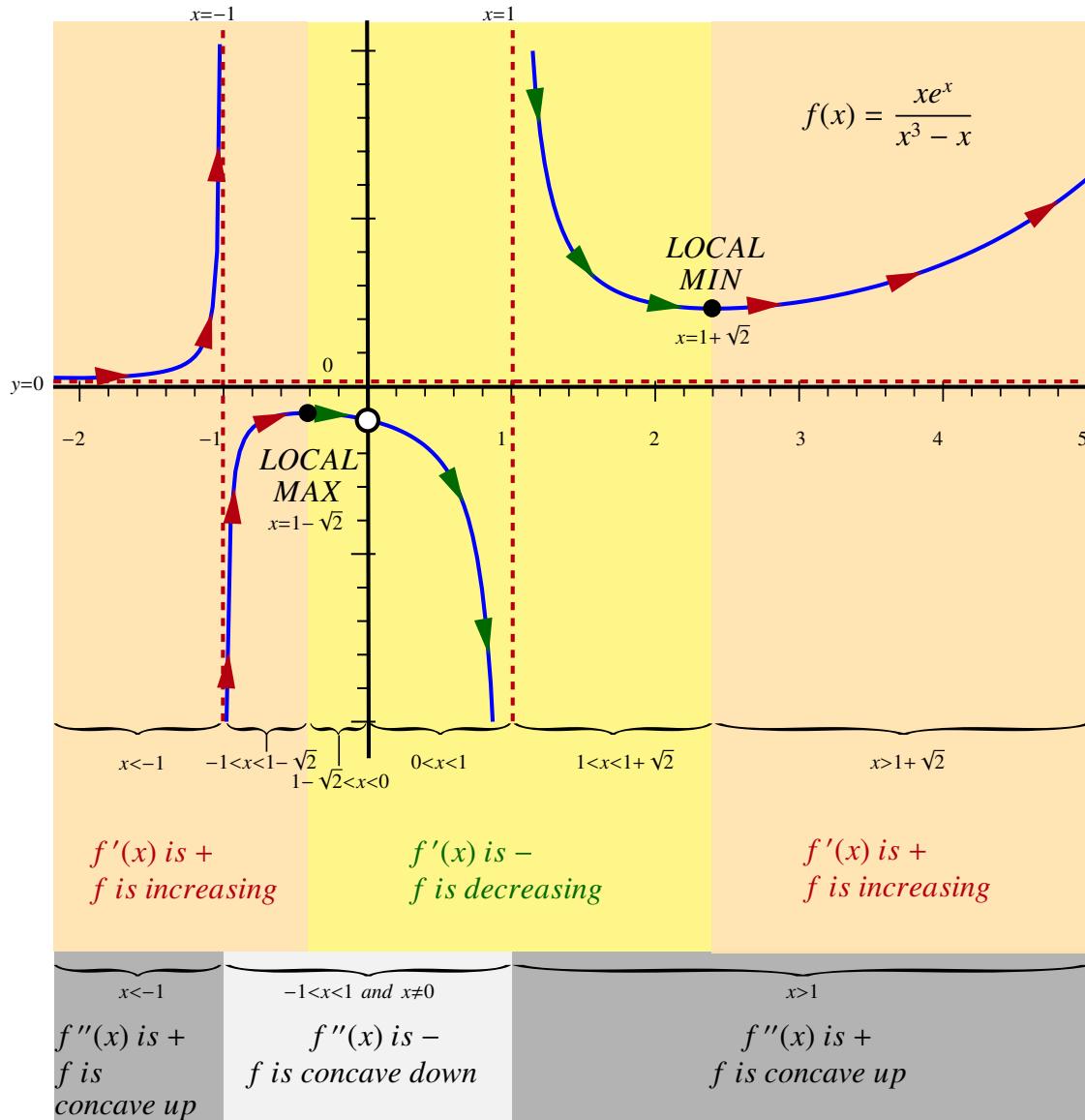
Steps 5-7: The second derivative of $f(x)$ is

$$f''(x) = e^x \frac{(x^4 - 4x^3 + 4x^2 + 4x + 3)}{(x^2 - 1)^3}$$

which is far to complicated to extract much useful information. However, the previous sketch suggests that $f''(x) < 0$ on the interval $(-1, 1)$ (since the graph of f appears concave downwards here except at $x = 0$) and that $f''(x) > 0$ everywhere else that it exists (since the graph of f appears concave upwards everywhere else). In fact, this can be verified by using a mathematical software program. It turns out that the polynomial in the numerator of $f''(x)$ has no real roots and as such the sign of the second derivative is the same as the sign of $x^2 - 1$, though you would not have been expected to deduce this from the information we have available. Since $f''(x)$

has no real roots, it cannot equal 0, and so there are no possible candidates for points of inflection.

Step 8: Enhance the sketch of the function using the information acquired from completing these steps.



EXAMPLE 23

Determine if the function $f(x) = x^2 e^x$ has any points of inflection.

SOLUTION To locate the points of inflection of a function, you must complete two tasks:

1. Find all candidates for points of inflection by locating where the second derivative $f''(x)$ is zero or does not exist, and

2. Confirm that the concavity of the function changes on either side of these candidates.

In other words, you must confirm that $f''(x)$ changes sign from positive to negative or from negative to positive on either side of every candidate for a point of inflection. Finding the zeros of $f''(x)$ is *not* sufficient to conclude that they are points of inflection.

Step 1: Find the candidates for inflection points (if any): Set $f''(x) = 0$ and solve for x , and locate where $f''(x)$ does not exist.

We have that

$$f'(x) = x^2 e^x + 2x e^x$$

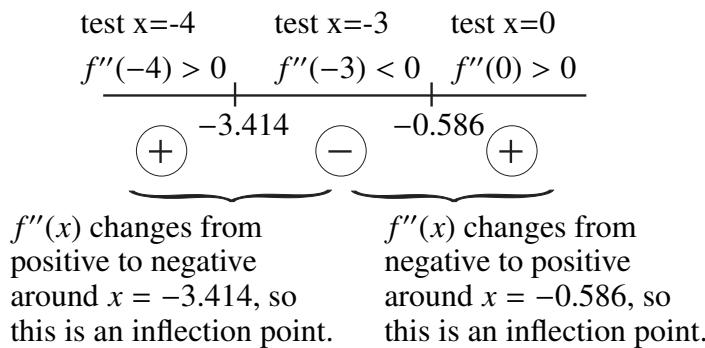
and so

$$f''(x) = x^2 e^x + 4x e^x + 2e^x = e^x(x^2 + 4x + 2).$$

Since $f''(x)$ exists everywhere we only need to solve $f''(x) = x^2 e^x + 4x e^x + 2e^x = e^x(x^2 + 4x + 2) = 0$. The factor e^x is always non-zero so we need to determine when $x^2 + 4x + 2 = 0$. Using the quadratic formula, we get that $x = -2 - \sqrt{2} \approx -3.414$ or $x = -2 + \sqrt{2} \approx -0.586$ are the only candidates for points of inflection.

Step 2: Confirm whether the concavity of the function changes on either side of these candidates by testing the sign of $f''(x)$.

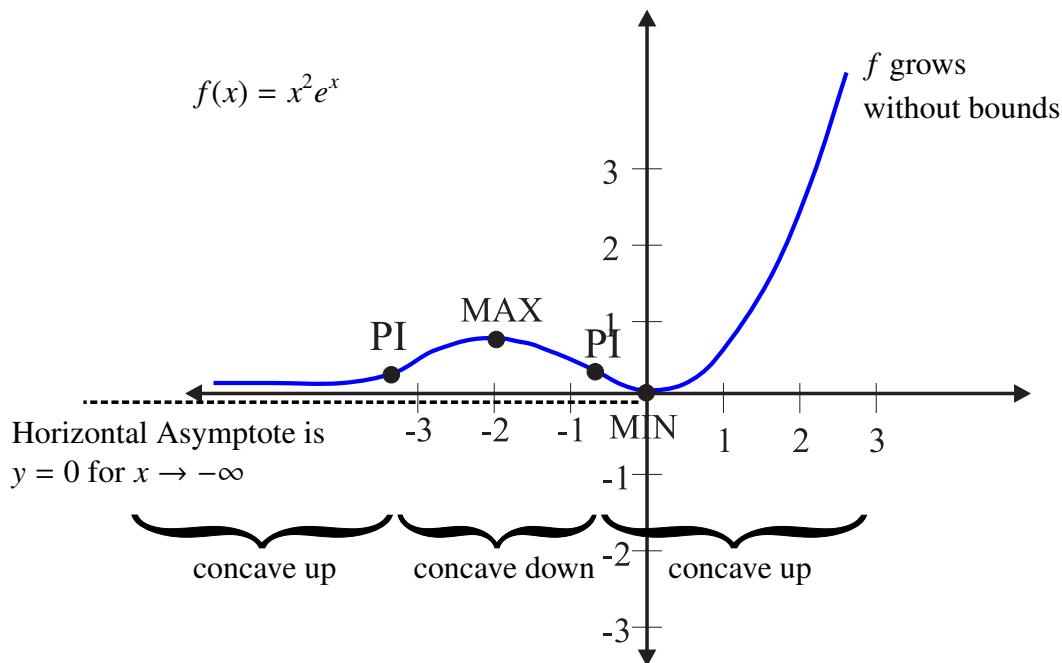
Use $x = -2 - \sqrt{2} \approx -3.414$ and $x = -2 + \sqrt{2} \approx -0.586$ to divide the domain of $f''(x)$ into three intervals. We need to check the sign of $f''(x)$ in each of these intervals.



Since the sign of $f''(x)$ changes on either side of each candidate, they must be points of inflection.

To find the y -coordinate of each point substitute these x -values back into $f(x)$. We have $f(-3.414) \approx 0.38$ and $f(-0.586) \approx 0.19$. Thus we can conclude that the points of inflection for this function are approximately $(-3.414, 0.38)$ and $(-0.586, 0.19)$.

As an additional exercise you should complete all of the steps for Curve Sketching to confirm that the graph of this function appears as follows.


EXAMPLE 24

Show that the function $f(x) = x^{\frac{1}{3}}$ has an inflection point at $x = 0$, but that $f''(x)$ does not exist at $x = 0$.

SOLUTION You should confirm that

$$f'(x) = \frac{1}{3x^{\frac{2}{3}}}$$

and

$$f''(x) = \frac{-2}{9x^{\frac{5}{3}}}.$$

Step 1: Find any candidates for inflection points: Set $f''(x) = 0$ and solve for x , and locate where $f''(x)$ does not exist.

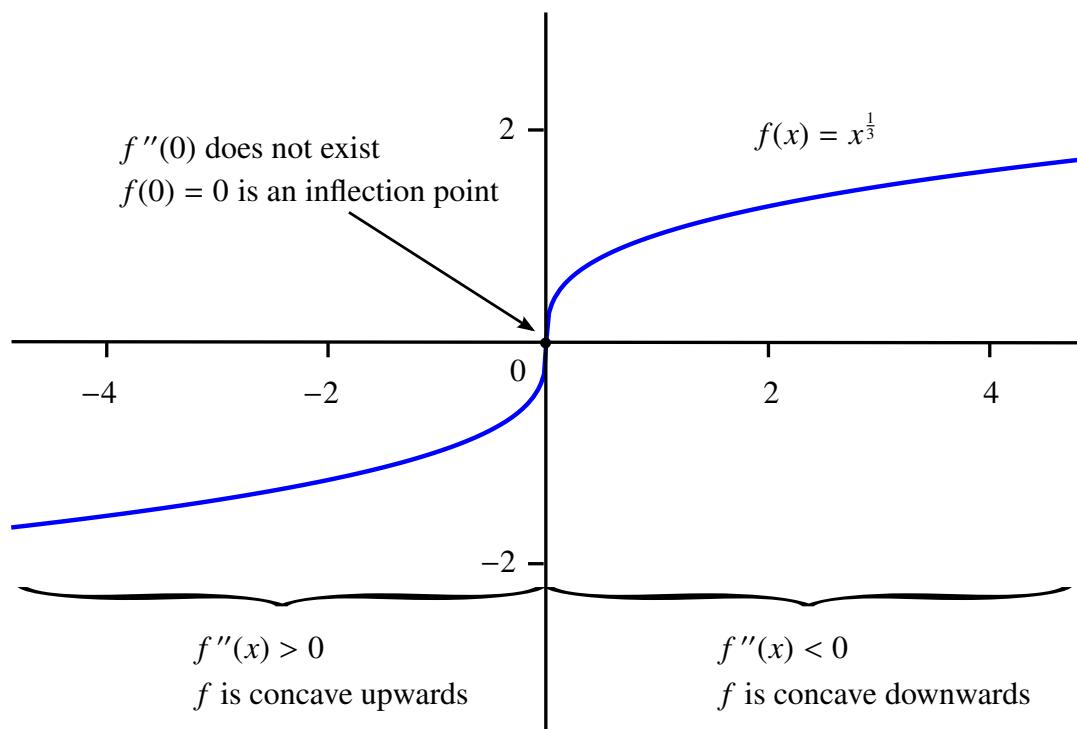
Note that $f''(x)$ has a numerator that is non-zero, so $f''(x) \neq 0$ for all x . However, $f''(x)$ does not exist when the denominator is zero; that is, when $x = 0$.

Step 2: Confirm whether the concavity of the function changes on either side of the candidate $x = 0$.

Case $x < 0$: Choose a test point, say $x = -1$. Since $f''(-1) = \frac{-2}{9(-1)^{\frac{5}{3}}} = +\frac{2}{9} > 0$, it follows that f is concave upwards when $x < 0$.

Case $x > 0$: Choose a test point, say $x = +1$. Since $f''(1) = \frac{-2}{9(1)^{\frac{5}{3}}} = -\frac{2}{9} < 0$, it follows that f is concave downwards when $x > 0$.

In fact, the graph of $f(x) = x^{\frac{1}{3}}$ appears as follows.



Notice that this function has an inflection point at $x = 0$ even though the second derivative $f''(x)$ does not exist at $x = 0$.



Chapter 8

Taylor Polynomials and Taylor's Theorem

In this Chapter we will see that if a function f has derivatives of higher orders, then we can not only construct the linear approximation, but we can also approximate f with higher order polynomials that encode the information provided by these higher derivatives.

8.1 Introduction to Taylor Polynomials and Approximation

Recall that if f is differentiable at $x = a$, then if $x \cong a$

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

Cross-multiplying gives us

$$f'(a)(x - a) \cong f(x) - f(a)$$

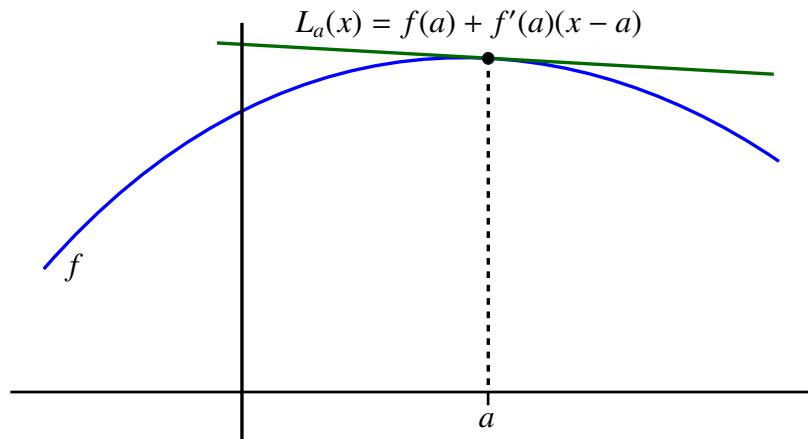
and finally that

$$f(x) \cong f(a) + f'(a)(x - a).$$

This led us to define the *linear approximation to f* at $x = a$ to be the function

$$L_a(x) = f(a) + f'(a)(x - a).$$

We saw that the geometrical significance of the linear approximation is that its graph is the tangent line to the graph of f through the point $(a, f(a))$.



Recall also that the linear approximation has the following two important properties:

1. $L_a(a) = f(a)$.
2. $L'_a(a) = f'(a)$.

In fact, amongst all polynomials of degree at most 1, that is functions of the form

$$p(x) = c_0 + c_1(x - a),$$

the linear approximation is the only one with both properties (1) and (2) and as such, the only one that encodes both the value of the function at $x = a$ and its derivative.

We know that for x near a that

$$f(x) \cong L_a(x).$$

This means that we can use the simple linear function L_a to approximate what could be a rather complicated function f at points near $x = a$. However, any time we use a process to approximate a value, it is best that we understand as much as possible about the **error** in the procedure. In this case, the error in the linear approximation is

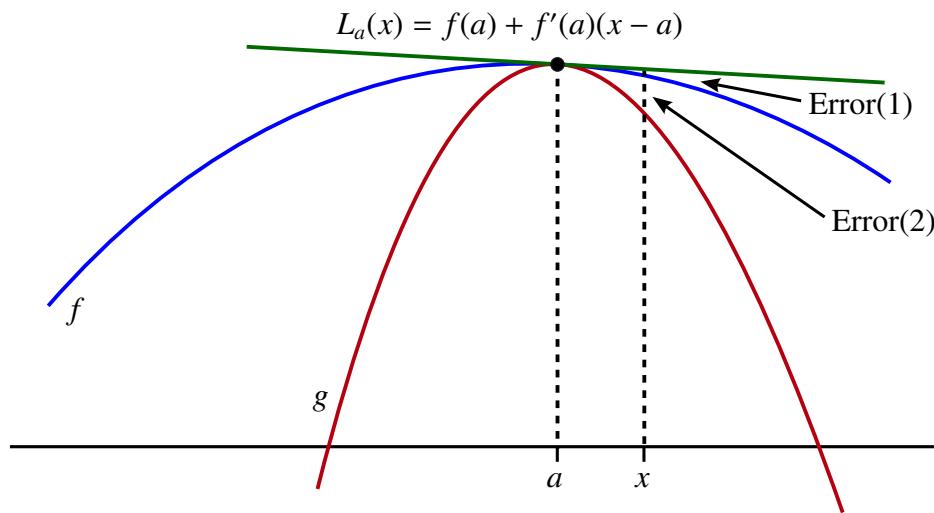
$$\text{Error}(x) = |f(x) - L_a(x)|$$

and at $x = a$ the estimate is exact since $L_a(a) = f(a)$.

There are two basic factors that affect the potential size of the error in using linear approximation. These are

1. The distance between x and a . That is, how large is $|x - a|$?
2. How curved the graph is near $x = a$?

Note that the larger $|f''(x)|$ is, the more rapidly the tangent lines turn, and hence the more curved the graph of f . For this reason the second factor affecting the size of the error can be expressed in terms of the size of $|f''(x)|$. Generally speaking, the further x is away from a and the more curved the graph of f , the larger the potential for error in using linear approximation. This is illustrated in the following diagram which shows two different functions, f and g , with the same tangent line at $x = a$. The error in using the linear approximation is the length of the vertical line joining the graph of the function and the graph of the linear approximation.



Notice that in the diagram, the graph of g is much more curved near $x = a$ than is the graph of f . You can also see that at the chosen point x the error

$$\text{Error}(1) = |f(x) - L_a(x)|$$

in using $L_a(x)$ to estimate the value of $f(x)$ is extremely small, whereas the error

$$\text{Error}(2) = |g(x) - L_a(x)|$$

in using $L_a(x)$ to estimate the value of $g(x)$ is noticeably larger. The diagram also shows that for both f and g , the further away x is from a , the larger the error is in the linear approximation process.

In the case of the function g , its graph looks more like a parabola (second degree polynomial) than it does a line. This suggests that it would make more sense to try and approximate g with a function of the form

$$p(x) = c_0 + c_1(x - a) + c_2(x - a)^2.$$

(Notice that the form for this polynomial looks somewhat unusual. You will see that we write it this way because this form makes it easier to properly encode the information about f at $x = a$).

In constructing the linear approximation, we encoded the value of the function and of its derivative at the point $x = a$. We want to again encode this local information, but we want to do more. If we can include the second derivative, we might be able to capture the curvature of the function that was missing in the linear approximation. In summary, we would like to find constants c_0 , c_1 , and c_2 , so that

1. $p(a) = f(a)$,
2. $p'(a) = f'(a)$, and
3. $p''(a) = f''(a)$.

It may not seem immediately obvious that we can find such constants. However, this task is actually not too difficult. For example, if we want $p(a) = f(a)$, then by noting that

$$p(a) = c_0 + c_1(a - a) + c_2(a - a)^2 = c_0$$

we immediately know that we should let $c_0 = f(a)$.

We can use the standard rules of differentiation to show that

$$p'(x) = c_1 + 2c_2(x - a).$$

In order that $p'(a) = f'(a)$, we have

$$f'(a) = p'(a) = c_1 + 2c_2(a - a) = c_1.$$

Finally, since

$$p''(x) = 2c_2$$

for all x , if we let $c_2 = \frac{f''(a)}{2}$, we have

$$p''(a) = 2c_2 = 2\left(\frac{f''(a)}{2}\right) = f''(a)$$

exactly as required. This shows that if

$$p(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2}(x - a)^2,$$

then p is the unique polynomial of degree 2 or less such that

1. $p(a) = f(a)$,
2. $p'(a) = f'(a)$, and
3. $p''(a) = f''(a)$.

The polynomial p is called the *second degree Taylor polynomial for f centered at $x = a$* . We denote this Taylor polynomial by $T_{2,a}$.

EXAMPLE 1 Let $f(x) = \cos(x)$. Then,

$$f(0) = \cos(0) = 1, \quad \text{and} \quad f'(0) = -\sin(0) = 0,$$

and

$$f''(0) = -\cos(0) = -1.$$

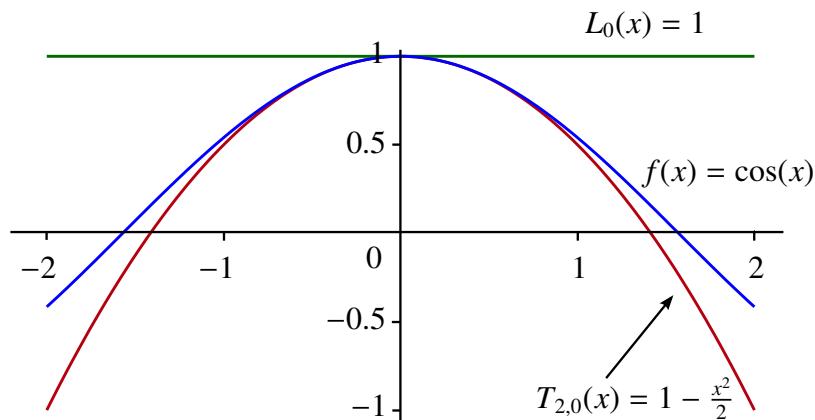
It follows that

$$L_0(x) = f(0) + f'(0)(x - 0) = 1 + 0(x - 0) = 1$$

for all x while

$$\begin{aligned}
 T_{2,0}(x) &= f(0) + f'(0)(x - 0) + \frac{f''(0)}{2}(x - 0)^2 \\
 &= 1 + 0(x - 0) + \frac{-1}{2}(x - 0)^2 \\
 &= 1 - \frac{x^2}{2}.
 \end{aligned}$$

The following diagram shows $\cos(x)$ with its linear approximation and its second degree Taylor polynomial centered at $x = 0$.



Notice that the second degree Taylor polynomial $T_{2,0}$ does a much better job approximating $\cos(x)$ over the interval $[-2, 2]$ than does the linear approximation L_0 .

We might guess that if f has a third derivative at $x = a$, then by encoding the value $f'''(a)$ along with $f(a)$, $f'(a)$ and $f''(a)$, we may do an even better job of approximating $f(x)$ near $x = a$ than we did with either L_a or with $T_{2,a}$. As such we would be looking for a polynomial of the form

$$p(x) = c_0 + c_1(x - a) + c_2(x - a)^2 + c_3(x - a)^3$$

such that

1. $p(a) = f(a)$,
2. $p'(a) = f'(a)$,
3. $p''(a) = f''(a)$, and
4. $p'''(a) = f'''(a)$.

To find such a p , we follow the same steps that we outlined before. We want $p(a) = f(a)$, but $p(a) = c_0 + c_1(a - a) + c_2(a - a)^2 + c_3(a - a)^3 = c_0$, so we can let $c_0 = f(a)$.

Differentiating p we get

$$p'(x) = c_1 + 2c_2(x - a) + 3c_3(x - a)^2$$

so that

$$p'(a) = c_1 + 2c_2(a - a) + 3c_3(a - a)^2 = c_1.$$

Therefore, if we let $c_1 = f'(a)$ as before, then we will get $p'(a) = f'(a)$.

Differentiating p' gives us

$$p''(x) = 2c_2 + 3(2)c_3(x - a).$$

Therefore,

$$p''(a) = 2c_2 + 3(2)c_3(a - a) = 2c_2.$$

Now if we let $c_2 = \frac{f''(a)}{2}$, we get

$$p''(a) = f''(a).$$

Finally, observe that

$$p'''(x) = 3(2)c_3 = 3(2)(1)c_3 = 3!c_3$$

for all x , so if we require

$$p'''(a) = 3!c_3 = f'''(a),$$

then we need only let $c_3 = \frac{f'''(a)}{3!}$.

It follows that if

$$p(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2}(x - a)^2 + \frac{f'''(a)}{3!}(x - a)^3,$$

then

1. $p(a) = f(a)$,
2. $p'(a) = f'(a)$,
3. $p''(a) = f''(a)$, and
4. $p'''(a) = f'''(a)$.

In this case, we call p the *third degree Taylor polynomial centered at $x = a$* and denote it by $T_{3,a}$.

Given a function f , we could also write

$$T_{0,a}(x) = f(a)$$

and

$$T_{1,a}(x) = L_a(x) = f(a) + f'(a)(x - a)$$

and call these polynomials the *zero-th degree* and the *first degree Taylor polynomials of f centered at $x = a$* , respectively.

Observe that using the convention where $0! = 1! = 1$ and $(x - a)^0 = 1$, we have the following:

$$\begin{aligned} T_{0,a}(x) &= \frac{f(a)}{0!}(x - a)^0 \\ T_{1,a}(x) &= \frac{f(a)}{0!}(x - a)^0 + \frac{f'(a)}{1!}(x - a)^1 \\ T_{2,a}(x) &= \frac{f(a)}{0!}(x - a)^0 + \frac{f'(a)}{1!}(x - a)^1 + \frac{f''(a)}{2!}(x - a)^2 \\ T_{3,a}(x) &= \frac{f(a)}{0!}(x - a)^0 + \frac{f'(a)}{1!}(x - a)^1 + \frac{f''(a)}{2!}(x - a)^2 + \frac{f'''(a)}{3!}(x - a)^3. \end{aligned}$$

Recall that $f^{(k)}(a)$ denotes the k -th derivative of f at $x = a$. By convention, $f^{(0)}(x) = f(x)$. Then using summation notation, we have

$$\begin{aligned} T_{0,a}(x) &= \sum_{k=0}^0 \frac{f^{(k)}(a)}{k!}(x - a)^k \\ T_{1,a}(x) &= \sum_{k=0}^1 \frac{f^{(k)}(a)}{k!}(x - a)^k \\ T_{2,a}(x) &= \sum_{k=0}^2 \frac{f^{(k)}(a)}{k!}(x - a)^k \end{aligned}$$

and

$$T_{3,a}(x) = \sum_{k=0}^3 \frac{f^{(k)}(a)}{k!}(x - a)^k.$$

This leads us to the following definition:

DEFINITION

Taylor Polynomials

Assume that f is n -times differentiable at $x = a$. The *n -th degree Taylor polynomial for f centered at $x = a$* is the polynomial

$$\begin{aligned} T_{n,a}(x) &= \sum_{k=0}^n \frac{f^{(k)}(a)}{k!}(x - a)^k \\ &= f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n \end{aligned}$$

NOTE

A remarkable property about $T_{n,a}$ is that for any k between 0 and n ,

$$T_{n,a}^{(k)}(a) = f^{(k)}(a).$$

That is, $T_{n,a}$ encodes not only the value of $f(x)$ at $x = a$ but all of its first n derivatives as well. Moreover, this is the *only* polynomial of degree n or less that does so! 

EXAMPLE 2 Find all of the Taylor polynomials up to degree 5 for the function $f(x) = \cos(x)$ with center $x = 0$.

We have already seen that $f(0) = \cos(0) = 1$, $f'(0) = -\sin(0) = 0$, and $f''(0) = -\cos(0) = -1$. It follows that

$$T_{0,0}(x) = 1,$$

and

$$T_{1,0}(x) = L_0(x) = 1 + 0(x - 0) = 1$$

for all x , while

$$T_{2,0} = 1 + 0(x - 0) + \frac{-1}{2!}(x - 0)^2 = 1 - \frac{x^2}{2}.$$

Since $f'''(x) = \sin(x)$, $f^{(4)}(x) = \cos(x)$, and $f^{(5)}(x) = -\sin(x)$, we get $f'''(0) = \sin(0) = 0$, $f^{(4)}(0) = \cos(0) = 1$ and $f^{(5)}(0) = -\sin(0) = 0$. Hence,

$$\begin{aligned} T_{3,0}(x) &= 1 + 0(x - 0) + \frac{-1}{2!}(x - 0)^2 + \frac{0}{3!}(x - 0)^3 \\ &= 1 - \frac{x^2}{2} \\ &= T_{2,0}(x) \end{aligned}$$

We also have that

$$\begin{aligned} T_{4,0}(x) &= 1 + 0(x - 0) + \frac{-1}{2!}(x - 0)^2 + \frac{0}{3!}(x - 0)^3 + \frac{1}{4!}(x - 0)^4 \\ &= 1 - \frac{x^2}{2} + \frac{x^4}{24} \end{aligned}$$

and

$$\begin{aligned} T_{5,0}(x) &= 1 + 0(x - 0) + \frac{-1}{2!}(x - 0)^2 + \frac{0}{3!}(x - 0)^3 + \frac{1}{4!}(x - 0)^4 + \frac{0}{5!}(x - 0)^5 \\ &= 1 - \frac{x^2}{2} + \frac{x^4}{24} \\ &= T_{4,0}(x) \end{aligned}$$

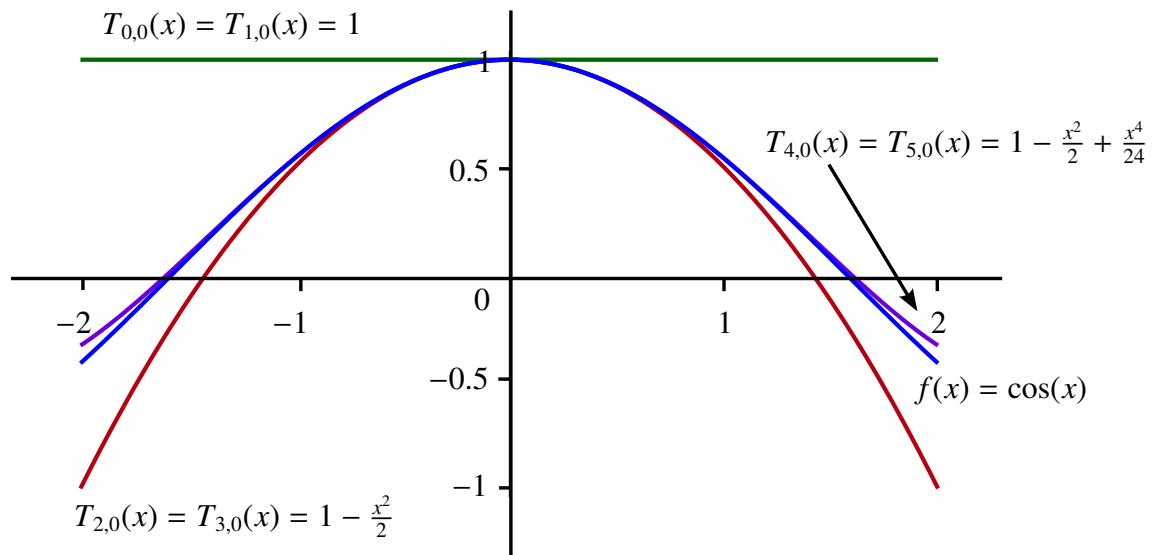


An important observation to make is that not all of these polynomials are distinct. In fact, $T_{0,0}(x) = T_{1,0}(x)$, $T_{2,0}(x) = T_{3,0}(x)$, and $T_{4,0}(x) = T_{5,0}(x)$. In general, this equality of different order Taylor polynomials happens when one of the derivatives is 0 at $x = a$. (In this example at $x = 0$.) This can be seen by observing that for any n

$$T_{n+1,a}(x) = T_{n,a}(x) + \frac{f^{(n+1)}(a)}{(n+1)!}(x-a)^{n+1}$$

so if $f^{(n+1)}(a) = 0$, we get $T_{n+1,a}(x) = T_{n,a}(x)$.

The following diagram shows $\cos(x)$ and its Taylor polynomials up to degree 5. You will notice that there are only four distinct graphs.



In the next example, we will calculate the Taylor Polynomials for $f(x) = \sin(x)$.

EXAMPLE 3 Find all of the Taylor polynomials up to degree 5 for the function $f(x) = \sin(x)$ with center $x = 0$.

We can see that $f(0) = \sin(0) = 0$, $f'(0) = \cos(0) = 1$, $f''(0) = -\sin(0) = 0$, $f'''(0) = -\cos(0) = -1$, $f^{(4)}(0) = \sin(0) = 0$, and $f^{(5)}(0) = \cos(0) = 1$. It follows that

$$T_{0,0}(x) = 0,$$

and

$$T_{1,0}(x) = L_0(x) = 0 + 1(x - 0) = x$$

and

$$\begin{aligned} T_{2,0}(x) &= 0 + 1(x - 0) + \frac{0}{2!}(x - 0)^2 \\ &= x \\ &= T_{1,0}(x). \end{aligned}$$

Next we have

$$\begin{aligned} T_{3,0}(x) &= 0 + 1(x - 0) + \frac{0}{2!}(x - 0)^2 + \frac{-1}{3!}(x - 0)^3 \\ &= x - \frac{x^3}{6} \end{aligned}$$

and that

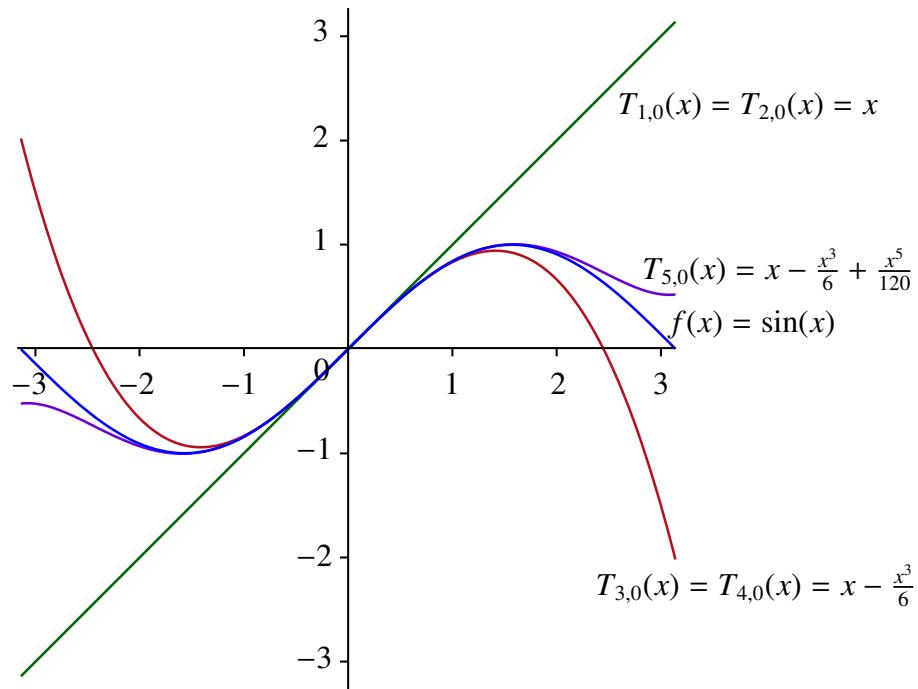
$$\begin{aligned} T_{4,0}(x) &= 0 + 1(x - 0) + \frac{0}{2!}(x - 0)^2 + \frac{-1}{3!}(x - 0)^3 + \frac{0}{4!}(x - 0)^4 \\ &= x - \frac{x^3}{6} \\ &= T_{3,0}(x). \end{aligned}$$

Finally,

$$\begin{aligned} T_{5,0}(x) &= 0 + 1(x - 0) + \frac{0}{2!}(x - 0)^2 + \frac{-1}{3!}(x - 0)^3 + \frac{0}{4!}(x - 0)^4 + \frac{1}{5!}(x - 0)^5 \\ &= x - \frac{x^3}{6} + \frac{x^5}{5!} \\ &= x - \frac{x^3}{6} + \frac{x^5}{120}. \end{aligned}$$



The following diagram includes the graph of $\sin(x)$ with its Taylor polynomials up to degree 5, excluding $T_{0,0}$ since its graph is the x -axis.

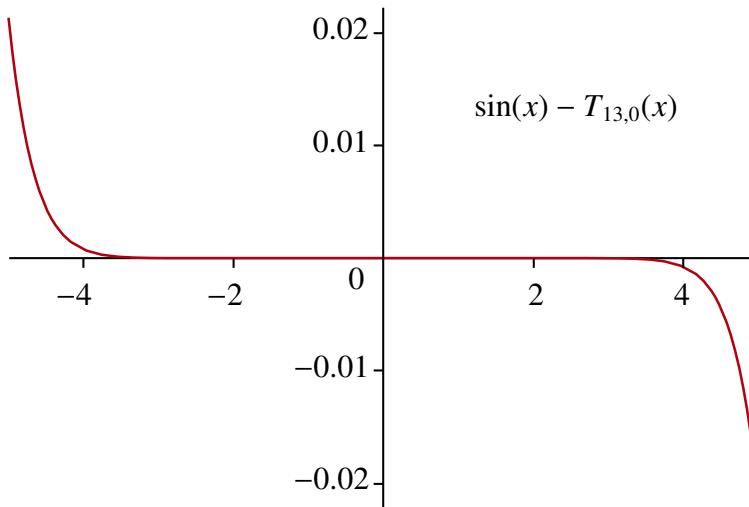


Notice again that the polynomials are not distinct though, in general, as the degree increases so does the accuracy of the estimate near $x = 0$.

To illustrate the power of using Taylor polynomials to approximate functions, we can use a computer to aid us in showing that for $f(x) = \sin(x)$ and $a = 0$, we have

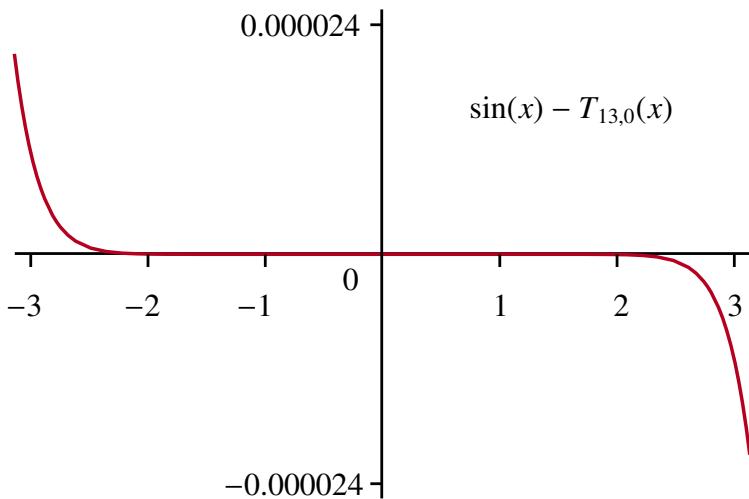
$$T_{13,0}(x) = x - \frac{1}{6}x^3 + \frac{1}{120}x^5 - \frac{1}{5040}x^7 + \frac{1}{362880}x^9 - \frac{1}{39916800}x^{11} + \frac{1}{6227020800}x^{13}$$

The next diagram represents a plot of the function $\sin(x) - T_{13,0}(x)$. (This represents the error between the actual value of $\sin(x)$ and the approximated value of $T_{13,0}(x)$.)



Notice that the error is very small until x approaches 4 or -4 . However, the y -scale is different from that of the x -axis, so even near $x = 4$ or $x = -4$ the actual error is still quite small. The diagram suggests that on the slightly more restrictive interval $[-\pi, \pi]$, $T_{13,0}(x)$ does an exceptionally good job of approximating $\sin(x)$.

To strengthen this point even further, we have provided the plot of the graph of $\sin(x) - T_{13,0}(x)$ on the interval $[-\pi, \pi]$.



Note again the scale for the y -axis. It is clear that near 0, $T_{13,0}(x)$ and $\sin(x)$ are essentially indistinguishable. In fact, we will soon have the tools to show that for $x \in [-1, 1]$,

$$|\sin(x) - T_{13,0}(x)| < 10^{-12}$$

while for $x \in [-0.01, 0.01]$,

$$|\sin(x) - T_{13,0}(x)| < 10^{-42}.$$

Indeed, in using $T_{13,0}(x)$ to estimate $\sin(x)$ for very small values of x , round-off errors and the limitations of the accuracy in floating-point arithmetic become much more significant than the true difference between the functions.

EXAMPLE 4

The function $f(x) = e^x$ is particularly well-suited to the process of creating estimates using Taylor polynomials. This is because for any k , the k -th derivative of e^x is again e^x . This means that for any n ,

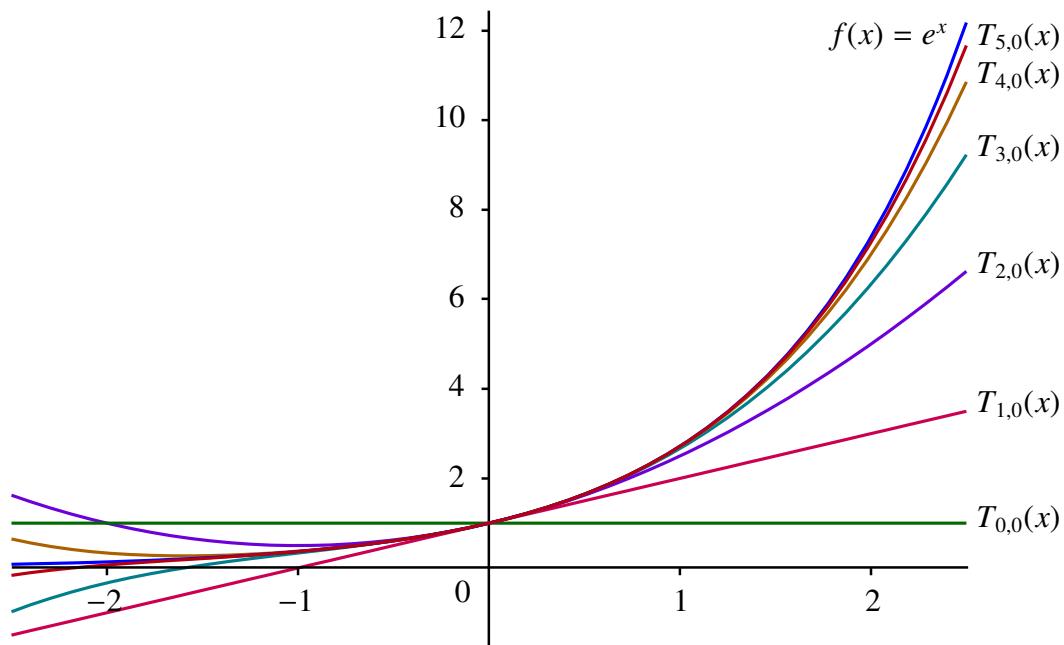
$$\begin{aligned} T_{n,0}(x) &= \sum_{k=0}^n \frac{f^{(k)}(a)}{k!} (x-a)^k \\ &= \sum_{k=0}^n \frac{e^0}{k!} (x-0)^k \\ &= \sum_{k=0}^n \frac{x^k}{k!}. \end{aligned}$$

In particular,

$$\begin{aligned} T_{0,0}(x) &= 1, \\ T_{1,0}(x) &= 1 + x, \\ T_{2,0}(x) &= 1 + x + \frac{x^2}{2}, \\ T_{3,0}(x) &= 1 + x + \frac{x^2}{2} + \frac{x^3}{6}, \\ T_{4,0}(x) &= 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24}, \text{ and} \\ T_{5,0}(x) &= 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120}. \end{aligned}$$

Observe that in the case of e^x , the Taylor polynomials are distinct since e^x , and hence all of its derivatives, is never 0. ◀

The next diagram shows the graphs of e^x and its Taylor polynomials up to degree 5.



8.2 Taylor's Theorem and Errors in Approximations

We have seen that using linear approximation and higher order Taylor polynomials enable us to approximate potentially complicated functions with much simpler ones with surprising accuracy. However, up until now we have only had qualitative information about the behavior of the potential error. We saw that the error in using Taylor polynomials to approximate a function seems to depend on how close we are to the center point. We have also seen that the error in linear approximation seems to depend on the potential size of the second derivative and that the approximations seem to improve as we encode more local information. However, we do not have any precise mathematical statements to substantiate these claims. In this section, we will correct this deficiency by introducing an upgraded version of the Mean Value Theorem called *Taylor's Theorem*.

We begin by introducing some useful notation.

DEFINITION

Taylor Remainder

Assume that f is n times differentiable at $x = a$. Let

$$R_{n,a}(x) = f(x) - T_{n,a}(x).$$

$R_{n,a}(x)$ is called the n -th degree Taylor remainder function centered at $x = a$.

The error in using the Taylor polynomial to approximate f is given by

$$\text{Error} = |R_{n,a}(x)|.$$

The following is the central problem for this approximation process.

Problem: Given a function f and a point $x = a$, how do we estimate the size of $R_{n,a}(x)$?

The following theorem provides us with the answer to this question.

THEOREM 1

Taylor's Theorem

Assume that f is $n + 1$ -times differentiable on an interval I containing $x = a$. Let $x \in I$. Then there exists a point c between x and a such that

$$f(x) - T_{n,a}(x) = R_{n,a}(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1}.$$

PROOF

Let $x \in I$ be such that $x \neq a$. Then there exists an M such that

$$R_{n,a}(x) = f(x) - T_{n,a}(x) = M(x-a)^{n+1}$$

Let

$$F(t) = f(t) + f'(t)(x-t) + \frac{f''(t)}{2!}(x-t)^2 + \cdots + \frac{f^{(n)}(t)}{n!}(x-t)^n + M(x-t)^{n+1}$$

Notice that $F(x) = f(x) = F(a)$. By the MVT, there exists some c between x and a such that $F'(c) = 0$.

We have that

$$\frac{d}{dt}\left(\frac{f^{(k)}(t)}{k!}(x-t)^k\right) = -\frac{f^{(k)}(t)}{(k-1)!}(x-t)^{(k-1)} + \frac{f^{(k+1)}(t)}{k!}(x-t)^k.$$

It follows that

$$F'(t) = \frac{f^{(n+1)}(t)}{n!}(x-t)^n - M(n+1)(x-t)^n.$$

This means that

$$0 = F'(c) = \frac{f^{(n+1)}(c)}{n!}(x-c)^n - M(n+1)(x-c)^n.$$

Solving for M yields that

$$M = \frac{f^{(n+1)}(c)}{(n+1)!}$$

exactly as desired. ■

We will make **three important observations about Taylor's theorem.**

- 1) First, since $T_{1,a}(x) = L_a(x)$, when $n = 1$ the absolute value of the remainder $R_{1,a}(x)$ represents the error in using the linear approximation. Taylor's Theorem shows that for some c ,

$$|R_{1,a}(x)| = \left| \frac{f''(c)}{2}(x-a)^2 \right|.$$

This shows explicitly how the error in linear approximation depends on the potential size of $f''(x)$ and on $|x-a|$, the distance from x to a .

- 2) The second observation involves the case when $n = 0$. In this case, the theorem requires that f be differentiable on I and its conclusion states that for any $x \in I$ there exists a point c between x and a such that

$$f(x) - T_{0,a}(x) = f'(c)(x-a).$$

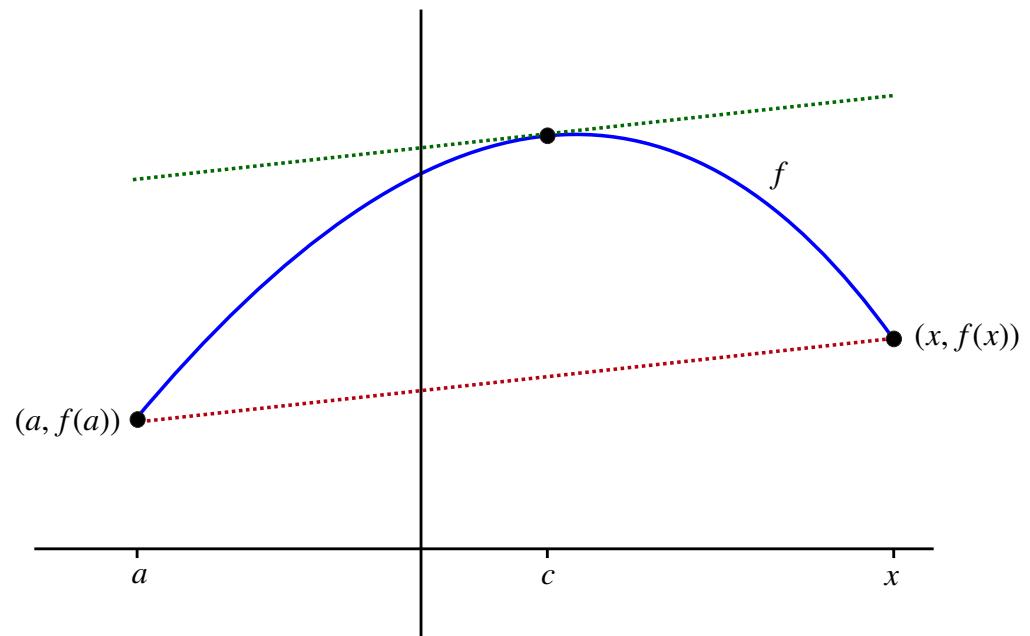
But $T_{0,a}(x) = f(a)$, so we have

$$f(x) - f(a) = f'(c)(x-a).$$

Dividing by $x-a$ shows that there is a point c between x and a such that

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

This is exactly the statement of the Mean Value Theorem. Therefore, Taylor's Theorem is really a higher-order version of the MVT.



- 3) Finally, Taylor's Theorem does not tell us how to find the point c , but rather that such a point exists. It turns out that for the theorem to be of any value, we really need to be able to say something intelligent about how large $|f^{(n+1)}(c)|$ might be without knowing c . For an arbitrary function, this might be a difficult task since higher order derivatives have a habit of being very complicated. However, the good news is that for some of the most important functions in mathematics, such as $\sin(x)$, $\cos(x)$, and e^x , we can determine roughly how large $|f^{(n+1)}(c)|$ might be and in so doing, show that the estimates obtained for these functions can be extremely accurate.

EXAMPLE 5 Use linear approximation to estimate $\sin(.01)$ and show that the error in using this approximation is less than 10^{-4} .

SOLUTION We know that $f(0) = \sin(0) = 0$ and that $f'(0) = \cos(0) = 1$, so

$$L_0(x) = T_{1,0}(x) = x.$$

Therefore, the estimate we obtain for $\sin(.01)$ using linear approximation is

$$\sin(.01) \cong L_0(.01) = .01$$

Taylor's Theorem applies since $\sin(x)$ is always differentiable. Moreover, if $f(x) = \sin(x)$, then $f'(x) = \cos(x)$ and $f''(x) = -\sin(x)$. It follows that there exists some c between 0 and .01 such that the error in the linear approximation is given by

$$|R_{1,0}(.01)| = \left| \frac{f''(c)}{2} (.01 - 0)^2 \right| = \left| \frac{-\sin(c)}{2} (.01)^2 \right|$$

Recall that the theorem does not tell us the value of c , but rather just that it exists. Not knowing the value of c may seem to make it impossible to say anything significant about the error, but this is actually not the case. The key observation in this example is that regardless the value of point c , $|\sin(c)| \leq 1$. Therefore,

$$\begin{aligned} |R_{1,0}(.01)| &= \left| \frac{-\sin(c)}{2} (.01)^2 \right| \\ &\leq \frac{1}{2} (.01)^2 \\ &< 10^{-4}. \end{aligned}$$

This simple process seems to be remarkably accurate. In fact, it turns out that this estimate is actually much better than the calculation suggests. This is true because not only does $T_{1,0}(x) = x$, but we also have that $T_{2,0}(x) = T_{1,0}(x) = x$. This means

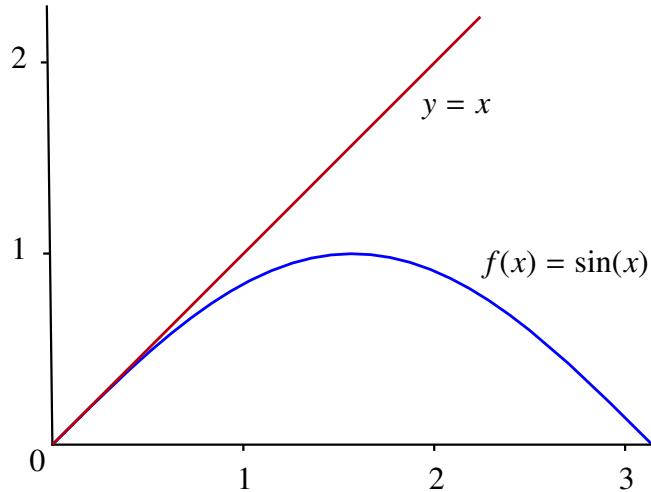
that there is a new number c between 0 and .01 such that

$$\begin{aligned}
 |\sin(.01) - .01| &= |R_{2,0}(.01)| \\
 &= \left| \frac{f'''(c)}{6} (.01 - 0)^3 \right| \\
 &= \left| \frac{-\cos(c)}{6} (.01)^3 \right| \\
 &< 10^{-6}
 \end{aligned}$$

since $|\cos(c)| \leq 1$ for all values of c .

This shows that the estimate $\sin(.01) \approx .01$ is accurate to six decimal places. In fact, the actual error is approximately $-1.666658333 \times 10^{-7}$.

Finally, we know that for $0 < x < \frac{\pi}{2}$, the tangent line to the graph of $f(x) = \sin(x)$ is above the graph of f since $\sin(x)$ is concave downward on this interval. (In fact, the Mean Value Theorem can be used to show that $\sin(x) \leq x$ for every $x \geq 0$.) Since the tangent line is the graph of the linear approximation, this means that our estimate is actually too large.



Taylor's Theorem can be used to confirm this because

$$\sin(x) - x = R_{1,0}(x) = \frac{-\sin(c)}{2} (x)^2 < 0$$

since $\sin(c) > 0$ for any $c \in (0, \frac{\pi}{2})$. ◀

In the next example we will see how Taylor's Theorem can help in calculating various limits. In order to simplify the notation, we will only consider limits as $x \rightarrow 0$.

EXAMPLE 6 Find $\lim_{x \rightarrow 0} \frac{\sin(x) - x}{x^2}$.

SOLUTION First notice that this is an indeterminate limit of the type $\frac{0}{0}$.

We know that if $f(x) = \sin(x)$, then $T_{1,0}(x) = T_{2,0}(x) = x$. We will assume that we are working with $T_{2,0}$. Then Taylor's Theorem shows that for any $x \in [-1, 1]$, there exists a c between 0 and x such that

$$|\sin(x) - x| = \left| \frac{-\cos(c)}{3!} x^3 \right| \leq \frac{1}{6} |x|^3$$

since $|\cos(c)| \leq 1$ regardless where c is located. This inequality is equivalent to

$$\frac{-1}{6} |x|^3 \leq \sin(x) - x \leq \frac{1}{6} |x|^3.$$

If $x \neq 0$, we can divide all of the terms by x^2 to get that for $x \in [-1, 1]$

$$\frac{-|x|^3}{6x^2} \leq \frac{\sin(x) - x}{x^2} \leq \frac{|x|^3}{6x^2}$$

or equivalently that

$$\frac{-|x|}{6} \leq \frac{\sin(x) - x}{x^2} \leq \frac{|x|}{6}.$$

We also know that

$$\lim_{x \rightarrow 0} \frac{-|x|}{6} = \lim_{x \rightarrow 0} \frac{|x|}{6} = 0$$

The Squeeze Theorem guarantees that

$$\lim_{x \rightarrow 0} \frac{\sin(x) - x}{x^2} = 0.$$



The technique we outlined in the previous example can be used in much more generality. However, we require the following observation.

Suppose that $f^{(k+1)}$ is a continuous function on $[-1, 1]$. Then so is the function

$$g(x) = \left| \frac{f^{(k+1)}(x)}{(k+1)!} \right|.$$

The Extreme Value Theorem tells us that g has a maximum on $[-1, 1]$. Therefore, there is an M such that

$$\left| \frac{f^{(k+1)}(x)}{(k+1)!} \right| \leq M$$

for all $x \in [-1, 1]$.

Let $x \in [-1, 1]$. Taylor's Theorem assures us that there is a c between x and 0 such that

$$|R_{k,0}(x)| = \left| \frac{f^{(k+1)}(c)}{(k+1)!} x^{k+1} \right|.$$

Therefore,

$$\begin{aligned} |f(x) - T_{k,0}(x)| &= |R_{k,0}(x)| \\ &= \left| \frac{f^{(k+1)}(c)}{(k+1)!} x^{k+1} \right| \\ &\leq M |x|^{k+1} \end{aligned}$$

since c is also in $[-1, 1]$.

It follows that

$$-M |x|^{k+1} \leq f(x) - T_{k,0}(x) \leq M |x|^{k+1}.$$

We summarize this technique as follows:

THEOREM 2

Taylor's Approximation Theorem I

Assume that $f^{(k+1)}$ is continuous on $[-d, d]$ for $d > 0$. Then there exists a constant $M > 0$ such that

$$|f(x) - T_{k,0}(x)| \leq M |x|^{k+1}$$

or equivalently that

$$-M |x|^{k+1} \leq f(x) - T_{k,0}(x) \leq M |x|^{k+1}$$

for each $x \in [-d, d]$.

This theorem is very helpful in calculating many limits.

EXAMPLE 7

Calculate $\lim_{x \rightarrow 0} \frac{\cos(x) - 1}{x^2}$.

SOLUTION We know that for $f(x) = \cos(x)$ we have $T_{2,0} = 1 - \frac{x^2}{2}$. Moreover, all of the derivatives of $\cos(x)$ are continuous everywhere. The Taylor Approximation Theorem tells us that there is a constant M such that

$$-M |x|^3 \leq \cos(x) - \left(1 - \frac{x^2}{2}\right) \leq M |x|^3$$

for all $x \in [-1, 1]$. Dividing by x^2 with $x \neq 0$ we have that

$$-M |x| \leq \frac{\cos(x) - \left(1 - \frac{x^2}{2}\right)}{x^2} \leq M |x|$$

for all $x \in [-1, 1]$. Simplifying the previous expression produces

$$-M|x| \leq \frac{\cos(x) - 1}{x^2} + \frac{1}{2} \leq M|x|$$

for all $x \in [-1, 1]$.

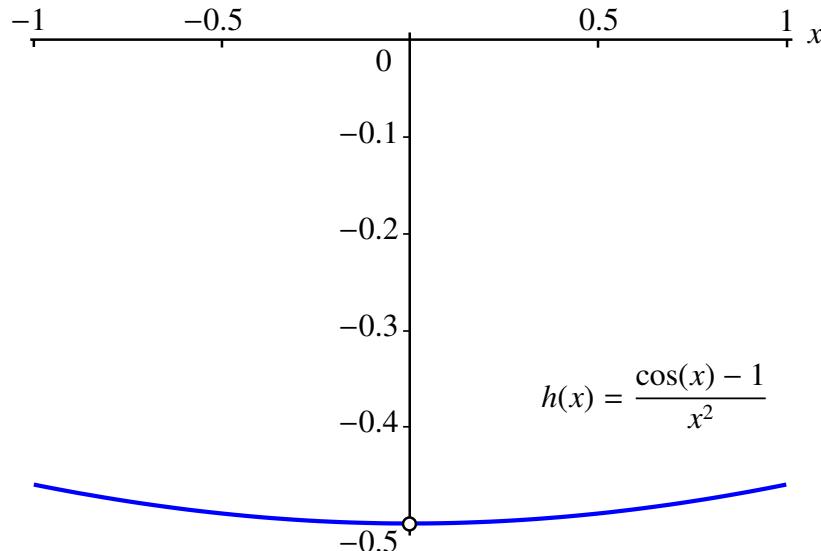
Applying the Squeeze Theorem we have that

$$\lim_{x \rightarrow 0} \frac{\cos(x) - 1}{x^2} + \frac{1}{2} = 0$$

which is equivalent to

$$\lim_{x \rightarrow 0} \frac{\cos(x) - 1}{x^2} = \frac{-1}{2}.$$

This limit is consistent with the behavior of the function $h(x) = \frac{\cos(x) - 1}{x^2}$ near 0. This is illustrated in the following graph.



The previous limit can actually be calculated quite easily using L'Hôpital's Rule. As an exercise, you should try to verify the answer using this rule. The next example would require much more work using L'Hôpital's Rule. It is provided to show you how powerful Taylor's Theorem can be for finding limits.

EXAMPLE 8 Find $\lim_{x \rightarrow 0} \frac{e^{\frac{x^4}{2}} - \cos(x^2)}{x^4}$.

SOLUTION This is an indeterminate limit of type $\frac{0}{0}$. We know from Taylor's Approximation Theorem that we can find a constant M_1 such that for any $u \in [-1, 1]$

$$-M_1 u^2 \leq e^u - (1 + u) \leq M_1 u^2$$

since $1 + u$ is the first degree Taylor polynomial of e^u . Now if $x \in [-1, 1]$, then $u = \frac{x^4}{2} \in [-1, 1]$. In fact, $u \in [0, \frac{1}{2}]$. It follows that if $x \in [-1, 1]$ and we substitute $u = \frac{x^4}{2}$, then we get

$$\frac{-M_1 x^8}{4} \leq e^{\frac{x^4}{2}} - (1 + \frac{x^4}{2}) \leq \frac{M_1 x^8}{4}.$$

We also can show that there exists a constant M_2 such that for any $v \in [-1, 1]$

$$-M_2 v^4 \leq \cos(v) - (1 - \frac{v^2}{2}) \leq M_2 v^4$$

since $1 - \frac{v^2}{2}$ is the third degree Taylor polynomial for $\cos(v)$.

If $x \in [-1, 1]$ then so is x^2 . If we let $v = x^2$, then we see that

$$-M_2 x^8 \leq \cos(x^2) - (1 - \frac{x^4}{2}) \leq M_2 x^8.$$

The next step is to multiply each term in the previous inequality by -1 to get

$$-M_2 x^8 \leq (1 - \frac{x^4}{2}) - \cos(x^2) \leq M_2 x^8.$$

(Remember, multiplying by a negative number reverses the inequality.)

Now add the two inequalities together:

$$-(\frac{M_1}{4} + M_2)x^8 \leq e^{\frac{x^4}{2}} - (1 + \frac{x^4}{2}) + (1 - \frac{x^4}{2}) - \cos(x^2) \leq (\frac{M_1}{4} + M_2)x^8.$$

If we let $M = \frac{M_1}{4} + M_2$ and simplify, this inequality becomes

$$-Mx^8 \leq e^{\frac{x^4}{2}} - \cos(x^2) - x^4 \leq Mx^8$$

for all $x \in [-1, 1]$. Dividing by x^4 gives us that

$$-Mx^4 \leq \frac{e^{\frac{x^4}{2}} - \cos(x^2)}{x^4} - 1 \leq Mx^4.$$

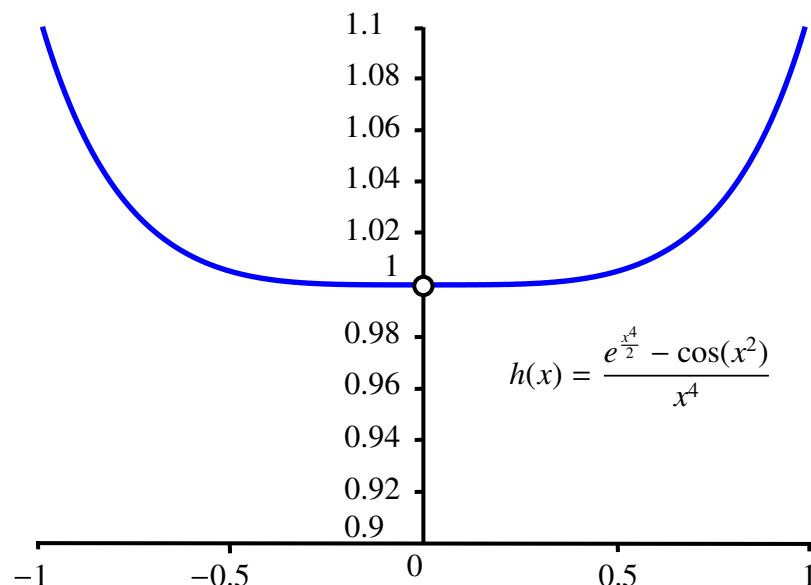
The final step is to apply the Squeeze Theorem to show that

$$\lim_{x \rightarrow 0} \frac{e^{\frac{x^4}{2}} - \cos(x^2)}{x^4} - 1 = 0$$

or equivalently that

$$\lim_{x \rightarrow 0} \frac{e^{\frac{x^4}{2}} - \cos(x^2)}{x^4} = 1.$$

This limit can be confirmed visually from the graph of the function $h(x) = \frac{e^{\frac{x^4}{2}} - \cos(x^2)}{x^4}$.



The previous example involved a rather complicated argument. However, with a little practice using Taylor polynomials and the mastery of a few techniques, limits like this can actually be done by inspection!

8.3 Big-O

Suppose that we know that

$$\lim_{x \rightarrow 0} f(x) = 0.$$

One question we might ask is: How quickly does $f(x)$ approach 0? For example, consider

$$\lim_{x \rightarrow 0} x^2 = 0 = \lim_{x \rightarrow 0} x^{17}.$$

It is easy to rationalize that x^{17} approaches 0 much more quickly than x^2 as x nears 0. In this section we will introduce notation to reflect the relative orders of magnitude of two functions, f and g , and use it to address this question. In particular, we will focus our attention on the case where a function f is of an order of magnitude no greater than x^n when x is near 0.

We begin with the following definition:

DEFINITION

Big-O Notation

We say that f is Big-O of g as $x \rightarrow a$ if there exists an $\epsilon > 0$ and an $M > 0$ such that

$$|f(x)| \leq M|g(x)|$$

for all $x \in (a - \epsilon, a + \epsilon)$ except possibly at $x = a$.

In this case, we write

$$f(x) = O(g(x)) \text{ as } x \rightarrow a$$

or simply $f(x) = O(g(x))$ if a is understood.

If f is Big-O of g as $x \rightarrow a$, then we say that $f(x)$ has *order of magnitude that is less than or equal to that of $g(x)$ near $x = a$* .

REMARK

In the definition above, once we find any positive ϵ that works, so will any smaller value of ϵ . As such we can always insist that $0 < \epsilon \leq 1$. 

Big-O notation is due to the German mathematician Edmund Landau. While we will use the notation to investigate the behavior of functions near a point, the notation is also commonly used in computer science for the analysis of the complexity of algorithms.

NOTE

In the following applications, we will always use $a = 0$ and $g(x) = x^n$ for some $n \in \mathbb{N}$. 

THEOREM 3

Suppose $f(x) = O(x^n)$ for some $n \in \mathbb{N}$. Then

$$\lim_{x \rightarrow 0} f(x) = 0.$$

PROOF

Suppose $f(x) = O(x^n)$ for some $n \in \mathbb{N}$. This implies that

$$-M|x^n| \leq f(x) \leq M|x^n|$$

on $(-\epsilon, \epsilon)$ except possibly at $x = 0$. Since

$$\lim_{x \rightarrow 0} -M|x^n| = 0 = \lim_{x \rightarrow 0} M|x^n|,$$

the Squeeze Theorem for functions guarantees that

$$\lim_{x \rightarrow 0} f(x) = 0.$$


REMARK

We have shown that every function that is Big-O of x^n as $x \rightarrow 0$ (for some $n \in \mathbb{N}$) converges to 0 as $x \rightarrow 0$. We denote this fact by writing

$$\lim_{x \rightarrow 0} O(x^n) = 0$$

for all $n \in \mathbb{N}$. ◀

Since our main interest will be to compare two functions f and g near a point $x = a$, we will require the following modification of the previous definition:

DEFINITION**Extended Big-O Notation**

Suppose that f , g and h are defined on an open interval containing $x = a$, except possibly at $x = a$. We write

$$f(x) = g(x) + O(h(x)) \text{ as } x \rightarrow a$$

if

$$f(x) - g(x) = O(h(x)) \text{ as } x \rightarrow a.$$

We may omit the $x \rightarrow a$ condition if a is understood.

REMARK

The notation $f(x) = g(x) + O(h(x))$ tells us that $f(x) \approx g(x)$ near $x = a$ with an error that is an order of magnitude at most that of $h(x)$. ◀

EXAMPLE 9

Consider $f(x) = \sin(x)$. Using Taylor's Theorem we get that if $x \in [-1, 1]$, there exists some c between x and 0 such that

$$|\sin(x) - T_{1,0}(x)| = |\sin(x) - x| = \left| \frac{f''(c)}{2!} x^2 \right| = \left| \frac{-\sin(c)}{2} x^2 \right| \leq \frac{1}{2} |x^2|.$$

Hence,

$$\sin(x) - x = O(x^2),$$

so that

$$\sin(x) = x + O(x^2).$$

This gives us a qualitative sense about how well the function $T_{1,0}(x) = x$ approximates the function $f(x) = \sin(x)$ near $x = 0$ by showing that the error is of an order of magnitude of at most x^2 .

In fact, since $T_{1,0}(x) = T_{2,0}(x)$ for $\sin(x)$, we can interpret x as $T_{2,0}(x)$ instead of $T_{1,0}(x)$. If we apply Taylor's Theorem again using $T_{2,0}$, we get that

$$\sin(x) = x + O(x^3).$$

This is a stronger statement because x^3 is an order of magnitude smaller than x^2 near $x = 0$ and as such this shows that the approximation $\sin(x) \approx x$ is even better than was suggested before. ◀

In the previous example we saw that if $f(x) = \sin(x)$, then $f(x) = T_{1,0}(x) + O(x^2)$ and $f(x) = T_{2,0}(x) + O(x^3)$. Both observations arose immediately from Taylor's Theorem. The next theorem shows that this phenomenon may be extended to many other functions.

THEOREM 4

Taylor's Approximation Theorem II

Let $r > 0$. If f is $(n+1)$ -times differentiable on $[-r, r]$ and $f^{(n+1)}$ is continuous on $[-r, r]$, then $f(x) = T_{n,0}(x) + O(x^{n+1})$ as $x \rightarrow 0$.

PROOF

By the Extreme Value Theorem, $f^{(n+1)}$ is bounded on $[-r, r]$. Let M be chosen so that $|f^{(n+1)}(x)| \leq M$ for all $x \in [-r, r]$. Taylor's Theorem implies that for any $x \in [-r, r]$, there exists a c between x and 0 so that

$$|f(x) - T_{n,0}(x)| = \left| \frac{f^{(n+1)}(c)}{(n+1)!} x^{n+1} \right| \leq \left| \frac{M}{(n+1)!} x^{n+1} \right| = \frac{M}{(n+1)!} |x^{n+1}|.$$

This shows that $f(x) - T_{n,0}(x) = O(x^{n+1})$ as $x \rightarrow 0$ and the result of the theorem follows. ■

Question: Assume that $f(x) = O(x^n)$ and $g(x) = O(x^m)$ as $x \rightarrow 0$. What can we say about $f(x) + g(x)$?

To answer this question we first observe that

$$\lim_{x \rightarrow 0} f(x) = 0 = \lim_{x \rightarrow 0} g(x).$$

It follows from the limit laws that

$$\lim_{x \rightarrow 0} f(x) + g(x) = 0$$

as well. But how quickly does the sum go to 0?

We can think of the Big-O symbols as representing the size of the error when approximating a function near 0. The triangle inequality tells us that the error in a sum

is at most the sum of the individual errors. But it is also the case that we cannot expect the error to be any smaller than the largest individual error. To make this more precise, we can find a $0 < \epsilon \leq 1$ and two constants M_1 and M_2 such that for all $x \in (-\epsilon, \epsilon) \subset [-1, 1]$, except possibly at $x = 0$, we have

$$|f(x)| \leq M_1|x^n|$$

and

$$|g(x)| \leq M_2|x^m|.$$

Therefore

$$|f(x) + g(x)| \leq M_1|x^n| + M_2|x^m|.$$

Next observe that the smaller the value of k , the more slowly x^k goes to 0. In fact, if

$$k = \min\{n, m\},$$

then for all $x \in [-1, 1]$ we have

$$|x^n| \leq |x^k| \quad \text{and} \quad |x^m| \leq |x^k|.$$

It follows that for all $x \in (-\epsilon, \epsilon)$, except possibly at $x = 0$, we have

$$|f(x) + g(x)| \leq M_1|x^n| + M_2|x^m| \leq M_1|x^k| + M_2|x^k| = (M_1 + M_2)|x^k|.$$

This shows that

$$f(x) + g(x) = O(x^k)$$

where

$$k = \min\{n, m\}.$$

In other words, *the potential error in a sum is at least as large as the error in either part*. We summarize this by writing

$$O(x^n) + O(x^m) = O(x^k)$$

where $k = \min\{n, m\}$.

The next theorem provides a summary of all of the arithmetic properties of Big-O notation. Aside for the second property, which we have just outlined above, the other properties follow almost immediately from the definition of Big-O. Therefore, the proofs are left as exercises.

THEOREM 5

Arithmetic of Big-O

Assume that $f(x) = O(x^n)$ and $g(x) = O(x^m)$ as $x \rightarrow 0$, for some $m, n \in \mathbb{N}$. Let $k \in \mathbb{N}$. Then we have the following:

- 1) $c(O(x^n)) = O(x^n)$. That is, $(cf)(x) = c \cdot f(x) = O(x^n)$.
- 2) $O(x^n) + O(x^m) = O(x^k)$, where $k = \min\{n, m\}$. That is, $f(x) \pm g(x) = O(x^k)$.
- 3) $O(x^n)O(x^m) = O(x^{n+m})$. That is, $f(x)g(x) = O(x^{n+m})$.
- 4) If $k \leq n$, then $f(x) = O(x^k)$.
- 5) If $k \leq n$, then $\frac{1}{x^k}O(x^n) = O(x^{n-k})$. That is, $\frac{f(x)}{x^k} = O(x^{n-k})$.
- 6) $f(u^k) = O(u^{kn})$. That is, we can simply substitute $x = u^k$.

EXAMPLE 10 Show that $f(x) = \cos(x^2) - 1 = -\frac{x^4}{2} + O(x^8)$. Use this result to evaluate

$$\lim_{x \rightarrow 0} \frac{\cos(x^2) - 1}{x^4}.$$

SOLUTION We begin by observing that if $g(u) = \cos(u)$, then since the third degree Taylor polynomial for g centered at $u = 0$ is

$$T_{3,0}(u) = 1 - \frac{u^2}{2}$$

the Taylor Approximation Theorem II gives us that

$$g(u) = 1 - \frac{u^2}{2} + O(u^4).$$

Arithmetic Rule (6) allows us to substitute x^2 for u to get

$$\cos(x^2) = g(x^2) = 1 - \frac{(x^2)^2}{2} + O((x^2)^4) = 1 - \frac{x^4}{2} + O(x^8).$$

Then

$$f(x) = g(x^2) - 1 = -\frac{x^4}{2} + O(x^8).$$

To evaluate

$$\lim_{x \rightarrow 0} \frac{\cos(x^2) - 1}{x^4}$$

we use the Arithmetic Rules to get

$$\begin{aligned} \lim_{x \rightarrow 0} \frac{\cos(x^2) - 1}{x^4} &= \lim_{x \rightarrow 0} \frac{-\frac{x^4}{2} + O(x^8)}{x^4} \\ &= \lim_{x \rightarrow 0} -\frac{1}{2} + O(x^4) \\ &= -\frac{1}{2} \end{aligned}$$



EXAMPLE 11 Evaluate

$$\lim_{x \rightarrow 0} \frac{x^2 \sin(x^2)(e^x - 1)}{(\cos(x) - 1)(\sin^2(x))(\sin(2x))}.$$

SOLUTION Using the arithmetic of Big-O, observe that $\sin(u) = u + O(u^3)$, and

so $\sin(x^2) = x^2 + O(x^6)$. Next, observe also that $e^x = 1 + x + O(x^2)$, and so $e^x - 1 = x + O(x^2)$. Then

$$\begin{aligned} x^2(e^x - 1)\sin(x^2) &= x^2(x + O(x^2))(x^2 + O(x^6)) \\ &= (x^3 + O(x^4))(x^2 + O(x^6)) = x^5 + O(x^9) + O(x^6) + O(x^{10}) \\ &= x^5 + O(x^6). \end{aligned}$$

Now $\cos(x) = 1 - \frac{x^2}{2} + O(x^4)$ and so $\cos(x) - 1 = \frac{-x^2}{2} + O(x^4)$; $\sin(u) = u + O(u^3)$ so $\sin(2x) = 2x + O(x^3)$; and $\sin^2(x) = (x + O(x^3))(x + O(x^3)) = x^2 + O(x^4) + O(x^4) + O(x^6) = x^2 + O(x^4)$. Then

$$\begin{aligned} (\cos(x) - 1)(\sin^2(x))(\sin(2x)) &= \left(\frac{-x^2}{2} + O(x^4)\right)(x^2 + O(x^4))(2x + O(x^3)) \\ &= \left(\frac{-x^4}{2} + O(x^6) + O(x^6) + O(x^8)\right)(2x + O(x^3)) \\ &= \left(\frac{-x^4}{2} + O(x^6)\right)(2x + O(x^3)) \\ &= -x^5 + O(x^7) + O(x^7) + O(x^9) \\ &= -x^5 + O(x^7). \end{aligned}$$

Substituting we get that

$$\begin{aligned} \lim_{x \rightarrow 0} \frac{x^2 \sin(x^2)(e^x - 1)}{(\cos(x) - 1)(\sin^2(x))(\sin(2x))} &= \lim_{x \rightarrow 0} \frac{x^5 + O(x^6)}{-x^5 + O(x^7)} \\ &= \lim_{x \rightarrow 0} \frac{1 + O(x)}{-1 + O(x^2)} \quad (\text{factoring out } x^5) \\ &= \lim_{x \rightarrow 0} \frac{1 + 0}{-1 + 0} \\ &= -1. \end{aligned}$$



8.3.1 Calculating Taylor Polynomials

Recall that the Taylor Approximation Theorem II tells us that if $f^{(n+1)}$ is continuous on a non-trivial closed interval containing $x = 0$, then

$$f(x) = T_{n,0}(x) + O(x^{n+1}).$$

In particular, if $f(x) = \cos(x^2) - 1$, then

$$\cos(x^2) - 1 = T_{7,0}(x) + O(x^8).$$

But in a previous example we also showed that

$$\cos(x^2) - 1 = -\frac{x^4}{2} + O(x^8)$$

leading us to ask: Is

$$T_{7,0}(x) = -\frac{x^4}{2}?$$

We could calculate the first seven derivatives of f to try and verify this result. However, this is a long and tedious calculation. Instead, it would be helpful if there was a converse to the Taylor Approximation Theorem II. That is, if p is a polynomial of degree n or less, and $f(x) = p(x) + O(x^{n+1})$, we would hope that $p(x) = T_{n,0}(x)$.

It turns out that if we can verify that $f^{(n+1)}$ is continuous on a non-trivial closed interval containing $x = 0$, and if p is a polynomial of degree n or less, such that $f(x) = p(x) + O(x^{n+1})$, then $p(x) = T_{n,0}(x)$. To prove that this is indeed the case, we begin with the following proposition:

PROPOSITION 6

If p is a polynomial with degree n or less (where $n \in \mathbb{N} \cup \{0\}$), and $p(x) = O(x^{n+1})$, then $p(x) = 0$ for all x .

PROOF

Let $Q(n)$ denote the statement:

If $p(x)$ is a polynomial with degree n or less, and $p(x) = O(x^{n+1})$, then $p(x) = 0$ identically.

We will proceed by induction to show that $Q(n)$ is true for all $n \in \mathbb{N} \cup \{0\}$.

First assume that $n = 0$. Hence $p(x) = c_0 = O(x)$ for some $c_0 \in \mathbb{R}$. We also know that since $p(x) = O(x)$ and $p(x)$ is continuous, that

$$c_0 = \lim_{x \rightarrow 0} f(x) = 0.$$

Hence $p(x) = 0$ identically, and $Q(0)$ holds.

Note: We could now proceed directly to our inductive step by beginning our induction at $n = 0$, but for clarity we will look at $Q(1)$ as well.

For $n = 1$, $p(x) = c_0 + c_1 x = O(x^2)$. This would again imply that

$$c_0 = \lim_{x \rightarrow 0} f(x) = 0$$

and hence that $p(x) = c_1x$.

Dividing $p(x) = c_1x$ by x , we obtain a new polynomial

$$q(x) := \frac{p(x)}{x} = c_1.$$

Moreover, since $p(x) = c_1x = O(x^2)$ we get that $q(x) = c_1 = O(x)$ by the Arithmetic Rules of Big-O. Now $q(x)$ is a polynomial of degree zero and $q(x) = O(x)$, so it follows as above that $q(x) = 0$ identically by the case $n = 0$. Hence $q(0) = c_1 = 0$ and so $p(x) = c_0 + c_1x = 0$ identically.

Suppose $Q(k)$ is true for some $k \geq 1$. Then let

$$p(x) := c_0 + c_1x + c_2x^2 + \cdots + c_kx^k + c_{k+1}x^{k+1}$$

be any polynomial with degree $k + 1$ or less and $p(x) = O(x^{k+2})$. Then we have once more that

$$c_0 = \lim_{x \rightarrow 0} p(x) = 0.$$

As above, we divide $p(x)$ by x to obtain,

$$q(x) := \frac{p(x)}{x} = c_1 + c_2x + c_3x^2 + \cdots + c_kx^{k-1} + c_{k+1}x^k,$$

and note that

$$q(x) = O(x^{k+2-1}) = O(x^{k+1}).$$

by arithmetic of Big-O. Now $q(x)$ is a polynomial of degree k or less and is Big-O of x^{k+1} , therefore by the inductive hypothesis, $q(x) = 0$ identically. Hence $p(x) = xq(x) = 0$ for all $x \neq 0$ and $p(0) = 0$ together prove $Q(k + 1)$.

By the Principle of Mathematical Induction, $Q(n)$ is true for all $n \in \mathbb{N} \cup \{0\}$.

■

THEOREM 7

Characterization of Taylor Polynomials

Assume that $r > 0$. Assume also that f is $(n + 1)$ -times differentiable on $[-r, r]$ and $f^{(n+1)}$ is continuous on $[-r, r]$. If p is a polynomial of degree n or less with

$$f(x) = p(x) + O(x^{n+1}),$$

then $p(x) = T_{n,0}(x)$.

PROOF

First note that by assumption

$$f(x) - p(x) = O(x^{n+1}).$$

We also know that the Taylor Approximation Theorem II guarantees that

$$f(x) - T_{n,0}(x) = O(x^{n+1}).$$

Using Big-O arithmetic, we have that

$$\begin{aligned} h(x) &= p(x) - T_{n,0}(x) \\ &= [f(x) - T_{n,0}(x)] - [f(x) - p(x)] \\ &= O(x^{n+1}) + O(x^{n+1}) \\ &= O(x^{n+1}). \end{aligned}$$

But h is a polynomial of degree n or less with $h(x) = O(x^{n+1})$, so it follows from the previous observation that

$$0 = h(x) = p(x) - T_{n,0}(x).$$

Therefore $p(x) = T_{n,0}(x)$. ■

EXAMPLE 12 We have previously shown that if $f(x) = \cos(x^2) - 1$, then

$$f(x) = -\frac{x^4}{2} + O(x^8).$$

This means that

$$T_{7,0}(x) = -\frac{x^4}{2}.$$



EXAMPLE 13 Let $f(x) = x^2(e^x - 1) \sin(x^2)$. Find $T_{5,0}(x)$ and in particular, find $f^{(4)}(0)$ and $f^{(5)}(0)$.

SOLUTION Previously we would have been discouraged by the prospect of finding the fourth and fifth derivatives of f at $x = 0$. Using Big-O arithmetic makes this problem easy.

Recall that we previously calculated $\sin(u) = u + O(u^3)$, and so $\sin(x^2) = x^2 + O(x^6)$. Observe also that $e^x = 1 + x + O(x^2)$, and so $e^x - 1 = x + O(x^2)$. Then

$$\begin{aligned} f(x) &= x^2(e^x - 1) \sin(x^2) = x^2(x + O(x^2))(x^2 + O(x^6)) \\ &= (x^3 + O(x^4))(x^2 + O(x^6)) \\ &= x^5 + O(x^9) + O(x^6) + O(x^{10}) \\ &= x^5 + O(x^6). \end{aligned}$$

The Characterization of Taylor Polynomials Theorem tells us that $x^5 = T_{5,0}(x)$. Since

$$T_{5,0}(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \frac{f^{(4)}(0)}{4!}x^4 + \frac{f^{(5)}(0)}{5!}x^5,$$

by matching coefficients, we get that $0 = \frac{f^{(4)}(0)}{4!}$ and $1 = \frac{f^{(5)}(0)}{5!}$. It follows that $f^{(4)}(0) = 0$ and $f^{(5)}(0) = 5! = 120$. 